U.PORTO

MESTRADO
NEUROBIOLOGIA

Mapping Minds with Machines:

Integrating Structural and Functional Brain Information into Explainable Graph-Informed Neurobiologically Plausible Models

Maria Beatriz Ramos

Supervisor: José Paulo Marques dos Santos

M

JULY, 2026

FACULDADE DE **MEDICINA**

Resumo

O cérebro é a estrutura mais complexa do corpo humano, composto por biliões de neurónios conectados entre si, formando uma rede vasta e dispersa para executar eximiamente as suas funções cognitivas. Devido à sua sofisticação, estudar, compreender e reproduzir esta estrutura biológica tem sido um persistente desafio neurocientífico. Diversas técnicas de neuroimagem emergiram com o intuito de contribuir para o avanço deste campo. Contudo, métodos tradicionais de análise de neuroimagem limitam a organização do cérebro a modelos puramente estatísticos, ignorando restrições biológicas impostas pela própria anatomia. O *machine learning* é um ramo da inteligência artificial que tem sido progressivamente aplicado ao estudo do cérebro humano devido à capacidade de aprender padrões complexos a partir de grandes quantidades de dados, sem impor restrições prévias. Além disso, métodos de inteligência artificial explicável (xAI) têm inovado este tipo de modelos ao adicionar um nível de explicabilidade que revela as características mais importantes para o resultado, ultrapassando as tradicionais arquiteturas opacas. Ainda assim, grande parte dos modelos de inteligência artificial para descodificar estados mentais continuam a depender de conexões funcionais que assentam em simplificações matemáticas não representativas da biologia do cérebro. Por estes motivos, o presente trabalho pretende desenvolver uma rede neural artificial (ANN) simples para, a partir de dados de ressonância magnética funcional, associar padrões de atividade cerebral ao respetivo processo cognitivo em execução, impondo restrições biológicas de modo a refletir o verdadeiro funcionamento do cérebro humano. A arquitetura proposta consiste numa rede neural rasa (SNN) com uma camada escondida composta por apenas 10 neurónios, para que seja possível manter o equilíbrio entre bom desempenho e interpretação do modelo. Os procedimentos de treino, remoção das ligações irrelevantes e de explicabilidade seguem os otimizados em estudos anteriores. O atlas HCP MMP1.0 é utilizado para segmentar o cérebro nas suas regiões independentes, correspondendo aos dados de entrada. As restrições biológicas são impostas de acordo com matrizes de conectividade, obtidas a partir de diferentes métodos de tractografia, através de um mecanismo de *message passing*. Enquanto a primeira análise é efetuada sob o paradigma motor, devido à localização bem caracterizada das regiões envolvidas, a transição para o paradigma social reflete os padrões distribuídos típicos deste tipo de função cognitiva. Os modelos de base obtiveram percentagens de proximidade entre a resposta correta e a classificação superiores a 80%, no entanto apenas restringiam o cérebro a regiões independentes e ignoravam a possibilidade ou impossibilidade de interações entre elas. O modelo não sofreu perda de desempenho preditivo com a representação em redes nem com a adição de estruturas subcorticais e conexões inter-hemisféricas. Por outro lado, a análise dos valores de SHAP revelam que o modelo não só faz predições corretas com um grau de confiança satisfatório, como também utiliza redes e regiões previamente associadas a essa mesma tarefa. Por último, a adição de regiões subcorticais e conexões inter-hemisféricas, através do atlas HCPex, demonstram que apesar da boa performance preditiva dos modelos obtidos, o contributo mais valioso do estudo é a aproximação mais fiel ao funcionamento e organização em larga escala do cérebro, através da imposição de restrições biológicas. O modelo final depende de redes amplamente distribuídas compostas por regiões que englobam ambos os hemisférios, e que têm sido consistentemente relacionadas na literatura com a respetiva tarefa cognitiva

em estudo. Progressos futuros deverão centrar-se na validação dos modelos aqui desenvolvidos, através de técnicas como a validação cruzada, para avaliar a consistência do desempenho dos modelos. Além disso, de modo a alargar a aplicabilidade da arquitetura aqui desenvolvida a paradigmas mais interativos e a abranger uma maior quantidade de sujeitos elegíveis, é importante procurar integrar progressivamente dados de espetroscopia funcional de infravermelho próximo (fNIRS) e ressonância magnética funcional (fMRI). Por fim, é necessária a implementação de novos paradigmas que tenham maior envolvência de regiões subcorticais, ou interações cortico-subcorticais mais preponderantes, para avaliar se o aumento da complexidade do modelo tem influência no seu desempenho preditivo.

Abstract

The human brain is the most complex structure of the human body, composed of billions of interconnected neurons, forming a vast and distributed network to expertly perform its cognitive functions. Due to its sophistication, studying, understanding, and reproducing this biological structure has been a persistent neuroscientific challenge. Several neuroimaging techniques have emerged with the aim of shedding some light on this field. However, traditional neuroimaging analysis methods limit the organization of the brain to purely statistical models, ignoring biological constraints imposed by anatomy. Machine learning is a branch of artificial intelligence that has been progressively applied to the study of the human brain due to its ability to learn patterns from large amounts of data, without any prior assumptions. Furthermore, explainable artificial intelligence (xAI) methods have innovated this type of model by adding a level of explainability that reveals the most important features for the result, overcoming traditional "black-box" architectures. However, many artificial intelligence models for mental states decoding still rely on functional relationships that are nothing more than mathematical simplifications unrepresentative of the biology of the brain. For these reasons, this work aims to develop a simple ANN to, from functional magnetic resonance imaging data, associate patterns of brain activity to its respective cognitive process in execution, imposing biological constraints to reflect the true functioning of the human brain. The chosen network is a shallow neural network (SNN) with only one hidden layer composed of 10 hidden nodes to maintain a balance between good predictive performance and model interpretability. The training, pruning mechanism and explainability methods follow those optimized in previous studies. The HCP MMP1.0 atlas is used to segment the brain into its independent regions, corresponding to the input data. Biological constraints are imposed according to connectivity matrices, obtained from different tractography methods, through a message passing mechanism. While the initial analysis is conducted under the motor paradigm, due to the well-characterized location of its inherent regions, the transition to the social paradigm reflects the typical distributed connection patterns of this type of cognitive function. The baseline models obtained accuracy values exceeding 80%, but they only restricted the brain to independent regions and ignored the possibility or impossibility of interactions between them. The model did not lose predictive performance with the network message-passing mechanism or with the addition of subcortical structures and interhemispheric connections. On the other hand, the analysis of SHAP values reveals that the model not only makes correct predictions with a satisfactory degree of confidence but also uses networks and regions previously associated with the respective task. Finally, the addition of subcortical regions and interhemispheric connections, through the HCPex atlas, demonstrates that despite the good predictive performance of the obtained models, the most valuable contribution of the study is the more faithful approximation to the large-scale functioning and organization of the brain, through the imposition of biological constraints. The final model relies on dense, distributed networks composed of regions across both hemispheres. These regions have been consistently correlated in literature with the respective cognitive task under study. Future progress should involve validating the models developed here through techniques such as cross-validation to ensure their generalizability. Furthermore, to broaden the applicability of the architecture developed

here to more interactive paradigms and to encompass a larger number of eligible subjects, it is important to progressively integrate fNIRS and fMRI data. Finally, it is necessary to implement new paradigms with greater involvement of subcortical regions, or more prevalent cortico-subcortical interactions, to assess whether increasing the model's complexity influences its predictive performance.

Acknowledgments

I would like to express my sincere gratitude to Professor José Paulo Marques dos Santos for his availability, guidance, and trust placed in me throughout the development of this dissertation. The concepts he transmitted to me were of great importance not only for this work but will undoubtedly be valuable for my future academic and professional course. I also thank José Diogo Marques dos Santos for his availability, patience, and promptness in addressing all the questions I raised. Finally, I would also like to express my warm gratitude to my family and friends for making the concretization of this goal possible, as well as for their support in all of the most challenging times.

Contents

List of Figures

List of Tables

Abbreviations

AI	Artificial Intelligence
ALL3	Automated Anatomical Labelling 3rd version
ANN	Artificial Neural Network
BOLD	Blood-Oxygen-Level Dependent
dMRI	Diffusion Magnetic Resonance Imaging
DTI	Diffusion Tensor Imaging
EEG	Electroencephalography
FA	Fractional Anisotropy
fMRI	Functional Magnetic Resonance Imaging
GLM	General Linear Model
GNN	Graph Neural Network
GSP	Graph Signal Processing
HCP	Human Connectome Project
HCP-MMP1.0	Human Connectome Project – Multimodal Parcellation
HOA	Harvard-Oxford Atlas
ICA	Independent Component Analysis
MEG	Magnetoencephalography
MRI	Magnetic Resonance Imaging
SHAP	SHapley Addictive exPlanations
SNN	Shallow Neural Network
ToM	Theory of Mind
xAI	Explainable Artificial Intelligence

Chapter 1
Introduction

The brain is the most complex biological structure of the human body, comprising approximately 86 billion neurons and trillions of connections (Soisoonthorn & Unger, 2025). Unraveling and replicating its intricate functioning has therefore remained one of the greatest challenges in Neuroscience. Research in the last centuries has enlightened some aspects of brain working at different levels of organization, from processes at the cell level to cognitive activities at the region and network levels (Amunts et al., 2024).

A classical milestone in the macroscopical study of the brain was the identification of Broca's area, in 1861. Broca hypothesized that the left inferior frontal gyrus played a central role in speech articulation, as lesions in this region were consistently linked to impaired language production. A decade later, in 1874, Wernicke observed that damage to the left superior posterior temporal gyrus resulted in a fluent yet semantically meaningless speech, leading to the association of this region with language comprehension. Together, these findings established a localizationist perspective of brain organization, according to which specific cognitive functions were believed to reside in specialized and anatomically distinct brain regions. Under this framework, damage to these specialized areas would result in permanent functional deficits, whereas lesions in less specialized regions were assumed to have minimal impact on overall brain functioning (Duffau, 2018; Soisoonthorn & Unger, 2025).

Although these discoveries represented crucial advances in the understanding of cerebral organization and enabled a more parcellated study of brain function, they overlooked an important aspect of cognition: cognitive processes rarely depend on isolated brain regions. Instead, they emerge from the dynamic interactions between distributed regions, organized into complex functional networks (Sporns, 2016).

Neuroimaging techniques have revolutionized neuroscience by driving a transition from a localizationist view of brain function to a network-based perspective of the human brain. In particular, functional magnetic resonance imaging (fMRI), through the measurement of blood oxygen level-dependent (BOLD) signals, enabled the identification of anatomically distant regions exhibiting temporally correlated activity, revealing complex, dynamic, and interacting neural networks underlying human behavior (Biswal et al., 1995). Alterations in these functional connectivity patterns have also been associated with several neurological and psychiatric disorders, including Alzheimer's disease, depression, and schizophrenia (Fox & Raichle, 2007).

The increasing availability of neuroimaging data stimulated the development of computational approaches for investigating brain organization. Traditionally, fMRI analyses have been dominated by statistical methods such as the General Linear Model (GLM) and Independent Component Analysis (ICA) (Bzdok, 2017). While GLM is a model-driven framework that represents voxel-wise signal changes according to predefined experimental conditions, ICA is a data-driven approach that decomposes brain activity into statistically independent components, allowing the identification of large-scale functional networks without imposing prior assumptions regarding their temporal profiles. However, these models often represent brain regions or voxels as largely independent entities and do not explicitly incorporate the anatomical relationships that

support communication between them. As a consequence, they are limited in their ability to capture the network architecture underlying information transfer and the dynamic interactions that emerge from it (Moeller et al., 2011).
More recently, diffusion magnetic resonance imaging (dMRI) further transformed the field by enabling the investigation of structural connectivity between brain regions. While functional connectivity reflects statistical dependencies between neural signals, structural connectivity describes the physical white-matter pathways that support information transfer across the brain (Axer et al., 2013). Together, these anatomical connections form the structural connectome, which constrains and shapes large-scale brain dynamics. Effective connectivity extends these concepts by modelling the causal influence that one brain region exerts over another, providing a mechanistic framework to explain functional interactions within the constraints imposed by the underlying anatomical architecture (Friston, 2011).
The human connectome comprises thousands of interconnected regions and pathways, resulting in a highly complex and high-dimensional system that is difficult to characterize using traditional analytical approaches. The growing availability of connectomic datasets, the reconstruction of large-scale electron microscopy volumes, and the prediction and quantification of human behavior increasingly require advanced computational methods. Artificial Intelligence (AI), particularly machine learning approaches, has played a crucial role in overcoming the scalability and complexity challenges associated with connectome mapping and analysis (Fan & Markram, 2019). Among these approaches, artificial neural networks (ANNs), inspired by the organizational principles of biological neural systems, process information through layers of interconnected computational units. By iteratively learning from data, these models can identify complex patterns and generate predictions without the need for explicitly programmed rules (Qamar & Zardari, 2023).
By incorporating biological knowledge into machine learning frameworks, it becomes possible to develop models that more closely reflect the organizational principles of the human brain, potentially paving the way towards digital twins: computational representations that replicate aspects of an individual's brain organization and cognitive processes. The present work contributes to this objective by explicitly constraining information flow according to the brain's anatomical connectivity, an aspect that remains underrepresented in many predictive models, including graph-based deep learning approaches (Li et al., 2021). Additionally, the framework adopted in this work seeks to balance performance and interpretability, by applying explainable artificial intelligence (xAI) methods to a simple ANN model, thereby moving towards the long term goal of a more biologically informed approximation of a digital twin of the brain.

1.1. Functional Magnetic Resonance Imaging

Functional neuroimaging comprises a set of non-invasive techniques used to measure and visualize brain activity in vivo. Among the most widely used methods are Electroencephalography (EEG) and Magnetoencephalography (MEG), which directly measure the electrical and magnetic fields generated by neuronal activity. These techniques provide millisecond temporal resolution, allowing the characterization of rapid neural dynamics, but suffer from relatively poor spatial resolution (Matthews & Jezzard, 2004).

The introduction of functional Magnetic Resonance Imaging (fMRI) in the early 1990s marked a major advance in neuroimaging by enabling the mapping of brain activity through changes in cerebral blood flow with spatial resolution on the order of millimeters. fMRI relies on the BOLD contrast, which reflects variations in blood oxygenation associated with neuronal activity. When a brain region becomes active, neurons consume additional oxygen to support increased metabolic demands. This rise in oxygen extraction is rapidly followed by a much larger increase in local cerebral blood flow, a process known as neurovascular coupling. Because the influx of oxygenated blood exceeds the amount of oxygen consumed by the tissue, the local concentration of deoxygenated hemoglobin (deoxyhemoglobin) decreases. Since oxyhemoglobin is diamagnetic whereas deoxyhemoglobin is paramagnetic, this reduction alters the local magnetic field and produces measurable changes in the fMRI signal, allowing the identification of functionally active brain regions (Chen & Glover, 2015).

Despite its widespread use, fMRI presents several limitations. Since the BOLD signal reflects a hemodynamic response rather than neural activity directly, it is delayed by several seconds relative to the underlying neuronal events, resulting in lower temporal resolution than EEG or MEG. Furthermore, fMRI data are particularly susceptible to noise and motion-related artefacts, and image acquisition requires participants to remain motionless within a noisy and confined scanner environment (Voineskos et al., 2024). Nevertheless, large-scale initiatives such as the Human Connectome Project (HCP) have generated and openly distributed extensive fMRI datasets, making the technique particularly valuable for the study of brain organization and connectivity at the population level.

1.2. Diffusion Tensor Imaging

Diffusion Tensor Imaging (DTI) is an advanced Magnetic Resonance Imaging (MRI) technique that enables the non-invasive characterization of water molecule diffusion in biological tissues *in vivo*. The method is based on the principle of Brownian motion, describing the random movement of water molecules within tissue microenvironments (Hagmann et al., 2006).

In the brain, the diffusion of water is constrained by structural barriers such as cell membranes, axons, and myelin sheaths. As a result, diffusion is not equal in all directions. In white matter, water molecules diffuse more readily along the longitudinal axis of axonal bundles than perpendicular to them, a phenomenon known as anisotropic diffusion. To characterize this directional behavior, DTI represents diffusion within each voxel using a tensor model, mathematically described by a 3 × 3 matrix. The tensor can be decomposed into three eigenvectors and three corresponding eigenvalues. The eigenvectors define the principal directions of diffusion, whereas the eigenvalues (λ_1, λ_2, λ_3) quantify the magnitude of diffusion along each direction. The principal eigenvector, associated with the largest eigenvalue, is generally assumed to align with the predominant orientation of local white matter fibers.

Several quantitative metrics can be derived from the diffusion tensor, the most widely used being fractional anisotropy (FA), mathematically calculated by comparing each eigenvalue with the average (λ) of all of them, according to the formula below:

$$FA = \sqrt{\frac{3}{2}}\sqrt{\frac{[\lambda_1-(\lambda)]^2+[\lambda_2-(\lambda)]^2+[\lambda_3-(\lambda)]^2}{{\lambda_1}^2+{\lambda_2}^2+{\lambda_3}^2}} \quad (1)$$

FA measures the degree of directional preference of water diffusion, ranging from 0 in isotropic environments, where diffusion occurs equally in all directions, to 1 in highly anisotropic tissues characterized by coherent fiber organization. Consequently, densely packed and highly organized white matter tracts exhibit elevated FA values (de Figueiredo et al., 2011).

FA is widely used as an indirect marker of white matter integrity. During brain development, FA generally increases as a result of progressive myelination and fiber organization. Conversely, reductions in FA are observed during normal aging and in several neurological disorders, including multiple sclerosis, chronic ischemia, and Wallerian degeneration. Despite its utility, FA is limited by the assumptions of the tensor model. In particular, the tensor describes diffusion using a single dominant orientation per voxel and therefore performs poorly in regions containing crossing, kissing, or branching fibers. In such cases, FA may be artificially reduced despite preserved tissue integrity, potentially leading to misleading interpretations (Tournier et al., 2011).

1.3. Tractography

Beyond the characterization of local microstructure, DTI enables the reconstruction of large-scale white matter pathways through tractography. By linking the estimated fiber orientations across neighboring voxels, tractography algorithms infer the trajectories of white matter bundles and provide the only non-invasive method for visualizing the three-dimensional architecture of structural brain connections *in vivo*. This capability has been fundamental for the development of structural connectomics, allowing the construction of network models describing the anatomical organization of the human brain. Tractography approaches are broadly divided into deterministic and probabilistic methods (Dong et al., 2004).

1.3.1. Deterministic Tractography

Deterministic tractography assumes a single fiber orientation at each voxel and reconstructs streamlines by continuously following the principal diffusion direction. Tracking is typically terminated when FA falls below a predefined threshold or when the trajectory exceeds a maximum angular deviation. This approach is computationally efficient and generates anatomically intuitive reconstructions, making it particularly attractive for clinical applications such as neurosurgical planning. However, because it follows only the most likely direction of diffusion, deterministic tractography is highly sensitive to noise and often fails in regions of complex fiber architecture, leading to false-negative connections.

1.3.2. Probabilistic Tractography

Probabilistic tractography explicitly models uncertainty in fiber orientation estimates. Instead of generating a single trajectory, the algorithm samples multiple possible directions from a probability distribution at each voxel, producing thousands of candidate streamlines from a given seed region. The resulting output is a connectivity probability map that reflects the likelihood of anatomical connections between brain regions. This approach is considerably more sensitive in areas containing crossing fibers, edema, or disrupted tissue architecture. Nevertheless, this increased sensitivity comes at the cost of

higher computational demands and an increased susceptibility to false-positive connections.

The choice between deterministic and probabilistic tractography therefore reflects a trade-off between specificity and sensitivity (Kierońska-Siwak et al., 2026). Deterministic approaches tend to produce more conservative reconstructions with fewer spurious streamlines, whereas probabilistic methods are better suited to capturing the uncertainty and complexity inherent to white matter organization. Both approaches have been widely employed in connectomics and play an important role in the reconstruction of structural brain networks.

1.4. Artificial Neural Networks

The human brain is a highly efficient and massively parallel information-processing system composed of billions of interconnected neurons. Inspired by its ability to learn from experience, recognize complex patterns, and adapt to new information, Artificial Neural Networks were developed as computational models capable of learning relationships directly from data without requiring explicitly programmed rules.

ANNs consist of interconnected computational units, known as nodes or artificial neurons, linked through weighted connections. These weights determine the strength of information transmitted between nodes and are progressively adjusted during training. A typical feedforward neural network is organized into three types of layers: an input layer, which receives the data; one or more hidden layers, where information is transformed through successive computations; and an output layer, which produces the final prediction. Through the interactions between these layers, the network learns increasingly complex representations of the input data (Wu & Feng, 2018).

At the level of an individual node, information processing begins by computing a weighted sum of the incoming inputs and adding a bias term (Figure 1), which allows the model to shift the output independently of the input values. This operation produces the net input:

$$Z = \sum_{i=0}^{n} w_i x_i + b \tag{2}$$

where x_i represents the input received from the previous layer, w_i the corresponding weight, and b the bias term. The resulting value is then passed through an activation function, introducing non-linearity into the network and enabling the modelling of complex relationships. Common activation functions include the hyperbolic tangent function, which maps values to the interval [-1,1], and the sigmoid function, which maps values to the interval [0,1] and is frequently used in binary classification tasks.

The learning process of an ANN is typically performed through supervised training. Initially, weights and biases are assigned random values (Han et al., 2018). During a forward pass, the input data propagate through the network and generate a prediction. The difference between the predicted and expected outputs is quantified using a loss function, which measures the network's error. This error is subsequently propagated backwards through the network using the backpropagation algorithm, allowing the contribution of each parameter to be estimated. Optimization algorithms, such as gradient descent, then iteratively adjust the weights and biases in the direction that minimizes the loss function. Repeating this process over multiple training iterations gradually improves the predictive performance of the model.

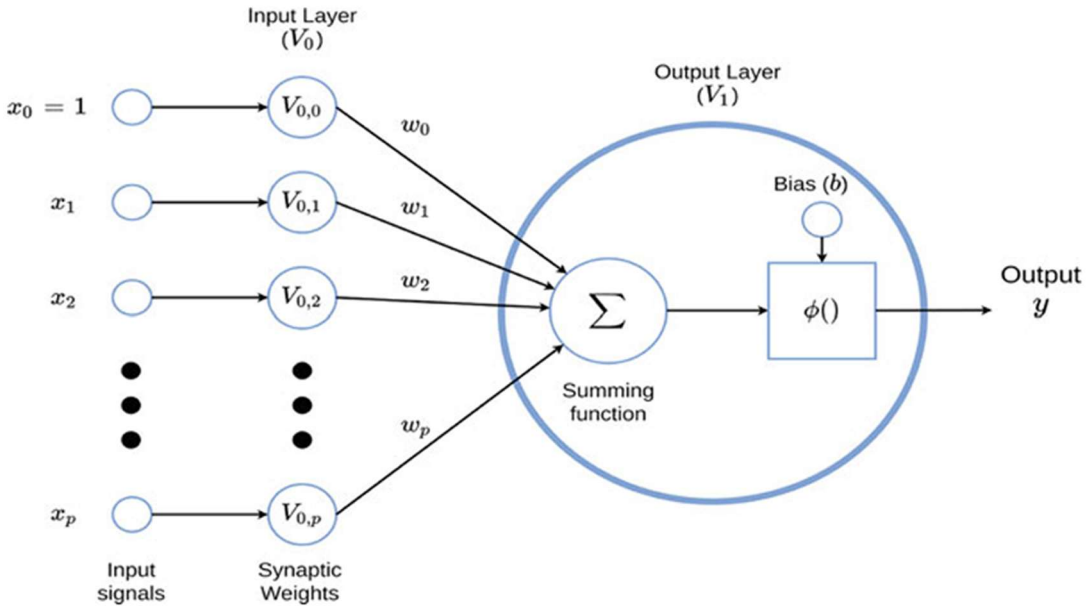

Figure 1: The sum function and bias computed for each node (Montesinos López et al., 2022).

