A framework for functional alignment applications in cognitive neuroscience

ELIZABETH DUPRE
Integrated Program in Neuroscience
McGill University, Montréal, QC, Canada
April 2022

A thesis submitted to McGill University in partial fulfillment
of the requirements of the degree of
Doctor of Philosophy

Abstract

Significant inter-individual variability has been noted at every scale of brain organization, from cellular features to large-scale anatomy to functional patterning. This variability presents fundamental challenges to identifying shared neural principles supporting complex cognition and behavior. With the advent of functional Magnetic Resonance Imaging (fMRI), researchers have gained unprecedented access to data from healthy participants engaged in a range of experimental tasks. However, this volume of data has only highlighted the challenge of inter-individual comparisons. This thesis explores a class of recently developed methods that aim to address this challenge through "functional alignment;" that is, directly aligning individual brains on functional features. In considering these methods, we aim to develop guidelines of direct relevance for cognitive neuroscience researchers. Chapters 1 and 2 provide a brief introduction to the challenge of inter-subject comparisons in neuroimaging research and lay out the primary objectives of the work presented in this thesis. Chapter 3 benchmarks five available functional alignment algorithms via inter-subject decoding performance across four open-access datasets. The presented results highlight the potential of functional alignment to improve inter-individual comparisons, while cautioning that algorithm choice may significantly impact performance. Chapter 4 explores additional experimental factors that impact functional alignment performance: namely, selected data to derive and to which to apply the alignment. We examine both of these dimensions across two publicly available datasets with extensive characterizations of individual subjects, presenting results for three unique performance metrics. Chapter 5 synthesizes these results into practical recommendations for cognitive neuroscience researchers who hope to use functional alignment in their own research questions. We also include interactive, online materials supporting these recommendations, leveraging recent developments in open publishing infrastructure. Chapter 6 then reviews the current landscape of open publishing infrastructure, highlighting its importance in developing future scientific objects such as those included in Chapter 5. We address ongoing infrastructure

work as well as potential areas for development, several of which have been advanced through this thesis. Finally, Chapter 7 closes with a discussion of the overall contributions of this thesis and highlights avenues for future work to develop functional alignment methods. Throughout this thesis, the presented work underscores the interaction between domain research and methods development.

Résumé en français

Une importante variabilité interindividuelle a été constatée à toutes les échelles de l'organisation cérébrale, à partir des caractéristiques cellulaires jusqu'à l'anatomie à grande échelle et au schéma fonctionnel. Cette variabilité présente des défis fondamentaux pour l'identification des principes neuronaux partagés par tous les individus et dont émergent la cognition et les comportements complexes. Avec l'avènement de l'imagerie par résonance magnétique fonctionnelle (IRMf), les chercheurs ont obtenu un accès sans précédent à des données provenant de participants sains engagés dans une série de tâches expérimentales. Cependant, ce volume de données n'a fait que souligner le défi des comparaisons interindividuelles. Cette thèse explore une classe de méthodes récemment développées qui visent à relever ce défi par le biais de "l'alignement fonctionnel," c'est-à-dire l'alignement direct des cerveaux individuels grâce à leurs caractéristiques fonctionnelles. En examinant ces méthodes, nous visons à développer des lignes directrices directement pertinentes pour les chercheurs en neurosciences cognitives. Les chapitres 1 et 2 présente une brève introduction au défi des comparaisons inter-sujets dans la recherche en neuroimagerie et exposent les objectifs principaux du travail présenté dans cette thèse. Le chapitre 3 compare la performance de décodage inter-sujets de cinq algorithmes d'alignement fonctionnel sur quatre ensembles de données ouvertes. Les résultats présentés mettent en évidence le potentiel de l'alignement fonctionnel pour améliorer les comparaisons inter-individuelles, tout en soulignant que le choix de l'algorithme peut avoir un impact significatif sur les performances. Le chapitre 4 explore d'autres facteurs expérimentaux qui ont un impact sur la performance de l'alignement fonctionnel, notamment les données sélectionnées pour dériver et celles auxquelles il est appliqué. Nous examinons ces deux dimensions à travers deux jeux de données publics, qui fournissent des caractérisations étendues des sujets individuels, présentant des résultats pour trois mesures de performance. Le chapitre 5 synthétise ces résultats sous forme de recommandations pratiques à l'intention

des chercheurs en neurosciences cognitives qui espèrent utiliser l'alignement fonctionnel pour leurs propres questions de recherche. Nous incluons également des documents interactifs en ligne à l'appui de ces recommandations, en tirant parti des récents développements en matière d'infrastructure de publication ouverte. Le chapitre 6 offre un aperçu de l'infrastructure de publications ouvertes, en soulignant son importance dans le développement de futurs objets scientifiques tels que ceux inclus dans le chapitre 5. Nous abordons les travaux d'infrastructure en cours ainsi que les domaines potentiels de développement, dont plusieurs ont été avancés dans le cadre de cette thèse. Enfin, le chapitre 7 se termine par une discussion sur les contributions globales de cette thèse et met en évidence les pistes de travail futures pour développer des méthodes d'alignement fonctionnel. Les travaux présentés par cette thèse soulignent l'interaction entre la recherche sur le domaine et le développement de méthodes.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Richard H Tomlinson, the Canadian Open Neuroscience Platform, Healthy Brains for Healthy Lives, and Unifying Neuroscience and Artificial Intelligence - Québec (UNIQUE) for their funding during my PhD.

I have been lucky to have the expert guidance of my supervisor Jean-Baptiste Poline as well as my collaborators Bertrand Thirion and Pierre Bellec, who together shepherded this work through sun and storms.

It's hard to imagine how this research would have happened without the involvement of Thomas Bazeille; indeed, I'd rather not imagine it. Thank you for being the best collaborator and friend one could ask for.

I am forever grateful for my colleagues from the ORIGAMI lab as well as the BrainHack and Organization for Human Brain Mapping (OHBM) communities, each of whom helped me to grow personally and professionally throughout my degree. Thank you in particular to Cameron Craddock, Chris Gorgolewski, Kirstie Whitaker, and Tal Yarkoni for inspiring me and so many others.

I've been lucky to work directly and indirectly with wonderful friends and colleagues throughout my dissertation. Thank you for both your encouragement and enthusiasm—as well as your patience and grace—in coming on this journey with me. Although it would be impossible to list everyone whose help I've benefited from, a few names stand out: Valérie Hayot-Sasson, Peer Herholz, Alexandre Hutton, Agah Karakuzu, Greg Kiar, Laetitia Mwilambwe-Tshilobo, Taylor Salo, Jessica Thompson, and Jake Vogel. Thank you all.

Finally, thank you to Ross: without your love and support, none of this would have been possible.

"They say you are not you except in terms of relation to other people."

- *All the King's Men*, Robert Penn Warren

Contribution of authors

As lead author on the manuscripts presented in chapters 3 through 6, I was responsible for all aspects of the research including: conceptualization, experimental design, data curation and processing, statistical analysis, visualization, as well as writing and revising the associated manuscripts.

Contributions of co-authors for these chapters are outlined below:

Chapter 3

- Thomas Bazeille: conceptualization, experimental design, data curation and processing, statistical analysis, visualization, as well as writing and editing of the manuscript
- Hugo Richard: methodology development, editing of the manuscript
- Jean-Baptiste Poline: conceptualization, editing of the manuscript
- Bertrand Thirion: conceptualization, editing of the manuscript

Chapter 4

- Thomas Bazeille: conceptualization, experimental design, data curation and processing, statistical analysis, visualization, editing of the manuscript
- Bertrand Thirion: conceptualization, editing of the manuscript
- Jean-Baptiste Poline: conceptualization, editing of the manuscript

Chapter 5

- Jean-Baptiste Poline: conceptualization, editing of the manuscript

Chapter 6

- Chris Holdgraf: editing of the manuscript

- Agah Karakuzu: editing of the manuscript
- Loïc Tetrel: editing of the manuscript
- Pierre Bellec: editing of the manuscript
- Nikola Stikov: editing of the manuscript
- Jean-Baptiste Poline: conceptualization, editing of the manuscript

Chapter 1

Introduction

1.1 General context

Individual differences dominate our lived experience, with each person showing a unique constellation of physical, cognitive, and behavioral traits. Cognitive neuroscience seeks the relationship between 'the brain' and 'the mind' (Churchland and Sejnowski, 1988), but the sheer diversity of individual brains and minds complicates this goal. While neuroscientists have adopted many strategies to overcome this challenge, a promising addition is 'functional alignment,' which compares individual brains directly on their functional activity rather than the underlying anatomy, the latter of which differs at the cellular-, circuit-, and macro-levels. Functional alignment is not a single method, however, but a broad class of algorithms each with unique assumptions.

The complexity of associated constraints means that these methods remain relatively inaccessible, even to researchers themselves. The aim of this thesis is to provide a general framework for understanding and evaluating functional alignment applications in cognitive neuroscience. To do so, I consider both the algorithms themselves as well as the experimental choices that impact alignment performance. I use these results to create guidelines for researchers and develop on existing publishing infrastructure to communicate these guidelines in accessible formats.

1.2 Objectives

Chapter 3 objectives

Chapter 3 presents a benchmarking analysis of five functional alignment algorithms across four open fMRI datasets. Using inter-subject decoding as an index of alignment performance, I compare the relative increase in decoding accuracy for each considered functional alignment algorithm against both within-subject decoding as well as anatomical alignment baselines. I generate estimates across the whole brain and within task-relevant regions-of-interest, providing researchers with general guidelines for selecting among available functional alignment algorithms in their own work.

Chapter 4 objectives

Chapter 4 of this thesis examines the impact of experimental factors on functional alignment performance. Beyond algorithm choice, alignment performance can be significantly impacted by the data on which alignment transformations are learnt and applied. Using two well-sampled, open access fMRI datasets, I compare the influence of training and application data across three performance metrics in use in the literature. I evaluate how these effects differ between task-relevant and task-irrelevant brain regions, underscoring the complexity of appropriately deploying these methods in cognitive neuroscience questions.

Chapter 5 objectives

Chapter 5 synthesizes current literature and results from Chapters 3 and 4 to familiarize cognitive neuroscience researchers with functional alignment and outline general recommendations for appropriate experimental design in adopting these methods. I additionally provide an online, executable series of tutorials to help researchers directly integrate these methods into their existing workflows.

Chapter 6 objectives

Chapter 6 of this thesis overviews the current publishing landscape for executable research objects, such as the tutorials included in Chapter 5. Appropriate usage of multivariate methods in cognitive neuroscience directly benefits from the availability of executable research objects, though these are difficult to publish and so disincentivized in the current

publishing landscape. I highlight current gaps in the supporting infrastructure and provide suggestions for future development. I draw on my experience in developing open publishing workflows to center the social and technical solutions likely to support future publishing platforms for executable research objects in neuroscience.

1.3 Contributions to original knowledge

Collectively, the work included in this thesis provides a general framework for functional alignment adoption across cognitive neuroscience and suggests future technical work to increase the accessibility of these and other multivariate methods. In doing so, the presented research explores how methodological choices and research context interact in characterizing individual variability of functional organization. Below, I briefly highlight the distinct conclusions and contributions of each included chapter.

Chapter 3

- Functional alignment recovers approximately half of the individual variability lost in anatomical alignment
- Piecewise aggregation schemes outperform the popular searchlight aggregation scheme
- Piecewise Shared Response Model is the best performing of all considered methods, although it does require an additional hyperparameter search

Chapter 4

- Stimuli used to derive and to apply alignment transformations both impact performance, with cross-modal (i.e., audio-only versus visual-only) pairings showing weakest results
- Functional alignment improves inter-individual similarity in regions with little task-relevant signal, but this effect is balanced by reduced similarity in highly task-relevant areas
- Functional alignment cannot be assumed to improve inter-subject similarity for a given application; instead, this effect must be evaluated with independent testing data

Chapter 5

- Functional alignment offers unique insight compared to other methods for improving inter-individual similarity such as individualized parcellations
- Researchers using functional alignment should ensure that data used to derive alignment transformations are kept separate from data on which alignment is applied
- Many kinds of stimuli can be used in learning alignment transformations, though these should be selected in conjunction with the desired experimental task

Chapter 6

- Publishing infrastructure is consolidated around static formats such as the PDF, with key barriers in archiving and reviewing other research objects
- Developing open standards for executable research objects is necessary to move forward future publishing formats
- Initiatives such as NeuroLibre provide a potential blueprint for new platforms to leverage composable infrastructure

Chapter 2

Review of the literature

Although neuroscience case studies such as Phineas Gage (Teles, 2020) and patient HM (Squire, 2009) have yielded foundational insights, the field remains fundamentally focussed on general organizational principles supporting neural activity and behavior. Uncovering these principles relies on making appropriate comparisons across individuals; however, this presents a core technical and conceptual challenge. Human brains differ along nearly every dimension: from cytoarchitectonic features (Amunts et al., 1999; Rademacher et al., 1993), to large-scale sulco-gyral anatomy (Galaburda et al., 1990; Marie et al., 2015), to functional organization (Benson et al., 2021; Frost and Goebel, 2012; Gordon et al., 2017a), to functional response profiles (Henriksen et al., 2012).

With the advent of Positron Emission Tomography (PET) imaging, researchers could non-invasively collect measurements from non-clinical populations, dramatically expanding our ability to characterize neural organization (Portnow et al., 2013). However, the need to minimize an individual participant's radiation exposure (Fedorenko, 2021) and to overcome the low signal-to-noise ratio of PET images (Crivello et al., 2009) necessitated that researchers focus on deriving group-level maps. The emergence of functional magnetic resonance imaging (fMRI) dramatically increased available temporal and spatial resolution—while removing radiation-exposure risks—however, the field remained largely focussed on group-level results (Raichle, 2009). In the early days of fMRI, this may have been driven by the large overlap in scientists with PET and fMRI expertise. Moreover, appropriate statistical methods for analyzing these data were jointly developed and distributed (Ashburner, 2012) encouraging harmonized workflows.

As fMRI grew to be a predominant modality for human neuroimaging (Poldrack and Farah, 2015), the effects of sample size in identifying brain-behavior relationships

came under greater scrutiny, with relationships derived from small sample sizes showing limited reproducibility (Button et al., 2013; Marek et al., 2020). Further, efforts such as MyConnectome (Poldrack et al., 2015) highlighted how the relatively small amount of participant-level data in traditional group studies prevented systematic investigation into dynamic cycles in sleep or mood, alongside other idiosyncratic factors. Researchers thus began to turn to extensive characterizations of a small number of individuals, sidestepping statistical power issues in group-level inference and opening up new scientific questions. As a result, recent years have seen a dramatic increase in the number of fMRI studies that "deeply" or "densely" phenotype individual participants (Naselaris et al., 2021; Poldrack, 2017). Thanks to open sharing initiatives such as OpenNeuro (Markiewicz et al., 2021; Poldrack et al., 2013), many of these "deep phenotyping" fMRI datasets are publicly available for re-analysis, including Courtois-NeuroMod (Boyle et al., 2020), the Individual Brain Charting initiative (Pinho et al., 2018), the Midnight Scan Club (Gordon et al., 2017b), and Study Forrest (Hanke et al., 2014), among others. These datasets promise significant insight in describing fine-scale functional organization; however, there is little consensus on how to use these resources to make appropriate comparisons across individuals while overcoming the technical and conceptual challenges noted above.

Here, I review how fMRI is currently used in cognitive neuroscience to investigate the functional architecture of the human brain as well as relevant ideas from cognitive science to compare across individual functional organizations. I first describe standard fMRI preprocessing pipelines and their methodological assumptions before reviewing the idea of functional representations as putatively captured by fMRI. I then introduce the idea of functional alignment, exploring its roots in cognitive science and particularly the connectionist philosophy of science. Finally, I consider current applications of and open challenges for functional alignment within cognitive neuroscience, which this thesis aims to address.

2.1 Characterizing fMRI measurements

Whereas PET imaging measures the neural uptake of tagged tracers such as radiolabeled-glucose (e.g. FDG-PET; Berti et al., 2014), fMRI measures the relative concentration of deoxyhemoglobin via its paramagnetic properties. This index of the magnetic susceptibility of circulating blood is known as the blood-oxygen level dependent (BOLD) signal (Greve, 2011) and provides an estimate of the uptake of oxygen in imaged tissue. Although the BOLD signal avoids invasive introduction of contrast agents, it provides a measure

of cerebrovascular—rather than directly neuronal—activity. Thus, it is commonly confounded with noise sources, both those shared with other neuroimaging modalities and those unique to cerebrovascular measures. For example, neuroimaging modalities such as (f)MRI, PET, and even calcium imaging (Robbins et al., 2021) all share a susceptibility to tissue motion, due to their fixed acquisition windows and relatively long sampling rates (Dubbs et al., 2016; Mukherjee et al., 2016). In the case of human brain imaging, participant motion can arise from both voluntary processes such as subject discomfort as well as involuntary, physiological motions such as respiratory oscillations and cardiac pulsations (Power et al., 2018). Additionally, as a measure of cerebrovascular activity, fMRI is uniquely sensitive to fluctuations in blood pressure (Wang et al., 2006) and cerebrovascular reactivity (Pinto et al., 2020). These latter properties position fMRI as a relative index of cerebrovascular activity, rather than a quantitative measure, meaning that absolute measured signal differs from session-to-session even within the same subject.

As a result of these diverse noise sources, many denoising methods exist for fMRI data, most of which I consider to be beyond the scope of the present review (e.g. multi-echo denoising; DuPre et al., 2021). Nonetheless, I describe these noise sources to highlight the unique properties of the fMRI signal which further challenge direct comparisons in measured activation patterns across individuals. To date, standard preprocessing pipelines have been developed for cleaning and comparing fMRI signals (e.g., fMRIPrep; Esteban et al., 2019), although the exact algorithms implemented at each step differ significantly across research groups (Carp, 2012; Li et al., 2021). Here, I review three stages of the standard preprocessing pipeline with strong relevance for inter-subject comparison: (1) anatomical-functional co-registration, (2) normalization to a reference template, and (3) Gaussian smoothing.

2.1.1 Anatomical-functional co-registration

In order to make comparisons across participant functional patterns, an initial co-registration between individual anatomical and functional images is necessary (c.f., Dohmatob et al., 2018). This co-registration allows for direct mapping between locations in each anatomical and functional image, though initial functional images may have significant distortions due to their unique acquisition parameters. For example, magnetic field inhomogeneity from the air-tissue interface in the sinuses commonly causes significant dropout of BOLD signal in regions including the orbitofrontal and anterior temporal cortices with standard echo-planar imaging (EPI) sequences, particularly when collected with anterior-posterior phase encoding. Thus, an initial susceptibility distortion correction (Hutton et al., 2002)

may be performed before co-registration to correct some of this distortion, though this procedure cannot create signal in areas with severe dropout artifacts. I note this only to emphasize the immediate challenges in making comparisons between function and anatomy.

Optionally following distortion correction, multiple algorithms exist to co-register functional and anatomical images. One of the most well-known is boundary-based registration (Greve and Fischl, 2009), as implemented in the widely-used software packages FSL (Jenkinson et al., 2012) and FreeSurfer (Fischl and Dale, 2000). This algorithm emphasizes the cross-modal nature of the alignment, delineating tissue-boundaries in a high-resolution anatomical reference image and then using these boundaries to drive subsequent alignment with a lower-resolution functional input image. Other linear algorithms such as local Pearson correlation (Saad et al., 2009) as implemented in the AFNI platform (Cox, 1996) similarly emphasize tissue boundaries in driving cross-modal alignment. Additional, non-linear approaches such as the symmetric diffeomorphic transformation model SyN (Avants et al., 2008) have recently seen an increase in popularity for situations where gold-standard estimates of anatomical-functional misregistration—such as fieldmaps—are not available. These non-linear approaches, however, can be especially sensitive to initial image masking (Huntenberg, 2014) and require careful visual inspection to ensure that functional images are not inappropriately warped to the reference anatomy. Thus, some degree of misalignments may persist across the two modalities, even within a single participant (Dukart and Bertolino, 2014). Even assuming a successful co-registration, however, additional processing is necessary to extend inferences to across participants.

2.1.2 Normalization to a reference template

Following anatomical-functional co-registration, participant anatomical images are commonly normalized to a reference template in a defined stereotaxic space. Two primary challenges exist when normalizing participant images to a reference anatomy: the definition of a standard template and the process by which images are normalized to it (Brett et al., 2002). Initial fMRI studies normalized acquired images to the "Talaraich brain" (Talairach and Tournoux, 1988), which was defined on only a single hemisphere of the post-mortem, ex vivo brain of a 60-year-old French woman. This allowed for the inclusion of additional histological information on Brodmann's cytoarchitectonic areas (Strotzer, 2009). However, a high degree of uncertainty arose in using Talaraich space as a stereotaxic reference due to its definition on a single subject. Thus, subsequent work to define reference templates primarily focussed on average anatomy across many individuals;

for example, the MNI152 (Mazziotta et al., 2001) which was defined on 152 participants. In those cases where participants have variable anatomical features (e.g. duplication of Heschel's gyrus; Marie et al., 2015) an ideal correspondence between individual anatomies may not exist, requiring a defined reference to only model a subset of the population or to blur the included anatomies into a single consensus image. Despite this inherent limitation, the definition of standard templates has significantly facilitated inter-subject comparisons with downstream impacts on knowledge aggregation across the field, enabling efforts such as meta-analysis (Wager et al., 2007) and broad adoption of standard parcellations (Lawrence et al., 2021).

To date, a wide variety of reference templates are in use for unique developmental (Fonov et al., 2011) or geographic populations (Liang et al., 2015) and supported by atlas sharing infrastructure such as TemplateFlow (Ciric et al., 2021). Two of the most commonly used are the MNI152 templates (for volumetric representations of anatomy) and the fsaverage templates (for surface-based representations of anatomy; Fischl et al., 1999b). Different registration algorithms are generally favored for these differing representations of anatomical representations. While volumetric normalization can be achieved using either affine (i.e., linear) or non-linear registration, surface-based normalization heavily relies on non-linear normalizationto project individual cortical meshes to a spherical representation (Fischl et al., 1999a). These individual cortical meshes can then be aligned using one or more reference features (e.g., both sulco-gyral patterning and functional activations as in MSM-all; Glasser et al., 2016). In general, these different normalization procedures allow for different relative weightings of relevant features that are expected to correspond across subjects, with significant consequences for downstream analysis workflows (Coalson et al., 2018). Wide variability in individual cortical anatomy and functional organization, however, mean that there is no registration algorithm which can perfectly map between all participants across all features of interest. Thus, preprocessing pipelines commonly include additional Gaussian smoothing to smooth over any remaining misalignments.

2.1.3 Gaussian smoothing

Many researchers working with fMRI data for the first time are confused why smoothing is included in standard preprocessing pipelines. Gaussian smoothing is implemented as an isotropic filter that indiscriminately blurs spatial information, resulting in a loss of fine detail. This loss, however, is often outweighed by the benefits that smoothing brings to fMRI analysis. I have already alluded to the first of these, which is to blur over remaining inter-subject variability following normalization. As this variability is on the

order of millimeters (Tahmasebi et al., 2009), applied smoothing kernels typically have a full-width-half-maximum (FWHM) of at least one voxel, if not more. Although less immediately relevant for inter-subject comparisons, Gaussian smoothing brings at least two important, additional benefits to fMRI analysis, which I briefly review below.

The first benefit is to improve statistical inference by reducing the number of resolution elements (or "resels") that must be accounted for when correcting for multiple comparisons. This is particularly important in the framework of Random Field Theory (RFT; Worsley et al., 1996) which assumes that data are reasonably represented by a smooth Gaussian field. While fMRI data does have an inherent spatial smoothness, the exact value is dependent on scanning acquisition parameters, with higher-resolution sequences generally showing lower intrinsic smoothness (Bollmann and Barth, 2020). Thus, smoothing fMRI data before statistical analysis ensures that data can be appropriately considered with a smooth Gaussian representation. Today, RFT is only one of many multiple comparison correction frameworks for fMRI data (Lindquist and Mejia, 2015), although it remains in active use across the community.

The second additional benefit of Gaussian smoothing is to increase statistical power both for individual voxels as well as for detection of supravoxel signals. These effects can be explained by two main factors. First, given the intrinsic spatial smoothness of fMRI, nearby voxels often share some overlapping signal. Smoothing increases the correlation between nearby voxels, thereby strengthening this shared signal while reducing the effects of asynchronous signals that might arise from partial volume effects (Dukart and Bertolino, 2014) or unshared noise sources. Smoothing also increases the signal-to-noise ratio for supravoxel signals of the same effective resolution as the applied smoothing kernel, as explained by the matched filter theorem (Rosenfeld, 2014). For example, cognitive neuroscience experiments commonly adopt an 8mm FWHM Gaussian kernel to capture distributed spatial activations such as those evoked during a visual oddball task (Mikl et al., 2008).

Despite these benefits, analytic frameworks have arisen that broadly reject smoothing to maintain the highest spatial resolution possible. Perhaps the most influential of these is Multi-Voxel Pattern Analysis (MVPA; Norman et al., 2006), which encourages researchers to directly leverage distributed voxelwise activation patterns for decoding participant mental states. The popularization of this approach accelerated a move away from large Gaussian kernel sizes in the hopes of uncovering fine-scaled functional organization.

2.2 Assessing individual correspondence in the analysis of fMRI data

Multi-voxel pattern analysis (MVPA) emerged as an alternative to univariate analyses—such as the general linear model—that separately model the relationship of each voxel to an experimental feature of interest. While univariate analyses continue to advance the study of brain organization, MVPA offers unique insights by examining how information is represented in the across voxelwise activations rather than solely within individual voxels (Weaverdyck et al., 2020). Many MVPA researchers have thus advised against Gaussian smoothing in order to maintain as much voxelwise information as possible (c.f., Beeck, 2010). As hundreds of thousands of voxels can be obtained in a single scanning session, cognitive neuroscientists have turned to increasingly sophisticated analytic techniques to explore these relationships.

Since its introduction in Haxby et al. 2001, encoding and decoding models (Naselaris et al., 2011) and pattern classification (Kragel et al., 2012), have emerged as predominant forms of MVPA in the field, with many successful applications (Haxby et al., 2014). This relative success has been suggested to reveal principles of neural coding (Guest and Love, 2017); i.e., the mechanism by which information is represented in neural activity (deCharms and Zador, 2000; c.f. Brette, 2018). Importantly, however, MVPA is typically conducted within individual subjects with summary information (e.g., individual classifier accuracy) carried forward for group-level comparisons. Thus, there is no direct comparison of voxelwise patterns across individuals, as variable correspondence between structural and functional features (Paquola et al., 2019a; Vázquez-Rodríguez et al., 2019) yields consistently poor performance with group-level MVPA models (Bilenko et al., 2010). Here, I motivate MVPA by briefly introducing the idea of functional representations, a principle of neural coding that has driven significant research within the field.

2.2.1 Distributed functional representations organize neural activity

Neuroscience has broadly adopted the term ”representation“ to describe any systematic patterns in the relationship between features of the world and neural activity (Poldrack, 2020). The widely observed organization of large portions of primary cortices to mirror their interacting sensory systems provides strong support for the idea of representations. For example, the retinotopic organization of primary visual cortex (Hubel and Wiesel,

1959), the tonotopic organization of auditory cortex (Humphries et al., 2010), or the somatotopic organization of primary motor cortex are all well-known principles of neural coding. Representations in successive cortical areas have been suggested to correspond to successive layers of abstraction (Eickenberg et al., 2017), motivating investigation into higher-order representations of complex stimuli such as faces (Jiahui et al., 2020). Within this broad organization, neurons are further organized into columnar structures which interact to create areal population responses (Panzeri et al., 2015). The relatively low spatial resolution of standard 3T fMRI sequences—commonly on the order of 2mm isotropic voxels—means that columnar information—on the order of 600 μm in diameter (Mountcastle, 1997)—is inaccessible. Instead, voxels irregularly sample underlying neuronal population activity through the complex spatiotemporal filter of the hemodynamic response (Kriegeskorte et al., 2010). Further, both noise correlations and tuning heterogeneity (i.e. those stimulus features that neurons respond most strongly to) in the underlying neuronal populations significantly influence what information that can be recovered from voxelwise activations (Zhang et al., 2020).

The relative success of MVPA applications, however, suggest that multivariate voxelwise information provides a useful, if coarse, measure of population-level activity supporting functional representations. Nonetheless, both the irregular sampling of cortical columns as well as the heterogeneity of their supporting neuronal populations throw the challenge of identifying inter-individual correspondence into stark relief. That is, we cannot assume that a given voxel provides information on the same neuronal population across individuals, even when located at identical coordinates in standard space. While their close spatial proximity means that they are likely to provide similar information, their unique sampling means that they require additional attention to align across participants. To directly leverage voxelwise response patterns across individuals, then, requires a method by which to compare distributed functional representations.

2.3 Origins of functional alignment in connectionism

The challenge of comparing distributed functional representations loomed in the late twentieth century as cognitive scientists sought a new approach to artificial intelligence. Connectionism was emerging as a promising framework to understand both the brain as well as the artificial neural networks modelled after it (Rumelhart et al., 1988). Classical models of artificial intelligence, by contrast, argued that cognition could be understood as operating over symbols, similar to natural language. Neural activity, then, was assumed to

involve tokenized representations of these symbols (Newell, 1980). Although this classical approach had dominated early discussions of artificial intelligence, it struggled to translate into successful applications. Connectionism offered an alternative approach with no direct reliance on symbolic representations. Instead, computation was assumed to be carried out on distributed activations across inter-connected "units;" i.e., neurons. This approach showed several initial successes (cf. single layer perceptrons; Minsky and Papert, 1988), providing the foundation for today's deep learning algorithms (Buckner and Garson, 2019).

Nonetheless, philosophers of science disagreed—then and now—as to whether connectionism could provide a meaningful framework for understanding the brain itself (Fodor and Pylyshyn, 1988). One strong challenge came from Jerry Fodor and Ernest Lepore, who argued that no "useful sense of notions of conceptual identity, or even conceptual similarity, [could be found] in the face of the enormous functional and structural diversity across individual networks that constitute human brains" (Churchland, 1998). That is, given the remarkable inter-individual variability in brain organization, it would be difficult if not impossible to find meaningful correspondence in the computation carried out across these different architectures, even at a single spatial scale. This challenge strongly overlaps with the difficulties I have described above in finding correspondence across activations from irregular sampling of individual brains.

To overcome this challenge, Paul Churchland advanced a specific view of connectionism which he coined "state-space semantics," wherein representations can be viewed as corresponding to specific points in activation space (Laakso and Cottrell, 2005). Figure 2.1, from Churchland (1998), demonstrates this idea: an identical activation pattern across three units can be viewed as either a histogram or distribution of activation values, a vector of length three, or a point in a three-dimensional space—with each dimension corresponding to a single unit. Within this framework, Churchland argued that correspondence could be identified by mapping between two or more defined activation spaces based on their labelled activations. Thus, although the specific arrangement of neurons across two brains may differ, the relative organization of content within them would likely be preserved; e.g., activations evoked by 'cat' and 'dog' stimuli would be closer to one another than to activation evoked by 'car.' Aligning functional spaces on these labelled representations, which we call "functional alignment," would allow for direct inferences across activation spaces.

While state-space semantics offered one solution to the Fodor-Lepore challenge, it came with its own criticisms. Laakso and Cottrell (2000) pointed out a fundamental "problem with Churchland's strict identification of content with a specific position in state space. It

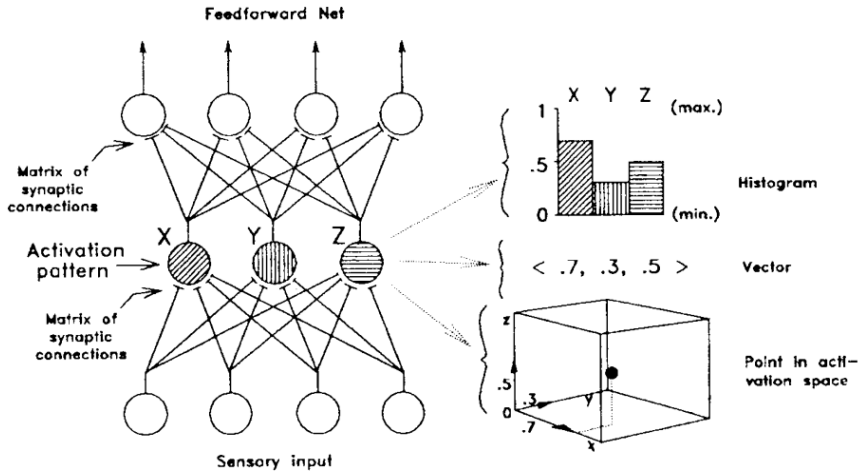

Figure 2.1: **Three ways of conceiving the activity of a specific neuronal population.** Neuronal population activity can be represented using many different formats. Here, we show a single activity pattern as a histogram of activation levels, as an activation vector, and as a point in an activation space. Figure reproduced from Churchland (1998)

.

is well known that networks with different numbers of hidden units can solve the same problem." Indeed, the exact number of neurons varies significantly across brains (Neves et al., 2020), and it is unclear how one could identify a direct mapping between activation spaces with a variable number of dimensions. Laasko and Cottrell thus proposed that comparisons between these activation spaces should not be direct, but instead take place on second-order isomorphisms (Shepard and Chipman, 1970). This is the approach adopted by Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008), where correlations between evoked activation patterns are compared rather than the activation patterns themselves. RSA has seen widespread adoption across cognitive and computational neuroscience (Kriegeskorte and Diedrichsen, 2019) and generated significant methodological development to improve measurement of second-order isomorphisms beyond simple correlations (e.g., Williams et al., 2021).

Despite Laakso and Cottrell's criticism, functional alignment remains a viable analytic approach even in networks with variable numbers of neurons. While Churchland argued for direct comparison using the smallest number of dimensions (Churchland, 1998), neuroscientists more commonly adopt techniques such as dimensionality reduction or manifold learning to learn shared subspace with desired statistical properties (Chen et al., 2015; Dabagia et al., n.d.). In the case of fMRI data, standard normalization and resampling procedures constrain each participant's measurements to the same number of voxels, with an approximate anatomical correspondence. This implicit mapping encouraged initial work

to adopt methods such as Procrustes analysis which assume a shared dimensionality when functionally aligning fMRI data (Haxby et al., 2011). Although these methods follow in the tradition of Churchland's state-space semantics, they were largely rediscovered by Haxby and colleagues from the "representational spaces" of RSA (Haxby et al., 2014). In the decade since its introduction to cognitive neuroscience, functional alignment has seen significant development—in both the range of algorithms and relevant applications—which I briefly review below.

2.4 Applications of functional alignment in cognitive neuroscience

While earlier methods for improving functional correspondence operated in three-dimensional anatomical space (e.g., rubber-sheet warping; Conroy et al., 2013; Sabuncu et al., 2010), perhaps the foundational functional alignment reference is Haxby et al. 2011 which introduced "hyperalignment." Hyperalignment is a specific application of generalized Procrustes analysis where fMRI data from three or more individuals are successively registered to a common activation space through linear transformations. Although the original experiments focused on activations in a single region of interest, hyperalignment has since been extended to whole-brain contexts (Guntupalli et al., 2016) and to functional connectivity data (Busch et al., 2020; Guntupalli et al., 2018).

Beyond Procrustes-based methods, a variety of other algorithms have also been proposed for functional alignment. At a broad level, these algorithms can be characterized along two dimensions. The first is whether they employ linear or nonlinear transformations. Linear methods only allow for aligning two or more activation spaces by rigid-body transformations, along with affine registrations such as stretching and shearing. Nonlinear methods, by contrast, allow for more variable transformations and are particularly promising in cases of weak correspondence across individuals, such as broad functional reorganization following stroke (Langs et al., 2010; Nenning et al., 2020) or when comparing across species (Xu et al., 2019). Throughout this thesis, I focus on linear algorithms, as these have the highest potential for interpretability and—unlike several field-standard non-linear algorithms (e.g., diffusion map embedding; Nenning et al., 2017)—they can be derived on co-occurring data of one task structure such as movie-watching and applied on another, unrelated task structure such as a traditional psychological paradigm.

The second dimension by which to characterize functional alignment algorithms is on whether they include dimensionality reduction. For example, algorithms such as gradient

hyperalignment (Xu et al., 2018), regularized canonical correlation analysis (Bilenko and Gallant, 2016; Xu et al., 2012; Yousefnezhad and Zhang, 2016) and the shared response model (Chen et al., 2015; Richard et al., 2019) include a dimensionality reduction to find shared latent factors across individual functional activations. Other methods such as hyperalignment, optimal transport (Bazeille et al., 2019), and ridge regression (Tavor et al., 2016), by contrast, operate directly on the provided voxel time series. Fundamentally, methods that include a dimensionality reduction make different assumptions as to the nature of information that is likely to correspond across individuals; that is, whether the high-dimensional, voxelwise representation is a faithful measure of shared information, or if it is a noisy measure of a lower-dimensional, shared latent process.

For each of these functional alignment methods, a variety of applications are possible with both synchronized and unsynchronized task data. From synchronized data, researchers first generate functional alignment parameters from high-engagement stimuli such as naturalistic audio-visual narratives (Sonkusare et al., 2019; Vanderwal et al., 2019). In the case of unsynchronized data—such as resting-state time series—functional alignment may be used to synchronize the time series itself (Joshi et al., 2017). More commonly, however, unsynchronized data are aligned using patterns of functional connectivity derived across the entire time series (Nastase et al., 2020b). Derived connectomes themselves can also be directly functionally aligned to find correspondence across different parcellation schemes (Dadashkarimi et al., 2021). Other relevant applications include approximating functional localizers from synchronized naturalistic data (Jiahui et al., 2020), deriving shared encoding models (Van Uden et al., 2018), and even evaluating individual differences (Feilong et al., 2018).

Typically, the success of functional alignment is then assessed by applying the derived transformations to unrelated task data and training classifiers to predict individual participant activity based on shared patterns across the group; i.e., inter-subject decoding. Researchers have primarily focused on improving inter-individual correspondence during low-engagement paradigms such as traditional psychological tasks (Nastase et al., 2020a; Vanderwal et al., 2017) as these are likely to most strongly benefit from improved mapping between individual activation spaces. In cases where separate psychological task data is not available in the same participants, however, researchers may instead create alternate classification benchmarks such as time-segment matching for a continuous, naturalistic narrative. In some of these applications, alignment has been seen to significantly boost inter-subject decoding accuracy, approaching or even exceeding accuracy from within-subject classification (Haxby et al., 2011). This is particularly impressive as rates of accuracy

for between-subject classification for functionally unaligned data are typically at or below chance.

Given these successes, interest in functional alignment has continued to grow. Potential applications in psychiatric research (Anderson et al., 2021) and connections with ongoing work in systems neuroscience (Chen et al., 2021; Stella and Treves, 2021) promise new opportunities to assess both how neural coding is preserved in health and disease as well as across species. Despite this enthusiasm, the proportion of cognitive neuroscience researchers who engage with functional alignment methods is relatively small, both due to challenges in defining appropriate use cases as well as providing accessible implementations.

