Prediction of Task-fMRI Activation Maps and Cognitive Traits From Resting State-fMRI To Enhance Clinical Applicability Of fMRI

BTP Thesis submitted to the

Indian Institute of Technology Bhubaneswar

by

Bathula Satwik

Under the guidance of

Dr. Himanshu Padole



SCHOOL OF ELECTRICAL AND COMPUTER SCIENCES

INDIAN INSTITUTE OF TECHNOLOGY BHUBANESWAR

May 2025

©2025 Bathula Satwik. All rights reserved.

CERTIFICATE

This is to certify that the thesis entitled **Prediction of task-fMRI activation maps and Cognitive Traits from resting state-fMRI to enhance clinical applicability of fMRI**, submitted by **Bathula Satwik** to Indian Institute of Technology Bhubaneswar, is a record of bonafide research work under my supervision and I consider it worthy of consideration for the award of the degree of Bachelor of Technology of the Institute.

Date :

Dr. Himanshu Pramod Padole
Assistant Professor
School of Electrical and Computer Sciences
Indian Institute of Technology Bhubaneswar
Bhubaneswar, India

DECLARATION

I certify that

a. The work contained in the thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other institute for any degree or diploma.

c. I have followed the guidelines provided by the institute in writing the thesis.

d. I have conformed to the norms and guidelines given in the ethical code of conduct of the institute.

e. Whenever I have used materials (data, theoretical analysis, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

f. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.



Bathula Satwik

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Himanshu Pramod Padole, for his constant guidance, encouragement, and valuable insights throughout the course of this thesis. Their mentorship has been instrumental in shaping both the direction and quality of this work.

I am sincerely thankful to the faculty and staff of the Department of Electronics and Communication engineering, Indian Institute of Technology Bhubaneswar, for providing a supportive academic environment and access to essential resources.

I would also like to acknowledge my lab mates and fellow researchers, especially Sasideep Pasumarthi, for the stimulating discussions and the collaborative spirit that greatly enriched this work.

Finally, I am profoundly grateful to my family and friends for their unwavering support, patience, and encouragement during the course of my academic journey. Their belief in me has been a constant source of strength.

Bathula Satwik

Abstract

Functional Magnetic Resonance Imaging (fMRI) is a cornerstone of neuroimaging, widely used in diagnosing neurological disorders, localizing brain regions for surgery, and assessing cognitive traits such as intelligence and brain age. However, the dependency on task-based fMRI (task-fMRI) limits its clinical applicability, particularly for noncompliant populations like infants and individuals with disabilities. This study introduces a novel framework to predict task activation maps and cognitive traits using resting-state fMRI (rs-fMRI), eliminating the need for task-based data acquisition.

The proposed framework integrates three advanced methodologies: **univariate time series regression**, which is employed for task activation map prediction by modeling voxel-level temporal dynamics; **multivariate time series regression**, applied to predict cognitive traits by capturing complex temporal interactions across multiple voxels. Since the use of fully connected graphs is impractical for graph-based learning due to computational inefficiency and redundancy, a novel **Enforced Graph Information Bottleneck** method is proposed for graph pruning, designed to preserve essential information relevant for regression tasks.

The framework addresses critical challenges in fMRI applications by increasing accessibility, accurately predicting task activation maps, and enabling cognitive trait estimation with minimal human intervention. These advancements aim to revolutionize the clinical and research utility of fMRI, making it a more accessible and versatile neuroimaging tool.

Contents

Certification of Approval i

Certificate i

Declaration ii

Acknowledgments iii

Abstract iv

1 Introduction 1

1.1 Background 1

1.2 Challenges 2

1.3 Literature Gap 3

1.4 Motivation 4

1.5 Objectives 7

2 Literature Review 8

2.1 IQ Score Prediction 8

2.2 Task Activation Map Prediction 10

2.3 Graph Pruning 11

3 Dataset Description **14**

3.1 ABIDE Dataset 14

3.2 HCP Dataset 15

3.2.1 Selected Task Contrasts and Their Functional Relevance 17

3.2.2 Parcellation Methodologies 18

4 Implementation and Methodology **20**

4.1 IQ Prediction 20

4.1.1 HYDRA: A Dictionary-Based Time Series Transformation Framework 20

4.1.2 MTSR-based IQ Prediction Model Using HYDRA 23

4.1.3 Similarity-Based IQ Prediction Model 24

4.1.4 Hybrid IQ Prediction Model 28

4.2 Task Activation Map Prediction 30

4.2.1 Time Series Feature Extraction 30

4.2.2 Univariate Time Series Regression 31

4.2.3 Performance Metrics Used: 32

4.2.4 List of all Catch22 Features [46]: 35

4.2.5 List of Summary Time Series Features: 38

4.2.6 Diversity-reduced Canonical Interval Forest (DrCIF) 38

4.3 Graph Pruning 41

4.3.1 Background 41

4.3.2 Mutual Information Neural Estimator - MINE 45

4.3.3 Information Bottleneck Theory and its Connection to Rate Distortion Theory . . . 47

4.3.4 A New Variational Principle for Information Bottleneck . . . 49

4.3.5 Graph Pruning using information bottleneck principle . . . 52

4.3.6 Architectural Information . . . 60

4.3.7 Graph Mutual Information Estimator (GraphMINE) Architecture 62

5 Results **64**

5.1 IQ Prediction . . . 64

5.1.1 Performance of Similarity-based CNN Prediction Model . . . 65

5.1.2 Performance of Hybrid IQ Prediction Model . . . 66

5.2 Task Activation Map prediction from rs-fMRI . . . 67

5.2.1 Results . . . 67

5.3 Graph Pruning . . . 100

5.3.1 Results . . . 100

5.3.2 Relational Contrast . . . 100

5.3.3 Social Contrast . . . 102

6 Future Scope **105**

7 Limitations **106**

7.1 IQ Prediction . . . 106

7.2 Task Activation Map prediction . . . 108

7.3 Graph Pruning . . . 110

Dissemination 112

Bibliography **113**

Appendix **117**

Chapter 1

Introduction

1.1 Background

Functional Magnetic Resonance Imaging (fMRI) has revolutionized the field of neuroscience and clinical diagnostics, offering unparalleled insights into brain activity and function. As a non-invasive and high-resolution imaging modality, fMRI has become indispensable for understanding neural mechanisms, localizing functional brain regions, and studying various cognitive and psychological traits. Its clinical applications extend from early diagnosis of neurological disorders, such as Alzheimer's disease, schizophrenia, and autism spectrum disorder, to precise brain region localization for surgical planning. Beyond clinical settings, fMRI has been pivotal in cognitive neuroscience, enabling the prediction of individual traits such as intelligence, brain age, and language abilities.

Despite these advancements, the utility of fMRI is significantly hindered by its reliance on task-based acquisitions. Task-based fMRI (task-fMRI) requires subjects to perform specific, predefined tasks inside an MRI scanner to evoke neural activity, which is then captured as activation maps. While this provides valuable information about brain functionality, it introduces several challenges. The process is often time-consuming, expensive, and dependent on precise task execution under expert su-

pervision. Moreover, designing optimal tasks to activate specific brain regions remains a complex and labor-intensive endeavor. These constraints are further amplified when dealing with noncompliant populations, such as infants, elderly individuals, or patients with physical or cognitive impairments, for whom task performance may not be feasible.

Resting-state fMRI (rs-fMRI), which measures brain activity while the subject is at rest, offers an alternative. Unlike task-fMRI, rs-fMRI does not require task execution, making it more practical and accessible for broader populations. Leveraging rs-fMRI to predict task activation maps and cognitive traits can address the limitations of task-fMRI, opening new possibilities for its application in clinical and research settings. This shift in approach has the potential to enhance the accessibility and applicability of fMRI, transforming it into a more inclusive and efficient neuroimaging modality.

1.2 Challenges

Despite its potential, the widespread adoption of fMRI in clinical and research domains is constrained by several challenges. Task-based fMRI (task-fMRI), which requires subjects to perform predefined tasks during scanning, is inherently complex, time-intensive, and costly. The dependency on task execution introduces significant logistical and methodological challenges, including the need for expert supervision and the difficulty in designing optimal tasks to localize specific brain regions.

These challenges are particularly pronounced when dealing with noncompliant populations such as infants, elderly individuals, and patients with neurological or physical disabilities. For such populations, task performance may be unreliable or entirely infeasible, further limiting the practicality of task-fMRI. Additionally, task-based data acquisition is resource-intensive, both in terms of time and cost, making it less viable for routine clinical use or large-scale studies. Addressing these limitations is essential for improving the accessibility and utility of fMRI in diverse populations and settings.

1.3 Literature Gap

- Traditional methods for IQ assessment typically involve task-based evaluations that are time-consuming, costly, and require expert supervision. To address these challenges, recent efforts have shifted toward utilizing neuroimaging data, particularly resting-state fMRI (rs-fMRI), due to its ability to capture intrinsic brain activity without task engagement.

  While several approaches have been proposed for predicting IQ from rs-fMRI, two critical limitations remain unaddressed. First, existing methods often ignore the multivariate temporal dependencies inherent in rs-fMRI signals, treating each voxel or region independently. Second, global relational information across different brain regions—captured through inter-series dependencies—is frequently overlooked, leading to suboptimal representations of brain connectivity.

  To bridge these gaps, this work proposes a hybrid framework that simultaneously extracts multivariate temporal patterns and relational structures from rs-fMRI signals. By combining a time series regression model with a relation-aware representation, the framework aims to provide a more comprehensive and accurate prediction of IQ scores.

- Despite notable progress in task-activation map prediction using resting-state fMRI (rs-fMRI), several methodological limitations remain. A majority of existing approaches rely heavily on resting-state functional connectivity (RSFC) maps as input features. While these maps capture statistical dependencies between brain regions, they discard the rich temporal dynamics inherent in the original BOLD time series. This results in the loss of transient, frequency-specific, and nonstationary patterns that may be critical for accurately predicting task-related activations. Voxel-based regression models often operate under linear assumptions, limiting their ability to model complex relationships. On the other hand, whole-cortex prediction frameworks, which apply a single regression model across the entire set of grayordinates, fail to account for spatial heterogeneity and localized variations in neural response. Furthermore, most models do not

treat the prediction task as a true time-series regression problem, thus missing the opportunity to leverage advanced univariate time-series modeling techniques. Finally, adaptive or ensemble-based methods that dynamically select the most suitable regressor for each grayordinate are rarely explored, leading to suboptimal performance in region-specific predictions.

- Despite significant advancements in graph simplification and sparsification techniques, most existing graph pruning methods are inherently task-agnostic and rely on heuristic or structural properties such as node degree, centrality, or spectral approximations. While effective in preserving general graph properties like connectivity and clustering, these methods often overlook the relevance of specific substructures to downstream predictive tasks. In contrast, the Information Bottleneck (IB) principle provides a theoretically grounded framework to retain only the most task-informative components of data. Recent applications of the Graph Information Bottleneck (GIB) have demonstrated its utility in learning compact and discriminative representations within Graph Neural Networks (GNNs). However, its potential remains underexplored in the context of graph pruning as a preprocessing tool—independent of any specific learning model. There exists a clear gap in leveraging the GIB framework for pruning the graph structure itself (nodes or edges) in a task-aware, model-agnostic manner. Addressing this gap could lead to principled pruning strategies that balance graph compression with the preservation of task-relevant information, offering a compelling alternative to heuristic-based pruning methods.

1.4 Motivation

- The ability to predict intelligence scores from non-invasive neuroimaging data, such as resting-state fMRI (rs-fMRI), offers transformative potential for personalized cognitive assessment, clinical diagnostics, and neurodevelopmental monitoring. Existing approaches, while demonstrating some predictive success, often rely on static connectivity representations or coarse brain parcellations, which fail

to capture the rich multivariate temporal dependencies and global inter-regional relationships present in rs-fMRI signals. This limits their ability to extract discriminative information that is critical for robust and interpretable intelligence prediction.

Motivated by these shortcomings, this study aims to explore a more comprehensive framework that integrates both temporal and spatial-relational aspects of rs-fMRI data. By formulating the prediction task as a multivariate time series regression problem and simultaneously modeling the relational structure among BOLD signals, the proposed hybrid architecture seeks to capture deeper neural representations that are more aligned with cognitive functioning. Leveraging modern time series modeling techniques and similarity-based graph representations, the approach aims to provide a scalable and generalizable solution for intelligence score prediction, while contributing to a better understanding of the neural correlates of intelligence.

- The prediction of task-activation maps from resting-state fMRI holds significant potential for enhancing the practical applicability of functional neuroimaging by eliminating the need for task compliance, expert supervision, and extensive scanning protocols. However, most existing models rely on static connectivity maps, which fail to capture the complex temporal dynamics embedded in the raw rs-fMRI signals. Capturing these dynamics is essential for identifying transient or frequency-specific neural patterns that are critical for accurate task activation inference. Moreover, the heterogeneity of functional responses across grayordinates necessitates a localized and tailored modeling approach, rather than relying on global regression strategies that overlook regional specificity.

  Motivated by these considerations, this study introduces a novel modeling framework that treats each grayordinate as a distinct predictive unit, utilizing univariate time series regression techniques to fully exploit the temporal structure of rs-fMRI data. This localized modeling strategy enhances interpretability and prediction accuracy. Furthermore, the integration of spatial connectivity with temporal features is pursued to provide a more comprehensive representation of

brain function. By leveraging hybrid feature extraction, dynamic model selection, and ensemble learning, the proposed approach aims to advance the state of the art in task activation map prediction and enable more robust applications in cognitive neuroscience and clinical diagnostics.

- Graph-structured data often contains redundant, noisy, or irrelevant connections that can obscure important patterns and inflate computational costs in downstream tasks. Traditional pruning techniques, while effective at reducing complexity, typically operate without considering the relevance of the retained subgraph to a specific predictive task. To address this, we explore a task-aware graph pruning approach grounded in the Graph Information Bottleneck (GIB) principle. The proposed framework formulates graph pruning as an optimization problem that balances two competing objectives: minimizing the mutual information between the original graph and its pruned counterpart to encourage compression, and maximizing the mutual information between the pruned graph and the task labels to preserve predictive relevance.

  This is operationalized using Mutual Information Neural Estimation (MINE) networks, which provide a scalable and differentiable mechanism to estimate and optimize information-theoretic quantities. A Lagrangian loss formulation incorporates both sparsity and mutual information constraints, enabling the dynamic learning of edge structures that are both compact and task-informative. Unlike existing methods that rely on fixed graph topologies or local heuristics, this approach adaptively refines the graph structure itself based on information-theoretic relevance, offering a principled and flexible framework for interpretable and efficient graph simplification.

1.5 Objectives

The proposed framework aims to:

- Predict task activation maps using resting-state fMRI (rs-fMRI) without relying on task-based acquisitions.
- Estimate cognitive traits such as intelligence scores and brain age with minimal human intervention.
- Develop advanced methodologies, including univariate time series regression, multivariate time series regression, and graph representation learning, to enhance prediction accuracy and model interpretability.
- Initially model the brain as a fully connected graph; however, challenges were encountered in defining an appropriate and tractable connectivity structure. As a result, current efforts focus on formulating a **novel mathematical variational principle inspired by the Information Bottleneck**, designed to **prune the graph by maximizing mutual information** under **dynamic sparsity constraints**.

Chapter 2

Literature Review

2.1 IQ Score Prediction

Lohmann et al. [1], in their innovative graph-based method, employed Ricci-Forman curvature for dimensionality reduction, effectively capturing the underlying topological features of the brain connectome. This approach was particularly significant in reducing the computational complexity associated with high-dimensional fMRI data. They then proposed a voxel-graph machine learning algorithm that modeled the relationships between voxels as a graph structure. This algorithm was applied to intelligence score measurement, achieving a correlation value of 0.37 between the actual scores and the predicted scores. The use of Ricci-Forman curvature highlights the potential of leveraging advanced geometric and graph-based methods for brain data analysis, paving the way for more robust graph-based models in cognitive trait prediction.

Tong et al. [2] advanced the field by employing connectome-based predictive modeling (CPM), which is a data-driven approach to identify brain-behavior relationships. In their method, CPM was utilized to extract key features from the rs-fMRI data, enabling a concise representation of the connectivity patterns that are predictive of intelligence scores. These extracted features were then subjected to a LASSO regressor, a linear model with built-in regularization to prevent overfitting. This approach

yielded an impressive correlation of 0.5573 between the predicted and actual IQ values, marking a significant improvement in prediction accuracy. The success of Tong et al.'s method underscores the effectiveness of combining feature extraction techniques like CPM with regularized regression models for cognitive score prediction.

Similar to the work of Tong et al. [2], Dryburgh et al. [3] also employed connectome-based predictive modeling (CPM) to extract subject-specific summary values from rs-fMRI data. By applying CPM, they derived connectivity features that served as input to a linear regression model for IQ score prediction. Despite using a simpler regression approach compared to Tong et al., their method achieved a correlation coefficient of 0.25. While the predictive performance was modest, their work demonstrated the applicability of CPM in summarizing rs-fMRI data for cognitive assessments, particularly in populations with autism spectrum disorder (ASD), where variability in brain connectivity is significant.

Gal et al. [4] proposed a unique two-step process for cognitive score prediction. First, they derived task activation maps, termed connTask maps, from rs-fMRI data using dimensionality reduction techniques such as independent component analysis (ICA) and principal component analysis (PCA), along with a linear regression model. These connTask maps were then processed using PCA to further reduce dimensionality, and the resulting features were fed into a linear regressor for final score prediction. This approach yielded a correlation value of 0.30 between the predicted and actual scores. The innovation in their work lies in the creation of synthetic task activation maps (connTask maps) from resting-state data, demonstrating an alternative strategy to bridge the gap between task-based and resting-state fMRI modalities.

Martin et al. [5] made significant contributions by leveraging advanced graph-based machine learning models to predict cognitive scores. They utilized rs-fMRI data to construct graph representations of brain connectivity and employed graph neural networks (GNNs) to analyze these graphs. To enhance the training process, they incorporated sample selection learning, a strategy that selectively emphasizes informative samples to improve model generalization. This combination of GNNs and sample selection

learning resulted in a root mean square error (RMSE) of 12.27, showcasing the potential of graph-based deep learning models for cognitive trait prediction. Their work also highlighted the importance of incorporating advanced neural network architectures tailored to the unique challenges of fMRI data, such as its high dimensionality and intricate connectivity patterns.

The aforementioned studies collectively underscore the progress in task activation map prediction and cognitive score estimation using rs-fMRI data. From graph-based methods like Ricci-Forman curvature and voxel-graph models to connectome-based approaches and deep learning frameworks, these methodologies reflect the diverse strategies employed to unlock the predictive potential of resting-state fMRI. Future work in this domain can benefit from integrating these approaches, further refining feature extraction, dimensionality reduction, and predictive modeling to enhance accuracy and clinical applicability.

2.2 Task Activation Map Prediction

Several recent studies have aimed to predict task-activation maps using only resting-state fMRI (rs-fMRI) data, thereby eliminating the need for actual task performance required in conventional task-based fMRI. In one of the early efforts, Tavor et al. [6] formulated the task activation map prediction as a regression problem, where resting-state functional connectivity (RSFC) maps served as predictors and the corresponding task-activation maps were the targets. A linear regression model was trained independently for each voxel, and the average of the learned coefficients was used for prediction. This method achieved a correlation range of 0.13 to 0.80 between the predicted and actual activation maps across different task contrasts. However, the approach was limited by the linearity assumption inherent to the model.

To address this, Cohen et al. [7] proposed a more advanced voxel-based regression framework using feed-forward neural networks and random forests, resulting in im-

proved correlations ranging from 0.54 to 0.804. Similarly, another study [8] applied ridge regression to the voxel-wise prediction task and reported an average correlation of $r = 0.582$ across contrast targets.

Beyond voxel-wise models, whole-cortex regression approaches have also been explored. In these methods, the entire RSFC map is used as input to predict the full activation map in one shot. For instance, Zhang et al. [9] trained a convolutional neural network (CNN) to learn the mapping between RSFC maps and task activation maps. Ellis et al. [10] extended this concept by using structural connectivity maps as predictors in a similar CNN framework. Deep learning-based models such as SurfCNN [11] have achieved an AUC of 0.3 and a DICE coefficient of 0.67. An ensemble learning approach presented in [12] was also used to enhance predictive performance. In the context of schizophrenia, Sheba et al. [13] employed a general linear model to predict individual variability in task activation, reporting a mean correlation of 0.54.

Despite these advances, several key limitations remain. Whole-cortex models apply a single regression model to all gray matter pixels (grayordinates), which overlooks spatial heterogeneity in neural activation patterns. Furthermore, nearly all existing models rely solely on RSFC maps that capture static statistical relationships (e.g., correlations), but omit the rich temporal dynamics present in rs-fMRI signals. This results in the loss of frequency-specific and transient patterns essential for accurately predicting task-induced activations, thereby limiting the expressive power and generalizability of these models.

2.3 Graph Pruning

Graph pruning is a critical technique in reducing the complexity of large-scale graphs while preserving their essential structural or task-relevant properties. Traditional graph pruning methods, such as k-core [14] and k-truss decomposition [15], rely on local connectivity or triangle-based cohesiveness to extract dense substructures, whereas

approaches like minimum spanning trees [16] and disparity filters [17] aim to retain global connectivity or statistically significant edges. These methods, while effective in preserving structural integrity, are typically task-agnostic and may overlook semantic relevance. In the context of deep learning, model compression techniques such as magnitude-based weight pruning [18], filter pruning [19], and dynamic network surgery [20] have been successfully employed to reduce network size and inference time, though they are usually designed for standard neural networks rather than graph structures. More recently, pruning techniques tailored to Graph Neural Networks (GNNs) have emerged, incorporating edge importance scores [21], attention mechanisms [22], and variational masking to selectively retain informative substructures. Notably, information-theoretic frameworks such as the Information Bottleneck [23] [24] and Mutual Information Neural Estimation (MINE) [25] have introduced a principled way to control the trade-off between structural compression and task-relevant information retention. However, the use of these frameworks has been largely confined to GNN training, and their potential for standalone graph pruning—prior to model learning—remains underexplored. This highlights a critical research gap in developing task-aware, model-agnostic graph pruning strategies that leverage mutual information to preserve only the most relevant components of the graph.

Recent advances in graph representation learning have explored the use of variational principles and information-theoretic objectives to guide structural sparsification and improve generalization. Particularly, the Information Bottleneck (IB) principle has inspired several efforts toward disentangling relevant from redundant topological components in graph data. Notable among them is the Variational Graph Information Bottleneck (VGIB) framework proposed by Zhang et al. [26], which encourages compression by minimizing mutual information between the learned graph representation and the input features. The Deep Graph Infomax (DGI) model by Veličković et al. [27] focuses on maximizing mutual information between node and global graph representations to encourage informative embeddings. Extensions like InfoGraph [28] scale this concept to whole-graph representations. On the sparsification side, works such as Graph Information Bottleneck by Wu et al. [29] and Variational Graph Prun-

ing [30] use variational constraints or reinforcement signals to prune redundant edges while preserving label-relevant structure. These ideas also resonate with earlier variational formulations of information constraints from classical literature, particularly in the context of rate-distortion theory [31] and optimal representation learning [24]. Our method builds on this foundation by incorporating a constrained Lagrangian formulation with learnable sparsity and threshold parameters, yielding an efficient, MI-regularized hard-thresholding framework for structure learning.

Chapter 3

Dataset Description

3.1 ABIDE Dataset

For the IQ prediction phase, we performed experimentations on the autism brain imaging data exchange (ABIDE) [32, 33] dataset, a meticulously selected compilation of neuroimaging data. The ABIDE dataset comprises 1,057 participants, consisting of 525 individuals with autism spectrum disorder (ASD) and 532 neurotypical controls, also known as typically developing controls (TDC). IQ scores were provided for each of these participants, although the present study was restricted to the data of the neurotypical controls. 80% of the data was used for training regression models while the remaining was used as test data for performance evaluation. All the experiments were performed using an Intel Xeon Gold processor with 128 GB RAM, and Nvidia RTX A6000 GPU with 48 GB RAM.

The data analysis and preprocessing pipeline for the ABIDE dataset comprised multiple essential stages that were performed using the configurable pipeline for the analysis of connectomes (C-PAC) [34]. Noise and artifacts in the fMRI data were eliminated and corrected using techniques such as global signal regression and band-pass filtering. Global signal regression was employed to eliminate broad, nonspecific fluctuations in signal, thereby enhancing the emphasis on pertinent brain activity.

Figure 3.1: ABIDE

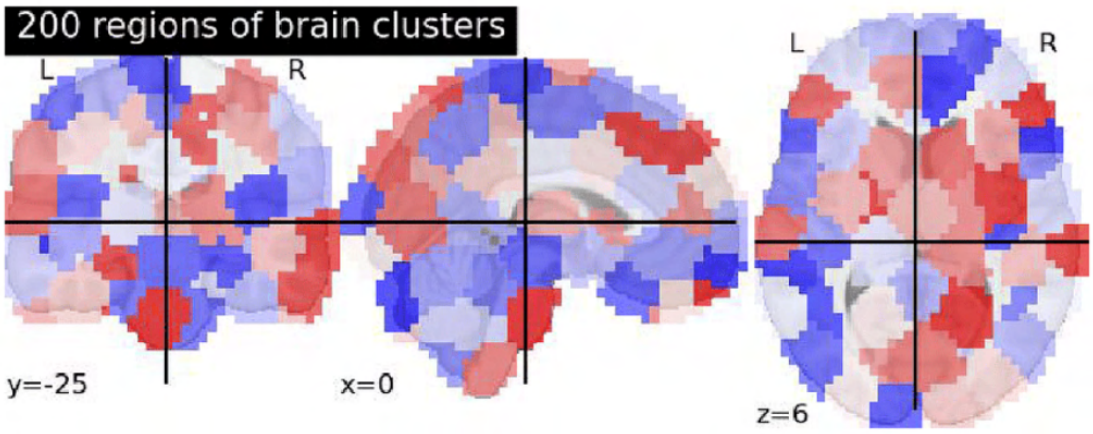


Figure 3.2: Cameron and Craddock 200 atlas.

Band-pass filtering was used to isolate and enhance specific frequency bands. Finally, the Cameron and Craddock 200 (CC200) atlas [35,36] was employed to extract regions of interest (ROI), as shown in Fig 3.2, offering a systematic method for analyzing brain connectivity and activity.

3.2 HCP Dataset

The study of prediction of task activation maps from rs-fMRI utilizes the **Human Connectome Project (HCP)** dataset, specifically the **WU–Minn HCP Data**,

which consists of neuroimaging data from a cohort of 1,200 subjects [37]. For this research, the analysis focuses on the **S900 New Release**, a subset of the HCP dataset, comprising 350 subjects. This dataset is known for its high-quality imaging data and comprehensive coverage of functional and structural connectivity.

The HCP dataset includes task-based functional Magnetic Resonance Imaging (fMRI) data collected during the performance of seven tasks, each designed to probe distinct cognitive and behavioral domains [38]. Each task comprises a set of contrasts to evaluate specific brain functions, detailed as follows:

1. **Working Memory:** 30 contrasts
2. **Gambling:** 6 contrasts
3. **Motor:** 26 contrasts
4. **Language:** 6 contrasts
5. **Social Cognition:** 6 contrasts
6. **Relational Processing:** 6 contrasts
7. **Emotion Processing:** 6 contrasts

The preprocessing of this dataset was performed using the HCP minimal preprocessing pipelines, which include essential steps such as spatial normalization, ICA-based denoising, and cortical surface extraction [39]. These pipelines ensure that the data is of high quality and suitable for advanced analyses, such as task activation map prediction and cognitive trait modeling.

The inclusion of both task-based and resting-state fMRI data in the HCP dataset allows for a detailed exploration of brain connectivity and activity across different states. This makes the dataset particularly well-suited for examining the relationships between resting-state dynamics and task-induced brain activity [38,39]. Such data are invaluable for advancing our understanding of functional connectivity and individual differences in cognitive traits.

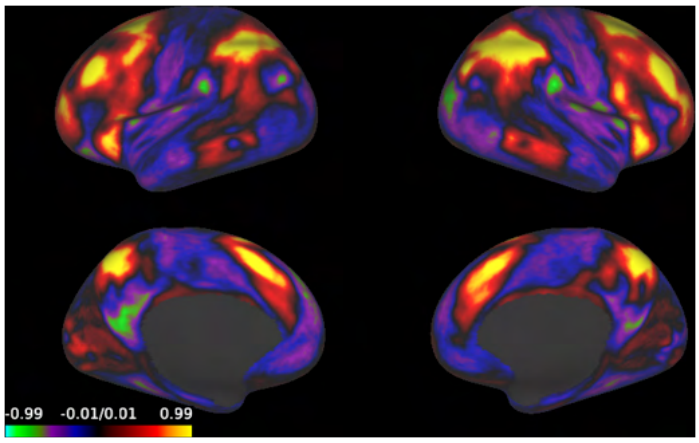


Figure 3.3: Average Task fMRI

3.2.1 Selected Task Contrasts and Their Functional Relevance

From the Human Connectome Project (HCP) dataset, a total of 10 contrasts were selected from the available 86 contrasts across 7 cognitive tasks. These selected contrasts were chosen to represent a diverse set of brain functions, including emotion processing, language comprehension, reasoning, reward-based decision making, social cognition, working memory, and motor control. Specifically, the selected contrasts include: **Emotion** (cope1: FACES), highlighting brain activation in response to viewing emotional facial expressions; **Language** (cope1: MATH), capturing regions involved in arithmetic processing; **Relational** (copes 1 & 3: MATCH, MATCH-REL), identifying activity during a relational matching task and contrasting it with abstract reasoning; **Gambling** (cope3: REWARD-PUNISH), emphasizing areas activated more during reward compared to punishment; **Social** (cope2: TOM), revealing activation related to theory-of-mind tasks involving interpretation of intentional movement; **Working Memory** (cope3: 2BK_PLACE), which isolates activation associated with a 2-back place-based working memory task; and **Motor** (copes 5: RH), representing right hand motor activity. Together, these contrasts span a wide spectrum of cognitive and sensorimotor functions, allowing for robust evaluation of task activation prediction models across multiple domains of brain function.

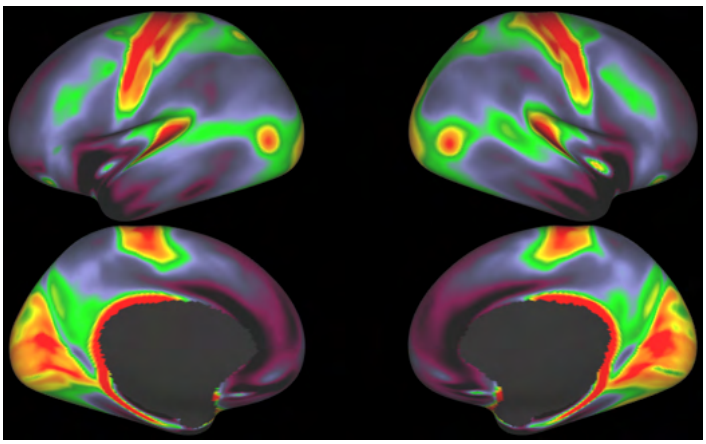


Figure 3.4: Group Average Maps

3.2.2 Parcellation Methodologies

For parcellation, we employed:

- **Multimodal Parcellation:** This method integrates structural and functional imaging modalities, enabling accurate identification of cortical regions with shared connectivity and functional properties [40].
- **Independent Component Analysis (ICA):** ICA was applied to decompose fMRI signals into spatially independent components, facilitating the identification of brain networks relevant to the selected task contrasts.