Depending on the number of hidden layers, neural networks are commonly classified as shallow or deep architectures. Networks containing a single hidden layer are typically referred to as shallow neural networks (Figure 2), whereas deep neural networks contain multiple hidden layers capable of learning highly complex hierarchical representations. Although deep architectures often achieve superior predictive performance, their increased complexity frequently reduces interpretability, making it difficult to understand the mechanisms underlying their predictions. In contrast, shallow neural networks offer a favorable compromise between predictive accuracy and interpretability, which is particularly valuable in scientific applications where understanding the learned relationships is as important as obtaining accurate predictions (Montesinos López et al., 2022).

Regardless of the chosen architecture, neural networks remain susceptible to two common learning problems: overfitting and underfitting (Krogh, 2008). Overfitting occurs when a model learns patterns that are specific to the training data, including noise and irrelevant variations, resulting in poor generalization to unseen data. Conversely, underfitting occurs when the model lacks sufficient complexity or training to capture the underlying structure of the data, leading to poor performance even on the training set. Balancing model complexity and generalization therefore represents a fundamental challenge in machine learning applications.

A wide variety of neural network architectures have been developed to address different computational problems. Nevertheless, shallow feedforward neural networks provide sufficient flexibility to learn meaningful patterns from large neuroimaging datasets while preserving a level of interpretability compatible with the biological objectives of the present study.

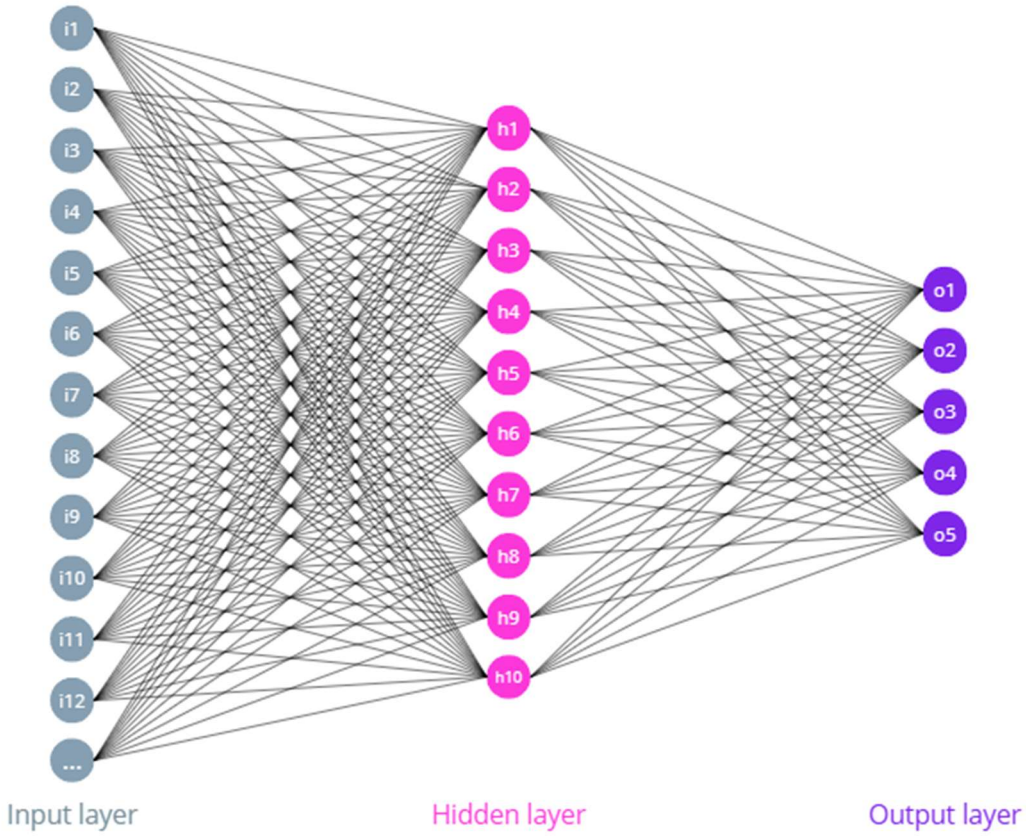

Figure 2: Example of a shallow neural network with 10 hidden nodes in the hidden layer and 5 output nodes in the output layer.

1.5. Graph-Based Neural Architectures

A graph is a mathematical structure used to model relationships between objects, composed of nodes connected by edges. The brain, also constituted by regions and their respective connections, possesses inherent characteristics such as preferential interaction between nearby regions to avoid unnecessary energy expenditure in information transmission, typically called small worldness, and grouping of specialized networks, forming hubs (Vecchio et al., 2017). These particularities make the graph an appropriate structure for studying the influence of brain organization in diverse cognitive tasks.

Traditional convolutional neural networks for studying the brain treat connectivity data as grids and often ignore non-Euclidean topology, leading to sub-optimal use of network structure. In recent years, the paradigm of neural networks has extended beyond Euclidean grid data to non-Euclidean domains, giving rise to Graph Neural Networks (GNNs) (Cui et al., 2022). In the context of neuroimaging, where the brain is naturally modelled as a network of interconnected regions, GNNs have emerged as the state-of-the-art for decoding brain states and disease classification. Frameworks such as *BrainGNN* (Li et al., 2021) utilize deep, highly parameterized architectures with trainable attention mechanisms to dynamically learn and weight the relationships between brain regions directly from fMRI data.

However, while these deep architectures achieve remarkable predictive performance, its complexness limits severely clinical interpretability. Furthermore, deep GNNs are highly susceptible to overfitting when trained on medical datasets, which are characteristically limited in sample size. Additionally, even these models typically rely solely on functional data and assume that all regions may be connected, neglecting structural constraints that are essential for biologically realistic modelling.

Graph Signal Processing (GSP) treats neuroimaging measurements as signals defined on graph nodes, analysed jointly with the structural or functional connectivity graph. GSP is based on mathematical operations such as Fourier transform applied to graphs, graph filtering, slepian functions, and surrogate data generation (Huang et al., 2018). From these operations, it is possible to predict what part of the functional signal is explained by structural connectivity.

On one hand, traditional GSP methods apply rigid, linear filters that often lack the flexibility to decode complex, non-linear cognitive states. On the other hand, deep GNNs offer high flexibility but disregard the fixed anatomical constraints of the brain, often creating abstract, uninterpretable pathways. There is a lack of intermediate frameworks that leverage the simplicity and biological grounding of fixed structural priors within a predictive neural network architecture. This dissertation addresses this gap by proposing a shallow, interpretability-focused network that utilizes the structural connectome as a fixed graph filter, bridging the gap between rigid anatomical anchoring and machine learning decoding capabilities.

1.6. Explainable Artificial Intelligence

As discussed previously, artificial neural networks learn complex relationships from data through the interactions of numerous weighted connections distributed across hidden layers. Although these architectures can achieve remarkable predictive performance, the mechanisms underlying their decisions are often difficult to interpret. This challenge is commonly referred to as the “black-box problem”, in which a model receives known inputs and produces observable outputs, while the internal processes leading to those predictions remain largely inaccessible to human understanding.

The lack of transparency associated with “black-box” models creates several important challenges. First, concerns regarding risk and ethics arise when AI systems are deployed in sensitive domains such as healthcare, finance, law, or autonomous vehicles, where decisions may have significant consequences for individuals and society. Second, issues of bias and fairness emerge when models inadvertently learn and perpetuate biases present in the training data, potentially leading to discriminatory outcomes (Arrieta et al., 2020). Third, legal and regulatory requirements increasingly demand greater transparency in automated decision-making processes. For example, the European Union's General Data Protection Regulation emphasizes the need for meaningful information regarding the logic underlying algorithmic decisions (Guidotti et al., 2018). Finally, trust and safety remain fundamental concerns, as clinicians, researchers, and other stakeholders may be reluctant to adopt AI systems whose decision-making processes cannot be adequately understood or validated.

To address these challenges, the field of explainable artificial intelligence has emerged as a collection of methods designed to improve the interpretability of machine learning

models. The primary objective of xAI is not only to increase trust in artificial intelligence systems, but also to facilitate model validation, debugging, bias detection, and scientific discovery. By revealing which features contribute most strongly to a prediction, xAI methods allow to move beyond performance metrics and gain insight into the underlying mechanisms learned by a model.

Among the most widely used xAI approaches are SHapley Additive exPlanations (SHAP), a framework based on concepts from game theory. SHAP quantifies the contribution of each feature to a model's prediction by assigning a Shapley value, representing the marginal contribution of that feature across all possible combinations of input features. In practice, the method estimates how much a model's prediction changes when a given feature is included or excluded from different subsets of variables. The resulting Shapley value can therefore be interpreted as the average contribution of a feature to the model's output (Fryer et al., 2021). Importantly, SHAP provides both global explanations, describing the overall importance of features across the dataset, and local explanations, revealing the factors driving individual predictions. Nevertheless, caution is required when interpreting SHAP values in the presence of highly correlated features, as importance may be distributed among related variables even when some of them provide redundant information.

1.7. Research Questions

The present dissertation is structured around three research questions:

- Is it possible to introduce biological informed structural constraints that restrict interactions between brain regions to existing anatomical pathways, enabling the transition from a region-level to a network-level decoding of motor tasks without loss in predictive performance?
- To what extent does a network-level representation of brain activity capture distributed patterns of social cognitive processes?
- Does the inclusion of subcortical regions and interhemispheric connections lead to more comprehensive and distributed representations of brain networks underlying motor and social cognitive tasks, compared to only cortical atlases or region-level models?

1.8. Objectives

Considering the concepts and methodologies previously described, the overarching objective of this dissertation is to develop a biologically informed machine learning framework capable of integrating structural and functional neuroimaging information for the interpretation of brain activity patterns. More specifically, the work aims to incorporate anatomical connectivity constraints into simple yet interpretable artificial neural network architectures, thereby moving towards a more realistic computational representation of brain organization, ultimately paving the way to produce digital twins of the human brain.

To achieve this objective, the problem is divided into several intermediate goals. First, different approaches for reconstructing structural brain connectivity are evaluated in order to identify the representation that most accurately captures the underlying anatomical organization. Second, structural and functional neuroimaging information are integrated into a unified graph-based representation of the brain. Third, this representation is

incorporated into a shallow artificial neural network framework and evaluated using a well-established motor paradigm, providing a robust benchmark for model validation. The framework is then extended to paradigms that rely more on distributed patterns of brain activity, such as social cognition, in order to assess its generalizability across different domains of brain function. Explainable artificial intelligence techniques are employed to investigate the neural features driving model predictions, increasing transparency and facilitating the biological interpretation of the learned representations. Finally, the inclusion of subcortical structures and interhemispheric connections is implemented into the model to evaluate the impact of a more robust representation of real brain organization in classification accuracies and interpretability.

Beyond its methodological objectives, this work seeks to contribute to the development of machine learning models that are not only predictive but also biologically meaningful. By integrating anatomical constraints and explainability into the modeling process, the proposed framework represents a step towards computational systems capable of capturing aspects of the relationship between brain structure, function, and behavior. Although the construction of a comprehensive digital twin of the human brain remains a long-term challenge that extends far beyond the scope of the present dissertation, the approaches explored here may contribute to future efforts aimed at developing more realistic and interpretable models of brain organization.

1.9. Dissertation Structure

This dissertation is organized into six chapters.

Chapter 1 introduces the theoretical background of the work, covering key concepts in neuroimaging, brain connectivity, and artificial neural networks. It also outlines the motivation of the study, defines the central research question, and presents the main objectives of the dissertation.

Chapter 2 reviews three studies previously developed by our research group that serve as the foundation for the present work. This chapter summarizes the main methodological advances achieved so far, including the development of a shallow neural network framework that balances predictive performance and interpretability. It also discusses the limitations of purely statistical approaches and the challenges associated with selecting an appropriate brain atlas for representing brain organization.

Chapters 3, 4, and 5 present the original contributions of this dissertation. Chapter 3 focuses on the incorporation of structural connectivity information into the machine learning framework. Several structural connectivity matrices obtained from the literature are evaluated and integrated into the previously established neural network model in order to identify the representation that most accurately captures the anatomical organization of the brain while maintaining satisfactory predictive performance.

Building upon these results, Chapter 4 extends the framework from a well-established motor paradigm to the more demanding domain of social cognition. Given the distributed nature of social processing, this chapter investigates whether the proposed architecture can generalize to more complex cognitive functions. Furthermore, explainable artificial intelligence techniques are employed to evaluate the biological plausibility of the learned representations and to identify the neural features driving model predictions.

Chapter 5 addresses a limitation present throughout the previous studies: the exclusive use of cortical regions from the HCP MMP1.0 atlas. To overcome this constraint, the

framework is expanded to include subcortical structures and interhemispheric connections through an extended atlas representation. The model is subsequently retrained and re-evaluated to determine whether a more complete representation of brain anatomy does not diminish predictive performance while simultaneously improving biological interpretability.

Finally, Chapter 6 summarizes the main findings of the dissertation and discusses their implications in light of the initial objectives. The chapter also highlights the limitations of the proposed framework and outlines potential future directions for the development of increasingly biologically informed models of brain organization and function.

Chapter 2
Defining the Pipeline

2.1. Introduction

Machine learning methods have become increasingly important tools for analyzing the large and complex datasets generated in modern neuroscience. Their ability to identify non-linear relationships and high-dimensional patterns has enabled the study of brain organization at levels of complexity that would be difficult to achieve using conventional statistical approaches alone.

Despite their success, the application of machine learning to neuroscience presents a fundamental challenge. Models optimized exclusively for predictive performance often become increasingly complex and difficult to interpret, limiting their usefulness for generating biologically meaningful insights. In contrast, simpler and more transparent models may provide greater explanatory value, even when associated with a modest reduction in predictive accuracy. Consequently, achieving an appropriate balance between performance and interpretability has become a central objective in the development of machine learning frameworks for neuroscience research.

Within this context, a series of studies developed by our research group explored the use of simple neural network architectures and explainability approaches to investigate the relationship between brain connectivity patterns and cognitive processes. Particular attention was given to the identification of model configurations capable of maintaining competitive predictive performance while preserving the interpretability required for neuroscientific applications. These studies also evaluated different anatomical atlases and knowledge-based explanation strategies, providing important insights into the biological relevance of the obtained results.

The developments achieved through this line of research constitute the foundation of the work presented here. Therefore, this chapter reviews the main methodological advances and findings obtained in previous studies, highlighting both their contributions and the limitations that motivated the subsequent developments addressed in the present work.

2.2. Path-Weight-Based Pruning and SHAP-Based Explanations

In an effort to develop an efficient yet interpretable machine learning framework for motor movement classification, Marques dos Santos and Marques dos Santos (2024) combined independent component analysis (ICA) of functional magnetic resonance imaging (fMRI) data with explainable artificial intelligence techniques. The study aimed to address one of the major challenges in machine learning applications to neuroscience: balancing predictive performance with model interpretability. Rather than maximizing accuracy alone, the proposed approach sought to retain only the most relevant information required for classification, enabling a clearer understanding of the features driving the model predictions.

The pipeline started by preprocessing raw data from the Human Connectome Project Young Adults database, more precisely, from the motor paradigm within the 100

Unrelated Subjects subset. Thirty subjects from this database were randomly allocated into train (20 subjects) and test (10 subjects) groups. The motor paradigm included five distinct movements: Left foot squeezing (LF), right foot squeezing (RF), left hand tapping (LH), right hand tapping (RH) and tongue movement (T). The most significant features for analysis were obtained by averaging the seventh, eighth, and ninth time points post-stimulus onset, reflecting the range in which the peak of the hemodynamic response was most likely found.

The study employed a shallow neural network (SNN), chosen for its simplicity and interpretability. The network consisted of 46 input nodes, corresponding to the ICA components, a single hidden layer with 10 hidden nodes, and 5 output nodes representing the five motor tasks.

To improve interpretability, the trained networks underwent a pruning procedure based on path-weight analysis, a more effective pruning strategy than other previous arbitrary attempts. This approach identifies the most influential input-output pathways and removes connections with minimal contribution to classification performance. Following pruning, the networks were retrained to recover part of the lost predictive accuracy while preserving the simplified architecture.

To further investigate feature importance, Shapley values were computed to quantify the contribution of each input feature to the classification outcome. Features were subsequently categorized according to whether they increased, decreased, or had negligible influence on class predictions.

Partial and global accuracies are represented for the fully connected, pruned, and retrained neural networks in Table 1, Table 2 and Table 3, respectively. In general, accuracies obtained were superior to mere chance. Notably, the fully connected neural network model achieved good global accuracy (83%), which dropped when the model was pruned (77.5%), but was able to recover, after retraining, to 80.5%. Interestingly, what concerns the partial accuracies it was possible to observe that the left foot had consistently lower accuracies across all models, when compared to the other 4 movements. This may be due to the position of the left foot representation in the motor cortex, which is very close to the right foot representation, only separated by the longitudinal fissure, leading to erroneous results for this class.

Table 1: Confusion matrix, partial and global accuracies and precision for the fully connected neural network.

Stimulus		**Prediction**					**Total**
		LF	LH	RF	RH	T	
Input	LF	27	1	6	5	1	40
	LH	3	36	0	1	0	40
	RF	8	0	31	1	0	40
	RH	0	2	1	37	0	40
	T	4	0	1	0	35	40
Total		42	39	39	44	36	200
Accuracy (%)		67.5	90.0	77.5	92.5	87.5	83.0
Precision (%)		64.3	92.3	79.5	84.1	97.2	

Table 2: Confusion Matrix, Partial and global accuracies and precision for the pruned network.

Stimulus		Prediction					Total
		LF	LH	RF	RH	T	
Input	LF	29	2	6	1	2	40
	LH	4	31	2	3	0	40
	RF	5	0	29	0	6	40
	RH	0	4	1	34	1	40
	T	4	0	4	0	32	40
Total		42	37	42	38	41	200
Accuracy (%)		72.5	77.5	72.5	85.0	80.0	77.5
Precision (%)		69.0	83.8	69.0	89.1	78.0	

Table 3: Confusion Matrix, Partial and global accuracies and precision for the retrained network.

Stimulus		Prediction					Total
		LF	LH	RF	RH	T	
Input	LF	19	6	13	0	2	40
	LH	1	38	0	1	0	40
	RF	6	0	34	0	0	40
	RH	0	4	1	35	0	40
	T	4	0	1	0	35	40
Total		30	48	49	36	37	200
Accuracy (%)		47.5	95.0	85.0	87.5	87.5	80.5
Precision (%)		63.3	79.2	69.4	97.2	94.6	

Shapley values results showed that there were 4 independent components that were the most relevant to the model: IC5; IC7; IC11 and IC12. Specifically, it was hypothesised that IC5 is responsible for tongue movements. IC7 is in control of the left hemisphere. IC11 is involved in movements of the right hemisphere and tongue and IC12 predicts feet movements, either on the left or right side. These results were consistent with the path-weight analysis and with established neuroscientific knowledge, showing the efficacy of the method in revealing the most important features of the model accurately.

Despite the promising results obtained, the use of ICA as the feature extraction method presented important limitations. Although ICA effectively identifies statistically independent patterns of brain activity, the resulting components are not constrained by known anatomical structures and therefore lack direct biological interpretability. Furthermore, ICA-derived networks do not correspond to predefined neuroanatomical regions, limiting their spatial specificity and making it difficult to relate the identified components to established brain organization frameworks.

Nevertheless, this work introduced several important methodological improvements. In particular, the elbow-based pruning strategy proved more effective than previously used

arbitrary thresholds, while the application of Shapley values provided explanations that were consistent with both path-weight analysis and current neuroscientific knowledge. These findings demonstrated that interpretable machine learning models can successfully identify meaningful brain-related patterns while maintaining satisfactory predictive performance.

However, the absence of an anatomically grounded representation remained an open challenge. This limitation motivated subsequent studies aimed at replacing ICA-derived features with atlas-based brain representations, allowing the optimized classification and explainability framework developed in this work to be evaluated within a biologically meaningful anatomical context.

2.3. Introducing Anatomical Brain Atlases

This subchapter investigates the use of anatomical brain atlases as an alternative to ICA-based dimensionality reduction for the development of knowledge-based neural network models using fMRI data. The approach proposed by Marques do Santos (2024) aimed to overcome some of the limitations of ICA, namely its reduced anatomical specificity and the difficulties associated with applying cross-validation without introducing data leakage.

Four anatomical atlases were evaluated: Automated Anatomical Labelling, 3rd version (AAL3); Brodmann; Harvard–Oxford atlas together with a modified version incorporating GLM-based sub-segmentation; and the Human Connectome Project Multimodal Parcellation atlas (HCP-MMP1.0).

The procedure was the same for every matrix and followed the optimized steps of the previous study. The time series of every region was extracted by averaging the value of every voxel in that region. The 7th, 8th and 9th points after stimulus onset are averaged. Training and testing input matrices were generated with lines corresponding to the training or test instances, respectively, while columns correspond to brain regions accordingly to each atlas.

The training process consisted of, firstly, applying a grid search for the best hyperparameters The best pair combination was subsequently used to train 10000 networks in order to find the starting weights that achieved the highest performance. The following steps consisted in pruning and retraining the network accordingly with the pipeline described previously.

The atlases differed mainly in their parcellation strategy and spatial resolution. The Brodmann atlas is based on cytoarchitectural divisions of the cerebral cortex and provides a relatively coarse parcellation with 78 regions. The Harvard–Oxford Atlas is a probabilistic anatomical atlas derived from structural MRI data, containing 99 cortical and subcortical regions, although its larger regions may include both task-relevant and irrelevant voxels. The AAL3 atlas offers a finer anatomical segmentation, dividing the brain into 170 regions based on structural landmarks. The HCP-MMP1.0 atlas provides the highest level of anatomical detail, combining multimodal imaging information to define 360 cortical regions with high specificity. Finally, the Harvard–Oxford + GLM approach extends the original Harvard–Oxford atlas by subdividing motor-related regions using task-specific statistical maps, thereby increasing regional specificity while incorporating functional information relevant to the studied paradigm.

Comparisons between atlases final accuracies are resumed in Table 4. Atlas-based parcellation can achieve performance comparable to, and in some cases exceeding, that obtained with ICA. Among all tested approaches, the HCP-MMP1.0 atlas yielded the highest classification accuracy (85.0%), surpassing the ICA-based model while providing improved anatomical interpretability and better discrimination between left and right motor activations.

Table 4: Number of regions, voxels and final accuracy for each atlas.

Atlas	Number Regions	Number of voxels	Final Accuracy
AAL3	170	185,751	78.5%
Brodmann	78	74,677	73.5%
HOA	99	199,998	68.0%
HOA+GLM	107	195,835	83.5%
HCP-MMP1.0	360	78,619	85.0%

The findings also suggested that atlas granularity and anatomical specificity play a critical role in model performance. Atlases containing large regions with substantial inclusion of non-task-relevant voxels, such as the original Harvard–Oxford atlas, appeared to suffer from signal dilution effects. This limitation was partially addressed through GLM-guided sub-segmentation, which substantially improved classification accuracy, highlighting the value of combining anatomical priors with task-specific functional information.

Overall, the results support the use of anatomically informed brain parcellations, particularly the HCP-MMP1.0 atlas, as a robust framework for constructing interpretable neural network models and functional digital twins of the human brain. Furthermore, because atlas-based segmentation is independent of the training dataset, it enables rigorous cross-validation procedures without the risk of data leakage.

2.4. Expanding to Social Cognition Paradigms

The studies described in the previous sections focused on optimizing the proposed machine learning pipeline through the refinement of network architecture, pruning strategies, explainability methods, and brain representations. The resulting framework demonstrated high classification performance in the Human Connectome Project motor paradigm while maintaining a level of interpretability that is uncommon in machine learning approaches applied to neuroimaging data.