2.5 Challenges in adopting functional alignment

While potential algorithms and applications for functional alignment have proliferated, the gap between theory and experiment has become increasingly obvious. This problem is not unique to functional alignment and instead reflects a relatively common dissociation between methodological and domain-oriented work, with only a small subset of developed algorithms taken up by the field. Many challenges exist for broader adoption, including minimal characterization of the appropriate applications for each new method. That is, while a method may perform well on the dataset used to benchmark its performance, will it perform similarly on other datasets with different characteristics? This is particularly concerning as methods tend to overfit to a given benchmark (Recht et al., 2018), and a small number of open datasets such as Study Forrest (Hanke et al., 2014) and Sherlock (Chen et al., 2017) are predominantly used to characterize new functional alignment methods. Further, providing methods in accessible software packages or other, computational resources significantly lowers the barrier to adopting a method. New methods, however, are still rarely distributed as usage-oriented code—if indeed they are distributed as code at all—but instead as software for re-executing the original experiments (Pradal et al., 2013).

As an interdisciplinary field, cognitive neuroscience faces unique, additional challenges beyond characterizing appropriate applications. Many researchers lack the methodological training necessary to engage with high-dimensional methods (Hauk, 2020), making it difficult to effectively re-implement functional alignment within individual research groups. This challenge has been recognized for multivariate methods in cognitive neuroscience more broadly (Cohen et al., 2017), leading to the development of many field-standard Python software libraries such as PyMVPA (Hanke et al., 2009) and Nilearn (Abraham

et al., 2014). Importantly, these libraries include extensive tutorials orienting researchers to their use. As the multivariate toolkit has continued to expand, new software libraries have emerged to meet community needs. Recently, the Python library BrainIAK (Kumar et al., 2020) released a series of tutorials on emerging multivariate methods for fMRI data, including for the Shared Response Model (SRM; Kumar et al., 2019) as well as other methods such as inter-subject correlation (Nastase et al., 2019). Thus, at present, PyMVPA supports hyperalignment while BrainIAK supports SRM, with no direct connection between the two. Other methods exist in standalone implementations developed for individual research projects, discouraging wider use (Benureau and Rougier, 2017). As a result, the relationship between alignment techniques and their appropriate applications remains poorly explored outside of the original methodological papers introducing each method.

There is a clear need for new educational materials to connect functional alignment methods implemented across software libraries, without being directly tied to a single library. One challenge for researchers who hope to develop these materials is to select the appropriate format. Tutorials that accompany software libraries are often written as Jupyter Notebooks (Kluyver et al., 2016), which allow for interweaving code, results, and supporting scientific narrative. Jupyter notebooks are a commonly used computational format (Rule et al., 2019); however, they may be relatively unfamiliar to cognitive neuroscience researchers who use other programming languages such as MATLAB. Further, these formats do not align with traditional publishing infrastructures which focus on PDFs. The BrainIAK tutorials, for example, are described in a brief piece published in *PLOS Computational Biology*, while the actual tutorials themselves are hosted directly by the authors as standalone Jupyter notebooks. Since only the short paper is archived, the availability of the tutorials themselves depends on the authors continued access to their current hosting infrastructure. It further means that relatively few journals consider these as publishable research objects—since the majority of content is outside of the review process—disincentivizing researchers from investing time to develop these materials. Alleviating this challenge requires the development of new publishing infrastructures that more directly support executable research objects.

2.6 Summary and conclusions

Although neuroscience aims to identify generalizable principles by which the brain supports cognition and behavior, to do so effectively requires robust inter-individual mappings. Variability in structural and functional features of brain organization challenge

these mappings, and the limitations of modern neuroimaging methods such as fMRI further complicate clear comparisons across individuals. From its origins in connectivist philosophy of science to its rapid development in cognitive neuroscience, functional alignment has emerged as a promising technique to improve mapping across the distributed representations that characterize neural activity. The work described in this thesis explores existing paradigms for functional alignment with fMRI data and develops new resources to address current challenges in these applications. First, in Chapter 3, I benchmark five functional alignment methods on four publicly available datasets to identify which algorithms show robust performance. Then, in Chapter 4, I examine how diverse experimental factors including data characteristics for deriving and applying functional alignment transformations impact algorithm performance. Next, in Chapter 5, I synthesize these results into general guidelines to guide cognitive neuroscience researchers who are interested in using functional alignment methods in their own experimental work. I also develop accompanying online materials for researchers to directly access code supporting each of the described alignment methods. Finally, in Chapter 6, I discuss current publishing infrastructure for executable research objects such as those included in Chapter 5, highlighting my recent work in this area as well as important next steps.

Overall, the current thesis develops an initial framework for the application of functional alignment in fMRI, examining how algorithm choice and experimental context impact relative performance, developing guidelines for experimental applications, and exploring how these guidelines may be better communicated through future developments on open publishing infrastructure.

Chapter 3

An empirical evaluation of functional alignment using inter-subject decoding

Thomas Bazeille[1,†], Elizabeth DuPre[2,†], Hugo Richard[1], Jean-Baptise Poline[2], Bertrand Thirion[1]

[1]Université Paris-Saclay, Inria, CEA, Palaiseau, 91120, France

[2]Montréal Neurological Institute, McGill University, Montréal, Canada

†These authors contributed equally to this work.

Published in:

NeuroImage: https://doi.org/10.1016/j.neuroimage.2021.118683

3.1 Preface

As a broad class of methods, functional alignment includes several algorithms which each place different constraints on the transformations that can be learned across individual functional patterns. Many of these algorithms are already in use in the cognitive neuroscience literature. However, they are rarely systematically compared outside of their initial introduction to the field. Even in these methodological papers, the number of algorithms and the range of considered datasets is limited, hindering a clear understanding of their application to functional magnetic resonance imaging (fMRI) data. In this chapter, I benchmark five unique functional alignment algorithms across four publicly available datasets. I use inter-subject decoding to evaluate the relative similarity across subjects before and after

functional alignment and characterize algorithm performance at both the region-of-interest and whole-brain level of analysis. This work was published in *NeuroImage* in 2021 (Bazeille et al., 2021).

3.2 Abstract

Inter-individual variability in the functional organization of the brain presents a major obstacle to identifying generalizable neural coding principles. Functional alignment—a class of methods that matches subjects' neural signals based on their functional similarity—is a promising strategy for addressing this variability. To date, however, a range of functional alignment methods have been proposed and their relative performance is still unclear. In this work, we benchmark five functional alignment methods for inter-subject decoding on four publicly available datasets. Specifically, we consider three existing methods: piecewise Procrustes, searchlight Procrustes, and piecewise Optimal Transport. We also introduce and benchmark two new extensions of functional alignment methods: piecewise Shared Response Modelling (SRM), and intra-subject alignment. We find that functional alignment generally improves inter-subject decoding accuracy though the best performing method depends on the research context. Specifically, SRM and Optimal Transport perform well at both the region-of-interest level of analysis as well as at the whole-brain scale when aggregated through a piecewise scheme. We also benchmark the computational efficiency of each of the surveyed methods, providing insight into their usability and scalability. Taking inter-subject decoding accuracy as a quantification of inter-subject similarity, our results support the use of functional alignment to improve inter-subject comparisons in the face of variable structure-function organization. We provide open implementations of all methods used.

3.3 Introduction

A core challenge for cognitive neuroscience is to find similarity across neural diversity (Churchland, 1998); that is, to find shared or similar neural processes supporting the diversity of individual cognitive experience. Anatomical variability and limited structure-function correspondence across cortex (Paquola et al., 2019; Vázquez-Rodríguez et al., 2019) make this goal challenging (Rademacher et al., 1993; Thirion et al., 2006). Even after state-of-the-art anatomical normalization to a standard space, we still observe differences in individual-level functional activation patterns that hinder cross-subject comparisons

(Langs et al., 2010; Sabuncu et al., 2010). With standard processing pipelines, it is therefore difficult to disentangle whether individuals are engaging in idiosyncratic cognitive experience *or* if they are engaging in shared functional states that are differently encoded in the supporting cortical anatomy.

To address this challenge, *functional alignment* is an increasingly popular family of methods for functional magnetic resonance imaging (fMRI) analysis: from the initial introduction of hyperalignment in Haxby et al. 2011, the range of associated methods has grown to include Shared Response Modelling (SRM; Chen et al., 2015) and Optimal Transport (Bazeille et al., 2019) with many variations thereof (see e.g. Xu et al. 2018; Yousefnezhad and Zhang 2017, among others). Although this class of methods is broadly referred to as both *functional alignment methods* and *hyperalignment methods*, we adopt the term *functional alignment methods* to better distinguish from the specific Procrustes-based hyperalignment implementation in use in the literature.

The conceptual shift from anatomically-based to functionally-driven alignment has opened new avenues for exploring neural similarity and diversity. In particular, by aligning activation patterns in a high-dimensional functional space (i.e., where each dimension corresponds to a voxel), we can discover shared representations that show similar trajectories in functional space but rely on unique combinations of voxels across subjects. For a review of current applications of functional alignment, see Haxby et al. 2020.

Nonetheless, it remains unclear how researchers should choose among the available functional alignment methods for a given research application. We therefore aimed to benchmark performance of existing functional alignment methods on several publicly accessible fMRI datasets, with the goal of systematically evaluating their usage for a range of research questions. We consider performance to include both (1) improving inter-subject similarity while retaining individual signal structure as well as (2) computational efficiency, as the latter is an important consideration for scientists who may not have access to specialized hardware. Here, we specifically focus on pairwise alignments wherein subjects are directly aligned to a target subject's functional activations. An alternative approach is known as template-based alignment, wherein a group-level functional template is first created and then used as a reference space to which individual functional activations are aligned. Although template-based approaches are an important area of research—particularly for datasets with a large number of subjects—the question of how best to generate the reference template is distinct from its alignment and beyond the scope of the current work. For all alignment methods considered here, technically up-to-date and efficient implementations to reproduce these results are provided at `https://github.com/`

neurodatascience/fmralign-benchmark.

3.3.1 Defining levels of analysis: region-of-interest or whole-brain

Functionally aligning whole-brain response patterns at the voxel level is computationally demanding and may yield biologically implausible transformations (e.g., aligning contralateral regions). Therefore, currently available functional alignment methods generally define transformations within a sub-region. This constraint acts as a form of regularization, considering local inter-subject variability rather than global changes such as large-scale functional reorganization. It also divides the computationally intractable problem of matching the whole-brain into smaller, more tractable sub-problems.

An important consideration, then, is how to define a local neighborhood. Broadly, two main strategies exist: (1) considering voxels within a given region of interest (ROI) that reflects prior expectations on the predictive pattern or (2) grouping or parcellating voxels into a collection of subregions across the whole-brain. Existing functional alignment methods have been proposed using both approaches. For example, the initial introduction of hyperalignment in Haxby et al. 2011 was evaluated within a ventral temporal cortex ROI and was later extended to aggregate many local alignments into larger transforms using a *Searchlight* scheme (Guntupalli et al., 2016). Other methods such as Optimal Transport have been evaluated on whole-brain parcellations (Bazeille et al., 2019), where transforms are derived for each parcel in parallel and then aggregated into a single whole-brain transform. Throughout this work, we therefore consider functional alignment methods at both the ROI and aggregated whole-brain level of analysis.

3.3.2 Quantifying the accuracy of functional alignment

3.3.2.1 Image-based statistics

A key question is how to objectively measure the performance of functional alignment. One approach is to consider alignment as a reconstruction problem, where we aim to learn a functional alignment transformation that allows us to impute missing images in a target subject using data from source subjects. These functionally aligned maps can then be compared with held-out ground-truth maps from the target subject. We can quantify this comparison using image-based statistics such as the correlation of voxel activity profiles across tasks (Guntupalli et al., 2016; Jiahui et al., 2020), spatial correlation or Dice coefficient between estimated and held-out brain maps (Langs et al., 2014) or other metrics such as *reconstruction ratio* (Bazeille et al., 2019). However, these image-based statistics are sensitive

to low-level image characteristics (e.g., smoothness, scaling), and their values can therefore reflect trivial image processing effects (such as the smoothness introduced by resampling routines) rather than meaningful activity patterns.

3.3.2.2 Adopting a predictive framework to quantify alignment accuracy

Rather than using image-based statistics, an alternative approach is to test functional alignment accuracy in a predictive framework. Prior work adopting this framework has used tests such as time-segment matching from held-out naturalistic data (e.g., Chen et al., 2015; Guntupalli et al., 2016). However, because time-segment matching relies on the same stimulus class to train and test the alignment, it is unclear whether the learnt functional transformations extend to other, unrelated tasks—particularly tasks with low inter-subject correlation (Nastase et al., 2019). We are therefore specifically interested in predictive frameworks that probe model validity by measuring accuracy on held-out data from a different stimulus class, with or without functional alignment.

Inter-subject decoding is a well-known problem in the literature aimed at uncovering generalizable neural coding principles. More in detail, in inter-subject decoding one learns a predictive model on a set of subjects and then test that model on held-out subjects, measuring the extent to which learned representations generalize across individuals. In an information-mapping framework (Kriegeskorte and Diedrichsen, 2019), decoding allows one to assess the mutual information between task conditions. Alternate information-mapping approaches include Representational Similarity Analysis (Kriegeskorte et al., 2008), which assesses similarities between relative patterns of activations across task conditions. In this context, functional alignment should facilitate information-mapping by increasing the similarity of condition-specific representations across subjects, thus improving their decoding.

Although the link between mutual information and decoding accuracy is non-trivial (Olivetti et al., 2011), we consider that measuring alignment with decoding accuracy on unseen subjects better fulfils neuroscientists' expectations of inter-subject alignment in two main ways. First, decoding accuracy provides a more interpretable assessment of performance than other measures such as mutual information estimates. Second, decoding accuracy on a held-out sample provides insight into the external validity and therefore generalizability of derived neural coding principles. Compared to image-based measures, decoding accuracy is thus a more rigorous measure of whether functional alignment improves the similarity of brain signals across subjects while also preserving their structure and usability for broader research use cases. In this work, we therefore quantify functional

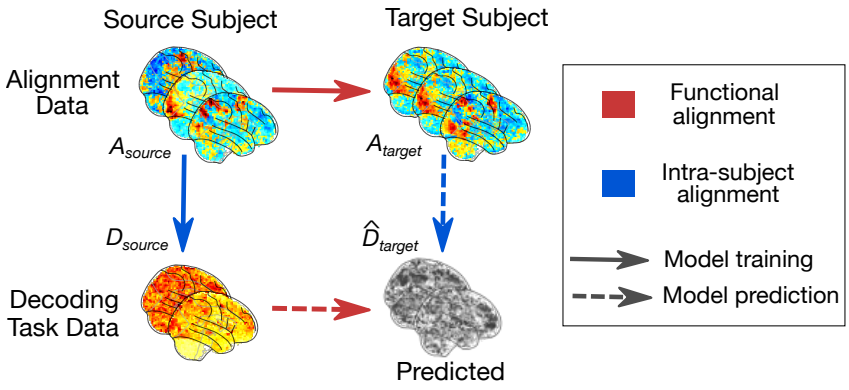

Figure 3.1: **Principle of functional alignment.** The goal of functional alignment is to learn correspondence between data drawn from two subjects: from a **source** subject to a **target** subject using their synchronized **alignment** data **A**. In this paper, each subject comes with additional **decoding** task data **D**. Red arrows describe functional alignment methods where correspondence is learnt from $\mathbf{A}^{source}$ to $\mathbf{A}^{target}$, while blue arrow describes intra-subject alignment method, where we learn correlation structure from $\mathbf{A}^{source}$ to $\mathbf{D}^{source}$. Solid arrows indicate a transformation learnt during training. Dashed arrows indicate when the previously learnt transformation is applied in prediction to estimate $\hat{\mathbf{D}}^{target}$.

alignment accuracy by assessing improvements in inter-subject decoding when using functional alignment over and above anatomical alignment. That is, the field-standard approach of normalizing subjects to a standardized anatomical template using diffeomorphic registrations, as implemented in e.g. fMRIPrep (Esteban et al., 2019).

3.3.3 The present study

Using this inter-subject decoding framework, we: (1) establish that functional alignment improves decoding accuracy above anatomical-only alignment, (2) investigate the impact of common methodological choices such as whether alignment is learned in subregions across the whole brain or in a pre-defined region-of-interest (ROI), and (3) compare the impact of specific alignment methods in whole-brain and ROI-based settings. We then provide a qualitative comparison of the transformations learnt by each method to "open the black box" and provide insights into how potential accuracy gains are achieved. Finally, we discuss the availability, usability and scalability of current implementations for each of the methods considered.

3.4 Materials and methods

In this section, we first consider frameworks for aggregating local functional alignment transformations into a single, larger transform (Section 3.4.1.1) that can be applied at a

whole-brain scale. We then proceed by introducing mathematical notations for functional alignment, as well as the alignment methods included in our benchmark (Section 3.4.2). We next describe our procedure to quantify alignment performance using inter-subject decoding (Section 3.4.3) and a series of experiments aimed at investigating the impact of functional alignment on decoding accuracy (Section 3.4.4). Finally, we describe the datasets (Section 3.4.5) and implementations used to run each experiment (Section 3.4.6).

3.4.1 Aggregating local alignments

3.4.1.1 Comparing searchlight and piecewise schemes

As discussed in Section 3.3.1, alignment methods are closely linked with the definition of local correspondence models. To align the entire cortex across subjects, two main frameworks have been proposed: searchlight and piecewise aggregation schemes. Each of these frameworks use functional alignment methods to learn local transformations and aggregate them into a single large-scale alignment; however, searchlight and piecewise differ in how they aggregate transforms, as illustrated in Figure 3.2. The *searchlight* scheme (Kriegeskorte et al., 2006), popular in brain imaging (Guntupalli et al., 2018, 2016), has been used as a way to divide the cortex into small overlapping spheres of a fixed radius. This method allows researchers to remain agnostic as to the location of functional or anatomical boundaries, such as those suggested by parcellation-based approaches. A local transform can then be learnt in each sphere and the full alignment is obtained by aggregating (e.g. summing as in Guntupalli et al., 2016 or averaging) across overlapping transforms. Importantly, the aggregated transformation produced is no longer guaranteed to bear the type of regularity (e.g orthogonality, isometry, or diffeomorphicity) enforced during the local neighborhood fit.

An alternative scheme, *piecewise alignment* (Bazeille et al., 2019), uses non-overlapping neighborhoods either learnt from the data using a parcellation method—such as k-means—or derived from an *a priori* functional or anatomical atlas. Local transforms are derived in each neighborhood and concatenated to yield a single large-scale transformation. Unlike searchlight, this returns a transformation matrix with the desired regularities. This framework might induce staircase effects or other functionally-irrelevant discontinuities in the final transformation due to the underlying boundaries.

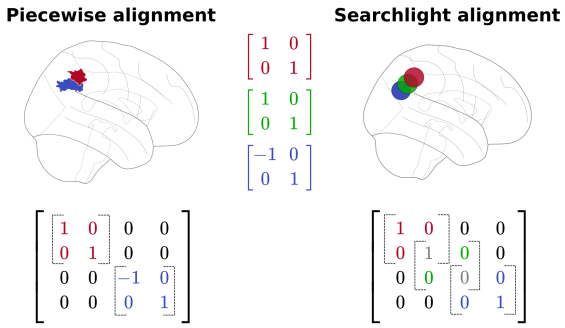

Figure 3.2: **Comparing piecewise and searchlight alignment.** In this illustration, transformations are derived for the blue, green, and red areas separately. Note that the piecewise alignment does not include a green area, as this corresponds to a searchlight overlapping both the red and blue areas. For non-overlapping parcels, these transformations are stacked into a larger orthogonal matrix. For the overlapping searchlight, these transformations are aggregated, with overlapping values averaged. Note that the final transformation for the searchlight alignment is no longer orthogonal in this example.

3.4.1.2 Aggregation schemes used in this benchmark

In the literature to date, searchlight and piecewise aggregation schemes have both been used in conjunction with Generalized Procrustes Analysis (detailed in section 3.4.2) under the names hyperalignment (Guntupalli et al., 2016) and scaled orthogonal alignment (Bazeille et al., 2019), respectively. We therefore include both searchlight Procrustes and piecewise Procrustes in our benchmark. Every other method is regularized at the whole-brain level of analysis through piecewise aggregation.

As piecewise alignment is learnt within a parcellation, an important question is: which brain atlas should be used for piecewise alignment? In Section S3.12 we compare results from the Schaefer et al. 2018 atlases to those from parcellations derived directly on the alignment data. By default, the results presented below are derived with the 300 ROI parcellation of the Schaefer atlas unless noted otherwise. In the case of searchlight Procrustes, we selected searchlight parameters to match those used in Guntupalli et al. 2016; that is, each searchlight had 5 voxel radius, with a 3 voxel distance between searchlight centers. All searchlight analyses were implemented using PyMVPA (Hanke et al., 2009).

3.4.2 Description of the benchmarked methods

As we use inter-subject decoding to compare functional alignment methods, we only consider methods that meet the following two criteria. First, the alignment transformations should be learnt on activations evoked during temporally synchronized (i.e., co-occuring) task data, or on contrasts matched across individuals. Second, the learnt transformations must be invertible or almost invertible linear mappings and applicable as-is on unseen

data with a different task structure. These two criteria exclude several methods currently used in the literature such as regularized canonical correlation analysis (rCCA; Bilenko and Gallant, 2016), gradient hyperalignment (Xu et al., 2018), connectivity hyperalignment (Guntupalli et al., 2018), and methods based on Laplacian embeddings (Langs et al., 2014).

In our whole-brain benchmark, we consider five different alignment methods: searchlight Procrustes (Guntupalli et al., 2016; Haxby et al., 2011), piecewise Procrustes, piecewise Optimal Transport (Bazeille et al., 2019), piecewise Shared Response Modelling (SRM; Chen et al., 2015), and intra-subject correlations across tasks (Tavor et al., 2016), here referred to as "intra-subject alignment." We provide a brief summary of these methods below.

3.4.2.1 General notations

Assume that for every subject we have alignment data $\mathbf{A} \in \mathbb{R}^{p \times n}$ and decoding task data $\mathbf{D} \in \mathbb{R}^{p \times d}$, where n is the number of alignment time points or frames, d is the number of decoding task images and p is the number of voxels. The alignment and decoding task data are collected for both *source* and *target* subjects, which we denote with superscripts.

In general, functional alignment methods learn a transformation matrix $\mathbf{R} \in \mathbb{R}^{p \times p}$ that best maps functional signals from a source subject to those of a target subject. To do so, $\mathbf{R}$ can be seen as a linear mixing of *source* voxels signals such that $\mathbf{R}\mathbf{A}^{source}$ best matches $\mathbf{A}^{target}$. $\mathbf{R}$ is then applied on separate, held-out data from the source subject, $\mathbf{D}^{source}$ to estimate $\mathbf{D}^{target}$. Because we are only learning an estimate of that held-out decoding task data, we denote this $\hat{\mathbf{D}}^{target}$. Thus, $\hat{\mathbf{D}}^{target} = \mathbf{R}\mathbf{D}^{source}$.

We consider one method, intra-subject alignment, which uses the same alignment and decoding task data to learn a different transformation than the one described above. Specifically, in intra-subject alignment we are interested in learning $\mathbf{R}^{intra} \in \mathbb{R}^{n \times s}$; that is, the "intra-subject" correlations between $\mathbf{A}^{source}$ and $\mathbf{D}^{source}$. We can then use $\mathbf{R}^{intra}$ to output $\hat{\mathbf{D}}^{target} = \mathbf{R}^{intra}\mathbf{A}^{target}$. Thus, the main distinction here is that intra-subject alignment does not learn a source-target mapping; instead, it learns a $\mathbf{A}$ to $\mathbf{D}$ mapping within-subjects. These notations are illustrated in Figure 3.1.

3.4.2.2 Procrustes

Generalized Procrustes analysis, introduced to the cognitive neuroscience literature as *hyperalignment* (Haxby et al., 2011), searches for an orthogonal local transformation $\mathbf{R}$ to align subject-level activation patterns such that:

$$\min_{\mathbf{R}=s\mathbf{M}} ||\mathbf{R}\mathbf{A}^{source} - \mathbf{A}^{target}||_F^2, \quad s \in \mathbb{R}^+, \quad \mathbf{M} \in \mathbb{R}^{p \times p} \tag{3.1}$$

where p is the number of voxels in a given region, such that

$$\mathbf{M}^\intercal \mathbf{M} = \mathbf{I}_p \tag{3.2}$$

This transform can be seen as a rotation matrix mixing signals of voxels in $\mathbf{A}^{\mathbf{source}}$ to reconstruct the signal of voxels in $\mathbf{A}^{\mathbf{target}}$ as accurately as possible. We note that hyperalignment as defined in (Haxby et al., 2011) uses a three stage alignment-and-averaging procedure to extend these Procrustes transformations into a group-level, template-based method. In the context of pairwise alignments, however, this method is naturally equivalent to Procrustes. Thus, in the rest of this work we use the terms "hyperalignment" and "Procrustes" interchangeably. As described in Section 3.4.1.2, we compare two whole-brain implementations of this method: piecewise Procrustes and searchlight Procrustes, that differ in the way local transformations are aggregated.

3.4.2.3 Optimal Transport

Optimal transport—first introduced as a functional alignment method in Bazeille et al. 2019—estimates a local transformation $\mathbf{R}$ that aligns subject-level activation patterns at a minimal overall cost. Specifically, we can compute the cost of aligning two subject-level activation patterns as $\mathrm{Tr}(\mathbf{R} \cdot \mathbf{C})$, where $\mathbf{C}$ is the functional dissimilarity—or difference in activation patterns—between source and target, as measured by a pairwise functional distance matrix. Thus, for voxel i in $\mathbf{A}^{\mathbf{source}}$ and voxel j in $\mathbf{A}^{\mathbf{target}}$:

$$\mathbf{C}_{i,j}(\mathbf{A}^{source}, \mathbf{A}^{target}) = ||\mathbf{A}_i^{source} - \mathbf{A}_j^{target}|| \tag{3.3}$$

Importantly, the resulting matching is constrained to exhaustively map all source voxels to all target voxels, with every voxel having an equal weight. This implicitly yields an invertible and strongly constrained transform, preserving signal structure as much as possible. To allow for a more efficient estimation, we slightly relax this constraint with an additional entropic smoothing term. As introduced in Cuturi, 2013, we can then find $\mathbf{R}$, the regularized Optimal Transport plan by finding a minimum for Equation 3.4 through the Sinkhorn algorithm.

$$\min_{\substack{\mathbf{R} \in {\mathbb{R}_+}^{p \times p}; \\ \mathbf{R}\mathbf{1}=1/p, \ \mathbf{1}\mathbf{R}^\top=1/p}} \mathrm{Tr}(\mathbf{R} \cdot \mathbf{C}) - \epsilon \mathbf{H}(\mathbf{R}) \tag{3.4}$$

where $\epsilon > 0$, and the discrete entropy of the transformation $\mathbf{H}(\mathbf{R})$ is defined as:

$$\mathbf{H}(\mathbf{R}) \stackrel{\text{def.}}{=} -\sum_{i,j} \mathbf{R}_{i,j}(\log(\mathbf{R}_{i,j}) - 1) \tag{3.5}$$

This method differs from Procrustes analysis in that it yields a sparser mapping between source and target voxels with high functional similarity, making it less sensitive to noisy voxels on both ends. The level of sparsity is controlled by ϵ, a user-supplied hyper-parameter, which we set to 0.1 throughout our experiments. For our implementation, we rely on the `fmralign` package. Optimal transport transformations are calculated in a piecewise fashion, following Bazeille et al., 2019.

3.4.2.4 Shared response model

The Shared Response Model (SRM), introduced in Chen et al. 2015, differs from Procrustes analysis and Optimal Transport in that it naturally provides a decomposition of all subjects's activity at once, rather than requiring pairwise transformations. Specifically, SRM (in its deterministic formulation) estimates a common shared response $\mathbf{S} \in \mathbb{R}^{k \times n}$ and a per-subject orthogonal basis $\mathbf{W}^i \in \mathbb{R}^{p \times k}$ from subject-level alignment data $\mathbf{A}^i$ such that:

$$\min_{\mathbf{W}^1,\ldots,\mathbf{W}^n,\mathbf{S}} \sum_i ||\mathbf{A}^i - \mathbf{W}^i\mathbf{S}||_F^2 \tag{3.6}$$

where n is the number of time points, p is the number of voxels, and k is a hyper-parameter indexing the dimensionality. The subject-specific basis $\mathbf{W}^i$ has orthonormal columns such that :

$$\mathbf{W}^{i\mathsf{T}}\mathbf{W}^i = \mathbf{I}_k \;\; \forall i \tag{3.7}$$

We specifically use the FastSRM implementation proposed by Richard et al. 2019 and available in the BrainIAK library (RRID: SCR_01 4824), that approximates this calculation with an emphasis on improved computational performance. For full details on the computational advantages of FastSRM, we direct the reader to their work.

In order to align our SRM implementation with the other considered alignment algorithms, we introduce a new piecewise SRM method to aggregate SRM transformations across the whole brain. Thus within each parcel or across an *a priori* ROI, SRM decomposes the signal of many subjects in a common basis, with the same orthogonality constraint as Procrustes. This ability to jointly fit inter-subject data through orthogonal transforms makes it reminiscent of Procrustes, with a caveat: SRM is effective if the number of components k is large enough to capture all distinct components in the signal.

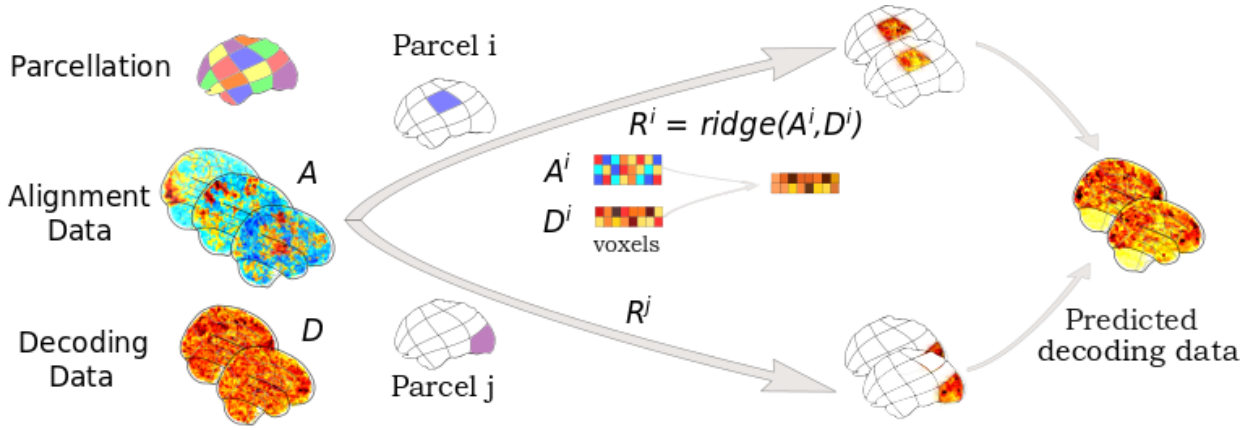

Figure 3.3: **Intra-subject alignment.** Using intra-subject alignment to learn piecewise correlations between a single subject's alignment and decoding task data. As with other piecewise methods, this mapping is learnt separately for all parcels $i \ldots j$ of the chosen parcellation. For the ith parcel, voxels are samples used to train a cross-validated ridge regression $\mathbf{R}^i$ to map between the two task conditions—alignment data $\mathbf{A_i}$ and independent decoding task data $\mathbf{D_i}$—for this source subject. We then aggregate these piecewise predictions into a single, whole-brain prediction $\hat{\mathbf{D}}$. In training, this prediction can be directly compared to the ground-truth decoding data, $\mathbf{D}$. When testing, we would have access to the target subject's alignment data $\mathbf{A}$ but not their decoding task data, $\mathbf{D}$.

Given the strong dependency of SRM performance on the selected hyper-parameter k, this parameter requires additional experimenter consideration. For piecewise SRM, we perform a grid search to select the relevant Schaefer parcellation resolution and number of components k (see Section S3.13). From these results, we chose to use Schaefer atlas 700 and run one SRM on each parcel searching for 50 components—or equal to the number of voxels if less than 50 voxels are in a given parcel. For ROI-based analyses, we set k to 50 components as in our piecewise analyses and matching the original SRM benchmarks provided in Chen et al., 2015.

3.4.2.5 Intra-subject alignment

Another alternative to pairwise functional alignment has been proposed in Tavor et al. 2016. In their paper, Tavor and colleagues show that while individual activity patterns in each task may appear idiosyncratic, correspondences learnt across different tasks using a general linear model (e.g., to predict task data from resting-state derived features) display less across-subject variability than individual activity maps. This provides an interesting twist on the typical functional alignment workflow: while most methods learn alignments within a single task and across subjects, we can instead learn within-subject correlations across tasks. The structure of learnt task-specific correlations should then hold in new,

unseen subjects. We include here a method for learning these intra-subject correlations in a piecewise fashion, which we call *intra-subject alignment*.

Figure 3.3 illustrates how we can learn the local-level correlation structure between two independent tasks $\mathbf{A}^{source} \in \mathbb{R}^{p \times n}$, $\mathbf{D}^{source} \in \mathbb{R}^{p \times d}$ within a single *source* subject. We denote the mapping between these tasks as $\mathbf{R}^{intra}$ to distinguish it from mappings that are learnt between pairs of subjects.

From preliminary analyses we observed that—unlike other piecewise techniques (Section S3.12)—the decoding accuracy for intra-subject alignment strictly improved with parcellation resolution so we use the highest resolution Schaefer atlas available (Schaefer et al., 2018). Thus, we first divide alignment and decoding data into 1000 parcels. On a local parcel i, each voxel is considered a sample and we train $\mathbf{R}_i^{intra} \in \mathbb{R}^{n \times d}$ through ridge regression:

$$\mathbf{R}_i^{intra} = \arg\min_{\mathbf{R}_i} ||\mathbf{A}_i^{source}\mathbf{R}_i - \mathbf{D}_i^{source}||_F^2 + \alpha||\mathbf{R}_i||_F^2 \tag{3.8}$$

The hyperparameter α is chosen with nested cross-validation among five values scaled between 0.1 and 1000 logarithmically.

After repeating this procedure for all *source* subjects, we then use $\mathbf{R}^{intra}$ to estimate decoding data for *target* subject as $\hat{\mathbf{D}}^{target} = \mathbf{A}^{target}\mathbf{R}^{intra}$. As with other functional alignment methods, we can evaluate the quality of our estimation using an inter-subject decoding framework.

3.4.3 Experimental procedure

For each dataset considered (described in Section 3.4.5), we calculated the inter-subject decoding accuracy for anatomical-only alignment and for each of the five considered functional alignment methods.

To calculate inter-subject decoding accuracy, we took the trial- or condition-specific beta maps generated for each dataset (see Section 3.4.5 for full details on beta-map generation) and fit a linear Support Vector Machine (SVM). In order to ensure fair comparisons of decoding accuracy across experiments, we chose a classifier with no feature selection and default model regularization ($C = 1.0$). Classifiers were implemented in scikit-learn (Pedregosa et al., 2011), and decoding accuracy was assessed using a leave-one-subject-out cross-validation scheme. That is, the linear SVM was trained to classify condition labels on all-but-one subject and the resulting trained classifier was used without retraining on the held-out subject, providing an accuracy score for that cross-validation fold.

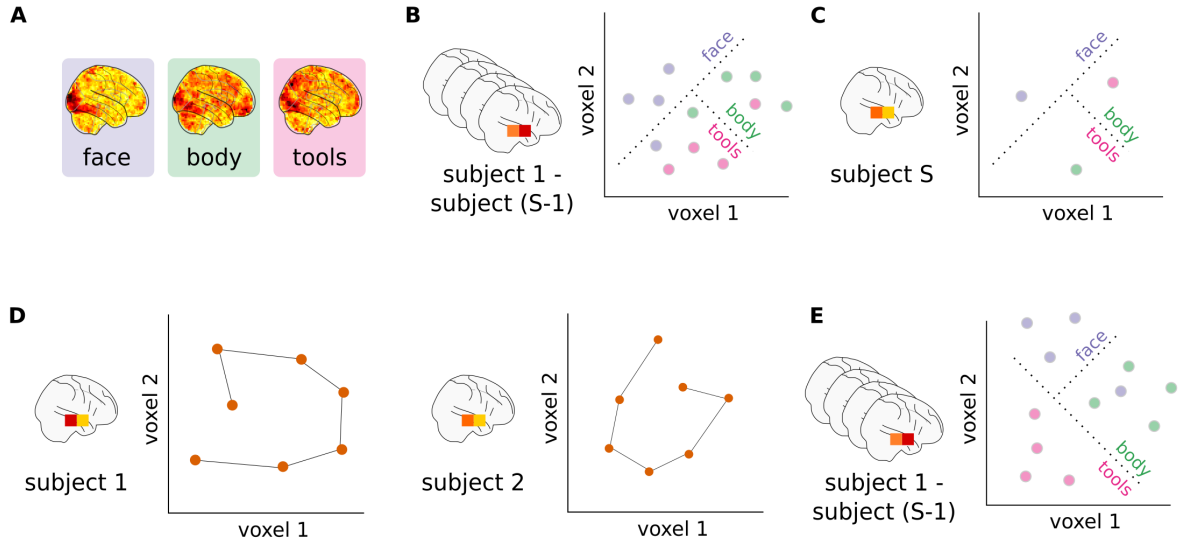

Figure 3.4: **Analysis pipeline.** (A) First-level general linear models are fit for each subject to derive trial- or condition-specific beta-maps for each session. (B) These beta maps and their matching condition labels are used to train a linear SVM on the training set of subjects. (C) The trained classifier is applied on a held-out test subject, and accuracy is assessed by comparing the predicted and actual condition labels. (D) On a separate task, we compare subject-level activation patterns as trajectories in the high-dimensional voxel space. This allows us to learn functional alignment transformations that maximize the similarity of these high-dimensional spaces. (E) These voxel-wise transformations are applied on the decoding beta maps, and a new linear SVC is trained to predict condition labels. This trained classifier can then be applied to the held-out test subject and decoding accuracy assessed as in (C).

For each dataset, we first calculated the inter-subject decoding accuracy using anatomical alignment. This served as a baseline accuracy against which we could compare each functional alignment method. Using alignment data, functional alignment transformations were then learnt for each pairwise method, where the left-out subject for that cross-validation fold was the target subject for functional alignment. Inter-subject decoding accuracy was then re-calculated after applying functional alignment transformations to the decoding beta maps.

In the special case of SRM—which allows for calculating an alignment from all provided subjects in a single decomposition—we withheld the left-out subject from the shared response estimation step to avoid data leakage. The projection of the left-out subject is then learnt from previously estimated shared space. Finally, the learnt projections are applied to the decoding data, and decoding is performed on the projected data.