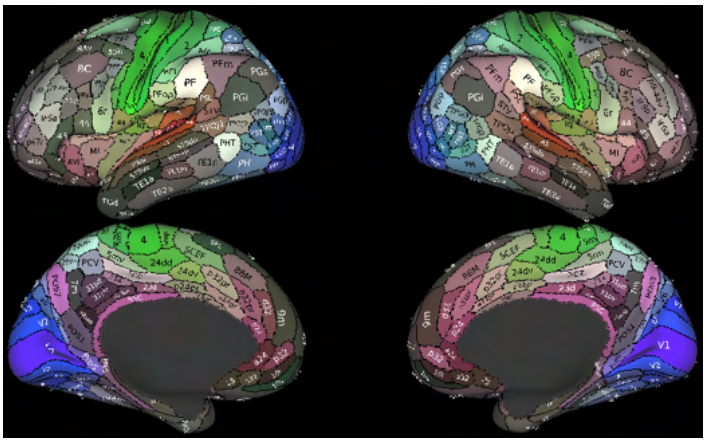


Figure 3.5: Cortical Parcellations

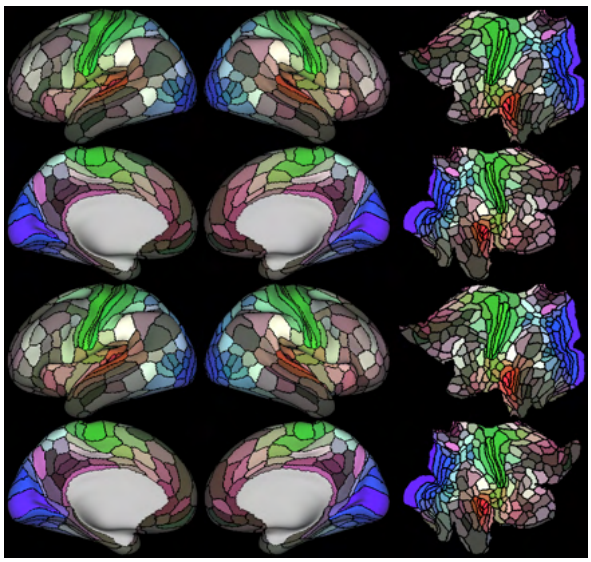


Figure 3.6: Multimodal Parcellation

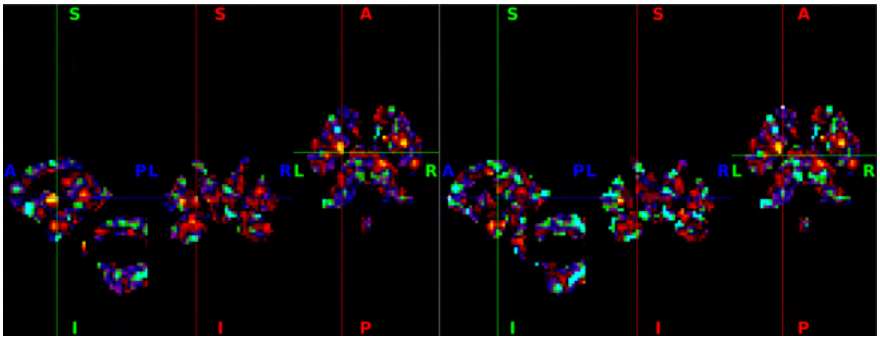


Figure 3.7: Original Vs Predicted Task Activation maps of Gambling Task (Left - Original, Right - Predicted).

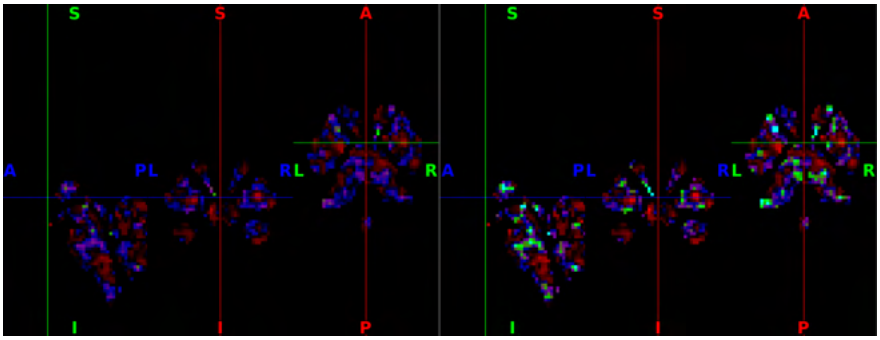


Figure 3.8: Original vs Predicted Task Activation maps of Social Task (Left - Original, Right - Predicted).

Chapter 4

Implementation and Methodology

4.1 IQ Prediction

4.1.1 HYDRA: A Dictionary-Based Time Series Transformation Framework

HYDRA is a dictionary-based method for time series classification that combines the strengths of convolutional kernel-based approaches (like ROCKET) with traditional dictionary learning techniques. It achieves this by transforming the input time series into a sparse feature representation based on the counts of randomly initialized convolutional kernels.

1. Core Design

HYDRA employs a large collection of convolutional kernels grouped into g groups, each containing k kernels. These kernels are initialized randomly and remain fixed throughout training. At each time point in the input series, HYDRA evaluates which kernel in each group most closely matches the current window and increments the

count for that kernel. This count serves as a feature.

2. Kernel Characteristics

The convolutional kernels used in HYDRA adopt the following attributes:

- Kernel length is fixed (default 9).
- Weights are drawn from a normal distribution $\mathcal{N}(0, 1)$.
- Exponential dilations $\{2^0, 2^1, 2^2, \ldots\}$ are used to capture multi-scale information.

Kernels are always zero-padded to allow valid convolutions over variable-length input series.

3. Feature Generation

At each time step, the kernel with the maximum response in each group is identified via an arg max operation. HYDRA supports two modes of counting:

- **Hard Counting:** Increments the count of the kernel with the maximum (or minimum) response.
- **Soft Counting:** Accumulates the value of the maximum (or minimum) response.

By default, HYDRA uses both modes simultaneously: hard counting for the minimum response and soft counting for the maximum.

4. Approximation Behavior

HYDRA's approximation capability is regulated by the number of kernels per group:

- **High k:** When more kernels compete within each group, the chance of accurately matching input subsequences increases. This behavior mimics classical dictionary methods.
- **Low k:** With fewer competing kernels, the matches become coarser and more approximate, capturing more abstract or general features of the signal.

This parameter acts as a trade-off between expressiveness and computational efficiency.

5. Clipping and First-Order Differences

To improve discriminative power and suppress noise, HYDRA supports clipping, where only positive maximum (or negative minimum) responses are counted. Moreover, HYDRA optionally includes the first-order difference of the time series alongside the original input, providing derivative-like features.

6. Final Representation and Classification

The resulting count vectors from all kernel groups and all dilations form a high-dimensional sparse feature vector. These vectors are passed to a linear classifier (ridge or logistic regression) for final prediction.

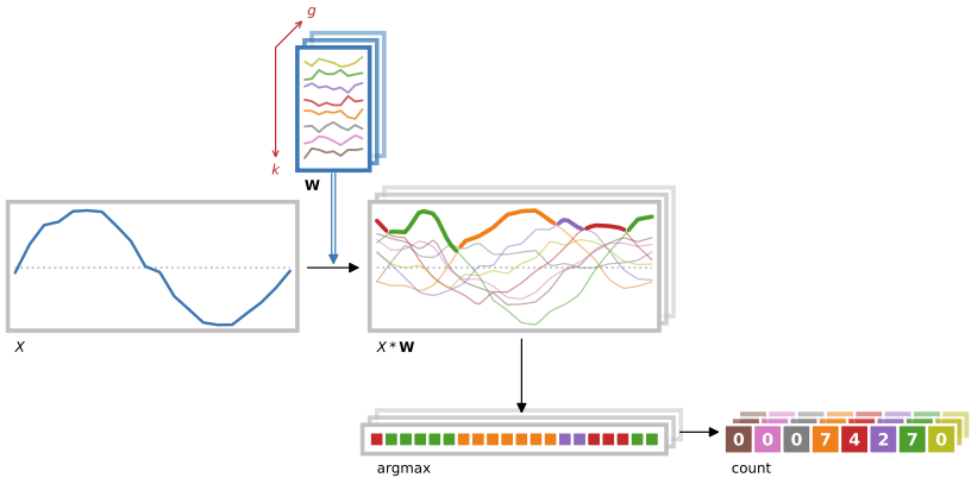


Figure 4.1: Hydra transforms each time series by convolving with random kernels grouped into g sets of k kernels and counts the closest matching kernel per group at each timepoint.

4.1.2 MTSR-based IQ Prediction Model Using HYDRA

The prediction problem in the proposed framework was formulated as a multivariate time series regression (MTSR) task, where the input comprised blood-oxygen-level-dependent (BOLD) time series extracted from resting-state fMRI data, and the output was the intelligence quotient (IQ) score. The rs-fMRI data from the ABIDE dataset underwent preprocessing that included noise and artifact removal. Subsequently, regions of interest (ROIs) were extracted using the CC200 atlas, yielding 200-dimensional time series for each subject.

These time series were provided as input to the HYDRA algorithm [41], which operates by convolving each input time series with a set of random convolutional kernels organized into $N_g = 128$ groups, each containing $N_k = 16$ kernels. At each time point, the kernels that provided the closest match to the input sequence were selected to generate group-wise summaries. These feature representations were then used for regression.

Ridge regression was identified as the optimal regressor for this task. Ridge regression, also known as Tikhonov regularization, addresses multicollinearity and overfitting

by adding an ℓ_2 penalty to the ordinary least squares objective. Formally, the objective is:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \|Y - X\beta\|^2 + \lambda \|\beta\|^2 \right\},$$

where X is the input feature matrix, Y is the target variable (IQ scores), β represents the regression coefficients, and $\lambda > 0$ is the regularization parameter.

This formulation encourages smaller coefficient values, improving model generalization. Using ridge regression, a Pearson correlation of 0.69 was obtained between the predicted and actual IQ scores, which outperformed other tested regression models.

While the HYDRA-based MTSR model exhibited high performance, it did not explicitly leverage the relational dependencies among different BOLD time series. To capture this complementary information, a similarity-based IQ prediction model employing convolutional neural networks (CNNs) was subsequently proposed, as detailed in the next section.

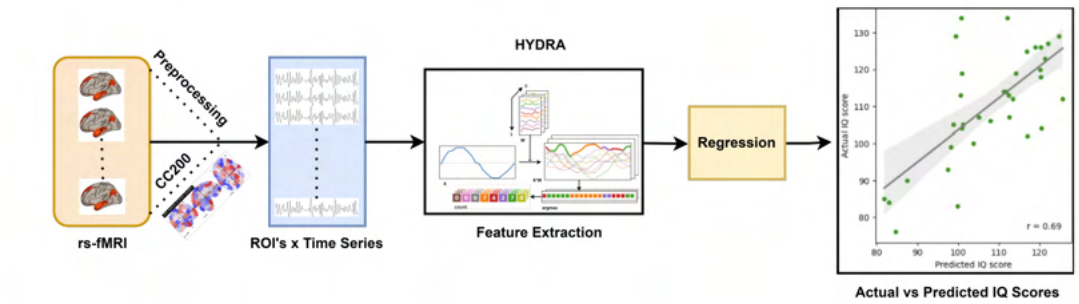


Figure 4.2: MTSR based IQ prediction using HYDRA

4.1.3 Similarity-Based IQ Prediction Model

To complement the temporal modeling of BOLD signals using multivariate time series regression, a relational learning framework was also explored. This approach focuses on capturing pairwise dependencies among different BOLD time series to enrich the feature space for IQ prediction. Unlike the previously proposed MTSR-based models that treat each brain region independently, the relational approach models inter-regional

dependencies by computing various similarity metrics between pairs of BOLD signals. These measures serve to encode the structure of interactions and synchrony in brain dynamics, offering complementary insights into cognitive function.

Specifically, the following six similarity metrics were used to quantify relationships between each pair of time series from 200 brain regions, resulting in six 200×200 similarity matrices for each subject. Each matrix was then fed to a Convolutional Neural Network (CNN) regressor for IQ score prediction.

1. Pearson Correlation Coefficient

The Pearson correlation captures the linear relationship between two time series x and y, and is computed as:

$$\rho(x, y) = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{T}(y_t - \bar{y})^2}}$$

2. Covariance

Covariance evaluates the degree to which two signals vary together and is computed as:

$$\mathrm{Cov}(x, y) = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})$$

3. Phase Lag Index (PLI) [42]

Let $\phi_x(t)$ and $\phi_y(t)$ denote the instantaneous phases obtained using the Hilbert transform. Then:

$$\mathrm{PLI}(x, y) = \left|\mathbb{E}\left[\mathrm{sign}\left(\sin(\phi_x(t) - \phi_y(t))\right)\right]\right|$$

4. Phase Locking Value (PLV) [42]

PLV measures phase synchrony between two signals:

$$\text{PLV}(x, y) = \left| \frac{1}{T} \sum_{t=1}^{T} e^{j(\phi_x(t) - \phi_y(t))} \right|$$

5. Longest Common Subsequence (LCSS) [43]

LCSS identifies the longest matching subsequence with tolerance ϵ for similarity and δ for time index flexibility:

$$\text{LCSS}_\epsilon^\delta(x, y) = \begin{cases} 0, & \text{if } x \text{ or } y \text{ is empty} \\ 1 + \text{LCSS}_\epsilon^\delta(x_{1:T-1}, y_{1:T-1}), & \text{if } |x_T - y_T| < \epsilon \text{ and } |T_x - T_y| < \delta \\ \max\{\text{LCSS}_\epsilon^\delta(x_{1:T-1}, y), \text{LCSS}_\epsilon^\delta(x, y_{1:T-1})\}, & \text{otherwise} \end{cases}$$

6. Dynamic Time Warping (DTW) [44]

DTW computes the minimal alignment cost by warping the time axis of sequences:

$$\text{DTW}(x, y) = \min_{\pi \in \mathcal{P}} \sum_{(i,j) \in \pi} \|x_i - y_j\|^2$$

Using dynamic programming:

$$D(i, j) = \|x_i - y_j\|^2 + \min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\}$$

Modeling

Each similarity matrix was input to a CNN model for IQ prediction. The CNN architecture consisted of three convolutional layers with progressively smaller filters (128,

64, and 32), each followed by max-pooling operations. A fully connected dense layer with 32 neurons was added before the output regression layer. The model was trained for 50 epochs using the Adam optimizer and mean squared error as the loss function.

Among individual similarity measures, the PLI-based CNN yielded the highest Pearson correlation of 0.42 between the predicted and actual IQ scores. To further improve prediction performance, integrated similarity matrices were created by combining multiple similarity measures. A CNN trained on these integrated matrices achieved a higher correlation of 0.50, outperforming the individual measure-based models.

Motivated by these results, a hybrid IQ prediction framework was developed by combining the HYDRA-based MTSR features with the relational features extracted from similarity-based CNNs. This fused model resulted in further improvements, indicating the complementary value of temporal and relational information in modeling cognitive traits. Detailed experimental evaluation and ablation studies are presented in subsequent sections.

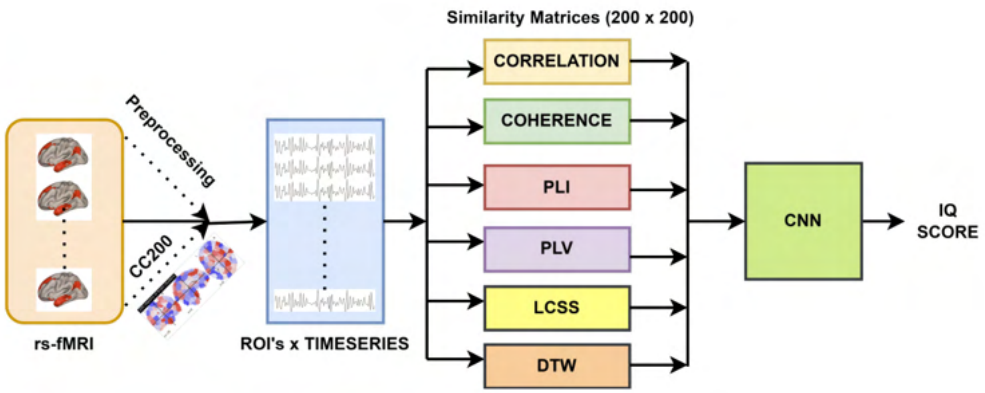


Figure 4.3: Proposed Similarity based IQ prediction model

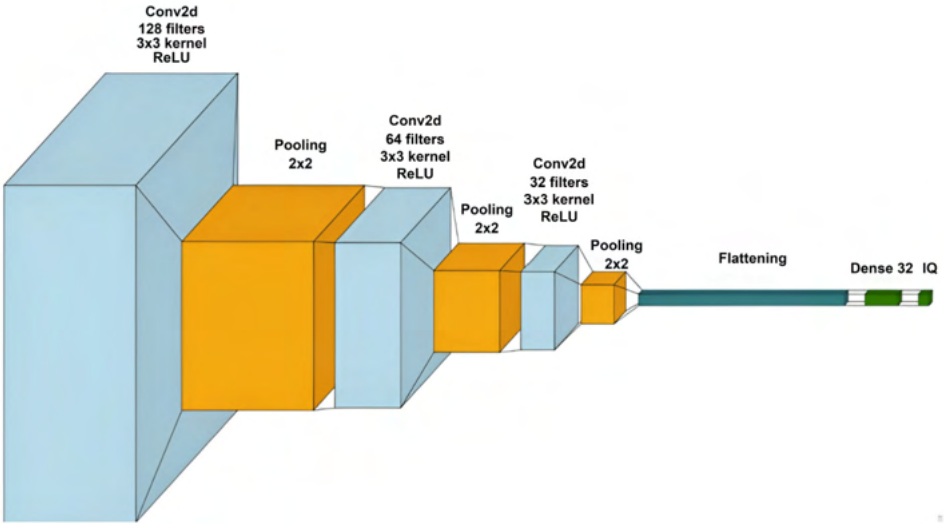


Figure 4.4: CNN Architecture for the proposed similarity based IQ prediction model

4.1.4 Hybrid IQ Prediction Model

Building upon the complementary strengths of the HYDRA-based multivariate time series regression (MTSR) model and the similarity-based relational model, a hybrid IQ prediction framework was developed to enhance prediction accuracy by integrating both temporal and structural information embedded in the rs-fMRI data. The core motivation behind this hybrid approach lies in the understanding that BOLD signals captured from different regions of the brain not only exhibit individual temporal dynamics but also participate in complex inter-regional interactions that encode cognitively relevant information.

In the hybrid framework, the preprocessed rs-fMRI signals were initially modeled as a 200-dimensional multivariate time series using the CC200 parcellation scheme. These time series were input into the HYDRA algorithm [41], which was configured with empirically optimized parameters—specifically, $N_g = 128$ kernel groups and $N_k = 16$ kernels per group. HYDRA performed convolution operations using randomly initialized kernels across all time series and aggregated feature maps based on proximity matching. The resulting feature embeddings captured fine-grained temporal patterns across multiple brain regions, forming the basis of the MTSR-based feature represen-

tation.

Simultaneously, a relational modeling stream was constructed to capture inter-regional dependencies by computing pairwise similarity measures among the 200 BOLD time series. Specifically, four similarity metrics—Phase Lag Index (PLI), Phase Locking Value (PLV), Longest Common Subsequence (LCSS), and Dynamic Time Warping (DTW)—were employed to quantify pairwise interactions. Each measure produced a 200×200 similarity matrix encoding unique aspects of functional relationships. To further enrich the relational representation, integrated similarity matrices were generated by combining the outputs of multiple measures.

Each of these similarity matrices was then passed through a convolutional neural network (CNN) designed to extract higher-order structural features. The CNN architecture consisted of three convolutional layers with progressively decreasing filter sizes (128, 64, and 32), followed by max-pooling operations and a dense layer with 32 neurons preceding the regression output. These convolutional layers were trained to extract abstract spatial patterns embedded in the similarity matrices, translating them into compact feature vectors.

Once both the HYDRA-derived temporal features and the CNN-derived relational features were obtained, they were concatenated to form a unified feature vector that combined both modalities. This fused feature vector was subsequently fed into a ridge regression model, which was previously identified as the best-performing regressor for this task. The combined model achieved a Pearson correlation of 0.75 between the predicted and actual IQ scores—surpassing the performance of models based solely on HYDRA features or similarity-based CNN features.

The observed improvement underscores the importance of intermediate-level fusion [45], wherein independently trained temporal and relational modules contribute complementary information that enhances generalization. By jointly leveraging temporal dynamics and relational structures, the hybrid framework provides a more holistic characterization of brain function, paving the way for more robust and interpretable

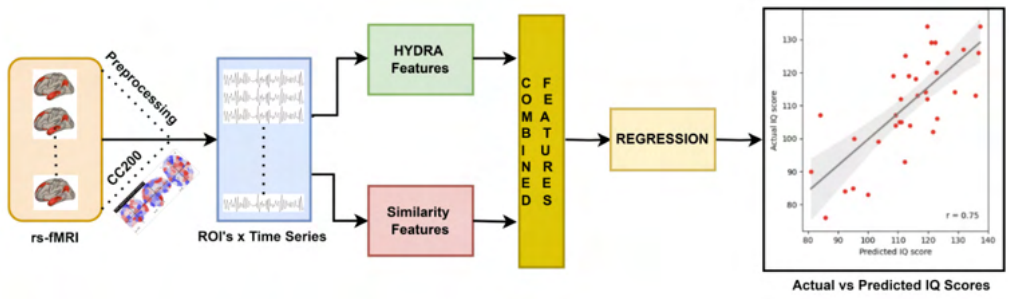


Figure 4.5: Proposed Hybrid IQ Prediction Model

models in cognitive trait prediction. Detailed evaluations, comparative performance metrics, and ablation studies on this hybrid architecture are presented in subsequent sections.

4.2 Task Activation Map Prediction

4.2.1 Time Series Feature Extraction

To encode the temporal characteristics of each signal representation into structured feature vectors, we utilized three categories of time series feature transformations: **CATCH22**, **Summary statistics**, and **Histograms**. These techniques are outlined below:

- **CATCH22**: This framework computes 22 statistically selected features that characterize canonical time-series behavior. These features span various domains such as autocorrelation, entropy, non-linearity, and distributional shape. CATCH22 offers a compact yet expressive representation of the temporal signal, reducing dimensionality while preserving critical information.
- **Summary Statistics**: We employed the Seven-Number Summary transformation, which encapsulates key distributional descriptors of the time series. Specifically, we extract the following features: mean, standard deviation, minimum,

maximum, and the 25^{th}, 50^{th}, and 75^{th} percentiles. This provides a succinct overview of the signal's central tendency, dispersion, and spread.

- **DrCIF (Diversity-reduced Canonical Interval Forest)**: DrCIF (Diversity-reduced Canonical Interval Forest) is a highly accurate and scalable ensemble classifier for time series classification that builds upon the Canonical Interval Forest (CIF) framework. It works by extracting simple yet effective statistical features—mean, standard deviation, and slope—from randomly selected intervals of the time series, and training an ensemble of decision trees on these interval-based features. Unlike CIF, DrCIF incorporates a diversity reduction mechanism to eliminate redundant or highly correlated intervals, thereby enhancing both accuracy and efficiency. This model is effective for both univariate and multivariate time series, and is particularly well-suited for datasets where interpretability, robustness, and computational scalability are important.

4.2.2 Univariate Time Series Regression

The rs-fMRI dataset exhibits extremely high dimensionality. For a single subject, the data comprises approximately 91,282 × 1,200 values (grayordinates by timepoints). Even when limited to cortical grayordinates, the dimensionality remains substantial at 59,412×1,200. To address this challenge and ensure meaningful regional interpretation, we employed the **Multimodal Parcellation (MMP)** scheme to group spatially and functionally coherent grayordinates into biologically meaningful parcels.

Within each parcel, we performed univariate time series regression independently for each grayordinate, resulting in a total of 59,412 regression models. Each model maps the extracted features to the target Z-score.

This regression strategy presents several key advantages:

- **Dimensionality Reduction**: By dividing the data into parcels and analyzing each grayordinate independently, we reduce computational complexity and avoid

overfitting.

- **Localized Interpretation**: The univariate modeling approach facilitates interpretability by preserving the spatial specificity of each grayordinate, enabling localized understanding of neural contributions.

- **Model Simplicity and Clarity**: Univariate regression inherently models the effect of a single predictor on a single outcome, making it well-suited for isolating direct and interpretable relationships between local brain activity and the target variable.

4.2.3 Performance Metrics Used:

- **Pearson Correlation Coefficient:** Measures the linear association between the predicted values $\hat{y}_i$ and the actual values y_i. It is defined as:

$$\text{Corr}(y, \hat{y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{4.1}$$

  where $\bar{y}$ and $\bar{\hat{y}}$ are the means of the actual and predicted values, respectively. A value close to ± 1 indicates a strong linear relationship.

- **Coefficient of Determination ($\mathbf{R}^2$ Score):** Indicates how well the predicted values approximate the actual values. It is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{4.2}$$

  A value of $R^2 = 1$ implies perfect predictions, while $R^2 = 0$ indicates that the model does no better than predicting the mean of the observed data.

- **Mean Squared Error (MSE):** Quantifies the average of the squares of the differences between predicted and actual values, penalizing larger errors more

heavily. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4.3}$$

Smaller values signify better model performance, and the metric is especially sensitive to large deviations (outliers).

For each grayordinate, the model yielding the best trade-off across these metrics was selected as the final model. This ensemble approach allowed us to create a grayordinate-wise optimized regression framework that accounts for localized variations in predictability, model suitability, and temporal signal behavior. The resultant predictions offered an individualized and fine-grained mapping of task activations from resting-state signals, fully learned through data-driven modeling pipelines.

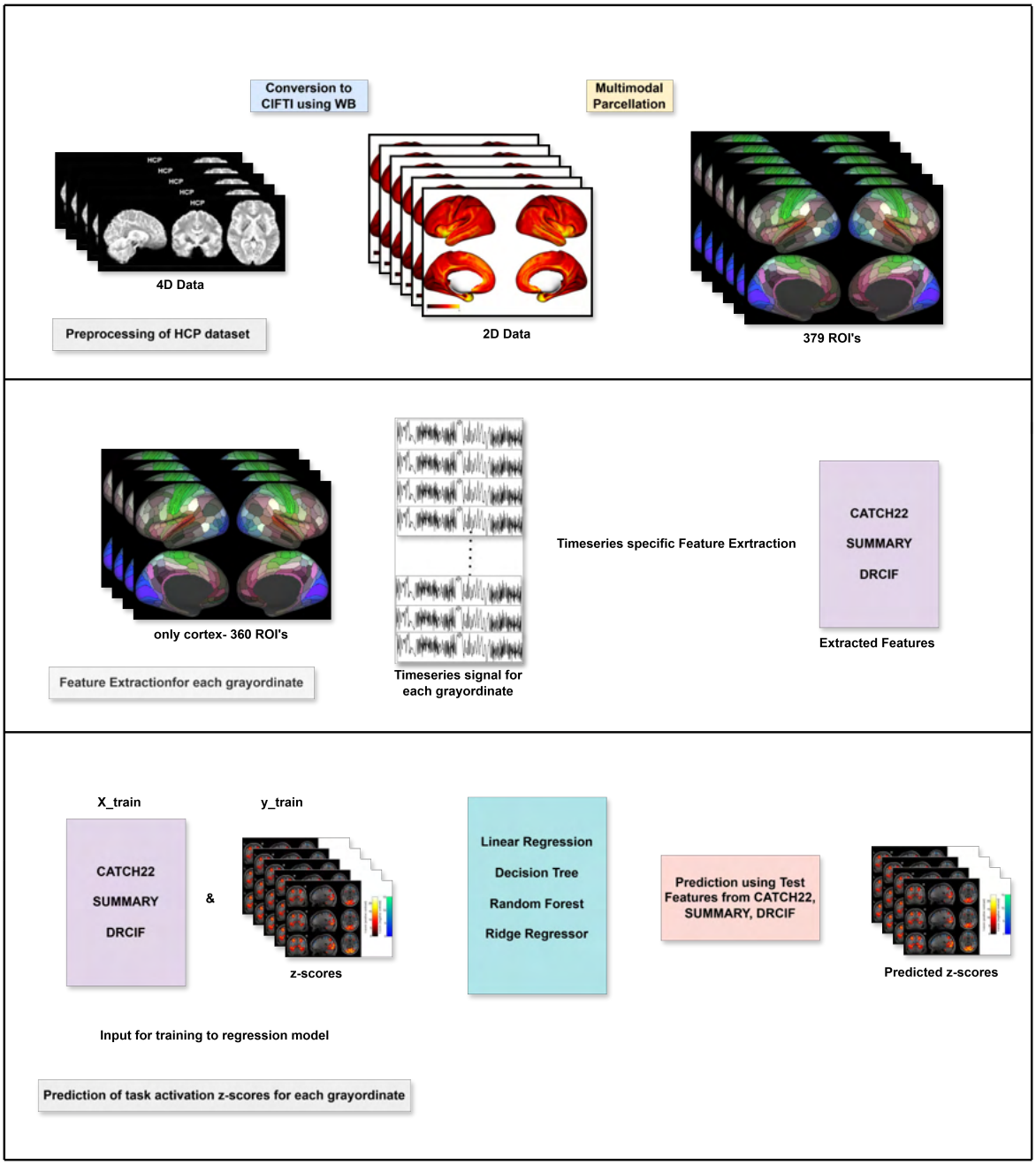


Figure 4.6: Pipeline for task activation map prediction.

4.2.4 List of all Catch22 Features [46]:

- **DN_HistogramMode_5 and DN_HistogramMode_10:** These describe the mode of the time series when divided into fixed-width histogram bins.

  Mathematical definition: Let $x = \{x_1, x_2, \ldots, x_T\}$ be a time series. Partition its range into B equal-width bins. Then,

  $$\text{mode}_B = \arg\max_b \#\{x_t \in \text{bin}_b\}, \quad b = 1, \ldots, B$$

  where $B = 5$ or 10.

- **DN_OutlierInclude_p_001_mdrmd and DN_OutlierInclude_n_001_mdrmd:** These features identify the timing of the most extreme positive and negative outliers, respectively.

  Mathematical description:

  $$z_t = \frac{0.6745(x_t - \text{median})}{\text{MAD}}, \quad \text{find } t : |z_t| > 3.5$$

  The output is t/T where T is the series length.

- **first1e_acf_tau:** The first time lag τ where the autocorrelation function drops below $1/e$:

  $$\text{ACF}(\tau) = \frac{\mathbb{E}[(x_t - \mu)(x_{t+\tau} - \mu)]}{\sigma^2}, \quad \text{find } \tau : \text{ACF}(\tau) < \frac{1}{e}$$

- **firstMin_acf:** The lag τ where the autocorrelation function reaches its first minimum:

  $$\tau_{\min} = \arg\min_{\tau > 0} \text{ACF}(\tau)$$

- **SP_Summaries_welch_rect_area_5_1 and SP_Summaries_welch_rect_centroid:**

  - *low_freq_power:* Total spectral power in the lowest 20% of frequencies.

  - *centroid_freq:* Center of mass of the power spectrum:

$$f_c = \frac{\sum_i f_i P(f_i)}{\sum_i P(f_i)}$$

- **FC_LocalSimple_mean3_stderr:** Standard deviation of the error between actual values and 3-point moving average predictions:

$$\text{error}_t = x_t - \frac{x_{t-1} + x_{t-2} + x_{t-3}}{3}, \quad \text{forecast_error} = \text{std}(\text{error}_t)$$

- **FC_LocalSimple_mean1_tauresrat:** Measures change in autocorrelation timescale after first-order differencing:

$$\text{whiten_timescale} = \frac{\tau_{\text{ACF}}(x)}{\tau_{\text{ACF}}(\nabla x)}$$

- **MD_hrv_classic_pnn40:** Proportion of absolute differences between successive values exceeding 40 ms:

$$\text{pNN40} = \frac{\#\{|x_{t+1} - x_t| > 40\}}{T - 1}$$

- **SB_BinaryStats_mean_longstretch1 and SB_BinaryStats_diff_longstretch0:**

  - *stretch_high:* Longest consecutive sequence of values above the mean.

  - *stretch_decreasing:* Longest decreasing sequence.

- **SB_MotifThree_quantile_hh:** Entropy of adjacent symbol pairs in a quantized version of the time series:

$$H = -\sum_{ij} p_{ij} \log p_{ij}, \quad p_{ij} = \mathbb{P}(s_t = i, s_{t+1} = j)$$

- **CO_HistogramAMI_even_2_5:** Estimates auto mutual information at lag 2

using a 5-bin histogram:

$$I(X_t; X_{t+\tau}) = \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

- **CO_trev_1_num:** Measures asymmetry in time:

$$\text{trev} = \mathbb{E}[(x_{t+1} - x_t)^3]$$

- **IN_AutoMutualInfoStats_40_gaussian_fmmi:** First minimum of the auto mutual information function, indicative of the independence timescale.