Although the motor paradigm provides an ideal benchmark due to its well-established anatomical organization, the ultimate goal is to develop models capable of addressing more complex aspects of brain function. Human cognition extends far beyond motor execution and encompasses higher-order processes such as social reasoning, communication, and the interpretation of others' intentions. Therefore, the next step in this line of research consisted of evaluating the optimized framework in a more challenging cognitive setting: the Theory of Mind paradigm.

The work in analysis (Marques dos Santos et al., 2025) used the exact same network structure as described in the previous studies. In the Theory of Mind paradigm, participants were presented with 20-second animations depicting geometric objects either

moving randomly or interacting in ways that resemble intentional human behaviour. When participants attribute intentions, emotions, or social meaning to these interactions, they engage in a cognitive process known as mentalizing. After each stimulus, participants classified the animation as random (rnd), mentalizing (mnt), or uncertain.
The Human Connectome Project multimodal parcellation (MMP1.0) atlas was used as template for segmenting the brain. A BOLD signal time series was obtained for each of the independent brain regions. Feature extraction consisted of extracting the mean time series for each region, and averaging the seventh, eighth and nineth time points after the stimulus onset, just like previously described. Only correct answers were considered, i.e. the participant answered mentalizing when objects were performing human interactions or answered random when the objects were moving randomly. Consequently, training input matrix consisted of 168 training instances in lines and 360 columns corresponding to the brain regions. The test input matrix consisted of 85 test instances lines and 360 columns.
For the fully connected neural network, the model achieved 88.2% of global accuracy, which dropped to 80% with pruning but regaining most of its performance after retraining (84.7%). Results are summarized in Table 5. SHAP analysis revealed distinct sets of regions contributing to each class. Regions positively associated with mentalizing tended to contribute negatively to random classifications and vice versa, suggesting that the model learned meaningful discriminative patterns between both cognitive states. The corresponding anatomical locations are summarized in Table 6 and include regions that have consistently been implicated in social cognition and Theory of Mind processing.

Table 5: Confusion matrix, partial and global accuracies and precision for the social paradigm.

Network	Stimulus		Prediction		Total
			mnt	rnd	
Fully Connected Network	Input	mnt	43	2	45
		rnd	8	32	40
	Total		51	34	85
	Accuracy (%)		95.6	80.0	88.2
	Precision (%)		84.3	94.1	
Pruned Network	Input	mnt	41	4	45
		rnd	13	27	40
	Total		54	31	85
	Accuracy (%)		91.1	67.5	80.0
	Precision (%)		75.9	87.1	
Retrained Network	Input	mnt	41	4	45
		rnd	9	31	40
	Total		50	35	85
	Accuracy (%)		91.1	77.5	84.7
	Precision (%)		82.0	77.5	

Table 6: Anatomical location of the most relevant features of the model according to SHAP analysis.

Region ID	|SHAP|	Region Name	Hemisphere	Lobe	Cortex ID	Cortex
			Random stimuli (rnd)			
96	1.506	Area 6 anterior L	Left	Fr	8	Premotor
175	1.824	Auditory 4 Complex L	Left	Temp	11	Auditory association
125	3.977	Auditory 5 Complex L	Left	Temp	11	Auditory association
151	1.580	Area PGs L	Left	Par	17	Inferior parietal
13	1.928	Area V3A L	Left	Occ	3	Dorsal stream visual
219	2.800	Area V3B R	Right	Occ	3	Dorsal stream visual
159	1.357	Area Lateral Occipital 3 L	Left	Occ	5	MT+ complex and neighboring visual areas

While the previous studies demonstrated the effectiveness of the proposed framework in a motor task, the present work extended its application to a considerably more complex cognitive paradigm. Successfully classifying Theory of Mind processes represents a more demanding challenge due to the distributed and highly integrated nature of the underlying neural mechanisms. Therefore, the ability of the model to maintain high classification accuracy while producing biologically meaningful explanations provides further evidence of the robustness of the proposed approach.

2.5. Discussion

Despite the promising results obtained in both motor and cognitive paradigms, an important limitation remained unresolved. Although the use of anatomical atlases introduced a biologically meaningful spatial representation of the brain, the neural network architecture itself remained agnostic to the underlying structural organization of the nervous system. All brain regions were treated as equally connected entities, regardless of the physical pathways that support information transfer in the brain.

Consequently, the resulting models continued to rely on a tabular representation of neuroimaging data, where interactions between regions were learned exclusively from the data without considering known anatomical constraints. This assumption contrasts with the biological reality of the brain, which is organized as a highly heterogeneous network shaped by structural connectivity patterns.

The success of the optimized pipeline developed throughout the previous studies therefore motivated the next stage of this research line: the incorporation of anatomical connectivity information directly into the machine learning architecture. By embedding biologically informed constraints into the model, it becomes possible to move beyond purely data-driven representations and towards computational frameworks that more faithfully reflect the organization of the human brain. The development and evaluation of such biologically constrained neural networks constitute the focus of the work presented in the following chapters.

Chapter 3
Network-Level Analysis of Motor Functions

3.1. Introduction

Functional interactions observed in fMRI data emerge within the constraints imposed by the underlying anatomical architecture of the brain. Although statistical dependencies between regional signals can reveal patterns of co-activation, neuronal communication ultimately relies on white matter pathways that enable information transfer between distant cortical areas. Consequently, incorporating structural connectivity information into computational models may provide a biologically grounded framework for constraining information flow and improving interpretability.

Throughout this chapter, several connectivity matrices, obtained through different tractography methods, are explored in order to evaluate how biological constraints contribute to the model's performance.

3.2. Datasets, Subjects and Experimental Paradigm

This study draws on a dataset of 30 subjects sourced from the Human Connectome Project (HCP) Young Adult database (https://www.humanconnectome.org/study/hcp-young-adult), specifically those with identifiers ranging from 100307 to 124422, all belonging to the *100 Unrelated Subjects* subset (Elam et al., 2021; Van Essen & Glasser, 2016; Van Essen et al., 2013). These subjects are randomly allocated into two groups: a training cohort of 20 subjects and a testing cohort containing the remaining 10. The motor task paradigm encompasses five distinct categories: left foot movement (LF), left-hand digit flexion (LH), right foot movement (RF), right-hand digit flexion (RH), and tongue movement (T). Since each of the two sessions relies on different acquisition sequences, every session is analysed separately. As a result, the training cohort amounts to 40 data files (20 subjects × 2 sessions), and the testing cohort to 20 data files (10 subjects × 2 sessions), with no overlap between the two groups.

3.3. Feature Extraction

To generate the features used as input for training the shallow neural network, the brain is parcellated using the HCP-MMP1.0 atlas (Human Connectome Project—multimodal parcellation) (Glasser et al., 2016). Mean regional time series are obtained by applying each atlas region as a mask on the subject's blood oxygen level–dependent (BOLD) data and then averaging the voxel-wise signals contained within that region. This procedure produces 360 regional time series per session, with each one representing the BOLD signal of a given brain region over the course of the full session.

Once the time series have been extracted, feature selection and data standardization are carried out. Each regional time series is first segmented according to the stimulus sequence, after which the seventh, eighth, and ninth time points following stimulus onset are averaged together to yield a single feature. This feature captures the class-specific neural response of each brain region. The training and testing input matrices are then assembled independently and standardized as a whole, treating each matrix in its entirety.

Both matrices share 360 columns, corresponding to the brain regions, and differ only in the number of samples: the training matrix holds 400 instances, whereas the testing matrix holds 200.

3.4. Structural Connectivity Matrices

Structural connectivity information is obtained from previously published connectomes based on the HCP-MMP1.0 atlas. Three distinct connectomes are selected to represent different tractography approaches and levels of anatomical specificity:

1. An anatomy-driven deterministic connectome (Baker et al., 2018);
2. A population-based deterministic connectome (Yeh, 2022);
3. A probabilistic connectome (Rosen & Halgren, 2021).

These connectomes are chosen because tractography methodology can substantially influence the resulting connectivity estimates. Deterministic approaches generally produce more conservative reconstructions of white matter pathways, whereas probabilistic approaches tend to capture a broader range of potential connections while being more susceptible to false positives. Additionally, probabilistic tractography is preferable in terms of ensuring reproducibility between subjects, while deterministic methods tend to be less consistent and noisier between subjects. However, the latter has the ability to preserve clinically relevant morphological variabilities, which would be masked by the averaging of the probabilistic method (Shuai et al., 2025). When comparing these connectomes, it is expected that the different representations of brain anatomy will affect the model's performance.

To ensure compatibility with the proposed modelling framework, all connectomes are converted into binary adjacency matrices indicating only the presence (1) or absence (0) of structural connections between pairs of cortical regions. When necessary, thresholding procedures are applied to transform weighted or probabilistic connectivity estimates into binary representations. Figure 3 illustrates the connectivity patterns associated with all structural connectivity matrices used throughout the study.

3.4.1. Anatomy-Driven Deterministic Connectome

The anatomy-driven deterministic connectome is derived from diffusion-weighted imaging data acquired from ten subjects. Structural pathways are reconstructed using Generalized Q-Sampling Imaging (GQI) combined with deterministic tractography. In contrast to fully automated approaches, tractography seeds are manually positioned to target well-established anatomical pathways, resulting in a connectome strongly informed by neuroanatomical knowledge.

Connectivity between cortical regions is defined by the presence of reconstructed streamlines linking parcels of the HCP-MMP1.0 atlas. The final group-averaged connectome contains 2.100 anatomical connections and represents a conservative estimate of cortical structural organization.

3.4.2. Population-Based Deterministic Connectome

The population-based deterministic connectome is generated using diffusion imaging data from 1.065 participants. Structural pathways are reconstructed using deterministic tractography with randomized tracking parameters and automated seeding procedures based on templates of major white matter bundles.

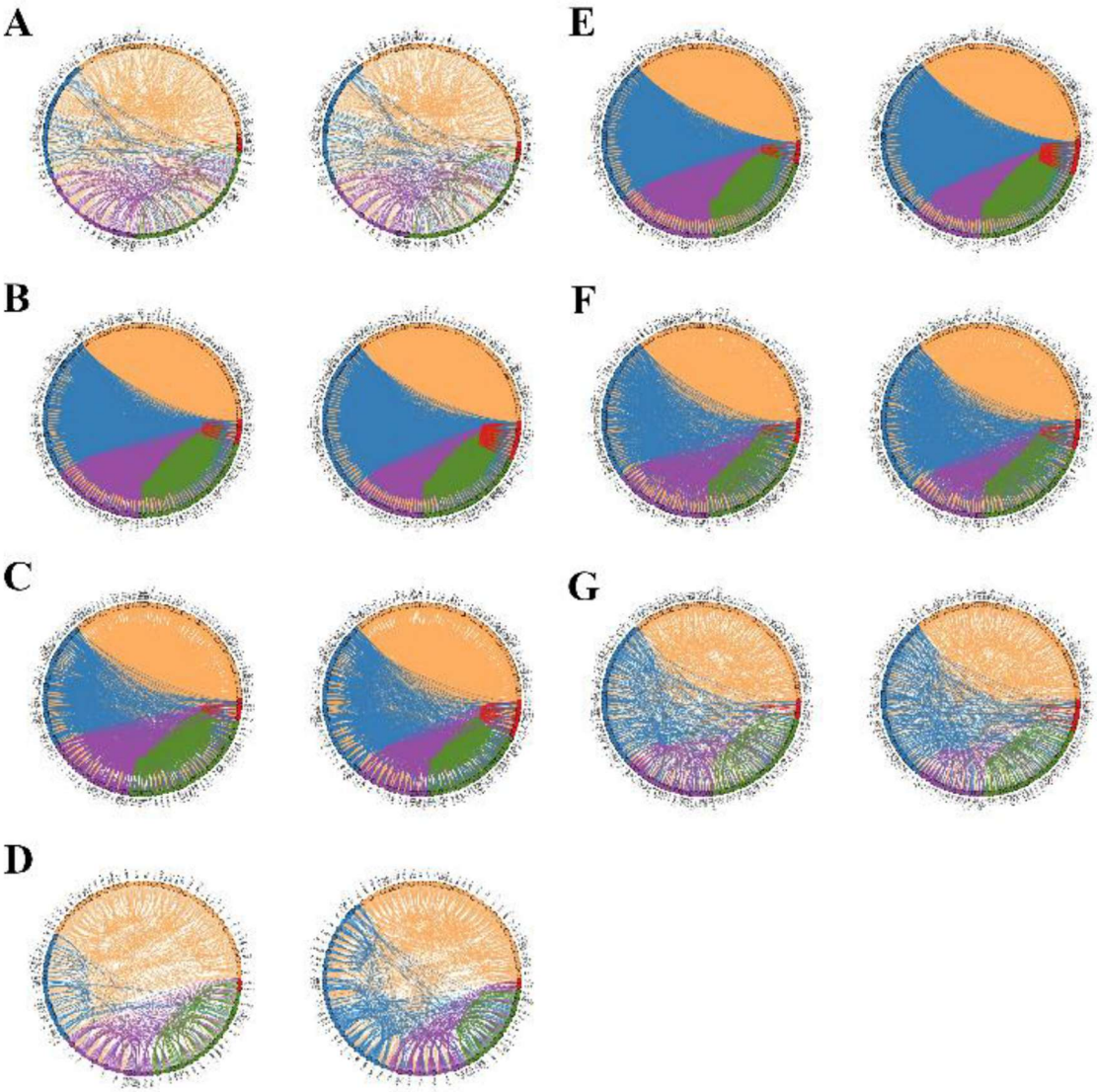

Figure 3: Connectivity Graph Representation for matrices: A - Anatomy-driven deterministic; B - Population-based deterministic (5% threshold); C - Population-based deterministic (70% threshold); D - Population-based deterministic (99% threshold); E – Probabilistic (-2 threshold); F – Probabilistic (-3 threshold); G - Probabilistic (-4 threshold)

Unlike the anatomy-driven connectome, connectivity estimates are expressed as population-level probabilities representing the proportion of subjects in whom a given tract was observed. To obtain binary region-by-region connectivity matrices, probability thresholds of 5%, 70%, and 99% are applied. These thresholds generate connectomes of varying sparsity, ranging from broad and inclusive representations of structural connectivity to highly conservative anatomical networks.

3.4.3. Probabilistic Connectome

The probabilistic connectome is also derived from diffusion imaging data obtained from 1,065 participants. In this case, structural pathways are reconstructed using probabilistic tractography implemented through FSL ProbtrackX.

Connectivity strength between cortical regions is quantified as the logarithm of the number of streamlines connecting each pair of regions. To generate binary adjacency matrices comparable to those obtained from deterministic approaches, group-averaged connectivity values are thresholded at −4, −3, and −2. Since connectivity values are

represented on a base-10 logarithmic scale, these thresholds correspond to increasingly restrictive criteria for defining structural connections.

For all connectomes, the identity matrix is added to the adjacency matrix prior to analysis. This operation introduces self-connections for every cortical region, ensuring that a region's own activity remains represented after the subsequent message-passing procedure.

From a neurobiological perspective, this step reflects the fact that regional activity is influenced not only by external inputs from connected areas but also by local processing occurring within the region itself. The inclusion of self-connections therefore preserves local information while enabling the integration of signals originating from structurally connected neighbors.

3.5. Overlap Between Connectivity Matrices

To quantify the similarity between the different structural connectivity matrices, pairwise comparisons are performed using the Jaccard index and the overlap coefficient.

The Jaccard index measures the proportion of shared connections ($| A \cap B |$) relative to the total number of unique connections ($| A \cup B |$) present across two matrices A and B:

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

Higher Jaccard values indicate greater similarity between connectomes.

To complement this analysis, the overlap coefficient is also calculated:

$$Overlap(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \quad (4)$$

Where $| A \cap B |$ represents the set of values that are in common in both matrices A and B, and $\min(| A |, |B|)$ is the set with fewer elements.

Unlike the Jaccard index, the overlap coefficient evaluates the extent to which the smaller connectome is contained within the larger one. An overlap coefficient equal to one indicates that all connections present in the smaller matrix are also found in the larger matrix.

Together, these measures provide complementary information regarding the degree of agreement between different structural connectivity matrices.

3.6. Message-Passing Mechanism

The message-passing mechanism constitutes the core operation through which structural connectivity information is incorporated into the modelling framework.

Conceptually, this procedure allows each cortical region to integrate information originating from anatomically connected neighbors. Rather than considering the activation of each brain region independently, the resulting representation incorporates the broader structural context in which each region is embedded.

In graph theory, message passing forms the basis of Graph Neural Networks (GNNs), where node representations are iteratively updated through information exchange with neighboring nodes. In the present work, a simplified implementation consisting of a single message-passing step is adopted.

Mathematically, the procedure corresponds to multiplying the feature matrix by the structural connectivity matrix:

$$Z = I \cdot (C + I_{360}) \quad (5)$$

Where I is the signal matrix obtained during feature extraction (an $n \times 360$ matrix), C is the binary anatomical adjacency (a 360×360 matrix), I_{360} is the identity matrix added to

introduce self-loops, and Z is the resulting transformed signal matrix (an n×360 matrix). The self-loop addition preserves each region's own signal during aggregation.

A correction factor is additionally evaluated to account for differences in network size across connectomes. This correction consists of dividing the aggregated signal of each region by the total number of regions contributing to that signal, including the self-loop. Because self-connections are included in all connectivity matrices, the updated value of each cortical region reflects the combined contribution of its own signal and the signals of all structurally connected regions. Consequently, each feature becomes a representation of network-level activity rather than purely local activation.

The rationale behind this correction stems from the possibility of signal dilution. When only a subset of regions within a structural network is actively involved in a task, averaging activity across large networks containing numerous inactive regions may reduce the distinction between task-related and baseline activity. The correction factor is therefore introduced to determine whether normalizing for network size improves the utility of broad connectivity matrices.

3.7. Neural Network Training, Pruning and Retraining

To facilitate direct comparison with previous studies, the neural network architecture consists of 360 input nodes, 10 hidden nodes arranged in a single hidden layer, and 5 output nodes corresponding to the five motor conditions included in the experimental paradigm.

The network is implemented using the AMORE package in R. Hidden nodes employ a hyperbolic tangent activation function ("tansig"), producing outputs within the interval]−1,1[, whereas output nodes use the sigmoid activation function, producing values within the interval]0,1[.

3.7.1. Initial Training

Network training begins with a grid search procedure aimed at identifying the optimal combination of learning rate and momentum parameters. To account for variability associated with random weight initialization, each hyperparameter combination is evaluated through 100 independent training runs.

Following hyperparameter selection, 50.000 networks are trained using the optimal parameter combination. The only source of variability between these networks is the random initialization of weights. The best-performing model resulting from this procedure is designated as the fully connected network.

3.7.2. Network Pruning

The pruning procedure, an essential step in ANN models, is based on the concept of path-weights, according to the following formula.

$$path - weight_{i,j,k} = \left| w_{ij} \times w_{jk} \right| \quad (6)$$

Where the path-weight from input i to output k is given by the multiplication of the weight connection from input node i to hidden node j with the weight of the connection from hidden node j to output node k. Path weights close to zero indicate connections with minimal influence on the network output and are therefore considered candidates for removal. Conversely, large positive or negative path-weights correspond to pathways

exerting stronger influence on network decisions. All path-weights are ranked according to magnitude, and elbow points are identified separately for positive and negative values. Connections falling between the two elbow points are considered weak contributors and are removed from the network architecture.
The resulting model is designated as the pruned network.

3.7.3. Retraining

Following pruning, the network is retrained to allow the remaining connections to adapt to the new architecture. During this stage, an additional grid search is performed to determine optimal training parameters. Unlike the initial training phase, only one training run is performed for each hyperparameter combination because the network weights are no longer randomly initialized.
The objective of retraining is to recover performance potentially lost during pruning while taking advantage of a simpler and more efficient architecture. The resulting model is designated as the retrained network.

3.8. Results

3.8.1. Comparisons Between Structural Connectivity Matrices

The similarity analyses reveal substantial differences among the structural connectivity matrices included in this study. As shown in Table 7, Jaccard index values are generally low across all pairwise comparisons, indicating that the proportion of shared connections between matrices is limited. This observation is not restricted to comparisons between connectomes generated using different tractography strategies but also evident among matrices obtained from different threshold levels of the same connectome.
The relatively low Jaccard values observed between thresholded versions of the same connectome can be explained by the characteristics of the metric itself. Although more restrictive matrices are nested within broader ones, increasing the number of included connections substantially enlarges the union of connections while leaving the intersection largely unchanged. Consequently, Jaccard similarity remains modest despite the hierarchical relationship between matrices.

Table 7: Jaccard indexes between connectivity matrices. Anatomy-driven deterministic is represented as "Anatomical"; Population-based deterministic matrices are truncated to "Pop." followed by the respective threshold and Probabilistic matrices are summarized as "Prob." Followed by its respective threshold

Jaccard	Anatomical	Pop. 5%	Pop. 70%	Pop. 99%	Prob. (-4)	Prob. (-3)	Prob. (-2)
Number of Connections	2,100	36,942	14,258	2,732	65,790	16,750	3,762
Anatomical		0.0525	0.101	0.193	0.024	0.039	0.077
Pop. 5%			0.385	0.073	0.342	0.179	0.065
Pop. 70%				0.191	0.141	0.124	0.079
Pop. 99%					0.031	0.054	0.099
Prob. (-4)						0.254	0.057
Prob. (-3)							0.224
Prob. (-2)							

Differences in network density also influence the observed values. Comparisons involving matrices with markedly different numbers of connections tend to produce particularly low Jaccard indices. In these cases, the larger union generated by the denser matrix outweighed the number of shared connections, resulting in reduced similarity estimates. Nevertheless, deterministic connectomes generally exhibit greater similarity to

one another than to the probabilistic connectome, suggesting that tractography methodology has a substantial impact on the resulting connectivity patterns.
A complementary perspective emerges from the overlap coefficient analysis (Table 8). For matrices derived from the same connectome but thresholded at different levels, overlap coefficients reach values of one. This result reflects the fact that more restrictive matrices constitute subsets of broader versions, such that every connection present in the sparse matrix is also contained within the denser representation.

Table 8: Overlap coefficients between connectivity matrices. Anatomy-driven deterministic is represented as "Anatomical"; Population-based deterministic matrices are truncated to "Pop." followed by the respective threshold and Probabilistic matrices are summarized as "Prob." Followed by its respective threshold

Overlap	Anatomical	Pop. 5%	Pop. 70%	Pop. 99%	Prob. (-4)	Prob. (-3)	Prob. (-2)
Number of Connections	2,100	36,942	14,258	2,732	65,790	16,750	3,762
Anatomical		0.928	0.719	0.373	0.758	0.343	0.200
Pop. 5%			1.000	1.000	0.708	0.488	0.669
Pop. 70%				1.000	0.697	0.240	0.354
Pop. 99%					0.757	0.366	0.215
Prob. (-4)						1.000	1.000
Prob. (-3)							1.000
Prob. (-2)							

When different connectomes are compared, overlap coefficients decrease considerably and vary according to the relative density of the matrices. Comparisons involving matrices with similar numbers of connections generally produce lower overlap values, further supporting the notion that the three connectomes capture distinct representations of structural brain organization.
Among all cross-connectome comparisons, the greatest overlap is observed between the anatomy-driven deterministic connectome and the population-based deterministic connectome. In particular, the broadest population-based deterministic matrix contains approximately 93% of the connections identified in the anatomy-driven deterministic connectome, indicating substantial correspondence between these two deterministic approaches despite their methodological differences.

3.8.2. Region-Level Classification

The region-level classification model uses the activity of each of the 360 HCP-MMP1.0 regions as independent inputs to the neural network. Classification accuracies obtained at the different stages of model development are summarized in Table 9.
The fully connected network achieves an overall accuracy of 90.0%, representing the best performance observed throughout the study. Following the pruning procedure, accuracy decreases to 83.0%, indicating that some information relevant for classification is removed together with low-contributing paths. Subsequent retraining partially compensates for this loss, increasing accuracy to 85.0%.
Analysis of class-specific performance reveals that left foot (LF) trials are consistently the most challenging condition to classify correctly. Examination of the confusion matrices show that LF samples are frequently misclassified as right foot (RF) movements, suggesting limited separability between the neural representations associated with these two motor conditions. While pruning negatively affects the classification of both RF and RH conditions, a slight improvement is observed for LF classification. Retraining alleviates most of these effects and contributes to the partial recovery of global performance.