For each cross-validation fold, we report the inter-subject decoding accuracy of a given functional alignment method after subtracting the baseline, anatomical-only accuracy for that same fold. An overview of the experimental procedures is provided in Figure 3.4.

3.4.4 Main experiments

Experiment 1 uses the experimental procedure described previously to assess accuracy gains provided by alignment methods with respect to anatomical alignment when applied on whole-brain images. We benchmarked the five methods described in Section 3.4.2: piecewise Procrustes, searchlight Procrustes, piecewise Optimal Transport, piecewise SRM, and intra-subject alignment, with relevant hyperparameters selected as described previously. Results of this benchmark (on five tasks from four datasets as described in Section 3.4.5) are presented in Section 3.5.1. For each method, we also assessed its computation time relative to piecewise Procrustes alignment. Piecewise Procrustes provides a reasonable computational baseline as it is the only considered alignment method that does not include a hyperparameter and therefore shows a stable computation time across experiments.

We estimate the noise ceiling for this task as within-subject decoding accuracy. Within-subject decoding was calculated separately for each subject as the average leave-one-session-out decoding accuracy. We can then directly compare this accuracy value to the inter-subject decoding accuracy when that subject is the target—that is, the left-out—subject. The difference between within- and anatomical inter-subject decoding accuracies, then, is a good approximation of the decoding accuracy lost due to inter-subject variability; therefore, it provides a range of possible accuracy gains that can be expected from functional alignment.

We then conducted *Experiment 2* to understand how whole-brain results compare to ROI-based analyses. Specifically, we replicated *Experiment 1* within selected ROIs, such that local alignment methods were applied directly without any aggregation scheme. ROIs were chosen based on *a priori* expectations of each decoding task (see Section 3.4.5 for details for each dataset). Results from *Experiment 2* are shown in Section 3.5.2.

Experiment 3 tackles the notoriously hard problem of understanding how each of the considered methods align subjects by examining qualitatively their impact on activity patterns across individuals. To “open the black-box," we reused IBC dataset full-brain alignments learnt in *Experiment 1*. Specifically, we consider the transformation to sub-04’s activity pattern from all other subjects’s functional data. With these transformations, we align two contrasts from each of the two decoding tasks of the IBC dataset: Rapid Serial Visual Presentation of words (RSVP language task) and sound listening. Finally, we run a group conjunction analysis (Heller et al., 2007) on these four aligned contrasts and visualize the results. This statistical analysis, more sensitive than its random effect equivalent on small samples, allows one to infer that every subject activated in the region with a proportion γ showing the effect considered. Here we use $\gamma = 0.25$ to recover all

regions selectively activated by at least a few subjects, and we show in Section 3.5.3 how this group functional topography is modified by alignment.

3.4.4.1 Control analyses

In addition to our three main experiments, we ran three additional control analyses on the IBC dataset. First, we aimed to assess the impact of the brain parcellation and its resolution on piecewise alignment by comparing whole-brain decoding accuracy for two IBC dataset tasks using piecewise Procrustes across both data-driven and pre-defined parcellations (Section S3.12). As piecewise SRM displays an interaction between parcellation resolution and the method-specific hyperparameter k, we ran an additional grid search for this algorithm to determine its optimal experimental parameters (Section S3.13).

Second, we calculated inter-subject decoding performance after applying Gaussian smoothing kernels of several widths on both IBC dataset decoding tasks (Section S3.14). Gaussian smoothing is of particular interest as a comparison to functional alignment, as it is commonly used to facilitate inter-subject comparisons by smoothing over residual variance in functional mappings. Finally, in a third control experiment, we assessed the impact of whether data is represented on the surface or the volume and resolution on decoding accuracy in the IBC RSVP language task (Section S3.15).

3.4.5 Datasets and preprocessing

In order to assess the performance of each functional alignment method in a range of applications, we searched for publicly accessible datasets that included both a task suitable to learn the alignment (e.g. naturalistic or localizer protocols) as well as an independent decoding task on which we could evaluate functional alignment performance. After discarding datasets where we could not obtain above-chance accuracy levels for within-subject decoding, we retained four datasets: BOLD5000 (Chang et al., 2019), Courtois-NeuroMod (Boyle et al., 2020), Individual Brain Charting (IBC; Pinho et al., 2018), and Study Forrest (Hanke et al., 2016). For the IBC dataset, we included both a language (RSVP language) and auditory (Sounds dataset) decoding task, yielding a total of five decoding tasks that probe visual, auditory and language systems. For a complete description of the alignment and decoding data included in each experiment, please see Table 3.1.

BOLD5000, StudyForrest and Courtois-NeuroMod were preprocessed with fMRIPrep (Esteban et al., 2019), while IBC data were preprocessed using an SPM-based pipeline as described in Pinho et al. 2018. A complete description of the fMRIPrep preprocessing

Dataset	S	Alignment data	p	Decoding task	Decoding categories	d
Individual Brain Charting	10	Contrast maps from HCP and ARCHI task batteries	53	RSVP Language	Words, Non-Words, Consonants, Sentences, Jabberwocky	360
				Sounds dataset	Voice, Nature, Animal, Music, Speech, Tools	72
BOLD5000	4	COCO, ImageNet, and Scenes images	300	Imagenet images content	Plant, Animal, Food, Artifact	350
Forrest	10	*Forrest Gump* audio-movie listening	1600	Music genre	Country, Metal, Ambient, Symphonic, Rock	200
Courtois Neuro-mod	6	*Life* movie watching	2008	Visual n-back condition	Body 0-back, Body 2-back, Face 0-back, Face 2-back, Place 0-back, Place 2-back, Tools 0-back, Tools 2-back	72

Table 3.1: **Datasets used to benchmark alignment methods.** The four datasets used in this benchmark, where each dataset consists of S subjects. We note the alignment data used for each dataset and p the number of timeframes it comprises. These datasets show the range of possible task structures which work for alignment—from static images for BOLD5000, to statistical contrast maps for IBC, to complex audio or audio-visual movies for Forrest and Courtois Neuromod. A full listing of included 53 contrast maps for IBC is included in Section S3.16. We also include the decoding task(s) used for each dataset. Each subject's decoding task data comprises d images evenly divided across the listed stimulus categories (except for BOLD5000 categories that are unbalanced). Of note, IBC dataset has two independent decoding tasks, bringing the total number of decoding tasks to five.

procedures is available in the appendix (Section S3.9). Preprocessed data were then masked using a grey matter mask, detrended, and standardized using Nilearn (Abraham et al., 2014a). To reduce the computational cost of functional alignment, we downsampled all included datasets to 3mm resolution. Both alignment and decoding task data were then additionally smoothed with a 5mm Gaussian kernel. A general linear model (GLM) was fit to each decoding task run to derive trial-specific beta maps (or condition-specific beta maps for the Courtois Neuromod and IBC Sounds tasks), which were carried forward for inter-subject decoding.

As described in Section 3.4.3, *Experiment 2* uses pre-defined regions of interest (ROIs). We selected large, task-relevant ROIs to ensure that sufficient signal was available when decoding. A large visual region, extracted from the Yeo7 (Buckner et al., 2011) atlas was used for the visual tasks in BOLD5000 and Courtois-NeuroMod. For Forrest and IBC Sounds—which are auditory tasks—we took the Neuroquery (Dockès et al., 2020) predicted response to the term “auditory". We then compared this predicted response with the BASC (Bootstrap Analysis of Stable Clusters) atlas (at scale 36; Bellec et al., 2010) and took the parcel most overlapping with the predicted response; namely, parcel 25. For IBC RSVP, which is a reading task, we extracted the BASC (at scale 20) atlas components most overlapping with MSDL (Multi-Subject Dictionary Learning; Varoquaux et al., 2011) atlas parcels labeled as left superior temporal sulcus, Broca and left temporo-parietal junction: namely, the 8 and 18 BASC components. We then kept only the largest connected component. All included ROIs are displayed in Figure 3.7.

3.4.6 Implementation

With the exception of Courtois Neuromod, all other included datasets are available on OpenNeuro (Poldrack et al., 2013) under the following identifiers: *ds000113* (Study Forrest), *ds001499* (BOLD5000), and *ds002685* (IBC). Courtois Neuromod `2020-alpha2` release will be available under a data usage agreement as outlined on `https://docs.cneuromod.ca`.

Our pipeline entirely relies on open-source Python software, particularly the SciPy stack (Virtanen et al., 2020). All included methods are implemented in `fmralign` or accessed through their original, open source implementations as described in Section 3.4.2. To ease replication and extension of the presented results, we have created the `fmralign-benchmark` repository under `https://github.com/neurodatascience/fmralign-benchmark`. This repository provides an implementation of the procedures adopted in these experiments, building on `fmralign` and previously cited tools.

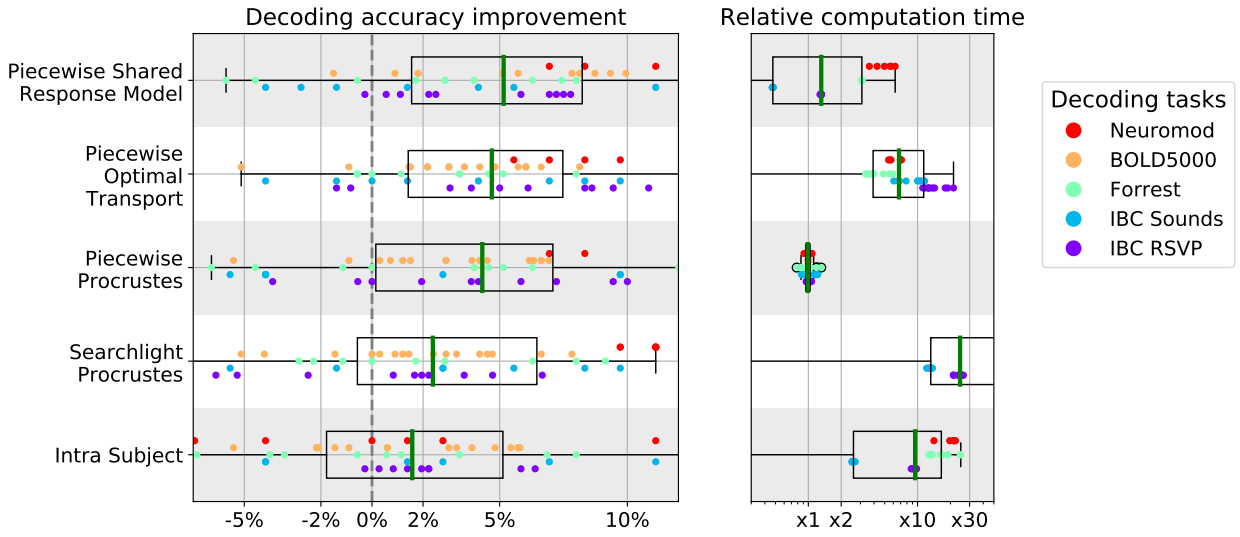

Figure 3.5: **Decoding accuracy improvement and computation time after whole-brain functional alignment.** In the *left panel*, we show decoding accuracy improvement for each of the considered functional alignment methods at the whole-brain level of analysis. Each dot represents a single subject, and subjects are colored according to their decoding task. To aggregate results across datasets, we show accuracy scores after subtracting inter-subject decoding accuracy for the same leave-one-subject-out cross-validation fold with anatomical-only alignment. While we depict relative accuracies here, absolute accuracy values are provided in Table S3.1. In the *right panel*, we show the computational time for each of the considered methods. All computation times are depicted as relative to piecewise Procrustes. For both panels, each box plot describes the distribution of values across datasets, where the green line indicates the median. All methods seem to improve decoding accuracy across datasets, especially piecewise Shared Response Model, piecewise Optimal Transport and piecewise Procrustes. We also see that piecewise Optimal Transport and searchlight Procrustes are respectively 7 and 25 times slower than piecewise Procrustes.

3.5 Results

3.5.1 Functional alignment improves inter-subject decoding

The *left panel* of Figure 3.5 displays absolute decoding accuracy change brought by each functional alignment method relative to anatomical alignment on whole-brain images. As every method is trained and tested on the same cross-validation folds, we report the fold-by-fold performance change. The *right panel* displays each method's relative computation time compared to piecewise Procrustes alignment. For each panel, each point displayed is the result for one leave-one-subject-out cross validation fold and each color corresponds to one of the five decoding tasks. Note that these timings are based on available implementations — `fmralign` for piecewise alignment methods, `pymvpa2` for searchlight, and `BrainIAK` for SRM— and are therefore subject to change as implementations improve. Nonetheless, these estimates provide insight into the current state-of-the-art.

3.5.1.1 Alignment substantially improves inter-subject decoding accuracy

Overall, we see that most functional alignment methods consistently improve decoding accuracy, with gains from 2-5% over baseline. This trend is relatively consistent across datasets and target subjects. Thus, alignment methods manage to reliably reduce individual signal variability while preserving task-relevant information in a variety of conditions. Although there is noticeable variance in performance across data sets, these methods generally show significant effects on inter-subject decoding accuracies. As reported in Table S3.1, baseline accuracy is around 20% above chance on average. In this setting, the observed 5% average improvement across datasets is a substantial increase in performance.

In order to provide further context for these results, we also estimated the noise ceiling for inter-subject decoding. Figure 3.6 reports that across datasets, the leave-one-session-out (i.e., within-subject) decoding accuracy for the target subject is on average 8.5% higher than the corresponding leave-one-subject-out (i.e., inter-subject) decoding accuracy after anatomical alignment for the same target subject. Thus, we expect that functional alignment methods will achieve at most an 8.5% increase in inter-subject decoding accuracy over anatomical alignment. In this light, we can see that the best functional alignment method recovers more than half of the decoding accuracy lost to inter-subject variability.

Additional control analyses suggest that this effect cannot be explained by smoothing (Section S3.14). We further find that the presented results are largely insensitive both to

whether the data is represented on the cortical surface or in volumetric space as well as to the parcellation resolution used (see section S3.15).

3.5.1.2 Piecewise methods show computational and accuracy advantages

Procrustes alignment results in better inter-subject decoding accuracies when performed in a piecewise as compared to a searchlight approach. Specifically, searchlight Procrustes shows lower decoding accuracies on average, suggesting that its internal averaging destroys part of the signal structure recovered by Procrustes. With respect to computational cost, we can see that searchlight Procrustes is 25 times slower on average than piecewise Procrustes. These results suggest that piecewise alignment is a better choice when calculating functional alignment transformations on full-brain data. Moreover, Section S3.12 shows that gains from piecewise alignment are largely insensitive to the resolution and type of parcellation used; i.e., taken from an atlas or learnt directly from subject data.

The two best performing alignment methods also use a piecewise aggregation scheme. Specifically, piecewise SRM and Optimal Transport yield the highest decoding scores, with a slightly lower standard deviation in accuracy scores than Procrustes.

Piecewise SRM is the best performing method and faster to train than piecewise Procrustes for a fixed set of hyperparameters; however, identifying the ideal hyperparameters for a new dataset requires a computationally costly grid-search. Our results (see Section S3.13) suggest that, in general, a large number of components k and a high-resolution parcellation are likely to give reasonable performance across datasets.

The second best performing method, Optimal Transport, gives non-trivial accuracy gains in most configurations and only rarely decreases decoding accuracy, likely because of the stronger constraints that it imposes. However, this extra-performance comes at a computational cost: it is on average 7 times slower than Procrustes. For data sets without sufficient data or computational power to perform a hyper-parameter grid search for piecewise SRM, we suggest that Optimal Transport offers robust decoding performance with little hyper-parameter tuning. It remains, however, more computationally costly than the reference implementation of piecewise Procrustes.

3.5.1.3 Task-specific mappings can be learnt within subjects

The intra-subject alignment approach differs from other considered functional alignment methods in that it learns mappings between the alignment data and decoding task data, with the assumption that these mappings can be generalized across subjects. Our results support this assumption, although this method yields gains half as large as the best

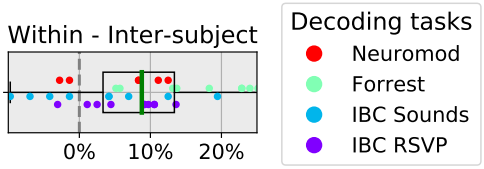

Figure 3.6: **Within-subject minus inter-subject decoding accuracy.** We show the difference between the average leave-one-session-out within-subject decoding accuracy and anatomically-aligned leave-one-subject-out inter-subject decoding accuracy, when that target subject is left-out. Thus, each dot corresponds to a single subject, and the dot's color indicates the decoding task. Of note, BOLD5000 was dropped as it did not have independent folds, and therefore could not be used for within-subject cross-validation. The box plot describes the distribution of differences, where the green line represents the median value. Considering that this difference approximates the effects of inter-individual variability, the best average accuracy improvement one can hope for using functional alignment is around 9%.

performing alignment method and comes with a significant computational cost. Part of this cost can be accounted for by the increase in the number of parcels that are used to preserve signal specificity. Nonetheless, using task-specific mappings as a functional alignment method suggests that future work on refining related methods may be a promising direction of research.

3.5.2 Whole-brain alignment outperforms ROI-based alignment

The *left panel* of Figure 3.7 displays the performance of each functional alignment method relative to anatomical alignment within task-relevant ROIs. The *right panel* displays each method's relative computation time compared to piecewise Procrustes alignment.

When visually compared to Figure 3.5, ROI-based decoding accuracies appear to be slightly lower than whole-brain decoding accuracies for most of the methods considered. We directly compare ROI-based and whole-brain alignment in a supplementary analysis, depicted in Figure S3.1, confirming that ROI-based decoding accuracies are in fact lower on average for the datasets considered in this work. Our results support previous work from the inter-subject decoding literature (Chang et al., 2015; Schrouff et al., 2018) and suggest that full-brain piecewise alignment yields the best overall decoding pipeline, though we note that this conclusion may change depending on the exact research context.

3.5.2.1 Optimal Transport and SRM show high ROI performance

Overall, we find that the best performing methods bring a 3-5% improvement in decoding accuracy at the ROI level of analysis. Specifically, Optimal Transport is on average the best performing method, with a median accuracy increase of 5% within task-relevant ROIs. Here, we see that baseline decoding accuracy is less than 10% above chance in all datasets

(with the exception of Courtois Neuromod; see Table S3.2 for exact accuracy values). Thus, the 5% accuracy increase brought by Optimal Transport represents a strong effect.

SRM yields the second best performance within ROIs, showing reasonable decoding accuracy gains on most datasets. It shows more variance across datasets, however, than the other considered methods. In particular, SRM decreases inter-subject decoding accuracy on the visual ROI for Courtois Neuromod, with accuracy values dropping by approximately -20% compared to anatomical alignment (see Table S3.2). Performance was not significantly improved by using a higher number (up to 600) components, highlighting the unique difficulty in identifying well-suited hyper-parameters for SRM. Interestingly, Procrustes shows substantially lower performance on average in the ROI compared to the whole-brain level of analysis, especially on large ROIs, possibly due to its weak regularization.

Computationally, we see that SRM is the fastest method and runs roughly 3 times faster than Procrustes, while Optimal Transport remains 10 times slower than Procrustes.

We also note that—on average—intra-subject alignment does not show increased inter-subject decoding accuracy within task-relevant ROIs. We suspect that this is likely because when restricting the learnt relationship between data types (e.g. movie-watching to classification task data) to a single ROI, the low number of predicted features precludes the identification of stable multivariate patterns that can transfer across subjects.

3.5.3 Qualitative display of transformations learnt by various methods

Understanding the effects of high-dimensional transformations—such as those used in functional alignment—is non-trivial. To aid in this process, we “open the black box" by functionally aligning a group of subjects to an individual target subject’s functional space and depict the resulting maps in Figure 3.8. Here, we reuse whole-brain alignments learnt in *Experiment 1*.

We also display the ground-truth individual activation maps in *panel A*, in order to better highlight how each method affects the signal distribution. As a reminder, the contrast data displayed here was not used to learn alignments, so it means that alignment learnt on various task data, not specifically related to language nor audition carried enough information for fine-grain registration of these networks.

We can see that overall, functional alignment methods enhance group-level contrasts compared to anatomical-only alignment; i.e., activation maps are more similar across functionally-aligned subjects. This result is not at the expense of signal specificity, since the aligned group topographies are still sharp. From the comparison between panels *A* and *B*, one can also conclude that alignment methods bring group topography much

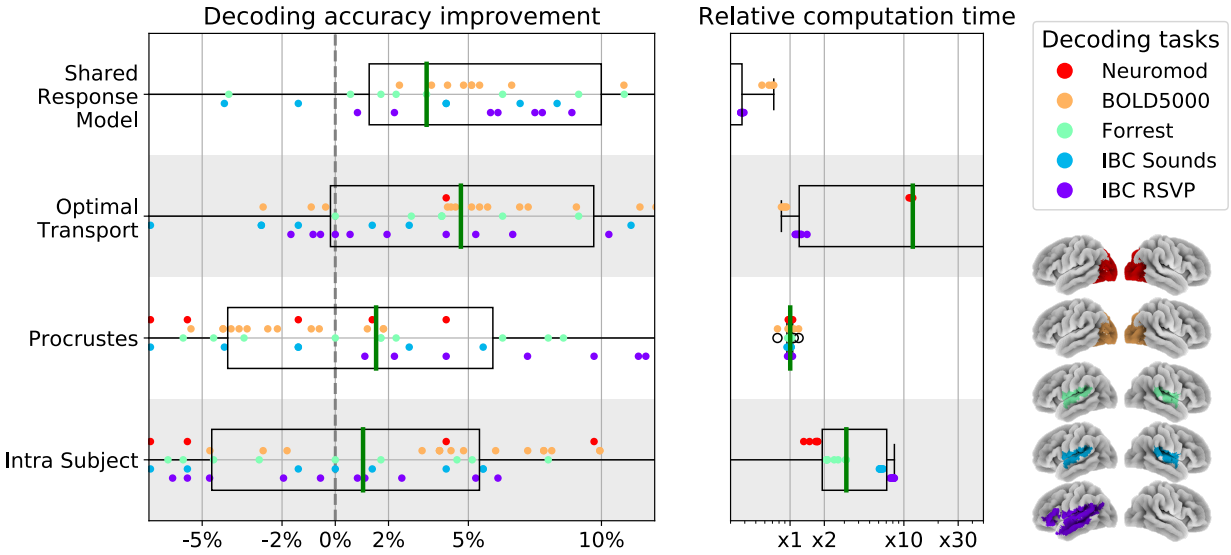

Figure 3.7: **Decoding accuracy improvement and computation time after ROI-based functional alignment.** In the *left panel*, we show decoding accuracy for each of the considered local functional alignment methods at the ROI level of analysis. The ROIs used for each dataset are displayed on the *far right*. Each dot represents a single subject, and subjects are colored according to their decoding task. Rather than raw values, we show accuracy scores after subtracting inter-subject decoding accuracy for the same leave-one-subject-out cross-validation fold with anatomical-only alignment. While we depict relative accuracies here, absolute accuracy values are provided in Table S3.2. Note that all methods are applied without aggregation, so only the method name is given. In the *right panel*, we show the computational time for each of the considered methods. All computation times are depicted as relative to piecewise Procrustes. For both panels, each box plot describes the distribution of values where the green line indicates the median.

closer to the targeted subject topography across the considered contrasts. Nonetheless, one can still observe that there seems to be a trade-off between sharpness of activation (low smoothness of image, due to low variance across aligned subjects) with Optimal Transport, and accuracy of their location compared to the target ones (low bias introduced by the matching) with searchlight Procrustes.

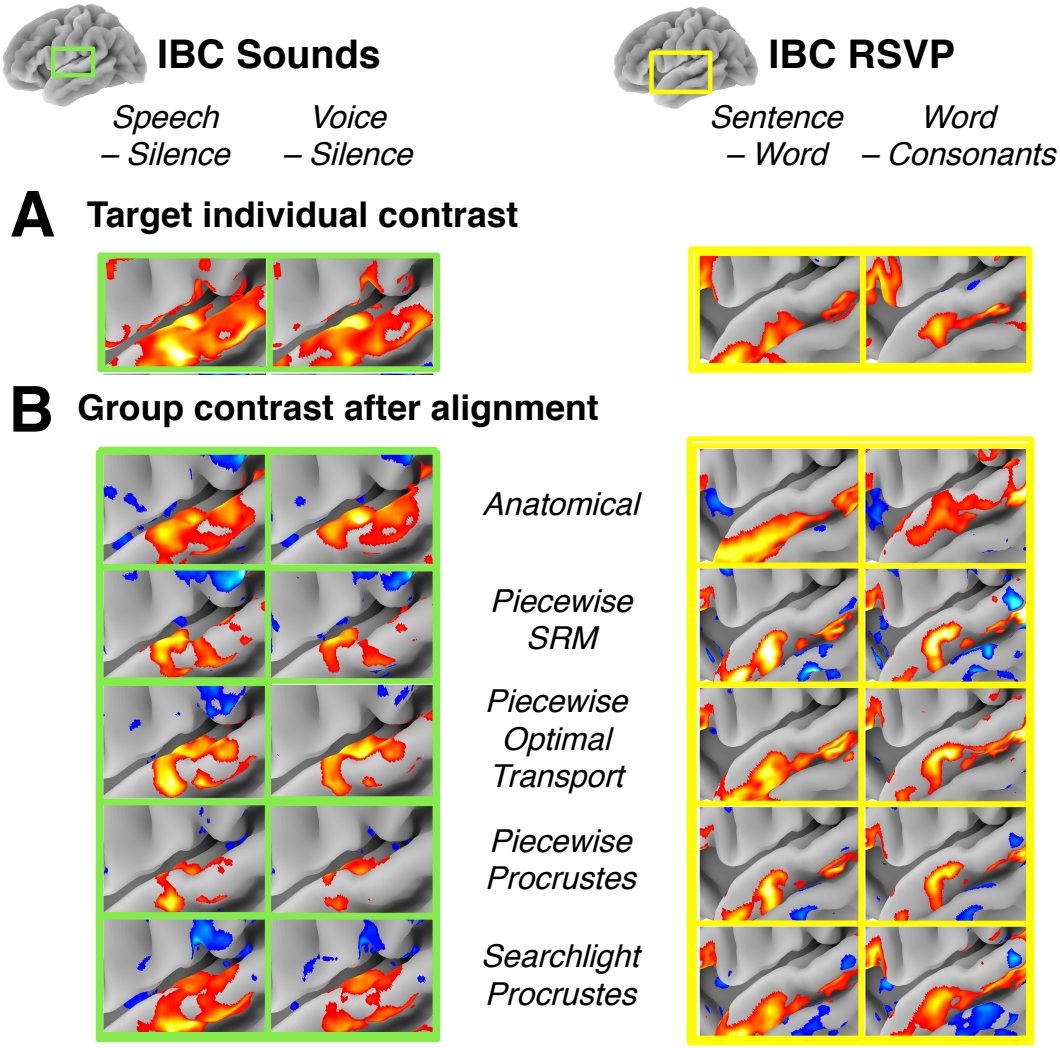

Figure 3.8: **Comparison of alignment methods geometrical effects.** (A) Activation patterns for the Target subject (IBC sub-04) for two contrasts from the IBC Sounds task (*Speech* > *Silence*, *Voice* > *Silence*) and IBC RSVP task (*Sentence* > *Word*, *Word* > *Consonants*). Here, we only show contrast maps from a sub-region of the temporal lobe containing contrast-relevant information. Note that this sub-region differs slightly between the Sounds and RSVP task. (B) Visualization of a group conjunction analysis of all IBC subjects after alignment to the target subject for each of the considered methods. We used a γ value of 0.25 in the group conjunction analysis, which corresponds to at least 25% of the IBC sample showing activation in this temporal region after alignment. For ease of comparison, the colorbar for each contrast and method was scaled to show the full range of values (i.e., the colorbar spans different interval across methods and contrasts) and so is not included here. All displayed maps were thresholded at 1/3 of their maximum value. We see that functional alignment yields stronger contrasts overall when compared to anatomical alignment. Piecewise Procrustes and piecewise Optimal Transport yield less smooth representations, better preserving signal specificity.

3.6 Discussion

In this work, we have proposed a new procedure to measure the information recovered through functional alignment using inter-subject decoding, and we subsequently used this framework to benchmark five functional alignment methods on five distinct decoding tasks across four publicly available datasets.

In general, we find that functional alignment improves inter-subject decoding accuracy in both whole-brain and ROI settings. These results, combined with our qualitative visualization of the effects of functional alignment on signal structure, suggest that functional alignment improves inter-subject correspondence while matching signal to realistic functional topographies. This finding extends and supports conclusions from earlier work (Güçlü and Gerven, 2015; Guntupalli et al., 2016).

At a whole-brain scale, the best performing methods are piecewise SRM, piecewise Optimal Transport, and piecewise Procrustes which each bring 5% improvement over baseline on average. As the baseline inter-subject decoding accuracy is roughly 20% above chance across datasets (Table S3.1), this 5% increase represents a substantial improvement. We also note that this represents recovering more than half of the accuracy lost to inter-subject variability.

The considered functional alignment methods also improve decoding performance when applied *without* an aggregation scheme (i.e., piecewise or searchlight aggregation) within task-relevant ROIs. Here, Optimal Transport and SRM bring 5% and 3% improvement in inter-subject decoding accuracy, respectively, over a baseline accuracy which is on average 10-15% above chance across datasets (Table S3.2).

From our control analyses, we observe that these increases in decoding accuracy were reliably greater than the effect of Gaussian smoothing (see section S3.14). In a minimalistic replication, this effect seems to hold for both volumetric and surface data and at different parcellation resolutions (see section S3.15; cf. Oosterhof et al., 2011).

Our benchmark also brings new evidence that the latent correspondences that can be learnt between different tasks display less inter-individual variability than the task-specific activation maps (Tavor et al., 2016). *Experiment 1* indeed showed that such correspondences could even be used at a whole-brain scale to transfer signals subjects to solve an inter-subject decoding problem, which is—to the best of our knowledge—an original experimental result. By releasing efficient and accessible implementations of these methods in the `fmralign` package, we hope to facilitate future cognitive neuroscience research using functional alignment methods.

3.6.1 Combining local alignment models

Across datasets, we find that the aggregation scheme for alignment significantly affects subsequent performance. Notably, piecewise Procrustes outperforms searchlight Procrustes, both in terms of accuracy as well as computational performance. The methodological difference between these aggregation schemes is whether alignment transformations are learnt within overlapping neighborhoods (as in searchlight Procrustes) or not (as in piecewise Procrustes). Searchlight alignment suffers in that the overlap between searchlights requires multiple computations for a given neighborhood, and the aggregated transformation is no longer guaranteed to reflect properties of the original transforms, e.g. orthogonality. Although piecewise aggregation may theoretically introduce discontinuities at parcel boundaries, in our results we do not find evidence of this effect and indeed find that piecewise aggregation overall benefits decoding performance. Importantly, we found that the improved performance of piecewise Procrustes was largely insensitive to parcel size and definition (see Figure S3.2).

3.6.2 Evaluating alignment performance with decoding

We use inter-subject decoding to quantify the amount of information recovered by functional alignment methods. In general, identifying publicly available datasets with tasks appropriate for both inter-subject decoding as well as functional alignment remains a challenge. Beyond the four datasets included in these results, we investigated several other publicly available datasets such as the Neuroimaging Analysis Replication and Prediction Study (NARPS; Botvinik-Nezer et al., 2020),the Healthy Brain Network Serial Scanning Initiative (HBN-SSI; O'Connor et al., 2017), the interTVA dataset (Aglieri et al., 2019, available as Openneuro *ds001771*) and the Dual Mechanisms of Cognitive Control Project (DMCC, Etzel et al., 2021).

We had difficulties in achieving sufficient baseline accuracy levels in these and other datasets, and we therefore chose not to include them in the present study. This suggests that the amount of signal discriminating complex experimental conditions is not strong enough to find inter-subject patterns robust to variability in many publicly available datasets, likely due to limited sample sizes and suboptimal experimental designs. We hope that broader recognition of the benefits of using inter-subject decoding to uncover neural coding principles across subjects—using functional alignment if necessary—will encourage investigators to collect and share more datasets supporting this type of analysis.

Greater data availability will encourage robust, principled comparisons of alignment methods and foster progress in the field.

3.6.3 Study limitations and future directions

Although our study provides a broad evaluation of the performance of several functional alignment methods, there are several dimensions which we hope future work will better address. Notably, we did not thoroughly investigate how alignment performance is impacted by image resolution and whether data are represented on the surface or the volume. Using volumetric images downsampled to a standard resolution of 3mm isotropic enabled us to make fair comparisons across datasets at a reasonable computational cost. We also show in Section S3.15 that results from piecewise Procrustes alignment on the IBC dataset hold in a higher resolution, surface-based setting. Nonetheless, other functional alignment methods might show different patterns of performance in this setting or at different resolution levels. Moreover, applying these methods on high-resolution images is an exciting perspective to better understand how brain function details vary across subjects. To progress in this direction, a stronger focus on developing computationally efficient methods will be needed. The use of high-resolution parcellations—combined with more efficient implementations of piecewise Optimal Transport or a piecewise Shared Response Model—seem to be particularly promising directions.

We have not examined either the impact of alignment data on the learnt transformations or whether this impact varies across cortex. That is, we could further ask whether certain kinds of stimuli may produce more accurate functional alignments for specialized functional regions. In general, the surveyed functional alignment methods view each subject alignment image as a sample, and the resulting transformation is trained to match corresponding samples across subjects. If some training images lack stable signal in a given ROI, functional alignment methods are unlikely to learn meaningful transformations in this region. Finally this benchmark largely focused on pairwise alignment models. Template-based models—beyond latent factor models as SRM—are an important area of research to further improve the usability of functional alignment methods, particularly in research settings with a large number of subjects. In future work, we intend to address the above questions to learn more about when functional alignment methods are most appropriate.

3.7 Conclusion

In the present work, we have provided an extensive benchmark of five popular functional alignment methods across five unique experimental tasks from four publicly available datasets. Assessing each method in an inter-subject decoding framework, we show that both Shared Response Modelling (SRM) and Optimal Transport perform well at a region-of-interest level of analysis, as well as at the whole-brain scale when aggregated through a piecewise scheme. Our results support previous work proposing functional alignment to improve across-subject comparisons, while providing nuance that some alignment methods may be most appropriate for a given research question. We further suggest that identified improvements in inter-subject decoding demonstrate the potential of functional alignment to identify generalizable neural coding principles across subjects.

3.8 Bibliography

Abraham, A. et al. (2014a). "Machine learning for neuroimaging with scikit-learn". *Front. Neuroinform.*, 8, p. 14.

Abraham, A. et al. (2014b). "Machine learning for neuroimaging with scikit-learn". *Frontiers in Neuroinformatics*, 8.

Aglieri, V. et al. (2019). *InterTVA. A multimodal MRI dataset for the study of inter-individual differences in voice perception and identification*. https://openneuro.org/datasets/ds001771/versions/1.0.2.

Avants, B. et al. (2008). "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain". *Medical Image Analysis*, 12(1), pp. 26–41.

Bazeille, T et al. (2019). "Local Optimal Transport for Functional Brain Template Estimation". In: *Information Processing in Medical Imaging*. Springer International Publishing, pp. 237–248.

Bazeille, T. et al. (2021). "An empirical evaluation of functional alignment using inter-subject decoding". *Neuroimage*, p. 118683.

Bellec, P. et al. (2010). "Multi-level bootstrap analysis of stable clusters in resting-state fMRI". *NeuroImage*, 51(3), pp. 1126 –1139.

Bilenko, N. Y. and J. L. Gallant (2016). "Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging". *Front. Neuroinform.*, 10, p. 49.

Botvinik-Nezer, R. et al. (2020). "Variability in the analysis of a single neuroimaging dataset by many teams". *Nature*, pp. 1–7.

Boyle, J. A. et al. (2020). *The Courtois project on neuronal modelling: 2020 data release*. https://docs.cneuromod.ca. Presented at the 26th annual meeting of the Organization for Human Brain Mapping.

Buckner, R. L. et al. (2011). "The organization of the human cerebellum estimated by intrinsic functional connectivity". *Journal of Neurophysiology*, 106(5). PMID: 21795627, pp. 2322–2345.

Chang, L. J. et al. (2015). "A Sensitive and Specific Neural Signature for Picture-Induced Negative Affect". *PLoS Biol.*, 13(6), e1002180.

Chang, N. et al. (2019). "BOLD5000, a public fMRI dataset while viewing 5000 visual images". *Sci Data*, 6(1), p. 49.

Chen, P.-H. et al. (2015). "A Reduced-Dimension fMRI Shared Response Model". In: *Advances in Neural Information Processing Systems 28*. Ed. by C Cortes et al. Curran Associates, Inc., pp. 460–468.

Churchland, P. M. (1998). "Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered". *J. Philos.*, 95(1), pp. 5–32.

Coalson, T. S., D. C. Van Essen, and M. F. Glasser (2018). "The impact of traditional neuroimaging methods on the spatial localization of cortical areas". *Proceedings of the National Academy of Sciences*, 115(27), E6356–E6365.

Cox, R. W. and J. S. Hyde (1997). "Software tools for analysis and visualization of fMRI data". *NMR in Biomedicine*, 10(4-5), pp. 171–178.

Cuturi, M. (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. J. C. Burges et al. Curran Associates, Inc., pp. 2292–2300.

Dockès, J. et al. (2020). "NeuroQuery, comprehensive meta-analysis of human brain mapping". *eLife*, 9. Ed. by C. Büchel, T. Yeo, and T. D. Wager, e53385.

Esteban, O. et al. (2018a). "fMRIPrep". *Software*.

Esteban, O. et al. (2018b). "fMRIPrep: a robust preprocessing pipeline for functional MRI". *Nature Methods*.

Esteban, O. et al. (2019). "fMRIPrep: a robust preprocessing pipeline for functional MRI". *Nat. Methods*, 16(1), pp. 111–116.

Etzel, J. A. et al. (2021). "The Dual Mechanisms of Cognitive Control dataset: A theoretically-guided within-subject task fMRI battery". *bioRxiv*.

Fonov, V. et al. (2009). "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". *NeuroImage*, 47, Supplement 1, S102.

Gorgolewski, K. et al. (2011). "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python". *Frontiers in Neuroinformatics*, 5, p. 13.

Gorgolewski, K. J. et al. (2018). "Nipype". *Software*.

Greve, D. N. and B. Fischl (2009). "Accurate and robust brain image alignment using boundary-based registration". *NeuroImage*, 48(1), pp. 63–72.

Güçlü, U. and M. A. van Gerven (2015). "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream". *Journal of Neuroscience*, 35(27), pp. 10005–10014.

Guntupalli, J. S., M. Feilong, and J. V. Haxby (2018). "A computational model of shared fine-scale structure in the human connectome". *PLoS Comput. Biol.*, 14(4), e1006120.

Guntupalli, J. S. et al. (2016). "A Model of Representational Spaces in Human Cortex". *Cereb. Cortex*, 26(6), pp. 2919–2934.

Hanke, M. et al. (2009). "PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data". *Neuroinformatics*, 7(1), pp. 37–53.

Hanke, M. et al. (2016). "A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation". *Sci Data*, 3, p. 160092.