- **SB_TransitionMatrix_3ac_sumdiagcov:** Variance of columns in the transition matrix derived from symbolic dynamics.

- **PD_PeriodicityWang_th001:** Periodicity score based on ACF:

$$P = \frac{\sum_\tau \text{ACF}(\tau)^2}{\sum_\tau \text{ACF}(\tau)}$$

- **CO_Embed2_Dist_tau_d_expfit_meandiff:** Evaluates how well the distribution of 2D embedding distances fits an exponential decay.

- **SC_FluctAnal_2_rsrangefit_50_1_logi_prop_r1:** Rescaled range (R/S) analysis for estimating Hurst exponent.

- **SC_FluctAnal_2_dfa_50_1_2_logi_prop_r1:** Detrended Fluctuation Analysis (DFA) to quantify long-range correlations:

$$F(n) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (X(i) - Y_n(i))^2}$$

where $Y_n(i)$ is the best linear fit in a local window.

4.2.5 List of Summary Time Series Features:

- **SevenNumber Summary:** In this feature extraction technique, the input time series is transformed into a set of descriptive statistical summaries. Specifically, the following features are computed:
  - *Mean* (μ): Average value of the series.
  - *Standard deviation* (σ): Measures the dispersion or spread.
  - *Minimum* ($\min(x)$): Smallest observed value.
  - *Maximum* ($\max(x)$): Largest observed value.
  - *25th percentile* ($Q_{0.25}$): Value below which 25% of the data fall.
  - *Median* ($Q_{0.5}$): Central value (50th percentile).
  - *75th percentile* ($Q_{0.75}$): Value below which 75% of the data fall.

  These features collectively capture the central tendency, spread, and range of the distribution, providing a compact and informative summary of the time series.

4.2.6 Diversity-reduced Canonical Interval Forest (DrCIF)

1. Introduction

The **Diversity-reduced Canonical Interval Forest (DrCIF)** is an ensemble learning algorithm for time series classification. It builds upon the Canonical Interval Forest (CIF) framework by incorporating a redundancy-reduction strategy to eliminate correlated interval features, thereby improving classification accuracy, generalization, and computational efficiency. DrCIF is applicable to both univariate and multivariate time series and leverages simple statistical descriptors (canonical features) computed over randomly selected intervals.

2. Canonical Feature Extraction

Let $x = [x_1, x_2, \ldots, x_T]$ be a univariate time series of length T. For an interval $[s, e] \subseteq [1, T]$, the following canonical features are computed:

- **Mean**:

$$\mu_{s:e} = \frac{1}{e - s + 1} \sum_{t=s}^{e} x_t$$

- **Standard Deviation**:

$$\sigma_{s:e} = \sqrt{\frac{1}{e - s + 1} \sum_{t=s}^{e} (x_t - \mu_{s:e})^2}$$

- **Slope** (from linear regression):

$$a_{s:e} = \frac{\sum_{t=s}^{e} (t - \bar{t})(x_t - \mu_{s:e})}{\sum_{t=s}^{e} (t - \bar{t})^2}, \quad \bar{t} = \frac{1}{e - s + 1} \sum_{t=s}^{e} t$$

For a multivariate time series with d channels, these features are computed independently per channel and concatenated:

$$\phi(x^{(1)}), \phi(x^{(2)}), \ldots, \phi(x^{(d)})$$

3. Interval Sampling and Diversity Reduction

For each decision tree, M random intervals $\{[s_i, e_i]\}_{i=1}^{M}$ are sampled. From these, the feature set $\Phi = \{\phi_1, \ldots, \phi_M\}$ is constructed. To reduce redundancy:

- Compute pairwise Pearson correlation $S(i, j) = \text{corr}(\phi_i, \phi_j)$.
- Set a threshold $\rho \in [0, 1]$ for correlation.
- Select features greedily: add ϕ_i to final set $\mathcal{F}$ only if $|\text{corr}(\phi_i, \phi_j)| < \rho \quad \forall \phi_j \in \mathcal{F}$.

This yields a diverse set of informative features per tree.

4. Ensemble Construction and Inference

Let N be the number of decision trees. Each tree is trained using:

- A bootstrap sample from the training set.
- A reduced, diverse set of interval features $\mathcal{F}$.

At inference time, for a test time series x, each tree T_i predicts a label, and the final class is chosen by majority voting:

$$\hat{y}(x) = \text{mode}\left\{T_i(x) \mid i = 1, 2, \ldots, N\right\}$$

5. Algorithm

Algorithm 1 DrCIF (for univariate time series)

Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, number of trees N, intervals per tree M, redundancy threshold ρ Predicted label $\hat{y}$ for test instance x

- **for** $i = 1$ N **do** Sample bootstrap dataset $\mathcal{D}_i \subset \mathcal{D}$ Sample M random intervals $\{I_1, \ldots, I_M\}$ Compute feature vectors $\Phi = \{\phi_k\}$ Apply redundancy filtering with threshold ρ to get $\mathcal{F}_i$ Train decision tree T_i on $\mathcal{F}_i$ For test input x, compute features $\phi(x)$ and compute prediction:

$$\hat{y} = \text{mode}\left\{T_i(x)\right\}_{i=1}^{N}$$

6. Computational Complexity

Let T be the time series length, n the number of samples, d the number of channels, M the number of intervals per tree, and N the number of trees:

- **Training Complexity**: $\mathcal{O}(N \cdot M \cdot d \cdot n \cdot T)$
- **Testing Complexity**: $\mathcal{O}(N \cdot M \cdot d \cdot T)$
- Redundancy filtering adds $\mathcal{O}(M^2)$ correlation checks per tree.

4.3 Graph Pruning

4.3.1 Background

Let $\mathcal{X}$ be a measurable space, and let Q be a probability measure on $\mathcal{X}$. Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function such that $\mathbb{E}_Q[e^{f(x)}] < \infty$. Define the log-partition function (log-moment generating function) as

$$\Lambda_Q(f) := \log \mathbb{E}_Q \left[e^{f(x)} \right].$$

Then, the Donsker–Varadhan variational formula [47] states:

$$\Lambda_Q(f) = \sup_{P \ll Q} \left\{ \mathbb{E}_P[f(x)] - D_{\mathrm{KL}}(P \| Q) \right\},$$

where $D_{\mathrm{KL}}(P\|Q)$ denotes the Kullback–Leibler divergence between P and Q, defined as

$$D_{\mathrm{KL}}(P\|Q) := \int_{\mathcal{X}} \log \left(\frac{dP}{dQ} \right) dP = \mathbb{E}_P \left[\log \left(\frac{dP}{dQ} \right) \right].$$

Mathematical Interpretation

For any $P \ll Q$, define the Radon–Nikodym derivative $\frac{dP}{dQ}$. Then the functional

$$\mathbb{E}_P[f] - D_{\mathrm{KL}}(P\|Q) = \int_{\mathcal{X}} f(x)\, dP(x) - \int_{\mathcal{X}} \log \left(\frac{dP}{dQ}(x) \right) dP(x)$$

expresses a trade-off between maximizing the expected value of f and minimizing divergence from the reference distribution Q.

Proof Sketch

Define the **Gibbs measure** P_f induced by f and Q as:

$$\frac{dP_f}{dQ}(x) := \frac{e^{f(x)}}{\mathbb{E}_Q[e^{f(x)}]}.$$

This gives:

$$dP_f(x) = \frac{e^{f(x)}}{Z}\, dQ(x), \quad \text{where } Z = \mathbb{E}_Q[e^{f(x)}].$$

Now compute the KL divergence:

$$\begin{aligned} D_{\mathrm{KL}}(P_f \| Q) &= \int_{\mathcal{X}} \log\left(\frac{dP_f}{dQ}(x)\right) dP_f(x) \\ &= \int_{\mathcal{X}} \log\left(\frac{e^{f(x)}}{Z}\right) \cdot \frac{e^{f(x)}}{Z}\, dQ(x) \\ &= \mathbb{E}_{P_f}[f] - \log \mathbb{E}_Q[e^{f(x)}]. \end{aligned}$$

Rearranging gives:

$$\log \mathbb{E}_Q[e^{f(x)}] = \mathbb{E}_{P_f}[f] - D_{\mathrm{KL}}(P_f \| Q).$$

Hence,

$$\log \mathbb{E}_Q[e^{f(x)}] = \sup_{P \ll Q} \left\{ \mathbb{E}_P[f(x)] - D_{\mathrm{KL}}(P \| Q) \right\},$$

with the supremum attained at $P = P_f$, completing the proof sketch.

Mutual Information and KL Divergence

Mutual information is a measure of how much knowing one random variable tells us about another. It can be defined using Shannon entropy as:

$$I(X; Z) = H(X) - H(X|Z),$$

where $H(X)$ is the entropy of X, and $H(X|Z)$ is the conditional entropy of X given Z. This expression tells us how much uncertainty in X is reduced when Z is known.

Mutual information can also be written using KL divergence between the joint distribution $\mathbb{P}_{XZ}$ and the product of the marginals $\mathbb{P}_X \otimes \mathbb{P}_Z$:

$$I(X; Z) = D_{\mathrm{KL}}(\mathbb{P}_{XZ} \,\|\, \mathbb{P}_X \otimes \mathbb{P}_Z).$$

KL divergence itself is a way to measure how different two probability distributions are. For distributions $\mathbb{P}$ and $\mathbb{Q}$ (with $\mathbb{P} \ll \mathbb{Q}$), it is defined as:

$$D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \mathbb{E}_{\mathbb{P}} \left[\log \left(\frac{d\mathbb{P}}{d\mathbb{Q}} \right) \right].$$

Dual Representations of KL Divergence

Modern information estimation methods often use alternative "dual" forms of KL divergence. These allow us to write KL divergence as a maximization problem over functions, which is especially useful for learning algorithms.

Donsker–Varadhan Representation. The KL divergence can be written as:

$$D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) = \sup_{T:\Omega \to \mathbb{R}} \left\{ \mathbb{E}_{\mathbb{P}}[T] - \log \left(\mathbb{E}_{\mathbb{Q}}[e^T] \right) \right\},$$

where the supremum is taken over all functions T such that both expectations are finite.

This means we can estimate KL divergence by finding a function T that maximizes the difference between the expected value of T under $\mathbb{P}$, and the log of the expected exponential of T under $\mathbb{Q}$.

Lower Bound Using Restricted Function Class. If we restrict T to a specific class of functions $\mathcal{F}$, we get a lower bound:

$$D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \left\{ \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T] \right\}.$$

Gibbs Density. The optimal function T^* that achieves the supremum is connected to the Gibbs distribution:

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{e^{T^*}}{\mathbb{E}_{\mathbb{Q}}[e^{T^*}]}.$$

Comparison with the f-Divergence Representation

Another popular dual form of KL divergence is the f-divergence representation, given by:

$$D_{\mathrm{KL}}(\mathbb{P} \,\|\, \mathbb{Q}) \geq \sup_{T \in \mathcal{F}} \left\{ \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[e^T - 1] \right\}.$$

Both this and the Donsker–Varadhan form are tight when the function class $\mathcal{F}$ is large enough. However, the Donsker–Varadhan bound is often *stronger* for a fixed T, meaning:

$$\mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T] \geq \mathbb{E}_{\mathbb{P}}[T] - \mathbb{E}_{\mathbb{Q}}[e^T - 1],$$

since $\log x \leq x - 1$ for $x > 0$.

This makes the Donsker–Varadhan form more desirable in practice, especially in mutual information estimators such as MINE. Theoretical connections between both forms can be understood via Fenchel duality, as explained in Ruderman et al. (2012).

4.3.2 Mutual Information Neural Estimator - MINE

Recall that mutual information between random variables X and Z can be expressed as:

$$I(X;Z) = D_{\mathrm{KL}}(\mathbb{P}_{XZ} \,\|\, \mathbb{P}_X \otimes \mathbb{P}_Z),$$

where $\mathbb{P}_{XZ}$ is the joint distribution, and $\mathbb{P}_X \otimes \mathbb{P}_Z$ is the product of the marginals. Using the Donsker–Varadhan variational representation of the KL-divergence, we obtain the lower bound:

$$I(X;Z) \geq I_\theta(X,Z),$$

where $I_\theta(X,Z)$ is referred to as the *neural information measure*, defined by:

$$I_\theta(X,Z) := \sup_{\theta\in\Theta} \left\{ \mathbb{E}_{\mathbb{P}_{XZ}}[T_\theta] - \log \mathbb{E}_{\mathbb{P}_X \otimes \mathbb{P}_Z}[e^{T_\theta}] \right\}.$$

Here, $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is a neural network parameterized by θ, often called the *statistics network*. This formulation defines a new family of lower bounds on mutual information that can be optimized using stochastic gradient-based methods.

Practical Estimation. In practice, the expectations in the above expression are computed using samples. Let $\hat{\mathbb{P}}^{(n)}$ denote the empirical distribution over n i.i.d. samples. Then, the MINE estimate becomes:

$$I(X;Z)_n := \sup_{\theta\in\Theta} \left\{ \mathbb{E}_{\hat{\mathbb{P}}^{(n)}_{XZ}}[T_\theta] - \log \mathbb{E}_{\hat{\mathbb{P}}^{(n)}_X \otimes \hat{\mathbb{P}}^{(n)}_Z}[e^{T_\theta}] \right\}.$$

Definition(MINE). Let $\mathcal{F} = \{T_\theta\}_{\theta\in\Theta}$ denote the family of neural network functions. Then, the Mutual Information Neural Estimator (MINE) is formally defined as:

$$I(X;Z)_n := \sup_{\theta\in\Theta} \left\{ \mathbb{E}_{\hat{\mathbb{P}}^{(n)}_{XZ}}[T_\theta] - \log \mathbb{E}_{\hat{\mathbb{P}}^{(n)}_X \otimes \hat{\mathbb{P}}^{(n)}_Z}[e^{T_\theta}] \right\}.$$

Optimization Algorithm

The MINE objective is differentiable with respect to θ, so it can be optimized using stochastic gradient descent (SGD). The practical optimization algorithm is summarized below.

Algorithm 1: MINE Training Procedure

1. Initialize neural network parameters θ.
2. Repeat until convergence:
   (a) Draw a minibatch of b samples $\{(x^{(i)}, z^{(i)})\}_{i=1}^{b}$ from the joint $\mathbb{P}_{XZ}$.
   (b) Draw independent samples $\{z'^{(i)}\}_{i=1}^{b}$ from the marginal $\mathbb{P}_Z$.
   (c) Compute the variational lower bound estimate:

$$\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(x^{(i)}, z^{(i)}) - \log\left(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(x^{(i)}, z'^{(i)})}\right).$$

   (d) Compute the gradient $\hat{G}(\theta) = \nabla_\theta \mathcal{V}(\theta)$.
   (e) Update network parameters: $\theta \leftarrow \theta + \eta \cdot \hat{G}(\theta)$, where η is the learning rate.

Correcting the Bias from Stochastic Gradients

A naive implementation of the above algorithm introduces bias due to the use of minibatches when estimating the denominator in the gradient of the log term. Specifically, the stochastic gradient is:

$$\hat{G}_B = \mathbb{E}_B[\nabla_\theta T_\theta] - \frac{\mathbb{E}_B[\nabla_\theta T_\theta e^{T_\theta}]}{\mathbb{E}_B[e^{T_\theta}]},$$

where expectations are taken over the samples in a minibatch B. This estimator is biased because the ratio of expectations is not equal to the expectation of the ratio.

To reduce this bias, an exponential moving average can be used to estimate the denominator more accurately. This correction has been shown to improve the stability and performance of MINE in practice, especially for small batch sizes.

4.3.3 Information Bottleneck Theory and its Connection to Rate Distortion Theory

Information Bottleneck Framework

The *Information Bottleneck* (IB) method, proposed by Tishby et al., formulates the problem of extracting the most relevant information from a random variable X about another variable Y. The goal is to find a compressed representation T of X that preserves maximal information about Y.

The trade-off is captured by the following constrained optimization problem:

$$\min_{p(t|x)} I(X;T) \quad \text{subject to} \quad I(T;Y) \geq \text{const.}$$

Using a Lagrangian formulation, this becomes:

$$\mathcal{L}_{\text{IB}} = I(X;T) - \beta I(T;Y),$$

where:

- $I(X;T)$ measures how much information the representation T preserves about the input X,
- $I(T;Y)$ quantifies how informative the representation T is about the target Y,
- $\beta > 0$ is a trade-off parameter that adjusts the balance between compression and relevance.

Rate Distortion Theory

Rate Distortion Theory (RDT) is a classical framework in information theory concerned with lossy data compression. It characterizes the minimum number of bits per symbol required to represent a source X such that the expected distortion remains within an acceptable level.

The optimization problem for RDT is formulated as:

$$R(D) = \min_{p(\hat{x}|x):\mathbb{E}[d(X,\hat{X})]\leq D} I(X;\hat{X}),$$

where:

- $\hat{X}$ is the reconstruction of X,
- $d(x,\hat{x})$ is a distortion function,
- D is the maximum allowable distortion.

This yields the *rate-distortion function* $R(D)$, which gives the minimal achievable rate (in bits per symbol) under a distortion constraint.

Connection Between Information Bottleneck and Rate Distortion Theory

The Information Bottleneck principle can be viewed as a generalization of Rate Distortion Theory. In particular:

- In RDT, the aim is to compress X while maintaining fidelity to X itself (through $\hat{X}$), whereas in IB, the aim is to compress X while preserving relevance to a different variable Y.
- The distortion measure $d(x,\hat{x})$ in RDT is replaced by the mutual information $I(T;Y)$ in IB, which measures relevance.

The parallel between the two frameworks is summarized by the following formulations:

Concept	Rate Distortion Theory (RDT)	Information Bottleneck (IB)
Compression	$I(X;\hat{X})$	$I(X;T)$
Relevance	$\mathbb{E}[d(X,\hat{X})]$	$I(T;Y)$
Objective	$\min I(X;\hat{X})$ s.t. $\mathbb{E}[d] \leq D$	$\min I(X;T)$ s.t. $I(T;Y) \geq \text{const}$
Dual Form	$R(D) = I(X;\hat{X}) + \beta\mathbb{E}[d]$	$\mathcal{L}_{\text{IB}} = I(X;T) - \beta I(T;Y)$

4.3.4 A New Variational Principle for Information Bottleneck

Formulation of the Variational Principle

We aim to compress a signal X into a quantized representation $\tilde{X}$, such that as much information as possible about a relevant variable Y is retained. Unlike rate-distortion theory where the focus is on fidelity to the input, here the focus is on retaining relevance. The mutual information between $\tilde{X}$ and Y is:

$$I(\tilde{X};Y) = \sum_{y}\sum_{\tilde{x}} p(y,\tilde{x}) \log \frac{p(y,\tilde{x})}{p(y)p(\tilde{x})} \leq I(X;Y).$$

We cannot increase information through compression, but we seek to find a balance between compressing X (minimizing $I(X;\tilde{X})$) and preserving information about Y (maximizing $I(\tilde{X};Y)$).

This leads to the variational objective:

$$\mathcal{L}[p(\tilde{x}|x)] = I(\tilde{X};X) - \beta I(\tilde{X};Y),$$

where β is a Lagrange multiplier controlling the trade-off between compression and relevance.

Self-Consistent Equations

The optimal solution is obtained by minimizing the Lagrangian above. Theorem 4 gives the optimal assignment:

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left(-\beta \sum_y p(y|x) \log \frac{p(y|x)}{p(y|\tilde{x})}\right),$$

where $Z(x,\beta)$ is a normalization term, and $p(y|\tilde{x})$ is defined using the Markov chain $\tilde{X} \leftarrow X \rightarrow Y$ as:

$$p(y|\tilde{x}) = \frac{1}{p(\tilde{x})} \sum_x p(y|x)p(\tilde{x}|x)p(x),$$

and marginalization conditions ensure consistency:

$$p(\tilde{x}) = \sum_x p(\tilde{x}|x)p(x), \quad p(y|\tilde{x}) = \sum_x p(y|x)p(x|\tilde{x}).$$

Derivation of Optimality Conditions

Differentiating the Lagrangian w.r.t. $p(\tilde{x}|x)$ gives:

$$\frac{\delta \mathcal{L}}{\delta p(\tilde{x}|x)} = p(x) \left\{ \log \left[\frac{p(\tilde{x}|x)}{p(\tilde{x})}\right] - \beta \sum_y p(y|x) \log \left[\frac{p(y|x)}{p(y|\tilde{x})}\right] - \frac{\lambda(x)}{p(x)} \right\}.$$

Absorbing terms independent of $\tilde{x}$ into $\lambda(x)$, and defining:

$$\tilde{\lambda}(x) = \frac{\lambda(x)}{p(x)} - \beta \sum_y p(y|x) \log \left[\frac{p(y|x)}{p(y)}\right],$$

we recover:

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left(-\beta D_{\mathrm{KL}}[p(y|x)\|p(y|\tilde{x})]\right),$$

with normalization:

$$Z(x,\beta) = \sum_{\tilde{x}} p(\tilde{x}) \exp\left(-\beta D_{\mathrm{KL}}[p(y|x)\|p(y|\tilde{x})]\right).$$

Iterative Information Bottleneck Algorithm

We define the free energy functional:

$$\mathcal{F}[p(\tilde{x}|x); p(\tilde{x}); p(y|\tilde{x})] = -\mathbb{E}_{p(x)}[\log Z(x,\beta)] = I(X;\tilde{X}) + \beta \mathbb{E}_{p(x,\tilde{x})}[D_{\mathrm{KL}}(p(y|x)\|p(y|\tilde{x}))].$$

At each iteration t, update as:

$$\begin{cases} p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp(-\beta d(x,\tilde{x})), \\ p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x), \\ p_{t+1}(y|\tilde{x}) = \sum_y p(y|x) p_t(x|\tilde{x}), \end{cases}$$

where $d(x,\tilde{x}) = D_{\mathrm{KL}}[p(y|x)\|p(y|\tilde{x})]$.

Convexity and Convergence

The convergence of this iterative method is guaranteed because each step minimizes the same functional independently. The functional is convex in each distribution individually, ensuring monotonic decrease of the free energy. However, the joint space is not convex, so uniqueness is not guaranteed.

Structure of the Solutions

The solutions of the self-consistent equations depend on the cardinality of $\tilde{X}$ and the value of β. For every β, the mutual information values $I(X;\tilde{X})$ and $I(\tilde{X};Y)$ trace out a curve in the "information plane".

The derivative of the curve is:

$$\frac{\delta I(\tilde{X};Y)}{\delta I(X;\tilde{X})} = \beta^{-1} > 0.$$

These form a family of convex curves (similar to rate-distortion theory), with

transitions at critical β, indicating phase changes in the structure of quantizations.

4.3.5 Graph Pruning using information bottleneck principle

Lagrangian Formulation with Penalty Design

We formulate the following Lagrangian for sparse graph learning:

$$\mathcal{L}(\hat{A}, \tau, \alpha) = I(\hat{A}; A) - \beta \cdot I(G'; Y) + \mu \cdot \text{Penalty}(\hat{A}, \tau, \alpha) \tag{4.4}$$

Term Explanations

- $I(\hat{A}; A)$: Mutual information between the learned sparse adjacency matrix $\hat{A}$ and the original adjacency matrix A. Encourages retention of key structural features from the original graph.
- $I(G'; Y)$: Mutual information between the graph representation G' (after pruning) and the labels Y. Penalizes overfitting by suppressing excessive label dependence.
- β: Trade-off parameter balancing structure preservation and label compression.
- μ: Regularization coefficient controlling the influence of the sparsity constraint.
- **Penalty term**:

$$\text{Penalty}(\hat{A}, \tau, \alpha) = \text{ReLU}\left(\sum_{i,j} \sigma\left(\gamma \cdot (A_{ij} - \tau)\right) - \alpha \cdot n^2\right) \tag{4.5}$$

  where:

  * $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function,
  * γ controls the sharpness of thresholding,

* τ is a learnable threshold,
* α sets the sparsity target,
* n is the number of nodes in the graph.

Design of the Penalty Term

The penalty term is constructed to softly enforce a constraint on the number of active edges in $\hat{A}$, by approximating the binary indicator function with a smooth sigmoid. This allows gradient-based optimization to be applied, while still enabling eventual hard pruning.

Why Sigmoid? The sigmoid provides a *differentiable approximation* to the discrete indicator $\mathbb{1}[A_{ij} > \tau]$, enabling backpropagation and continuous optimization. It allows the model to learn which edges are informative enough to retain based on their activations relative to τ.

Three-Phase Gamma Scheduling Strategy

To balance exploration and constraint during training, we apply a scheduled control on the sigmoid's steepness using the parameter γ, progressing through three phases:

Phase I: Free Exploration ($\gamma = 0.1$) The sigmoid is almost linear, resulting in a *soft penalty* where nearly all edges contribute partially. Enables *broad search* over the space of possible edge configurations. Promotes *diversity and flexibility* early in training.

Phase II: Constrained Refinement ($\gamma = 1$) The sigmoid curve becomes steeper, so only moderately confident edges significantly contribute. *Narrows the search space*, guiding the model toward informative structures while still allowing flexibility.

Phase III: Hard Commitment ($\gamma = 10$) Sigmoid approximates a step function, pushing the penalty toward *binary decisions*. Acts as a *soft-to-hard pruning* mechanism. Finalizes the sparse structure while preserving differentiability.

This *annealing-inspired schedule* allows the model to gradually shift from *free exploration* to *strict structural regularization*, ensuring both learning flexibility and convergence stability.

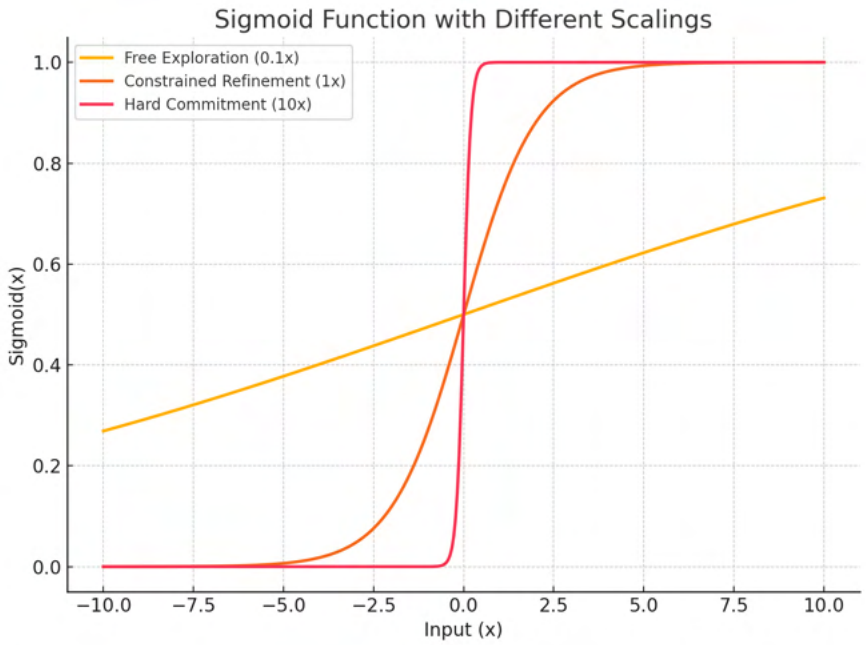


Figure 4.7: Visualization of the sigmoid function under different scaling regimes—*Free Exploration* ($0.1x$), *Constrained Refinement* ($1x$), and *Hard Commitment* ($10x$). As the scaling factor increases, the sigmoid curve transitions from a smooth, gradual slope to a sharp, almost binary step, reflecting increasing decisiveness in the activation behavior.

Information-Theoretic Interpretation of Scaling Regimes in the Sigmoid Function

To understand why the specified three γ values—namely *Free Exploration* ($\gamma = 0.1$), *Constrained Refinement* ($\gamma = 1$), and *Hard Commitment* ($\gamma = 10$)—are named as such, we analyze their information-theoretic properties using the concept of differential entropy.

We consider the sigmoid transformation $\sigma(x) = \frac{1}{1+e^{-x}}$ applied to a variable $x \in (-\varepsilon, +\varepsilon)$, where x is sampled from a uniform distribution, $x \sim \mathcal{U}(-\varepsilon, +\varepsilon)$

with $\varepsilon = 0.5$. The transformed output is $Y = \sigma(\gamma x)$, where γ controls the sharpness of the sigmoid.

Under variable transformation, the probability density function (PDF) of Y becomes:

$$f_Y(y) = \frac{1}{2\varepsilon} \cdot \frac{1}{\gamma \cdot y(1-y)} \quad \text{for } y \in (\sigma(-\gamma\varepsilon), \sigma(\gamma\varepsilon)) \tag{4.6}$$

The differential entropy of the continuous variable Y is then given by:

$$h(Y) = -\int f_Y(y) \log f_Y(y)\, dy \tag{4.7}$$

This integral is evaluated numerically for each γ value, yielding the results in Table 4.1.

Scaling Factor (γ)	**Interpretation**	**Differential Entropy (nats)**
0.1	Free Exploration	-3.689
1.0	Constrained Refinement	-1.407
10.0	Hard Commitment	-0.524

Table 4.1: Differential entropy of sigmoid-transformed variables in the interval $(-\varepsilon, +\varepsilon)$ with $\varepsilon = 0.5$

Interpretation of Negative Differential Entropy

Unlike discrete entropy, differential entropy can take negative values. This is because it depends not just on the spread of the probability distribution but also on its absolute scale. When a probability distribution becomes tightly concentrated within a small interval, its density values become high, and the integral of $-f(x) \log f(x)$ can result in a negative number.

In the case of the sigmoid transformation, increasing γ makes the function steeper, thereby compressing the output $Y = \sigma(\gamma x)$ into a narrow interval around $y = 0.5$. This leads to higher values of the PDF $f_Y(y)$ and hence lower (more negative) differential entropy values.

Justification of Nomenclature

The names assigned to the scaling factors—*Free Exploration*, *Constrained Refinement*, and *Hard Commitment*—are justified by their corresponding entropy values:

- **Free Exploration** ($\gamma = 0.1$) yields the highest entropy (-3.689), indicating a more uniform spread and less certainty in the output, allowing for exploratory behavior.
- **Constrained Refinement** ($\gamma = 1$) represents a moderate entropy level (-1.407), balancing exploration and commitment, which suits controlled refinement of decisions.
- **Hard Commitment** ($\gamma = 10$) has the lowest entropy (-0.524), signifying a sharply peaked distribution where the output is almost binary, thus modeling deterministic or committed choices.

Hence, the information-theoretic behavior of the sigmoid-transformed variable across different scalings validates the conceptual interpretation of each regime.

Algorithm: Sparse Graph Learning with Hard Thresholding

Input:

- Adjacency matrix A (weighted graph).
- Node features X and labels Y.
- Initial parameters:
  - Learnable: τ (threshold), α (sparsity ratio).
  - Hyperparameters: β, μ.
  - Scheduling parameter: γ (initial value, scheduling strategy).
  - Target sparsity ratio: α_{target}.

Output:

- Optimized adjacency matrix $\hat{A}$.
- Learned parameters τ, α.

Steps:

1. Initialize Parameters:

- Initialize τ, α, γ (based on chosen schedule), and hyperparameters β, μ.
- Initialize $\hat{A}$ as a random adjacency matrix where each entry is greater than τ:

$$\hat{A}_{ij} = \text{Uniform}(\tau, \tau + \epsilon), \quad \forall (i, j)$$

where ϵ is a small positive constant (e.g., 0.1) to ensure $\hat{A}_{ij} > \tau$.