3.8.3. Network-Level Classification

To investigate whether incorporating anatomical information improves the biological relevance of the models, classification is repeated after applying the message-passing procedure based on the different structural connectivity matrices.

For the anatomy-driven deterministic connectome without correction, the fully connected model achieves an accuracy of 80.0%. Performance decreases substantially following pruning, reaching 50.5%, but is almost entirely restored after retraining, which achieves 79.5% accuracy. The largest pruning-related reductions are observed for LH and RF classifications.

The population-based deterministic connectomes display a clear relationship between matrix sparsity and classification performance. The broadest matrix (5% threshold) achieves 53.5% accuracy in the fully connected stage, whereas the intermediate (70%) and restrictive (99%) matrices achieve 68.0% and 82.5%, respectively. Although pruning produces substantial performance reductions in all cases, retraining recovers most or all of the lost accuracy.

A similar pattern is observed for the probabilistic connectome. The broadest matrix (−4 threshold) produces the lowest classification accuracy of all tested connectomes, reaching only 31.5% in the fully connected stage. Performance increases progressively as connectivity became more restrictive, with accuracies of 39.0% and 76.5% for thresholds of −3 and −2, respectively. Interestingly, for the −3 threshold matrix, retraining produces a final model that outperforms the original fully connected network.

The introduction of the correction factor generally improves performance. For the anatomy-driven deterministic connectome, accuracy increases from 80.0% to 83.0%. Similar improvements are observed for the broad probabilistic and deterministic matrices, where normalization substantially enhances classification results.

Overall, three major trends emerge. First, the anatomy-driven deterministic matrix consistently produces the highest classification accuracies. Second, more restrictive connectivity matrices tend to outperform broader ones. Third, the correction factor generally improves model performance, particularly for matrices containing large numbers of connections.

3.9. Discussion

The present results demonstrate that the choice of structural connectivity representation has a substantial impact on classifier performance. Among all tested connectomes, the anatomy-driven deterministic connectome produces the highest overall accuracy, reaching 83% when combined with the correction factor. Other high-performing models include the restrictive population-based deterministic matrix (99% threshold) and the restrictive probabilistic matrix (−2 threshold).

A consistent finding across both deterministic and probabilistic connectomes is the relationship between network sparsity and classification performance. As connectivity matrices become more restrictive, classifier accuracy generally increases. This trend suggests that limiting information exchange to the strongest or most reliable anatomical connections may preserve task-relevant information while reducing the influence of unrelated regions.

Table 9: Global and partial accuracies of the fully connected, pruned and retrained networks for all classifiers, "Anatomical" represents the anatomy-driven deterministic connectivity matrix, "Pop." indicates the population-based deterministic connectivity matrix followed by the respective threshold, and "Prob." indicates the probabilistic connectivity matrix followed by the respective threshold.

		LF	LH	RF	RH	T	Total
HCP-MMP1.0	Fully Conn.	0.700	1.000	0.875	0.950	0.975	0.900
	Pruned	0.725	0.975	0.750	0.775	0.925	0.830
	Retrained	0.700	0.975	0.825	0.825	0.925	0.850
Anatomical Uncorrected	Fully Conn.	0.625	0.975	0.700	0.800	0.900	0.800
	Pruned	0.500	0.550	0.250	0.650	0.575	0.505
	Retrained	0.625	0.975	0.700	0.800	0.875	0.795
Pop. 5% Uncorrected	Fully Conn.	0.400	0.675	0.475	0.475	0.650	0.535
	Pruned	0.525	0.500	0.125	0.350	0.600	0.420
	Retrained	0.375	0.750	0.225	0.525	0.800	0.535
Pop. 70% Uncorrected	Fully Conn.	0.525	0.850	0.500	0.725	0.800	0.680
	Pruned	0.425	0.350	0.150	0.575	0.600	0.420
	Retrained	0.475	0.850	0.500	0.700	0.775	0.660
Pop. 99% Uncorrected	Fully Conn.	0.625	0.950	0.825	0.850	0.875	0.825
	Pruned	0.350	0.600	0.450	0.300	0.375	0.415
	Retrained	0.550	0.975	0.850	0.725	0.825	0.785
Prob. -4 Uncorrected	Fully Conn.	0.125	0.775	0.600	0.050	0.025	0.315
	Pruned	0.600	0.750	0.000	0.025	0.025	0.280
	Retrained	0.075	0.775	0.625	0.025	0.025	0.305
Prob. -3 Uncorrected	Fully Conn.	0.125	0.675	0.600	0.375	0.175	0.390
	Pruned	0.275	0.125	0.475	0.450	0.500	0.365
	Retrained	0.150	0.700	0.650	0.500	0.200	0.440
Prob. -2 Uncorrected	Fully Conn.	0.600	0.975	0.650	0.725	0.875	0.765
	Pruned	0.525	0.475	0.450	0.850	0.600	0.580
	Retrained	0.575	0.875	0.700	0.725	0.875	0.750
Anatomical Corrected	Fully Conn.	0.675	0.950	0.775	0.875	0.875	0.830
	Pruned	0.725	0.850	0.700	0.775	0.875	0.785
	Retrained	0.675	0.975	0.725	0.825	0.925	0.825
Pop. 5% Corrected	Fully Conn.	0.525	0.900	0.650	0.575	0.750	0.680
	Pruned	0.325	0.250	0.250	0.425	0.575	0.365
	Retrained	0.500	0.875	0.675	0.525	0.800	0.675
Pop. 70% Corrected	Fully Conn.	0.550	0.700	0.475	0.775	0.750	0.650
	Pruned	0.525	0.375	0.300	0.475	0.500	0.435
	Retrained	0.425	0.775	0.375	0.650	0.775	0.600
Pop. 99% Corrected	Fully Conn.	0.600	0.925	0.750	0.750	0.900	0.785
	Pruned	0.750	0.900	0.550	0.600	0.825	0.725
	Retrained	0.625	0.875	0.725	0.725	0.925	0.775
Prob. -4 Corrected	Fully Conn.	0.625	0.775	0.675	0.750	0.700	0.705
	Pruned	0.425	0.275	0.125	0.675	0.125	0.325
	Retrained	0.650	0.800	0.600	0.725	0.700	0.695
Prob. -3 Corrected	Fully Conn.	0.525	0.925	0.650	0.900	0.850	0.770
	Pruned	0.400	0.825	0.175	0.725	0.225	0.470
	Retrained	0.550	0.975	0.675	0.800	0.775	0.755
Prob. -2 Corrected	Fully Conn.	0.675	0.925	0.775	0.875	0.800	0.810
	Pruned	0.475	0.850	0.625	0.650	0.475	0.615
	Retrained	0.625	0.925	0.800	0.875	0.825	0.810

Despite these improvements, none of the network-level models surpassed the performance of the original region-based classifier. One possible explanation is the

phenomenon of signal dilution. During message passing, activity from strongly task-responsive regions is combined with activity from neighboring regions that may not be involved in the task. Consequently, distinctive activation peaks become less pronounced, making class separation more difficult for the neural network.

This observation highlights a trade-off between predictive performance and biological interpretability. Region-level models achieve the highest accuracies but provide information primarily about individual brain areas. In contrast, network-level models incorporate anatomical context and therefore offer a more biologically meaningful representation of large-scale brain organization, albeit at some cost in classification performance.

The correction factor further reveals interesting differences between connectomes. Its beneficial effects are most evident for probabilistic matrices and for the anatomy-driven deterministic matrix, particularly when networks contain a large number of connections. By normalizing the summed activity according to network size, the correction factor appears to mitigate the signal dilution introduced by broad connectivity matrices.

Analysis of pruning effects also suggest that the different connectomes encode relevant information in distinct ways. Models exhibiting large performance drops after pruning likely distribute informative features across many pathways, whereas models showing smaller decreases appear to concentrate relevant information within a smaller subset of highly informative connections. The reduced impact of pruning observed in several corrected models supports the hypothesis that normalization promotes a more efficient concentration of task-relevant information.

Across all models, motor conditions involving the feet consistently represent the most difficult classification problem. In particular, LF trials show reduced performance throughout the analysis and are frequently confused with RF trials. This pattern suggests that the challenge is intrinsic to the underlying functional data rather than being introduced by the connectivity modelling procedure itself.

Taken together, these findings support three principal conclusions. First, anatomy-driven deterministic connectivity provides the most effective structural framework for the message-passing approach explored in this study. Secondly, increased sparsity of structural connectivity matrices often benefits classification performance. Third, applying a correction factor tends to enhance model performance by reducing the influence of network size on signal integration.

Most importantly, although network-level models do not surpass region-level classifiers in predictive accuracy, they provide a biologically informed representation of functional activity that may facilitate the identification of distributed brain networks involved in motor behavior. This additional level of neurobiological interpretation represents a significant advantage when the ultimate goal extends beyond classification and includes understanding the organization of brain function.

Motor control depends strongly on well-characterized cortical regions within the precentral gyrus, as classically described by Penfield and Rasmussen (1950). However, many complex cognitive and social processes cannot be attributed to isolated brain regions, instead arising from coordinated activity across distributed functional networks. In this context, the findings presented in this chapter raise new questions concerning the ability of the current model to generalize to paradigms involving neural mechanisms that are less spatially constrained and more dependent on whole-brain integration.

Chapter 4
From Motor Function to Social Cognition

4.1. Introduction

The previous chapter explored the incorporation of structural connectivity information into the classification of motor-task fMRI data obtained from the Human Connectome Project. Motor paradigms provide an advantageous framework for methodological development because they evoke robust and highly reproducible activation patterns within well-characterized sensorimotor systems. Consequently, they offer an appropriate setting in which to evaluate the effects of incorporating structural connectivity constraints into machine-learning models.

However, many cognitive processes of neuroscientific interest cannot be adequately explained by the activity of isolated cortical regions. Higher-order functions such as social cognition emerge from coordinated interactions across distributed large-scale brain networks.

One of the central components of social cognition is Theory of Mind (ToM), defined as the capacity to attribute beliefs, intentions and emotions to other individuals or inanimate objects. Neuroimaging studies have consistently demonstrated that ToM depends on the coordinated recruitment of a distributed mentalizing network, including regions such as the medial prefrontal cortex, temporoparietal junction, posterior superior temporal sulcus and precuneus/posterior cingulate cortex (Frith & Frith, 2006; Saxe et al., 2004; Schurz et al., 2014; Van Overwalle, 2009). Importantly, these regions do not operate independently but rather interact dynamically to support social-cognitive processing.

The present chapter investigates whether the message-passing framework introduced previously can provide biologically meaningful representations of these distributed social-cognitive processes. This paradigm constitutes a more demanding test of the hypothesis that anatomically constrained message passing can improve the biological interpretability of machine-learning models without substantially compromising predictive performance.

4.2. Methods

The dataset, preprocessing procedures, feature extraction methods, neural network architecture, pruning strategy, and message-passing implementation are identical to those described in the previous chapter.

The present study employed the Social Cognition task from the Human Connectome Project. This task assesses aspects of Theory of Mind (ToM), i.e., the ability to attribute mental states to other individuals. During the task, participants view short animations depicting the movement of geometric shapes. In the mentalizing condition, the shapes interact in a manner suggestive of intentional social behaviour. In contrast, the random condition consists of similar visual stimuli in which the shapes move unpredictably, lacking any apparent intentional interaction. Following each animation, participants

indicate whether the observed movements represent a social interaction or random motion.

To make the model interpretable and allow for neuroscientific validation, Shapley values are applied to every feature on every output, according to the formula below:

$$\varphi_i = \frac{1}{|N|!}\sum_{S\subseteq N\setminus i}|S|!\,(|N| - |S| - 1)!\,[f(S \cup \{i\}) - f(S)] \quad (7)$$

Where φ_i is the Shapley value of feature i and tells how much it has contributed to the prediction; N is the full set of features and |N| is how many there are; S can be any subset of features that does not include i ; $\frac{1}{|N|!}$ divides by the total number of orderings of all features, turning the weighted sum into an average; $f(S)$ is the prediction produced by subset S alone. $f(S \cup \{i\}) - f(S)$ is the marginal contribution of feature i and $|\,S\,|\,!$ ($|\,N\,| - |\,S\,| - 1)!$ is a combinatorial weight that counts how many of those orderings put feature i in a specific spot.

Shapley values segregated networks into three groups per ToM class: those with a positive impact (increasing the predicted probability of that movement), those with a negative impact (decreasing it), and those with negligible impact.

The connectivity matrix is the anatomy-driven connectivity matrix as it is the one that achieves better accuracies in the standard motor model. The message passing mechanism on the SNN classifier is represented in Figure 4.

In this study, the SHAP analysis ranks highly the brain networks whose putative absence would mostly impact a drop in accuracy. However, these must be compared to established neuroscientific knowledge. To establish the baseline for this comparison a GLM-based statistical parametric map of a similar mentalizing condition using Frith-Happé animations is extracted from the literature (Dureux et al., 2023). To obtain a quantitative measure of the overlap between this established baseline and the identified relevant networks from the SHAP analysis, the overlap coefficient is used, according to formula 4.

To evaluate the contribution of each region within each network, the Cohen's D factor is computed for every pair of regions in every network for mentalizing and random classes, according to the following formula:

$$d = \frac{\overline{x_1} - \overline{x_2}}{s} \quad (8)$$

Where d is the Cohen's D values, $\overline{x_1}$ is the mean of the first set of data, $\overline{x_2}$ is the mean of the second set of data points, and s is the pooled standard deviation. Cohen's D measures the distance between two sets of data points. A value closer to 0 means both sets of data points are close together, while a higher or lower value means they are far apart.

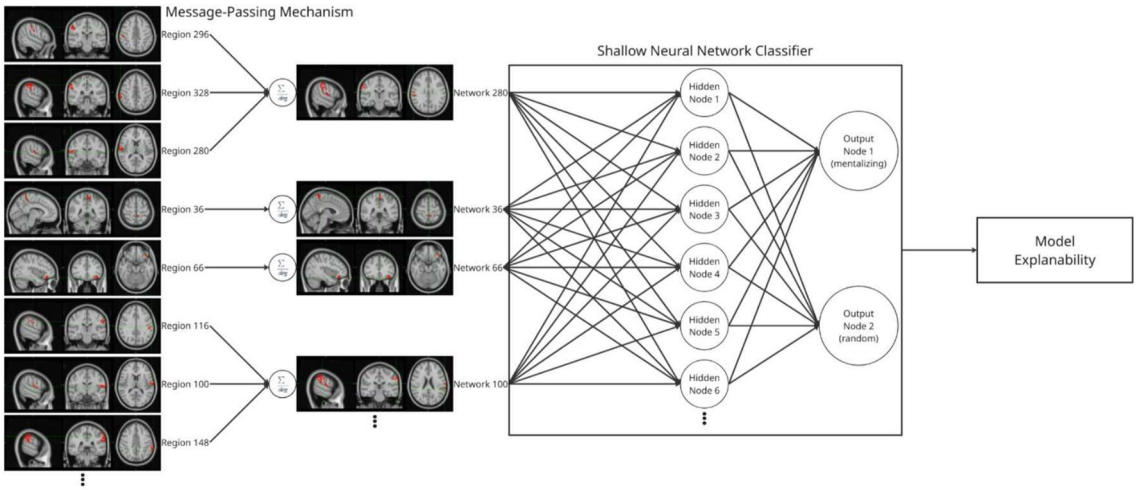

Figure 4: Message-passing mechanism on the SNN classifier. The first two layers represent the process of transforming isolated regions into brain networks through the connectivity matrix. The subsequent layers represent the SNN classifier, while the right portion represents the explainability steps.

4.3. Results

The region-level classification, which serves as the baseline for this study, is described in subchapter 2.4. The confusion matrix, precisions and accuracies of the fully connected, pruned and retrained networks using the message-passing mechanism for the social paradigm are presented in Table 10. Just as observed previously, the fully connected network achieves the highest global accuracy (87.1%), which drops significantly after pruning (77.6%), but is able to recover most of its performance after retraining (84.7%). These results are comparable to those without the message passing mechanism, where the accuracy starts at 88.2% for the fully connected network, drops to 80% with pruning and ends with the exact same value as the model currently in analysis after retraining (84.7%). The mentalizing class achieves consistently higher partial accuracies when compared to the random class in all stages of network development, in agreement with the previous case. The pruning step has a higher negative impact for the mnt class (-11.2%), than for the rnd class (-7.5%), which is in contradiction to the baseline where accuracy of the rnd class decreases more (-12.5%) with pruning, than mnt (-4.5%). In the retrained network, partial accuracies favour the mnt with the message-passing mechanism when comparing to the baseline. This pattern suggests that the model in analysis is particularly sensitive to mentalizing-related network patters, potentially providing meaningful insights into stimulus-induced network recruitment.

Table 10: Confusion matrix, precisions and accuracies for the fully connected, pruned and retrained message passing networks using ToM paradigm.

Network	Stimulus		Prediction		Total
			mnt	rnd	
Fully Connected Network	Input	mnt	43	2	45
		rnd	9	31	40
	Total		52	33	85
	Accuracy (%)		95.6	77.5	87.1
	Precision (%)		82.7	93.9	
Pruned Network	Input	mnt	38	7	45
		rnd	12	28	40
	Total		50	35	85
	Accuracy (%)		84.4	70	77.6
	Precision (%)		76	80	
Retrained Network	Input	mnt	43	2	45
		rnd	11	29	40
	Total		54	31	85
	Accuracy (%)		95.6	72.5	84.7
	Precision (%)		79.6	93.5	

4.3.1. SHAP Values

The beeswarm plots for only the top 20 highest-ranked SHAP values are presented in Figure 5, for ease of interpretation.

Because SHAP explanations are additive and the classification problem involves only two categories, the contribution of each brain network is inherently symmetric between classes. Consequently, a feature that increases the likelihood of classification into the mentalizing (mnt) condition simultaneously decreases the likelihood of classification into the random (rnd) condition, and vice versa.

Based on the magnitude of the SHAP values, networks 109, 40, 291, 48, 228, 277, 85, and 351 are identified as the strongest positive contributors to the rnd class. In contrast, networks 280, 36, 66, 100, 192, 266, 8, 282, 335, 139, 120, and 301 show positive contributions toward the mnt class. The anatomical composition of these networks, together with the class to which they contribute positively, is summarized in Table 11.

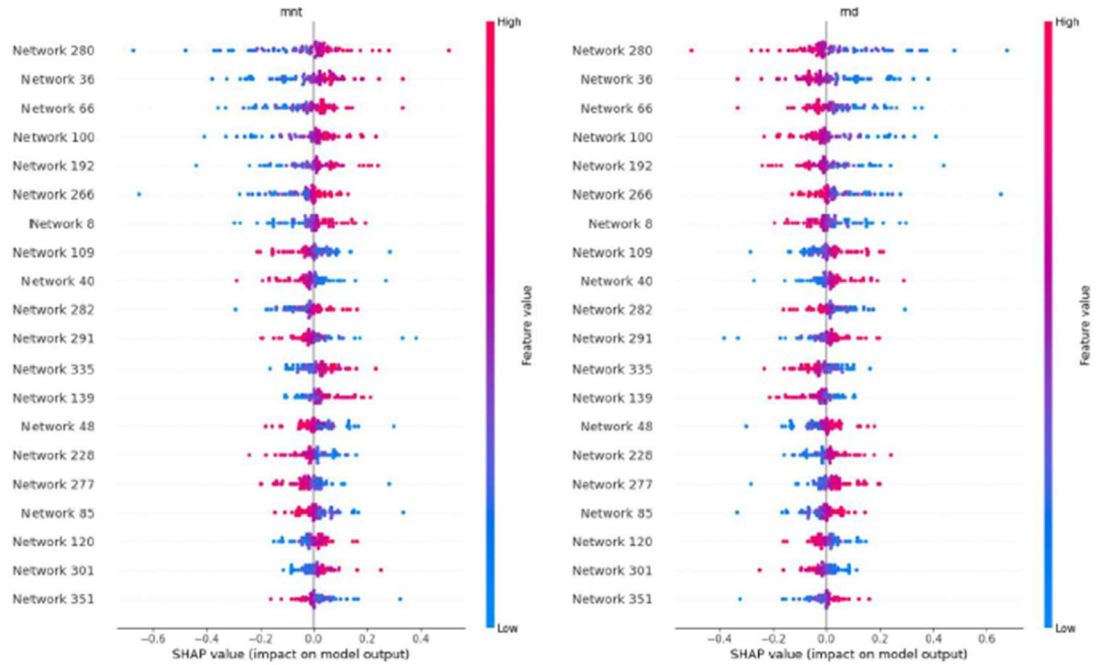

Figure 5: Distribution of the twentieth highest ranked SHAP values of the retrained network for mentalizing (left) and random (right) classes. A higher rank means greater influence on the output computation. Red on the right means positive contribution whereas blue on the right means negative contribution.

The identification of networks that positively contribute to mnt class naturally prompts a comparison with the brain regions most commonly reported in neuroscientific research. Figure 6 illustrates this comparison with the GLM-based statistical parametric map obtained from Dureux et al. (2023). In this representation, regions identified in the literature are shown in green, networks highlighted by the SNN-SHAP framework are displayed in red, and areas of spatial overlap are indicated in yellow.

The literature-derived GLM map encompasses 47.840 voxels, whereas the networks selected by the proposed approach account for 7.664 voxels. The intersection between both sets comprises 1.910 voxels, corresponding to an overlap coefficient of 0.249. While this coefficient suggests a relatively modest degree of spatial correspondence, visual inspection of Figure 6 indicates that much of the overlap is concentrated within regions that are consistently implicated in ToM processing. Furthermore, applying the same statistical threshold used in the original study ($z > 2.57$) results in extensive clusters that extend into white-matter regions. Considering that the BOLD signal primarily reflects neural activity within gray matter, such white-matter activations are unlikely to have a meaningful biological interpretation and are more plausibly explained by artifacts introduced through spatial smoothing procedures.

To provide a more biologically relevant comparison, the literature-derived GLM map is therefore constrained using the MMP1.0 atlas, which contains only cortical gray-matter parcels. Apart from this masking step, all processing parameters remain unchanged. The resulting map is shown in Figure 7.

After applying the MMP1.0 mask, the number of voxels contained within the GLM map decreases from 47.840 to 17.875, indicating that 29.965 voxels originally classified as active are located outside cortical gray matter and therefore have limited biological plausibility. The masking procedure also facilitates a clearer visualization of the overlap between the two approaches. The most prominent intersections are observed in region 36

(area 5m within the left paracentral lobular and mid-cingulate cortex, postcentral gyrus; panel A of Figure 7), region 79 (area IFJa in the left inferior frontal cortex, inferior frontal gyrus; panel B), region 317 (area PHT in the right lateral temporal cortex, inferior temporal gyrus; panel C), region 328 (area PF complex in the right inferior parietal cortex, supramarginal gyrus; panel D), and region 148 (area PF complex in the left inferior parietal cortex, supramarginal gyrus; panel E). Additional noteworthy overlaps are identified in regions 8, 78, 80, 116, 139, 148, 191, 192, 280, 296, 324, and 329.

Table 11: Important networks, according to SHAP analysis, for the classification process and their respective composition in terms of regions. Their absence during Shapley values estimation leads to a significant drop in classification performance.