Haxby, J. V. et al. (2011). "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron*, 72(2), pp. 404–416.

Haxby, J. V. et al. (2020). "Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies". *Elife*, 9, e56601.

Heller, R. et al. (2007). "Conjunction group analysis: An alternative to mixed/random effect analysis". *NeuroImage*, 37, pp. 1178–85.

Jenkinson, M. et al. (2002). "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images". *NeuroImage*, 17(2), pp. 825–841.

Jiahui, G. et al. (2020). "Predicting individual face-selective topography using naturalistic stimuli". *Neuroimage*, 216, p. 116458.

Kriegeskorte, N. and J. Diedrichsen (2019). "Peeling the Onion of Brain Representations". *Annu. Rev. Neurosci.*, 42(1), pp. 407–432.

Kriegeskorte, N., R. Goebel, and P. Bandettini (2006). "Information-based functional brain mapping". *Proc. Natl. Acad. Sci. U. S. A.*, 103(10), pp. 3863–3868.

Kriegeskorte, N., M. Mur, and P. A. Bandettini (2008). "Representational similarity analysis-connecting the branches of systems neuroscience". *Frontiers in systems neuroscience*, 2, p. 4.

Lanczos, C. (1964). "Evaluation of Noisy Data". *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1), pp. 76–85.

Langs, G. et al. (2010). "Functional Geometry Alignment and Localization of Brain Areas". *Adv. Neural Inf. Process. Syst.*, 1, pp. 1225–1233.

Langs, G. et al. (2014). "Decoupling function and anatomy in atlases of functional connectivity patterns: language mapping in tumor patients". *Neuroimage*, 103, pp. 462–475.

Nastase, S. A. et al. (2019). *Measuring shared responses across subjects using intersubject correlation*.

Olivetti, E., S. Veeramachaneni, and P. Avesani (2011). "Testing for Information with Brain Decoding". In: *2011 International Workshop on Pattern Recognition in NeuroImaging*, pp. 33–36.

Oosterhof, N. N. et al. (2011). "A comparison of volume-based and surface-based multivoxel pattern analysis". *Neuroimage*, 56(2), pp. 593–600.

O'Connor, D. et al. (2017). "The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions". *Gigascience*, 6(2), giw011.

Paquola, C. et al. (2019). "Microstructural and functional gradients are increasingly dissociated in transmodal cortices". *PLOS Biology*, 17(5), e3000284.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Pinho, A. L. et al. (2018). "Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping". *Sci Data*, 5, p. 180105.

Poldrack, R. A. et al. (2013). "Toward open sharing of task-based fMRI data: the OpenfMRI project". *Frontiers in neuroinformatics*, 7, p. 12.

Rademacher, J et al. (1993). "Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology". *Cereb. Cortex*, 3(4), pp. 313–329.

Richard, H. et al. (2019). "Fast shared response model for fMRI data". *arXiv preprint arXiv:1909.12537*.

Sabuncu, M et al. (2010). "Function-based intersubject alignment of human cortical anatomy". *Cerebral Cortex*, 20, pp. 130–140.

Schaefer, A. et al. (2018). "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI". *Cereb. Cortex*, 28(9), pp. 3095–3114.

Schrouff, J. et al. (2018). "Embedding Anatomical or Functional Knowledge in Whole-Brain Multiple Kernel Learning Models". *Neuroinformatics*, 16(1), pp. 117–143.

Tavor, I et al. (2016). "Task-free MRI predicts individual differences in brain activity during task performance". *Science*, 352(6282), pp. 216–220.

Thirion, B. et al. (2006). "Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets". *Hum. Brain Mapp.*, 27(8), pp. 678–693.

Tustison, N. J. et al. (2010). "N4ITK: Improved N3 Bias Correction". *IEEE Transactions on Medical Imaging*, 29(6), pp. 1310–1320.

Varoquaux, G. et al. (2011). "Multi-subject dictionary learning to segment an atlas of brain spontaneous activity". In: *Information Processing in Medical Imaging*. Vol. 6801. Lecture Notes in Computer Science. Gábor Székely, Horst Hahn. Kaufbeuren, Germany: Springer, pp. 562–573.

Vázquez-Rodríguez, B. et al. (2019). "Gradients of structure–function tethering across neocortex". *Proceedings of the National Academy of Sciences*, 116(42), pp. 21219–21227.

Virtanen, P. et al. (2020). "SciPy 1.0: fundamental algorithms for scientific computing in Python". *Nature methods*, 17(3), pp. 261–272.

Xu, T., M. Yousefnezhad, and D. Zhang (2018). "Gradient Hyperalignment for multi-subject fMRI data alignment". In: *Pacific Rim International Conference on Artificial Intelligence*. Springer, pp. 1058–1068.

Yousefnezhad, M. and D. Zhang (2017). "Deep Hyperalignment". In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 1604–1612.

Zhang, Y., M. Brady, and S. Smith (2001). "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm". *IEEE Transactions on Medical Imaging*, 20(1), pp. 45–57.

S3.9 fMRIPrep preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.1.1+38.g8480eabb (Esteban et al., 2018b; Esteban et al., 2018a; RRID:SCR_016216), which is based on *Nipype* 1.5.0 (Gorgolewski et al., 2011; Gorgolewski et al., 2018; RRID:SCR_002502).

S3.9.1 Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

S3.9.2 Functional data preprocessing

For each subject's BOLD runs (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated by aligning and averaging 1 single-band references (SBRefs). A B0-nonuniformity map (or *fieldmap*) was estimated based on two (or more) echo-planar imaging (EPI) references with opposing phase-encoding directions, with `3dQwarp` Cox and Hyde (1997) (AFNI 20160207). Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters

with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002).

First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*.

All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964).

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014b, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

S3.9.3 Copyright Waiver

The above boilerplate text was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts *unchanged*. It is released under the CC0 license.

S3.10 Absolute decoding accuracy of various methods

Tables S3.1 and S3.2 report absolute decoding accuracies for *Experiment 1* and *Experiment 2*, to bring a different view of results presented in Figures 3.5 and 3.7, as relative improvements brought over anatomical registration by various alignment methods. This "per dataset view" highlight that gains brought by best methods are substantial improvement over baseline, especially when compared to chance.

Methods/Dataset	IBC RSVP	IBC Sounds	Forrest	BOLD5000	Neuromod
Chance	16.7	16.7	20	25.5	12.5
Anatomical	38.2 ± 4.1	32.7 ±4.8	31.4 ± 4.7	33.3 ± 2.9	54.1 ± 6.7
Intra-subject	39.6 ± 3.1	36.4 ± 8.6	31.9 ± 5.4	34.9 ± 2.8	55.6 ± 7.7
Searchlight Procrustes	39.0 ± 5.1	32.6 ± 6.5	**33.6** ± 6.1	35.2 ± 2.0	65.5 ± 6.6
Piecewise Procrustes	42.0 ± 4.7	36.6 ± 5.5	**33.8** ± 6.4	36.4 ± 1.8	**67.4** ± 10.6
Piecewise Optimal Transport	**43.5** ± 5.5	**38.0** ± 9.5	**33.8** ± 5.6	36.6 ± 2.1	65.3 ± 9.1
Piecewise Shared Response Model	42.4 ± 4.0	37.0 ± 6.8	**33.7** ± 7.2	**39.6** ± 2.9	66.2 ± 7.0

Table S3.1: **Fullbrain benchmark absolute decoding accuracy (%).**

Methods/Dataset	IBC RSVP	IBC Sounds	Forrest	BOLD5000	Neuromod
Chance	16.7	16.7	20	25.5	12.5
Anatomical	22.3 ± 3.3	26.9 ± 6.9	28.6 ± 6.0	33.5 ± 3.3	50.2 ± 5.3
Intra-subject	22.0 ± 1.7	25.6 ± 2.7	27.5 ± 3.2	38.0 ± 2.6	51.4 ± 10.2
Procrustes	**31.0** ± 4.3	**32.7** ± 8.3	30.1 ± 7.0	30.1 ± 2.2	46.5 ± 7.8
Optimal Transport	24.9 ± 2.5	29.9 ± 5.4	28.9 ± 2.2	39.7 ± 2.7	**64.4** ± 8.9
Shared Response Model	30.4 ± 4.4	30.9 ± 6.3	**34.3** ± 5.2	**40.9** ± 3.6	32.6 ± 5.5

Table S3.2: **ROI benchmark absolute decoding accuracy (%).**

S3.11 Whole-brain decoding provides better accuracy than ROI-based decoding

In Figure S3.1, we compare ROI-based and whole-brain inter-subject decoding accuracy improvements for piecewise Procrustes alignment above anatomical-only alignment. We see that whole-brain alignment generally shows higher inter-subject decoding improvements compared to ROI-based alignment. As mentioned in the main text, this result supports previous work from the inter-subject decoding literature (Chang et al., 2015; Schrouff et al., 2018), and it suggests that full-brain piecewise alignment yields the best overall decoding pipeline.

S3.12 Parcellation has limited impact on decoding accuracy

To assess the impact of the parcellation used on piecewise alignment results, we compared decoding accuracy gains while varying the parcellation kind and resolution. First, we consider the multi-resolution Schaefer et al. (2018) atlas, which was learnt using a gradient weighted markov random field on resting state data from 1489 subjects. We compare this *a priori* parcellation to two parcellations learnt directly on the subject's alignment data after 5mm FWHM Gaussian smoothing: K-means or Hierarchical K-means. All these parcellations were taken at ten resolutions from 100 to 1000 parcels.

As hierarchical K-means may be less familiar to readers, we briefly describe it in more detail here. This method is a variant of K-means aimed specifically at obtaining more balanced parcels. To identify k parcels, we first apply K-means to cluster the voxels in $\sqrt{k}$ big clusters. Each of these "big clusters" is then clustered again in $\sqrt{k}$ to obtain a total of

k smaller well-balanced parcels. In this experiment, K-means and Hierarchical K-means implementations used are respectively from `scikit-learn` and `fmralign`, and fitted as part of `fmralign` alignment functions on the source subject data.

We plot piecewise Procustes accuracy improvements for these three parcellation methods and ten resolutions in Figure S3.2. Here, we only show the IBC Sounds and IBC RSVP decoding tasks to ease in interpretation. Overall, we observe on these two tasks that the type and resolution of parcellation used does not have a strong impact on accuracy improvements above anatomical-only alignment. We therefore suggest that *piecewise alignment* can be used with confidence that the parcellation choice won't strongly impact its results.

S3.13 Grid-search of Piecewise SRM hyperparameters

As a piecewise implementation of SRM is a novel contribution from this work, we had no prior knowledge on how to properly set hyperparameters from this methods (clustering type and resolution as well as the number of components to set for each parcelwise SRM).

For the type of clustering, we limited ourselves to a pre-computed parcellation (the Schaefer atlas) available at various resolutions. This is based on the intuitions acquired on Procrustes (see Section S3.12) that the parcellation type was not of utmost importance to decoding results. We ran a cross-validation on the two remaining parameters. We used Schaefer atlas at resolution : [100,300,500,700] while our number of components ranged in [5,25,35,50]. Figure S3.3 present the results of this cross-validation, that led us to chose Schaefer atlas 700 and 50 components as hyper parameters for our main experiments.

S3.14 Functional alignment is not merely smoothing

Gaussian smoothing is a common preprocessing step in neuroimaging group studies, which reconciles dissimilar subject-level signals by smoothing over inter-individual variability. Our qualitative results (section 3.5.3) show that best performing alignment methods do not seem to smooth the signal across voxels, but instead preserve the signal specificity while matching its geometry with the target subject functional topography. Specifically, we compared decoding gains from six different Gaussian smoothing kernels to those obtained through the reference method piecewise Procrustes alignment.

The results displayed in Figure S3.4 clearly support previous findings (Guntupalli et al., 2016) that smoothing does not improve inter-subject decoding performance—and therefore recover mutual information—in the same way as functional alignment.

S3.15 Impact of the data representation and resolution

Oosterhof et al. 2011 argued that functional alignment benefits from working with a representation of the fMRI signal on the cortical surface (Coalson et al., 2018). Relatedly, we would also expect that the resolution of the data representation—whether in the surface or the volume—will impact the quality of the alignment learnt.

To assess the dependence of our 3mm volumetric results presented in the main text on sampling parameters, we replicated our inter-subject decoding framework with the IBC RSVP language task data on a high-resolution cortical surface representation(*fsaverage7*)(obtained through `freesurfer` surface projection of full-resolution raw images in their respective subject space, later on mapped to the common surface template). This surface mesh includes 168k cortical nodes per hemisphere, which we divided into 350 parcels per hemisphere using *Schaefer* atlas at scale 700.

We provide results for the inter-subject decoding accuracy gains seen with the reference functional alignment method of piecewise Procrustes over standard, anatomical-only alignment. We had to limit to this setting because (i) replicating this analysis on every dataset would represent an important amount of processing work, and (ii) working on other methods than piecewise Procrustes on this very large data is computationally prohibitive.

The results displayed in Figure S3.5 show that although decoding gains are a little higher using high-resolution surface-based representation, they remain in the same range as the volume-based representation. This shows that a 10-fold higher resolution can help match more precisely topographies across subjects (and reduce the decoding variance as a consequence), but no important marginal gains can be expected from it. In the end the signal available for use is bounded by the same rough limitations: test-retest reliability in each subject.

S3.16 IBC alignment data explained

In this work, 53 contrasts were pulled together are used as alignment for IBC dataset. The contrasts are common to all subjects and taken from both HCP and ARCHI protocol.

In order they are labelled : *audio left button press, audio right button press, video left button press, video right button press, horizontal checkerboard, vertical checkerboard, audio sentence, video sentence, audio computation, video computation, saccades, rotation hand, rotation side, object grasp, object orientation, mechanistic audio, mechanistic video, triangle mental, triangle random, false belief audio, false belief video, speech sound, non speech sound, face gender, face control, face trusty, expression intention, expression gender, expression control, shape, face, punishment, reward, left hand, right hand, left foot, right foot, tongue, cue, story, math,,relational, match, mental, random, 0back body, 2back body, 0back face, 2back face, 0back tools, 2back tools, 0back place, 2back place*.

To know more, please visit the relevant IBC documentation.

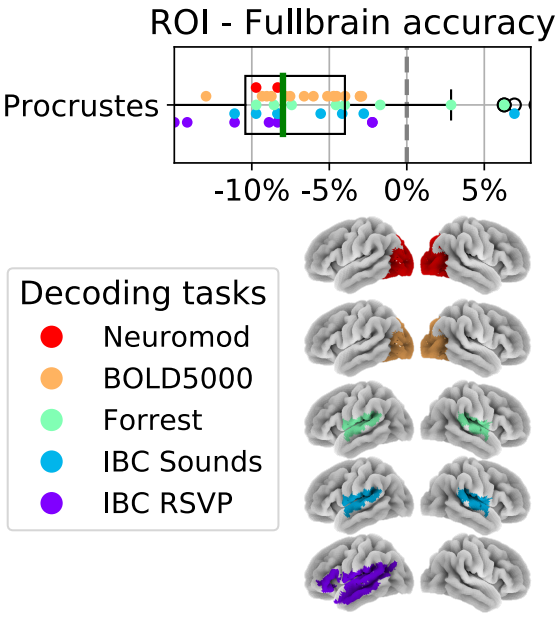

Figure S3.1: **Comparing ROI and whole-brain decoding accuracy after piecewise Procrustes alignment.** The ROIs used for each dataset are displayed on the *lower panel*. In the *upper panel*, we show the distribution of differences in decoding accuracy scores between ROI-based and whole-brain piecewise Procrustes alignment. Each dot represents a single subject, and subjects are colored according to their decoding task. Each difference score is calculated by subtracting the inter-subject decoding accuracy for whole-brain piecewise Procrustes alignment from the ROI-based piecewise Procrustes alignment accuracy score—for the same leave-one-subject-out cross-validation fold. The box plot thus describes the distribution of differences, where the green line represents the median value. We see that decoding accuracy is lower when performed within ROIs than when performed on the whole-brain data.

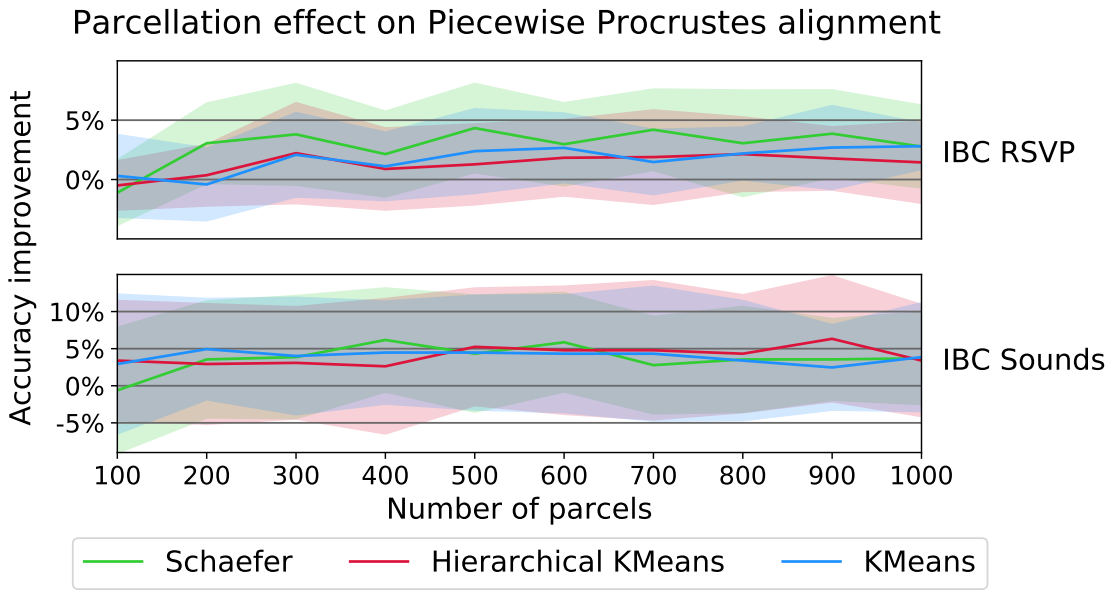

Figure S3.2: **Effect of parcellation type, resolution on Piecewise Procrustes decoding accuracy improvement over anatomical alignment.** We consider the impact of parcellation type (the *a priori* Schaefer atlas or learned directly on the data with k-means or hierarchical k-means) and resolution (from 100 to 1000 parcels). Results are shown for the IBC RSVP and IBC Sounds decoding tasks. Each line represents the average accuracy improvement for piecewise Procruses over standard, anatomical-only alignment, and the confidence band represents the range of accuracy improvements seen across all IBC subjects. Accuracy improvements are calculated by subtracting anatomical-only inter-subject decoding accuracy scores for the same leave-one-subject-out cross-validation fold. We see that parcellation type and resolution show limited impact on accuracy gains.

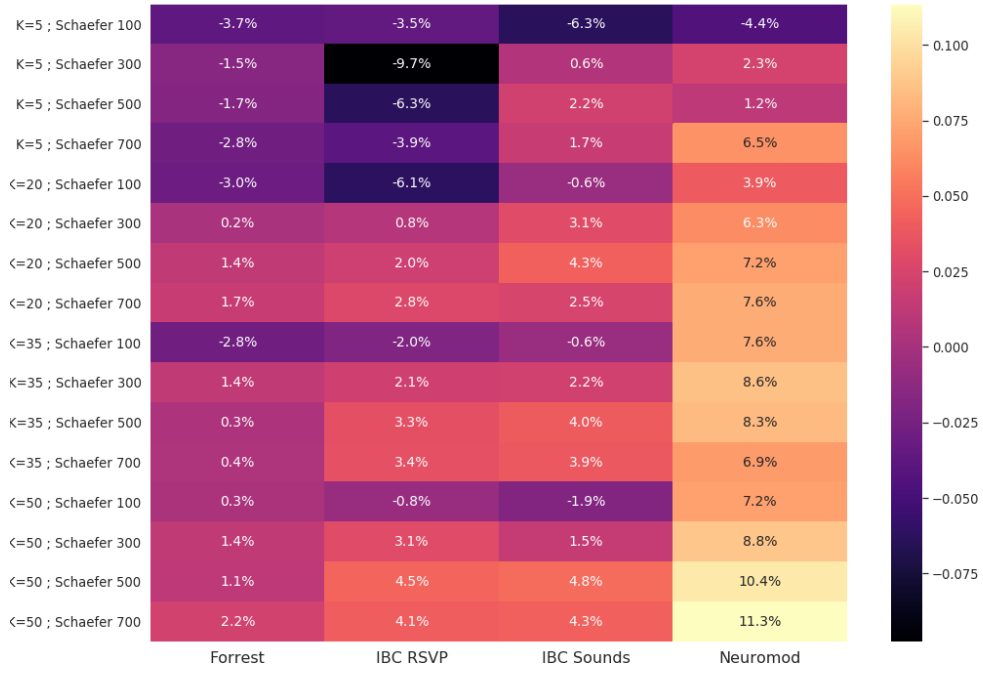

Figure S3.3: **Grid search of Piecewise SRM hyperparameters impact on decoding accuracy across datasets.** We considered a grid of 4 parcellations (Schaefer atlas at resolution 100, 300, 500 and 700) and 4 different values of k (number of SRM components) for each model fitted on a parcel. We ran our inter-subject decoding pipeline on four inter-subject decoding task (in columns, among those used in the main benchmark). We report here the decoding accuracy improvement over anatomical baseline across datasets for each set of parameter (in line). Altough we didn't have the computational means to run an extensive grid-search, we can already conclude that high-resolution parcelations (and thus more fitted local SRMs) yield a higher decoding gain as long as they come with enough component. Decoding accuracy is also positively linked with K, probably up to a plateau that we did not clearly reached with our limited grid. For the main benchmark we retained the last line model ($K = 50$, Schaefer 700).

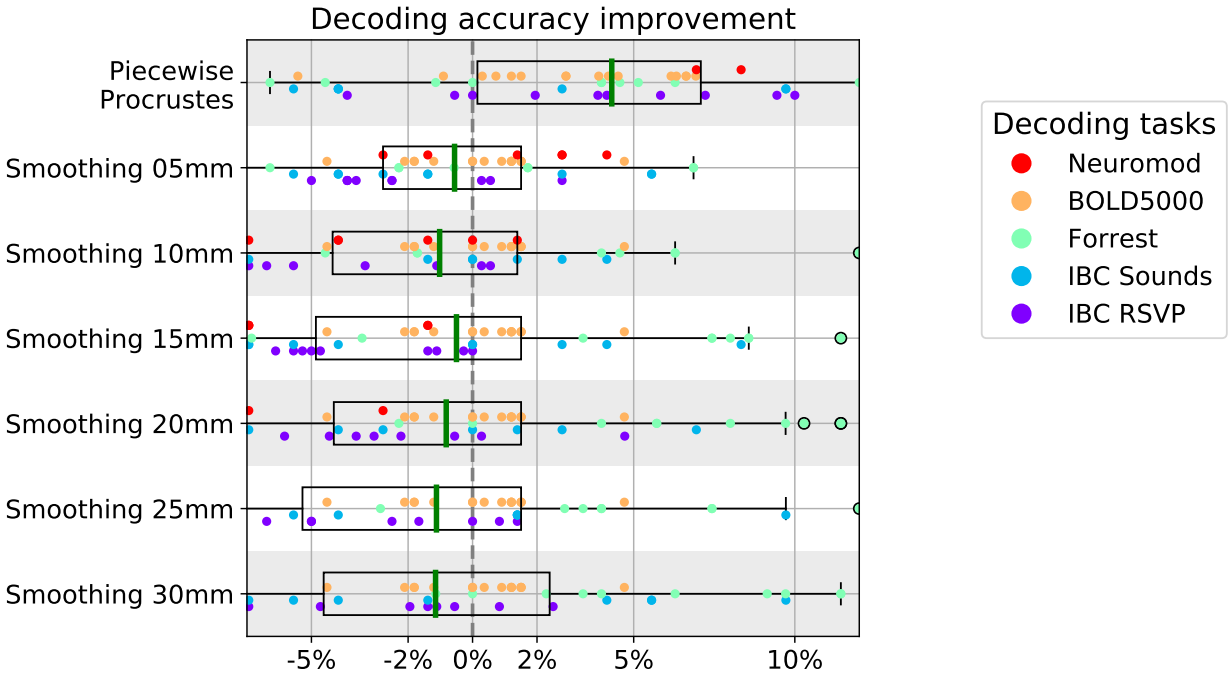

Figure S3.4: **Decoding accuracy does not improve after Gaussian smoothing over anatomical alignment.** For six smoothing kernels, we show inter-subject decoding accuracy scores after subtracting anatomical-only inter-subject decoding accuracy for the same leave-one-subject-out cross-validation fold. Each dot represents a single subject, and subjects are colored according to their decoding task. We also show differences in decoding accuracy scores for the reference functional alignment method piecewise Procrustes, again as compared to anatomical-only alignment. Each box plot describes the distribution of values for that smoothing kernel or alignment method, and the green line indicates the median. We see that Gaussian smoothing does not show the same pattern of decoding accuracy differences as the reference functional alignment method.

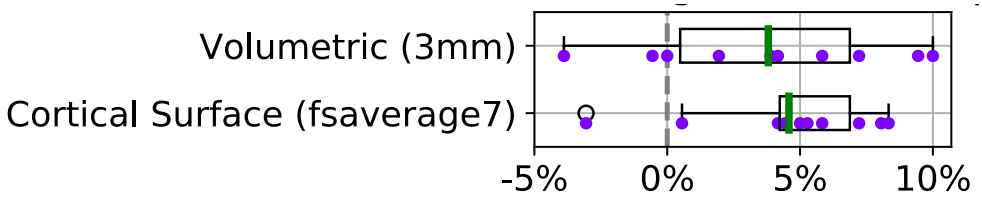

Figure S3.5: **Comparing piecewise Procrustes accuracy improvements across volumetric and surface data representations.** For the IBC RSVP task, we compare piecewise Procrustes decoding accuracy scores to anatomical-only alignment. Each dot represents an IBC subject, where their difference score is calculated by subtracting the inter-subject decoding accuracy for anatomical-only alignment from the piecewise Procrustes alignment accuracy score for the same leave-one-subject-out cross-validation fold; i.e., where they are the left-out subject. We compare these difference scores as calculated using data in the volume (3mm resolution), to data on the high-resolution cortical surface (*fsaverage7*). Each box plot describes the distribution of values for that data representation, and the green line indicates the median. We see that the high-resolution surface representation yields a moderate gain of decoding accuracy, compared to 3mm isotropic volumetric representation.

Chapter 4

When is functional alignment useful? Examining the impact of experimental context

Elizabeth DuPre[1,†], Thomas Bazeille[2,†], Bertrand Thirion[2], Jean-Baptise Poline[1]

[1]Montréal Neurological Institute, McGill University, Montréal, Canada

[2]Université Paris-Saclay, Inria, CEA, Palaiseau, 91120, France

†These authors contributed equally to this work.

4.1 Preface

While the work presented in Chapter 3 provides important insight into the relative performance of available functional alignment algorithms, it does not explore other factors relevant for cognitive neuroscience applications. For example: What data should I use to learn functional alignment applications? How can I assess whether functional alignment has effectively improved inter-subject similarity? These other experimental dimensions are often most pressing for investigators. In this chapter, I characterize three experimental dimensions relevant for cognitive neuroscience research: the data used to train functional alignment transformations, the data on which learned transformations are applied, and the metric by which functional alignment performance is assessed. By comparing these dimensions across two publicly available, well-sampled datasets, we assess their relative impacts in a range applications. This project will be submitted for publication in 2022.

4.2 Introduction

Individual differences in cortical anatomy and organization challenge group-level inferences in human brain mapping. Standard neuroimaging reference spaces—such as MNI space—are designed to provide a common coordinate system for mapping structural and functional characteristics across individual brains. In doing so, they necessarily represent a consensus or average anatomy across many participants. In areas where participants may have variable anatomical features (e.g., duplication of Heschel's gyrus; Marie et al. 2015), however, such an average anatomy may not map accurately to any individual subject. Thus, there is no ideal correspondence in these regions between individual anatomy and a reference template.

This concern is not limited to a small number of regions; instead, individual-level variability has emerged as a dominant principle of human brain organization, particularly across cortex. Within primary somatosensory areas, we see significant individual variability in cytoarchitectonically- (Rademacher et al., 1993), topographically- (Marie et al., 2015), and functionally-defined regions (Benson et al., 2021). Higher-order association cortices show still weaker relationships between structure and function (Paquola et al., 2019; Vázquez-Rodríguez et al., 2019), suggesting an even greater variability across individuals in these areas. This weak structure-function mapping may be particularly noticeable in relatively low engagement tasks, such as traditional psychological paradigms (Hasson et al., 2010; Sonkusare et al., 2019).

Resulting challenges for human brain mapping have long been known within the field (Brett et al., 2002; Thirion et al., 2006), and they continue to complicate direct comparisons of functional activation patterns (Bilenko et al., 2010; Michel et al., 2012; Raizada and Connolly, 2012). Traditionally, functional magnetic resonance imaging (fMRI) studies have met this challenge by introducing additional, artificial smoothness through Gaussian blurring. Even as smoothing minimizes the impact of mis-registration across subjects, it may obscure meaningful information (Coalson et al., 2018). Thus, we have gained a significant understanding of cognition at a group-level, but often at the sacrifice of individual-specific mappings.

Recently, however, researchers have become increasingly interested in mapping individual-level organization, as reflected in the rise of novel "deep phenotyping" acquisitions such as the Midnight Scan Club (Gordon et al., 2017) and Individual Brain Charting initiative (Pinho et al., 2018; Pinho et al., 2020). Building on these new, richly sampled datasets, methods that align directly on individual functional activation—rather than solely on

anatomy—have been developed over the past decade. This class of "functional alignment" methods aims to preserve the individual-specific information that is commonly lost in cross-subject comparisons. Beginning with the introduction of *hyperalignment* in Haxby et al. (2011), functional alignment has expanded to a range of algorithms that are increasingly used in cognitive neuroscience research.

In our previous work Bazeille et al. (2021), we benchmarked four of these algorithms to align individual pairs of subjects across a range of publicly available datasets. While algorithm choice outlines the kinds of transformations that can be learned and therefore substantially impacts derived results, experimental context constrains the kinds of data available for alignment transformations. In particular, the relationship between the data in which the alignment is learned and the data to which the alignment is applied—as well as the brain region in which this data is evaluated—all are likely to significantly influence the success of functional alignment. Here, we extend on our previous results to better capture the performance of functional alignment in cognitive neuroscience applications. We consider how a single functional alignment algorithm interacts with a range of experimental factors to assess their relative impacts.

4.2.1 Experimental dimensions

Functional alignment is a complex transformation that relies on at least three experimental factors: (*1*) an alignment stimulus on which to learn the alignment transformation parameters, (*2*) a downstream application on which to apply the learned alignment transformation, and (*3*) a performance metric to evaluate the success of alignment in improving inter-subject similarity. While these three factors cannot be considered entirely independently, they broadly outline most applications of functional alignment in cognitive neuroscience to date. Although a number of studies in the current literature apply functional alignment to fMRI data, they have little overlap in these described experimental factors. For example, in Nastase et al. (2017), the authors learn functional alignment transformations using whole-brain searchlight hyperalignment on *Life* documentary viewing and then apply these transformation parameters to an independent visual attention task. In Chen et al. (2017), by contrast, the authors use the Shared Response Model (SRM; Chen et al., 2015) to learn a shared functional mapping between participants watching BBC's *Sherlock* within a posterior medial cortex region-of-interest. They then examine whether the dimensionality of this mapping affects scene-classification accuracy both within movie-viewing as well as during movie-recall. In each study, all three of the described experimental factors are tailored to the research question, making it difficult to disentangle their relative impacts

on functional alignment performance. Understanding these impacts will help to develop more general guidelines for functional alignment applications.

We thus consider each of these experimental dimensions as important methodological choices for applications of functional alignment. To evaluate these experimental factors, we use two openly-available deep phenotyping datasets, the Courtois Project on Neuronal Modelling (Courtois-NeuroMod; Boyle et al. 2020) and Individual Brain Charting (IBC; Pinho et al. 2018; Pinho et al. 2020) datasets, each of which include a range of alignment stimuli and downstream experimental applications. In directly building on our previous results, we draw on synchronized alignment data across both datasets and and apply linear, pairwise functional alignment algorithms as in Bazeille et al. (2021). Using these richly-sampled datasets, we can systematically compare the impacts of our three experimental dimensions: the data on which functional alignment is learned, the data to which it is applied, and the metric by which alignment success is evaluated. We will provide analytic Python code to re-create all of our analyses, available at https://github.com/neurodatascience/cog-align.

4.3 Methods

In our previous work, we explored the impact of alignment algorithms on inter-subject decoding accuracy across a range of datasets, each with a single alignment stimulus (Bazeille et al., 2021). Here, we build on this work and shift our focus from alignment algorithm to explore the impact of experimental factors. Specifically, we consider (*1*) the alignment data on which a functional alignment transformation is learned, (*2*) the downstream application data on which functional alignment is applied, and (*3*) associated performance metrics such as inter-subject decoding. Throughout, we use the piecewise Procrustes alignment algorithm as our previous results suggest that it performs well across a range of datasets with a low computational cost.

4.3.1 Performance metrics

Across the existing literature, we have identified three metrics used to evaluate the success of functional alignment in improving inter-subject similarity: inter-subject decoding, time segment matching, and spatial inter-subject correlation. A graphical summary of these metrics is available in Figure 4.1, and we briefly review each in turn below.

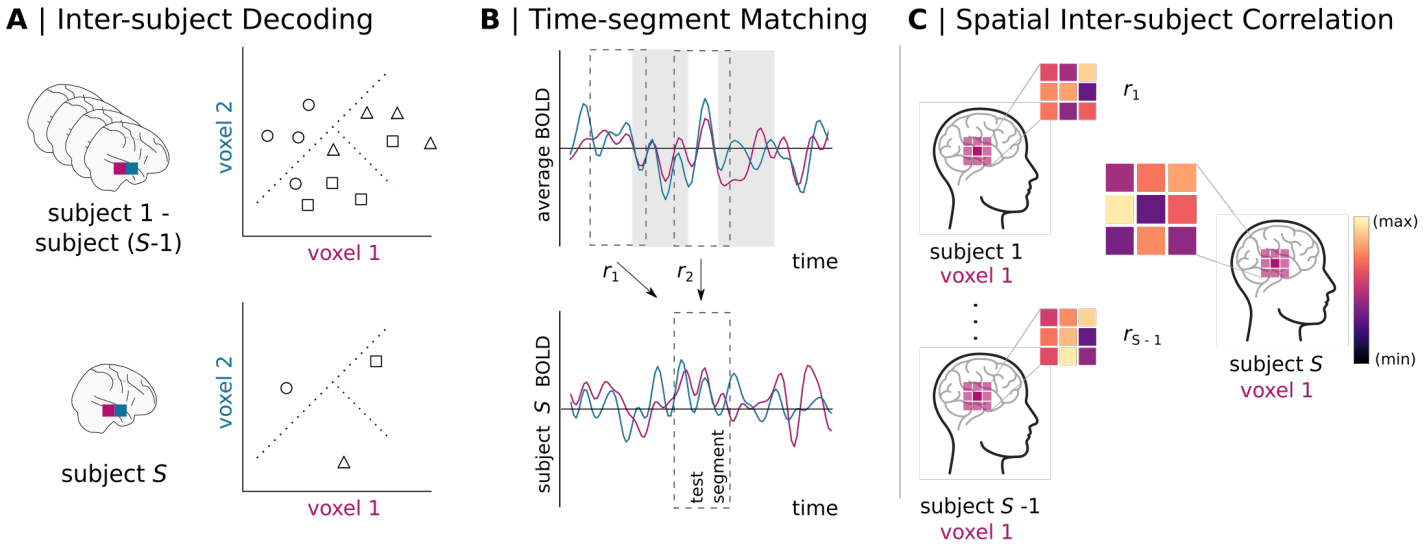

Figure 4.1: **Graphical overview of considered performance metrics.** We focus on two voxels to ease visualization, with voxel 1 depicted in pink and voxel 2 in blue. *Panel A* depicts inter-subject decoding, used on data with clear task labels for supervised learning. In our experiments, we learn a logistic regression classifier on all-but-one subject and then test the trained classifier on the single, held-out subject S. We report the accuracy of the trained classifier on subject S as the accuracy for that inter-subject decoding cross-validation fold. *Panel B* schematizes time-segment matching, which can be used on unlabelled data unfolding over time. Here, we take a sliding window approach to divide a time series into N segments, each containing a pre-defined number of TRs. We correlate each window from our single, held-out subject S with all non-overlapping windows—depicted in grey—from the average time series across all other subjects. We then learn a maximum-correlation classifier to identify each time-segment. The overall accuracy for the time-segment matching cross-validation fold is the average accuracy across all N segments for subject S. In *Panel C*, we show spatial inter-subject correlation (ISC) which can be calculated on both labelled and unlabelled data. This metric differs from the two previous metrics in that we calculate it between all possible pairs of subjects rather than in a leave-one-subject-out fashion. First, we define a set radius around the considered voxel, within which we separately average each subjects' time series to derive their unique spatial pattern. The average of all possible pairwise correlations between a target subject S and all other subjects' spatial patterns is reported as the spatial inter-subject correlation for that voxel in subject S. The same procedure is repeated for all voxels and all target subjects to derive individual spatial ISC maps.

4.3.1.1 Inter-subject decoding

First, and as motivated by our previous work (Bazeille et al., 2021), we use inter-subject decoding in the case of downstream tasks with well-defined labels appropriate for supervised learning. To calculate inter-subject decoding accuracy, we learned a logistic regression classifier with no feature selection on fixed-effects contrast maps from each downstream application (see Section 4.3.4.1 for details on contrast map generation). Classifiers were implemented in scikit-learn (Pedregosa et al., 2011), and decoding accuracy was assessed using a leave-one-subject-out cross-validation scheme. That is, the logistic regression classifier was trained to classify condition labels on all-but-one subject and the resulting trained classier was used without retraining on the held-out subject, providing an accuracy score for that cross-validation fold.

4.3.1.2 Time-segment matching

For downstream tasks without well-defined experimental labels, inter-subject decoding is not supported. This lack of labelled data may occur, for example, when working with naturalistic audio-visual data without annotated features, such as the presence or absence of a face in the video. In this case, time-segment matching—as introduced in Haxby et al., 2011—is a useful adaptation of inter-subject classification. Here, the stimulus is divided into overlapping segments using a sliding window of experimenter-specified window size. In this work, we adopt 30 TR windows, corresponding to one minute of acquisition in IBC and 45 seconds in Courtois-NeuroMod.

For all-but-one subject, we calculate the average activity across subjects during each segment. Iterating through all available time segments, we then calculate the correlation between the held-out subject's activity during that segment and every segment of the average time series of all other subjects, discarding the correlations for segments that overlap with the test segment. Note that overlapping segments are excluded from the correlation to avoid duplicating data between our train and test sets. If the maximum correlation corresponds to the same time segment in the training and held-out subject's data, we consider that time segment to be accurately matched. The overall accuracy for a given subject is then the number of correct matches divided by the total number of time segments. As in inter-subject decoding, we can iterate over each subject to calculate the average accuracy for the given data set.