2. For Each Epoch (Training Loop):

Step 2.1: Forward Pass

1. Compute the **Mutual Information Terms**:
   - $I(\hat{A}; A)$: Using MINE to estimate mutual information between $\hat{A}$ and A.
   - $I(G'; Y)$: Using GraphMINE to estimate mutual information between G' (graph derived from $\hat{A}$) and node labels Y.
2. Compute the **Penalty Term**:

$$\text{Penalty}(\hat{A}, \tau, \alpha) = \text{ReLU}\left(\sum_{i,j} \sigma(\gamma \cdot (A_{ij} - \tau)) - \alpha \cdot n^2\right)$$

3. Compute the **Lagrangian**:

$$\mathcal{L} = I(\hat{A}; A) - \beta \cdot I(G'; Y) + \mu \cdot \text{Penalty}(\hat{A}, \tau, \alpha)$$

Step 2.2: Backward Pass

1. Compute gradients:
   - Update τ (threshold) and α (sparsity ratio).
2. Update scheduling parameter γ based on the chosen strategy:
   - Linear: $\gamma \leftarrow \gamma + k$ (constant increment).
   - Exponential: $\gamma \leftarrow \gamma \cdot e^{r \cdot t}$ (growth rate r).
   - Stepwise: Increment γ at specific epochs.

Step 2.3: Gradient Descent

- Update learnable parameters τ, α using their gradients.

3. End-of-Epoch Hard Thresholding:

$$\hat{A}_{ij} = \begin{cases} A_{ij}, & \text{if } A_{ij} > \tau, \\ 0, & \text{otherwise.} \end{cases}$$

4. Repeat Until Convergence:

- Monitor the Lagrangian and sparsity level to ensure convergence.

Variational Analysis of Sparse Graph Learning with Hard Thresholding

We have the variational objective:

$$\mathcal{L}(\hat{A}, \tau, \alpha) = I(\hat{A}; A) - \beta \cdot I(G'; Y) + \mu \cdot \text{Penalty}(\hat{A}, \tau, \alpha), \tag{4.8}$$

where the penalty is defined as:

$$\text{Penalty}(\hat{A}, \tau, \alpha) = \text{ReLU}\left(\sum_{i,j} \sigma\left(\gamma(A_{ij} - \tau)\right) - \alpha n^2\right), \tag{4.9}$$

and the sigmoid function $\sigma(x)$ is:

$$\sigma(x) = \frac{1}{1 + e^{-x}}. \tag{4.10}$$

Constraint Reformulation

We define an auxiliary soft constraint function:

$$\phi(\hat{A}, \tau, \alpha) = \sum_{i,j} \sigma\left(\gamma(A_{ij} - \tau)\right) - \alpha n^2. \tag{4.11}$$

Then the objective becomes:

$$\mathcal{L} = I(\hat{A}; A) - \beta I(G'; Y) + \mu \cdot \max(0, \phi(\hat{A}, \tau, \alpha)). \tag{4.12}$$

Derivatives of the Lagrangian

We compute the derivatives with respect to $\hat{A}_{ij}, \tau, \alpha$.

Gradient w.r.t. $\hat{A}_{ij}$: Let $S = \sum_{i,j} \sigma(\gamma(A_{ij} - \tau))$, then:

$$\frac{\partial \mathcal{L}}{\partial \hat{A}_{ij}} = \frac{\partial I(\hat{A}; A)}{\partial \hat{A}_{ij}} - \beta \cdot \frac{\partial I(G'; Y)}{\partial \hat{A}_{ij}} + \mu \cdot \mathbb{I}_{[S > \alpha n^2]} \cdot \frac{\partial S}{\partial \hat{A}_{ij}}. \tag{4.13}$$

Using the chain rule:

$$\frac{\partial S}{\partial \hat{A}_{ij}} = \gamma \cdot \sigma(\gamma(A_{ij} - \tau)) \cdot (1 - \sigma(\gamma(A_{ij} - \tau))). \tag{4.14}$$

Gradient w.r.t. τ:

$$\frac{\partial \mathcal{L}}{\partial \tau} = -\mu \cdot \mathbb{I}_{[S > \alpha n^2]} \cdot \sum_{i,j} \gamma \cdot \sigma(\gamma(A_{ij} - \tau))(1 - \sigma(\gamma(A_{ij} - \tau))). \tag{4.15}$$

Gradient w.r.t. α:

$$\frac{\partial \mathcal{L}}{\partial \alpha} = -\mu \cdot \mathbb{I}_{[S > \alpha n^2]} \cdot n^2. \tag{4.16}$$

Scheduling of γ

The scheduling rule for γ_t over training epochs t is defined in phases:

$$\gamma_t = \begin{cases} 0.1, & t < \frac{T}{3} \\ 1, & \frac{T}{3} \leq t < \frac{2T}{3} \\ 10, & t \geq \frac{2T}{3} \end{cases} \tag{4.17}$$

This gradually sharpens the sigmoid $\sigma(\gamma(A_{ij} - \tau))$, approximating a step function.

End-of-Epoch Hard Thresholding

At the end of each epoch, the adjacency matrix is thresholded as:

$$\hat{A}_{ij} = \begin{cases} A_{ij}, & A_{ij} > \tau \\ 0, & \text{otherwise} \end{cases} \tag{4.18}$$

Gradient Descent Updates

Update the learnable parameters using their gradients:

$$\tau \leftarrow \tau - \eta_\tau \cdot \frac{\partial \mathcal{L}}{\partial \tau}, \tag{4.19}$$

$$\alpha \leftarrow \alpha - \eta_\alpha \cdot \frac{\partial \mathcal{L}}{\partial \alpha}. \tag{4.20}$$

4.3.6 Architectural Information

To estimate the mutual information $I(\hat{A}; A)$ between the learned adjacency matrix $\hat{A}$ and the original matrix A, a neural estimation approach based on the Donsker–Varadhan variational representation of Kullback–Leibler divergence is employed. A neural network, referred to as the Mutual Information Neural Estimator (MINE), is used to approximate a variational function $T(x, y)$ capable of

distinguishing samples drawn from the joint distribution $P_{\hat{A},A}$ and the product of marginals $P_{\hat{A}} \otimes P_A$.

$I(\hat{A}; A)$ Estimation Network Architecture

The MINE model is implemented as a three-layer feedforward neural network composed of fully connected layers and non-linear activations. The overall architecture is described as follows:

- **Input Layer:**

  The input to the network consists of concatenated edge feature pairs from $\hat{A}$ and A, with a total dimensionality denoted as `input_size`.

- **First Hidden Layer:**

  A linear transformation projects the input into a `hidden_size`-dimensional space. An Exponential Linear Unit (ELU) activation function is applied to the output:

$$h_1 = \text{ELU}(W_1 x + b_1)$$

- **Second Hidden Layer:**

  The hidden representation is further transformed through another linear layer, followed by a second ELU activation:

$$h_2 = \text{ELU}(W_2 h_1 + b_2)$$

- **Output Layer:**

  The final scalar output $T(x, y)$ is obtained via a final linear transformation:

$$T(x, y) = W_3 h_2 + b_3$$

This output is utilized in the Donsker–Varadhan lower bound to estimate mutual information:

$$\widehat{I}(\hat{A}; A) = \mathbb{E}_{P_{\hat{A},A}}[T(x, y)] - \log \mathbb{E}_{P_{\hat{A}} \otimes P_A}[e^{T(x,y)}]$$

Weight Initialization

All weight matrices W_i are initialized using a normal distribution with mean zero and standard deviation 0.02:

$$W_i \sim \mathcal{N}(0, 0.02)$$

Bias terms b_i are initialized to zero. This initialization scheme is adopted to ensure stable gradient flow and effective early-stage learning.

Activation Function

The ELU (Exponential Linear Unit) activation function is employed due to its ability to mitigate the "dying neuron" problem and provide smooth gradients, particularly in regions with negative input values. This facilitates effective training of the mutual information estimator.

4.3.7 Graph Mutual Information Estimator (GraphMINE) Architecture

To estimate the mutual information $I(G'; Y)$ between the learned graph structure G' and the task labels Y, a Graph Mutual Information Estimator (GraphMINE) model is utilized. This model combines graph convolution operations with deep fully connected layers to produce a discriminative representation of the graph conditioned on label structure.

Network Architecture

The GraphMINE model is composed of graph convolutional layers followed by dense feedforward layers:

- **Input:**

  The model takes node features x, edge indices `edge_index`, edge weights

`edge_weight`, and binary label information y.

– **Graph Convolutional Layers:**

Two stacked GCN layers are applied sequentially:

$$h_1 = \text{ELU}(\text{GCNConv}_1(x, \texttt{edge_index}, \texttt{edge_weight}))$$

$$h_2 = \text{ELU}(\text{GCNConv}_2(h_1, \texttt{edge_index}, \texttt{edge_weight}))$$

– **Graph-Level Pooling:**

The node-level features are averaged to create a global graph representation:

$$h_{\text{graph}} = \frac{1}{n} \sum_{i=1}^{n} h_2[i]$$

– **Fully Connected Layers:**

Three fully connected layers refine the graph-level embedding:

$$z_1 = \text{ELU}(W_1 h_{\text{graph}} + b_1)$$

$$z_2 = \text{ELU}(W_2 z_1 + b_2)$$

$$T(g, y) = W_3 z_2 + b_3$$

The scalar output $T(g, y)$ is then used within the mutual information lower bound:

$$\widehat{I}(G'; Y) = \mathbb{E}_{P_{G',Y}}[T(g, y)] - \log \mathbb{E}_{P_{G'} \otimes P_Y}[e^{T(g,y)}]$$

Weight Initialization

All linear layers are initialized with weights drawn from a normal distribution $\mathcal{N}(0, 0.02)$, and all biases are initialized to zero.

Activation Function

The ELU activation function is used after each intermediate transformation to improve gradient flow and promote stable training.

Chapter 5

Results

5.1 IQ Prediction

The prediction performance of the first proposed IQ prediction model, which uses the multivariate time series regression (MTSR) approach via HYDRA, was evaluated by systematically varying the hyperparameters N_g and N_k, which determine the granularity of spatial and temporal features respectively. Table 5.1 presents the correlation coefficients achieved for different combinations of these parameters. The configuration with $N_g = 128$ and $N_k = 16$ yielded the highest correlation value of $r = 0.69$, and was therefore adopted in the final implementation.

The extracted HYDRA features were then used with several standard regression models. As shown in Table 5.2, ridge regression outperformed other models such as XGBoost, Random Forest, and SVR, achieving the highest correlation of $r = 0.69$. Consequently, ridge regression was selected for use in the final version of the HYDRA-based IQ prediction model.

Table 5.1: Correlation Using HYDRA-based Model for Different N_g and N_k

Correlation (r)	$N_k = 8$	$N_k = 16$	$N_k = 32$	$N_k = 64$
$N_g = 16$	0.615	0.53	0.51	0.54
$N_g = 32$	0.54	0.49	0.54	0.55
$N_g = 64$	0.62	0.56	0.6	0.61
$N_g = 128$	0.52	**0.69**	0.61	0.59

Table 5.2: Correlation Using HYDRA Features for Different Regressors

Regressor	**Correlation (r)**
XGBoost	0.40
Random Forest	0.47
SVR	0.50
Ridge	**0.69**

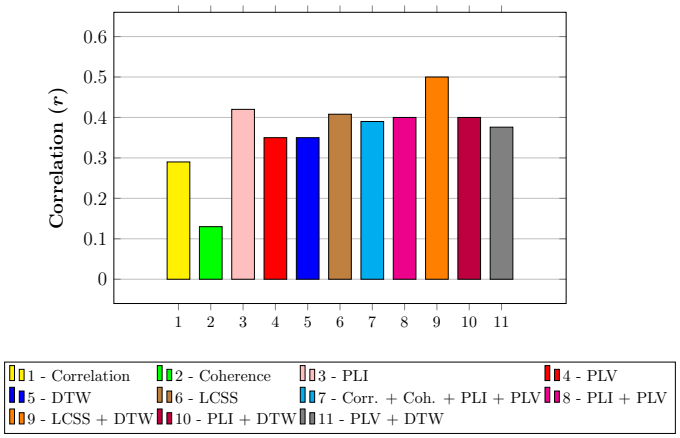


Figure 5.1: Correlation using similarity-based model for different similarity measures and their combinations.

To provide a comparative baseline, a standard CNN regressor was trained directly on the 200-dimensional BOLD time series input. Despite optimization, the CNN model could only achieve a correlation of $r = 0.40$, reinforcing the efficacy of the MTSR formulation via HYDRA in capturing the temporal structure of rs-fMRI data.

5.1.1 Performance of Similarity-based CNN Prediction Model

The second proposed model utilized similarity-based representations derived from time series measures and fed them into a CNN architecture. Initially, individual similarity measures were evaluated for their predictive power. Subsequently, combinations of these similarity matrices were considered to leverage complementary information. As shown in Figure 5.1, combinations such as DTW + LCSS yielded the best performance with $r = 0.50$, indicating the advantage of using unconventional temporal similarity measures in modeling brain relational dynamics.

Table 5.3: Correlation Using Hybrid Model for Different Regressors

Regressor	**Correlation (r)**
XGBoost	0.46
Random Forest	0.51
SVR	0.47
Ridge	**0.75**

5.1.2 Performance of Hybrid IQ Prediction Model

The final model integrated both the HYDRA-based and similarity-based features into a hybrid representation using an intermediate-level fusion strategy. Various regression models were evaluated using this combined feature vector. As summarized in Table 5.3, ridge regression again demonstrated superior performance, achieving a state-of-the-art correlation of $r = 0.75$.

The consistent improvement in prediction performance from individual models to the hybrid model supports the effectiveness of feature fusion. The results validate the hypothesis that combining temporal and relational information extracted from rs-fMRI significantly enhances IQ prediction capabilities.

The performance of various methods for predicting intelligence scores is summarized in the below, which compares previously established techniques with our proposed models, including HYDRA, Similarity-based modeling, and the Hybrid model. The *Lohmann2021* method achieved a correlation of approximately 0.4, utilizing traditional statistical approaches [1]. The *Tong2022* model improved the correlation to 0.5 by incorporating advanced machine learning techniques [2], while *Dryburgh2020* and *Gal2022* achieved correlations of 0.3 and 0.4, respectively, reflecting early-stage explorations of intelligence prediction using simpler modeling strategies [3, 4]. More recently, *tokuhiro2023* leveraged deep learning approaches, achieving a correlation of 0.6 [48]. In comparison, our proposed HYDRA model (Heterogeneity through Discriminative Regression Analysis) demonstrated a superior correlation of 0.7 by capturing variability across subjects and tasks through discriminative modeling. The Similarity-based method, designed to incorporate relational metrics between subjects in the feature space, achieved a correlation of 0.65, emphasizing the importance of relational information in

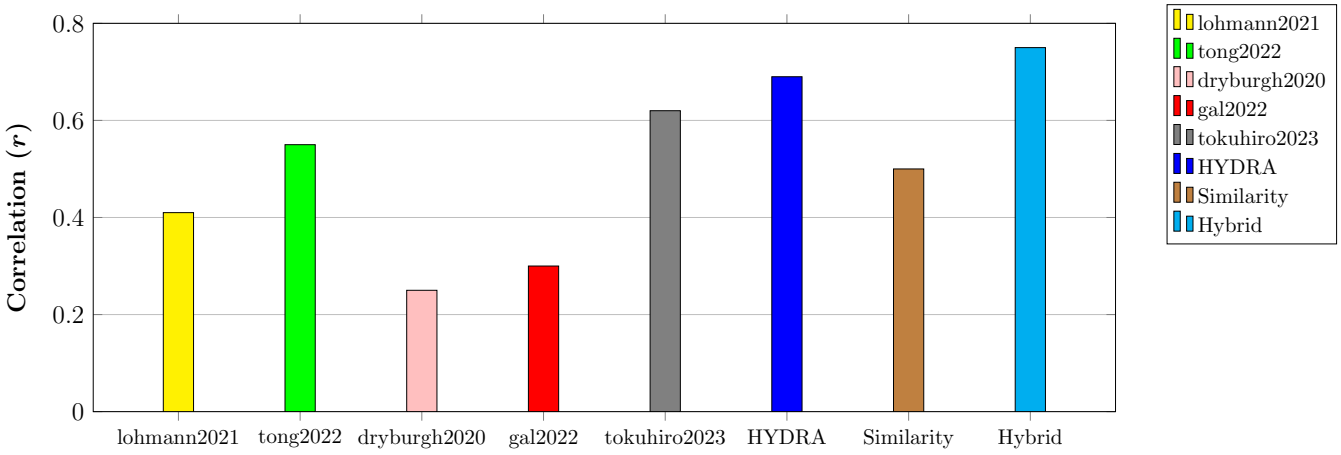


Figure 5.2: Correlation values obtained using different intelligence score prediction methods. References: [1], [2], [3], [4], [48].

predictive tasks. Finally, the Hybrid model, which combines the strengths of HYDRA and Similarity-based approaches, achieved the highest correlation of 0.75, demonstrating the efficacy of integrating discriminative and relational methodologies. These results underscore the superiority of our proposed methods in enhancing prediction accuracy compared to existing techniques and highlight the potential of combining multiple modeling perspectives for more robust and interpretable intelligence score predictions.

5.2 Task Activation Map prediction from rs-fMRI

5.2.1 Results

This section presents the brain surface heatmaps corresponding to each selected contrast, visualizing the spatial distribution of activation associated with their respective COPEs. These heatmaps offer an intuitive depiction of task-related cortical activity across the brain, highlighting regions that exhibit significant signal changes under each contrast condition.

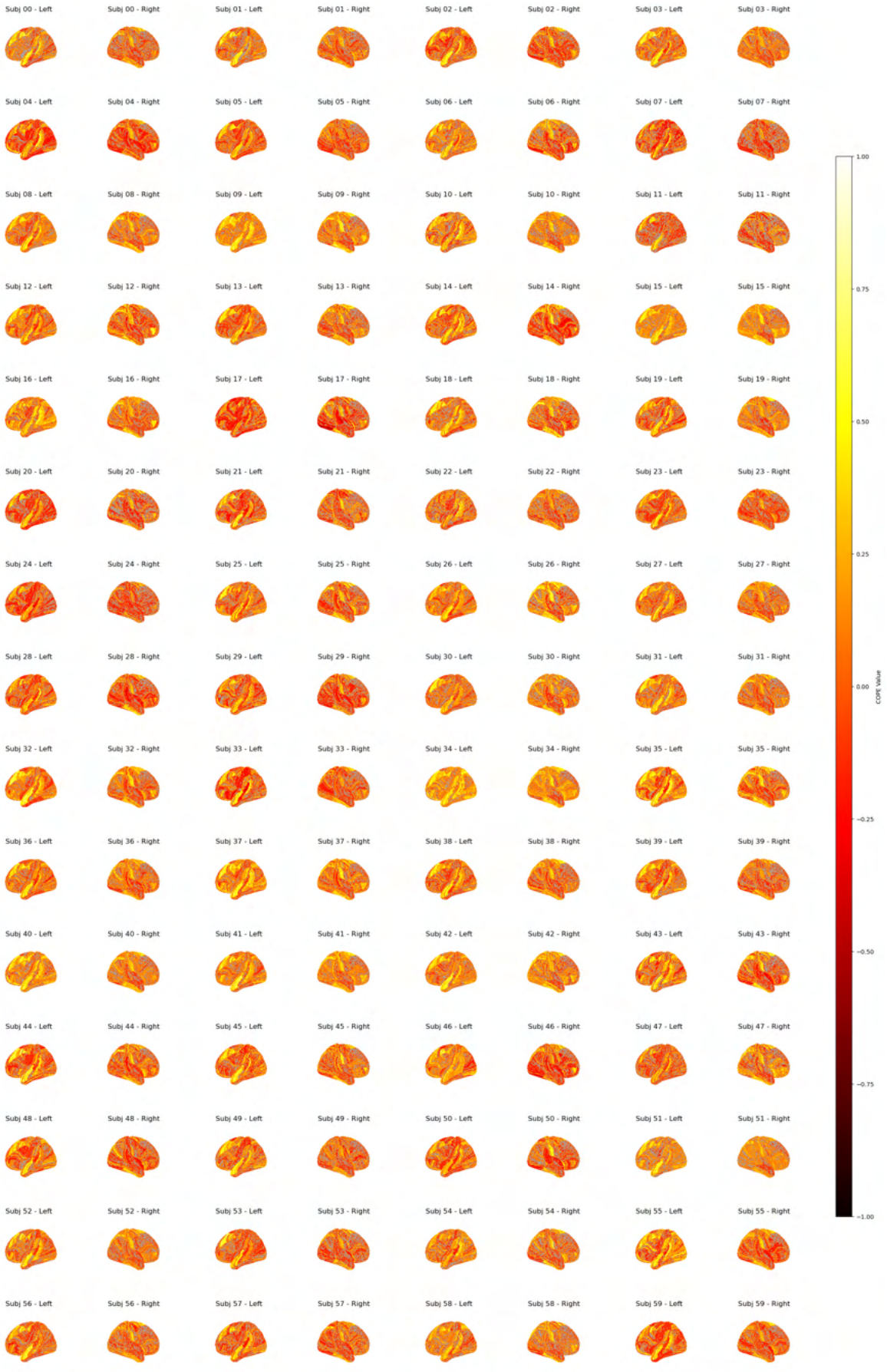


Figure 5.3: Actual Task Activation Map of Relational Contrast COPE 1.

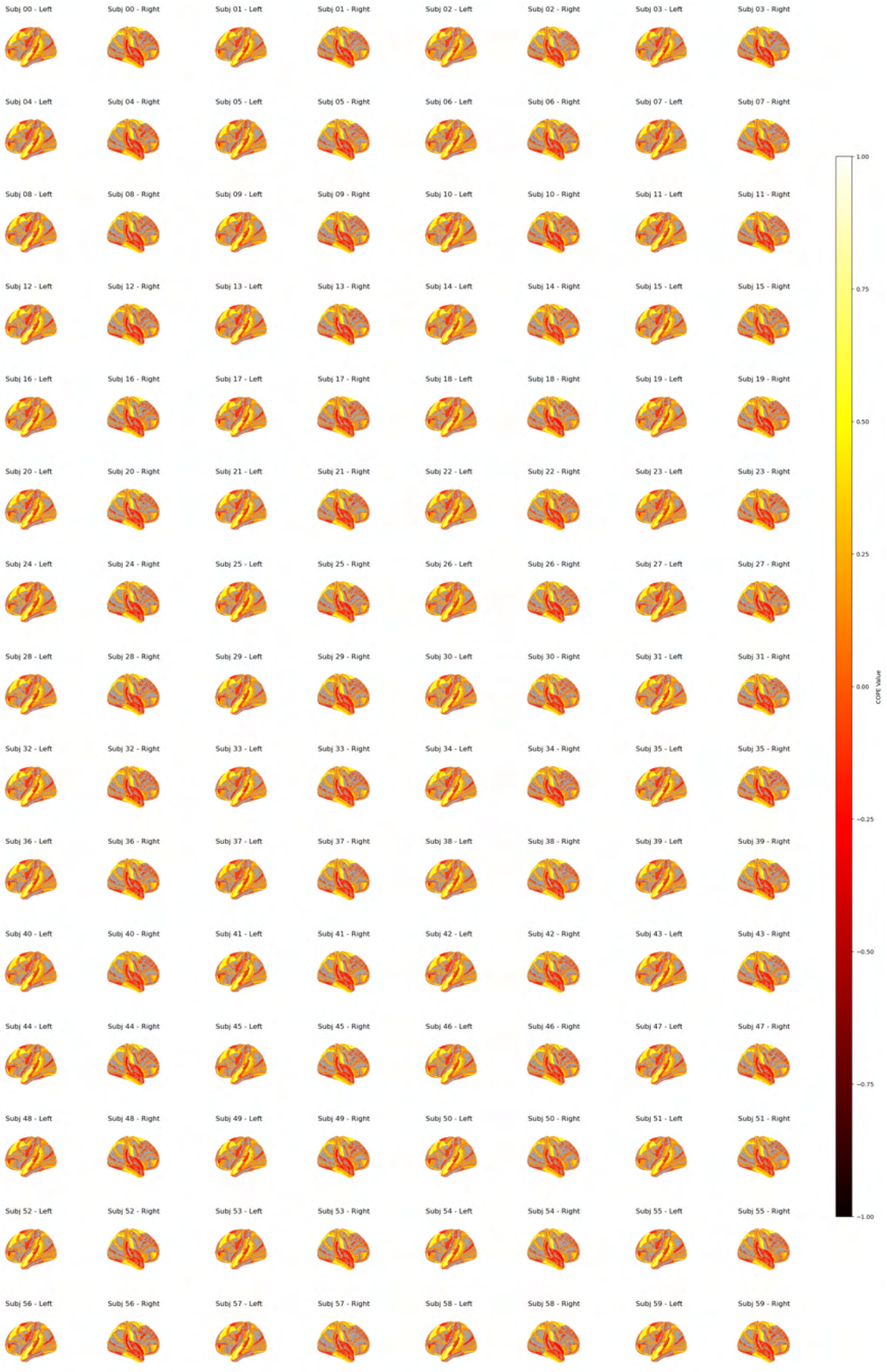


Figure 5.4: Predicted Task Activation Map of Relational Contrast COPE 1 using Catch22 features and Linear Regression.

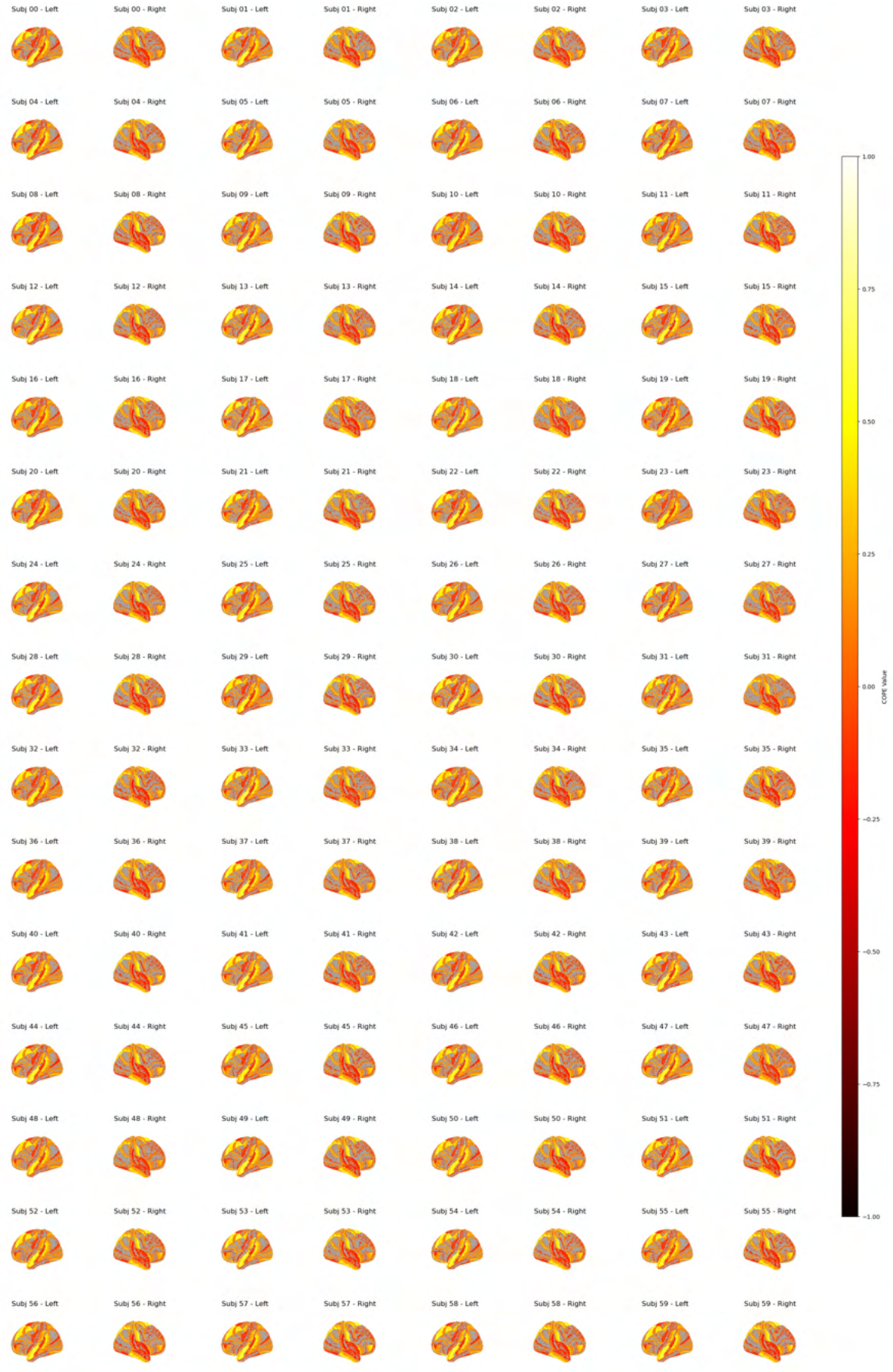


Figure 5.5: Predicted Task Activation Map of Relational Contrast COPE 1 using Summary Features.

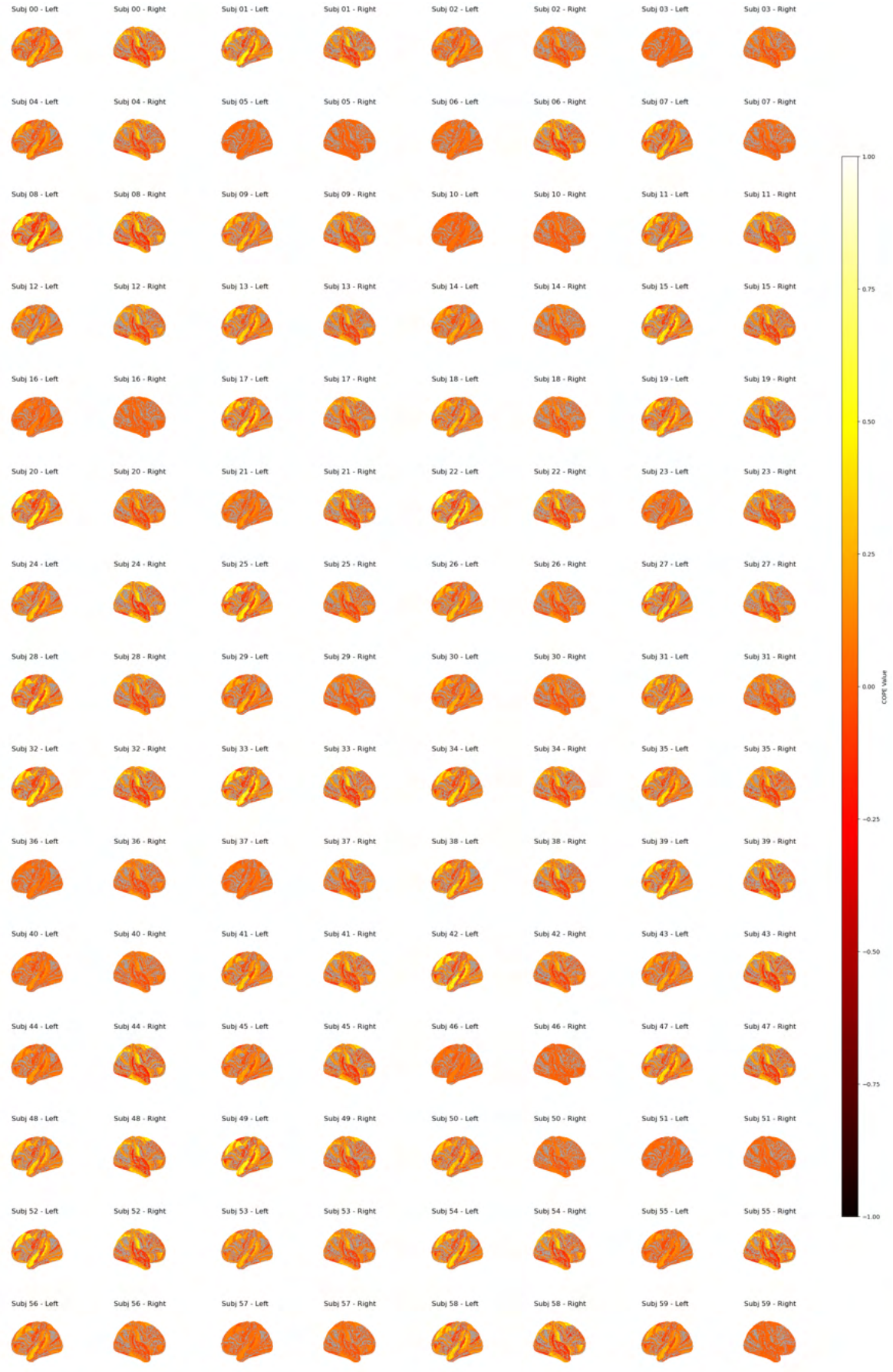


Figure 5.6: Predicted Task Activation Map of Relational Contrast COPE 1 using DrCIF.