Network	Region ID	Region name	Lobe	Cortex
280	280	Area_OP4-PV_R	Par	Posterior_Opercular
	296	Area_PFt_R	Par	Inferior_Parietal
	328	Area_PF_Complex_R	Par	Inferior_Parietal
36	36	Area_5m_L	Par	Paracentral_Lobular_and_Mid_Cingulate
66	66	Area_47m_L	Fr	Orbital_and_Polar_Frontal
100	100	Area_OP4-PV_L	Par	Posterior_Opercular
	116	Area_PFt_L	Par	Inferior_Parietal
	148	Area_PF_Complex_L	Par	Inferior_Parietal
192	192	Area_55b_R	Fr	Premotor
	275	Area_Lateral_IntraParietal_dorsal_R	Par	Superior_Parietal
	311	Area_TG_dorsal_R	Temp	Lateral_Temporal
	317	Area_PHT_R	Temp	Lateral_Temporal
	324	Area_IntraParietal_2_R	Par	Inferior_Parietal
	325	Area_IntraParietal_1_R	Par	Inferior_Parietal
	329	Area_PFm_Complex_R	Par	Inferior_Parietal
266	266	Area_9-46d_R	Fr	Dorsolateral_Prefrontal
8	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	82	Area_IFSa_L	Fr	Inferior_Frontal
	116	Area_PFt_L	Par	Inferior_Parietal
	148	Area_PF_Complex_L	Par	Inferior_Parietal
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
109	26	Superior_Frontal_Language_Area_L	Fr	Dorsolateral_Prefrontal
	44	Area_6m_anterior_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	70	Area_8B_Lateral_L	Fr	Dorsolateral_Prefrontal
	104	RetroInsular_Cortex_L	Par	Early_Auditory
	109	Middle_Insular_Area_L	Ins	Insular_and_Frontal_Opercular
	125	Auditory_5_Complex_L	Temp	Auditory_Association
	175	Auditory_4_Complex_L	Temp	Auditory_Association
40	40	Dorsal_Area_24d_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
282	282	Area_OP2-3-VS_R	Par	Posterior_Opercular
291	291	Anterior_Ventral_Insular_Area_R	Ins	Insular_and_Frontal_Opercular
335	302	Perirhinal_Ectorhinal_Cortex_R	Temp	Medial_Temporal
	335	ParaHippocampal_Area_2_R	Temp	Medial_Temporal
139	78	Rostral_Area_6_L	Fr	Premotor
	79	Area_IFJa_L	Fr	Inferior_Frontal
	80	Area_IFJp_L	Fr	Inferior_Frontal
	139	Area_TemporoParietoOccipital_Junction_1_L	Temp	Temporo-Parieto_Occipital_Junction
48	48	Area_Lateral_IntraParietal_ventral_L	Par	Superior_Parietal
	124	ParaBelt_Complex_L	Temp	Early_Auditory
	175	Auditory_4_Complex_L	Temp	Auditory_Association
228	228	Area_Lateral_IntraParietal_ventral_R	Par	Superior_Parietal
	304	ParaBelt_Complex_R	Temp	Early_Auditory
	355	Auditory_4_Complex_R	Temp	Auditory_Association
277	277	Inferior_6-8_Transitional_Area_R	Fr	Dorsolateral_Prefrontal
85	85	Area_anterior_9-46v_L	Fr	Dorsolateral_Prefrontal
120	120	Hippocampus_L	Temp	Medial_Temporal
301	301	ProStriate_Area_R	Par	Posterior_Cingulate
	302	Perirhinal_Ectorhinal_Cortex_R	Temp	Medial_Temporal
351	351	Area_posterior_47r_R	Fr	Inferior_Frontal

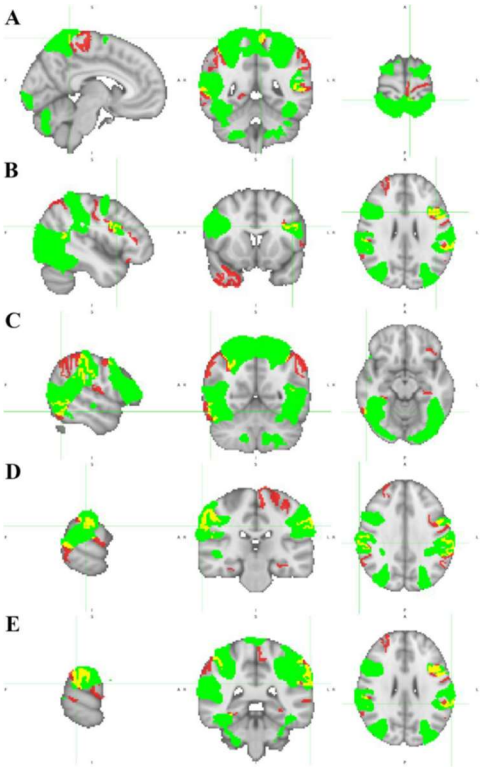

Figure 6: Networks in the top twenty SHAP values that contribute positively to mentalizing overlapped with the GLM-based statistical parametric map obtained from (Dureux et al., 2023). Regions that belong only to the identified networks are in red, areas that are present only in the literature are in green, and regions that are present in both (overlap) are in yellow. Panels display sagittal, coronal, and axial views in MNI152 standard space using radiological convention.

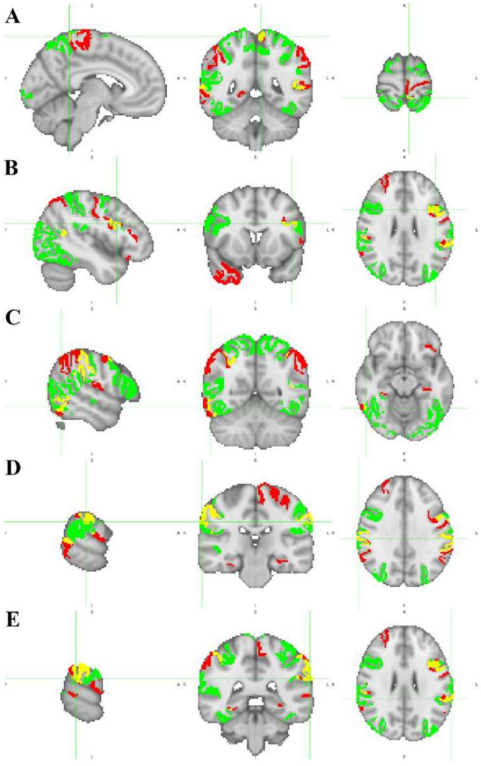

Figure 7: Networks corresponding to the top twenty positive SHAP values for the mentalizing condition overlapped with the GLM-based statistical parametric map reported by Dureux et al. (2023), after masking both datasets with the MMP1.0 atlas to include only cortical gray-matter voxels. Regions identified exclusively by the SHAP-based approach are shown in red, regions reported exclusively in the literature are shown in green, and overlapping regions are shown in yellow. Panels display sagittal, coronal, and axial views in MNI152 standard space using radiological convention.

4.3.2. Identification of Regions within Networks

To better understand which specific regions within the identified networks drive the classifier's distinction between mentalizing and random processes, an additional analysis is performed at the regional level. For this purpose, Cohen's *d* values are calculated separately for the mentalizing and random conditions. Because Cohen's *d* quantifies the standardized difference between group means, it provides a measure of the relative magnitude of the signal associated with each region. This makes it possible to rank the regions within a network according to their contribution to the overall network signal. When combined with the SHAP-based network importance analysis, these rankings provide further insight into how individual regions may influence the classifier's final decision.

Table 12: Cohen's d values for mentalizing-condition instances across regions within the selected networks included in this analysis. Values closer to 0 indicate more similar signal magnitudes between regions, whereas higher absolute values indicate greater differences in regional signal magnitude.

Network		Cohen's D mentalizing							
280	regions		280	296	328				
		280	0.000	-0.831	-0.458				
		296	0.831	0.000	0.386				
		328	0.458	-0.386	0.000				
100	regions		100	116	148				
		100	0.000	-0.755	-0.558				
		116	0.755	0.000	0.191				
		148	0.558	-0.191	0.000				
192	regions		192	275	311	317	324	325	329
		192	0.000	-0.170	-0.216	0.104	-0.178	-0.090	0.384
		275	0.170	0.000	-0.086	0.253	-0.034	0.069	0.596
		311	0.216	0.086	0.000	0.285	0.050	0.135	0.561
		317	-0.104	-0.253	-0.285	0.000	-0.255	-0.180	0.226
		324	0.178	0.034	-0.050	0.255	0.000	0.091	0.551
		325	0.090	-0.069	-0.135	0.180	-0.091	0.000	0.473
		329	-0.384	-0.596	-0.561	-0.226	-0.551	-0.473	0.000
8	regions		8	82	116	148	149		
		8	0.000	-0.568	-0.756	-0.561	-0.184		
		82	0.568	0.000	-0.150	0.032	0.359		
		116	0.756	0.150	0.000	0.191	0.523		
		148	0.561	-0.032	-0.191	0.000	0.342		
		149	0.184	-0.359	-0.523	-0.342	0.000		
109	regions		26	44	70	104	109	125	175
		26	0.000	-0.352	0.267	-0.034	-0.093	0.404	0.385
		44	0.352	0.000	0.630	0.294	0.266	0.739	0.718
		70	-0.267	-0.630	0.000	-0.285	-0.368	0.159	0.142
		104	0.034	-0.294	0.285	0.000	-0.052	0.414	0.396
		109	0.093	-0.266	0.368	0.052	0.000	0.499	0.480
		125	-0.404	-0.739	-0.159	-0.414	-0.499	0.000	-0.016
		175	-0.385	-0.718	-0.142	-0.396	-0.480	0.016	0.000

Since this approach is not informative for networks composed of a single region, the analysis is restricted to the five highest-ranked networks identified through SHAP that contain at least two regions. Networks 36 and 66 are excluded because they are

disconnected according to the connectivity matrix. The resulting regional rankings are presented in Table 12 for the mentalizing condition and Table 13 for the random condition. Examining each network individually reveals distinct patterns across the two experimental conditions. In network 280, the regions are ordered according to decreasing signal magnitude as 296, 328, and 280 during mentalizing. In the random condition, however, the ordering changes slightly to 296, 280, and 328. Although the same regions remain dominant, the differences between them become considerably smaller.

Table 13: Cohen's d values for random-condition instances across regions within the selected networks included in this analysis. Values closer to 0 indicate more similar signal magnitudes between regions, whereas higher absolute values indicate greater differences in regional signal magnitude.

Network		Cohen's D random							
280	regions		280	296	328				
		280	0.000	-0.186	0.009				
		296	0.186	0.000	0.206				
		328	-0.009	-0.206	0.000				
100	regions		100	116	148				
		100	0.000	-0.163	-0.036				
		116	0.163	0.000	0.131				
		148	0.036	-0.131	0.000				
192	regions		192	275	311	317	324	325	329
		192	0.000	-0.160	0.171	0.353	-0.221	-0.097	0.307
		275	0.160	0.000	0.402	0.599	-0.104	0.048	0.511
		311	-0.171	-0.402	0.000	0.214	-0.419	-0.282	0.174
		317	-0.353	-0.599	-0.214	0.000	-0.587	-0.460	-0.019
		324	0.221	0.104	0.419	0.587	0.000	0.127	0.521
		325	0.097	-0.048	0.282	0.460	-0.127	0.000	0.404
		329	-0.307	-0.511	-0.174	0.019	-0.521	-0.404	0.000
8	regions		8	82	116	148	149		
		8	0.000	-0.214	-0.210	-0.075	-0.081		
		82	0.214	0.000	-0.005	0.131	0.127		
		116	0.210	0.005	0.000	0.131	0.126		
		148	0.075	-0.131	-0.131	0.000	-0.006		
		149	0.081	-0.127	-0.126	0.006	0.000		
109	regions		26	44	70	104	109	125	175
		26	0.000	-0.176	-0.038	-0.387	0.100	-0.044	-0.276
		44	0.176	0.000	0.116	-0.222	0.248	0.120	-0.113
		70	0.038	-0.116	0.000	-0.309	0.123	-0.003	-0.211
		104	0.387	0.222	0.309	0.000	0.429	0.322	0.101
		109	-0.100	-0.248	-0.123	-0.429	0.000	-0.132	-0.333
		125	0.044	-0.120	0.003	-0.322	0.132	0.000	-0.219
		175	0.276	0.113	0.211	-0.101	0.333	0.219	0.000

A similar trend can be observed in network 100. During mentalizing, regions 116, 148, and 100 display a clear hierarchical organization in terms of signal magnitude. In contrast, the random condition is characterized by a reduced separation between regions, resulting in a less pronounced ranking structure.

Network 192 exhibits a somewhat different pattern. Under the mentalizing condition, most regions show relatively similar signal magnitudes, with region 329 standing out as consistently lower than the others. During the random condition, the distribution becomes more differentiated, with regions 324 and 275 displaying the highest values, while regions 317 and 329 occupy the lower end of the ranking.

For network 8, a clear ordering is again evident during mentalizing. Region 116 shows the strongest signal contribution, followed by regions 82 and 148, both of which maintain positive average Cohen's *d* values. Regions 149 and 8 contribute to a lesser extent. As observed in previous networks, these differences become substantially attenuated in the random condition.

Finally, network 109 presents a distinct profile. In the mentalizing condition, region 44 displays the largest signal magnitude by a considerable margin. This is followed by regions 109 and 104, which show similar values, and then by region 26. Region 70 contributes less strongly, while regions 125 and 175 are clearly separated from the remaining regions by their lower signal magnitudes. Once again, the random condition is associated with a marked reduction in these differences.

It is worth noting that networks 280, 100, 192, and 8 are identified through SHAP as contributing positively to the mentalizing classification. In contrast, network 109 is the only network among the five that contributed positively to the random classification.

4.4. Discussion

The results demonstrate that, although the proposed framework initially produces a slightly lower global accuracy than the region-level baseline model (87.1% versus 88.2%), both approaches converge to the same final accuracy after pruning and retraining (84.7%). Importantly, this comparable predictive performance is obtained while substantially shifting the interpretative focus of the model. Rather than identifying isolated brain regions independently associated with classification, the proposed framework highlights structurally connected subnetworks that are biologically more plausible representations of large-scale cognitive organization.

This distinction is particularly relevant in the context of cognitive neuroscience. Cognitive functions such as ToM are generally understood not as localized phenomena emerging from a single cortical area, but rather as distributed processes involving coordinated interactions among multiple brain systems. By constraining information propagation according to diffusion-based structural connectivity, the proposed model aligns more closely with this systems-level view of brain organization. In this framework, explanatory units correspond to anatomically grounded networks rather than isolated regions, reducing the likelihood that the classifier exploits biologically implausible or spurious patterns unrelated to known brain architecture.

Interestingly, the retrained model achieves particularly high performance for the mentalizing condition, reaching an accuracy of 95.6%. This finding suggests that structurally organized network recruitment may be especially informative for socially relevant cognitive processes. The SHAP analysis further supports this interpretation, as most of the highest-ranked subnetworks contribute positively toward the mentalizing classification. Specifically, 12 of the 20 most relevant networks favor the mentalizing condition, including all seven top-ranked subnetworks.

The biological relevance of these findings becomes clearer when comparing the identified networks with regions traditionally associated with ToM processing in literature. Several of the subnetworks highlighted by the proposed framework involve regions within the temporo-parietal junction (TPJ), superior temporal sulcus (STS), and inferior frontal gyrus (IFG), all of which have consistently been implicated in mentalizing and social cognition studies (Frith & Frith, 2006; Lavoie et al., 2016; Ogawa & Matsuyama, 2023; Otsuka et al., 2009; Qin et al., 2023; Schurz et al., 2014; Takahashi et al., 2004; Young et

al., 2010). The overlap analysis reveals that the correspondence between the proposed framework and literature-derived GLM activations occurred primarily within regions that are widely recognized as core components of the ToM network. This overlap becomes substantially clearer after masking the GLM maps with the MMP1.0 cortical atlas, which removes a large number of white-matter voxels unlikely to reflect biologically meaningful BOLD activity.

Indeed, one of the most revealing aspects of the overlap analysis concerns the limitations of conventional GLM-based approaches. A considerable proportion of the voxels identified by the literature-derived GLM maps are located within white matter. From a neurophysiological perspective, this is difficult to justify, since the BOLD signal primarily reflects metabolic activity associated with gray matter. White matter contains lower vascular density, weaker neurovascular coupling, and reduced metabolic demand, making meaningful task-related BOLD fluctuations less likely. These apparent activations are therefore more plausibly explained by methodological artifacts, particularly the effects of spatial smoothing, which can blur gray-matter activations into neighboring white-matter regions through partial-volume effects. In contrast, the structurally constrained framework proposed here inherently restricts information propagation to anatomically plausible pathways, thereby improving biological realism.

The comparison between the present network-level analysis and the previous region-based SHAP analysis also provides important insights into how information is represented within the classifier. Surprisingly, most of the top-ranked regions identified in the previous region-level study do not reappear within the most relevant subnetworks identified here. Only region 125, previously ranked as the most important individual feature, appears within network 109, while region 80, previously ranked fourth, appears within network 139. Conversely, several regions that showed only moderate importance in the earlier regional analysis emerge prominently within the structurally informed subnetworks identified in the present study.

This observation suggests that network-level organization may capture distributed patterns of activity that are not evident when regions are analyzed independently. Groups of regions with relatively modest individual contributions may collectively provide stronger information regarding the distinction between mentalizing and random processes than any isolated region alone. In other words, the discriminative information may emerge from coordinated interactions between regions rather than from localized activity itself.

Several mechanisms may explain why some regions gain or lose prominence in the network-based framework. One possible explanation relates to signal dilution during message passing. Regions that appear highly informative in a region-level analysis may be connected to multiple regions unrelated to the task. As information propagates across the network, task-relevant signals may become partially diluted by baseline activity and noise originating from these additional connections. Another possibility concerns the role of the hidden layer within the classifier. During training, hidden nodes may integrate information from multiple structurally connected subnetworks, generating higher-level representations that are more informative for classification than any single regional feature. Under this interpretation, the relevance of a region depends not only on its own activity but also on how it interacts with the surrounding network architecture.

The analysis of Cohen's *d* values provides complementary information regarding the internal organization of the identified subnetworks. Overall, the subnetworks display

larger differences in signal magnitude during the mentalizing condition than during the random condition. This pattern is particularly noteworthy given that most of the subnetworks identified through SHAP contributed positively toward mentalizing classification. One possible interpretation is that mentalizing recruits functionally selective subnetworks in which certain regions become strongly engaged while others remain closer to baseline activity. Such heterogeneous recruitment would naturally produce larger differences in signal magnitude across regions within the same network. By contrast, during the random conditions, activity may remain more homogeneous and closer to baseline levels across the entire network, leading to smaller Cohen's *d* differences. Although this interpretation remains speculative and would require validation using analyses of absolute signal amplitudes, the current results nevertheless suggest the existence of structurally constrained, yet functionally selective subnetworks involved in social cognition.

By incorporating structural constraints, the model reduces the hypothesis space to anatomically feasible configurations, thereby limiting the possibility that the classifier relies on biologically implausible correlations. At the same time, the inclusion of a diagonal self-connection term preserves intra-regional contributions during propagation, while degree normalization mitigates the dilution effects that could emerge in highly connected nodes. Together, these mechanisms allow both local and distributed contributions to remain meaningfully represented within the model.

In summary, the findings of this study show that it is possible to impose anatomical structural constraints on an interpretable decoder without sacrificing accuracy, shifting the explanatory unit from isolated regions to distributed and anatomically grounded subnetworks that correspond to previously established ToM networks.

Despite these promising results, several limitations remain. One important limitation concerns the adjacency matrix used for structural propagation. The current framework relies on anatomy-driven deterministic tractography derived from a limited number of subjects and restricted primarily to cortical intra-hemispheric connectivity. Consequently, important aspects of brain organization remain absent from the model. In particular, subcortical structures and interhemispheric connections are known to play fundamental roles in higher-order cognitive processes, including social cognition. Their exclusion therefore limits the biological completeness of the current framework.

This limitation naturally motivates the next stage of development explored in the following chapter. Incorporating subcortical regions and interhemispheric pathways into the adjacency matrix would allow the model to represent brain organization more realistically and capture interactions that are essential for large-scale cognitive integration. Such extensions may improve not only biological validity but also the capacity of the framework to identify distributed neural signatures underlying complex cognitive processes.

Chapter 5
Introducing Subcortical Regions and Interhemispheric Connections

5.1. Introduction

The studies described in the previous chapters demonstrate that structurally constrained message-passing can be successfully integrated into shallow neural network classifiers to produce biologically interpretable brain decoding models without compromising predictive performance. By shifting interpretation from isolated regions toward anatomically grounded subnetworks, the framework developed provides a representation of cognitive organization that more closely reflects contemporary systems-level neuroscience. The identified subnetworks show meaningful correspondence with established Theory of Mind literature while also revealing distributed patterns that are not apparent in region-level analyses alone. However, the structural matrix that represents the biological constraints of the model developed only includes cortical regions and intra-hemispheric connections. Neuroscience literature elucidates the importance of subcortical regions for certain cognitive processes, functioning as an integration center (Schulte et al., 2025). Additionally, interhemispheric connections can act as a shield to reduce interference between hemispheres, as well as ensure the sharing and integration of resources when task demands are high, enabling parallel processing (Schulte & Müller-Oehring, 2010; van der Knaap & van der Ham, 2011). For this reason, the current study expands the framework previously optimized to a more complete representation of the brain, by incorporating subcortical regions included in the expanded HCP MMP1.0 atlas (HCPex).

5.2. Methods

The HCP-MMP 1.0 atlas is a broadly used parcellation of human cortical areas, combining several organization criteria. However, many functional and structural connectome investigations require the analysis of deep brain structures. The HCPex is an extended version of this atlas, created to fill this gap, adding 66 subcortical areas (33 per hemisphere), including thalamic nuclei, amygdala, putamen, among others (Huang et al., 2022). The additional areas included in the HCPex atlas are represented in Table 14.

Cabrera-Álvarez et al. (2025) implemented a connectivity matrix based on this expanded atlas. The pipeline consisted of the reconstruction of the diffusion data using generalized q-sampling imaging. To map the connections, a deterministic fiber tracking algorithm with augmented strategies was used to ensure reproducibility, similarly to the one implemented in the deterministic population-based connectivity matrix described in 3.4.2. One million seeds were placed throughout the brain to initiate the tracking. Fibers with lengths shorter than 10 mm or longer than 300 mm were discarded to ensure anatomical quality. Two matrices were constructed based on fiber counts (streamlines) and the average length of these fibers connecting each pair of regions.

As this connectivity matrix is composed of counts of streamlines connections between two regions, additionally processing is necessary to transform it into a binary format. The first step is to organize the 360 first labels accordingly with the order of the matrices previously used, keeping the original organization for the following 66 regions.

A threshold is applied with the intention of preserving the same proportion of connections as the best-performing threshold from the population-based deterministic matrix achieved in subsection 3.4.2. The 99% population-based deterministic matrix is composed of 2732 connections out of 129,600 possibilities, which corresponds to approximately 2.11% of connections. This is the percentage of the strongest connections to maintain in the expanded connectivity matrix, resulting in 3826 edges.
The remaining pipeline follows the exact same steps as described for the previous studies, including dataset preprocessing, application of the correction factor for the connectivity matrices, message-passing mechanism, SNN model implementation, pruning and retraining pipeline.

Table 14: Additional subcortical regions added to the HCP MMP1.0 atlas.