4.3.1.3 Spatial inter-subject correlation

Finally, spatial inter-subject correlation (spatial ISC; Nastase et al., 2019) can be used to assess the performance of functional alignment both with and without well-labelled data. Spatial ISC is calculated as the correlation between a target subject's and all other subject's spatial patterns at each time point within a pre-defined radius. That is, for each voxel, we calculate the correlation of all voxel activity patterns within a sphere of some pre-specified radius centered on that voxel, and we repeat this procedure for each time point. In this work, we use a radius of 5mm resulting in a sphere of approximately 20 voxels. By averaging across all subjects, we can derive a single spatial ISC value for a given target subject's voxel, repeating the procedure for all voxels and target subjects. Because spatial ISC as calculated for each voxel, we can also derive voxelwise spatial ISC maps showing the distribution of spatial ISC values over the cortex. Note that in the case of labelled task data—such as fixed-effects contrast maps—each test sample corresponds to a single contrast map.

4.3.2 Datasets and preprocessing

To rigorously investigate the impacts of alignment stimulus and downstream application, we focus on two datasets which include extensive characterization of participant responses across a range of experimental conditions. Specifically, we use 10 subjects from the Individual Brain Charting (IBC; Pinho et al. 2018) and six subjects from the Courtois Project on Neuronal Modelling (Courtois-NeuroMod; Boyle et al. 2020) datasets. Courtois-NeuroMod was preprocessed with fMRIPrep LTS v20.2 (Esteban et al., 2018b); a complete description of the fMRIPrep preprocessing procedures is included in the appendix (Section S4.8). IBC was preprocessed using an SPM-based pipeline as described in (Pinho et al., 2018).

To reduce the computational cost of functional alignment, we downsampled all included datasets to 3mm isotropic resolution. For both datasets, preprocessed data were masked using a grey matter mask, corrected for the six motion regressors and 10 high-variance (i.e., CompCorr) components, and smoothed with a 5mm Gaussian kernel using Nilearn (Abraham et al., 2014). Audio-visual alignment data were additionally trimmed to approximately 2000 time frames such that stimulus-type was not confounded with stimulus-length.

4.3.3 Data description

Here we review the exact data included across our experiments to first learn and then to apply functional alignment transformations. When evaluating how these learned transformations influence inter-subject similarity, we adopt the performance metrics detailed in Section 4.3.1. Importantly, inter-subject decoding requires labelled data to derive an accuracy score, while time-segment matching relies on unlabelled data. We therefore separately present the downstream application datasets for these two metrics, noting that inter-subject correlation can be applied in either case.

4.3.4 Alignment data

Courtois-NeuroMod and IBC emphasize different experimental paradigms in their scanning protocols. Courtois-NeuroMod includes a broader range of naturalistic stimuli, including repetitions of full audio-visual movies. IBC, meanwhile, emphasizes traditional psychological paradigms while also including additional naturalistic stimuli.

For Courtois-NeuroMod, we derived alignments using the `movie10` sub-dataset which includes four different audio-visual movies: *Bourne Supremacy*, *Hidden Figures*, *Life*, and *Wolf of Wall Street*. For IBC, we derived alignments from four different stimuli: a *Raiders of the Lost Ark* audio-visual movie, *Le Petit Prince* audio-only recording, short visual-only clips introduced in Nishimoto et al. (2011), and 151 non-overlapping contrasts collected using non-overlapping task paradigms (i.e., excluding HCP, RSVP, and Sounds tasks). For a complete description of the alignment stimuli included in each data set, please see Table 4.1.

4.3.4.1 Downstream tasks for inter-subject decoding

Both the IBC and Courtois-NeuroMod data sets include the six Human Connectome Project (HCP, Van Essen et al. 2013) tasks in their experimental protocols. We therefore calculated fixed-effects contrasts for the 24 task conditions (i.e., non-oppositional contrasts) pooled across all six HCP tasks and used this set of fixed-effects maps as a downstream application for both Courtois-NeuroMod and IBC.

As IBC further contains a range of task-based protocols covering unrelated cognitive domains, we also incorporated two additional experimental tasks as downstream applications. Specifically, we included the Rapid Serial Visual Presentation (RSVP) language and Sounds auditory (Tonotopy) paradigms. Although the HCP task protocol does include a language task, this differs from RSVP language along several dimensions: while HCP

Dataset	Alignment stimulus	*n*	Modality
Individual Brain Charting	Task-battery contrasts	151	Fixed-effects statistical maps
	Clips dataset	2000	Visual-only clips
	Le Petit Prince	2000	Audio-only narrative
	Raiders of the Lost Ark	2000	Audio-visual narrative film
Courtois-NeuroMod	*Bourne Identity*	2023	Audio-visual narrative film
	Hidden Figures	2038	
	Life documentary	2008	
	Wolf of Wall Street	2030	

Table 4.1: **Included alignment stimuli.** The range of alignment stimuli used for each dataset, the number of timeframes *n* that it contains, and the modalities that it covers (e.g. audio only, audio-visual, narrative-based). For the IBC contrasts, a full listing of included 151 contrast maps is provided in Supplemental Section S4.9.

language focuses on contrasting an auditory narrative with spoken math problems, RSVP language engages syntactic and semantic processing of visually presented words and non-words. These downstream applications thus provide additional, unique cognitive contexts in which experimenters may wish to improve inter-subject similiarity.

For a complete description of the downstream tasks for inter-subject decoding considered with each data set, please see Table 4.2.

4.3.4.2 Downstream tasks for time-segment matching

For time-segment matching, we rely on the same stimuli described in Section 4.3.4. For a given experiment, we exclude overlapping stimuli when training and testing the alignment. That is, if functional alignment transformations are calculated using one naturalistic stimulus, then the learned transformation would be applied on all other naturalistic stimuli within the same dataset and evaluated. For example, if alignment transformations are learned on the *Life* dataset in Courtois-NeuroMod, time-segment matching would be calculated separately on *Bourne Supremacy*, *Hidden Figures*, and *Wolf of Wallstreet*.

Importantly, we exclude the 151 contrast maps used in IBC as these do not have a defined temporal structure. For a complete description of the downstream tasks for time-segment matching, please see Table 4.2.

Dataset	Downstream task	*d*	*n*	Cognitive domains
Individual Brain Charting	RSVP	360	72	Language
	Tonotopy	72	12	Auditory
	HCP 24	48	2	Emotion, Gambling, Language, Motor, Socio-relational, Working Memory
Courtois-Neuromod	HCP 24	216	9	Emotion, Gambling, Language, Motor, Socio-relational, Working Memory

Table 4.2: **Downstream inter-subject decoding tasks.** The downstream inter-subject decoding task(s) used for each dataset. Each subject's decoding task data comprises *d* images evenly divided across the *n* cross-validation folds. We note the cognitive domains covered by each downstream task. In the case of the *HCP 24* task, these include all six domains targeted by the six Human Connectome Project (HCP) task protocol.

4.3.5 Experimental procedure

For both the Courtois-NeuroMod and IBC datasets, we calculate piecewise Procrustes alignment transformations within the 300 region Schaefer parcellation (Schaefer et al., 2018). Pairwise alignment transformations were calculated using each of the dataset-specific alignment stimuli described in Section 4.3.4.

Alignment transformations were applied to each of the labelled downstream tasks described in Section 4.3.4.1 and the unlabelled downstream tasks described in Section 4.3.4.2. We then calculate each of the three performance metrics described in Section 4.3.1 at both the whole-brain and parcel-wise level. In all cases, we evaluate the performance of functional alignment as a change in each metric's values after piecewise Procrustes alignment as compared to anatomical-only alignment.

In addition to the qualitative spatial patterns we observe for each metric, we also derive additional quantitative indicators of functional alignment performance. Specifically, we correlate observed distributions to assess similarity (*1*) across alignment stimuli and (*2*) across performance metrics. We also conduct ANOVAs to assess the relative influence of alignment stimuli vs downstream task within each metric.

4.4 Results

4.4.1 Impacts of downstream task and alignment stimuli

We first examine the relative impacts of alignment stimuli and downstream task on performance. We assess these at a whole-brain level to establish general effects when pooling over all possible brain regions before examining in a parcelwise fashion. Note that for the whole-brain level-of-analysis, we only consider inter-subject decoding and spatial inter-subject correlation (spatial ISC) as relevant performance metrics as, to date, time-segment matching has not been developed and applied at a whole-brain scale.

4.4.1.1 Inter-subject decoding at a whole-brain scale

For inter-subject decoding, we compare all possible alignment stimuli for each of dataset-specific decoding tasks. Figure 4.2 shows the distribution of inter-subject decoding accuracy values relative to anatomical alignment. In general, we see that alignment stimuli within the same modality (e.g. audio-visual movies) show relatively similar performance within a given decoding task. However, across tasks, the same alignment stimulus can have different effects on downstream inter-subject decoding accuracy. For example, *Le Petit Prince* decreases inter-subject decoding accuracy in the RSVP language task but increases inter-subject decoding accuracy above baseline for the Tonotopy task.

To quantify this effect, we performed a two-way ANOVA to analyze the effect of task and alignment stimulus on change in inter-subject decoding accuracy following Procrustes alignment. The ANOVA revealed that there was a statistically significant interaction between the effects of task and alignment ($F(14, 84) = 5.43$, $p < 0.001$). Simple main effects analyses showed that both task ($F(2, 84) = 5.09$, $p = 0.008$) and alignment ($F(7, 84) = 3.66$, $p = 0.002$) have a statistically significant effect on accuracy change.

4.4.1.2 Spatial inter-subject correlation patterns

We generate spatial inter-subject correlation (spatial ISC) values for each decoding task following anatomical alignment and compare them with derived spatial ISC patterns following piecewise Procrustes alignment on each alignment stimulus.

Figure 4.3 shows the distribution of changes in spatial ISC values for the HCP24 decoding task in Courtois-NeuroMod after functional alignment on each of the four audio-visual alignment stimuli. In general, we note that a majority of regions show decreased spatial ISC after functional alignment, while a few regions—particularly in visual and

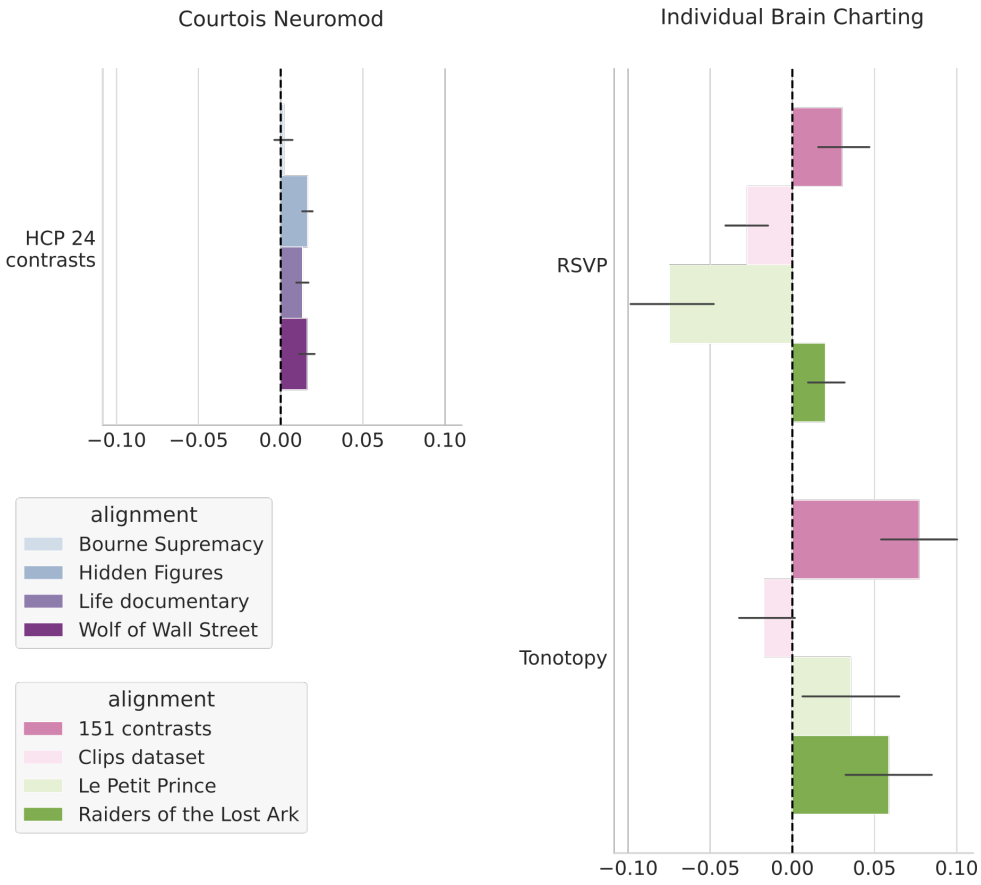

Figure 4.2: **Full brain inter-subject decoding accuracy changes across alignment stimuli.** We plot the change in full-brain inter-subject decoding accuracy against baseline, anatomical-only alignment for each of the four considered alignment stimuli in both the Courtois-NeuroMod (left panel) and IBC (right panel) datasets. Here, the zero indicates no-change from anatomical only alignment, whereas positive values indicate a relative increase in decoding performance and negative values indicate a relative decrease. Error bars represent the standard error across cross validation folds.

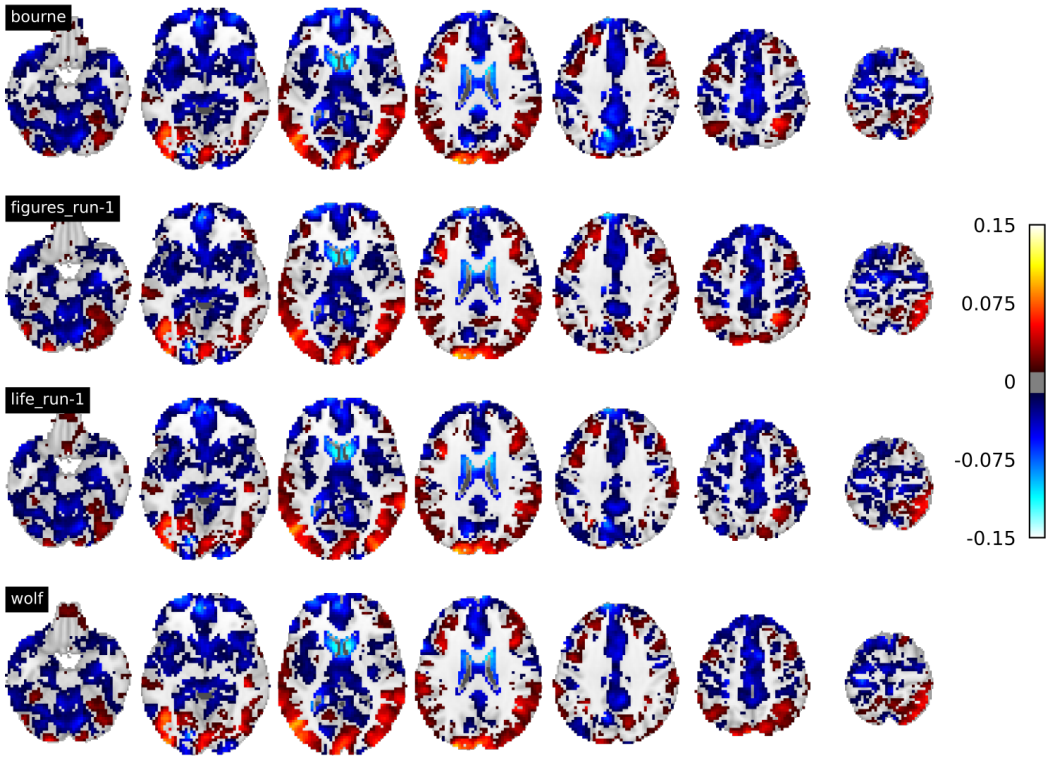

Figure 4.3: **Spatial inter-subject correlation in Courtois-NeuroMod.** Change in spatial ISC distributions following functional alignment across two naturalistic audio-visual films. Results are shown for a single representative cross-validation fold. Spatial ISC values are thresholded at 0.01 and shown from -0.15 to 0.15.

dorsolateral prefrontal cortex—show increased spatial ISC. This pattern is remarkably consistent across alignment stimuli even as individual regions show slight differences in the spatial extent of this effect.

4.4.2 Parcelwise analyses

Although spatial ISC provides voxelwise information on the relative effects of functional alignment, we conduct additional, parcelwise analyses to more directly examine regional differences in the impacts of functional alignment on our considered performance metrics.

In order to summarize general trends across parcels, we calculate the correlation between the original, baseline value of each performance metric with its value after functional alignment. For example, we calculate inter-subject decoding accuracy scores for all parcels with baseline, anatomical-alignment and then correlate these values with the inter-subject decoding accuracies computed for each parcel after functional alignment. By repeating this procedure for each cross-validation fold, we obtain a distribution of

Dataset	Alignment stimulus	24 contrasts	RSVP	Tonotopy
Individual Brain Charting	151 contrasts	-0.60 ± 0.05	-0.39 ± 0.10	-0.57 ± 0.07
	Clips dataset	-0.48 ± 0.08	-0.53 ± 0.07	-0.67 ± 0.04
	Le Petit Prince	-0.73 ± 0.02	-0.65 ± 0.07	-0.61 ± 0.05
	Raiders of the Lost Ark	-0.54 ± 0.06	-0.40 ± 0.11	-0.61 ± 0.06
Courtois-NeuroMod	*Bourne Identity*	-0.43 ± 0.05		
	Hidden Figures	-0.41 ± 0.06		
	Life documentary	-0.43 ± 0.03		
	Wolf of Wall Street	-0.41 ± 0.05		

Table 4.3: **Correlation between baseline and change in parcelwise decoding following functional alignment.** Correlation between parcelwise gains in inter-subject decoding of task-relevant information following Procrustes alignment with the baseline parcelwise decoding accuracy following anatomical alignment. Parcelwise correlations were calculated separately for each cross-validation fold (i.e., each left-out subject), and the average overall rho value is reported here.

correlation values for our sample, each of which represent the relative consistency of our performance metrics before and after functional alignment. Here, we focus on inter-subject decoding and time-segment matching accuracies, rather than re-calculating voxelwise spatial ISC values (see Section 4.4.1.2).

4.4.2.1 Inter-subject decoding within individual parcels

Table 4.3 shows the correlation between inter-subject decoding accuracies for each task following anatomical only alignment with inter-subject decoding accuracies after piecewise Procrustes alignment. Across all considered decoding tasks and alignment stimuli, we observe a negative relationship between parcelwise decoding accuracies at baseline and as changed after functional alignment. Thus, parcels with higher baseline inter-subject decoding accuracy values show the largest decreases following functional alignment on average.

To investigate the exact spatial pattern of these results, we sub-select two downstream decoding tasks from IBC: the Tonotopy and RSVP Language tasks. In Figure 4.4, we show the spatial pattern of inter-subject decoding accuracies across these two tasks at baseline (in purple) and as changed following functional alignment—with two unique alignment stimuli—for a single cross-validation fold.

These results highlight that observed changed in inter-subject decoding with functional alignment are primarily dependent on the unique spatial locations that are informative for a given downstream application rather than being driven by the considered alignment stimulus. We further visualize the correlation between these baseline values and the

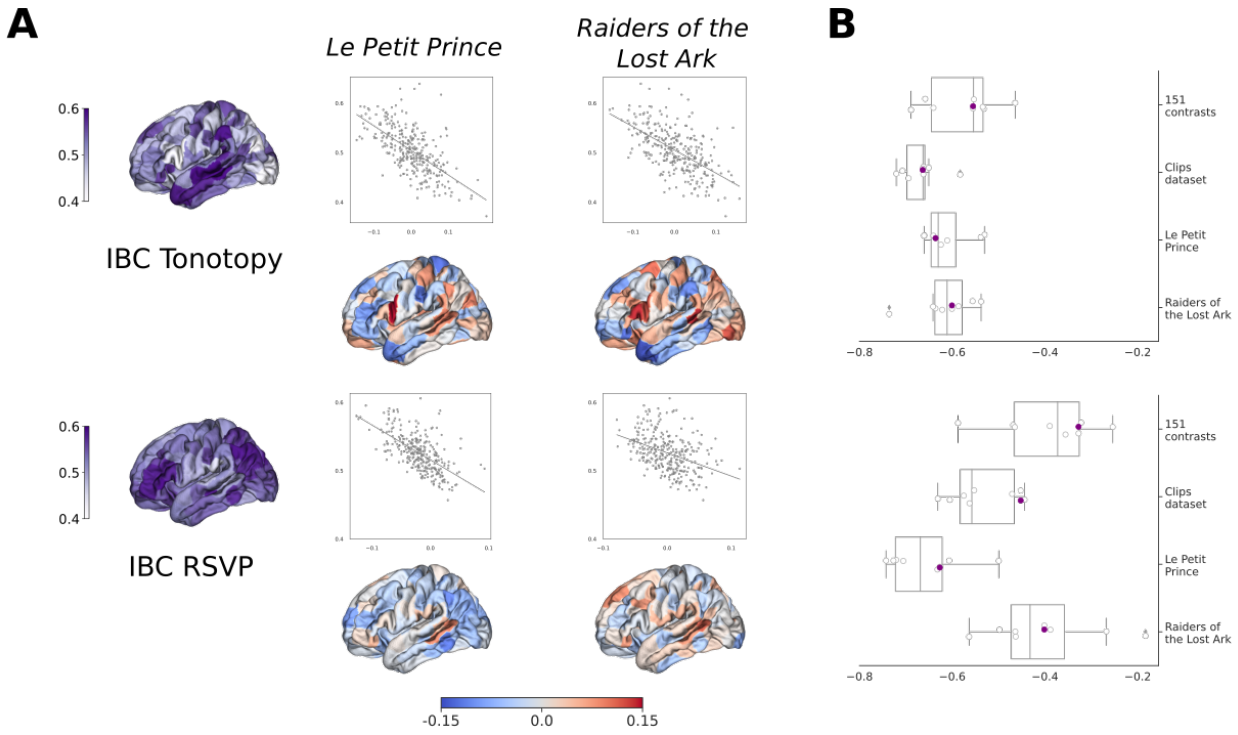

Figure 4.4: **Parcelwise relationship between baseline decoding accuracy and improvement following functional alignment.** In Panel A, results are shown for a representative cross-validation fold. In Panel B, each dot represents a separate cross-validation fold, with the fold shown in Panel A highlighted in purple.

Alignment stimulus	**Clips dataset**	***Le Petit Prince***	***Raiders of the Lost Ark***
151 contrasts	0.02 ± 0.39	-0.42 ± 0.11	-0.27 ± 0.07
Clips dataset		-0.58 ± 0.08	-0.35 ± 0.24
Le Petit Prince	-0.95 ± 0.03		-0.82 ± 0.07
Raiders of the Lost Ark	0.30 ± 0.33	-0.40 ± 0.15	

Table 4.4: **Parcelwise time-segment matching in IBC.** Correlation between parcelwise gains in time-segment matching following Procrustes alignment with baseline parcelwise time-segment matching accuracy after anatomical alignment. Parcelwise correlations were calculated separately for each cross-validation fold (i.e., each left-out subject), and the average overall rho value is reported here.

change in inter-subject decoding accuracy following functional alignment—on each of the two considered stimuli—showing the same negative relationship observed in Table 4.3.

4.4.2.2 Time-segment matching within individual parcels

Table 4.4 shows the average parcelwise correlations within IBC for time-segment matching accuracy values at baseline (i.e., following anatomical alignment) and accuracy gains following Procrustes alignment (i.e., subtracting the baseline accuracy value for that cross-validation fold from the accuracy value following Procrustes alignment).

In general, we observe a similar pattern to that seen with inter-subject decoding, where regions that had weakest performance with anatomical alignment show the largest relative increase in time-segment matching accuracy following piecewise Procrustes functional alignment. We note, however, that there is significantly more variability in these results: while the maximum standard deviation in Table 4.3 is 0.11, the same value in Table 4.4 is 0.39. This variability suggests that the relationship between the spatial pattern of baseline time-segment matching accuracies and parcelwise change in time-segment matching accuracies following functional alignment is weaker than that seen for inter-subject decoding.

4.4.2.3 Comparing spatial inter-subject correlation to other performance metrics

Spatial inter-subject correlation (spatial ISC) is unique among our considered performance metrics in that it can be applied on both labelled as well as unlabelled data. It is therefore interesting to ask whether there is a significant relationship between spatial ISC and our two other considered performance metrics. If, in fact, spatial ISC is strongly related to both metrics, it would provide a convenient mechanisms to compare results between labelled and unlabelled downstream applications.

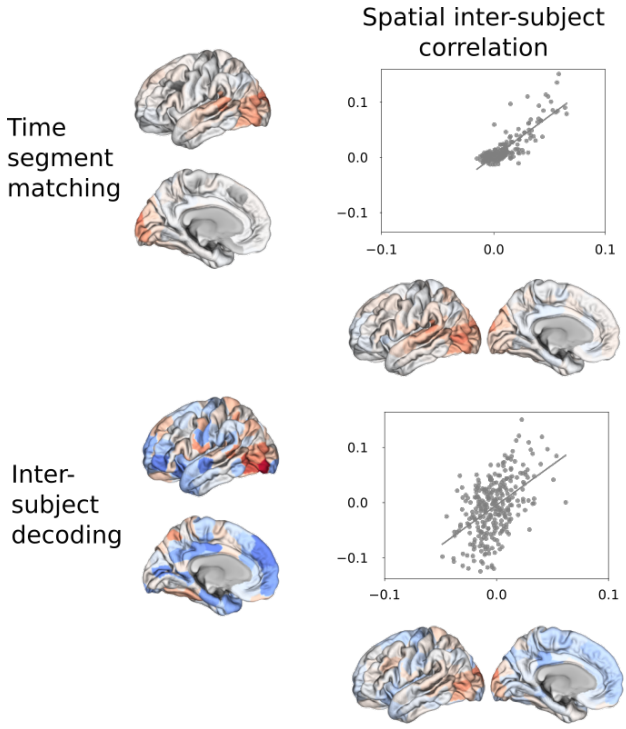

Figure 4.5: **Parcelwise correlation between change in spatial ISC and other performance metrics.** We compare parcelwise changes in spatial ISC following functional alignment on *Bourne Supremacy* with the parcelwise changes in time-segment matching accuracy for *Hidden Figures* (upper plot) and in inter-subject decoding for the HCP 24 contrasts task (lower plot) in Courtois-NeuroMod. Results are shown for a representative cross-validation fold.

We calculate and depict this relationship for a representative cross-validation fold of Courtois-NeuroMod in Figure 4.5. That is, we show the parcelwise relationship between change in spatial ISC following functional alignment and change in each of the two leave-one-out performance metrics for a single alignment stimulus, *Bourne Supremacy*. In general, we see a weak, positive relationship between changes in spatial ISC and each of the other metrics. That is, while spatial ISC generally shares a similar spatial pattern with both inter-subject decoding as well as time-segment matching, the observed values differ significantly. We note, though, that this relationship is strongest when training and testing alignment on two non-overlapping stimuli of the same structure; e.g., two naturalistic audio-visual films, as in the upper panel of Figure 4.5.

4.5 Discussion

In this work, we evaluated the impacts of experimental context on the performance of functional alignment in cognitive neuroscience applications. We defined performance using three unique metrics—inter-subject decoding, time-segment classification, and spatial inter-subject correlation (spatial ISC)—and calculate these metrics across two openly available, well-sampled deep phenotyping datasets for a variety of alignment and application tasks. In general, we find that there is a broad consistency across these metrics and that functional alignment improves performance in a majority of downstream applications. Importantly, however, we also find that in some contexts functional alignment can impair rather than improve observed inter-subject similarity. Further, even in cases where functional alignment improves inter-subject similarity on average, it may decrease the inter-subject similarity of individual brain regions.

4.5.1 Evaluating functional alignment performance across performance metrics

Assessing the effect of functional alignment on inter-subject similarity is non-trivial. In this work, we adopted three metrics commonly used in the cognitive neuroscience literature to try and quantify these effects: inter-subject decoding, time-segment matching, and spatial ISC. While both inter-subject decoding and time-segment matching provide quantitative estimates of inter-subject similarity via classification accuracy, they are intended for labelled and unlabelled data, respectively. Spatial ISC, meanwhile, estimates the voxelwise correlation of evoked activity patterns across subjects and can therefore be applied across differently structured datasets. Our results indicate that change in spatial ISC following functional alignment roughly matches the change in spatial patterns seen with either inter-subject decoding or time-segment matching. However, the magnitude of derived values differs significantly between spatial ISC and each of the two classification metrics—as well as across datasets—cautioning against its use as an independent metric. Indeed, this concern in using spatial ISC to evaluate inter-subject similarity across datasets aligns with more general concerns around image-based metrics (e.g. Dice coefficient), as these metrics may be sensitive to low-level image characteristics induced by processing routines or different acquisition parameters.

While the two classification-based performance metrics cannot be directly compared, they show unique performance in their sensitivity to functional alignment changes. There

is a consistent, negative relationship between baseline inter-subject decoding accuracy and the change in accuracy following functional alignment across the 300 considered Schaefer parcels. Time-segment matching similarly shows a negative relationship between baseline and change in classification accuracy; however, this relationship is notably weaker, showing more variability across parcels. This decoupling between baseline and observed changes suggests that time-segment classification accuracy may not serve as an effective replacement for inter-subject decoding, encouraging researchers to collect separate, task-based data to assess the impacts of functional alignment.

4.5.2 The importance of matching alignment and application data

At a whole-brain scale, we note that unique alignment stimuli differentially impact the performance of both inter-subject decoding and spatial ISC metrics. Nonetheless, as suggested by both metrics in Courtois-NeuroMod (see Figures 4.2 and 4.3), these differences are relatively minor within a given alignment modality (e.g., audio-visual narrative film). Across modalities, by contrast, we see more significant differences in performance. For example, in the IBC whole-brain inter-subject decoding results depicted in Figure 4.2, the visual-only Clips dataset impairs inter-subject decoding performance for both the RSVP Language and Tonotopy tasks. This effect is likely because Clips primarily contains structured signal in visual regions.

The audio-only narrative in the *Le Petit Prince* dataset, by contrast, improves inter-subject decoding accuracy in the Tonotopy but not RSVP Language application. We can infer that this is due to the RSVP Language task relying on more distributed regions outside of auditory context, as both 151 contrasts and *Raiders of the Lost Ark*—which evoke structured signal across the whole-brain—successfully improve inter-subject decoding. This result highlights the difficulty in selecting an appropriate functional alignment stimulus. That is, even when calculating functional alignment transformations on a language stimulus, there is no guarantee that applying those transformations to a non-overlapping language task will improve inter-subject similarity. In general, however, we expect that functional alignment applications will have a higher probability of successfully improving similarity when data on which the transformations are learned and applied share a similar structure or pattern of evoked activations.

4.5.3 Functional alignment improves information gain in ill-fitted areas

When comparing across individual parcels, we see that the regions which show the most change following functional alignment are largely dependent on the downstream application rather than the alignment dataset. For example, we see relative consistency in panel A of Figure 4.4 between the Tonotopy and RSVP Language tasks, with more similarity across the two alignment stimuli considered for each decoding task.

Functional alignment tends to improve information gain in areas that showed low inter-subject similarity at baseline. Importantly, this improvement comes at the expense of high inter-subject similarity regions, which show decreased inter-subject similarity after functional alignment. From the voxelwise spatial ISC patterns shown in Figure 4.3, however, many of these regions with high inter-subject similarity at baseline may be driven by shared noise sources rather than shared task-relevant signal. For example, functional alignment reduces the spatial ISC in both the ventricles as well as along the cortical mid-line, where we expect pulsation-related artifacts.

4.5.4 Study limitations and future directions

Although our study provides important insight into the experimental dimensions which influence functional alignment performance, there are of course more relevant dimensions than we can systematically assess here. For example, we have chosen to focus on whole-brain parcellations rather than task-specific regions-of-interest to more easily make comparisons across different downstream applications. It is nonetheless possible that a single alignment transformation derived from a large task-specific region-of-interest will be more dramatically influenced by signal structure differences between training and downstream application data. Aggregating learned transformations over smaller parcels, by contrast, may mediate these differences, allowing smaller regions with shared signal structure to partially correct for poor overlap in the training and application data. Given our previous findings showing relatively consistent performance of specific functional alignment algorithms between task-specific regions-of-interest and whole-brain parcellations (Bazeille et al., 2021), however, it is unclear exactly how strong this effect would be and it is likely to differ across both regions-of-interest and included data.

Further, while we have tried to include a broad range of downstream applications—including constructing a new HCP24 inter-subject decoding task, spanning six cognitive domains—it is possible that other cognitive domains not included here may show differential impacts of functional alignment. This is particularly interesting for complex social

cognition relying on regions such as the ventromedial prefrontal cortex, as the cortical mid-line shows no clear increase after functional alignment in any of the three considered performance metrics. Future work evaluating alignment specifically in these more targeted tasks will be useful to examine the limitations of this class of methods, and whether additional alignment in time (as in Xie et al., 2021) is also necessary to better capture these regions. In general, our current results suggest that training data that evoke structured signal across the whole brain—such as naturalistic audio-visual narratives, or statistical contrast maps from a range of cognitive domains—are likely to allow researchers to learn broadly generalizable alignment transformations. These results caution, however, that we cannot expect current techniques to maximize inter-subject similarity across all brain regions, even when applying these more generalizable transformations.

4.6 Conclusion

In the present work, we have investigated the impact of three experimental dimensions on a single functional alignment algorithm. We find that the relative success of the method is dependent on the exact combination of data used to train and test alignment transformations. These results highlight the importance of assessing alignment performance in cognitive neuroscience applications and suggest that caution is required when using these techniques to improve inter-subject similarity, as we do not find consistent improvement across downstream applications and identical alignment stimuli and vice versa. It is thus clear that independent testing data is required when applying functional alignment and that this testing data should share a similar signal structure to the data used to train the alignment; i.e., it should evoke structured signal in overlapping regions. Our results further suggest that this data should ideally be amenable to inter-subject decoding as this performance metric shows consistent spatial relationships between baseline task-relevant signal and signal following functional alignment. We argue that improvement in inter-subject decoding in a related task should therefore be considered a minimum requirement for future applications of functional alignment in cognitive neuroscience.

4.7 Bibliography

Abraham, A. et al. (2014). "Machine learning for neuroimaging with scikit-learn". *Frontiers in Neuroinformatics*, 8.

Avants, B. et al. (2008). "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain". *Medical Image Analysis*, 12(1), pp. 26–41.

Bazeille, T. et al. (2021). "An empirical evaluation of functional alignment using inter-subject decoding". *NeuroImage*, p. 118683.

Benson, N. C. et al. (2021). "Variability of the Surface Area of the V1, V2, and V3 Maps in a Large Sample of Human Observers".

Bilenko, N. Y. et al. (2010). "How much tuning information is lost when we average across subjects in fMRI experiments?" *Journal of Vision*, 10, p. 917.

Boyle, J. A. et al. (2020). *The Courtois project on neuronal modelling: 2020 data release*. https://docs.cneuromod.ca. Presented at the 26th annual meeting of the Organization for Human Brain Mapping.

Brett, M., I. S. Johnsrude, and A. M. Owen (2002). "The problem of functional localization in the human brain". *Nat. Rev. Neurosci.*, 3(3), pp. 243–249.

Chen, J. et al. (2017). "Shared memories reveal shared structure in neural activity across individuals". *Nat. Neurosci.*, 20(1), pp. 115–125.

Chen, P.-H. et al. (2015). "A Reduced-Dimension fMRI Shared Response Model". In: *Advances in Neural Information Processing Systems 28*. Ed. by C Cortes et al. Curran Associates, Inc., pp. 460–468.

Coalson, T. S., D. C. Van Essen, and M. F. Glasser (2018). "The impact of traditional neuroimaging methods on the spatial localization of cortical areas". *Proc. Natl. Acad. Sci. U. S. A.*, 115(27), E6356–E6365.

Cox, R. W. and J. S. Hyde (1997). "Software tools for analysis and visualization of fMRI data". *NMR in Biomedicine*, 10(4-5), pp. 171–178.

Esteban, O. et al. (2018a). "fMRIPrep". *Software*.

Esteban, O. et al. (2018b). "fMRIPrep: a robust preprocessing pipeline for functional MRI". *Nature Methods*.

Fonov, V. et al. (2009). "Unbiased nonlinear average age-appropriate brain templates from birth to adulthood". *NeuroImage*, 47, Supplement 1, S102.

Gordon, E. M. et al. (2017). "Precision Functional Mapping of Individual Human Brains". *Neuron*, 95(4), 791–807.e7.

Gorgolewski, K. et al. (2011). "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python". *Frontiers in Neuroinformatics*, 5, p. 13.

Gorgolewski, K. J. et al. (2018). "Nipype". *Software*.

Greve, D. N. and B. Fischl (2009). "Accurate and robust brain image alignment using boundary-based registration". *NeuroImage*, 48(1), pp. 63–72.

Hasson, U., R. Malach, and D. J. Heeger (2010). "Reliability of cortical activity during natural stimulation". *Trends Cogn. Sci.*, 14(1), pp. 40–48.

Haxby, J. V. et al. (2011). "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron*, 72(2), pp. 404–416.

Jenkinson, M. et al. (2002). "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images". *NeuroImage*, 17(2), pp. 825–841.

Lanczos, C. (1964). "Evaluation of Noisy Data". *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1), pp. 76–85.

Marie, D et al. (2015). "Descriptive anatomy of Heschl's gyri in 430 healthy volunteers, including 198 left-handers". *Brain Struct. Funct.*, 220(2), pp. 729–743.

Michel, V. et al. (2012). "A Comparative Study of Algorithms for Intra- and Inter-subjects fMRI Decoding". In: *Machine Learning and Interpretation in Neuroimaging*. Springer Berlin Heidelberg, pp. 1–8.

Nastase, S. A. et al. (2017). "Attention Selectively Reshapes the Geometry of Distributed Semantic Representation". *Cereb. Cortex*, 27(8), pp. 4277–4291.

Nastase, S. A. et al. (2019). "Measuring shared responses across subjects using intersubject correlation". *Soc. Cogn. Affect. Neurosci.*, 14(6), pp. 667–685.

Nishimoto, S. et al. (2011). "Reconstructing visual experiences from brain activity evoked by natural movies". *Current biology*, 21(19), pp. 1641–1646.

Paquola, C. et al. (2019). "Microstructural and functional gradients are increasingly dissociated in transmodal cortices". *PLoS Biol.*, 17(5), e3000284.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*, 12, pp. 2825–2830.

Pinho, A. L. et al. (2018). "Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping". *Sci Data*, 5, p. 180105.

Pinho, A. L. et al. (2020). "Individual Brain Charting dataset extension, second release of high-resolution fMRI data for cognitive mapping". *Scientific Data*, 7(1).