Figure 5.7: Actual Task Activation Map of Relational Contrast COPE 3



Figure 5.8: Predicted Task Activation Map of Relational Contrast COPE 3 using Summary Features.



Figure 5.9: Predicted Task Activation Map of Relational Contrast COPE 3 using DrCIF.



Figure 5.10: Actual Task Activation Map of Language Contrast.



Figure 5.11: Predicted Task Activation Map of Language Contrast using Catch22 features and Linear Regression.



Figure 5.12: Predicted Task Activation Map of Language Contrast using DrCIF.

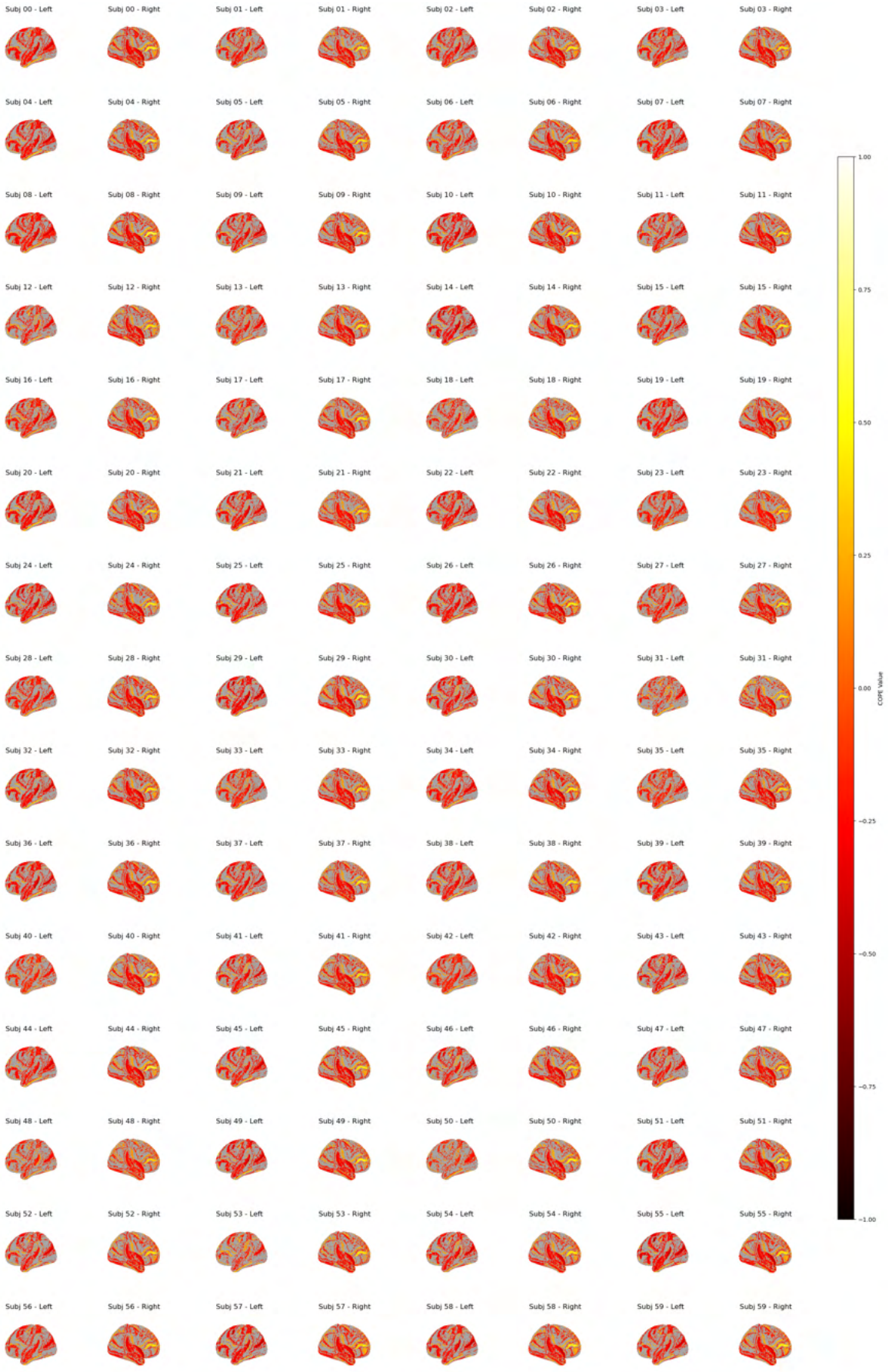


Figure 5.13: Predicted Task Activation Map of Language Contrast using Summary features.

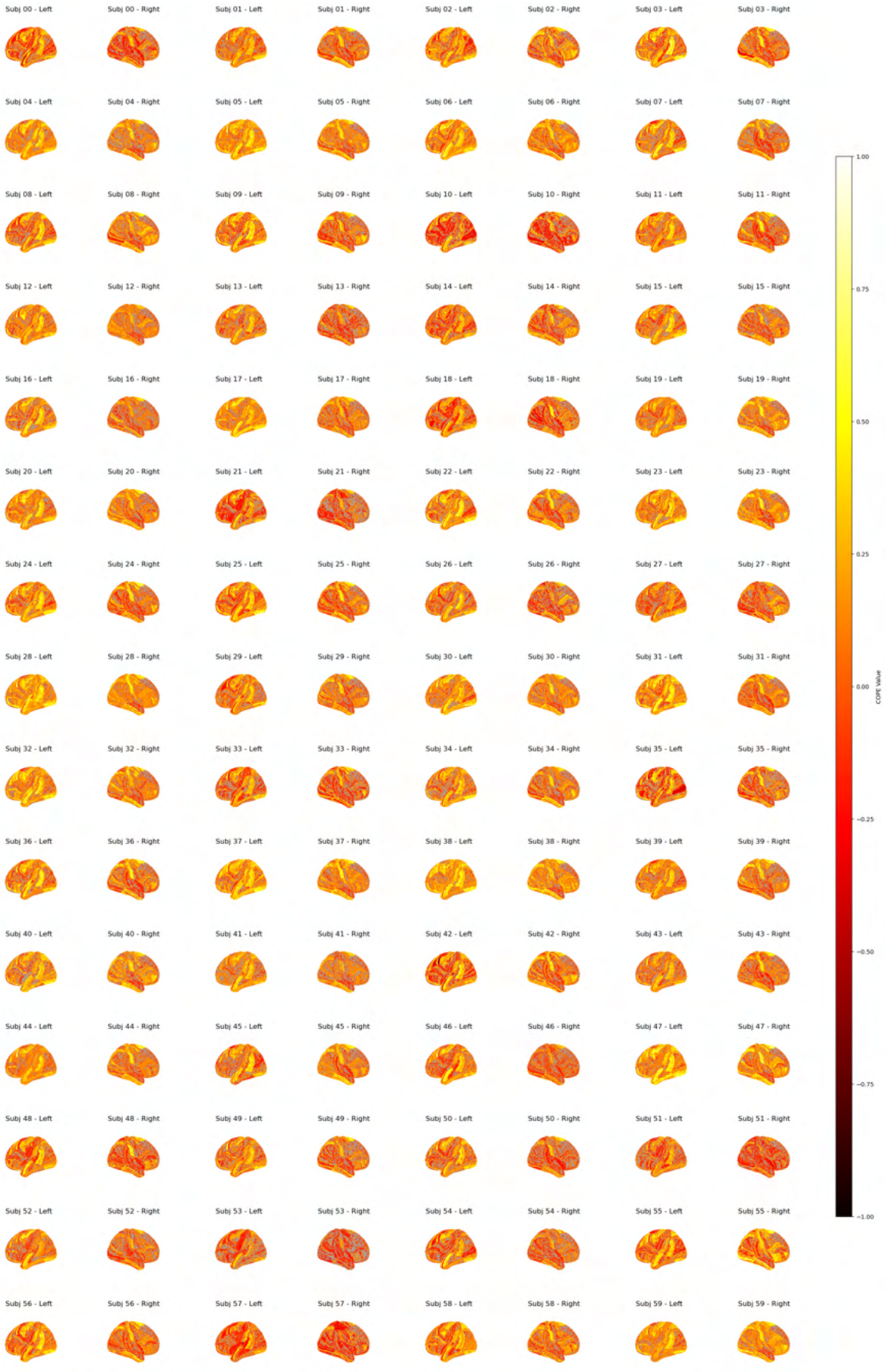


Figure 5.14: Actual Task Activation Map of Social Contrast.

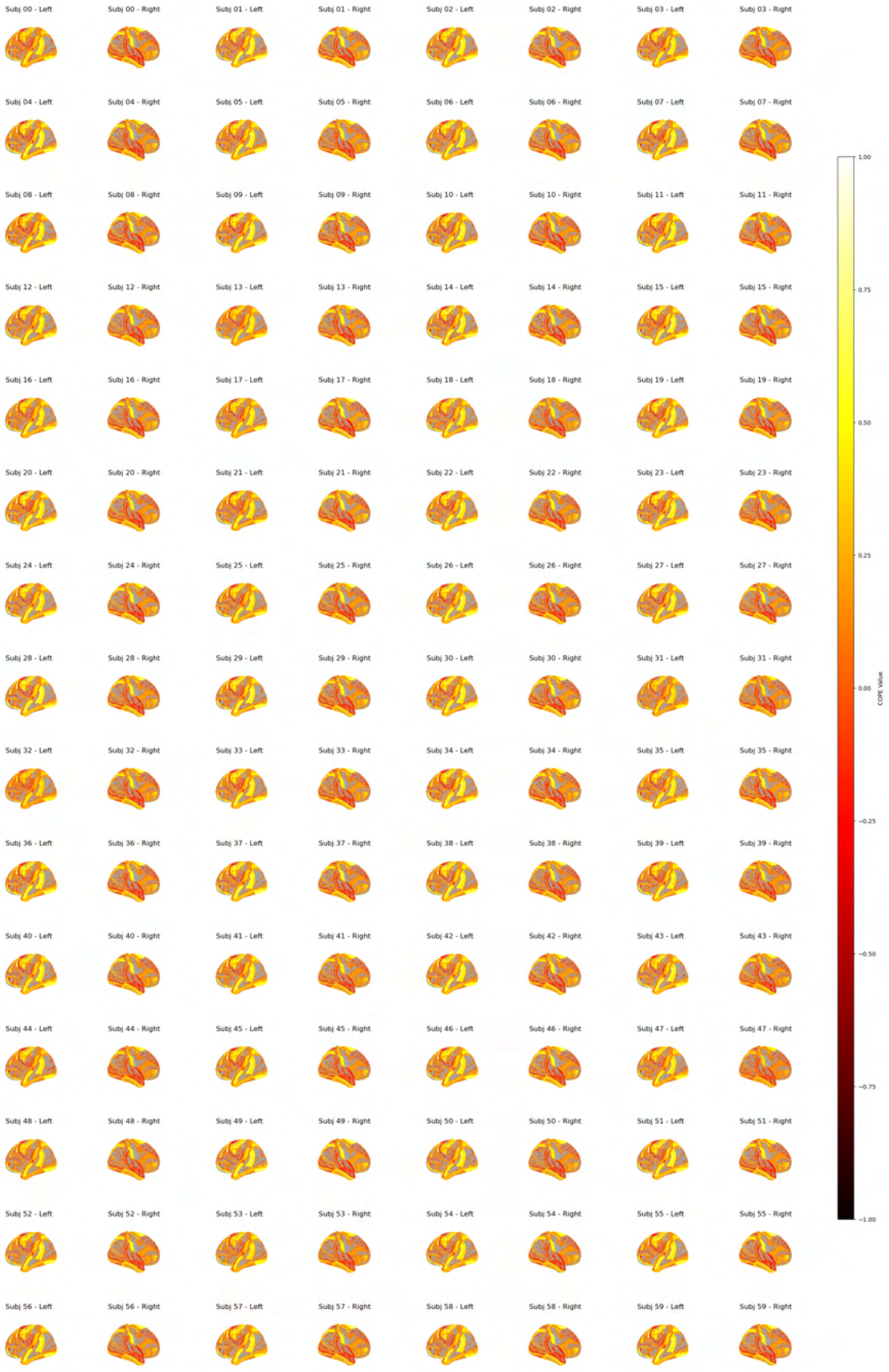


Figure 5.15: Predicted Task Activation Map of Social Contrast using Summary features.

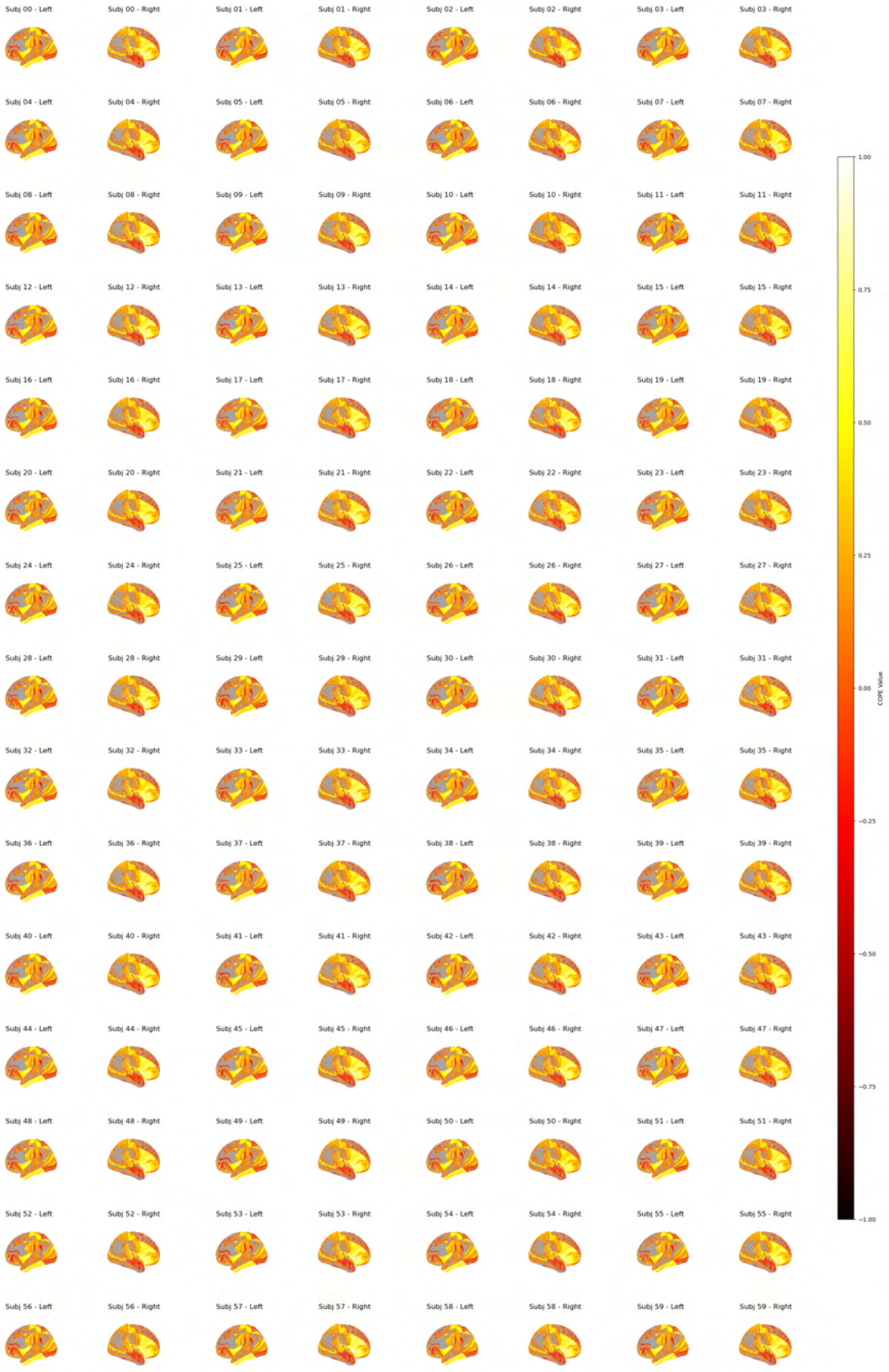


Figure 5.16: Predicted Task Activation Map of Social Contrast using catch22 features and Linear Regression.

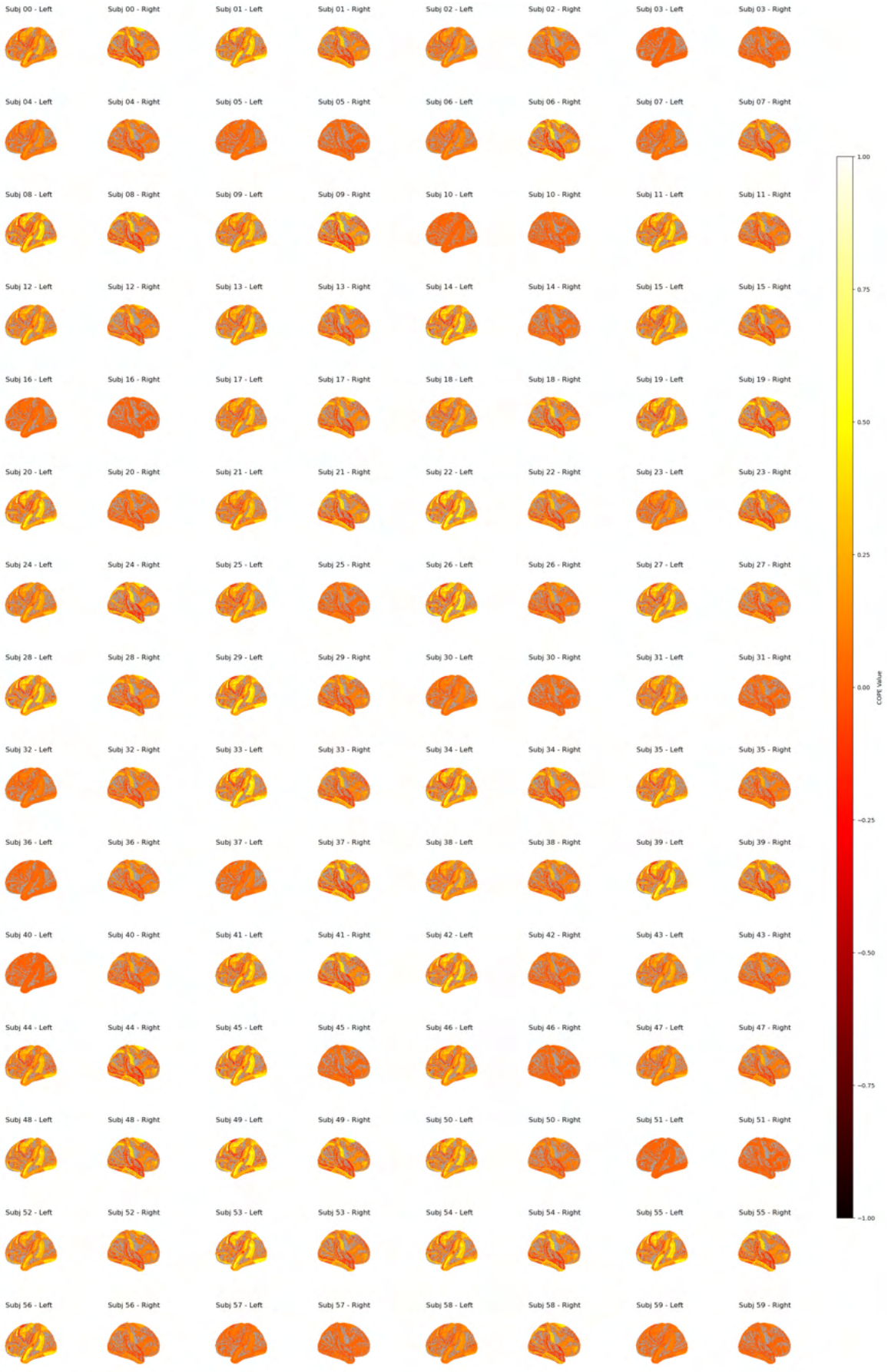


Figure 5.17: Predicted Task Activation Map of Social Constrast using DrCIF.

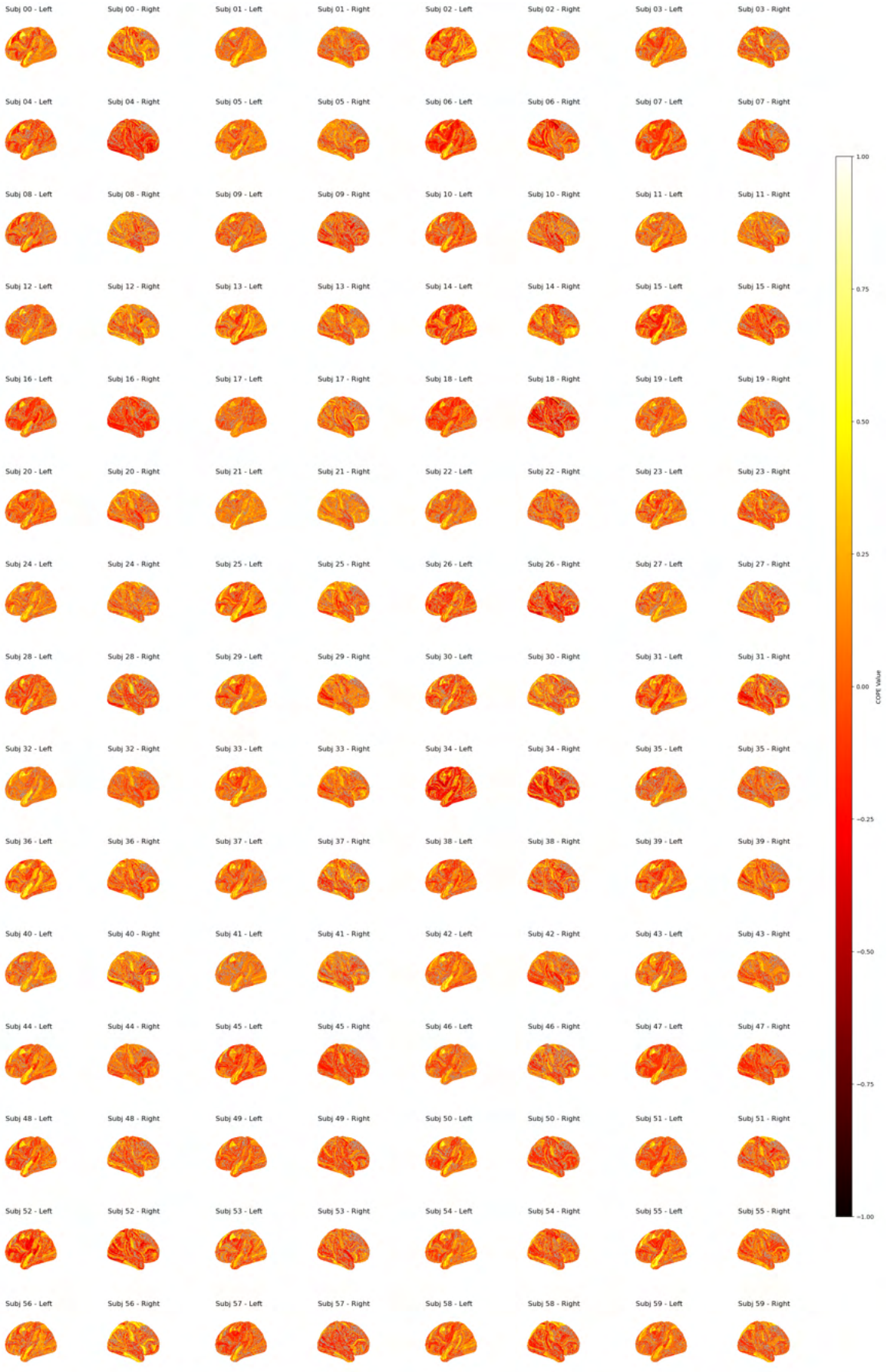


Figure 5.18: Actual Task Activation Map of Working Memory Contrast.

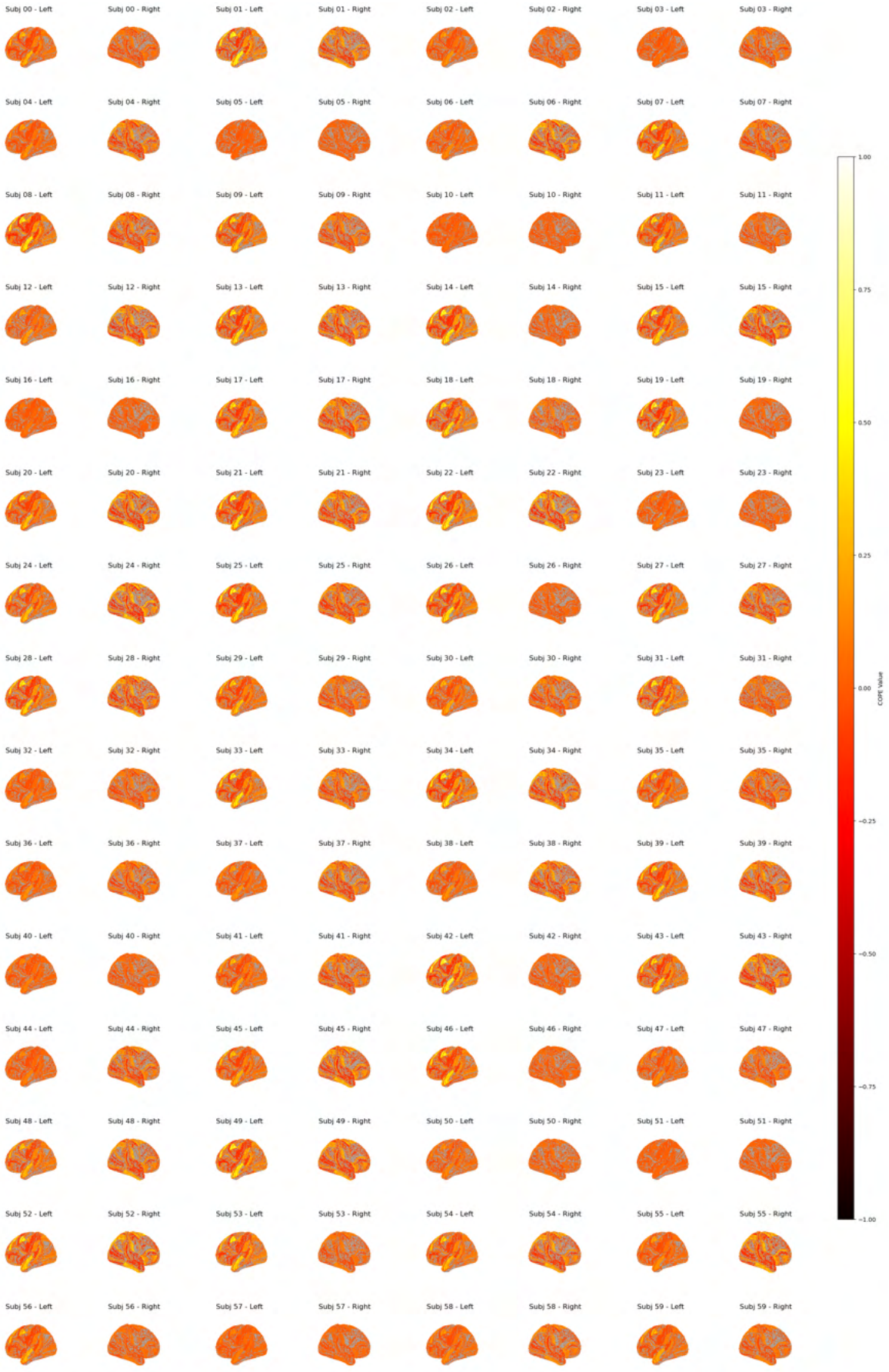


Figure 5.19: Predicted Task Activation Map of Working Memory Contrast using Dr-CIF.

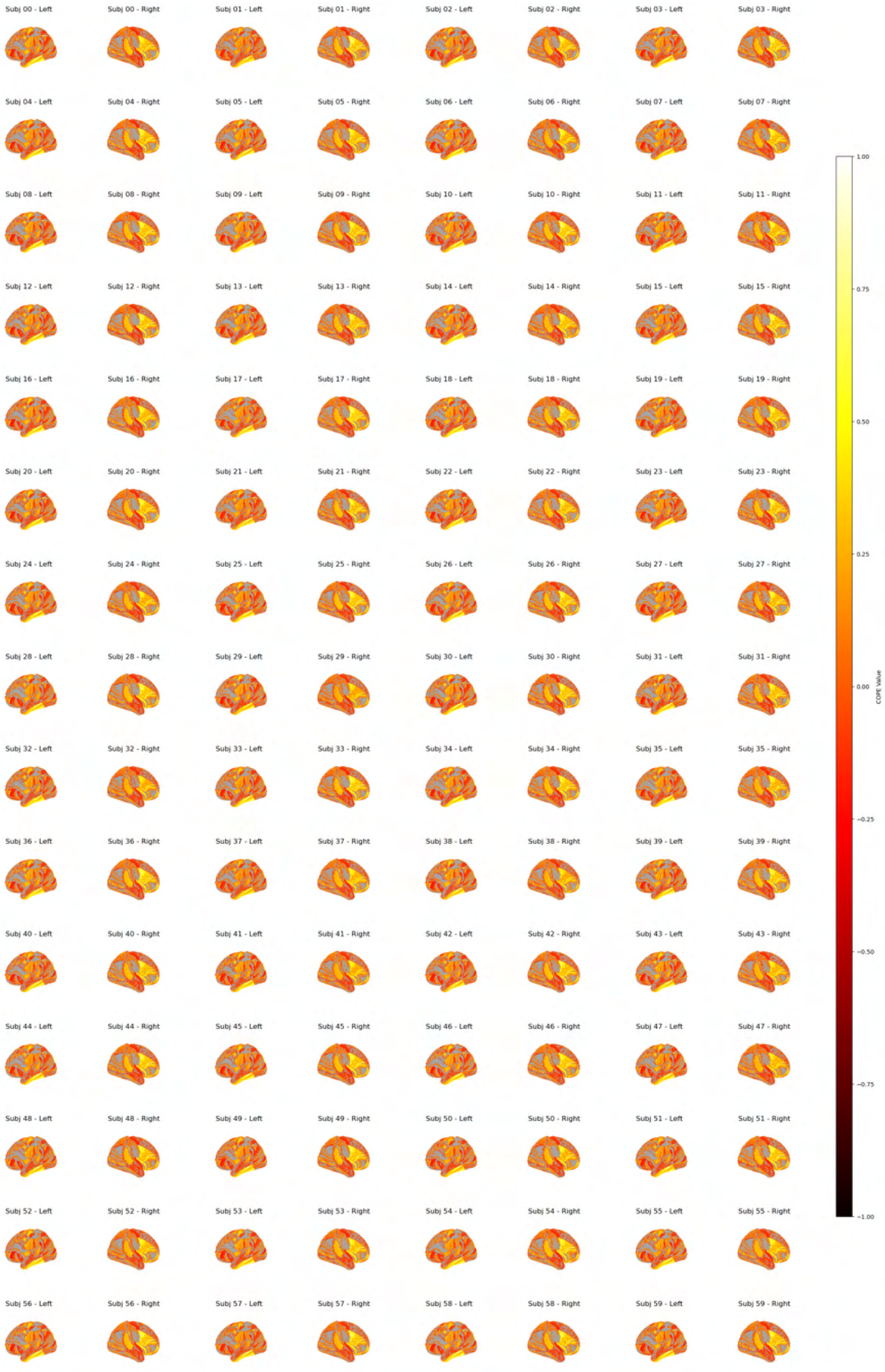


Figure 5.20: Predicted Task Activation Map of Working Memory Contrast using catch22 features and Linear Regression.