ID (L,R)	Abbreviation	Full name	Voxel numbers ($1\ mm^3$) (L, R)
361, 394	AV	Thalamus: Anteroventral Nucleus	256, 280
362, 395	CeM	Thalamus: Central medial	96, 88
363, 396	CL	Thalamus: Central lateral	16, 16
364, 397	CM	Thalamus: Centralmedian	424, 376
365, 398	LD	Thalamus: Laterodorsal	48, 8
366, 399	LGN	Thalamus: Lateral Geniculate	328, 256
367, 400	LP	Thalamus: Lateral Posterior	256, 264
368, 401	L-Sg	Thalamus: Limitans Suprageniculate	32, 16
369, 402	MDl	Thalamus: Mediodorsolateral parvocellular	328, 320
370, 403	MDm	Thalamus: Mediodorsomedial magnocellular	1104, 1192
371, 404	MGN	Thalamus: Medial Geniculate	160, 168
372, 405	MV(Re)	Thalamus: Reuniens	8, 8
373, 406	Pf	Thalamus: Parafascicular	56, 72
374, 407	PuA	Thalamus: Pulvinar anterior	320, 280
375, 408	PuI	Thalamus: Pulvinar inferior	336, 280
376, 409	PuL	Thalamus: Pulvinar lateral	304, 256
377, 410	PuM	Thalamus: Pulvinar medial	1904, 1680
378, 411	VA	Thalamus: Ventral Anterior	608, 616
379, 412	VLa	Thalamus: Ventral Lateral Anterior	880, 856
380, 413	VLp	Thalamus: Ventral Lateral Posterior	1384, 1288
381, 414	VPL	Thalamus: Ventral posterolateral	1600, 1368
382, 415	Putam	Putamen	7896, 7744
383, 416	Caud	Caudate	6896, 6952
384, 417	NAc	Nucleus Accumbens	600, 632
385, 418	Gpe	Globus pallidus externalis	1232, 1144
386, 419	Gpi	Globus pallidus internalis	632, 624
387, 420	Amyg	Amygdala	1608, 1632
388, 421	SNpc	Substantia nigra pars compacta	184, 200
389, 422	SNpr	Substantia nigra pars reticulata	408, 440
390, 423	VTA	Ventral tegmental area	40, 48
391, 424	MB	Mammillary bodies	112, 104
392, 425	Septum	Septal nuclei	248, 136
393, 426	Nb	Nucleus basalis	584, 600

5.3. Results

5.3.1. Region-Level Classification

The results for the motor paradigm using the HCPex atlas without network representations i.e. before the message passing mechanism, are represented in Table 15. The fully connected network achieves a global accuracy of 85.5%, which decreases with pruning to 81%, as seen previously. The model is able to recover all of its original accuracy after retraining. Still, the final accuracy of this model is 0.5 percentual points superior to the final accuracy when using the original atlas, even though the initial accuracy in that case started at 90%. Partial accuracies show that pruning and retraining preferably favours the left foot class, while all of the other classes are negatively affected, just as seen in the previous cases. Precision follows a general pattern of decreasing with pruning and then recovering almost of its accuracy when retrained for most of the classes except for the right hand, which increases after pruning.

Table 15: Region-level results for the fully connected, pruned and retrained networks for the motor paradigm using the HCPex atlas.

Network	Stimulus		Prediction					Total
			LF	LH	RF	RH	T	
Fully Connected Network	Input	LF	25	1	8	2	4	40
		LH	1	39	0	0	0	40
		RF	4	0	33	3	0	40
		RH	1	1	0	38	0	40
		T	1	0	2	1	36	40
	Total		32	41	43	44	40	200
	Accuracy (%)		62.5	97.5	82.5	95	90	85.5
	Precision (%)		78.1	95.1	76.7	86.4	90	
Pruned Network	Input	LF	32	2	3	0	3	40
		LH	2	37	1	0	0	40
		RF	13	0	25	1	1	40
		RH	2	1	3	34	0	40
		T	3	0	2	1	34	40
	Total		52	40	34	36	38	200
	Accuracy (%)		80	92.5	62.5	85	85	81
	Precision (%)		61.5	92.5	73.5	94.4	89.5	
Retrained Network	Input	LF	31	0	6	0	3	40
		LH	1	38	1	0	0	40
		RF	7	0	30	3	0	40
		RH	1	1	2	36	0	40
		T	2	0	1	1	36	40
	Total		42	39	40	40	39	200
	Accuracy (%)		77.5	95	75	90	90	85.5
	Precision (%)		73.8	97.4	75	90	92.3	

Table 16 shows the results of partial and global accuracies and precisions for the fully connected, pruned and retrained neural networks using the extended version of the MMP1.0 atlas, still at a regional level. In this case, pruning causes a decrease in accuracy a bit more accentuated, going from 89.4% to 81.2%, and recovering after retraining is not total, only reaching 84.7%. Interestingly, partial accuracies for the mnt and rnd classes as well as the global accuracy for the retrained network are exactly the same as the ones obtained for the region-level classification using the original HCP MMP1.0 atlas. The rest of the partial accuracies and precisions follow the same pattern for both mnt and rnd classes, never recovering totally the original network accuracy, unlike the last example.

Table 16: Region-level results for the fully connected, pruned and retrained networks, for the ToM paradigm using the HCPex atlas.

Network	Stimulus		Prediction		Total
			mnt	rnd	
Fully Connected Network	Input	mnt	43	2	45
		rnd	7	33	40
	Total		50	35	85
	Accuracy (%)		95.6	82.5	89.4
	Precision (%)		86	94.3	
Pruned Network	Input	mnt	40	5	45
		rnd	11	29	40
	Total		51	34	85
	Accuracy (%)		88.9	72.5	81.2
	Precision (%)		78.4	85.3	
Retrained Network	Input	mnt	41	4	45
		rnd	9	31	40
	Total		50	35	85
	Accuracy (%)		91.1	77.5	84.7
	Precision (%)		82	88.6	

5.3.2. Network-Level Classification

At the network-level, the results for the fully connected, pruned and retrained networks using the HCPex atlas are represented in Table 17. The initial network achieves a global accuracy of 80.5%, a very close value to the one obtained in the uncorrected version of the original atlas (80%), but a smaller value that the one obtained for the region-level using the same atlas (85.5%). Pruning drops drastically the performance to 69% but still not as drastically as the same model but with the original atlas. Retraining fixes accuracy at 81.5%, still a superior value to the one obtained when using the cortical atlas, but

inferior to the region-level classification. In terms of partial accuracies, the left foot is the class that presents worst performance. When comparing with the other models, the right foot is not as much negatively impacted with pruning. Precisions follow the same pattern of high-decrease-recover as accuracies do.

Table 17: Network-level confusion matrices, accuracies and precisions for the fully connected, pruned and retrained networks for the motor paradigm using he HCPex atlas.

Network	Stimulus		Prediction					Total
			LF	LH	RF	RH	T	
Fully Connected Network	Input	LF	27	0	8	1	4	40
		LH	0	38	0	1	1	40
		RF	3	0	29	7	1	40
		RH	1	4	4	31	0	40
		T	2	1	0	1	36	40
	Total		33	43	41	41	42	200
	Accuracy (%)		67.5	95	72.5	77.5	90	80.5
	Precision (%)		81.8	88.4	70.7	75.6	85.7	
Pruned Network	Input	LF	21	2	12	2	3	40
		LH	6	32	0	1	1	40
		RF	3	0	29	5	3	40
		RH	0	6	9	24	1	40
		T	4	1	3	0	32	40
	Total		34	41	53	32	40	200
	Accuracy (%)		52.5	80	72.5	60	80	69
	Precision (%)		61.8	78	54.7	75	80	
Retrained Network	Input	LF	27	0	10	1	2	40
		LH	1	38	0	0	1	40
		RF	3	0	30	7	0	40
		RH	0	5	4	31	0	40
		T	2	1	0	0	37	40
	Total		33	44	44	39	40	200
	Accuracy (%)		67.5	95	75	77.5	92.5	81.5
	Precision (%)		81.8	86.4	68.2	79.5	92.5	

Table 18 summarizes the partial and global accuracies for the three networks obtained using the HCPex atlas for the motor paradigm with the correction factor. The global accuracy starts with a value of 77% which is 8.5 percentual points inferior to the value obtained for the region-level and 6 percentual points inferior than the accuracy of the fully connected network using the HCP MMP1.0 atlas. In this case, pruning has a larger impact than the region level and original atlas cases, both for the global and partial accuracies. Retraining is able to recover accuracies, even though not completely. The final accuracy

of this model is 75%, which is considerably inferior to the region level and the original atlas (82.5%) accuracies. Generally, the correction factor implemented in this model has a negative impact in global accuracies for the three models, which is due mainly to the right foot partial accuracies. Precisions show the exact same behaviour as accuracies. In this case, it is possible to see that the worst performing class is evidently the left foot, which was also evident in the original atlas. However, the most affected class with pruning is the right foot, with a decrease of -22.5 percentual points, but with a recovery with retraining surpassing the original network. Interestingly, this is often the most impacted class with pruning.

Table 18: Network-level confusion matrices, accuracies and precisions for the fully connected, pruned and retrained networks for the motor paradigm using he HCPex atlas with the correction factor.

Network	Stimulus		Prediction					Total
			LF	LH	RF	RH	T	
Fully Connected Network	Input	LF	24	1	14	1	0	40
		LH	2	36	0	2	0	40
		RF	4	2	30	3	1	40
		RH	0	6	3	30	1	40
		T	0	1	2	3	34	40
	Total		30	46	49	39	36	200
	Accuracy (%)		60	90	75	75	85	77
	Precision (%)		80	78.3	61.2	76.9	94.4	
Pruned Network	Input	LF	17	2	12	9	0	40
		LH	9	28	0	2	1	40
		RF	5	3	21	10	1	40
		RH	0	8	2	28	2	40
		T	2	0	2	6	30	40
	Total		33	41	37	55	34	200
	Accuracy (%)		42.5	70	52.5	70	75	62
	Precision (%)		51.5	68.3	56.8	50.9	88.2	
Retrained Network	Input	LF	23	1	15	0	1	40
		LH	4	33	1	1	1	40
		RF	4	0	32	4	0	40
		RH	0	7	4	28	1	40
		T	0	1	2	3	34	40
	Total		31	42	54	36	37	200
	Accuracy (%)		57.5	82.5	80	70	85	75
	Precision (%)		74.2	78.6	59.3	77.8	91.9	

Confusion matrices, as well as accuracies and precisions for the social paradigm using the HCPex atlas are summarized in Table 19, for the fully connected, pruned and retrained networks. The model obtains an initial accuracy of 83.5%, a value inferior to the one obtained when using the original atlas (87.1%). The pruning stage, as expected, produces a decrease in accuracy to 76.5%, even though not as accentuated as the cortical atlas case, which dropped to 77.6%. Retraining is able to recover all of the original accuracy, unlike the anterior case, which only recovered to 84.7%. Despite this, the final HCPex model still has a very small reduction in accuracy when compared with the final HCP MMP1.0 model, with a difference of 1.2 percentual points. The same difference is observed between the region-level and the network level classifications for the same extended atlas version.

Partial accuracies follow a similar pattern as global accuracies of dropping and recovering after pruning and retraining, respectively. However, this time, and in comparison with the original atlas, the final partial accuracies are more balanced between classes, with a difference of only 6.7 percentual points (when compared with the 23.1 difference in the previous case). Precision is way more balanced between the two classes in this case too but still favouring the random class.

Table 19: Network-level confusion matrices, precisions and accuracies for the fully connected, pruned and retrained networks for the social paradigm using the HCPex atlas.

Network	Stimulus		Prediction		Total
			mnt	rnd	
Fully Connected Network	Input	mnt	41	4	45
		rnd	10	30	40
	Total		51	34	85
	Accuracy (%)		91.1	75	83.5
	Precision (%)		80.4	88.2	
Pruned Network	Input	mnt	36	9	45
		rnd	11	29	40
	Total		47	38	85
	Accuracy (%)		80	72.5	76.5
	Precision (%)		76.6	76.3	
Retrained Network	Input	mnt	39	6	45
		rnd	8	32	40
	Total		47	38	85
	Accuracy (%)		86.7	80	83.5
	Precision (%)		83	84.2	

5.3.3. SHAP analysis

In Figure 8 are summarized the top 20 most relevant features for each of the five classes in the motor paradigm when using the HCPex atlas. For the left foot, features more relevant for the correct classification include regions 216, 235, 188, 219, 282, 279, 322, 221 and 104. Regions 231, 189, 234, 232, 76, 220, 6 and 281 are the most relevant for the correct classification of left hand movements. The right foot classifications rely heavily on regions 36, 55, 8, 382, 363 and 144. Features that most contribute positively for right hand class are regions 54, 9, 51, 40, 52, 168, 42, 353 and 47. Finally, regions 53, 236, 100, 116, 280, 189, 8, 223, 233, 76, 231 are among the twentieth more relevant features classifying correctly tongue movements.

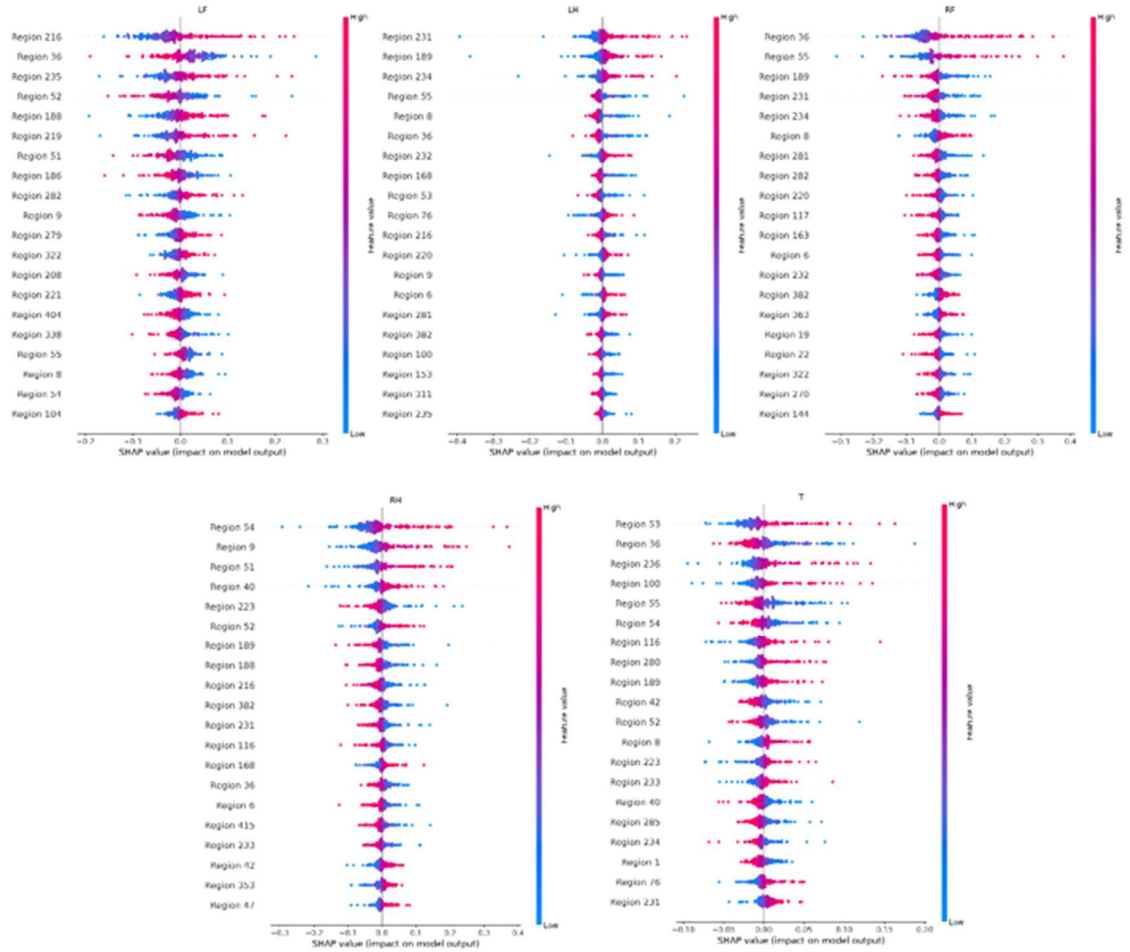

Figure 8: Distribution of the twentieth highest ranked SHAP values of the retrained network, using the HCPex atlas, for left foot (LF), left hand (LH), right foot (RF), right hand (RH) and tongue (T) classes. Red on the right means positive contribution whereas blue on the right means negative contribution.

The top 20 features that most contribute to the social cognition model's predictions are represented in Figure 9. Regions 148, 116, 153, 295, 323, 80, 120, 328 are the ones that most contribute positively for mentalizing processes, while contributing negatively for the opposite class. The features that positively contribute to the random class include regions 13, 199, 177, 355, 133, 304, 372, 125, 415, 16, 292, 276. These results show a predominance of most important features contributing to random processes.

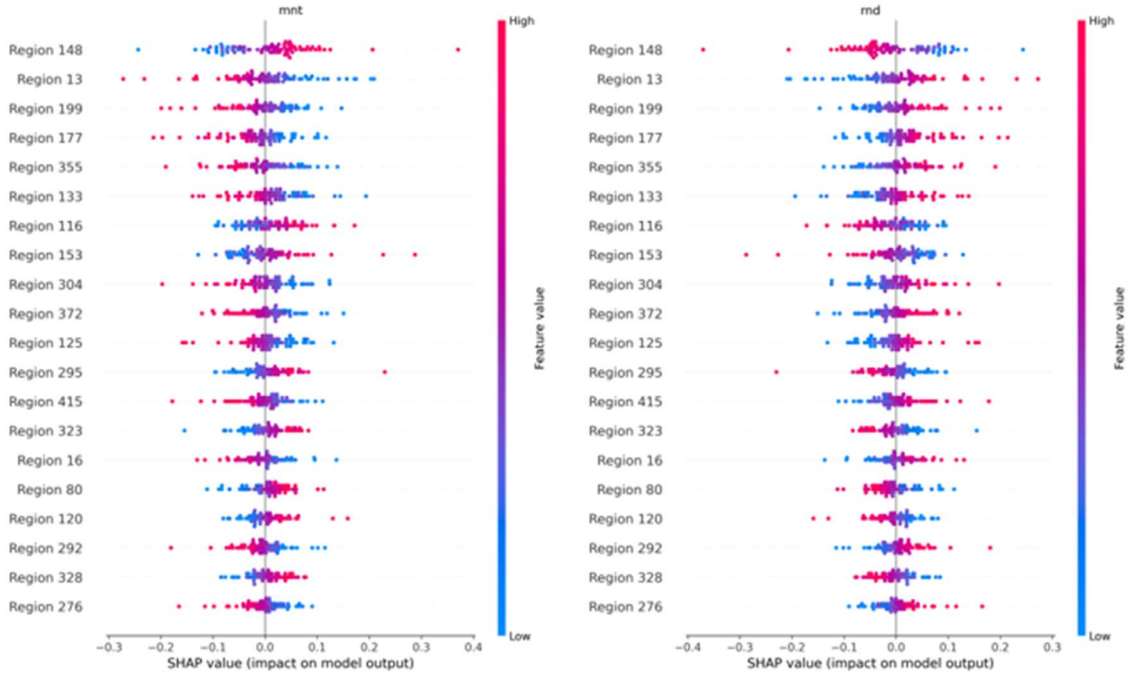

Figure 9: Distribution of the twentieth highest ranked SHAP values of the retrained network, using the HCPex atlas, for mentalizing (left) and random (right) classes. A higher rank means greater influence on the output computation. Red on the right means positive contribution whereas blue on the right means negative contribution.

The composition of the five most relevant networks for the LF, LH, RF, RH and T classes are summarized in Table 20, Table 21, Table 22, Table 23 and Table 24, respectively. The composition of the five most relevant networks for the correct classification of mentalizing and random processes is shown in Table 25. The first thing that is evident to see when comparing to the original atlas case is that, this time, the most relevant networks are composed of a greater number of regions. Consequently, only the top 5 networks are represented for ease of interpretation.

5.4. Discussion

The inclusion of subcortical structures led to a modest increase of 0.5 percentage points in the accuracy of the region-level model compared with the HCP MMP1.0 atlas. Although this difference is too small to be considered a meaningful improvement in predictive performance, it demonstrates that extending the atlas with subcortical regions does not compromise the model's ability to decode task-related brain activity. The resulting accuracy of 85.5% remains comparable to that achieved with the original atlas, indicating that the proposed framework maintains a robust predictive performance despite the increased anatomical complexity. These findings provide support for applying the extended atlas to more complex paradigms, such as social cognition.

Similarly, for the ToM classification, accuracies of the retrained models reached 85% accuracy for the original atlas and 84.7% for the extended atlas, representing a slight reduction of 0.3 percentual points. Together, these results indicate that the inclusion of subcortical regions preserves the decoding performance of the framework while enabling a more comprehensive representation of the organization of the brain.

Table 20: The five most important networks, according to SHAP analysis, for the left foot (LF) classification and their respective composition in terms of regions

Network	Region ID	Region name	Lobe	Cortex
216	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	216	Area_5m_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	217	Area_5m_ventral_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	218	Area_23c_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	221	Ventral_Area_24d_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
36	36	Area_5m_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
235	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	43	Supplementary_and_Cingulate_Eye_Field_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	54	Dorsal_area_6_L	Fr	Premotor
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	236	Ventral_Area_6_R	Fr	Premotor
	250	Area_8B_Lateral_R	Fr	Dorsolateral_Prefrontal
52	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	51	Area_1_L	Par	Somatosensory_and_Motor
	52	Area_2_L	Par	Somatosensory_and_Motor
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
	151	Area_PGs_L	Par	Inferior_Parietal
188	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	36	Area_5m_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	37	Area_5m_ventral_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	38	Area_23c_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	53	Area_3a_L	Fr	Somatosensory_and_Motor
	54	Dorsal_area_6_L	Fr	Premotor
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	221	Ventral_Area_24d_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	231	Area_1_R	Par	Somatosensory_and_Motor
	233	Area_3a_R	Fr	Somatosensory_and_Motor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	329	Area_PFm_Complex_R	Par	Inferior_Parietal
	415	Putamen_R		
	418	Globus palidus externalis_R		
	426	Nucleus basalis_R		

Table 21: The five most important networks, according to SHAP analysis, for the left hand (LH) classification and their respective composition in terms of regions

Network	Region ID	Region name	Lobe	Cortex
231	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	190	Frontal_Eye_Fields_R	Fr	Premotor
	192	Area_55b_R	Fr	Premotor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	227	Area_7PC_R	Par	Superior_Parietal
	229	Ventral_IntraParietal_Complex_R	Par	Superior_Parietal
	230	Medial_IntraParietal_Area_R	Par	Superior_Parietal
	231	Area_1_R	Par	Somatosensory_and_Motor
	232	Area_2_R	Par	Somatosensory_and_Motor
	234	Dorsal_area_6_R	Fr	Premotor
	236	Ventral_Area_6_R	Fr	Premotor
	247	Area_8Av_R	Fr	Dorsolateral_Prefrontal
	254	Area_44_R	Fr	Inferior_Frontal
	261	Area_IFSp_R	Fr	Inferior_Frontal
	263	Area_posterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	265	Area_anterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	277	Inferior_6-8_Transitional_Area_R	Fr	Dorsolateral_Prefrontal
	288	Frontal_Opercular_Area_4_R	Fr	Insular_and_Frontal_Opercular
	325	Area_IntraParietal_1_R	Par	Inferior_Parietal
	329	Area_PFm_Complex_R	Par	Inferior_Parietal
	331	Area_PGs_R	Par	Inferior_Parietal
	349	Area_Frontal_Opercular_5_R	Ins	Insular_and_Frontal_Opercular
189	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	37	Area_5m_ventral_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	39	Area_5L_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	231	Area_1_R	Par	Somatosensory_and_Motor
	234	Dorsal_area_6_R	Fr	Premotor
	247	Area_8Av_R	Fr	Dorsolateral_Prefrontal
	415	Putamen_R		
	418	Globus palidus externalis_R		
234	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	26	Superior_Frontal_Language_Area_L	Fr	Dorsolateral_Prefrontal
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	224	Area_6m_anterior_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	231	Area_1_R	Par	Somatosensory_and_Motor
	234	Dorsal_area_6_R	Fr	Premotor
	250	Area_8B_Lateral_R	Fr	Dorsolateral_Prefrontal

	254	Area_44_R	Fr	Inferior_Frontal
	257	Area_anterior_47r_R	Fr	Inferior_Frontal
	258	Rostral_Area_6_R	Fr	Premotor
	261	Area_IFSp_R	Fr	Inferior_Frontal
	263	Area_posterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	265	Area_anterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	276	Area_6_anterior_R	Fr	Premotor
	277	Inferior_6-8_Transitional_Area_R	Fr	Dorsolateral_Prefrontal
	349	Area_Frontal_Opercular_5_R	Ins	Insular_and_Frontal_Opercular
55	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	12	Area_55b_L	Fr	Premotor
	39	Area_5L_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	42	Lateral_Area_7A_L	Par	Superior_Parietal
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	56	Ventral_Area_6_L	Fr	Premotor
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	78	Rostral_Area_6_L	Fr	Premotor
	85	Area_anterior_9-46v_L	Fr	Dorsolateral_Prefrontal
	96	Area_6_anterior_L	Fr	Premotor
	99	Area_43_L	Fr	Posterior_Opercular
	151	Area_PGs_L	Par	Inferior_Parietal
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	190	Frontal_Eye_Fields_R	Fr	Premotor
	192	Area_55b_R	Fr	Premotor
	194	RetroSplenial_Complex_R	Par	Posterior_Cingulate
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	223	Supplementary_and_Cingulate_Eye_Field_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	234	Dorsal_area_6_R	Fr	Premotor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
8	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	10	Frontal_Eye_Fields_L	Fr	Premotor
	43	Supplementary_and_Cingulate_Eye_Field_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	51	Area_1_L	Par	Somatosensory_and_Motor
	53	Area_3a_L	Fr	Somatosensory_and_Motor
	54	Dorsal_area_6_L	Fr	Premotor
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	133	Area_TE1_posterior_L	Temp	Lateral_Temporal
	137	Area_PHT_L	Temp	Lateral_Temporal
	148	Area_PF_Complex_L	Par	Inferior_Parietal
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	194	RetroSplenial_Complex_R	Par	Posterior_Cingulate
	209	Medial_Area_7P_R	Par	Superior_Parietal
	213	Area_ventral_23_a+b_R	Par	Posterior_Cingulate

	216	Area_5m_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	225	Medial_Area_7A_R	Par	Superior_Parietal
	233	Area_3a_R	Fr	Somatosensory_and_Motor
	234	Dorsal_area_6_R	Fr	Premotor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	377	Thalamus: Pulvinar medial_L		
	382	Putamen_L		

For the network level classification, accuracies for both the original atlas and its extended version also achieved close accuracy values. However, in this case for the motor paradigm, the uncorrected version of the connectivity matrix with subcortical regions exceeds the uncorrected anatomical connectivity matrix by 2 percentual points. While when applying a correction factor, accuracies increased previously (from 79.5% to 82.5% in the retrained network), for the extended atlas version, applying a correction factor hinders performance, as this step decreases accuracies from 81.5% to 75%. In the other hand, the anterior population-based connectivity matrix used also suffered a decrease in accuracies with the application of a correction factor. This is not surprising, as the methods used to obtain the connectivity matrix for the HCPex atlas are similar to the ones used to obtain that previous matrix.