Rademacher, J et al. (1993). "Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology". *Cereb. Cortex*, 3(4), pp. 313–329.

Raizada, R. D. S. and A. C. Connolly (2012). "What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decoding". *J. Cogn. Neurosci.*, 24(4), pp. 868–877.

Schaefer, A. et al. (2018). "Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI". *Cereb. Cortex*, 28(9), pp. 3095–3114.

Sonkusare, S., M. Breakspear, and C. Guo (2019). "Naturalistic Stimuli in Neuroscience: Critically Acclaimed". *Trends Cogn. Sci.*

Thirion, B. et al. (2006). "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets". *Hum. Brain Mapp.*, 27(8), pp. 678–693.

Tustison, N. J. et al. (2010). "N4ITK: Improved N3 Bias Correction". *IEEE Transactions on Medical Imaging*, 29(6), pp. 1310–1320.

Van Essen, D. C. et al. (2013). "The WU-Minn Human Connectome Project: an overview". *Neuroimage*, 80, pp. 62–79.

Vázquez-Rodríguez, B. et al. (2019). "Gradients of structure-function tethering across neocortex". *Proc. Natl. Acad. Sci. U. S. A.*, 116(42), pp. 21219–21227.

Xie, T. et al. (2021). "Minimal functional alignment of ventromedial prefrontal cortex intracranial EEG signals during naturalistic viewing".

Zhang, Y., M. Brady, and S. Smith (2001). "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm". *IEEE Transactions on Medical Imaging*, 20(1), pp. 45–57.

S4.8 fMRIPrep preprocessing

Results included in this manuscript come from preprocessing performed using *fMRIPrep* 20.2 (Esteban et al., 2018b; Esteban et al., 2018a; RRID:SCR_016216), which is based on *Nipype* 1.5.0 (Gorgolewski et al., 2011; Gorgolewski et al., 2018; RRID:SCR_002502).

S4.8.1 Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a Nipype implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, RRID:SCR_002823, Zhang et al., 2001). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009, RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym).

S4.8.2 Functional data preprocessing

For each subject's BOLD runs (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated by aligning and averaging 1 single-band references (SBRefs). A B0-nonuniformity map (or *fieldmap*) was estimated based on two (or more) echo-planar imaging (EPI) references with opposing phase-encoding directions, with `3dQwarp` Cox and Hyde (1997) (AFNI 20160207). Based on the estimated susceptibility distortion, a corrected EPI (echo-planar imaging) reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters

with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al., 2002).

First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*.

All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964).

Many internal operations of *fMRIPrep* use *Nilearn* 0.6.2 (Abraham et al., 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

S4.8.3 Copyright Waiver

The above boilerplate text was automatically generated by fMRIPrep with the express intention that users should copy and paste this text into their manuscripts *unchanged*. It is released under the CC0 license.

S4.9 Listing of 151 contrasts used as IBC alignment stimuli

Here, we include 151 contrasts as alignment stimuli for the IBC dataset. The contrasts are common to all subjects and taken from various protocols. For completeness, we provide their full listing here. To know more about the supporting protocols, please visit the project documentation. Contrasts include :

- ARCHI battery of tasks: *'left - right button press','reading - listening','motor - cognitive','reading - checkerboard','computation - sentences','horizontal - vertical','saccades','hand*

- side','grasp - orientation','rotation side','object orientation','triangle mental - random','false belief - mechanistic audio','triangle random','mechanistic audio','mechanistic video','non speech sound','false belief - mechanistic video','speech - non speech','expression gender - control','expression intention - gender','face gender - control','face trusty - gender'

- Mental Time Travel tasks: *'we average reference','we all space - time cue','we average event','we space - time event','westside - eastside event','we before - after event','sn average reference','sn all space - time cue','sn average event','sn space - time event','southside - northside event','sn before - after event'*
- Preference tasks: *'preference constant','preference linear','preference quadratic'*
- Theory-of-Mind and Pain Matrices localizer tasks: *'photo','belief - photo','physical pain','emotional - physical pain','movie pain','movie mental - pain'*
- Visual Short-Term Memory and Enumeration tasks: *'vstm constant','vstm linear','vstm quadratic','enumeration constant','enumeration linear','enumeration quadratic'*
- Self task: *'encode other','encode self - other','recognition other hit','recognition self - other','correct rejection','recognition hit - correct rejection'*
- Lyon tasks battery: *'unattend','attend - unattend','attend','tear - silence','suomi - silence','yawn - silence','human - silence','silence','music - silence','reverse - silence','speech - silence','alphabet - silence','cough - silence','environment - silence','laugh - silence','animals - silence','scrambled','face - scrambled','characters - scrambled','scene - scrambled','house - scrambled','animal - scrambled','pseudoword - scrambled','tool - scrambled','random string','word - pseudoword','word - random string','pseudoword - random string','2 letters different - same','4 letters different - same','6 letters different - same','6 letters different - 2 letters different','2 dots - 2 dots control','4 dots - 4 dots control','6 dots - 6 dots control','6 dots - 2 dots','low - high salience','salience left - right'*
- Stanford tasks battery: *'spatial cue - double cue','spatial cue','incongruent - congruent','double incongruent - double congruent','double congruent','double incongruent','double cue','spatial incongruent - spatial congruent','spatial incongruent','spatial congruent','go','stop','stop - go','task stay cue stay','task switch cue switch','task switch cue stay','task stay cue switch','task switch - stay','cue switch - stay','delay','amount','go critical','go noncritical','stop','ignore','go critical - stop','go noncritical - ignore','ignore - stop','stop - ignore','congruent','incongruent','incongruent*

- congruent','num loss cards','loss','gain','cue','correct cue correct probe','correct cue incorrect probe','incorrect cue correct probe','incorrect cue incorrect probe','correct cue incorrect probe - correct cue correct probe','incorrect cue incorrect probe - incorrect cue correct probe','correct cue incorrect probe - incorrect cue correct probe','incorrect cue incorrect probe - correct cue incorrect probe','incorrect probe - correct probe','ambiguous intermediate','unambiguous direct','intermediate - direct','ambiguous direct','unambiguous intermediate','ambiguous - unambiguous'

- Biological Motion task: *'global upright - global inverted','natural upright - natural inverted','global upright - natural upright','modified upright - modified inverted','natural upright - natural inverted','natural upright - modified upright'*

Chapter 5

Finding similarity across participants using functional alignment

Elizabeth DuPre[1], Jean-Baptiste Poline[1]

[1]NeuroDataScience - ORIGAMI laboratory, McGill University, Montreal, Quebec, Canada

5.1 Preface

The gap between methodological and domain-oriented science persists in part because of the difficulty in translating controlled benchmarks into more general recommendations. To do so effectively requires synthesizing experimental results such as those presented in Chapter 3 and Chapter 4 with domain-relevant literature. Ideally, these recommendations should be provided as actionable steps for researchers to consider in their own studies. Although several tutorials and reviews exist for functional alignment to date, these are focused on a specific algorithm or provide a single analysis pipeline without more general considerations. In this chapter, I develop an accessible tutorial for functional alignment geared towards domain-oriented researchers. This work refines the research presented in Chapters 3 and 4 into a more general framework for considering functional alignment applications in cognitive neuroscience. Accompanying interactive materials provide opportunities to directly translate these recommendations into Python research workflows. This project will be submitted for publication in 2022.

5.2 Abstract

Finding shared neural principles that support the diversity of human experience is a core challenge for social and cognitive neuroscience. By leveraging data collected from different individuals, we can identify similar functional patterns supporting behavior and cognition. Although many functional magnetic resonance imaging (fMRI) studies perform group-level analyses, these largely assume that mapping across neuroanatomical features ensures functional correspondence. In recent years, however, this assumption has been increasingly challenged, and new methods have been introduced to find similarity across different functional organizations. In this tutorial, we introduce *functional alignment*, a class of methods that search for correspondence across participants from their function rather than their anatomy. Throughout, we focus on providing an accessible treatment of methodological choices specific to functional alignment and outline current best practices for applying these methods to fMRI data.

5.3 Introduction

Challenges in comparing fMRI data from different individuals have been well-recognized since fMRI was first introduced as a human brain imaging method (Rademacher et al., 1993) and remain a fundamental question in any multi-subject imaging study. While traditional preprocessing pipelines focus on equating individual anatomies, there are clear limitations to this approach (Brett et al., 2002; Thirion et al., 2006). Over the past decade, social and cognitive neuroscientists have thus turned to new methods for aligning individuals on function rather than structure. Following the introduction of hyperalignment in Haxby et al. (2011), the availability of these "functional alignment" methods has dramatically expanded, matching growing enthusiasm in the field. Despite this enthusiasm, these methods remain relatively inaccessible, with little shared understanding of the relationship between available methods or their appropriate applications. This tutorial aims to address this need by providing a practical introduction to functional alignment methods, their use in social and cognitive neuroscience, their limitations, and their relationship to other available methods for improving correspondence across individuals.

Our presentation is organized around an example implementation, highlighting available decision points in adopting functional alignment. At each described step, we discuss the choices it implicitly involves and available alternatives. While other tutorials (Hanke

et al., 2009; Kumar et al., 2019) similarly focus on illustrative examples, these are specific to a single functional alignment method and available Python software library. We expand on this idea to include multiple classes of functional alignment methods, underscoring shared considerations. We conclude by characterizing both functional alignment applications as well as how they compare against other techniques for improving inter-individual comparisons in fMRI, such as individualized parcellations. To increase the practical relevance of this tutorial, we further develop an accompanying online resource, https://neurodatascience.github.io/fmralign-tutorials, to provide formalisms and interactive Python code examples for five commonly-used functional alignment methods.

While functional alignment promises new insights into the shared principles supporting diverse functional organizations, the relative inaccessibility of these methods means that they remain untested in some experimental applications. It is thus important to acknowledge that in many cases, there is insufficient evidence to base detailed recommendations. Whenever possible, however, we derive recommendations from controlled empirical work. In those cases where experimental evidence is unavailable, we provide general heuristics based on our own experience in applying functional alignment to fMRI data. To ground our discussion, we first introduce the idea of functional alignment and briefly overview its application in social and cognitive neuroscience studies using fMRI data.

5.3.1 What is functional alignment?

Functional alignment is a family of methods for aligning individual activations directly on evoked functional response patterns, rather than on the underlying neuroanatomy. As a result, it is sometimes called "anatomy-free mapping," although this does not imply that anatomical alignment is not additionally performed—and indeed, some functional alignment algorithms can also incorporate anatomical information (Rustamov and Guibas, 2016). Perhaps the clearest contrast between functional alignment and anatomical alignment methods is in the space in which they operate. Anatomical alignment methods operate largely in three-dimensional, physical space, with X, Y, and Z coordinates for each location of measured fMRI data. Functional alignment methods, by contrast, require defining a new "activation space." The dimensions of a given activation space are set by the number of voxels included within it. For example, the activation space defined by two voxels will be a plane, while the activation space defined by three voxels will be a cube. Importantly, activation spaces are simply another way to conceptualize a measured functional response pattern, as illustrated in Figure 5.1.

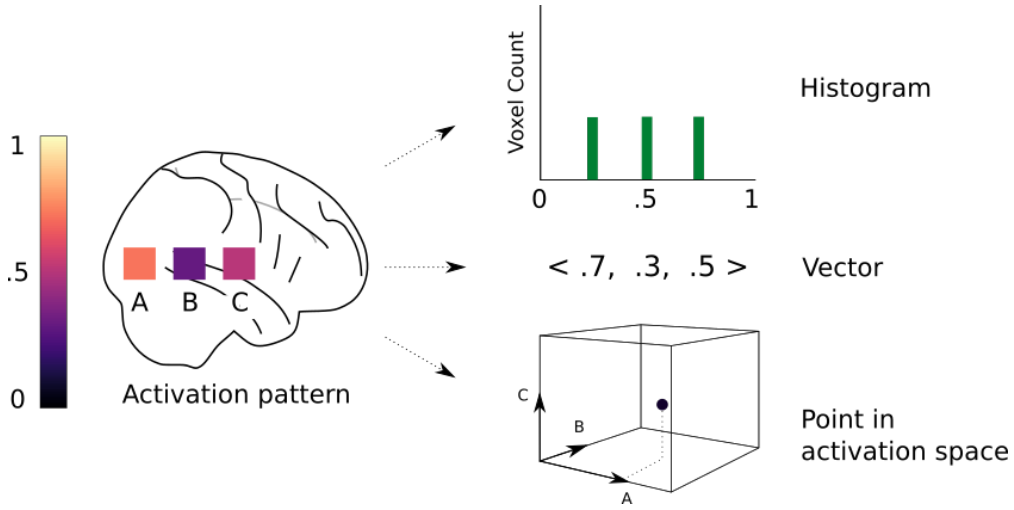

Figure 5.1: **Multiple ways to represent a voxelwise activity pattern.** Voxel activity patterns from voxels *A*, *B*, and *C* can be represented as a histogram of the distribution of values, an activity vector whose length is equal to the number of voxels, or as a point in activation space. In the latter case, the dimensions of the space are determined by the number of voxels. Note that we could also define these spaces using alternative measures such as connectivity. Figure adapted from Churchland (1998).

Although we lose our ability to intuitively visualize activation patterns in these new, high-dimensional spaces, describing fMRI data in an activation space also brings some important advantages. For example, we can easily incorporate information across many voxels, rather than focusing on each voxel independently. As a result of this focus on multi-voxel patterns, functional alignment may feel familiar for researchers who have previously used other Multi-Voxel Pattern Analysis (MVPA; Weaverdyck et al., 2020) methods such as pattern classifiers (Pereira et al., 2009) or Representational Similarity Analysis (RSA: Popal et al., 2019). While these spaces can be defined by other measures such as functional connectivity, for the sake of simplicity we focus on activation. We discuss connectivity and other, alternative measures in Section 5.6.

Functional alignment differs from MVPA methods, however, in that we are not interested in information within a single individual's fMRI data but in the relationship between their fMRI activations and activations from one or more unrelated individuals. That is, functional alignment aims to find correspondence across two or more participant's fMRI data and then use this correspondence to increase their inter-subject similarity. To do so, we learn an alignment transformation for each participant using set-aside fMRI data. This data may be known as alignment data or *training data*, as in other MVPA methods. The learned transformations can then be applied on a different set of non-overlapping *test data* acquired for each participant. Importantly for the methods considered here, this test data should cover the same region or regions of the brain, but it does not have to have the

same task structure. For example, we could learn functional alignment transformations on a movie-viewing alignment dataset and then apply these transformations on test data collected during a traditional psychological paradigm, such as a working-memory or language task.

Formally, we are interested in learning a transformation $\mathbf{R}^{source \rightarrow target}$ that maps an alignment dataset from one source participant $\mathbf{A}^{source}$ to an alignment dataset from a target participant $\mathbf{A}^{target}$, assuming that each dataset contains the same number of n time points for each of p voxels. Note that the target may either be another participant in the same dataset or a reference participant created by averaging across datasets; we discuss the difference between these two approaches further in Section 5.5.3.1. Most commonly, these training datasets are collected while participants watch synchronized stimuli such as a naturalistic film (Finn et al., 2019); however, we assume that we are learning a more general mapping between the functional organization of our source and target participants. In this case, the learned transformation $\mathbf{R}$ can then be applied to another dataset $\mathbf{D}^{source}$, and we can test if the learned relationship improves inter-individual correspondence. Thus, we are interested in whether $\mathbf{RD}^{source}$ is more similar to $\mathbf{D}^{target}$ than $\mathbf{D}^{source}$.

When the number of voxels p is very large, the mapping defined by $\mathbf{R}$ is difficult to reason about intuitively. We can visualize a simple case with three voxels, however, to illustrate the basic idea behind learning this transformation. In Figure 5.2, we show activation spaces defined by two individual's measured fMRI activity patterns across voxels A, B, and C. Within each activation space we define labelled points indicating relative voxelwise activation to the four stimuli: "face," "place," "body," and "tool." This labelling creates an important, implicit correspondence between the two activation spaces even though we can see from visual inspection that there is not a direct correspondence in position; i.e., most stimuli are not located in identical positions across the two activation spaces. When using functional alignment we assume, however, that the relative relationship between each stimulus is preserved across the two spaces, as in this example. The learned transformation $\mathbf{R}$ then reflects bringing the activations defined in one space into alignment with the activations defined in a different space. For example, we can align the activations in Activation space #2 with those in Activation space #1.

There are multiple constraints that we can place on our learned transformation $\mathbf{R}$, reflecting different assumptions about how similar the relationship between labelled activations are across activation spaces. In this tutorial, we focus on alignment methods which learn linear relationships for reasons we discuss later in Section 5.5.3.

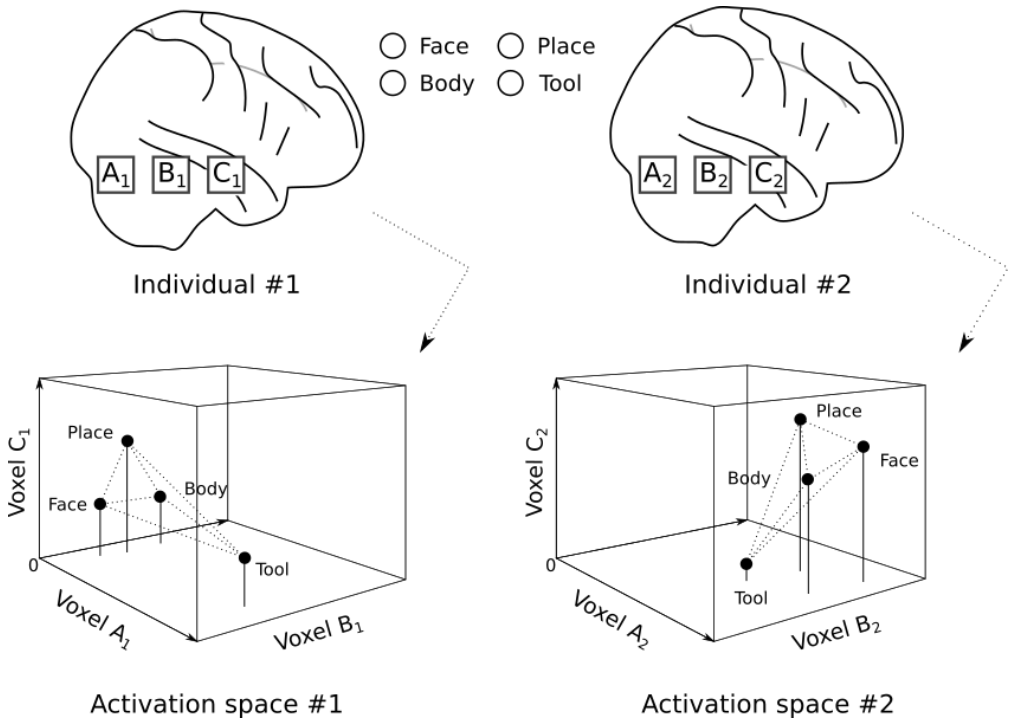

Figure 5.2: **Comparing across individual activation spaces.** For two different participants, we define an activation space using voxels *A*, *B*, and *C*. We then plot the evoked voxelwise activation for four example stimuli within each of these activations. In this case, only the 'Tool' stimulus is in the same place in each activation space; i.e., it evokes the same relative activity pattern across the two participants. The relationship between each of the four stimuli, however, is preserved across both spaces. Figure adapted from Churchland (1998).

5.4 An illustrative example

Here we describe an example application of functional alignment using the publicly available Individual Brain Charting (IBC; Pinho et al., 2018) dataset, accessed on OpenNeuro (Markiewicz et al., 2021) as `ds000244` and `ds002685`. Although IBC contains dozens of task protocols spanning a wide range of cognitive domains, here we focus on just two tasks from nine participants. First, we used a *Raiders of the Lost Ark* audio-visual film viewing during which participants watched the film in its entirety over ten runs, with no in-scanner behavioral responses. Second, we used an experiment designed to assess differences in syntactic and semantic processing of visually presented words and non-words. In each trial, the participant was rapidly visually presented with a series of words or non-words and then asked to indicate whether they recognized a test stimulus from the previous presentation. The exact experimental protocol for all six runs of this Rapid Serial Visual Presentation (RSVP) language task is available on GitHub. Note that as IBC protocols were designed for native French speakers, all stimuli were presented in French.

IBC data acquisition and preprocessing details are described in Pinho et al. (2018), with preprocessing code available from GitHub. Following preprocessing, data were masked using a grey matter mask, detrended, and standardized before being downsampled to 3mm isotropic resolution and smoothed with a 5mm FWHM Gaussian kernel. For the RSVP language task, we then fit a general linear model for each participant to generate trial-wise beta maps. Each run generated 60 beta maps evenly divided between five experimental categories: words, non-words, consonants, sentences, and jabberwocky. To define the data included in our alignment, we used a left lateralized, language-relevant region-of-interest defined in Bazeille et al. (2021) and depicted in Figure 5.3. All subsequently described steps are considering only the 3084 voxels within this region.

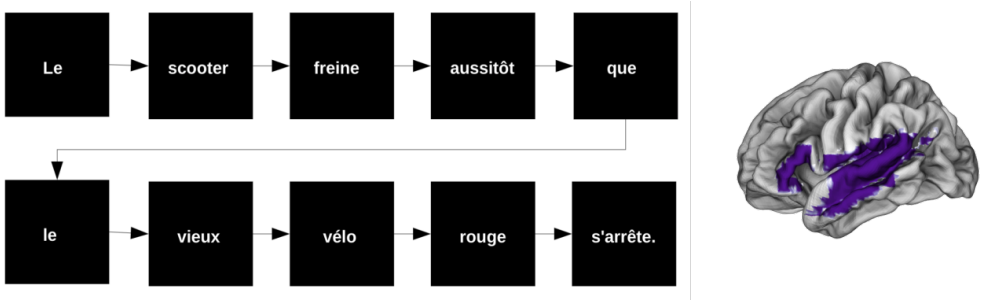

Figure 5.3: **The RSVP Language Task.** A schematic overview of data included in the RSVP language task. On the left, an example experimental trial with a sequentially, visually presented French sentence. On the right, the left-lateralized region of interest included in our analysis.

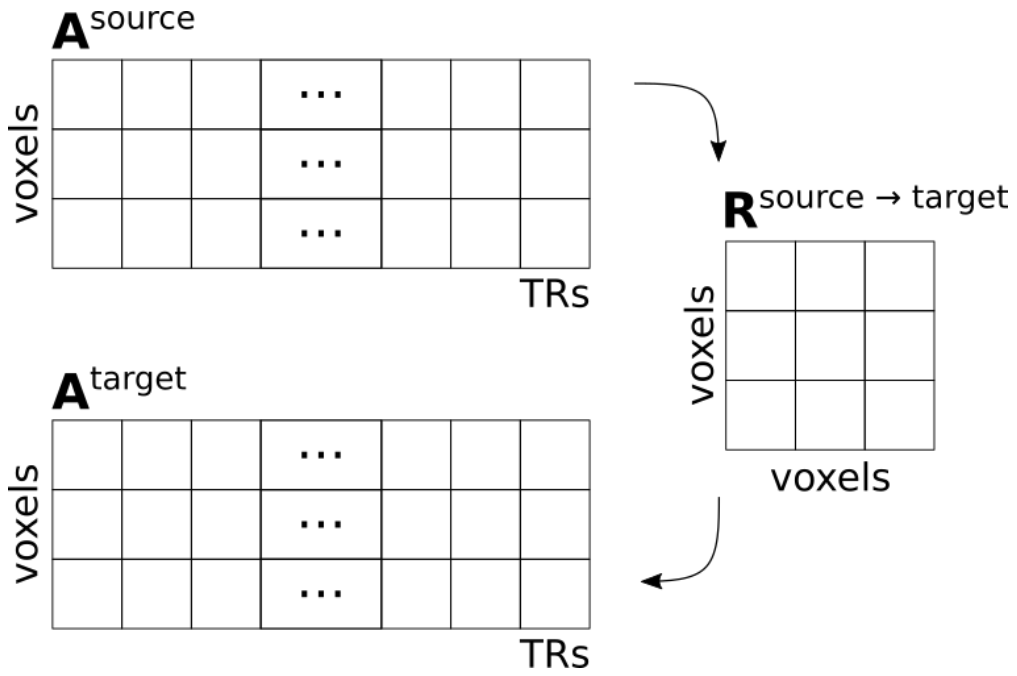

Figure 5.4: **Learning a transformation between participants.** Using alignment data **A** from a source and target participant, we can create data matrices $\mathbf{A}^{source}$ and $\mathbf{A}^{target}$, each of which is of the shape voxels by TRs. In our example, these alignment data are taken from a presentation of *Raiders of the Lost Ark*. We then learn a transformation $\mathbf{R}^{source \rightarrow target}$ between these two matrices. This transformation can then be applied on any dataset from the source subject with the same number of voxels; i.e., on the same ROI.

To train our alignment transformations, we concatenate 2000 TRs of *Raiders of the Lost Ark* viewing into a 2000 TRs by 3084 voxels matrix, with one matrix for each participant's data. Using Procrustes alignment (see Section 5.5.3), we align all-but-one source participants to a single target participant. This alignment learns a transformation matrix **R** for each source-target pair; in our case, this means that we learn eight **R** matrices. This process is depicted in Figure 5.4. We then apply the relevant learned **R** matrix to each of the nine source participant's RSVP language data.

We use inter-subject decoding to evaluate whether functional alignment successfully improves inter-subject similarity in this RSVP language task. Specifically, we train a linear Support Vector Machine to decode experimental categories across participants; for a review of using linear Support Vector Machines and other classifiers with fMRI data, we recommend Pereira et al. (2009). Here, we use the aligned RSVP language statistical maps as features, where each participant has 60 maps for each of six runs. We trained the model to classify each of the five experimental conditions on all source participants; i.e., on 60 x 6 x 8 = 2880 statistical maps. We then tested the trained Support Vector Machine on the trial-wise maps from the target participant's RSVP language data. Note that no

transformation matrix R has been applied to these data, since all other participants were aligned to this target. We repeated this procedure 9 times, iterating the choice of target participant each time. We find an average inter-subject decoding accuracy of 28% after functional alignment.

We can compare this value against the average inter-subject decoding accuracy on unaligned data; that is, by training on the statistical maps from all-but-one participant and testing on the held-out target participant without applying any of the derived R matrices. In this case, we find an average inter-subject decoding accuracy of 28%. Functionally aligning on *Raiders of the Lost Ark* thus brings an average gain of 0% decoding accuracy in this left-lateralized region-of-interest for the RSVP language task.

5.5 Implementation and available decision points

We can consider the preceding example as a series of steps for moving from data processing through algorithm choice and finally to result evaluation. At each of these steps, we have described a single analytic choice. In this section, we detail each step separately and lay out other potential alternative choices, consider how those choices interact, and describe available guidelines for selecting among available alternatives in a given experimental application.

Specifically, we consider: how researchers should define training and testing data, how to select spatial context within these datasets, choosing a functional alignment algorithm to apply, and evaluating the success of the learned alignment.

5.5.1 Defining training and testing data

As with other MVPA methods, functional alignment transformations must be learned on independent training data since training and testing on the same data will yield excessively optimistic estimates of performance. We review the effects of—as well as targeted use cases for—using overlapping data in Section 5.5.1.1. Beyond this requirement, there are several general guidelines driving choice of training dataset: (1) synchronization between presentations, (2) amount of available data, as well as (3) relevance to the downstream application.

Many traditional psychological paradigms may include counterbalancing session-specific stimuli across participants to account for ordering effects. To train functional alignment, however, we want data that are as consistent between participants as possible such that we learn transformations between individual-specific functional organization

rather than between different task conditions. Functional alignment training data should thus be synchronized or co-occurring; i.e., not unconstrained tasks such as resting-state or participant-driven tasks such as video-game play. While we do not consider them here, we note that connectivity-based alignment strategies (e.g., Guntupalli et al., 2018; Nastase et al., 2020b) should be a better option for desynchronized datasets. Importantly, training data do not have to be co-occuring in the original tasks but only when used to train the alignment. For example, Bazeille et al. (2021) use 53 statistical contrast maps to train functional alignment within the IBC dataset. Although these contrast maps are drawn from multiple experimental protocols, they are stacked in an identical order to create a single training matrix of p voxels by 53 conditions. This identical ordering is important in creating an implicit correspondence between $\mathbf{A}^{source}$ and $\mathbf{A}^{target}$, which is assumed across all of the functional alignment algorithms discussed in Section 5.5.3.

The amount of synchronized training data necessary largely depends on the amount of task-relevant signal available in each image for the specific combination of region-of-interest and test data. While as few as 53 statistical contrast maps have been used to successfully drive an alignment (Bazeille et al., 2021), Guntupalli et al. (2016) estimated that at least 30 minutes of *Raiders of the Lost Ark* movie data is necessary to learn robust transformations, with most applications falling somewhere in-between (Geerligs et al., 2021). In general, it is difficult to assess the amount of relevant signal outside of a specific application and associated performance metric (see Section 5.5.4). Researchers leverage previous literature, however, to assess the magnitude of data that should be used for training. In those cases where no relevant guidelines can be established, we recommend using as much data as possible.

Ideally, investigators would jointly select training and testing data that cover similar cognitive domains. For example, using train and test data of similar motor imagery (Al-Wasity et al., 2020) or sub-sampled visual categories (Ho et al., 2022) ensures that transformations learned on the training data will be relevant to the test data. Researchers should avoid learning transformations on data with no clear relevance to the task data; for example, learning alignment on a visual stimulus depicting natural scenes and then applying the learned alignment to a language task. These cross-modal pairings are likely to share very little task-relevant signal, limiting the success of alignment. Naturalistic stimuli are a popular choice for functional alignment, as their multimodal nature, complex temporal structure, and high ecological validity (Nastase et al., 2020a; Sonkusare et al., 2019) are relevant to many higher-order cognitive processes such as memory and attention (Baldassano et al., 2017; Nastase et al., 2017).

5.5.1.1 Learning and applying the alignment on overlapping data

Many investigators who are interested in applying functional alignment may be working with pre-collected datasets that do not include independent task data that can be used to derive an alignment. For example, investigators working with very short movie clips may be unable to effectively learn and apply transformations using independent subsets of the movie if insufficient data is available. In these cases, it is particularly appealing to directly train and apply functional alignment transformations on the same data. Indeed, this procedure has been adopted in the literature as in Geerligs et al. (2021), where hyperalignment transformations were learned and then applied on the same 8-minute movie clip. Using the aligned movie, the authors then segmented neural events within the movie timecourse and compared the consistency of event boundaries across participants. Importantly, the authors themselves acknowledge the limitations of this approach for deriving supervised classifiers, but argue that overlapping data will not influence the temporal structure of the data, which is their measure of interest.

For this and related measures, it is possible that using overlapping data may be appropriate to answer the research question of interest. We caution, however, that the derived transformation cannot be assumed to have successfully improved inter-individual similarity without an explicit test, and it may therefore be appropriate to report the derived measure of interest with and without functional alignment. Regardless, we encourage researchers to explicit report whether alignment was calculated and applied on overlapping data, as this will significantly influence the interpretation of derived transformations.

5.5.2 Selecting spatial context

Without constraints, functional alignment transformations can learn biologically implausible relationships; for example, finding correspondence between the visual cortex of one participant and the frontal cortex of a different participant. To avoid this, functional alignment is commonly constrained to learn transformations within a specific spatial context or neighborhood. In its first introduction in Haxby et al. (2011), functional alignment was learnt within a single region-of-interest as in our preceding example. Since then, other spatial contexts such as searchlights (Guntupalli et al., 2016; Kriegeskorte et al., 2006) and non-overlapping, deterministic parcellations (Bazeille et al., 2021) have also been used to constrain learned alignments.

While region-of-interests are generally chosen based on their relevance to the application task, searchlights and parcellations are used when investigators need to define a

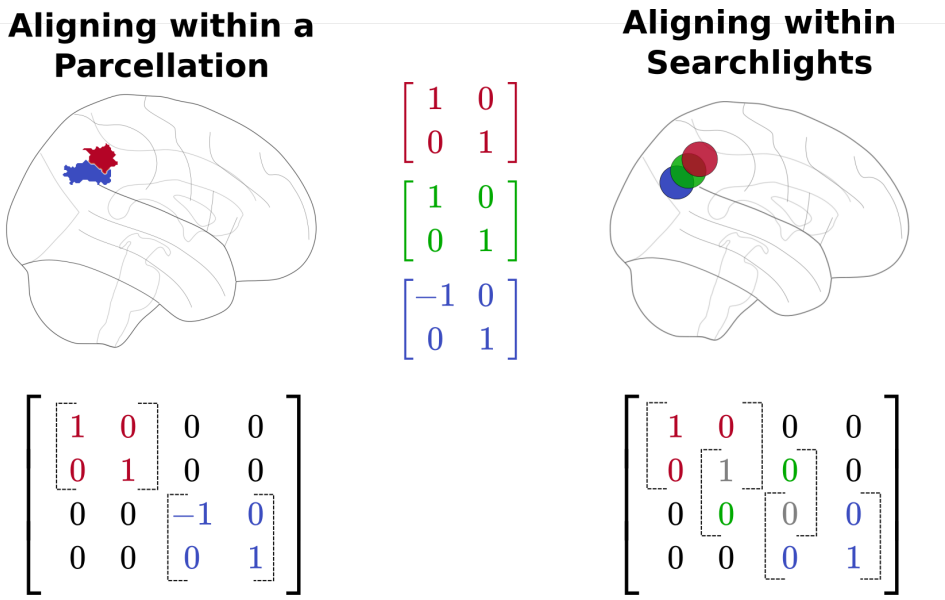

Figure 5.5: **Comparing non-overlapping and overlapping spatial contexts.** On the left, two non-overlapping parcels from a single parcellation. On the right, three overlapping searchlights. The blue transformation differs between the two spatial contexts, resulting in a different overall learned functional alignment. Figure adapted from Bazeille et al. (2021).

whole-brain transformation for function alignment. These methods differ, however, in how individual, neighborhood transformations are combined into a whole-brain context. Importantly, searchlights average overlapping transformations, as shown in Figure 5.5. While this allows investigators to avoid defining areal boundaries, it means that overlapping transformations must be aggregated in some way; for example, by summing or by averaging. The aggregated transformation, then, is no longer guaranteed to reflect the original transformations from which it was calculated.

As a result of this aggregation, we generally recommend that investigators first use either individual regions-of-interest or non-overlapping parcellations, unless their experimental question requires a searchlight approach. In this case, the ideal radius of the searchlight is largely dependent on the research question; however, we recommend the general guidelines for searchlight analyses provided in Etzel et al. (2013). For non-overlapping parcellations, we have found that many function alignment algorithms (see Section 5.5.3) are not strongly dependent on the exact parcellation scheme (Bazeille et al., 2021), though predictions from alignment can be further improved by using bootstrap aggregation across multiple parcellations (Dohmatob et al., 2021). We therefore recommend that researchers use their preferred parcellation for whole-brain functional alignment.

5.5.3 Choosing a functional alignment algorithm

At a high-level, the goal of functional alignment is to make two or more participant activity patterns look as similar as possible. The difference between available functional alignment algorithms are the ways by which we define "as similar as possible," and the exact constraints under which we can transform a given source activity pattern to increase its similarity to a target pattern. In this tutorial, we focus on methods which generate linear mappings for two main reasons. The first is interpretability: since we are constrained to linear transformations, we retain as much information as possible on each individual's activity pattern without more subtle, non-linear changes. The second reason is for downstream applications: because linear-only transformations do not fit as closely to the training data, we can use the same transformations in a new task context with a different structure. That is, we can learn a relationship between $\mathbf{A}^{source}$ and $\mathbf{A}^{target}$ using *Raiders of the Lost Ark* movie data and apply the resulting transformation on $\mathbf{D}^{source}$, statistical contrast maps from an RSVP language task.

In general, we suggest that the advantages of linear-only transformations outweigh the disadvantages of excluding non-linear changes when our goal is to apply the derived transformations in new contexts. In some applications, however, researchers may only be able to assume a weak correspondence between each participant's activity patterns; for example, in functional reorganization following stroke (Langs et al., 2010, 2014) or when making comparisons across species, such as human and non-human primates (Xu et al., 2019). In these cases, researchers may prefer to use non-linear alignment methods, sacrificing interpretability for a more effective alignment. Nonetheless, because the transformations resulting from these non-linear methods are difficult to apply and validate using the framework described above, we do not consider them here.

Although linear methods for functional alignment include a wide variety of algorithms, we can consider them to fall within two general families: direct alignment methods and latent factor methods. Latent factor methods assume that the observed activity patterns are generated by underlying latent factors and that these latent factors—rather than the observed activity—should be shared across participants. Direct alignment methods, on the other hand, assume that observed activity patterns are shared, if differently encoded in the supporting anatomy, across participants. Direct alignment methods include the popular hyperalignment algorithm, which is based on Procrustes analysis (Schönemann, 1966). Other methods in this family include ridge regression (Ho et al., 2022; Tavor et al., 2016) and optimal transport (Bazeille et al., 2019). Latent factor methods include the Shared Response Model (Chen et al., 2015) and regularized Canonical Correlation Analysis (rCCA;

Bilenko and Gallant, 2016).

We provide more mathematical detail on each of these methods in the online material. In brief, within a given family, each of these methods make slightly different constraints on the possible transformations. For example, Procrustes analysis only allows for rigid-body transformations (e.g., rotations and translations), stretching, and shearing. In this way, it matches our intuitive ideas of comparing two geometrical shapes. Optimal transport, by contrast, tries to maximize similarity for each observation based on some predefined cost metric; for example, their anatomical distance on the cortical surface. Importantly, however, these methods—particularly the direct-fit methods—are primarily defined between a single pair of subjects. In order to consider multiple subjects, we need to introduce the idea of functional templates.

5.5.3.1 Contrasting pairwise and template-based alignment

In our illustrative example, we functionally align a source and target participant by learning a direct transformation from source to target. This procedure is broadly known as "pairwise alignment," since transformations are learnt between pairs of participants. This approach is very useful when inferences can be drawn on a single participant; for example, in deep phenotyping datasets with a small number of subjects. By contrast, in many research contexts we may have a large number of participants that we would like to make direct comparisons between. To do so, investigators generally use "template-based alignment." In this case, a functional template or reference space is constructed and all available participants are then transformed into this new space.

Different researchers have adopted different approaches to construct these templates, which can have significant downstream effects. In the hyperalignment algorithm (Guntupalli et al., 2016), a three-pass iterative procedure is used. In the first stage, a single participant is arbitrarily chosen as the initial target. A second participant is aligned to this target using Procrustes analysis. The average of the transformed images from the aligned second participant and the target participant is then used as the target for aligning the third participant, and so on until all participants have been aligned. In the second stage, all participants—including the initial target—are aligned to the final, average target from the first stage, known as the intermediate common space. In the third and final stage, the resulting transformations are applied to each image, and the second step is repeated with the new target space serving as the final reference space. While this approach draws from Generalized Procrustes Analysis (Gower, 1975), it is sensitive to order effects, including in the choice of initial reference participant (Al-Wasity et al., 2020).