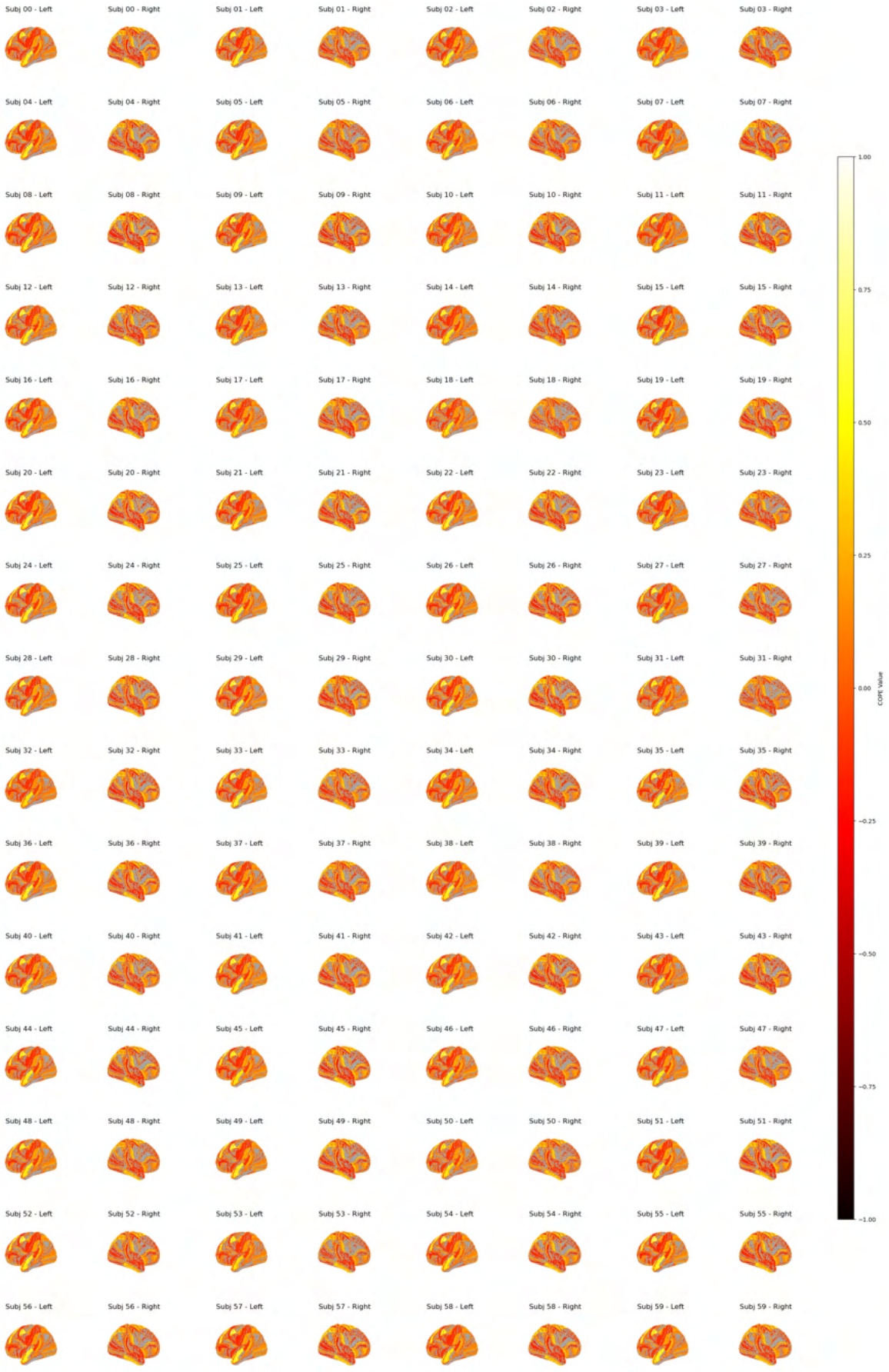


Figure 5.21: Predicted Task Activation Map of Working Memory Contrast using Summary features.

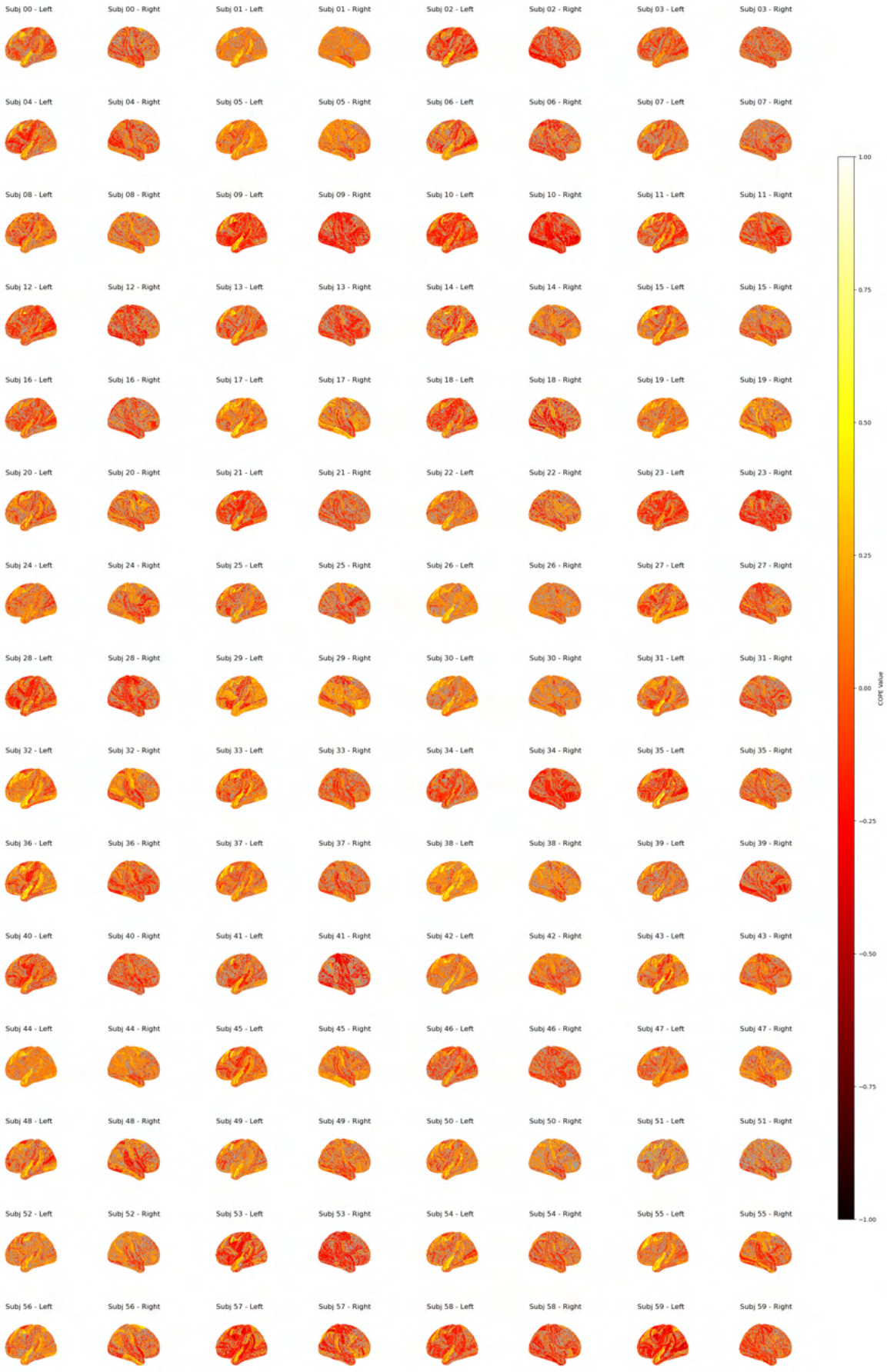


Figure 5.22: Actual Task Activation Map of Emotional Contrast.



Figure 5.23: Predicted Task Activation Map of Emotional Contrast using DrCIF.

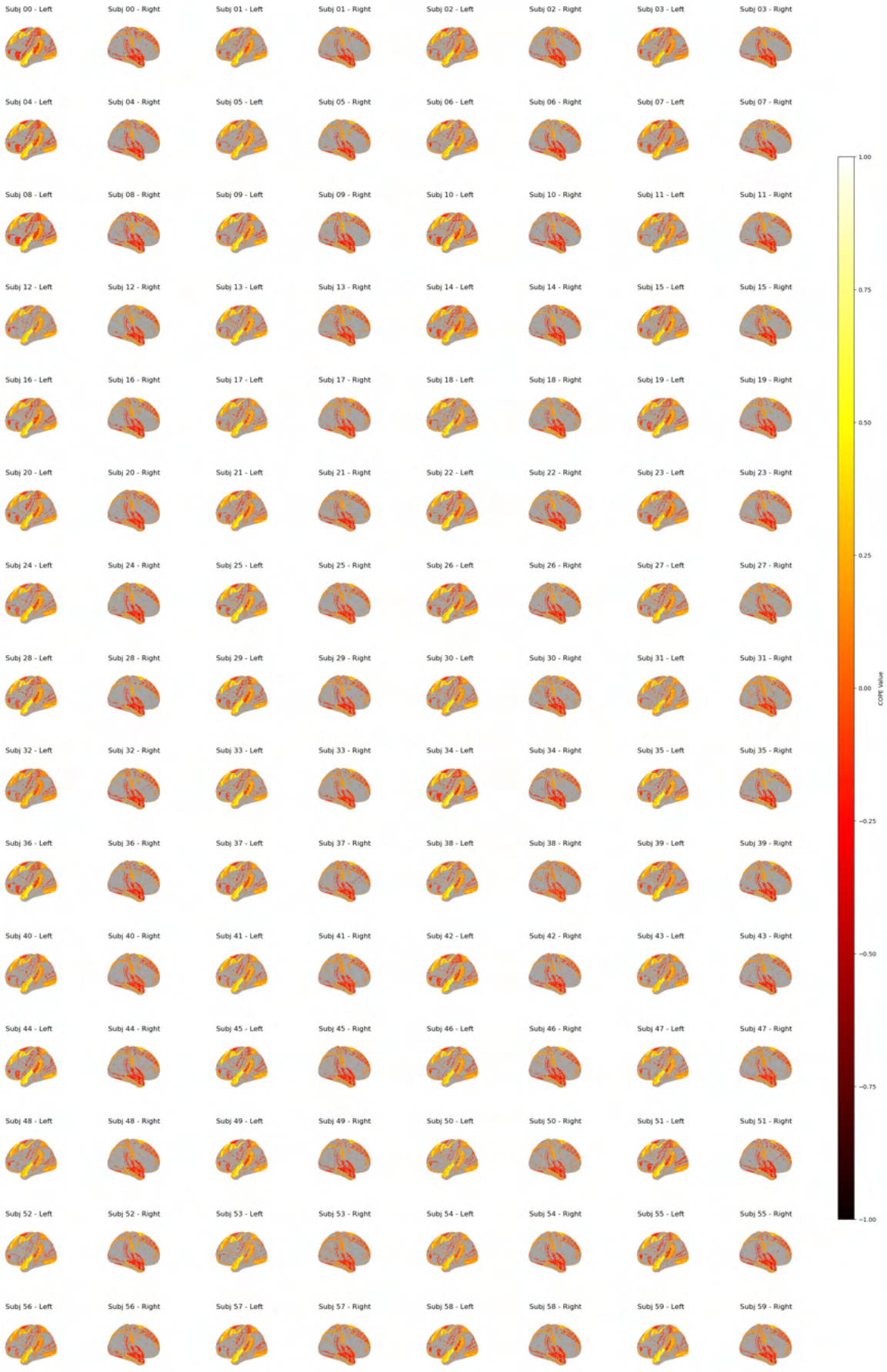


Figure 5.24: Predicted Task Activation Map of Emotional Contrast using Summary Features.

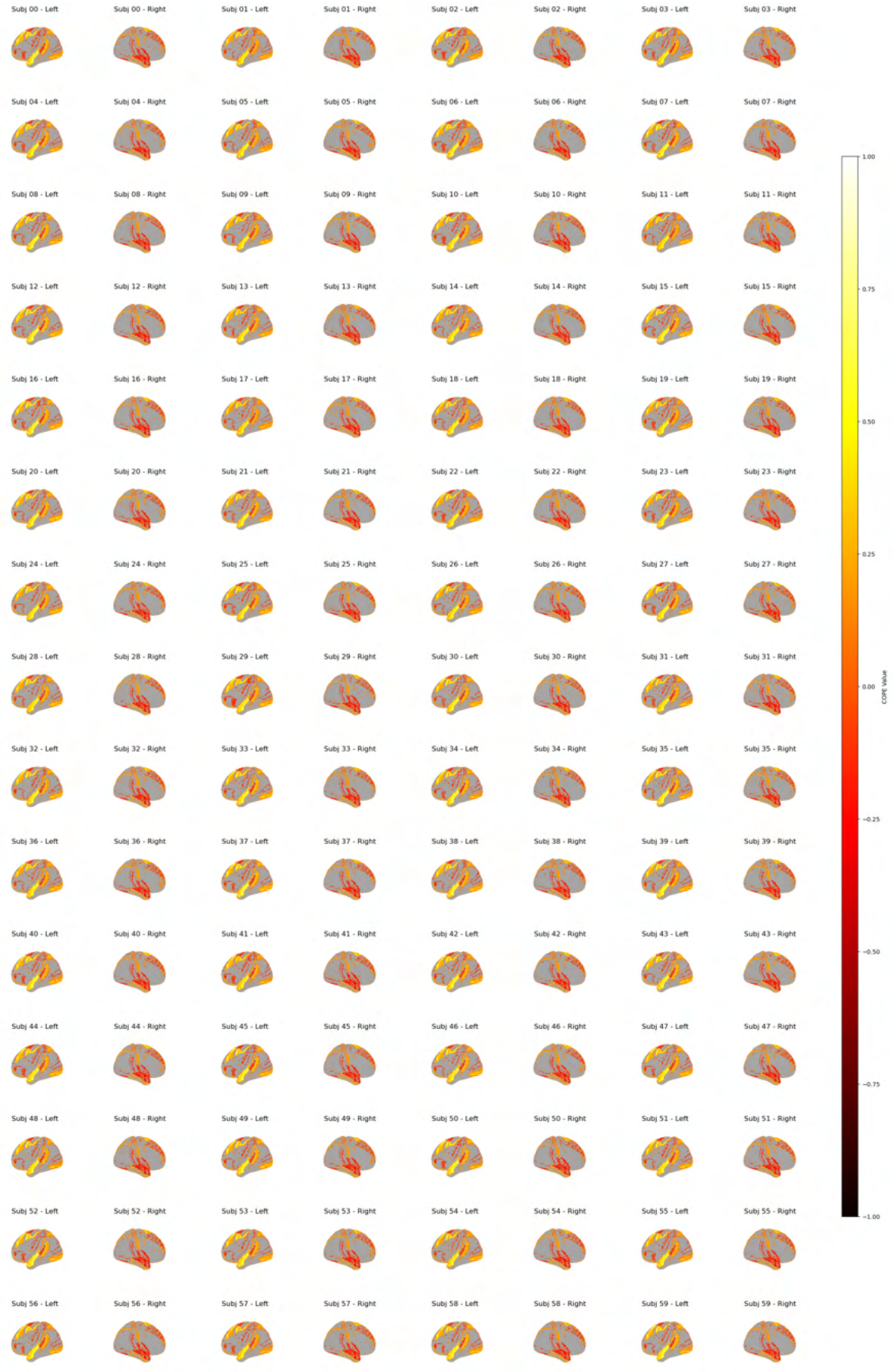


Figure 5.25: Predicted Task Activation Map of Emotional Contrast using catc22 features and Linear Regression.



Figure 5.26: Actual Task Activation Map of Gambling Contrast.

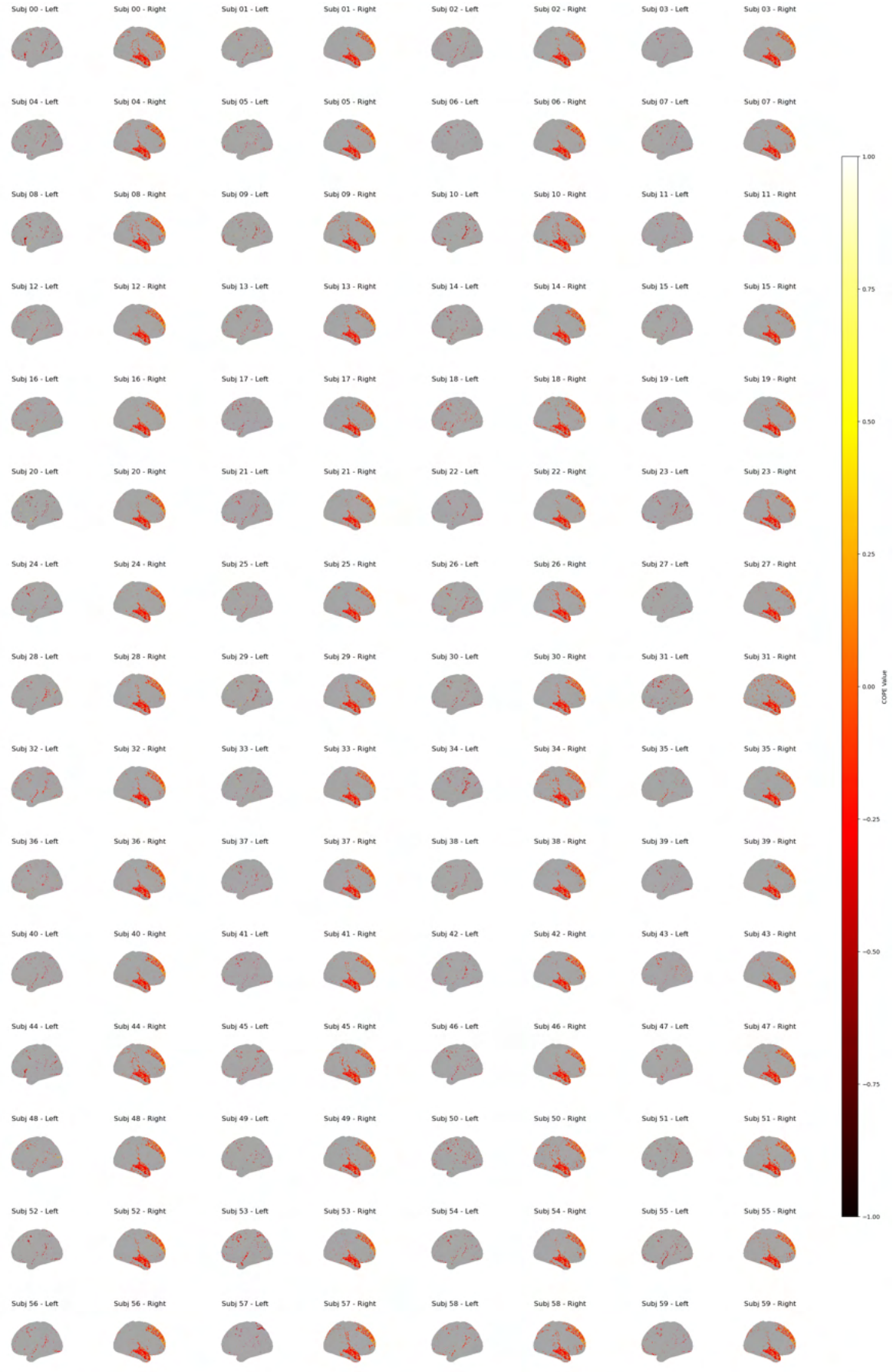


Figure 5.27: Predicted Task Activation Map of Gambling Contrast using Summary Features.

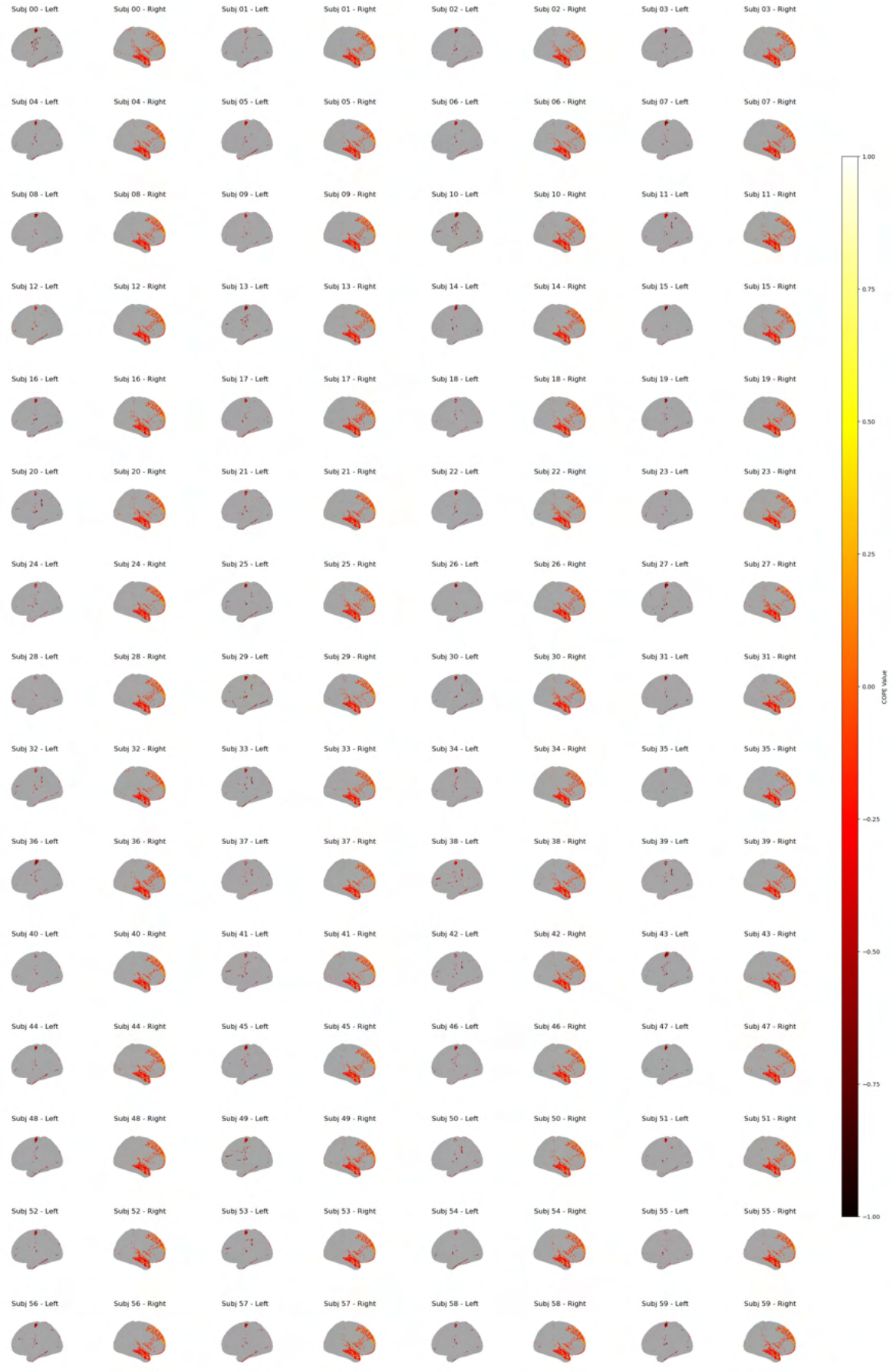


Figure 5.28: Predicted Task Activation Map of Gambling Contrast using catch22 and linear regression.

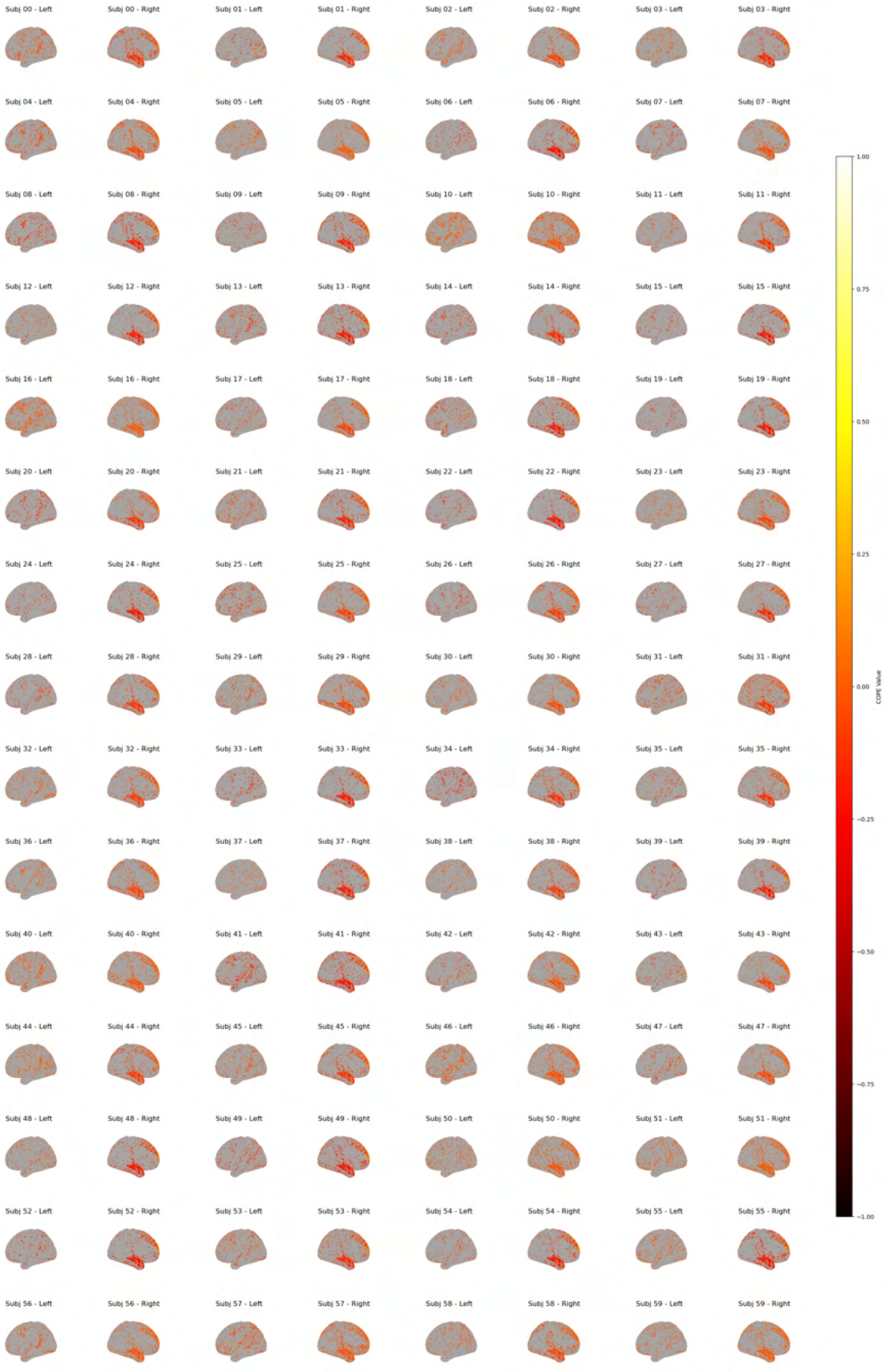


Figure 5.29: Predicted Task Activation Map of Gambling Contrast using DrCIF.

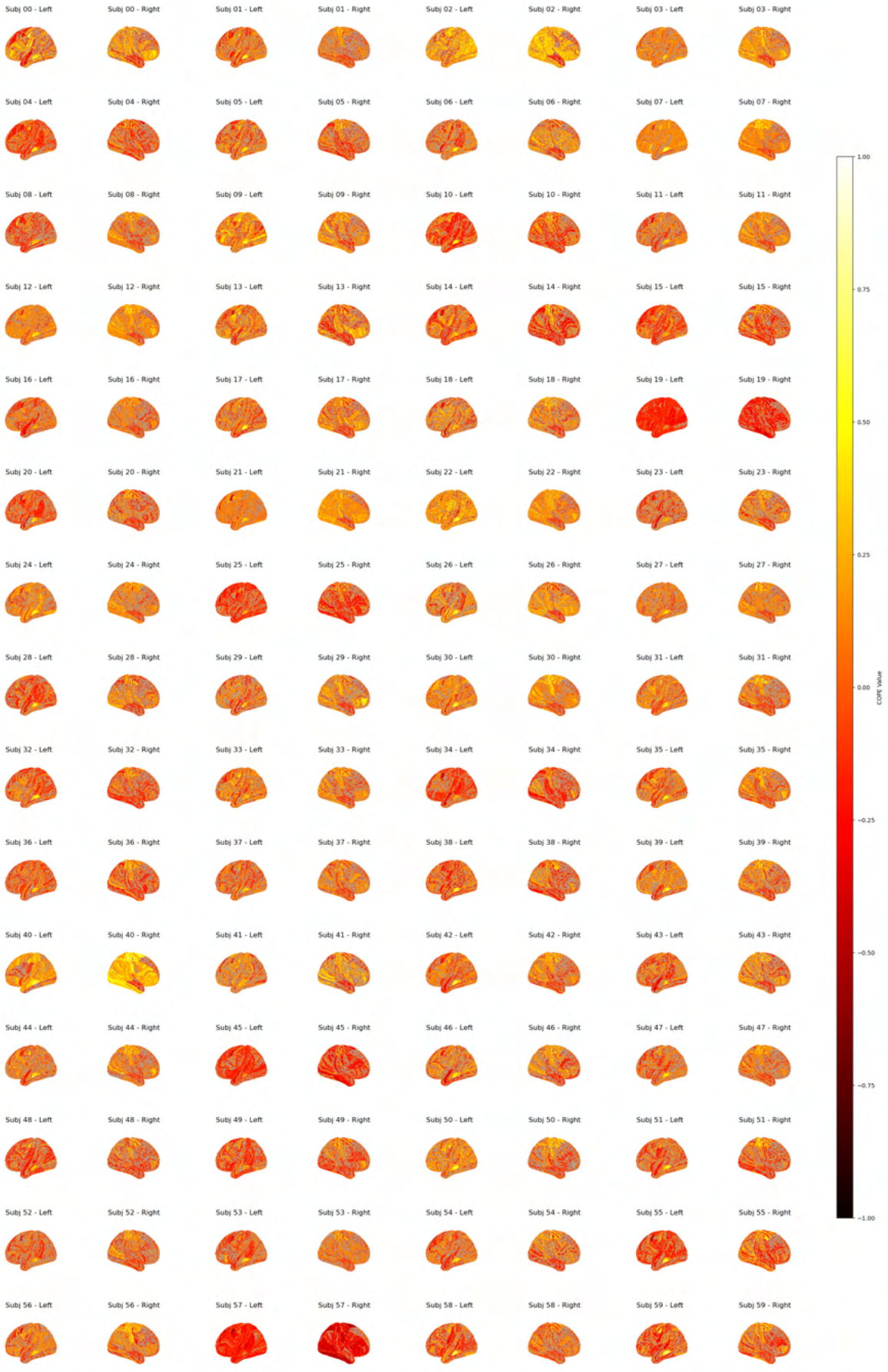


Figure 5.30: Actual Task Activation Map of Motor Contrast.



Figure 5.31: Predicted Task Activation Map of Motor Contrast using DrCIF.