For the social paradigm, increasing the number of regions resulted in a small reduction of 1.2 percentual points in accuracy for the retrained model. However, this decrease is too small to rule out the possibility of applying this model.

The differences seen between the only cortical atlas and the more complete atlas never exceed 5 percentual points (except for the corrected anatomical matrices). In fact, the model introduces 66 new subcortical regions, which increases the input set to 426 input nodes and 4260 connections between the first and second layers. Consequently, the complexity of the model increases substantially. Despite this increase, the same ANN architecture was retained, consisting of a single hidden layer with 10 hidden nodes, and the training procedure remained unchanged, with no optimization of hyperparameters to accommodate the larger and more complex input space.

Remarkably, the extended models achieved predictive performances comparable to those obtained with the cortical atlas. These findings suggest that the incorporation of subcortical structures and interhemispheric connections does not provide a substantial improvement in decoding accuracy for the motor and social paradigms investigated in this work. However, they also demonstrate that a more biologically comprehensive representation of brain organization can be incorporated without compromising model performance. While the extended framework has increased computational complexity, training burden, and interpretational challenges, it preserves predictive accuracy while offering a more complete representation of the underlying neuroanatomy.

The SHAP analysis revealed that the networks most relevant for the correct classification of motor movements include mainly regions from the premotor, somatosensory and motor areas from the frontal and parietal lobes. The regions prioritized by the model are responsible for visuomotor mapping, sensorimotor integration, planning and execution, respectively (Hardwick et al., 2013; Wei et al., 2018). Additionally, some subcortical regions are also consistently present in these main networks across the different motor

movements, including the thalamus, putamen and globus pallidus. The thalamus is responsible for somatosensory and motor integration, while the putamen and globus pallidus are known to be related with the correct movement choice and inhibition of undesired movements, contributing to the execution of fluid actions (Lacourse et al., 2005; Nambu, 2007).

The fact that the model uses typical literature regions associated with motor functions indicates that the accuracies achieved are not random. In reality, the model demonstrates an approximation to the true functioning of the human brain. However, given that the final model becomes considerably more complex, while the predictive performance remains equivalent to the simplistic model, these paradigms are not the most suitable for evaluating the impact of adding subcortical regions. Motor functions, in particular, are very well localized in the motor cortex, responsible for their planning, control, and execution, so it was already expected that the addition of basal ganglia, although equally important for the coordinated execution of movements, would not have such a visible weighting in the model.

The most striking result when observing the networks most important for the correct classifications of the model is the considerable increase in the number of regions in each network. In addition to sparser networks, it is possible to observe that many of these networks include regions from both hemispheres, which would not be possible using the previous matrices due to the exclusion of interhemispheric connections. A concrete example is the second most important region for the social paradigm, network 13, which is composed of 22 regions in the right and left hemispheres. Compared to the previous model that used only cortical regions and no commissural representations, the network with the highest number of connections in the top 20 was composed of only 7 regions, and 10 of these networks were connected only to themselves. These results highlight the typical activation pattern of the human brain, characterized by distributed networks of simultaneous activations and suppressions to perform perfectly its executive tasks.

The networks that most importantly distinguish mentalizing from random classes are distributed across all lobes. Regions from the temporo-parietal junction, superior temporal sulcus, and inferior frontal gyrus, responsible for perspective taking, social cues processing and understanding actions and emotional expressions through mirroring mechanisms, respectively, are considered relevant for the task, just as seen previously (Balgova et al., 2024; Kanske et al., 2015; Schurz et al., 2014).

Additionally, the top 5 most relevant features reveal a considerable number of subcortical regions, including several subdivisions of the thalamus, putamen and globus pallidus. The thalamus has been previously related with social cognition through interactions with the prefrontal cortex, being essential for learning tasks and social decision-making (Koshiyama et al., 2018). While no direct influence of the putamen and globus pallidus in social cognition has been identified, they have been described as fundamental hubs that, through their cortico-subcortical circuits, influence executive functions and social behavior (Okada et al., 2023).

The results obtained with this work confirm once more the ability of the model proposed to decode mental brain states. The accuracies above 80% indicate that the ANN proposed here has learned the underlying neurophysiological signatures of the mental states in evaluation rather than simply memorizing the training data.

However, all of the studies used a single partition between training and test subjects. Consequently, although more complete and biologically informed models obtained higher

Table 22: The five most important networks, according to SHAP analysis, for the right foot (RF) classification and their respective composition in terms of regions

Network	Region ID	Region name	Lobe	Cortex
36	36	Area_5m_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
55	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	12	Area_55b_L	Fr	Premotor
	39	Area_5L_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	42	Lateral_Area_7A_L	Par	Superior_Parietal
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	56	Ventral_Area_6_L	Fr	Premotor
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	78	Rostral_Area_6_L	Fr	Premotor
	85	Area_anterior_9-46v_L	Fr	Dorsolateral_Prefrontal
	96	Area_6_anterior_L	Fr	Premotor
	99	Area_43_L	Fr	Posterior_Opercular
	151	Area_PGs_L	Par	Inferior_Parietal
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	190	Frontal_Eye_Fields_R	Fr	Premotor
	192	Area_55b_R	Fr	Premotor
	194	RetroSplenial_Complex_R	Par	Posterior_Cingulate
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	223	Supplementary_and_Cingulate_Eye_Field_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	234	Dorsal_area_6_R	Fr	Premotor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
189	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	37	Area_5m_ventral_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	39	Area_5L_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	231	Area_1_R	Par	Somatosensory_and_Motor
	234	Dorsal_area_6_R	Fr	Premotor
	247	Area_8Av_R	Fr	Dorsolateral_Prefrontal
	415	Putamen_R		
	418	Globus palidus externalis_R		
231	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	190	Frontal_Eye_Fields_R	Fr	Premotor
	192	Area_55b_R	Fr	Premotor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	227	Area_7PC_R	Par	Superior_Parietal
	229	Ventral_IntraParietal_Complex_R	Par	Superior_Parietal
	230	Medial_IntraParietal_Area_R	Par	Superior_Parietal
	231	Area_1_R	Par	Somatosensory_and_Motor
	232	Area_2_R	Par	Somatosensory_and_Motor

	234	Dorsal_area_6_R	Fr	Premotor
	236	Ventral_Area_6_R	Fr	Premotor
	247	Area_8Av_R	Fr	Dorsolateral_Prefrontal
	254	Area_44_R	Fr	Inferior_Frontal
	261	Area_IFSp_R	Fr	Inferior_Frontal
	263	Area_posterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	265	Area_anterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	277	Inferior_6-8_Transitional_Area_R	Fr	Dorsolateral_Prefrontal
	288	Frontal_Opercular_Area_4_R	Fr	Insular_and_Frontal_Opercular
	325	Area_IntraParietal_1_R	Par	Inferior_Parietal
	329	Area_PFm_Complex_R	Par	Inferior_Parietal
	331	Area_PGs_R	Par	Inferior_Parietal
	349	Area_Frontal_Opercular_5_R	Ins	Insular_and_Frontal_Opercular
	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	26	Superior_Frontal_Language_Area_L	Fr	Dorsolateral_Prefrontal
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	222	Lateral_Area_7A_R	Par	Superior_Parietal
	224	Area_6m_anterior_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	231	Area_1_R	Par	Somatosensory_and_Motor
	234	Dorsal_area_6_R	Fr	Premotor
234	250	Area_8B_Lateral_R	Fr	Dorsolateral_Prefrontal
	254	Area_44_R	Fr	Inferior_Frontal
	257	Area_anterior_47r_R	Fr	Inferior_Frontal
	258	Rostral_Area_6_R	Fr	Premotor
	261	Area_IFSp_R	Fr	Inferior_Frontal
	263	Area_posterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	265	Area_anterior_9-46v_R	Fr	Dorsolateral_Prefrontal
	276	Area_6_anterior_R	Fr	Premotor
	277	Inferior_6-8_Transitional_Area_R	Fr	Dorsolateral_Prefrontal
	349	Area_Frontal_Opercular_5_R	Ins	Insular_and_Frontal_Opercular

predictive performance, it is not possible to say whether there are clear improvements or whether the results only reflect the effects of the partitioning used. Even so, the present model is a more faithful representation of the true organization and functioning of the human brain, shifting the interpretation from isolated brain regions to distributed networks across the whole brain without decreased predictive performance.

Overall, the inclusion of interhemispheric connections brought a clear improvement in the identification of large-scale distributed networks. Still, for motor movement classification or mentalizing and random processes differentiation, this model is not appropriate once its increased complexity does not bring any substantial benefit in terms of predictive performance. To further investigate the influence of subcortical structures in the present model, other paradigms that recruit more prominently subcortical structures, such as reward-based and habit learning or working and episodic memory, should be tested.

Table 23: The five most important networks, according to SHAP analysis, for the right hand (RH) classification and their respective composition in terms of regions

Network	Region ID	Region name	Lobe	Cortex
54	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	10	Frontal_Eye_Fields_L	Fr	Premotor
	44	Area_6m_anterior_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	51	Area_1_L	Par	Somatosensory_and_Motor
	54	Dorsal_area_6_L	Fr	Premotor
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	84	Area_46_L	Fr	Dorsolateral_Prefrontal
	85	Area_anterior_9-46v_L	Fr	Dorsolateral_Prefrontal
	96	Area_6_anterior_L	Fr	Premotor
	97	Inferior_6-8_Transitional_Area_L	Fr	Dorsolateral_Prefrontal
	151	Area_PGs_L	Par	Inferior_Parietal
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	190	Frontal_Eye_Fields_R	Fr	Premotor
	192	Area_55b_R	Fr	Premotor
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	223	Supplementary_and_Cingulate_Eye_Field_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
9	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	10	Frontal_Eye_Fields_L	Fr	Premotor
	12	Area_55b_L	Fr	Premotor
	48	Area_Lateral_IntraParietal_ventral_L	Par	Superior_Parietal
	51	Area_1_L	Par	Somatosensory_and_Motor
	52	Area_2_L	Par	Somatosensory_and_Motor
	53	Area_3a_L	Fr	Somatosensory_and_Motor
	54	Dorsal_area_6_L	Fr	Premotor
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	75	Area_45_L	Fr	Inferior_Frontal
	85	Area_anterior_9-46v_L	Fr	Dorsolateral_Prefrontal
	106	Posterior_Insular_Area_2_L	Ins	Insular_and_Frontal_Opercular
	108	Frontal_Opercular_Area_4_L	Fr	Insular_and_Frontal_Opercular
	109	Middle_Insular_Area_L	Ins	Insular_and_Frontal_Opercular
	110	Pirform_Cortex_L	Temp	Insular_and_Frontal_Opercular
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
	151	Area_PGs_L	Par	Inferior_Parietal
	169	Area_Frontal_Opercular_5_L	Fr	Insular_and_Frontal_Opercular
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	189	Primary_Sensory_Cortex_R	Fr	Somatosensory_and_Motor
	194	RetroSplenial_Complex_R	Par	Posterior_Cingulate
	213	Area_ventral_23_a+b_R	Par	Posterior_Cingulate
	216	Area_5m_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	217	Area_5m_ventral_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	219	Area_5L_R	Par	Paracentral_Lobular_and_Mid_Cingulate
	225	Medial_Area_7A_R	Par	Superior_Parietal

	233	Area_3a_R	Fr	Somatosensory_and_Motor
51	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	10	Frontal_Eye_Fields_L	Fr	Premotor
	12	Area_55b_L	Fr	Premotor
	42	Lateral_Area_7A_L	Par	Superior_Parietal
	47	Area_7PC_L	Par	Superior_Parietal
	51	Area_1_L	Par	Somatosensory_and_Motor
	52	Area_2_L	Par	Somatosensory_and_Motor
	53	Area_3a_L	Fr	Somatosensory_and_Motor
	54	Dorsal_area_6_L	Fr	Premotor
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	73	Area_8C_L	Fr	Dorsolateral_Prefrontal
	74	Area_44_L	Fr	Inferior_Frontal
	75	Area_45_L	Fr	Inferior_Frontal
	78	Rostral_Area_6_L	Fr	Premotor
	84	Area_46_L	Fr	Dorsolateral_Prefrontal
	106	Posterior_Insular_Area_2_L	Ins	Insular_and_Frontal_Opercular
	108	Frontal_Opercular_Area_4_L	Fr	Insular_and_Frontal_Opercular
	109	Middle_Insular_Area_L	Ins	Insular_and_Frontal_Opercular
	110	Pirform_Cortex_L	Temp	Insular_and_Frontal_Opercular
	112	Anterior_Agranular_Insula_Complex_L	Ins	Insular_and_Frontal_Opercular
	134	Area_TE2_anterior_L	Temp	Lateral_Temporal
	137	Area_PHT_L	Temp	Lateral_Temporal
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
	151	Area_PGs_L	Par	Inferior_Parietal
	169	Area_Frontal_Opercular_5_L	Fr	Insular_and_Frontal_Opercular
40	40	Dorsal_Area_24d_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	364	Thalamus: Centralmedian_L		
	370	Thalamus: Mediodorsomedial magnocellular_L		
	377	Thalamus: Pulvinar medial_L		
	379	Thalamus: Ventral Lateral Anterior_L		
	380	Thalamus: Ventral Lateral Posterior_L		
	382	Putamen_L		
	385	Globus palidus externalis_L		
	389	Substantia nigra pars reticulata_L		
223	10	Frontal_Eye_Fields_L	Fr	Premotor
	26	Superior_Frontal_Language_Area_L	Fr	Dorsolateral_Prefrontal
	43	Supplementary_and_Cingulate_Eye_Field_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	54	Dorsal_area_6_L	Fr	Premotor
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	223	Supplementary_and_Cingulate_Eye_Field_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	225	Medial_Area_7A_R	Par	Superior_Parietal
	236	Ventral_Area_6_R	Fr	Premotor
	240	Area_p32_prime_R	Fr	Anterior_Cingulate_and_Medial_Prefrontal

	258	Rostral_Area_6_R	Fr	Premotor
	415	Putamen_R		
	422	Substantia nigra pars reticulata_R		

Table 24: The five most important networks, according to SHAP analysis, for the tongue (T) classification and their respective composition in terms of regions

Network	Region ID	Region name	Lobe	Cortex
53	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	9	Primary_Sensory_Cortex_L	Par	Somatosensory_and_Motor
	51	Area_1_L	Par	Somatosensory_and_Motor
	53	Area_3a_L	Fr	Somatosensory_and_Motor
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
36	36	Area_5m_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
236	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	223	Supplementary_and_Cingulate_Eye_Field_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	224	Area_6m_anterior_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	231	Area_1_R	Par	Somatosensory_and_Motor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	236	Ventral_Area_6_R	Fr	Premotor
	253	Area_8C_R	Fr	Dorsolateral_Prefrontal
	254	Area_44_R	Fr	Inferior_Frontal
	257	Area_anterior_47r_R	Fr	Inferior_Frontal
	261	Area_IFSp_R	Fr	Inferior_Frontal
	329	Area_PFm_Complex_R	Par	Inferior_Parietal
	415	Putamen_R		
100	100	Area_OP4-PV_L	Par	Posterior_Opercular
	148	Area_PF_Complex_L	Par	Inferior_Parietal
55	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	12	Area_55b_L	Fr	Premotor
	39	Area_5L_L	Par	Paracentral_Lobular_and_Mid_Cingulate
	42	Lateral_Area_7A_L	Par	Superior_Parietal
	55	Area_6mp_L	Fr	Paracentral_Lobular_and_Mid_Cingulate
	56	Ventral_Area_6_L	Fr	Premotor
	67	Area_8Av_L	Fr	Dorsolateral_Prefrontal
	78	Rostral_Area_6_L	Fr	Premotor
	85	Area_anterior_9-46v_L	Fr	Dorsolateral_Prefrontal
	96	Area_6_anterior_L	Fr	Premotor
	99	Area_43_L	Fr	Posterior_Opercular
	151	Area_PGs_L	Par	Inferior_Parietal
	188	Primary_Motor_Cortex_R	Fr	Somatosensory_and_Motor
	190	Frontal_Eye_Fields_R	Fr	Premotor
	192	Area_55b_R	Fr	Premotor
	194	RetroSplenial_Complex_R	Par	Posterior_Cingulate
	206	Superior_Frontal_Language_Area_R	Fr	Dorsolateral_Prefrontal
	223	Supplementary_and_Cingulate_Eye_Field_R	Fr	Paracentral_Lobular_and_Mid_Cingulate
	234	Dorsal_area_6_R	Fr	Premotor
	235	Area_6mp_R	Fr	Paracentral_Lobular_and_Mid_Cingulate

Table 25: The five most important networks, according to SHAP analysis, for the social cognition classification and their respective composition in terms of regions

Network	Region ID	Region name	Lobe	Cortex
148	8	Primary_Motor_Cortex_L	Fr	Somatosensory_and_Motor
	47	Area_7PC_L	Par	Superior_Parietal
	99	Area_43_L	Fr	Posterior_Opercular
	100	Area_OP4-PV_L	Par	Posterior_Opercular
	134	Area_TE2_anterior_L	Temp	Lateral_Temporal
	148	Area_PF_Complex_L	Par	Inferior_Parietal
13	1	Primary_Visual_Cortex_L	Occ	Primary_Visual
	6	Fourth_Visual_Area_L	Occ	Early_Visual
	13	Area_V3A_L	Occ	Dorsal_Stream_Visual
	18	Fusiform_Face_Complex_L	Temp	Ventral_Stream_Visual
	22	Posterior_InferoTemporal_complex_L	Occ	Ventral_Stream_Visual
	77	Area_anterior_47r_L	Fr	Inferior_Frontal
	91	Area_11l_L	Fr	Orbital_and_Polar_Frontal
	123	Area_STGa_L	Temp	Auditory_Association
	128	Area_STSd_anterior_L	Temp	Auditory_Association
	135	Area_TF_L	Temp	Medial_Temporal
	169	Area_Frontal_Opercular_5_L	Fr	Insular_and_Frontal_Opercular
	176	Area_STSv_anterior_L	Temp	Auditory_Association
	181	Primary_Visual_Cortex_R	Occ	Primary_Visual
	183	Sixth_Visual_Area_R	Occ	Dorsal_Stream_Visual
	184	Second_Visual_Area_R	Occ	Early_Visual
	185	Third_Visual_Area_R	Occ	Early_Visual
	193	Area_V3A_R	Occ	Dorsal_Stream_Visual
	195	Parieto-Occipital_Sulcus_Area_2_R	Par	Posterior_Cingulate
	197	IntraParietal_Sulcus_Area_1_R	Par	Dorsal_Stream_Visual
	267	Area_9_anterior_R	Fr	Dorsolateral_Prefrontal
	322	Dorsal_Transitional_Visual_Area_R	Par	Posterior_Cingulate
	350	Area_posterior_10p_R	Fr	Orbital_and_Polar_Frontal
199	196	Seventh_Visual_Area_R	Occ	Dorsal_Stream_Visual
	199	Area_V3B_R	Occ	Dorsal_Stream_Visual
	200	Area_Lateral_Occipital_1_R	Occ	MT+_Complex_and_Neighboring_Visual_Areas
	201	Area_Lateral_Occipital_2_R	Occ	MT+_Complex_and_Neighboring_Visual_Areas
	257	Area_anterior_47r_R	Fr	Inferior_Frontal
	307	ParaHippocampal_Area_3_R	Temp	Medial_Temporal
	322	Dorsal_Transitional_Visual_Area_R	Par	Posterior_Cingulate
	332	Area_V6A_R	Par	Dorsal_Stream_Visual
177	74	Area_44_L	Fr	Inferior_Frontal
	84	Area_46_L	Fr	Dorsolateral_Prefrontal
	149	Area_PFm_Complex_L	Par	Inferior_Parietal
	177	Area_TE1_Middle_L	Temp	Lateral_Temporal
355	229	Ventral_IntraParietal_Complex_R	Par	Superior_Parietal
	257	Area_anterior_47r_R	Fr	Inferior_Frontal
	323	Area_PGp_R	Par	Inferior_Parietal
	331	Area_PGs_R	Par	Inferior_Parietal
	355	Auditory_4_Complex_R	Temp	Auditory_Association
	403	Thalamus: Mediodorsomedial magnocellular_R		
	406	Thalamus: Parafascicular_R		
	410	Thalamus: Pulvinar medial_R		
	411	Thalamus: Ventral Anterior_R		
	415	Putamen_R		
	418	Globus palidus externalis_R		

Chapter 6
Conclusion

The studies presented in this dissertation constitute a progressive effort to develop biologically informed models capable of decoding mental states from fMRI data, moving toward the broader objective of a digital representation of brain function. The baseline model that motivated this work achieved high predictive performance but largely ignored the large-scale network organization of the human brain, limiting its neurobiological interpretability. To address this limitation, successive refinements progressively incorporated biologically informed constraints, including the selection of a more appropriate structural connectivity matrix for motor-state classification, the validation on the more functionally distributed social cognition paradigm with explainable artificial intelligence approaches, and the expansion of regional representations to include subcortical structures. As Wang et al. (2024) stated, a digital twin of the brain should be "as simple as possible, but as complex as necessary". Consequently, the model developed here seeks to approximate the organization and functioning of the human brain as closely as possible, attempting to obtain results as if it were the biological structure itself, but maintaining sufficient simplicity to allow for its interpretation, maintenance, and reliability.