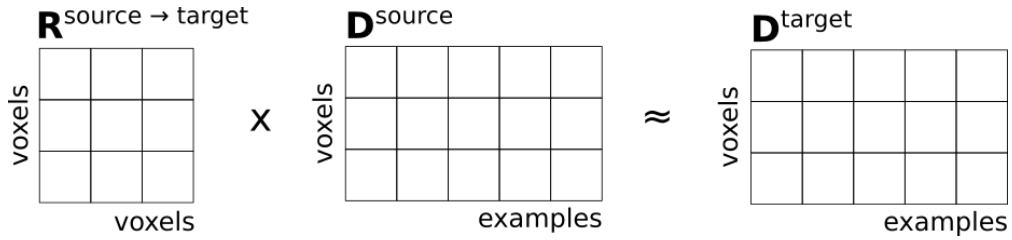

Figure 5.6: **Using learned transformations to improve similarity between the source and target participants.** Using the mapping **R** learned on our alignment data, we can now transform any data from our source participant with the same number of voxels. In this case, we apply it to their collected **D** dataset. $\mathbf{RD}^{source}$ can now be considered an approximation of $\mathbf{D}^{target}$, or a mapping of $\mathbf{D}^{source}$ into the activation space of $\mathbf{D}^{target}$.

Alternative procedures to generate functional templates are in active development (e.g., Bazeille et al., 2019). Importantly, however, latent factor methods such as the Shared Response Model (SRM) avoid this question of template definition since they include a shared decomposition across participants. In this case, new subjects can be projected into the shared decomposition (Chen et al., 2015) to identify their correspondence with existing participants, with the identified latent factors serving as a low-dimensional template for aligning new participants.

5.5.4 Evaluating results

For any transformation, an important question is how to evaluate the results. While anatomical alignments can be visually inspected (as in MRIQC reports; Esteban et al., 2017), functional alignment transformations take place in high-dimensional activation spaces. This high dimensionality complicates visualization; as a result, we need to define useful quantitative metrics.

When applying the calculated transformation **R** to $\mathbf{D}^{source}$, we can consider $\mathbf{RD}^{source}$ as an approximation of $\mathbf{D}^{target}$, as depicted in Figure 5.6. One metric to evaluate the success of functional alignment, then, is to directly compare our $\mathbf{RD}^{source}$ and $\mathbf{D}^{target}$ images using methods such as Pearson correlation or the Dice coefficient.

While these provide a useful estimate of how successfully **R** aligns the functional organization of source and target participants, the range of possible values of e.g. Dice coefficients are likely to vary across different datasets as a result of different acquisition and preprocessing pipelines. For example, in datasets processed with a larger smoothing kernel, source and target images will have slightly higher Dice values than images from a different dataset processed with a smaller smoothing kernel—even without any functional alignment. Thus, these image-based metrics provide useful comparisons within a given

dataset but may be hard to compare across datasets, complicating attempts to make comparisons across different studies or sites.

Alternative quantitative metrics can be defined based on prediction accuracy. That is, using predictive frameworks such as inter-subject decoding allows us to derive a set accuracy value for a given functional alignment application. Unlike image-based metrics, however, these accuracy values can be more directly compared across datasets when properly evaluated against chance and against relevant benchmarks; for example, within-subject alignment or standard, anatomical-only preprocessing. We adopt an inter-subject decoding framework in the illustrative example included in Section 5.4, though other predictive frameworks are also possible. For example, time-segment matching (Haxby et al., 2011; Kumar et al., 2019) is a predictive framework defined for naturalistic data where stimulus labels are not available. In this case, continuous stimuli are divided into small time bins using a sliding window approach. The average time course for all source subjects in this bin is compared against all possible, non-overlapping time bins from a single target subject. The time bin with the maximum correlation is selected. If this bin corresponds to the same time points in both the average source and individual target time course, then that time bin is marked as correct. This procedure is repeated for all time bins, and the average accuracy is reported.

We encourage researchers to adopt predictive models to evaluate the success of functional alignment in those cases where they have access to sufficient held-out test data to train a successful predictive model. If too little data is available, however, image-based metrics such as the Dice coefficient may be a better choice, though researchers should be aware that these metrics may be difficult to compare across datasets. Regardless of the exact metric used, however, we emphasize the importance of adopting some metric to evaluate alignment performance in addition to using the transformation for a downstream scientific question. Much as researchers visually evaluate the success of anatomical alignment before using fMRI data in standard preprocessing pipelines, quantitatively evaluating the success of functional alignment provides insight into how well the calculated transformations improve inter-subject similarity, which can then be used in a range of applications.

5.6 Applications and extensions

While our preceding sections overview functional alignment as a technique—and important implementation choices—we have not yet addressed a fundamental issue: what

kinds of scientific questions are amenable to functional alignment? How does functional alignment interact as a processing technique with different downstream scientific analyses?

Our illustrative example pairs functional alignment with a multivariate classification analysis The same example can also be considered with other univariate methods of analysis; for example, we can analyze the RSVP language task with a General Linear Model (GLM) after functionally aligning using transformations learned from the *Raiders of the Lost Ark* dataset. In both cases, however, we are learning functional alignment transformations using evoked activity in the training dataset. Importantly, activity is not the only measure on which we can learn these transformations. For example, several research groups have adopted connectivity-based functional alignment, building on algorithms discussed in Section 5.5.3 including hyperalignment (Busch et al., 2020; Guntupalli et al., 2018) and the Shared Response Model (Nastase et al., 2020b). For connectivity-based alignment, we replace the "activation spaces" discussed in Section 5.3.1 with new "functional spaces," where each point in functional space is the strength of functional connectivity between two spatial neighborhoods (e.g., searchlights in Guntupalli et al., 2018). Defining functional spaces in this way significantly increases the kinds of stimuli we can use to learn functional alignment from shared, synchronized stimuli to unsynchronized data such as resting-state or even non-overlapping narrative stimuli. Dadashkarimi et al. (2021) and colleagues took this idea one step further, using optimal transport to map between connectomes derived from different parcellations rather than functionally aligning the voxelwise time courses directly.

In addition to considering new measures for functional alignment, we can also broaden our potential use cases. Beyond individual-level mappings, similar questions of correspondence underlie processes unfolding over time such as attention, memory, and learning. In systems neuroscience, for example, functional alignment is already used to stabilize readouts across multi-session recordings (Gallego et al., 2020). As the amount of data available for individual participants continues to increase in new deep-phenotyping datasets (Naselaris et al., 2021; Poldrack, 2017)—such as the Individual Brain Charting dataset included in our example analysis (Pinho et al., 2018)—these multi-session alignments may become similarly accessible in social and cognitive neuroscience applications. We caution, though, that fMRI lacks the high temporal resolution of many systems neuroscience measurements, meaning that we may be less able to align the dynamics supporting attention and learning. Nonetheless, applications of functional alignment to other modalities such as intracranial electroencephalography (iEEG) as in Xie et al. (2021) confirm the potential of this approach.

5.7 Relationship to other alignment methods

While functional alignment promises continued insight into mappings between individual's fMRI measurements, other approaches have also been pioneered to improve inter-individual correspondence. To better situate alignment within the field, we briefly consider its relationship to two other methods: surface-based registration and individualized parcellations. We provide brief summaries of each method below.

5.7.1 Surface-based registration

Although volumetric registration remains widely-used within the community, this may lead to functionally significant mis-alignment of sulco-gyral patterns (Coalson et al., 2018). Since sulco-gyral patterns are closely tied to functional areas (Desai et al., 2005), volumetric registration may exacerbate misaligned functional activations across participants. Surface registration techniques have been proposed to instead align directly on sulco-gyral patterning, representing cortex as either a two-dimensional sheet (Van Essen et al., 1998) or as a three-dimensional sphere (Fischl et al., 1999; Yeo et al., 2010). Surface-based alignment has been shown to improve group-level analyses on the cortical surface in both univariate (Tucholka et al., 2012) and multivariate (Oosterhof et al., 2011) contexts.

While surface-based registration provides important improvements in mapping across neuroanatomy, it does not broadly address variable structure-function correspondence, instead improving correspondence only for functional areas that closely track sulco-gyral patterning. Several researchers have therefore advocated for pairing surface-based registration with functional alignment (Guntupalli et al., 2016; cf. Bazeille et al., 2021). Additionally, functional alignment can uniquely address functional variability within subcortical or allocortical structures such as the hippocampus (Chen et al., 2021) which are commonly excluded from surface-based registration.

5.7.2 Individualized parcellations

Unlike functional alignment, which can be applied across a range of spatial contexts, individualized parcellations work within a given parcellation scheme. For example, the individualized parcellation method proposed by Kong et al. (2018) works by assigning a given voxel to one parcel or another to maximize similarity of network-level functional connectivity patterns across participants. In creating functional connectomes, individualized parcellations have been shown to better account for individual-level differences in

cognition, emotion, and personality, reflecting the idea that functional areas captured by these parcellations show stable inter-individual differences (Gratton et al., 2018).

While individualized parcellations could in theory be used as spatial contexts for functional alignment, many of these methods yield a variable number of voxels per parcel across participants. The functional alignment algorithms considered here assume a consistent number of voxels across participants, limiting their use with individualized parcellations. Bootstrap aggregation across different parcellations, however, has been shown to improve the ability of alignment to predict individual task statistical maps (Dohmatob et al., 2021). Future work combining individualized parcellations and functional alignment is therefore likely to be of significant interest. In cases when only one method can be adopted, however, we recommend that researchers use individualized parcellations for analyses where they intend to extract parcel-level measures, while functional alignment may be more useful for MVPA or general-linear model analyses which rely on voxel-level information.

5.8 Conclusions

In this tutorial, we have introduced functional alignment as an accessible method for social and cognitive neuroscience through an illustrative example. We explained the intuitions supporting functional alignment as well as the available choices to guide its application. We further provided a few general recommendations for researchers, and we compared alignment to other methods for improving inter-individual correspondence to help situate its potential impact within the field.

Over the last decade, functional alignment has emerged as a powerful method for finding inter-individual similarity across social and cognitive neuroscience. The next decade promises more work in this direction, with new algorithms and applications for this toolset. While we have tried to point to several of the most promising developments in this area, other reviews such as Haxby et al. (2020) provide additional detail on several of the methods that we discussed here. We further direct readers to our accompanying online resource `https://neurodatascience.github.io/fmralign-tutorials` for formalisms and interactive Python code examples for five commonly-used functional alignment methods.

5.9 Bibliography

Al-Wasity, S. et al. (2020). "Hyperalignment of motor cortical areas based on motor imagery during action observation". *Sci. Rep.*, 10(1), p. 5362.

Baldassano, C. et al. (2017). "Discovering Event Structure in Continuous Narrative Perception and Memory". *Neuron*, 95(3), 709–721.e5.

Bazeille, T et al. (2019). "Local Optimal Transport for Functional Brain Template Estimation". In: *Information Processing in Medical Imaging*. Springer International Publishing, pp. 237–248.

Bazeille, T. et al. (2021). "An empirical evaluation of functional alignment using inter-subject decoding". *Neuroimage*, p. 118683.

Bilenko, N. Y. and J. L. Gallant (2016). "Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging". *Front. Neuroinform.*, 10, p. 49.

Brett, M., I. S. Johnsrude, and A. M. Owen (2002). "The problem of functional localization in the human brain". *Nat. Rev. Neurosci.*, 3(3), pp. 243–249.

Busch, E. L. et al. (2020). "Hybrid Hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity".

Chen, H.-T., J. R. Manning, and M. A. A. van der Meer (2021). "Between-subject prediction reveals a shared representational geometry in the rodent hippocampus". *Curr. Biol.*, 31(19), 4293–4304.e5.

Chen, P.-H. et al. (2015). "A Reduced-Dimension fMRI Shared Response Model". In: *Advances in Neural Information Processing Systems 28*. Ed. by C Cortes et al. Curran Associates, Inc., pp. 460–468.

Churchland, P. M. (1998). "Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered". *J. Philos.*, 95(1), pp. 5–32.

Coalson, T. S., D. C. Van Essen, and M. F. Glasser (2018). "The impact of traditional neuroimaging methods on the spatial localization of cortical areas". *Proc. Natl. Acad. Sci. U. S. A.*, 115(27), E6356–E6365.

Dadashkarimi, J., A. Karbasi, and D. Scheinost (2021). "Data-driven mapping between functional connectomes using optimal transport".

Desai, R. et al. (2005). "Volumetric vs. surface-based alignment for localization of auditory cortex activation". *Neuroimage*, 26(4), pp. 1019–1029.

Dohmatob, E. et al. (2021). "Brain topography beyond parcellations: Local gradients of functional maps". *Neuroimage*, 229, p. 117706.

Esteban, O. et al. (2017). "MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites". *PLoS One*, 12(9), e0184661.

Etzel, J. A., J. M. Zacks, and T. S. Braver (2013). "Searchlight analysis: promise, pitfalls, and potential". *Neuroimage*, 78, pp. 261–269.

Finn, E. S. et al. (2019). "Idiosynchrony: From shared responses to individual differences during naturalistic neuroimaging".

Fischl, B et al. (1999). "High-resolution intersubject averaging and a coordinate system for the cortical surface". *Hum. Brain Mapp.*, 8(4), pp. 272–284.

Gallego, J. A. et al. (2020). "Long-term stability of cortical population dynamics underlying consistent behavior". *Nat. Neurosci.*, 23(2), pp. 260–270.

Geerligs, L., M. van Gerven, and U. Güçlü (2021). "Detecting neural state transitions underlying event segmentation". *Neuroimage*, 236, p. 118085.

Gower, J. C. (1975). "Generalized procrustes analysis". *Psychometrika*, 40(1), pp. 33–51.

Gratton, C. et al. (2018). "Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation". *Neuron*, 98(2), 439–452.e5.

Guntupalli, J. S., M. Feilong, and J. V. Haxby (2018). "A computational model of shared fine-scale structure in the human connectome". *PLoS Comput. Biol.*, 14(4), e1006120.

Guntupalli, J. S. et al. (2016). "A Model of Representational Spaces in Human Cortex". *Cereb. Cortex*, 26(6), pp. 2919–2934.

Hanke, M. et al. (2009). "PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data". *Neuroinformatics*, 7(1), pp. 37–53.

Haxby, J. V. et al. (2011). "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron*, 72(2), pp. 404–416.

Haxby, J. V. et al. (2020). "Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies". *Elife*, 9, e56601.

Ho, J. K. et al. (2022). "Inter-individual deep image reconstruction".

Kong, R. et al. (2018). "Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion". *Cereb. Cortex*.

Kriegeskorte, N., R. Goebel, and P. Bandettini (2006). "Information-based functional brain mapping". *Proc. Natl. Acad. Sci. U. S. A.*, 103(10), pp. 3863–3868.

Kumar, M. et al. (2019). "BrainIAK tutorials: user-friendly learning materials for advanced fMRI analysis".

Langs, G. et al. (2010). "Functional Geometry Alignment and Localization of Brain Areas". *Adv. Neural Inf. Process. Syst.*, 1, pp. 1225–1233.

Langs, G. et al. (2014). "Decoupling function and anatomy in atlases of functional connectivity patterns: language mapping in tumor patients". *Neuroimage*, 103, pp. 462–475.

Markiewicz, C. J. et al. (2021). "The OpenNeuro resource for sharing of neuroscience data". *Elife*, 10.

Naselaris, T., E. Allen, and K. Kay (2021). "Extensive sampling for complete models of individual brains". *Current Opinion in Behavioral Sciences*, 40, pp. 45–51.

Nastase, S. A., A. Goldstein, and U. Hasson (2020a). "Keep it real: rethinking the primacy of experimental control in cognitive neuroscience". *Neuroimage*, 222, p. 117254.

Nastase, S. A. et al. (2017). "Attention Selectively Reshapes the Geometry of Distributed Semantic Representation". *Cereb. Cortex*, 27(8), pp. 4277–4291.

Nastase, S. A. et al. (2020b). "Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space". *Neuroimage*, 217, p. 116865.

Oosterhof, N. N. et al. (2011). "A comparison of volume-based and surface-based multi-voxel pattern analysis". *Neuroimage*, 56(2), pp. 593–600.

Pereira, F., T. Mitchell, and M. Botvinick (2009). "Machine learning classifiers and fMRI: a tutorial overview". *Neuroimage*, 45(1 Suppl), S199–209.

Pinho, A. L. et al. (2018). "Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping". *Sci Data*, 5, p. 180105.

Poldrack, R. A. (2017). "Precision Neuroscience: Dense Sampling of Individual Brains". *Neuron*, 95(4), pp. 727–729.

Popal, H., Y. Wang, and I. R. Olson (2019). "A Guide to Representational Similarity Analysis for Social Neuroscience". *Soc. Cogn. Affect. Neurosci.*, 14(11), pp. 1243–1253.

Rademacher, J et al. (1993). "Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology". *Cereb. Cortex*, 3(4), pp. 313–329.

Rustamov, R. M. and L. Guibas (2016). "Hyperalignment of Multi-subject fMRI Data by Synchronized Projections". In: *Machine Learning and Interpretation in Neuroimaging*. Springer International Publishing, pp. 115–121.

Schönemann, P. H. (1966). "A generalized solution of the orthogonal procrustes problem". *Psychometrika*, 31(1), pp. 1–10.

Sonkusare, S., M. Breakspear, and C. Guo (2019). "Naturalistic Stimuli in Neuroscience: Critically Acclaimed". *Trends Cogn. Sci.*

Tavor, I et al. (2016). "Task-free MRI predicts individual differences in brain activity during task performance". *Science*, 352(6282), pp. 216–220.

Thirion, B. et al. (2006). "Dealing with the shortcomings of spatial normalization: Multi-subject parcellation of fMRI datasets". *Hum. Brain Mapp.*, 27(8), pp. 678–693.

Tucholka, A. et al. (2012). "An empirical comparison of surface-based and volume-based group studies in neuroimaging". *Neuroimage*, 63(3), pp. 1443–1453.

Van Essen, D. C. et al. (1998). "Functional and structural mapping of human cerebral cortex: solutions are in the surfaces". *Proc. Natl. Acad. Sci. U. S. A.*, 95(3), pp. 788–795.

Weaverdyck, M. E., M. D. Lieberman, and C. Parkinson (2020). "Tools of the Trade Multivoxel pattern analysis in fMRI: a practical introduction for social and affective neuroscientists". *Soc. Cogn. Affect. Neurosci.*, 15(4), pp. 487–509.

Xie, T. et al. (2021). "Minimal functional alignment of ventromedial prefrontal cortex intracranial EEG signals during naturalistic viewing".

Xu, T. et al. (2019). "Cross-species Functional Alignment Reveals Evolutionary Hierarchy Within the Connectome".

Yeo, B. T. T. et al. (2010). "Spherical demons: fast diffeomorphic landmark-free surface registration". *IEEE Trans. Med. Imaging*, 29(3), pp. 650–668.

Chapter 6

Beyond advertising: New infrastructures for publishing integrated research objects

Elizabeth DuPre[1], Chris Holdgraf[2,3], Agah Karakuzu[4,5], Loïc Tetrel[6], Pierre Bellec[6,7], Nikola Stikov[4,5], Jean-Baptiste Poline[1]

[1]NeuroDataScience - ORIGAMI laboratory, McGill University, Montreal, Quebec, Canada

[2]The International Interactive Computing Collaboration (2i2c), Berkeley, California, USA

[3]International Computer Science Institute, Berkeley, California, USA

[4]NeuroPoly Lab, Polytechnique Montreal, Montreal, Quebec, Canada

[5]Montreal Heart Institute, Montreal, Quebec, Canada

[6]Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montreal, Quebec, Canada

[7]Department of Psychology, Université de Montréal, Montreal, Quebec, Canada

Published in:

PLOS Computational Biology: https://doi.org/10.1371/journal.pcbi.1009651

6.1 Preface

Publishing interactive code and data significantly facilitates exploration and adoption of methods such as functional alignment. Being able to directly build on another researcher's work—without re-implementing their described processing and analyses—allows for more

rapid iteration across the field. For example, the experimental code used in Chapters 3 and 4 benefited directly from Python packages such as PyMVPA and BrainIAK. Despite the clear utility of such resources, it is difficult to publish interactive research articles. As the infrastructure necessary to support them is not yet available, these materials largely exist outside of the traditional publishing platforms, often appearing instead as documentation for specific software packages. This limits their scope to those algorithms directly implemented in a given tool, preventing broader consideration of a given method. Further, it disincentivizes researchers from developing scientific articles with integrated code and data, such as the framework for functional alignment presented in Chapter 5. In this chapter, I review the current status of scientific publishing infrastructure and propose novel directions for development. Several of the future directions suggested in this piece have been actively developed to support the interactive materials accompanying Chapter 5. This work was published in *PLOS Computational Biology* in 2022 (DuPre et al., 2022).

6.2 Abstract

Moving beyond static text and illustrations is a central challenge for scientific publishing in the twenty-first century. As early as 1995, Donoho and Buckheit paraphrased John Claerbout that "an article about [a] computational result is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result" (Donoho, 2010). Awareness of this problem has only grown over the last 25 years; nonetheless, scientific publishing infrastructures remain remarkably resistant to change (Piotrowski, 2016). Even as these infrastructures have largely stagnated, the internet has ushered in a transition "from the wet lab to the web lab" (Keshavan and Poline, 2019). New expectations have emerged in this shift, but these expectations must play against the reality of currently available infrastructures and associated sociological pressures. Here, we compare current scientific publishing norms against those associated with online content more broadly, and we argue that meeting the "Claerbout challenge" of providing the full software environment, code, and data supporting a scientific result will require open infrastructure development to create environments for authoring, reviewing and accessing interactive research objects.

6.3 Publishing as curating, promoting, and archiving content

Scientific publishing platforms—traditionally, scientific journals—fulfill a variety of roles in their communities. Three of the most prominent of these are curating, promoting, and archiving research. Although these roles have adapted to online spaces, they have not been fundamentally reshaped. Indeed, contemporary scientific articles are disseminated primarily as PDFs, directly translating paper-based workflows into digital workspaces. Here, we briefly review how publishing fulfills these roles today: curation via peer review, short-term promotion via online dissemination, and long-term access via archiving.

Across many kinds of media, curating online content is challenging both due to its scale and its style of interaction, which often blurs the boundary between creating and consuming information. For scientific publishing, formal and independent peer review is widely considered to be a key demarcation (Mulligan et al., 2013) and provides an immediate mechanism to curate research objects. Curation in peer review involves checks on a submission's ethical and scientific rigor, in addition to its relevance to a particular research community. Even as many other forms of curation are possible—including crowd-sourced or algorithmically-driven (Yarkoni, 2012)—these remain relatively uncommon in neuroscience (cf. `arxiv-sanity.com`).

In addition to curating (i.e. reviewing and selecting) research objects, publishing also serves an important role in promoting and archiving content. This occurs in the short term through activities such as website hosting and advertising on social media platforms (Klar et al., 2020). Ongoing promotion to an ever evolving scientific community is enabled through the long term archiving and the references system. These roles can be fulfilled independently or in an arbitrary order. For example, online interactions have allowed peer review to expand into post-publication peer review on platforms such as PubPeer (`https://pubpeer.com`) and Sciety (`https://sciety.org`; Stern and O'Shea, 2019).

Even as scientific publishers have successfully moved online, they have not yet embraced the full potential of web-first workflows. We briefly review how two norms of online content, connectivity and interactivity, are currently reflected in scientific publishing before arguing for infrastructure that allows for more directly interactive and re-usable content.

6.4 Rich linking for research objects: Connecting through hybrid content types

Much of the rich, content-driven interactivity of the web depends on access to structured data such as user content on social media platforms. To separate out this content from its presentation, data formats such as XML have been developed to link online content with its supporting resources (Guha et al., 2015). Although scientific publishing workflows are largely built around the XML format, the need to output PDF documents means that resources that cannot be directly embedded—such as executable code or supporting data—have been largely excluded from academic publishing. Thus, the scientific narrative has historically been detached from its other associated research objects.

Recently, growing awareness of this problem has led to an increase in publishing what we term 'hybrid research objects.' Hybrid research objects are distinct from traditional publications in that they make multiple content types available in the same object; that is, they contain narrative text and at least one or more examples of code, data, and computation (e.g., Eglen et al., 2017). Multiple paths exist to make these objects available. One path is to include direct links to each resource such as through data and code availability statements (Colavizza et al., 2020), without constraining their format or content. Alternatively, some publishers require that linked research objects adhere to specified standards and are explicitly included in the review process. For example, the journal Scientific Data from Nature Research publishes descriptors of datasets (Poline, 2019) that include links to dedicated, domain-relevant data hosting infrastructure such as OpenNeuro (`https://openneuro.org`). Importantly, this raises new questions on how to appropriately handle their peer review; questions for which there is no current consensus (Carpenter, 2017).

As hybrid research objects have become more prominent, best practices in publishing these objects continue to evolve. We hope to see more hybrid research objects where each linked object is formatted with domain-relevant standards (e.g., neuroimaging data organized according to a domain-standard such as the Brain Imaging Data Structure [BIDS]; Gorgolewski et al., 2016) and bi-directionally linked using persistent identifiers. Nonetheless, because the linked research objects are hosted on unique platforms without clear checks on interoperability across the hybrid object components, it can be difficult to interact with the code, data, or their combination; for example, when trying to perform minimal quality checks on a dataset. It further prevents eventual readers from assessing the reproducibility or generalizability (The Turing Way Community, 2021) of presented

results. Enhancing this experience requires making these research objects interoperable, improving their reusability. Here, we introduce the idea of 'integrated' research objects to explicitly test the interaction of included research objects in reproducing a scientific result.

6.5 Bridging the gaps: Interactive and integrated research objects

Interactivity is an attractive feature of online content, and one that scientists have been especially eager to adopt (Perkel, 2021). This enthusiasm has spurred development of platforms such as Bokeh (https://bokeh.org) and Plotly (https://plotly.com), enabling scientists to provide multiple views of their data through interactive figures and dashboards. Although this work is impressive, it is limited: researchers remain unable to modify or re-execute the code used to generate these figures when shared through HTML documents. This hinders deep engagement with the presented results.

Achieving deeper interactivity requires interaction between the code, data, and computation supporting a scientific result. One approach to achieve this is to focus on what we call 'integrated research objects.' Integrated research objects not only make multiple kinds of research objects available and tightly coupled, but they do so in formats (e.g. computational notebooks) that foreground their interaction by allowing re-execution. In doing so, they offer a clear answer to the Claerbout challenge.

There are limits on the kinds of experiments that can be supported through integrated research objects; for example, experiments relying on cell cultures or other biological samples may only have digital representations of the statistical analyses and end results rather than the experiments themselves. Nonetheless, researchers should be encouraged to provide access to research objects that can be digitized. This is particularly important for computational work, where experiments are carried out in silico and so computation and the resulting narrative are closely linked.

Despite their immediate appeal, the infrastructure required to support integrated research objects is less straightforward. In particular, authoring, curating, and archiving these research objects all introduce significant challenges. Further, requiring that these objects be archivable imposes strong constraints on the kinds of technologies that can be used. Most archival services discourage submitting complex HTML objects with external dependencies as these documents are unlikely to retain their full functionality with evolving versions of HTML, JavaScript and web browsers (Davis, 2011).

To sidestep this concern, current pilots for publishing integrated research objects consider them as secondary to a traditional, archivable article. For example, eLife authors can develop additional, web-first materials to accompany their accepted research articles. Co-developed with Stencila (https://stenci.la), these Executable Research Articles (ERAs) inherit their structure from the Jupyter notebook (Kluyver et al., 2016) format. ERA development has explicitly focussed on improving the authoring experience, and authors are supported in ensuring that all relevant code and data files are included in the ERA environment. While this support reduces the technical barrier in creating integrated research objects, it also means that ERAs are necessarily only developed at the end of the publication process after scientific analyses are finalized. In this way, the traditional, narrative-text-based document remains privileged as the primary research object.

Centering integrated research objects will require infrastructure development to both ease the authoring experience as well as represent these objects in an archivable format. Although several standards for integrated research objects could serve as potential starting points, we argue that sustainable development demands open standards with multi-stakeholder governance and leadership to ensure that resulting specifications are not driven by a single stakeholder.

6.6 Authoring integrated research objects with open standards

Perhaps the two most broadly adopted standards for integrated research objects are the RMarkdown (https://rmarkdown.rstudio.com) and Jupyter notebook (Kluyver et al., 2016) formats. Both technologies allow researchers to create integrated research objects that include narrative text, code, and computation, though they do so using different internal implementations. Specifically, RMarkdown is based on YAML and markdown formats, while Jupyter notebook is based on the JSON format.

Recent development on Jupyter Book (https://jupyterbook.org) has led to the creation of a MyST markdown format (https://myst-parser.readthedocs.io) that extends Jupyter to build from a combination of YAML and markdown, improving handling for scientific publishing use cases. Thus, RMarkdown and MyST allow researchers to directly describe their scholarship—the code, data, and computation that support a given scientific result—such that it can be easily source-controlled and archived. They each also enable generation of user-focused HTML and PDF documents, including PDFs formatted for

several major scientific journals (using e.g. 'rticles', RStudio, 2021), from user-provided markdown content.

These technologies differ, however, in that RMarkdown development is controlled by a single-stakeholder, RStudio. Although its product is openly licensed, developed with community consultation, and freely available, decision-making power rests with RStudio employees. This model is distinct from multi-stakeholder governance, in which formats are not controlled by individual entities but instead benefit from consensus across organizations. We thus focus on standards developed within the Jupyter ecosystem.

Open standards development within Jupyter has enabled other initiatives such as Stencila and Curvenote (https://curvenote.com) to overlay with additional views and functionality. Integrating these technologies into existing standards (e.g. the Journal Article Tag Suite [JATS] XML format) via translation or conversion processes remains an active area of work. Perhaps their largest departure from existing formats, however, is that they can be re-executed in an integrated computational environment that includes the supporting data files.

6.7 Centering complex objects in scientific publishing with cloud infrastructure

Cloud infrastructure enables browser-based access to computational environments. A major challenge in extending these cloud infrastructures for scientific publishing is the associated cost, both for initial peer review as well as for the long term preservation of included research objects. User-focused cloud technologies such as Binder (https://mybinder.org; Project Jupyter et al., 2018) enable easy access to these environments, but they do not directly address dataset storage. Neuroscience datasets may involve terabytes of data and hundreds of CPU hours of compute time, making cloud computing and data hosting non-trivial. Including multiple versions of a given dataset—from raw data to analysis-ready derivatives—only compounds this problem.

Creating economically viable, non-commercial options will likely involve the coordination of multiple academic and non-profit groups such as the International Interactive Computing Collaboration (2i2c, https://2i2c.org) as well as explicit funding calls for projects advancing open standards through modular, composable infrastructure. Large field-standard datasets, such as those provided by the Allen Institute for Brain Science (https://alleninstitute.org) or the International Brain Laboratory (International Brain Laboratory, 2017), are likely to further benefit from centralized data and computation. This

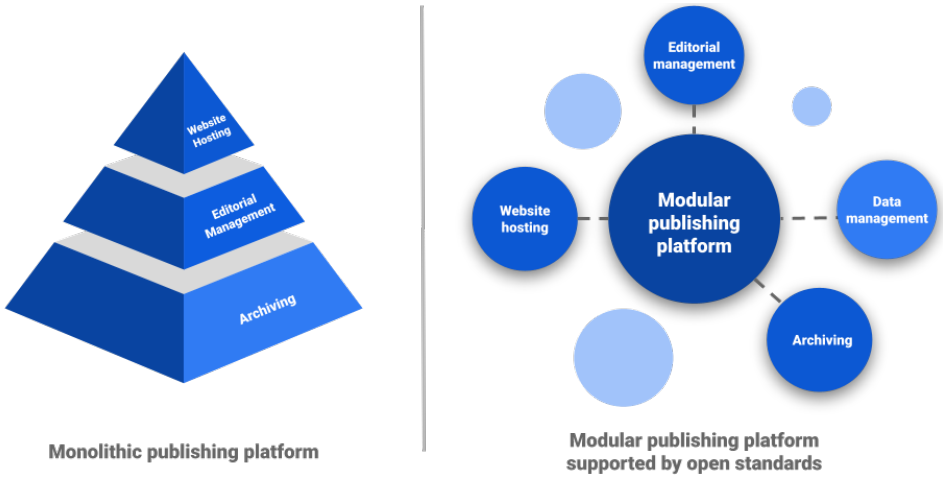

Figure 6.1: **Contrasting monolithic and modular publishing platforms.** While monolithic publishing platforms are self-contained, modular publishing platforms rely on open standards across composable infrastructure. In doing so, they create space for additional functionality such as data management that better supports scientific communities.

approach has been pioneered in geosciences by the Pangeo project (Odaka et al., 2020), which provides centralized access to and computation on field-standard climatology data via JupyterHubs hosted on commercial clouds. Recently, Rokem et al. 2021 have prototyped this approach in neuroscience through the development of a Pan-neuro initiative, encouraging optimism about future adoption in other scientific communities.

Smaller datasets collected by individual research groups, however, may require alternative approaches; in particular, decentralized data management offers a promising route forward to minimize reliance on a central hosting service in those cases where datasets are small enough to be duplicated (Hanke et al., 2021). NeuroLibre (`https://neurolibre.com`) provides one example of this model and relies on non-profit support to host a curated collection of datasets, each of which support one or more NeuroLibre publications through hosted environments for re-executing the described analyses.

Although different in scale, we argue that both Pangeo and NeuroLibre share a core approach that should be more broadly adopted. By investing in infrastructure for integrated research objects that heavily relies on open, modular components, we can make strong contributions in individual research domains while still ensuring that these investments can be easily re-tooled and extended. Fig 6.1 contrasts this modular, composable infrastructure with more traditional publishing platforms developed on a monolithic technology stack.

NeuroLibre, for example, relies on a combined technology stack from the Journal of Open Source Software (JOSS; Katz et al., 2018), Jupyter Book, and BinderHub. Each of these projects independently combines modular technologies to meet existing community

needs, and their combination—while currently unique to neuroscience—can easily be repurposed for other research communities, such as the development of Pan-neuro from the Pangeo model.

As scientists increasingly recognize the value in sharing their code and data (Boudreau et al., 2021), this approach could facilitate an important transition in scientific publishing. By leveraging MyST as an emerging standard for integrated research objects, alongside modular components for their hosting and re-execution through BinderHub and other open technologies, scientists will be better positioned to author articles which center all the research objects supporting a scientific result, in addition to the underlying narrative.

As science increasingly depends on digital infrastructure, it is clear that scientific publishing is at an inflection point. Reckoning with the Claerbout challenge will require providing access to the research objects supporting the actual scholarship rather than the "advertising" of static scientific articles. Adopting web-based technologies provides the strongest possible path forward, but managing this transition in the face of economic and sociological pressure requires academic communities to advocate for open and sustainable infrastructure development, as seen in the Pan-neuro and NeuroLibre initiatives. We argue that community-based efforts around open standards, modular and composable infrastructures, and new research object types will underpin the full potential of web-driven publishing.

6.8 Bibliography

Boudreau, M. et al. (2021). "On the open-source landscape of PLOS Computational Biology". *PLoS Comput. Biol.*, 17(2), e1008725.

Carpenter, T. A. (2017). "What Constitutes Peer Review of Data: A survey of published peer review guidelines". *CoRR*, abs/1704.02236.

Colavizza, G. et al. (2020). "The citation advantage of linking publications to research data". *PloS one*, 15(4), e0230416.

Davis, R. C. (2011). *Five Tips for Designing Preservable Websites*. https://siarchives.si.edu/blog/five-tips-designing-preservable-websites. Accessed: 2021-11-19.

Donoho, D. L. (2010). "An invitation to reproducible computational research". *Biostatistics*, 11(3), pp. 385–388.

DuPre, E. et al. (2022). "Beyond advertising: New infrastructures for publishing integrated research objects". *PLoS Comput. Biol.*, 18(1), e1009651.

Eglen, S. J. et al. (2017). "Toward standard practices for sharing computer code and programs in neuroscience". *Nat. Neurosci.*, 20(6), pp. 770–773.

Gorgolewski, K. J. et al. (2016). "The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments". *Scientific Data*, 3, p. 160044.

Guha, R. V., D. Brickley, and S. MacBeth (2015). "Schema.org: Evolution of Structured Data on the Web: Big data makes common schemas even more necessary". *Queueing Syst.*, 13(9), pp. 10–37.

Hanke, M. et al. (2021). "In defense of decentralized research data management". *Neuroforum*, 27(1), pp. 17–25.

International Brain Laboratory (2017). "An International Laboratory for Systems and Computational Neuroscience". *Neuron*, 96(6), pp. 1213–1218.

Katz, D. S., K. E. Niemeyer, and A. M. Smith (2018). "Publish your software: introducing the journal of open source software (JOSS)". *Comput. Sci. Eng.*

Keshavan, A. and J.-B. Poline (2019). "From the Wet Lab to the Web Lab: A Paradigm Shift in Brain Imaging Research". *Front. Neuroinform.*, 13, p. 3.

Klar, S. et al. (2020). "Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work". *PLoS One*, 15(4), e0229446.

Kluyver, T. et al. (2016). "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Amsterdam, NY: IOS Press, pp. 87–90.

Mulligan, A., L. Hall, and E. Raphael (2013). "Peer review in a changing world: An international study measuring the attitudes of researchers". *J. Am. Soc. Inf. Sci. Technol.*, 64(1), pp. 132–161.

Odaka, T. E. et al. (2020). "The Pangeo Ecosystem: Interactive Computing Tools for the Geosciences: Benchmarking on HPC". In: *Tools and Techniques for High Performance Computing*. Springer International Publishing, pp. 190–204.

Perkel, J. M. (2021). "Reactive, reproducible, collaborative: computational notebooks evolve". *Nature*, 593(7857), pp. 156–157.

Piotrowski, M. (2016). "Future Publishing Formats". In: *Proceedings of the 2016 ACM Symposium on Document Engineering*. DocEng '16. New York, NY, USA: Association for Computing Machinery, pp. 7–8.

Poline, J.-B. (2019). "From data sharing to data publishing". *MNI Open Res*, 2.

Project Jupyter et al. (2018). "Binder 2.0 - Reproducible, interactive, sharable environments for science at scale". In: *Proceedings of the 17th Python in Science Conference*, pp. 113–120.

Rokem, A. et al. (2021). "Pan-neuro: interactive computing at scale with BRAIN datasets".

RStudio (2021). *rticles*. https://github.com/rstudio/rticles.

Stern, B. M. and E. K. O'Shea (2019). "A proposal for the future of scientific publishing in the life sciences". *PLoS Biol.*, 17(2), e3000116.

The Turing Way Community (2021). *The Turing Way: A handbook for reproducible, ethical and collaborative research*.

Yarkoni, T. (2012). "Designing next-generation platforms for evaluating scientific output: what scientists can learn from the social web". *Front. Comput. Neurosci.*, 6, p. 72.