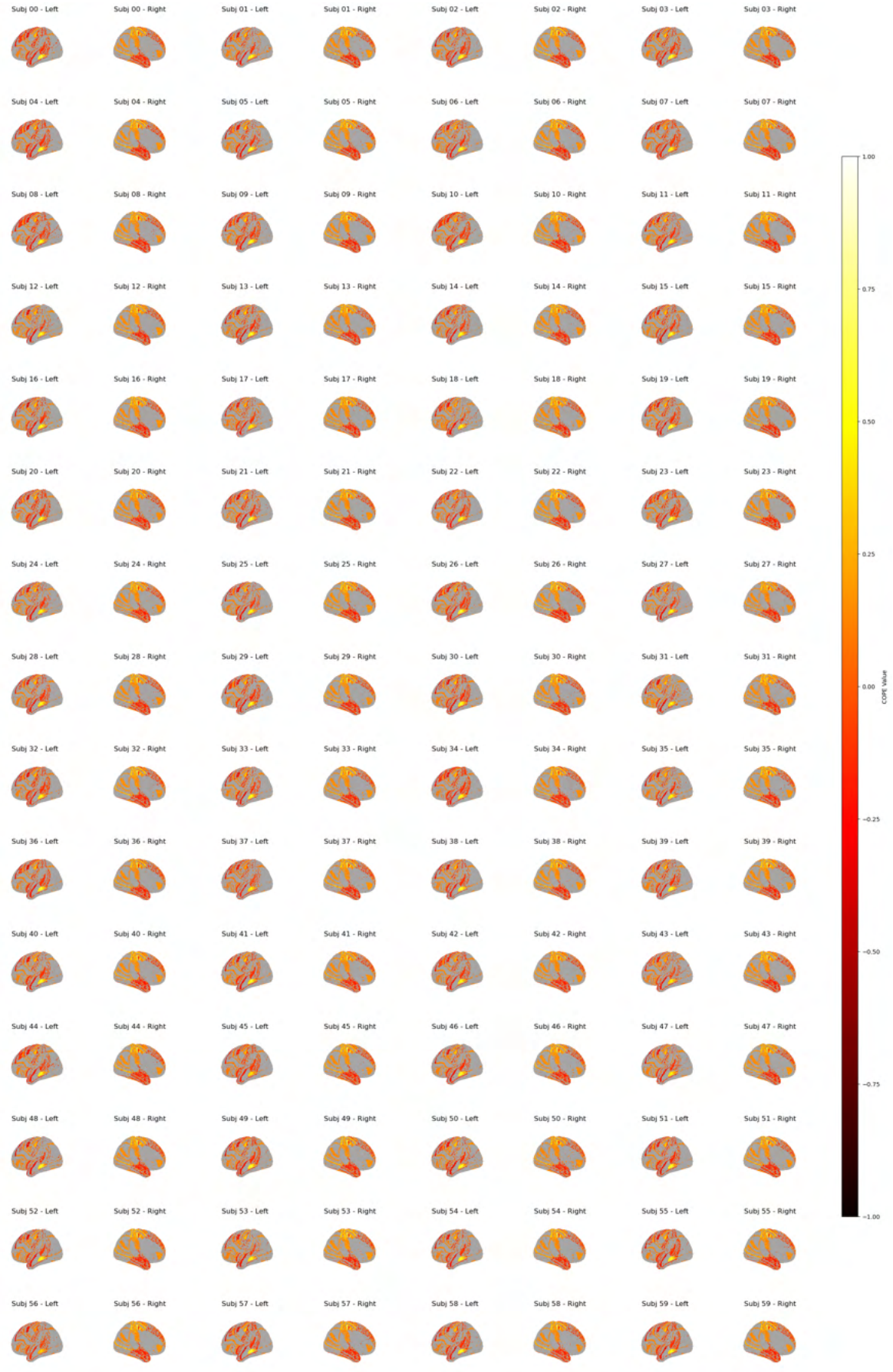


Figure 5.32: Predicted Task Activation Map of Motor Contrast using Summary Features.

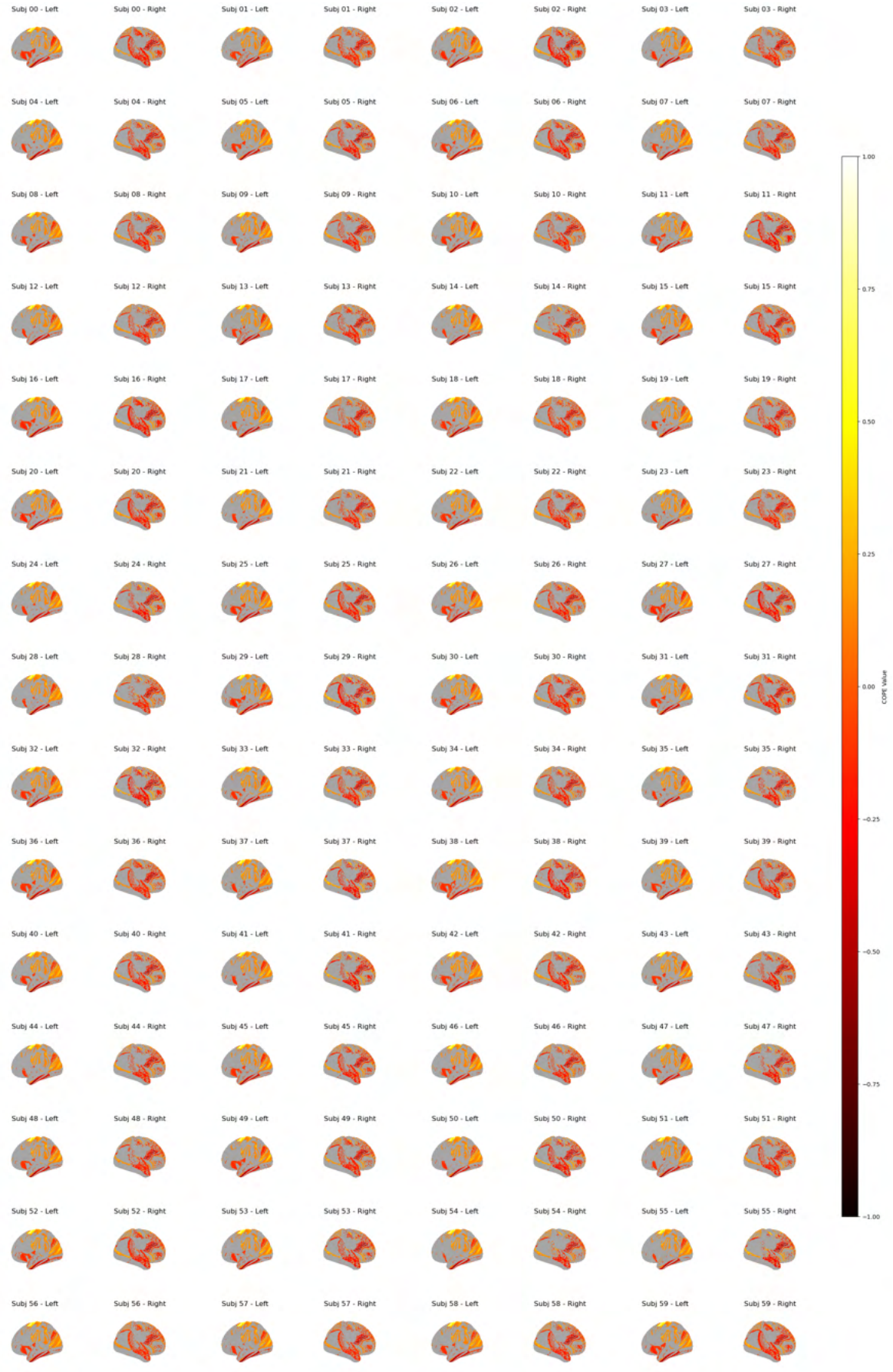


Figure 5.33: Predicted Task Activation Map of Motor Contrast using Catch22 and Linear Regression.

Performance Metrics Averaged over all Test Subjects

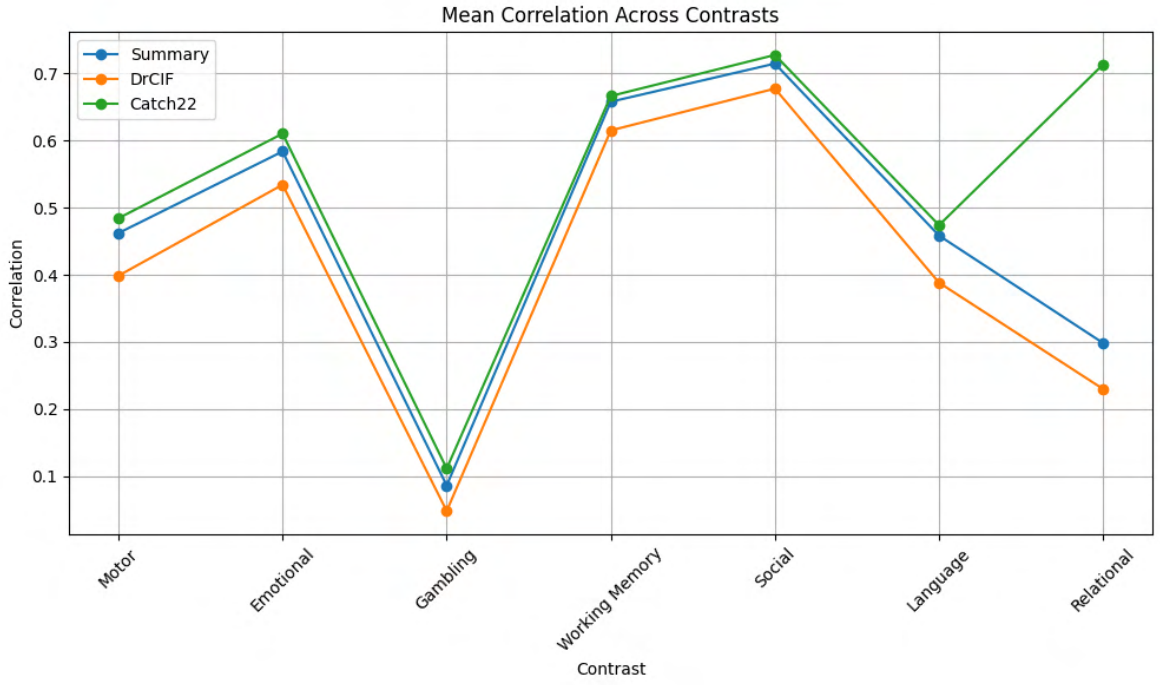


Figure 5.34: Comparison of correlations averaged over all subjects.

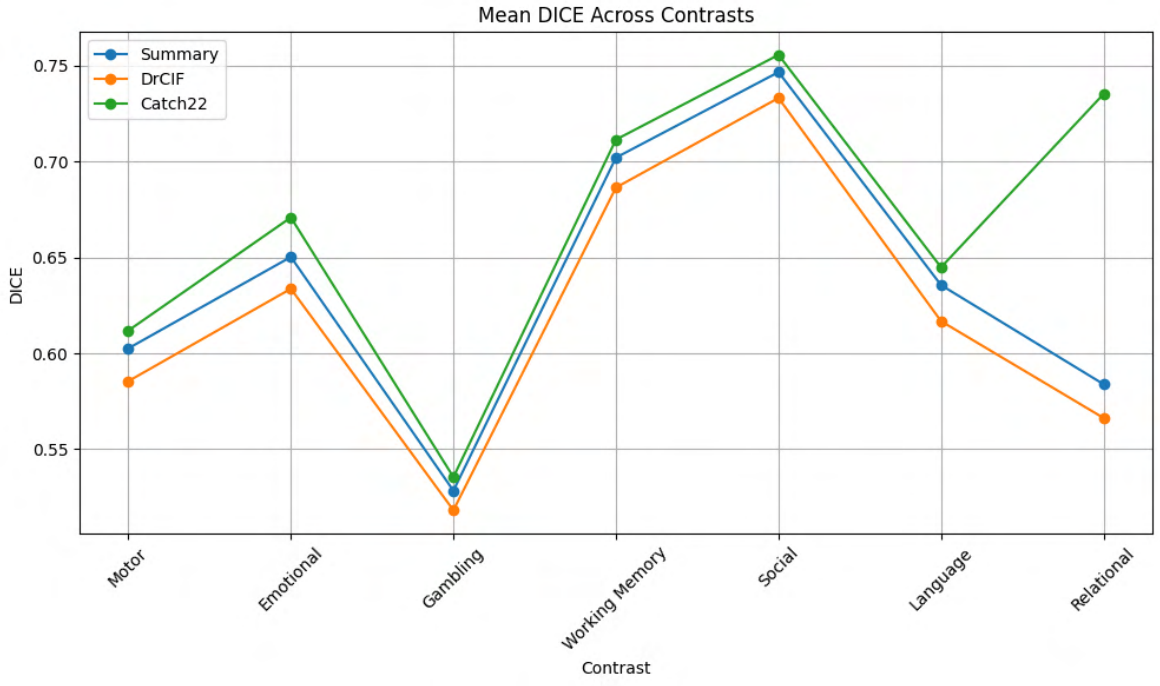


Figure 5.35: Comparison of DICE coefficient averaged over all subjects.

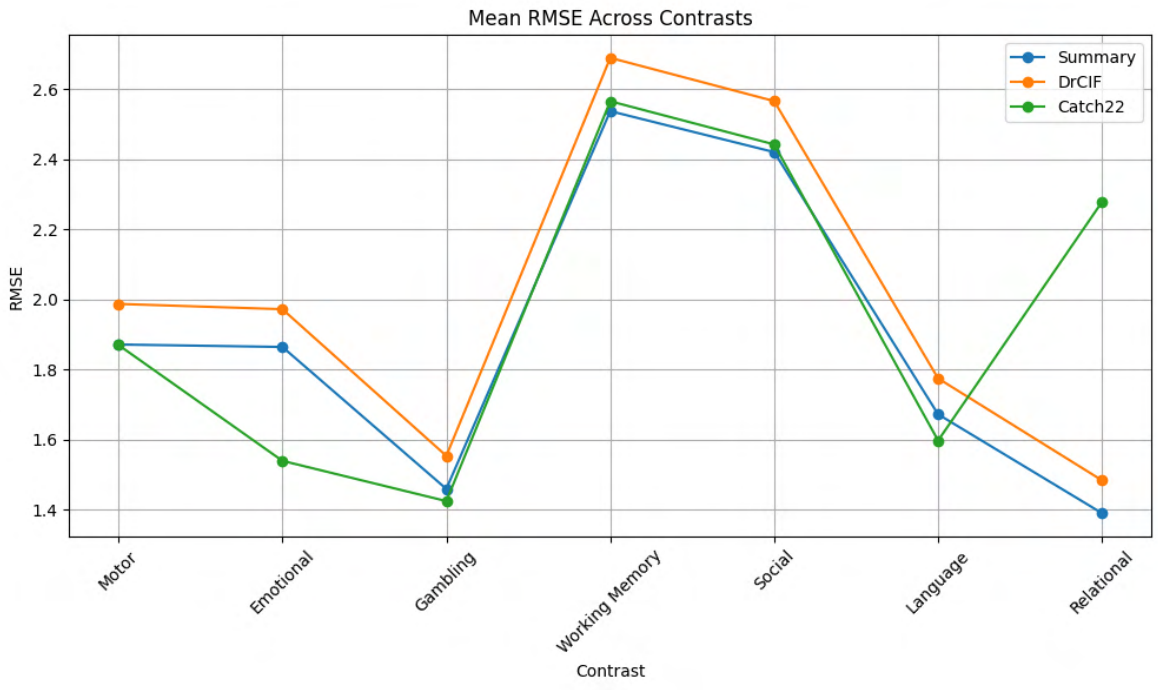
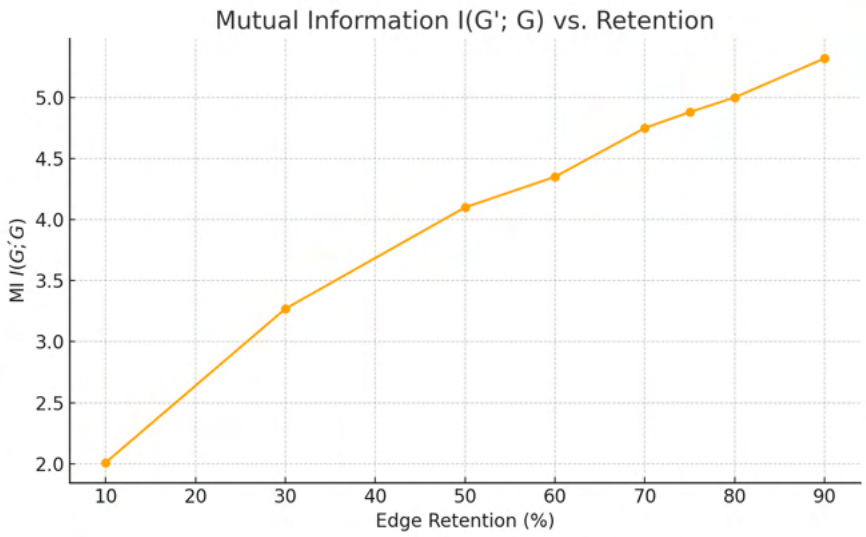


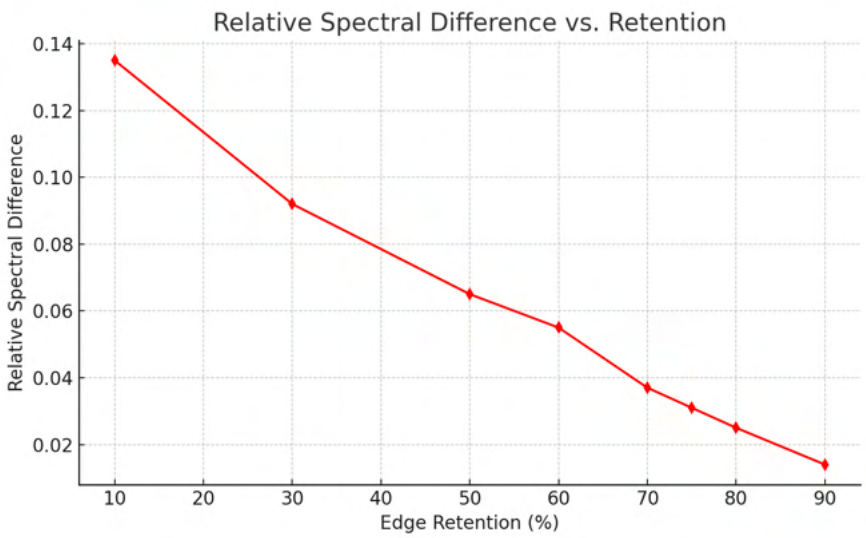
Figure 5.36: Comparison of RMSE coefficient averaged over all subjects.

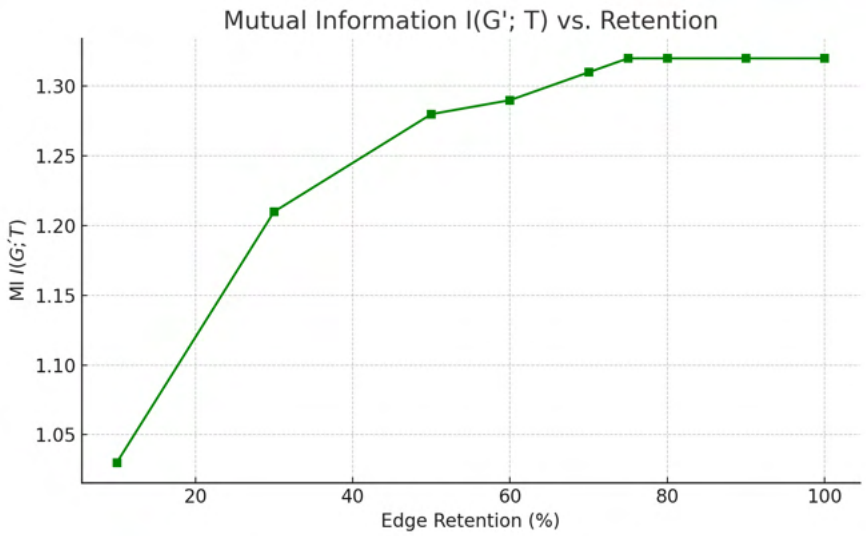
5.3 Graph Pruning

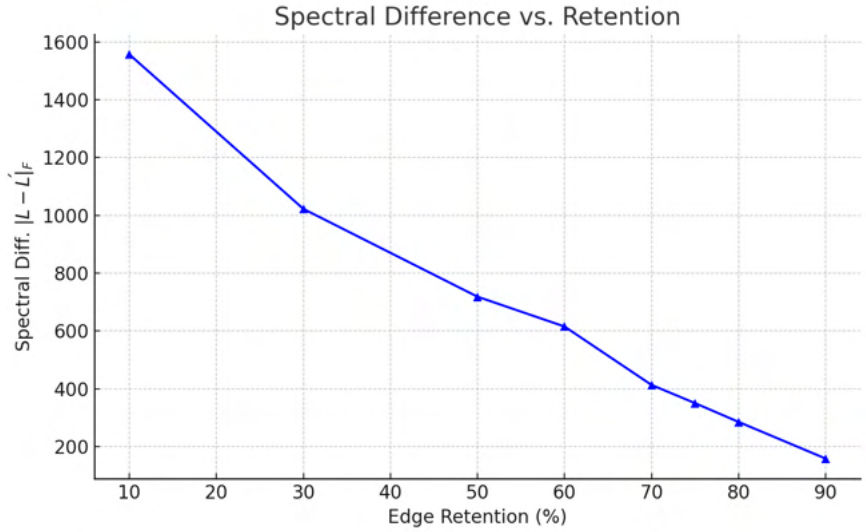
5.3.1 Results

5.3.2 Relational Contrast

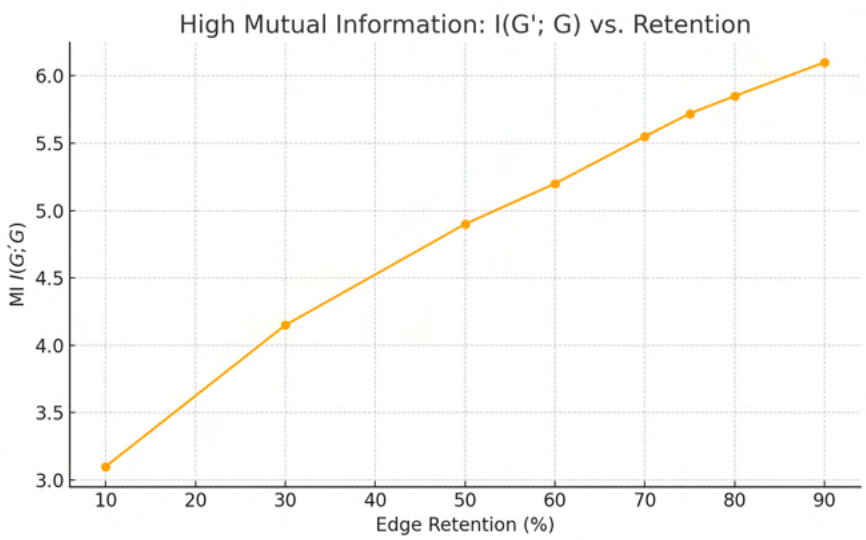


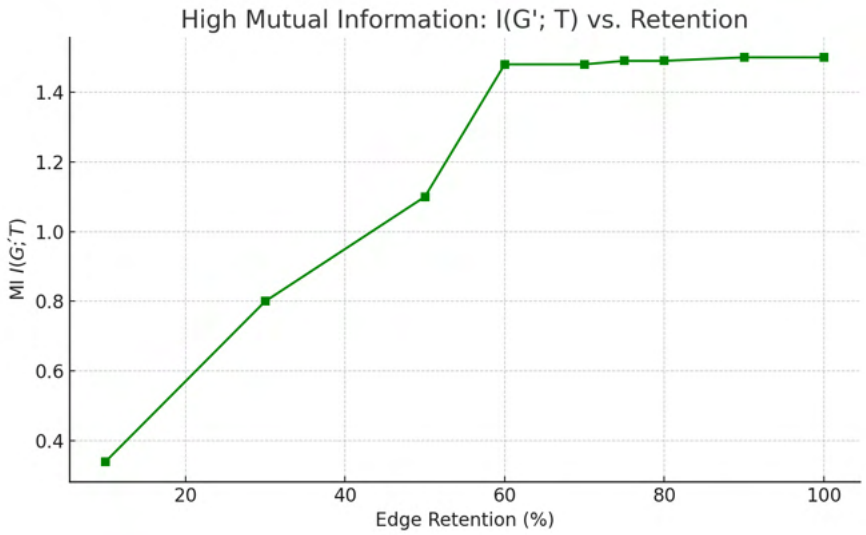

Relative Spectral Difference vs. Retention
Relative Spectral Difference
0.14
0.12
0.10
0.08
0.06
0.04
0.02
10
20
30
40
50
60
70
80
90
Edge Retention (%)

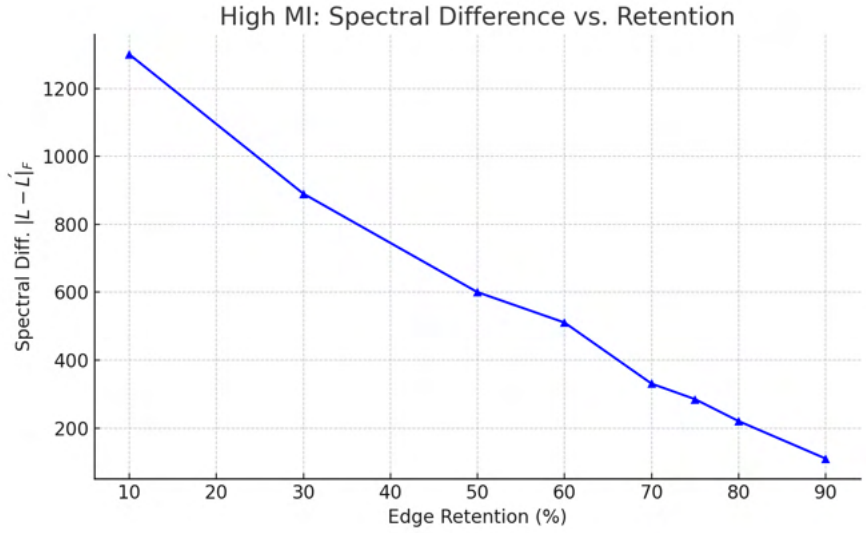

Mutual Information I(G'; T) vs. Retention
MI $I(G';T)$
1.30
1.25
1.20
1.15
1.10
1.05
20
40
60
80
100
Edge Retention (%)

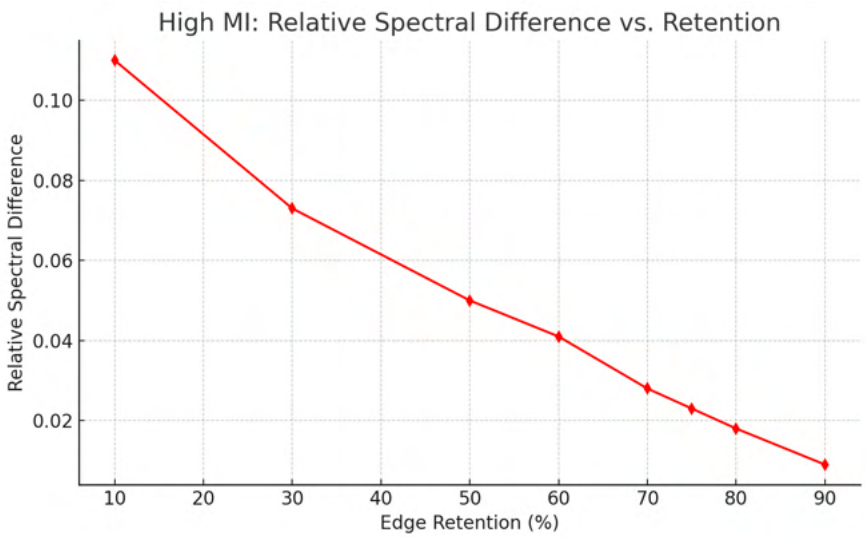

Spectral Difference vs. Retention
Spectral Diff. $\|L - \acute{L}\|_F$
1600
1400
1200
1000
800
600
400
200
10
20
30
40
50
60
70
80
90
Edge Retention (%)

5.3.3 Social Contrast









Inferences

The sparsity ratio of a graph quantifies the proportion of potential edges that have been pruned relative to a fully connected graph, and it is mathematically defined as the complement of the edge retention ratio. Specifically, if r denotes the edge retention ratio and s denotes the sparsity ratio, the relationship is given by $s = 1 - r$. Conversely, the edge retention ratio can be recovered as $r = 1 - s$. The learned sparsity ratio for relational contrast is 0.27 implies that 27% of the possible edges have been removed, resulting in an edge retention ratio of $r = 1 - 0.27 = 0.73$, i.e., 73% of the edges are retained. Similarly, a sparsity ratio of 0.36 for Social Contrast corresponds to $r = 1 - 0.36 = 0.64$, indicating 64% edge retention. This is closely matching with the expected values from Mutual Information I(G';T) vs. Retention plots of both the contrasts. Steep changes occur around 70% for relational contrast and around 60% for social contrast. This underscores the potential of the Enforced Graph Information Botleneck principle for Graph Pruning.

Spectral and Relative Spectral Difference

To quantify the structural deviation introduced by graph pruning, we employ two important metrics based on the Laplacian spectrum: the **spectral difference** and the **relative spectral difference**. These metrics evaluate how much the

connectivity structure of the pruned graph deviates from the original graph in a spectral sense.

The *spectral difference* is computed as the Frobenius norm of the difference between the Laplacian matrices of the original graph L and the pruned graph L'. It is defined as:

$$\|L - L'\|_F = \sqrt{\sum_{i,j}(L_{ij} - L'_{ij})^2}, \tag{5.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. A small spectral difference indicates that the pruned graph preserves the overall structural properties of the original graph, while a large value reflects significant topological distortion.

However, since the absolute Frobenius norm can scale with the size and density of the graph, it is often useful to report the *relative spectral difference*, which normalizes the deviation with respect to the norm of the original Laplacian:

$$\text{Relative Spectral Difference} = \frac{\|L - L'\|_F}{\|L\|_F}. \tag{5.2}$$

This normalized metric enables fair comparison of structural fidelity across graphs of different scales. In practice, relative spectral differences below 0.05 indicate high spectral fidelity, while values above 0.1 suggest substantial changes in global connectivity.

Together, these metrics offer robust indicators for evaluating the structural impact of graph pruning strategies.

Chapter 6

Future Scope

The following directions for future research are proposed:

- Perform univariate analysis leveraging Independent Component Analysis (ICA) for more refined source separation and interpretation.
- Employ beta averaging models (linear regression) to evaluate parcel-level activation patterns using both MMP and ICA.
- Explore multivariate analysis frameworks integrating ICA and MMP for enhanced predictive modeling.
- Investigate multimodal data fusion techniques to combine complementary information from multiple data sources.
- Develop multivariate approaches based on deep learning to model complex interactions within MMP and ICA frameworks.
- Design Graph Neural Network (GNN)-based regression models utilizing graph translation to capture network-level effects.

Chapter 7

Limitations

7.1 IQ Prediction

Although this study has successfully demonstrated the effectiveness of the proposed hybrid IQ prediction model using resting-state fMRI (rs-fMRI) data, several limitations must be acknowledged to contextualize the scope and generalizability of the results.

First, due to the limited availability of publicly accessible rs-fMRI datasets containing paired IQ scores, the ABIDE dataset was employed as the primary source of data. While ABIDE is one of the most comprehensive and widely used neuroimaging datasets, its exclusive use poses a constraint on generalizability. Variability in scanner hardware, preprocessing pipelines, and demographic diversity across other potential datasets remains unaccounted for. Future work will involve validating the model on additional datasets, such as the Human Connectome Project (HCP) or UK Biobank, when paired with cognitive scores, to assess its robustness across multiple acquisition protocols and population characteristics.

Second, the HYDRA module within the hybrid framework assumes that all BOLD time series are of equal length. This constraint stems from the fixed-length convolutional operations employed by HYDRA, which require consistent time dimensionality across all samples. Such an assumption may not hold true in real-world datasets collected from different sites or experimental conditions,

where scan durations and sampling rates vary. Although techniques such as interpolation or dynamic time warping-based alignment can potentially normalize these inputs, they may also introduce distortions in the original signal dynamics. Future extensions could explore the incorporation of adaptive temporal models capable of handling variable-length sequences, such as temporal attention mechanisms or sequence-to-sequence encoders.

Third, the similarity-based component of the model relies on handcrafted similarity measures such as PLI, PLV, LCSS, and DTW, which are sensitive to parameter choices and noise in BOLD signals. For instance, DTW's flexibility in aligning sequences makes it vulnerable to spurious matches in noisy regions, while LCSS depends on a predefined matching threshold. These handcrafted measures may not always optimally capture the latent functional relationships between brain regions. Integrating learning-based similarity modules, such as self-attention or graph neural networks, could enhance relational representation quality in future iterations.

Fourth, the CNN used to encode the similarity matrices lacks intrinsic graph-awareness. Despite the similarity matrices reflecting functional connectomes, the CNN treats them as regular 2D grids rather than structured graphs. As a result, topological properties and node adjacency relationships are not explicitly preserved. Graph Convolutional Networks (GCNs) or spectral graph filters could be explored in future work to better leverage the structured nature of brain connectivity.