The first study demonstrated that the anatomy-driven deterministic matrix yielded the best classification results when compared with the population-based deterministic and the probabilistic models. Although region-based classifiers remained superior in predictive accuracy, network-level approaches provided a more biologically meaningful representation of brain function by incorporating anatomical context and large-scale interactions between regions. The results showed that increased structural connectivity sparsity improved classification performance, suggesting that simplicity is preferred as long as the strongest or most reliable anatomical connections are preserved. Importantly, these results supported the use of anatomically informed network models to understand the distributed organization of brain function. While this framework was evaluated using motor tasks, whose neural substrates are relatively localized, questions arise on whether social paradigms that depend on large-scale brain integration would benefit more from this distributed analysis.

Changing to a cognitive paradigm, characterized by distributed large-scale processing, network-level modeling, did not substantially improved classification accuracies of the region-level approaches. In fact, both frameworks achieved comparable predictive performances after pruning and retraining. Nevertheless, the network-based approach provided a richer neurobiological characterization of the underlying cognitive processes, identifying structurally grounded subnetworks that closely correspond to established Theory of Mind circuitry. These findings suggest that the principal advantage of network-level modeling lies not in improved classification performance, but in its ability to capture the coordinated activity of anatomically connected systems that support social cognition. However, the model had limitations regarding the representation of anatomical structures, since only cortical regions were present in the atlas used and the connectivity matrices excluded interhemispheric connections. As long as brain regions are excluded from the model, its biological plausibility is contestable, which motivated the subsequent

expansion to a more complete representation of the structure and organization of the brain.

The inclusion of subcortical regions and interhemispheric connections increased the biological completeness of the proposed framework without producing substantial improvements in classification performance. Across both motor and social paradigms, the extended atlas achieved accuracies comparable to those obtained with the original cortical atlas, indicating that the additional anatomical information does not provide a clear predictive advantage for the tasks investigated. Nevertheless, the expanded connectivity representation enabled the identification of larger and more distributed subnetworks, particularly in the Theory of Mind paradigm, where relevant features spanned both hemispheres and incorporated several subcortical structures. Furthermore, SHAP analyses highlighted cortical and subcortical regions that are strongly supported by the literature, reinforcing the neurobiological validity of the identified networks. These findings suggest that, while the incorporation of subcortical and interhemispheric connectivity increases model complexity without markedly improving classification accuracy, it provides a more realistic representation of brain organization. This may have a beneficial impact in cognitive paradigms that rely more heavily on cortico-subcortical interactions.

The differences in accuracy observed between the models cannot be interpreted as definitive improvements or deteriorations in predictive performance, given the absence of cross-validation, which limits the ability to generalize the results. In this context, the observed test performance may be influenced by the specific characteristics of the selected train-test partition and may not reflect a stable pattern across different data splits. Nevertheless, while robust conclusions regarding relative model performance cannot be established, the results suggest that progressively more complex and biologically informed models provide increasingly comprehensive representations of brain structure and contribute to the identification of regions involved in well-characterized cognitive processes. This progression supports their potential value as tools for approximating brain organization.

Ultimately, these insights indicate that incorporating anatomical constraints into an interpretable decoding framework does not inherently compromise predictive accuracy. Instead, this approach allows the primary explanatory focus to transition from isolated, individual structures to systems distributed across the whole brain and anatomically grounded, such as the motor and ToM networks.

Overall, the studies presented in this work collectively represent a continuous effort to enhance a decoder of brain activity patterns derived from fMRI data. However, this complex organ is unique for every human being, as gene expressions and environment have an impact on the differential maturation of this structure. For this reason, a machine replica of the brain may never fully capture the intricate variations that may alter communication patterns and cognitive processes.

The results obtained throughout this dissertation demonstrate that incorporating structural connectivity information into machine learning frameworks can improve the neurobiological interpretability of brain-state decoding without hampering predictive performance. Specifically, the findings successfully address the formulated research questions across three main dimensions. First, they confirm that introducing biologically informed structural constraints enables a seamless transition from region-level to

network-level decoding of motor tasks, forcing the model to rely strictly on existing anatomical pathways without sacrificing accuracy. Second, this network-level representation proves effectiveness in capturing the distributed patterns of activation restricted by anatomical constraints, underlying complex social cognitive processes, such as Theory of Mind. Finally, the inclusion of subcortical regions and interhemispheric connections demonstrated a significantly more comprehensive and anatomically grounded representation of the brain networks involved in both motor and social tasks, outperforming more restrictive cortical-only or region-level alternatives. The close correspondence between the networks highlighted by the model and those consistently reported in neuroscience literature suggests that the learned representations reflect meaningful functional organization rather than purely statistical patterns, supporting the use of structurally informed approaches for studying large-scale brain function.

Although the present work focuses primarily on methodological development and validation, the resulting models may have important applications in both neuroscience research and clinical practice.

One immediate application lies in the study of neurological and psychiatric disorders associated with alterations in brain connectivity, including Alzheimer's disease, schizophrenia, and other neurodegenerative or neurodevelopmental conditions. By comparing models derived from healthy populations with those constructed from patient datasets, it may become possible to identify how disruptions in structural networks affect cognitive performance and behavior. Such approaches could contribute to a better understanding of disease mechanisms and the neural substrates underlying specific symptoms, establishing relevant biomarkers for early identification and intervention.

Beyond the characterization of pathological processes, biologically informed network models may provide a platform for investigating the relationship between brain structure and function. By identifying the anatomical pathways and functional interactions most relevant to a given cognitive process, these models could contribute to the construction of increasingly detailed maps linking distributed brain networks to specific behavioral functions. In this context, explainable artificial intelligence techniques may play a central role by revealing the neural features that drive model predictions and highlighting previously unrecognized patterns of organization.

As these models become increasingly accurate and individualized, they may support the development of subject-specific computational representations of the brain. Such models could eventually be used to simulate the effects of structural lesions, disease progression, pharmacological interventions, or rehabilitation strategies before their application in real-world settings. The ability to perform virtual experiments in a computational environment may facilitate hypothesis testing while reducing costs, risks, and dependence on animal models.

From a broader perspective, these developments constitute a step towards the long-term goal of creating digital twins of the human brain: computational systems capable of reproducing aspects of an individual's structural organization, functional dynamics, and cognitive processes. While such a goal remains far beyond the current state of the art, advances as the ones achieved with the current work continue to move the field towards increasingly realistic representations of brain function.

The implications of such technologies extend beyond neuroscience and medicine. Highly accurate computational models of cognition could eventually contribute to the

development of more adaptive human-machine interfaces and personalized healthcare systems. However, these possibilities also raise important ethical, legal, and societal questions regarding privacy, autonomy, accountability, and the limits of cognitive modelling. Addressing these challenges will be essential as future generations of brain-inspired computational systems become increasingly sophisticated.

Although the present dissertation represents only a modest step within this broader scientific endeavor, it highlights the potential of combining anatomical constraints, machine learning, and explainable artificial intelligence to produce models that are both predictive and biologically meaningful.

6.1. Limitations and Future Perspectives

The results present in this dissertation demonstrate the potential of structurally informed neural network models to decode motor and social cognitive processes while providing biologically interpretable representations of brain activity. However, several limitations remain that should be taken into consideration towards the development of more comprehensive and neurobiologically plausible approaches for studying large-scale brain function.

A major limitation of the present study is that model performance was evaluated using a single train-test partition, without cross validation. As a result, the observed differences between model accuracies may not necessarily reflect robust improvements in predictive performance but rather effects specific to the particular data split used, thereby limiting the generalizability of the findings.

Cross-validation would allow a more reliable assessment of model stability across different data partitions, reducing the dependence on a single split and providing a more robust estimate of general performance. Since the feature extraction process is independent of the training data and does not introduce data leakage, the application of cross-validation is a valid and important next step. Under such a framework, only consistently higher performance across multiple partitions can be interpreted as evidence of genuine model improvement.

Additionally, the present work uses a relatively small sample size of 30 subjects. Given the substantial inter-individual variability of functional brain organization and the inherent noise present in fMRI measurements, a limited sample may not adequately capture the full range of functional patterns, potentially restricting the generalizability of the proposed framework. Furthermore, small datasets increase the risk of overfitting, particularly in machine learning applications involving high-dimensional neuroimaging data. For this reason, increasing the dataset is an important future direction.

Another promising direction involves the integration of complementary neuroimaging modalities. Functional MRI offers high spatial resolution but is constrained by relatively low temporal resolution. Additionally, the highly controlled acquisition environments characteristic of this technique limits its application in interactions between individuals or complex movements, such as walking, studies. Functional near-infrared spectroscopy (fNIRS), in contrast, enables more naturalistic experimental settings, higher sampling frequencies, and broader accessibility while relying on a physiological principle similar to the BOLD signal, namely hemodynamic changes associated with neural activity. Combining fMRI and fNIRS data could therefore provide complementary information, potentially improving both model robustness and ecological validity.

Finally, although the inclusion of subcortical structures increased the biological completeness of the framework, it did not improve performance in either the motor or social cognition paradigms investigated. This finding may indicate that these tasks rely predominantly on cortical processing or that the selected paradigms do not strongly engage cortico-subcortical interactions. Future studies should therefore evaluate the framework using cognitive paradigms that are more explicitly dependent on subcortical circuitry.

A candidate example is the brain's reward system which relies on subcortical structures like the striatum, globus pallidus and cerebellum (Pierce & Péron, 2020). Including the HCP Gambling task in the analysis would therefore constitute an important validation step to exploit the present model. Additionally, evaluating the model under the working memory HCP paradigm would provide a more stringent assessment of the efficacy of the current framework, as activity in subcortical structures is able to decode most of the working memory cognitive tasks (Nakai & Nishimoto, 2022).

Taken together, these limitations highlight the need for larger and more diverse datasets, more robust validation strategies, and the exploration of additional paradigms that better engage distributed cortico-subcortical networks. Addressing these aspects will be essential to further improve the generalizability and neurobiological relevance of the proposed framework. Nevertheless, the progress achieved with this work demonstrates the model's ability to correctly perform the tasks to which it was subjected, and the possibility of expanding it to increasingly demanding and complex contexts.

Supplementary Materials

Due to the high dimensionality and scale of the generated figures and tables, presenting the complete tabular results within the main body of this dissertation would severely compromise text readability.

To ensure full academic transparency and support reproducibility, the repository of extended results including complete SHAP, Cohen's d and regions within networks analysis as well as the connectivity matrices have been made publicly available. These resources can be accessed and inspected at the following link: https://github.com/MariaBFCBRamos/MMwM-Supplementary_Materials.git

Bibliography

Amunts, K., Axer, M., Banerjee, S., Bitsch, L., Bjaalie, J. G., Brauner, P., Brovelli, A., Calarco, N., Carrere, M., & Caspers, S. (2024). The coming decade of digital brain research: A vision for neuroscience at the intersection of technology and computing. *Imaging Neuroscience*, *2*, imag–2–00137.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., & Benjamins, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Axer, H., Klingner, C. M., & Prescher, A. (2013). Fiber anatomy of dorsal and ventral language streams. *Brain and language*, *127*(2), 192–204.

Baker, C. M., Burks, J. D., Briggs, R. G., Conner, A. K., Glenn, C. A., Sali, G., McCoy, T. M., Battiste, J. D., O'Donoghue, D. L., & Sughrue, M. E. (2018). A connectomic atlas of the human cerebrum—Chapter 1: Introduction, methods, and significance. *Operative Neurosurgery*, *15*(suppl_1), S1–S9.

Balgova, E., Diveica, V., Jackson, R. L., & Binney, R. J. (2024). Overlapping neural correlates underpin theory of mind and semantic cognition: Evidence from a meta-analysis of 344 functional neuroimaging studies. *Neuropsychologia*, *200*, 108904.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic resonance in medicine*, *34*(4), 537–541.

Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Frontiers in neuroscience*, *11*, 543.

Cabrera-Álvarez, J., del Cerro-León, A., Carvajal, B. P., Carrasco-Gómez, M., Alexandersen, C. G., Bruña, R., Maestú, F., & Susi, G. (2025). The fluctuations of alpha power: bimodalities, connectivity, and neural mass models. *Imaging Neuroscience*, *3*, IMAG. a. 64.

Chen, J. E., & Glover, G. H. (2015). Functional magnetic resonance imaging methods. *Neuropsychology review*, *25*(3), 289–313.

Cui, H., Dai, W., Zhu, Y., Kan, X., Gu, A. A. C., Lukemire, J., Zhan, L., He, L., Guo, Y., & Yang, C. (2022). Braingb: a benchmark for brain network analysis with graph neural networks. *IEEE transactions on medical imaging*, *42*(2), 493–506.

de Figueiredo, E. H., Borgonovi, A. F., & Doring, T. M. (2011). Basic concepts of MR imaging, diffusion MR imaging, and diffusion tensor imaging. *Magnetic Resonance Imaging Clinics*, *19*(1), 1–22.

Dong, Q., Welsh, R. C., Chenevert, T. L., Carlos, R. C., Maly-Sundgren, P., Gomez-Hassan, D. M., & Mukherji, S. K. (2004). Clinical applications of diffusion tensor imaging. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *19*(1), 6–18.

Duffau, H. (2018). The error of Broca: From the traditional localizationist concept to a connectomal anatomy of human brain. *Journal of chemical neuroanatomy*, *89*, 73–81.

Dureux, A., Zanini, A., Selvanayagam, J., Menon, R. S., & Everling, S. (2023). Gaze patterns and brain activations in humans and marmosets in the Frith-Happé theory-of-mind animation task. *Elife*, *12*, e86327.

Elam, J. S., Glasser, M. F., Harms, M. P., Sotiropoulos, S. N., Andersson, J. L., Burgess, G. C., Curtiss, S. W., Oostenveld, R., Larson-Prior, L. J., & Schoffelen, J.-M. (2021). The human connectome project: a retrospective. *NeuroImage*, *244*, 118543.

Fan, X., & Markram, H. (2019). A Brief History of Simulation Neuroscience [Review]. *Frontiers in Neuroinformatics*, *Volume 13 - 2019*. https://doi.org/10.3389/fninf.2019.00032

Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature reviews neuroscience*, *8*(9), 700–711.

Friston, K. J. (2011). Functional and effective connectivity: a review. *Brain connectivity*, *1*(1), 13–36.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534.

Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*, *9*, 144352–144360.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., & Jenkinson, M. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1–42.

Hagmann, P., Jonasson, L., Maeder, P., Thiran, J.-P., Wedeen, V. J., & Meuli, R. (2006). Understanding diffusion MR imaging techniques: from scalar diffusion-weighted imaging to diffusion tensor imaging and beyond. *Radiographics*, *26*(suppl_1), S205–S223.

Han, S.-H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial neural network: understanding the basic concepts without mathematics. *Dementia and neurocognitive disorders*, *17*(3), 83–89.

Hardwick, R. M., Rottschy, C., Miall, R. C., & Eickhoff, S. B. (2013). A quantitative meta-analysis and review of motor learning in the human brain. *NeuroImage*, *67*, 283–297.

Huang, C.-C., Rolls, E. T., Feng, J., & Lin, C.-P. (2022). An extended Human Connectome Project multimodal parcellation atlas of the human cortex and subcortical areas. *Brain Structure and Function*, *227*(3), 763–778.

Huang, W., Bolton, T. A., Medaglia, J. D., Bassett, D. S., Ribeiro, A., & Van De Ville, D. (2018). A graph signal processing perspective on functional brain imaging. *Proceedings of the IEEE*, *106*(5), 868–885.

Kanske, P., Böckler, A., Trautwein, F.-M., & Singer, T. (2015). Dissecting the social brain: Introducing the EmpaToM to reveal distinct neural networks and brain–behavior relations for empathy and Theory of Mind. *NeuroImage*, *122*, 6–19.

Kierońska-Siwak, S., Rudaś, M., Meder, G., Baumgart, M., Grzonkowska, M., & Grzanka, D. (2026). Deterministic or probabilistic tractography? Choosing the appropriate approach in clinical neurosurgery and neuroradiology practice. *Polish Journal of Radiology*, *91*, e187.

Koshiyama, D., Fukunaga, M., Okada, N., Yamashita, F., Yamamori, H., Yasuda, Y., Fujimoto, M., Ohi, K., Fujino, H., & Watanabe, Y. (2018). Role of subcortical structures on cognitive and social function in schizophrenia. *Scientific reports*, *8*(1), 1183.

Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology*, *26*(2), 195–197.

Lacourse, M. G., Orr, E. L., Cramer, S. C., & Cohen, M. J. (2005). Brain activation during execution and motor imagery of novel and skilled sequential hand movements. *NeuroImage*, *27*(3), 505–519.

Lavoie, M.-A., Vistoli, D., Sutliff, S., Jackson, P. L., & Achim, A. M. (2016). Social representations and contextual adjustments as two distinct components of the

Theory of Mind brain network: Evidence from the REMICS task. *Cortex*, *81*, 176–191.

Li, X., Zhou, Y., Dvornek, N., Zhang, M., Gao, S., Zhuang, J., Scheinost, D., Staib, L. H., Ventola, P., & Duncan, J. S. (2021). Braingnn: Interpretable brain graph neural network for fmri analysis. *Medical image analysis*, *74*, 102233.

Marques do Santos, J. D. (2024). *Explainable Artificial Intelligence for Brain Functional Magnetic Resonance Imaging Analysis* [Dissertation, Faculdade de Engenharia da Universidade do Porto]. Repositório Aberto da Universidade do Porto. https://hdl.handle.net/10216/160965

Marques dos Santos, J. D., & Marques dos Santos, J. P. (2024). Path-Weight-Based Pruning and SHAP-Based Explanations of an ANN with fMRI Data. International Conference on Machine Learning, Optimization, and Data Science,

Marques dos Santos, J. D., Reis, L. P., & Marques dos Santos, J. P. (2025). Decoding Mental States in Social Cognition: Insights from Explainable Artificial Intelligence on HCP fMRI Data. *Machine Learning and Knowledge Extraction*, *7*(1), 17.

Matthews, P. M., & Jezzard, P. (2004). Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, *75*(1), 6–12.

Moeller, F., LeVan, P., & Gotman, J. (2011). Independent component analysis (ICA) of generalized spike wave discharges in fMRI: Comparison with general linear model-based EEG-fMRI. *Human brain mapping*, *32*(2), 209–217.

Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In *Multivariate statistical machine learning methods for genomic prediction* (pp. 379–425). Springer.

Nakai, T., & Nishimoto, S. (2022). Representations and decodability of diverse cognitive functions are preserved across the human cortex, cerebellum, and subcortex. *Communications Biology*, *5*(1), 1245.

Nambu, A. (2007). Globus pallidus internal segment. *Progress in brain research*, *160*, 135–150.

Ogawa, K., & Matsuyama, Y. (2023). Heterogeneity of social cognition between visual perspective-taking and theory of mind in the temporo-parietal junction. *Neuroscience letters*, *807*, 137267.

Okada, N., Fukunaga, M., Miura, K., Nemoto, K., Matsumoto, J., Hashimoto, N., Kiyota, M., Morita, K., Koshiyama, D., & Ohi, K. (2023). Subcortical volumetric alterations in four major psychiatric disorders: a mega-analysis study of 5604 subjects and a volumetric data-driven approach for classification. *Molecular Psychiatry*, *28*(12), 5206–5216.

Otsuka, Y., Osaka, N., Ikeda, T., & Osaka, M. (2009). Individual differences in the theory of mind and superior temporal sulcus. *Neuroscience letters*, *463*(2), 150–153.

Penfield, W., & Rasmussen, T. (1950). The cerebral cortex of man; a clinical study of localization of function.

Pierce, J. E., & Péron, J. (2020). The basal ganglia and the cerebellum in human emotion. *Social cognitive and affective neuroscience*, *15*(5), 599–613.

Qamar, R., & Zardari, B. A. (2023). Artificial neural networks: An overview. *Mesopotamian Journal of Computer Science*, *2023*, 124–133.

Qin, X., Huang, H., Liu, Y., Zheng, F., Zhou, Y., & Wang, H. (2023). Increased functional connectivity involving the parahippocampal gyrus in patients with schizophrenia during Theory of Mind processing: A psychophysiological interaction study. *Brain Sciences*, *13*(4), 692.

Rosen, B. Q., & Halgren, E. (2021). A whole-cortex probabilistic diffusion tractography connectome. *eneuro*, *8*(1).

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.*, *55*(1), 87–124.

Schulte, J., Senden, M., Deco, G., Kobeleva, X., & Zamora-López, G. (2025). The global communication pathways of the human brain transcend the cortical-subcortical-cerebellar division. *arXiv preprint arXiv:2505.22893*.

Schulte, T., & Müller-Oehring, E. M. (2010). Contribution of callosal connections to the interhemispheric integration of visuomotor and cognitive processes. *Neuropsychology review*, *20*(2), 174–190.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34.

Shuai, Y., Chandio, B. Q., Feng, Y., Villalon-Reina, J. E., Nir, T. M., Ching, C. R., Gari, I. B., Thomopoulos, S. I., Alibrando, J. D., & John, J. P. (2025). Deterministic versus probabilistic tractography: Impact on white matter bundle shape. 2025 21st International Symposium on Biomedical Image Processing and Analysis (SIPAIM),

Soisoonthorn, T., & Unger, H. (2025). The Human Brain. In *A Brain-Inspired Approach to Natural Language Processing* (pp. 1–30). Springer.

Sporns, O. (2016). *Networks of the Brain*. MIT press.

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., & Okubo, Y. (2004). Brain activation associated with evaluative processes of guilt and embarrassment: an fMRI study. *NeuroImage*, *23*(3), 967–974.

Tournier, J.-D., Mori, S., & Leemans, A. (2011). Diffusion tensor imaging and beyond. *Magnetic resonance in medicine*, *65*(6), 1532.

van der Knaap, L. J., & van der Ham, I. J. (2011). How does the corpus callosum mediate interhemispheric transfer? A review. *Behavioural brain research*, *223*(1), 211–221.

Van Essen, D. C., & Glasser, M. F. (2016). The human connectome project: Progress and prospects. Cerebrum: the Dana forum on brain science,

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

Van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human brain mapping*, *30*(3), 829–858.

Vecchio, F., Miraglia, F., & Rossini, P. M. (2017). Connectome: Graph theory application in functional brain network architecture. *Clinical neurophysiology practice*, *2*, 206–213.

Voineskos, A. N., Hawco, C., Neufeld, N. H., Turner, J. A., Ameis, S. H., Anticevic, A., Buchanan, R. W., Cadenhead, K., Dazzan, P., & Dickie, E. W. (2024). Functional magnetic resonance imaging in schizophrenia: current evidence, methodological advances, limitations and future directions. *World Psychiatry*, *23*(1), 26–51.

Wang, H. E., Triebkorn, P., Breyton, M., Dollomaja, B., Lemarechal, J.-D., Petkoski, S., Sorrentino, P., Depannemaecker, D., Hashemi, M., & Jirsa, V. K. (2024). Virtual brain twins: from basic neuroscience to clinical use. *National Science Review*, *11*(5), nwae079.

Wei, P., Bao, R., Lv, Z., & Jing, B. (2018). Weak but critical links between primary somatosensory centers and motor cortex during movement. *Frontiers in human neuroscience*, *12*, 1.

Wu, Y.-c., & Feng, J.-w. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, *102*(2), 1645–1656.

Yeh, F.-C. (2022). Population-based tract-to-region connectome of the human brain and its hierarchical topology. *Nature communications*, *13*(1), 4933.

Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, *48*(9), 2658–2664.