Chapter 7

Discussion

7.1 Summary of findings and contributions

The body of work presented in this thesis explores functional alignment applications in cognitive neuroscience: Chapter 3 benchmarks different alignment algorithms on functional Magnetic Resonance Imaging (fMRI) datasets; Chapter 4 examines the impact of experimental factors on the performance of a single functional alignment algorithm; Chapter 5 integrates these results with the current literature into an accessible, interactive introduction to functional alignment for social and cognitive neuroscientists; and Chapter 6 highlights ongoing infrastructure work to create platforms to directly embed these interactive materials in future publications, rather than authoring them in separate environments.

In Chapter 3, I collate and expand on existing functional alignment algorithms to quantitatively assess their performance on a range of public fMRI datasets. I demonstrate how the same datasets and tasks can show a range of inter-subject similarity values after functional alignment depending on the assumptions of the algorithm. I observe that the best-performing functional alignment algorithms can recover approximately half of the individual variability usually lost in anatomical alignment. These results suggest that functional alignment can be used to successfully improve inter-subject similarity but caution that its performance must be carefully assessed.

In Chapter 4, I then examine other factors—beyond algorithm choice—that are likely to influence the performance of functional alignment. Specifically, I use a single functional alignment algorithm to assess the impact of experimental factors including both the data used to learn and to apply functional alignment transformations as well as the exact

evaluation metric used. I find that functional alignment performance depends on the relationship between the alignment dataset and its application, highlighting the importance of careful evaluation when adopting functional alignment in experimental contexts. These results confirm that functional alignment performance is strongly influenced by its context and suggest that researchers should incorporate functional alignment into their initial experimental designs for the best chance of successful applications.

Synthesizing these experimental results, in Chapter 5 I develop an accessible introduction to functional alignment for researchers in social and cognitive neuroscience. While functional alignment has been increasingly adopted in these fields, relatively few resources exist for orienting researchers to its use. I generate accompanying interactive materials allowing researchers to directly incorporate these methods into their own Python workflows.

Finally, in Chapter 6, I review the current state of publishing for interactive research objects such as those developed in Chapter 5. I propose new publishing infrastructure for research objects that integrate across scientific narrative, code, and data, pointing to our ongoing work in this area. Future developments supporting these integrated research objects will improve adoption of methods such as functional alignment across the field.

While each of these chapters distinctly contributes to considering functional alignment applications in cognitive neuroscience, collectively they provide a general framework for its adoption and offer a template for bridging the gap between methods development and domain science.

7.2 The challenges of individual variability for human brain mapping

While many fMRI studies include group-level analyses, their results are sensitive to any remaining variability in anatomical features—from cytoarchitectonic composition (Amunts et al., 1999; Rademacher et al., 1993) to large-scale sulco-gyral patterning (Galaburda et al., 1990; Marie et al., 2015)—after normalization to a reference template. Although standard fMRI preprocessing pipelines incorporate Gaussian smoothing to spatially blur this between-subject variability, smoothing also discards fine-scale functional information. Further, it cannot account for variable structure-function mapping across cortex (Paquola et al., 2019b; Vázquez-Rodríguez et al., 2019), making it difficult to directly compare functional organizations across individuals. While this challenge to human brain mapping

has been well-recognized for over two decades (Brett et al., 2002), standard practice has lagged behind ongoing technical developments.

In this thesis, I explored the potential of functional alignment to address the challenge of inter-individual variability. Since its introduction to cognitive neuroscience over a decade ago in Haxby et al. (2011), functional alignment has expanded into a broad class of methods for aligning on functional responses rather than supporting anatomical features. The work presented here confirms that functional alignment can improve comparisons between different individual's functional patterns; however, its success is heavily dependent on both algorithm choice (Chapter 3) as well as experimental context (Chapter 4). These results caution that functional alignment cannot guarantee inter-subject similarity and that researchers should be well-informed when incorporating these methods into their own studies. Importantly, this does not mean that functional alignment should be discarded from the cognitive neuroscience toolkit; rather, it suggests that we need a clearer framework for evaluating the impact of these methods on individual results. We proposed one such framework in Chapter 5, encouraging researchers to adopt predictive tasks to benchmark the relative success of functional alignment in improving inter-subject similarity.

The need for a clear evaluation framework is not limited to functional alignment, instead reflecting a more general challenge for integrating emerging methods into standard practice. New publishing formats offer an important path forward in making these frameworks accessible to the widest possible community. By integrating the code, data, and computational environment supporting a given result, other scientists will be able to more directly iterate and build on previous research. While the appeal of these integrated research objects is obvious, there is relatively little publishing infrastructure to support them. In Chapter 6, we discussed the current landscape for publishing integrated research objects and highlight our recent work in this space, some of which has been used to advance the interactive materials provided in Chapter 5.

Together, these results provide a general framework for evaluating individual functional alignment applications and suggest a template for adopting emerging methods and communicating their usage to the broader research community. As we continue to grapple with the challenge of individual variability in human brain mapping, this template promises to be of broad use for the field. The tension between finding correspondence across individuals while retaining unique information will likely continue to drive significant methodological development over the coming years. Here, I briefly review some promising directions for this development and discuss its relationship to the work presented in this thesis.

7.3 Future directions for recognizing functional diversity within and between brains

Functional alignment—as currently applied—is unlikely to be a singular solution for the problem of individual variability in human brain mapping. Rather, the work introduced in this thesis supplements a growing awareness of the problem and provides a framework for researchers to explore additional methods in approaching this challenge. To date, two main areas of research appear particularly promising for evaluating inter-individual correspondence. The first is continued methodological development to incorporate other important sources of variability beyond individual functional organizations. The second is technical and sociological support to more effectively leverage emerging deep phenotyping datasets. I consider each of these directions in turn.

7.3.1 Intra-individual variability across brain regions and trials

Recent years have seen remarkable strides in creating models of neural systems whose internal representations closely resemble biological measurements (Hassabis et al., 2017; Schrimpf et al., 2018). Importantly, different neural systems are best captured by entirely different classes of models, rather than simply by different parameterizations. For example, motor cortex is commonly modelled as a dynamical system (Sussillo et al., 2015; Vyas et al., 2020) while the ventral visual stream is modelled as a hierarchical series of convolutions (Eickenberg et al., 2017; Issa et al., 2018). This diversity of successful models highlights the variety of information processing strategies used throughout the brain, as well as the different time scales at which they operate (Honey et al., 2012; Jain et al., 2021). To date, functional alignment applications such as those explored in Chapters 3 and 4 employ a single algorithm across all considered brain regions. It is likely, however, that different transformations may be better suited to different brain systems.

Further complicating the question of correspondence is trial-by-trial variability in processing stimuli (Donnet et al., 2006; Westfall et al., 2016). In some higher-order cognitive regions, even synchronized stimuli may be processed in an unsynchronized fashion across individuals and trials. For example, Xie et al. (2021) found that responses in the ventromedial prefrontal cortex to emotionally salient stimuli were not well-aligned with the Shared Response Model (Chen et al., 2015), even when participants occupied the same cognitive state labelled by a Hidden Markov Model, confirming that the Shared Response Model is not able to effectively accommodate time-varying cognitive events. In recognition

of this challenge, some extensions of the algorithms used in this thesis focus on aligning information over time rather than space; for example, the BrainSync algorithm (Joshi et al., 2017) uses Procrustes alignment to find corresponding timepoints across individuals for the same set of voxels. However, it is more likely that effective inter-individual alignment in these regions will require finding correspondence across both time and space.

Rather than searching for correspondence across both dimensions in their original resolution, low-dimensional representations of well-labelled states offer a tractable path forward (Williams and Linderman, 2021). These representations can be used both to establish correspondence across individuals (Dabagia et al., n.d.) as well as to create effective models for re-analysis (Musall et al., 2019). To date, these labelled low-dimensional representations have been primarily developed in systems neuroscience applications, particularly as technological developments have allowed for new, large-scale neural recording methods (Urai et al., 2022). Translating these ideas to human brain mapping will require access to large volumes of well-labelled data from single-subjects, underscoring the continued importance of emerging deep phenotyping datasets in this work.

7.3.2 Drawing insights from deep data

The experimental work presented in Chapters 3 and 4 of this thesis heavily relies on currently available deep phenotyping datasets, particularly the Individual Brain Charting initiative (Pinho et al., 2018) and the Courtois Project on Neuronal Modelling (Courtois-NeuroMod; Boyle et al., 2020). Both of these datasets, however, are still actively being collected. Indeed, despite the increasing popularity of deep phenotyping acquisitions (Naselaris et al., 2021), these datasets are still relatively rare in the field. Further, those that do exist are designed with markedly divergent scientific goals: from mapping individual differences in intrinsic activity during resting-state (e.g., Midnight Scan Club; Gordon et al., 2017b) to comparing fine-scale differences in visual activity to naturalistic scenes (e.g., Natural Scenes Dataset; Allen et al., 2022).

While deep phenotyping datasets hold immense potential for advancing the field, they must be approached with care. As we currently lack a clear ontology of psychological processes (Poldrack et al., 2011), there is little guidance for which acquisitions to prioritize, and currently available acquisitions therefore vary widely between datasets. There is thus a strong risk that individual datasets will become the sole benchmarks for their scientific domains of interest. These concerns have already come to light across the fields of machine learning and artificial intelligence, where an over-reliance on benchmark datasets (Recht et al., 2018; Thompson et al., 2020) obscures real-world performance of new algorithms.

Avoiding such a situation in computational neuroscience will require a cohesive framework for analyzing and reporting results from these datasets.

I suggest that at least two approaches will be necessary for successful adoption of deep phenotyping datasets across human brain mapping. The first approach relies on improved technical solutions for accessing, analyzing, and reporting on these datasets. Current tools are piecemeal or dedicated to only a small number of datasets (e.g., PanNeuro, Rokem et al., 2021). The developing deep phenotype literature therefore lacks coherence—obscuring both the robustness of individual claims as well as the potential for harmonization across these resources. The publishing infrastructure suggested in Chapter 6 provides one potential solution, allowing investigators to report not only their conclusions but also each of the supporting research objects, increasing their possibility of re-use.

The framework presented in Chapter 5 suggests a second approach for integrating deep phenotyping datasets in human brain mapping: namely, evaluating predictive models for individual subjects. While training many models on a single dataset carries a strong risk of "dataset decay" (Thompson et al., 2020), deep phenotyping datasets are unique in their ability to support individual-level models. That is, as the volume of individual-level data continues to increase, it will become increasingly tractable to create complex models on individual subjects using, for example, deep learning methods. Indeed, this is an explicit goal of several ongoing data collection efforts such as Courtois-NeuroMod (Boyle et al., 2020). Work from artificial intelligence suggests that there will likely be important inter-individual variability in these models (Mehrer et al., 2020) and that reporting this variability may help to counteract model overfit. Further, functional alignment may be useful for directly comparing individual-level models within a dataset (as in Ho et al., 2022), providing more effective estimates of variance.

This work will help to extend the utility of individual predictive models beyond their original implementation. The potential of predictive models to improve theory in psychology and cognitive neuroscience is increasingly well-recognized (Bzdok et al., 2020; Rocca and Yarkoni, 2020; Varoquaux and Poldrack, 2019). Studying the models themselves—abstracted away from the originating datasets—may provide investigators the opportunity to look for common computational mechanisms in related but non-overlapping cognitive tasks. This is currently relatively popular in computational cognitive neuroscience, where successful predictive models are often directly compared to biological measurements (Schrimpf et al., 2018). The assumption in this line of research is that by comparing the accuracy and learned representations of different predictive models, we may be able to infer the underlying information-processing strategies (c.f. Thompson, 2021). This strategy

is already in active development for deep learning models (Kornblith et al., 2019; Yang et al., 2019) and may further support ongoing efforts to develop new cognitive ontologies for human brain mapping (Eisenberg et al., 2019; Varoquaux et al., 2018). Incorporating functional alignment into this work will help to provide clearer understanding of the variability around individual model estimates and may provide more a more generalizable understanding of information-processing strategies.

7.4 Conclusions

Individual variability in brain structure and function presents a core challenge to human brain mapping, and it is likely to drive continued research for the foreseeable future. Leveraging this variability appropriately, however, may provide us with substantial scientific insight into the mechanisms supporting individual outcomes. The work presented in this thesis focuses on functional alignment as a method for finding similarity across the "sensory and neural diversity" that define this variability. While functional alignment is unlikely to serve as a definitive solution to individual-level variability, it can effectively improve similarity across subjects in a variety of experimental use cases. The projects presented here describe both the relative performance of functional alignment as a method as well as a more general framework for considering other future methods that may be developed to meet this challenge. Taken together, this thesis serves to scaffold future work recognizing and addressing the impacts of individual variability on human brain mapping.

Bibliography

Abraham, A. et al. (2014). "Machine learning for neuroimaging with scikit-learn". *Front. Neuroinform.*, 8, p. 14.

Allen, E. J. et al. (2022). "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence". *Nat. Neurosci.*, 25(1), pp. 116–126.

Amunts, K et al. (1999). "Broca's region revisited: cytoarchitecture and intersubject variability". *J. Comp. Neurol.*, 412(2), pp. 319–341.

Anderson, Z., C. Gratton, and R. Nusslock (2021). "The Value of Hyperalignment to Unpack Neural Heterogeneity in the Precision Psychiatry Movement". *Biol Psychiatry Cogn Neurosci Neuroimaging*, 6(9), pp. 935–936.

Ashburner, J. (2012). "SPM: a history". *Neuroimage*, 62(2), pp. 791–800.

Avants, B. et al. (2008). "Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain". *Medical Image Analysis*, 12(1). Special Issue on The Third International Workshop on Biomedical Image Registration – WBIR 2006, pp. 26–41.

Bazeille, T et al. (2019). "Local Optimal Transport for Functional Brain Template Estimation". In: *Information Processing in Medical Imaging*. Springer International Publishing, pp. 237–248.

Beeck, H. P. Op de (2010). "Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses?" *Neuroimage*, 49(3), pp. 1943–1948.

Benson, N. C. et al. (2021). "Variability of the Surface Area of the V1, V2, and V3 Maps in a Large Sample of Human Observers".

Benureau, F. C. Y. and N. P. Rougier (2017). "Re-run, Repeat, Reproduce, Reuse, Replicate: Transforming Code into Scientific Contributions". *Front. Neuroinform.*, 11, p. 69.

Berti, V., L. Mosconi, and A. Pupi (2014). "Brain: normal variations and benign findings in fluorodeoxyglucose-PET/computed tomography imaging". *PET Clin.*, 9(2), pp. 129–140.

Bilenko, N. Y. et al. (2010). "How much tuning information is lost when we average across subjects in fMRI experiments?" *Journal of Vision*, 10, p. 917.

Bilenko, N. Y. and J. L. Gallant (2016). "Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging". *Front. Neuroinform.*, 10, p. 49.

Bollmann, S. and M. Barth (2020). "New acquisition techniques and their prospects for the achievable resolution of fMRI". *Prog. Neurobiol.*, p. 101936.

Boyle, J. A. et al. (2020). *The Courtois project on neuronal modelling: 2020 data release*. https://docs.cneuromod.ca. Presented at the 26th annual meeting of the Organization for Human Brain Mapping.

Brett, M., I. S. Johnsrude, and A. M. Owen (2002). "The problem of functional localization in the human brain". *Nat. Rev. Neurosci.*, 3(3), pp. 243–249.

Brette, R. (2018). "Is coding a relevant metaphor for the brain?" *Behav. Brain Sci.*, 42, e215.

Buckner, C. and J. Garson (2019). "Connectionism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University.

Busch, E. L. et al. (2020). "Hybrid Hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity".

Button, K. S. et al. (2013). "Power failure: why small sample size undermines the reliability of neuroscience". *Nat. Rev. Neurosci.*, 14(5), pp. 365–376.

Bzdok, D., D. Engemann, and B. Thirion (2020). "Inference and Prediction Diverge in Biomedicine". *Patterns (N Y)*, 1(8), p. 100119.

Carp, J. (2012). "On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments". *Front. Neurosci.*, 6, p. 149.

Chen, H.-T., J. R. Manning, and M. A. A. van der Meer (2021). "Between-subject prediction reveals a shared representational geometry in the rodent hippocampus". *Curr. Biol.*, 31(19), 4293–4304.e5.

Chen, J. et al. (2017). "Shared memories reveal shared structure in neural activity across individuals". *Nat. Neurosci.*, 20(1), pp. 115–125.

Chen, P.-H. et al. (2015). "A Reduced-Dimension fMRI Shared Response Model". In: *Advances in Neural Information Processing Systems 28*. Ed. by C Cortes et al. Curran Associates, Inc., pp. 460–468.

Churchland, P. S. and T. J. Sejnowski (1988). "Perspectives on Cognitive Neuroscience". *Science*, 242, pp. 741–745.

Churchland, P. M. (1998). "Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered". *J. Philos.*, 95(1), pp. 5–32.

Ciric, R. et al. (2021). "TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models".

Coalson, T. S., D. C. Van Essen, and M. F. Glasser (2018). "The impact of traditional neuroimaging methods on the spatial localization of cortical areas". *Proc. Natl. Acad. Sci. U. S. A.*, 115(27), E6356–E6365.

Cohen, J. D. et al. (2017). "Computational approaches to fMRI analysis". *Nat. Neurosci.*, 20(3), pp. 304–313.

Conroy, B. R. et al. (2013). "Inter-subject alignment of human cortical anatomy using functional connectivity". *Neuroimage*, 81, pp. 400–411.

Cox, R. W. (1996). "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages". *Comput. Biomed. Res.*, 29(3), pp. 162–173.

Crivello, F et al. (2009). "Chapter 31 - Biological Underpinnings of Anatomic Consistency and Variability in the Human Brain". In: *Handbook of Medical Image Processing and Analysis (Second Edition)*. Ed. by I. N. Bankman. Burlington: Academic Press, pp. 525–540.

Dabagia, M., K. P. Kording, and E. L. Dyer (n.d.). "Comparing high-dimensional neural recordings by aligning their low-dimensional latent representations". *Nature Biomedical Engineering* ().

Dadashkarimi, J., A. Karbasi, and D. Scheinost (2021). "Data-Driven Mapping Between Functional Connectomes Using Optimal Transport". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Springer International Publishing, pp. 293–302.

deCharms, R. C. and A Zador (2000). "Neural representation and the cortical code". *Annu. Rev. Neurosci.*, 23, pp. 613–647.

Dohmatob, E., G. Varoquaux, and B. Thirion (2018). "Inter-subject Registration of Functional Images: Do We Need Anatomical Images?" *Front. Neurosci.*, 12, p. 64.

Donnet, S., M. Lavielle, and J.-B. Poline (2006). "Are fMRI event-related response constant in time? A model selection answer". *Neuroimage*, 31(3), pp. 1169–1176.

Dubbs, A., J. Guevara, and R. Yuste (2016). "moco: Fast Motion Correction for Calcium Imaging". *Front. Neuroinform.*, 10, p. 6.

Dukart, J. and A. Bertolino (2014). "When structure affects function–the need for partial volume effect correction in functional and resting state magnetic resonance imaging studies". *PLoS One*, 9(12), e114227.

DuPre, E. et al. (2021). "TE-dependent analysis of multi-echo fMRI with tedana". *J. Open Source Softw.*, 6(66), p. 3669.

Eickenberg, M. et al. (2017). "Seeing it all: Convolutional network layers map the function of the human visual system". *Neuroimage*, 152, pp. 184–194.

Eisenberg, I. W. et al. (2019). "Uncovering the structure of self-regulation through data-driven ontology discovery". *Nat. Commun.*, 10(1), p. 2319.

Esteban, O. et al. (2019). "fMRIPrep: a robust preprocessing pipeline for functional MRI". *Nat. Methods*, 16(1), pp. 111–116.

Fedorenko, E. (2021). "The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience". *Current Opinion in Behavioral Sciences*, 40, pp. 105–112.

Feilong, M. et al. (2018). "Reliable individual differences in fine-grained cortical functional architecture". *Neuroimage*, 183, pp. 375–386.

Fischl, B and A. M. Dale (2000). "Measuring the thickness of the human cerebral cortex from magnetic resonance images". *Proc. Natl. Acad. Sci. U. S. A.*, 97(20), pp. 11050–11055.

Fischl, B, M. I. Sereno, and A. M. Dale (1999a). "Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system". *Neuroimage*, 9(2), pp. 195–207.

Fischl, B et al. (1999b). "High-resolution intersubject averaging and a coordinate system for the cortical surface". *Hum. Brain Mapp.*, 8(4), pp. 272–284.

Fodor, J. A. and Z. W. Pylyshyn (1988). "Connectionism and cognitive architecture: a critical analysis". *Cognition*, 28(1-2), pp. 3–71.

Fonov, V. et al. (2011). "Unbiased average age-appropriate atlases for pediatric studies". *Neuroimage*, 54(1), pp. 313–327.

Frost, M. A. and R. Goebel (2012). "Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment". *Neuroimage*, 59(2), pp. 1369–1381.

Galaburda, A. M., G. D. Rosen, and G. F. Sherman (1990). "Individual variability in cortical organization: its relationship to brain laterality and implications to function". *Neuropsychologia*, 28(6), pp. 529–546.

Glasser, M. F. et al. (2016). "A multi-modal parcellation of human cerebral cortex". *Nature*, 536(7615), pp. 171–178.

Gordon, E. M. et al. (2017a). "Individual Variability of the System-Level Organization of the Human Brain". *Cereb. Cortex*, 27(1), pp. 386–399.

Gordon, E. M. et al. (2017b). "Precision Functional Mapping of Individual Human Brains". *Neuron*, 95(4), 791–807.e7.

Greve, D. N. and B. Fischl (2009). "Accurate and robust brain image alignment using boundary-based registration". *Neuroimage*, 48(1), pp. 63–72.

Greve, J. M. (2011). "The BOLD effect". *Methods Mol. Biol.*, 771, pp. 153–169.

Guest, O. and B. C. Love (2017). "What the success of brain imaging implies about the neural code". *Elife*, 6.

Guntupalli, J. S., M. Feilong, and J. V. Haxby (2018). "A computational model of shared fine-scale structure in the human connectome". *PLoS Comput. Biol.*, 14(4), e1006120.

Guntupalli, J. S. et al. (2016). "A Model of Representational Spaces in Human Cortex". *Cereb. Cortex*, 26(6), pp. 2919–2934.

Hanke, M. et al. (2009). "PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data". *Neuroinformatics*, 7(1), pp. 37–53.

Hanke, M. et al. (2014). "A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie". *Sci Data*, 1, p. 140003.

Hassabis, D. et al. (2017). "Neuroscience-Inspired Artificial Intelligence". *Neuron*, 95(2), pp. 245–258.

Hauk, O. (2020). "Human Cognitive Neuroscience as It Is Taught". *Front. Psychol.*, 11, p. 587922.

Haxby, J. V., A. C. Connolly, and J. S. Guntupalli (2014). "Decoding neural representational spaces using multivariate pattern analysis". *Annu. Rev. Neurosci.*, 37, pp. 435–456.

Haxby, J. V. et al. (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex". *Science*, 293(5539), pp. 2425–2430.

Haxby, J. V. et al. (2011). "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron*, 72(2), pp. 404–416.

Henriksen, O. M. et al. (2012). "Estimation of intersubject variability of cerebral blood flow measurements using MRI and positron emission tomography". *J. Magn. Reson. Imaging*, 35(6), pp. 1290–1299.

Ho, J. K. et al. (2022). "Inter-individual deep image reconstruction".

Honey, C. J. et al. (2012). "Slow cortical dynamics and the accumulation of information over long timescales". *Neuron*, 76(2), pp. 423–434.

Hubel, D. H. and T. N. Wiesel (1959). "Receptive fields of single neurones in the cat's striate cortex". *J. Physiol.*, 148, pp. 574–591.

Humphries, C., E. Liebenthal, and J. R. Binder (2010). "Tonotopic organization of human auditory cortex". *Neuroimage*, 50(3), pp. 1202–1211.

Huntenberg, J. (2014). "Evaluating nonlinear coregistration of BOLD EPI and T1w images". MA thesis. University of Berlin.

Hutton, C. et al. (2002). "Image distortion correction in fMRI: A quantitative evaluation". *Neuroimage*, 16(1), pp. 217–240.

Issa, E. B., C. F. Cadieu, and J. J. DiCarlo (2018). "Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals". *Elife*, 7.

Jain, S. et al. (2021). "Interpretable multi-timescale models for predicting fMRI responses to continuous natural speech".

Jenkinson, M. et al. (2012). "FSL". *Neuroimage*, 62(2), pp. 782–790.

Jiahui, G. et al. (2020). "Predicting individual face-selective topography using naturalistic stimuli". *Neuroimage*, 216, p. 116458.

Joshi, A. A., M. Chong, and R. M. Leahy (2017). "BrainSync: An Orthogonal Transformation for Synchronization of fMRI Data Across Subjects". *Med. Image Comput. Comput. Assist. Interv.*, 10433, pp. 486–494.

Kluyver, T. et al. (2016). "Jupyter Notebooks – a publishing format for reproducible computational workflows". In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Amsterdam, NY: IOS Press, pp. 87–90.

Kornblith, S. et al. (2019). "Similarity of Neural Network Representations Revisited".

Kragel, P. A., R. M. Carter, and S. A. Huettel (2012). "What makes a pattern? Matching decoding methods to data in multivariate pattern analysis". *Front. Neurosci.*, 6, p. 162.

Kriegeskorte, N., R. Cusack, and P. Bandettini (2010). "How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter?" *Neuroimage*, 49(3), pp. 1965–1976.

Kriegeskorte, N. and J. Diedrichsen (2019). "Peeling the Onion of Brain Representations". *Annu. Rev. Neurosci.*, 42(1), pp. 407–432.

Kriegeskorte, N., M. Mur, and P. Bandettini (2008). "Representational similarity analysis - connecting the branches of systems neuroscience". *Front. Syst. Neurosci.*, 2, p. 4.

Kumar, M. et al. (2019). "BrainIAK tutorials: user-friendly learning materials for advanced fMRI analysis".

Kumar, M. et al. (2020). "BrainIAK: The brain imaging analysis kit".

Laakso, A. and G. Cottrell (2000). "Content and cluster analysis: Assessing representational similarity in neural systems". *Philos. Psychol.*, 13(1), pp. 47–76.

Laakso, A. and G. W. Cottrell (2005). "Churchland on Connectionism". In: *Paul Churchland*. Cambridge University Press, pp. 113–153.

Langs, G. et al. (2010). "Functional Geometry Alignment and Localization of Brain Areas". *Adv. Neural Inf. Process. Syst.*, 1, pp. 1225–1233.

Lawrence, R. M. et al. (2021). "Standardizing human brain parcellations". *Sci Data*, 8(1), p. 78.

Li, X. et al. (2021). "Moving Beyond Processing and Analysis-Related Variation in Neuroscience".

Liang, P. et al. (2015). "Construction of brain atlases based on a multi-center MRI dataset of 2020 Chinese adults". *Sci. Rep.*, 5, p. 18216.

Lindquist, M. A. and A. Mejia (2015). "Zen and the art of multiple comparisons". *Psychosom. Med.*, 77(2), pp. 114–125.

Marek, S et al. (2020). "Towards reproducible brain-wide association studies". *BioRxiv*.

Marie, D et al. (2015). "Descriptive anatomy of Heschl's gyri in 430 healthy volunteers, including 198 left-handers". *Brain Struct. Funct.*, 220(2), pp. 729–743.

Markiewicz, C. J. et al. (2021). "The OpenNeuro resource for sharing of neuroscience data". *Elife*, 10.

Mazziotta, J et al. (2001). "A probabilistic atlas and reference system for the human brain: International Consortium for Brain Mapping (ICBM)". *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 356(1412), pp. 1293–1322.

Mehrer, J. et al. (2020). "Individual differences among deep neural network models".

Mikl, M. et al. (2008). "Effects of spatial smoothing on fMRI group inferences". *Magn. Reson. Imaging*, 26(4), pp. 490–503.

Minsky, M. L. and S. A. Papert (1988). "Perceptrons: expanded edition".

Mountcastle, V. B. (1997). "The columnar organization of the neocortex". *Brain*, 120 (Pt 4), pp. 701–722.

Mukherjee, J. M. et al. (2016). "Improved frame-based estimation of head motion in PET brain imaging". *Med. Phys.*, 43(5), p. 2443.

Musall, S. et al. (2019). "Harnessing behavioral diversity to understand neural computations for cognition". *Curr. Opin. Neurobiol.*, 58, pp. 229–238.

Naselaris, T., E. Allen, and K. Kay (2021). "Extensive sampling for complete models of individual brains". *Current Opinion in Behavioral Sciences*, 40, pp. 45–51.

Naselaris, T. et al. (2011). "Encoding and decoding in fMRI". *Neuroimage*, 56(2), pp. 400–410.

Nastase, S. A., A. Goldstein, and U. Hasson (2020a). "Keep it real: rethinking the primacy of experimental control in cognitive neuroscience". *Neuroimage*, 222, p. 117254.

Nastase, S. A. et al. (2019). "Measuring shared responses across subjects using intersubject correlation". *Soc. Cogn. Affect. Neurosci.*, 14(6), pp. 667–685.

Nastase, S. A. et al. (2020b). "Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space". *Neuroimage*, 217, p. 116865.

Nenning, K.-H. et al. (2017). "Diffeomorphic functional brain surface alignment: Functional demons". *Neuroimage*, 156, pp. 456–465.

Nenning, K.-H. et al. (2020). "Joint Embedding: A scalable alignment to compare individuals in a connectivity space". *Neuroimage*, p. 117232.

Neves, K. et al. (2020). "The relationship between the number of neurons and behavioral performance in Swiss mice". *Neurosci. Lett.*, 735, p. 135202.

Newell, A. (1980). "Physical symbol systems". *Cogn. Sci.*, 4(2), pp. 135–183.

Norman, K. A. et al. (2006). "Beyond mind-reading: multi-voxel pattern analysis of fMRI data". *Trends Cogn. Sci.*, 10(9), pp. 424–430.

Panzeri, S. et al. (2015). "Neural population coding: combining insights from microscopic and mass signals". *Trends Cogn. Sci.*, 19(3), pp. 162–172.

Paquola, C. et al. (2019a). "Microstructural and functional gradients are increasingly dissociated in transmodal cortices". *PLoS Biol.*, 17(5), e3000284.

Paquola, C. et al. (2019b). "Microstructural and functional gradients are increasingly dissociated in transmodal cortices". *PLOS Biology*, 17(5), e3000284.

Pinho, A. L. et al. (2018). "Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping". *Sci Data*, 5, p. 180105.

Pinto, J. et al. (2020). "Cerebrovascular Reactivity Mapping Without Gas Challenges: A Methodological Guide". *Front. Physiol.*, 11, p. 608475.

Poldrack, R. et al. (2011). "The Cognitive Atlas: Toward a Knowledge Foundation for Cognitive Neuroscience". *Front. Neuroinform.*, 5, p. 17.

Poldrack, R. A. (2017). "Precision Neuroscience: Dense Sampling of Individual Brains". *Neuron*, 95(4), pp. 727–729.

— (2020). "The physics of representation".

Poldrack, R. A. and M. J. Farah (2015). "Progress and challenges in probing the human brain". *Nature*, 526(7573), pp. 371–379.

Poldrack, R. A. et al. (2013). "Toward open sharing of task-based fMRI data: the OpenfMRI project". *Front. Neuroinform.*, 7, p. 12.

Poldrack, R. A. et al. (2015). "Long-term neural and physiological phenotyping of a single human". *Nat. Commun.*, 6, p. 8885.

Portnow, L. H., D. E. Vaillancourt, and M. S. Okun (2013). "The history of cerebral PET scanning: from physiology to cutting-edge technology". *Neurology*, 80(10), pp. 952–956.

Power, J. D. et al. (2018). "Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data". *Proc. Natl. Acad. Sci. U. S. A.*

Pradal, C., G. Varoquaux, and H. P. Langtangen (2013). "Publishing scientific software matters". *J. Comput. Sci.*, 4(5), pp. 311–312.

Rademacher, J et al. (1993). "Topographical variation of the human primary cortices: implications for neuroimaging, brain mapping, and neurobiology". *Cereb. Cortex*, 3(4), pp. 313–329.

Raichle, M. E. (2009). "A brief history of human brain mapping". *Trends Neurosci.*, 32(2), pp. 118–126.

Recht, B. et al. (2018). "Do CIFAR-10 Classifiers Generalize to CIFAR-10?"

Richard, H. et al. (2019). "Fast shared response model for fMRI data". *arXiv preprint arXiv:1909.12537*.

Robbins, M. et al. (2021). "Calcium imaging analysis - how far have we come?" *F1000Res.*, 10, p. 258.

Rocca, R. and T. Yarkoni (2020). "Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction".

Rokem, A. et al. (2021). "Pan-neuro: interactive computing at scale with BRAIN datasets".

Rosenfeld, A. (2014). *Digital Picture Processing*. Elsevier.

Rule, A. et al. (2019). "Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks". *PLoS Comput. Biol.*, 15(7), e1007007.

Rumelhart, D. E. et al. (1988). *Parallel distributed processing*. Vol. 1. IEEE Massachusetts.

Saad, Z. S. et al. (2009). "A new method for improving functional-to-structural MRI alignment using local Pearson correlation". *Neuroimage*, 44(3), pp. 839–848.

Sabuncu, M et al. (2010). "Function-based intersubject alignment of human cortical anatomy". *Cerebral Cortex*, 20, pp. 130–140.

Schrimpf, M. et al. (2018). "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?"

Shepard, R. N. and S. Chipman (1970). "Second-order isomorphism of internal representations: Shapes of states". *Cogn. Psychol.*, 1(1), pp. 1–17.

Sonkusare, S., M. Breakspear, and C. Guo (2019). "Naturalistic Stimuli in Neuroscience: Critically Acclaimed". *Trends Cogn. Sci.*

Squire, L. R. (2009). "The legacy of patient H.M. for neuroscience". *Neuron*, 61(1), pp. 6–9.

Stella, F. and A. Treves (2021). "Hyper-alignment: Great mice think alike". *Curr. Biol.*, 31(19), R1138–R1140.

Strotzer, M. (2009). "One century of brain mapping using Brodmann areas". *Klin. Neuroradiol.*, 19(3), pp. 179–186.

Sussillo, D. et al. (2015). "A neural network that finds a naturalistic solution for the production of muscle activity". *Nat. Neurosci.*, 18(7), pp. 1025–1033.

Tahmasebi, A. M. et al. (2009). "Reducing inter-subject anatomical variation: effect of normalization method on sensitivity of functional magnetic resonance imaging data analysis in auditory cortex and the superior temporal region". *Neuroimage*, 47(4), pp. 1522–1531.

Talairach, J. and P. Tournoux (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-dimensional Proportional System : an Approach to Cerebral Imaging*. G. Thieme.

Tavor, I et al. (2016). "Task-free MRI predicts individual differences in brain activity during task performance". *Science*, 352(6282), pp. 216–220.

Teles, R. V. (2020). "Phineas Gage's great legacy". *Dement Neuropsychol*, 14(4), pp. 419–421.

Thompson, J. A. F. (2021). "Forms of explanation and understanding for neuroscience and artificial intelligence". *J. Neurophysiol.*, 126(6), pp. 1860–1874.

Thompson, W. H. et al. (2020). "Dataset decay and the problem of sequential analyses on open datasets". *Elife*, 9.

Urai, A. E. et al. (2022). "Large-scale neural recordings call for new insights to link brain and behavior". *Nat. Neurosci.*, 25(1), pp. 11–19.

Van Uden, C. E. et al. (2018). "Modeling Semantic Encoding in a Common Neural Representational Space". *Front. Neurosci.*, 12, p. 437.

Vanderwal, T., J. Eilbott, and F. X. Castellanos (2019). "Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging". *Dev. Cogn. Neurosci.*, 36, p. 100600.

Vanderwal, T. et al. (2017). "Individual differences in functional connectivity during naturalistic viewing conditions". *Neuroimage*, 157, pp. 521–530.

Varoquaux, G. and R. A. Poldrack (2019). "Predictive models avoid excessive reductionism in cognitive neuroimaging". *Curr. Opin. Neurobiol.*, 55, pp. 1–6.

Varoquaux, G. et al. (2018). "Atlases of cognition with large-scale human brain mapping". *PLoS Comput. Biol.*, 14(11), e1006565.

Vázquez-Rodríguez, B. et al. (2019). "Gradients of structure-function tethering across neocortex". *Proc. Natl. Acad. Sci. U. S. A.*, 116(42), pp. 21219–21227.

Vyas, S. et al. (2020). "Computation Through Neural Population Dynamics". *Annu. Rev. Neurosci.*, 43(1), pp. 249–275.

Wager, T. D., M. Lindquist, and L. Kaplan (2007). "Meta-analysis of functional neuroimaging data: current and future directions". *Soc. Cogn. Affect. Neurosci.*, 2(2), pp. 150–158.

Wang, R. et al. (2006). "Transient blood pressure changes affect the functional magnetic resonance imaging detection of cerebral activation". *Neuroimage*, 31(1), pp. 1–11.

Weaverdyck, M. E., M. D. Lieberman, and C. Parkinson (2020). "Tools of the Trade Multivoxel pattern analysis in fMRI: a practical introduction for social and affective neuroscientists". *Soc. Cogn. Affect. Neurosci.*, 15(4), pp. 487–509.

Westfall, J., T. E. Nichols, and T. Yarkoni (2016). "Fixing the stimulus-as-fixed-effect fallacy in task fMRI". *Wellcome Open Res*, 1, p. 23.

Williams, A. H. and S. W. Linderman (2021). "Statistical Neuroscience in the Single Trial Limit".

Williams, A. H. et al. (2021). "Generalized Shape Metrics on Neural Representations".

Worsley, K. J. et al. (1996). "A unified statistical approach for determining significant signals in images of cerebral activation". *Hum. Brain Mapp.*, 4(1), pp. 58–73.

Xie, T. et al. (2021). "Minimal functional alignment of ventromedial prefrontal cortex intracranial EEG signals during naturalistic viewing".

Xu, H et al. (2012). "Regularized hyperalignment of multi-set fMRI data". In: *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 229–232.

Xu, T. et al. (2019). "Cross-species Functional Alignment Reveals Evolutionary Hierarchy Within the Connectome".

Xu, T., M. Yousefnezhad, and D. Zhang (2018). "Gradient Hyperalignment for multi-subject fMRI data alignment". In: *Pacific Rim International Conference on Artificial Intelligence*. Springer, pp. 1058–1068.

Yang, G. R. et al. (2019). "Task representations in neural networks trained to perform many cognitive tasks". *Nat. Neurosci.*, 22(2), pp. 297–306.

Yousefnezhad, M. and D. Zhang (2016). "Local Discriminant Hyperalignment for multi-subject fMRI data alignment".

Zhang, R.-Y., X.-X. Wei, and K. Kay (2020). "Understanding multivariate brain activity: Evaluating the effect of voxelwise noise correlations on population codes in functional magnetic resonance imaging". *PLoS Comput. Biol.*, 16(8), e1008153.