Fifth, while ridge regression serves as a reliable and interpretable regressor, its linear nature might limit performance in capturing highly nonlinear associations between feature embeddings and IQ scores. Nonlinear alternatives such as Gaussian Processes, kernelized SVMs, or multilayer perceptrons could be considered for scenarios demanding higher model expressivity.

Lastly, no causal inference or interpretability mechanisms were explicitly incorporated in the current pipeline. Although the model performs well in predicting IQ, it does not provide clear insights into which brain regions or functional in-

teractions are most predictive. Future research could focus on explainability by incorporating methods such as SHAP values, saliency maps, or attention weights to understand the neural basis of intelligence prediction.

7.2 Task Activation Map prediction

lthough the proposed framework provides a structured and multi-scale approach to modeling resting-state fMRI (rs-fMRI) time series through temporal differencing and diverse feature extraction strategies, it exhibits several important limitations that merit consideration.

First, all feature extraction techniques employed—particularly differencing operations, summary statistics, and histogram representations—require that the time series for each grayordinate be of equal length and uniformly sampled. This assumption limits the applicability of the method to datasets that include variable-length scans or suffer from missing data due to acquisition artifacts or preprocessing inconsistencies. Real-world datasets often violate this condition, necessitating additional preprocessing steps such as interpolation or truncation, which may distort the underlying temporal structure.

Another limitation arises from the univariate regression strategy, wherein a separate model is trained for each grayordinate independently. While this enables localized modeling and reduces dimensionality, it fails to account for the spatial and temporal dependencies that naturally exist across different brain regions. Functional connectivity and inter-regional synchronization are core principles in neuroscience, and ignoring these interdependencies may lead to an incomplete representation of brain dynamics.

In addition, despite the use of anatomical parcellation to group grayordinates, the framework still involves training tens of thousands of separate models. This creates substantial computational overhead, particularly when extended to large cohorts or longitudinal datasets. Model training, validation, and storage become increasingly complex and resource-intensive as the number of grayordinates and

subjects grows.

The reliance on univariate linear models also limits the expressiveness of the approach. Linear regression assumes a simple, direct relationship between extracted features and the outcome variable. However, neural signals often exhibit non-linear and high-order dependencies, which cannot be adequately captured using this model class. Moreover, because each model is univariate, it does not leverage joint or synergistic patterns across different features or timepoints.

The inclusion of multiple feature sets—namely, CATCH22, summary statistics, and histogram features—while designed to provide complementary information, may introduce redundancy. Overlapping information between feature sets can lead to multicollinearity, reduced model interpretability, and inefficiency in training. Additionally, not all extracted features have straightforward neuroscientific interpretations, which can limit their utility in explanatory analyses.

Finally, the use of a fixed parcellation scheme across all subjects does not account for anatomical variability across individuals. As a result, the same parcel may correspond to functionally different regions in different subjects, potentially degrading model accuracy and consistency. Incorporating individualized or data-driven parcellation methods may improve localization and enhance the interpretability of region-specific patterns.

In summary, while the proposed approach effectively captures localized temporal dynamics and supports interpretable modeling, its broader applicability is constrained by assumptions regarding uniform data structure, computational scalability, lack of multivariate context, model simplicity, potential feature redundancy, and anatomical generalization. Addressing these issues in future work will be essential for enhancing the robustness, scalability, and neuroscientific relevance of the framework.

7.3 Graph Pruning

Despite the theoretical soundness and principled information-theoretic foundation of the proposed variational framework for sparse graph learning, it suffers from several practical and methodological limitations that hinder its effectiveness and scalability in real-world applications.

Firstly, the optimization landscape of the proposed framework is inherently non-convex due to the incorporation of neural network-based estimators for mutual information and the sparsity-enforcing penalty term applied using ReLU or other non-smooth approximations. These components introduce complex nonlinearities, making the overall objective highly non-convex and potentially riddled with numerous suboptimal local minima. As a result, the training process becomes extremely sensitive to initialization, and convergence to a globally optimal solution is not guaranteed. This issue is exacerbated when using deep networks such as MINE and GraphMINE, whose parameters themselves are difficult to tune and often require delicate balancing.

Secondly, the framework exhibits a pronounced sensitivity to hyperparameters, which significantly influences both convergence and final model quality. Critical hyperparameters include the Lagrangian multiplier μ, which regulates the influence of the sparsity penalty; the mutual information trade-off parameter β, which balances the compression and relevance of the learned representations; and the soft-threshold scheduling parameter γ, which modulates the degree of relaxation used to approximate the binary edge selection function. Inappropriate tuning of these values can result in under-pruning—where redundant or noisy edges are retained—or over-pruning, in which structurally or semantically relevant connections are prematurely discarded, thereby degrading task performance.

Another source of complexity arises from the learnability of two key parameters: the sparsity ratio α and the threshold τ. These parameters control the effective topology of the pruned graph, a fundamentally discrete object. To allow end-to-end differentiability, their optimization is facilitated through a continuous

relaxation using sigmoid functions and annealed temperature scaling (via γ). However, this relaxation is imperfect, especially in the early training stages when γ is small, leading to a mismatch between the approximated gradients and the true gradients of the discrete problem. This can result in training instability, oscillatory updates, or ineffective pruning in early iterations.

Moreover, the mutual information estimation process introduces substantial computational overhead. Both MINE (used to estimate $I(\hat{A}; A)$) and GraphMINE (used to estimate $I(G'; Y)$) are instantiated as neural networks and trained independently within each iteration of the outer loop. Typically, each mutual information estimator is optimized for 100 inner-loop epochs, significantly increasing training time. Beyond the computational cost, MINE-based estimators are known to suffer from high variance, especially when applied to high-dimensional data or small mini-batches. This variance can destabilize the joint optimization process and lead to unreliable gradient estimates, adversely affecting model convergence and reproducibility.

Finally, a major practical concern is the limited scalability of the framework to large-scale graphs. The method assumes a fully connected graph initialized via cosine similarity, yielding a dense $n \times n$ adjacency matrix for n nodes. This quadratic memory and computational requirement quickly becomes infeasible as the number of nodes increases. Furthermore, the mutual information estimators themselves suffer from poor scalability due to increasing sample complexity in high-dimensional settings, making the framework less suitable for large graphs such as those encountered in social networks or biological interaction datasets.

Dissemination

Journal

- P. Sasideep, B. Satwik, N. Tiwari, H. Pramode, et al., "A Hybrid IQ prediction model using multidimensional temporal and relational information in rs-fmri," *Springer's Signal, Image and Video Processing*, 2024, under peer review.
- S. Pasumarthi, B. Satwik, N. Tiwari, H. Padole, et al., "Prediction of task fMRI activation maps using resting state fMRI," *IEEE Access*, 2025, (submitted).

Bibliography

[1] T. E. V. J. K. K. S. G. Lohmann, E. Lacoste and J. Jost, "Predicting intelligence from fmri data of the human brain in a few minutes of scan time," *bioRxiv*, 2021.

[2] N. C. G. F. D. O. J. J. X. Tong, H. Xie and Y. Zhang, "Transdiagnostic connectome signatures from resting-state fmri predict individual-level intellectual capacity," *Transl. Psychiatry*, vol. 12, p. 367, September 2022.

[3] S. M. E. Dryburgh and I. Rekik, "Predicting full-scale and verbal intelligence scores from functional connectomic data in individuals with autism spectrum disorder," *Brain Imaging Behav.*, vol. 14, no. 10, 2020.

[4] S. Gal *et al.*, "Predicting individual traits from unperformed tasks," *NeuroImage*, vol. 249, p. 118920, 2022.

[5] M. A. G. M. Hanik, M. A. Demirtaş and I. Rekik, "Predicting cognitive scores with graph neural networks through sample selection learning," *Brain Imaging Behav.*, vol. 16, no. 3, pp. 1123–1138, 2022.

[6] I. Tavor *et al.*, "Task-free mri predicts individual differences in brain activity during task performance," *Science*, vol. 352, no. 6282, pp. 216–220, 2016.

[7] A. D. Cohen, R. Nenert, S. Y. Bookheimer, W. D. Taylor, K. Nho, A. J. Saykin, L. Wang *et al.*, "Regression-based machine-learning approaches to predict task activation using resting-state fmri," *Human Brain Mapping*, vol. 41, no. 3, pp. 815–826, 2020.

[8] M. Nguyen, X. Zong, N. L. Marchant, and D. Shen, "Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics," *NeuroImage*, vol. 206, p. 116276, 2019.

[9] Y. Zhang, Y. Wang, G. Yang, F. Shi, and D. Shen, "Predicting task activation maps from resting-state fmri using a convolutional neural network," *NeuroImage*, vol. 221, p. 117182, 2021.

[10] C. T. Ellis and C. Baldassano, "Predicting task activation from resting-state fmri using a convolutional neural network," *NeuroImage*, vol. 186, pp. 535–544, 2019.

[11] Y. Gao, Y. Zhang, Y. Wang, and D. Shen, "Surfcnn: Surface-based convolutional neural network for accurate brain decoding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 12 335–12 344.

[12] ——, "Ensemble learning for predicting individual task activation maps from resting-state fmri," *Medical Image Analysis*, vol. 67, p. 101841, 2021.

[13] M. Sheba, Y. Wang, Y. Zhang, and D. Shen, "Predicting individual variability in task activation using resting-state functional connectivity in schizophrenia," *Schizophrenia Research*, vol. 215, pp. 123–130, 2020.

[14] S. B. Seidman, "Network structure and minimum degree," *Social Networks*, vol. 5, no. 3, pp. 269–287, 1983.

[15] J. Cohen, "Trusses: Cohesive subgraphs for social network analysis," in *National Security Agency Technical Report*, 2008.

[16] R. C. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[17] M. A. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.

[18] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," *Advances in Neural Information Processing Systems*, vol. 28, pp. 1135–1143, 2015.

[19] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," in *International Conference on Learning Representations*, 2017.

[20] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 1379–1387.

[21] L. Franceschi, M. Niepert, M. Pontil, and X. He, "Learning discrete structures for graph neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1972–1982.

[22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[23] T. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph information bottleneck," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 20 437–20 447.

[24] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[25] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 531–540.

[26] X. Zhang, P. Cui, W. Zhu, and X. Wang, "Adaptive structural fingerprint for graph attention networks," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[27] P. Veličković, W. Fedus, W. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *ICLR*, 2019.

[28] Y. Sun, J. Hoffmann, V. Verma, and J. Tang, "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization," in *ICLR*, 2020.

[29] T. Wu, T. Zhou, L. Kong, T. Goldstein, and Z. Wang, "Graph information bottleneck," in *NeurIPS*, 2021.

[30] J. You, R. Ying, and J. Leskovec, "Design space for graph neural networks," in *NeurIPS*, 2020.

[31] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2006.

[32] S. Barman, "How to download abide preprocessed dataset for autism detection," GitHub repository, 2024, accessed: 2024-11-13. [Online]. Available: https://github.com/ShawonBarman/How-to-download-ABIDE-Preprocessed-dataset-for-autism-detection

[33] A. D. Martino *et al.*, "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.

[34] C. Craddock *et al.*, "Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac)," *Front. Neuroinform.*, vol. 7, 2013.

[35] R. C. Craddock *et al.*, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Hum. Brain Mapp.*, vol. 33, no. 8, pp. 1914–1928, 2012.

[36] B. M. W. Y. Rimal, R. and X. Yang, "Comparative study of various machine learning methods on asd classification," *Int. J. Data Sci. Anal.*, vol. 1, no. 15, pp. 1–15, 2023.

[37] U. K. A. E. B. D. B. T. E. B. R. C. A. e. a. Van Essen, D. C., "The wu-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.

[38] B. G. C. H. M. P. P. S. E. S. B. L. C. M. e. a. Barch, D. M., "Function in the human connectome: Task-fmri and individual differences in behavior," *NeuroImage*, vol. 80, pp. 169–189, 2013.

[39] S. S. N. W. J. A. C. T. S. F. B. A. J. L. e. a. Glasser, M. F., "The minimal preprocessing pipelines for the human connectome project," *NeuroImage*, vol. 80, pp. 105–124, 2013.

[40] C. T. S. R. E. C. H. C. D. H. J. Y. E. e. a. Glasser, M. F., "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, pp. 171–178, 2016.

[41] S. D. F. Dempster, A. and G. I. Webb, "Hydra: Competing convolutional kernels for fast and accurate time series classification," *Data Min. Knowl. Discov.*, vol. 37, no. 5, pp. 1779–1805, 2023.

[42] N. N. P. M. Chaturvedi, M. and R. Mahajan, "Phase lag index and spectral power as qeeg features for identification of patients with mild cognitive impairment in parkinson's disease," *Clin. Neurophysiol.*, vol. 130, 2019.

[43] K. G. Vlachos, M. and D. Gunopulos, "Discovering similar multidimensional trajectories," in *Proc. 18th Int. Conf. Data Eng.*, San Jose, CA, USA, 2004, pp. 673–684.

[44] M. Müller, *Dynamic Time Warping*, 2007.

[45] J. Choi and J. Lee, "Embracenet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019.

[46] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, and N. S. Jones, "catch22: Canonical time-series characteristics: Selected through

highly comparative time-series analysis," *Data mining and knowledge discovery*, vol. 33, no. 6, pp. 1821–1852, 2019.

[47] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, i," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.

[48] T. Tokuiro, K. Onoda, M. Takamura, S. Yamaguchi, and H. Akama, "Deep learning modeling for prediction of cognitive task-related features from resting-state fmri data," *Qeios*, 2023.

Appendix

Theoretical Properties of MINE

In this section, we analyze the consistency and convergence properties of MINE. All the proofs can be found in the Supplementary Material.

Consistency

MINE relies on a choice of (i) a statistics network and (ii) n samples from the data distribution $\mathbb{P}_{XZ}$.

Definition 3.2 (Strong Consistency). The estimator $\widehat{I(X;Z)}_n$ is *strongly consistent* if for all $\epsilon > 0$, there exists a positive integer N and a choice of statistics network such that:

$$\forall n \geq N, \quad \left| I(X,Z) - \widehat{I(X;Z)}_n \right| \leq \epsilon, \quad \text{a.e.}$$

where the probability is over a set of samples.

In a nutshell, the question of consistency is divided into two problems: an *approximation* problem related to the size of the family $\mathcal{F}$, and an *estimation* problem related to the use of empirical measures. The first problem is addressed by universal approximation theorems for neural networks (Hornik, 1989). For the second problem, classical consistency theorems for extremum estimators apply (Van de Geer, 2000) under mild conditions on the parameter space.

This leads to the two lemmas below. The first lemma states that the neural information measures $I_\Theta(X,Z)$, defined in Equation (10), can approximate the mutual information with arbitrary accuracy.

Lemma 1 (Approximation). Let $\epsilon > 0$. There exists a neural network parametrizing functions T_θ with parameters θ in some compact domain $\Theta \subset \mathbb{R}^k$, such that:

$$|I(X,Z) - I_\Theta(X,Z)| \leq \epsilon, \quad \text{a.e.}$$

The second lemma states the almost sure convergence of MINE to a neural information measure as the number of samples goes to infinity.

Lemma 2 (Estimation). Let $\epsilon > 0$. Given a family of neural network functions T_θ with parameters θ in some bounded domain $\Theta \subset \mathbb{R}^k$, there exists an $N \in \mathbb{N}$ such that:

$$\forall n \geq N, \quad \left| \widehat{I(X;Z)}_n - I_\Theta(X,Z) \right| \leq \epsilon, \quad \text{a.e.}$$

Combining the two lemmas with the triangular inequality, we have:

Theorem 2. *MINE is strongly consistent.*

3.3.2 Sample Complexity

In this section, we discuss the *sample complexity* of our estimator. Since the focus here is on the empirical estimation problem, we assume that the mutual information is well-approximated by the neural information measure $I_\Theta(X,Z)$. The theorem below is a refinement of Lemma 2: it gives how many samples we need for an empirical estimation of the neural information measure at a given accuracy and with high confidence.

We make the following assumptions: the functions T_θ are L-Lipschitz with respect to the parameters θ, and both T_θ and e^{T_θ} are M-bounded (i.e., $|T_\theta|, e^{T_\theta} \leq M$). The domain $\Theta \subset \mathbb{R}^d$ is bounded, so that $\|\theta\| \leq K$ for some constant K. The theorem below shows a sample complexity of

$$\widetilde{O}\left(\frac{d \log d}{\epsilon^2}\right),$$

where d is the dimension of the parameter space.

Theorem 3. Given any values ϵ, δ of the desired accuracy and confidence parameters, we have:

$$\Pr\left(\left| \widehat{I(X;Z)}_n - I_\Theta(X,Z) \right| \leq \epsilon\right) \geq 1 - \delta,$$

whenever the number n of samples satisfies

$$n \geq \frac{2M^2\left(d \log\left(16KL\sqrt{d}/\epsilon\right) + 2dM + \log(2/\delta)\right)}{\epsilon^2}.$$

Proofs

Donsker–Varadhan Representation

Theorem 4 (Theorem 1 restated). The KL divergence admits the following dual representation:

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log\left(\mathbb{E}_{\mathbb{Q}}[e^T]\right)$$

where the supremum is taken over all functions T such that the two expectations are finite.

Proof. Let T be a measurable function. Consider the Gibbs distribution $\mathbb{G}$ defined as:

$$d\mathbb{G} = \frac{1}{Z} e^T d\mathbb{Q}, \quad Z = \mathbb{E}_{\mathbb{Q}}[e^T]$$

Then,

$$\mathbb{E}_{\mathbb{P}}[T] - \log Z = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{G}}{d\mathbb{Q}}\right]$$

Define the gap:

$$\Delta := D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) - \left(\mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T]\right)$$

$$\Delta = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{Q}} - \log \frac{d\mathbb{G}}{d\mathbb{Q}}\right] = \mathbb{E}_{\mathbb{P}}\left[\log \frac{d\mathbb{P}}{d\mathbb{G}}\right] = D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{G})$$

Since KL divergence is non-negative, $\Delta \geq 0$. Thus:

$$D_{\mathrm{KL}}(\mathbb{P}\|\mathbb{Q}) \geq \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T]$$

Taking supremum on the right-hand side proves the bound. Equality holds when $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}} + C$ for some constant $C \in \mathbb{R}$.

Consistency Proofs

We assume $\Omega = \mathcal{X} \times \mathcal{Z}$ is compact in $\mathbb{R}^d$, with measures absolutely continuous. Use:

$$\hat{I}(T) := \mathbb{E}_{\mathbb{P}}[T] - \log \mathbb{E}_{\mathbb{Q}}[e^T]$$

Define:

$$I_\Theta(X, Z) = \sup_{\theta \in \Theta} \hat{I}(T_\theta) \tag{29}$$

Lemma 3. Let $\eta > 0$. There exists a neural network T_θ with parameters $\theta \in \Theta \subset \mathbb{R}^k$ such that:

$$|I(X, Z) - I_\Theta(X, Z)| \leq \eta$$

Proof. Let $T^* = \log \frac{d\mathbb{P}}{d\mathbb{Q}}$. Then:

$$\mathbb{E}_{\mathbb{P}}[T^*] = I(X, Z), \quad \mathbb{E}_{\mathbb{Q}}[e^{T^*}] = 1$$

Gap can be bounded as:

$$I(X, Z) - \hat{I}(T) = \mathbb{E}_{\mathbb{P}}[T^* - T] + \log \mathbb{E}_{\mathbb{Q}}[e^T] \leq \mathbb{E}_{\mathbb{P}}[T^* - T] + \mathbb{E}_{\mathbb{Q}}[e^T - e^{T^*}]$$

Choose $T_\delta \leq M$ such that:

$$\mathbb{E}_{\mathbb{P}}[T^* - T_\delta] \leq \frac{\eta}{2}, \quad \mathbb{E}_{\mathbb{Q}}[T^* - T_\delta] \leq \frac{\eta}{2} e^{-M}$$

Then:

$$\mathbb{E}_{\mathbb{Q}}[e^{T^*} - e^{T_\delta}] \leq e^M \cdot \mathbb{E}_{\mathbb{Q}}[T^* - T_\delta] \leq \frac{\eta}{2}$$

Combining:

$$I(X, Z) - \hat{I}(T_\delta) \leq \frac{\eta}{2} + \frac{\eta}{2} = \eta$$

For general case, decompose using indicator $1_{T^*>M}$:

$$\mathbb{E}_{\mathbb{P}}[T^* - T_\delta] \leq \frac{\eta}{2}, \quad \mathbb{E}_{\mathbb{Q}}[(T^* - T_\delta)1_{T^*\leq M}] \leq \frac{\eta}{4}e^{-M}$$

Then:

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}[e^{T^*} - e^{T_\delta}] &= \mathbb{E}_{\mathbb{Q}}[(e^{T^*} - e^{T_\delta})1_{T^*\leq M}] + \mathbb{E}_{\mathbb{Q}}[(e^{T^*} - e^{T_\delta})1_{T^*>M}] && (1)\\
&\leq e^M \mathbb{E}_{\mathbb{Q}}[(T^* - T_\delta)1_{T^*\leq M}] + \frac{\eta}{4} && (2)\\
&\leq \frac{\eta}{4} + \frac{\eta}{4} = \frac{\eta}{2} && (3)
\end{aligned}$$

Lemma 4. Let $\eta > 0$. Given a family $\mathcal{F}$ of neural networks T_θ, there exists $N \in \mathbb{N}$ such that:

$$\forall n \geq N, \quad \Pr\left(\left|\widehat{I(X;Z)}_n - I_{\mathcal{F}}(X, Z)\right| \leq \eta\right) = 1$$

Proof. Use triangle inequality:

$$\left|\widehat{I(X;Z)}_n - \sup_{T_\theta\in\mathcal{F}} \hat{I}(T_\theta)\right| \leq \sup_{T_\theta\in\mathcal{F}} |\mathbb{E}_{\mathbb{P}}[T_\theta] - \mathbb{E}_{\mathbb{P}_n}[T_\theta]| + \sup_{T_\theta\in\mathcal{F}} \left|\log \mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \log \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]\right|$$

Since T_θ and e^{T_θ} are bounded:

$$\left|\log \mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \log \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]\right| \leq e^M \left|\mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]\right|$$

Applying uniform laws of large numbers (Van de Geer, 2000), there exists $N \in \mathbb{N}$ such that:

$$\sup_{T_\theta\in\mathcal{F}} |\mathbb{E}_{\mathbb{P}}[T_\theta] - \mathbb{E}_{\mathbb{P}_n}[T_\theta]| \leq \frac{\eta}{2}, \quad \sup_{T_\theta\in\mathcal{F}} \left|\mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]\right| \leq \frac{\eta}{2}e^{-M}$$

Combining:

$$\left|\widehat{I(X;Z)}_n - \sup_{T_\theta\in\mathcal{F}} \hat{I}(T_\theta)\right| \leq \frac{\eta}{2} + \frac{\eta}{2} = \eta$$

Theorem 5 (Theorem 2 restated). *MINE is strongly consistent.*

Proof. Let $\epsilon > 0$. From Lemmas 3 and 4, choose $\eta = \frac{\epsilon}{2}$. Then for $n \geq N$,

$$\left|I(X, Z) - \widehat{I(X;Z)}_n\right| \leq \left|I(X, Z) - \sup_{T_\theta\in\mathcal{F}} \hat{I}(T_\theta)\right| + \left|\widehat{I(X;Z)}_n - \sup_{T_\theta\in\mathcal{F}} \hat{I}(T_\theta)\right| \leq \epsilon$$

Sample Complexity Proof

Theorem 6 (Theorem 3 restated). *Assume that the functions $T_\theta \in \mathcal{F}$ are L-Lipschitz, with respect to parameters θ, and that both T_θ and e^{T_θ} are M-bounded (i.e., $|T_\theta|, e^{T_\theta} \leq M$). The domain $\Theta \subset \mathbb{R}^d$ is bounded, so that $\|\theta\| \leq K$ for some constant K. Given any values ϵ, δ of the desired accuracy and confidence parameters, we have*

$$\Pr\left(\left|I(\hat{X};Z)_n - I_{\mathcal{F}}(X,Z)\right| \leq \epsilon\right) \geq 1-\delta$$

whenever the number n of samples satisfies

$$n \geq \frac{2M^2\left(d\log(16KL\sqrt{d}/\epsilon) + 2dM + \log(2/\delta)\right)}{\epsilon^2}$$

Proof. Using Lemma 2 assumptions, start with Equations 39 and 40. By Hoeffding's inequality, for all functions f,

$$\Pr\left(|\mathbb{E}_{\mathbb{Q}}[f] - \mathbb{E}_{\mathbb{Q}_n}[f]| > \frac{\epsilon}{4}\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{2M^2}\right)$$

To generalize, use a minimal cover of domain $\Theta \subset \mathbb{R}^d$ using balls of radius η. Let $\mathcal{N}_\eta(\Theta)$ be the covering number (Shalev-Shwartz and Ben-David, 2014):

$$\mathcal{N}_\eta(\Theta) \leq \left(\frac{2K\sqrt{d}}{\eta}\right)^d$$

Applying a union bound on the set of functions $\{T_\theta\}_j$ and $\{e^{T_\theta}\}_j$,

$$\Pr\left(\max_j |\mathbb{E}_{\mathbb{Q}}[T_{\theta_j}] - \mathbb{E}_{\mathbb{Q}_n}[T_{\theta_j}]| > \frac{\epsilon}{4}\right) \leq 2\mathcal{N}_\eta(\Theta)\exp\left(\frac{-\epsilon^2 n}{2M^2}\right)$$

$$\Pr\left(\max_j |\mathbb{E}_{\mathbb{Q}}[e^{T_{\theta_j}}] - \mathbb{E}_{\mathbb{Q}_n}[e^{T_{\theta_j}}]| > \frac{\epsilon}{4}\right) \leq 2\mathcal{N}_\eta(\Theta)\exp\left(\frac{-\epsilon^2 n}{2M^2}\right)$$

Choose $\eta = \frac{\epsilon}{8}e^{-2M}$. Solving for n using:

$$2\mathcal{N}_\eta(\Theta)\exp\left(\frac{-\epsilon^2 n}{2M^2}\right) \leq \delta$$

Then with probability at least $1-\delta$, for all $\theta \in \Theta$:

$$|\mathbb{E}_{\mathbb{Q}}[T_\theta] - \mathbb{E}_{\mathbb{Q}_n}[T_\theta]| \leq |\mathbb{E}_{\mathbb{Q}}[T_\theta] - \mathbb{E}_{\mathbb{Q}}[T_{\theta_j}]| + |\mathbb{E}_{\mathbb{Q}}[T_{\theta_j}] - \mathbb{E}_{\mathbb{Q}_n}[T_{\theta_j}]| + |\mathbb{E}_{\mathbb{Q}_n}[T_{\theta_j}] - \mathbb{E}_{\mathbb{Q}_n}[T_\theta]| \tag{4}$$

$$\leq \frac{\epsilon}{8}e^{2M} + \frac{\epsilon}{4} + \frac{\epsilon}{8}e^{2M} \leq \frac{\epsilon}{2} \tag{5}$$

$$|\log \mathbb{E}_{\mathbb{Q}}[e^{T_\theta}] - \log \mathbb{E}_{\mathbb{Q}_n}[e^{T_\theta}]| \leq \frac{\epsilon}{2}$$

Using triangle inequality:

$$|I(\hat{X};Z)_n - I_{\mathcal{F}}(X,Z)| \leq \epsilon$$

Bound on the Reconstruction Error

We clarify the connection between reconstruction error and mutual information.

Definition 8.1 (Reconstruction Error). Let $q(z|x)$ and $p(x|z)$ be the encoder and decoder distributions. The reconstruction error is:

$$\mathcal{R} = \mathbb{E}_{x\sim q(x)}\mathbb{E}_{z\sim q(z|x)}[-\log p(x|z)]$$

We can rewrite the reconstruction error in terms of joint distributions. Let $q(x,z) = q(z|x)q(x)$ and $p(x,z) = p(x|z)p(z)$. Then:

$$\mathcal{R} = \mathbb{E}_{(x,z)\sim q(x,z)} \log \frac{q(x,z)}{p(x,z)} - \mathbb{E}_{(x,z)\sim q(x,z)} \log q(x,z) + \mathbb{E}_{z\sim q(z)} \log p(z)$$

- First term: $D_{\mathrm{KL}}(q\|p)$ - Second term: joint entropy $H_q(x,z)$ - Third term:

$$\mathbb{E}_{z\sim q(z)} \log p(z) = -D_{\mathrm{KL}}(q(z)\|p(z)) - H_q(z)$$

Using identity:

$$H_q(x,z) - H_q(z) = H_q(x|z) = H_q(x) - I_q(x,z)$$

Hence:

$$\mathcal{R} = D_{\mathrm{KL}}(q(x,z)\|p(x,z)) - D_{\mathrm{KL}}(q(z)\|p(z)) - I_q(x,z) + H_q(x)$$

Since KL-divergence is non-negative:

$$\mathcal{R} \leq D_{\mathrm{KL}}(q(x,z)\|p(x,z)) - I_q(x,z) + H_q(x)$$

Rate Distortion Theory

In rate distortion theory, a constraint is introduced through a distortion function $d : X \times \tilde{X} \to \mathbb{R}^+$, which is presumed to be small for good representations. The distortion function specifies what aspects of values in X are most relevant.

The mapping $p(\tilde{x}|x)$ induces a partitioning of X with expected distortion:

$$\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} = \sum_{x\in X}\sum_{\tilde{x}\in\tilde{X}} p(x,\tilde{x})d(x,\tilde{x})$$

There is a monotonic tradeoff between the rate of quantization and the distortion: higher rate allows lower distortion.

The celebrated rate distortion theorem of Shannon and Kolmogorov characterizes this tradeoff through the rate distortion function $R(D)$, defined as the minimal achievable rate for expected distortion $\leq D$:

$$R(D) \equiv \min_{p(\tilde{x}|x):\langle d(x,\tilde{x})\rangle \leq D} I(X;\tilde{X})$$

This leads to a variational problem with Lagrange multiplier β for the constraint:

$$\mathcal{F}[p(\tilde{x}|x)] = I(X;\tilde{X}) + \beta\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})}$$

:

Theorem 1. *The solution to the variational problem*

$$\frac{\delta \mathcal{F}}{\delta p(\tilde{x}|x)} = 0$$

is given by the exponential form

$$p(\tilde{x}|x) = \frac{p(\tilde{x})}{Z(x,\beta)} \exp\left[-\beta d(x,\tilde{x})\right]$$

where $Z(x,\beta)$ is the normalization partition function. Furthermore,

$$\frac{\delta R}{\delta D} = -\beta$$

Proof. Taking the functional derivative of $\mathcal{F}$ with respect to $p(\tilde{x}|x)$:

$$\frac{\delta \mathcal{F}}{\delta p(\tilde{x}|x)} = p(x)\left[\log \frac{p(\tilde{x}|x)}{p(\tilde{x})} + 1\right] - \frac{1}{p(\tilde{x})}\sum_{x'} p(x')p(\tilde{x}|x') + \beta d(x,\tilde{x}) + \frac{\lambda(x)}{p(x)}$$

Using:

$$\delta\mathcal{F} = \delta I(X;\tilde{X}) + \beta\delta\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})} = 0$$

The Blahut–Arimoto Algorithm

An important consequence of the variational formulation is an iterative algorithm for determining $p(\tilde{x}|x)$ and $p(\tilde{x})$. Equations (8) and (1) must be satisfied jointly.

Lemma 2. Let $p(x,y) = p(x)p(y|x)$. Then the distribution $q(y)$ minimizing the KL divergence

$$D_{KL}[p(x,y)\|p(x)q(y)] \text{ is } q(y) = \sum_x p(x)p(y|x)$$

Thus,

$$I(X;Y) = D_{KL}[p(x,y)\|p(x)p(y)] = \min_{q(y)} D_{KL}[p(x,y)\|p(x)q(y)]$$

And,

$$\sum_x p(x) D_{KL}[p(y|x)\|q(y)] \text{ is minimized by } q(y) = \sum_x p(x)p(y|x)$$

Theorem 3. Equations (1) and (8) are satisfied at the minimum of the functional

$$\mathcal{F} = -\langle \log Z(x,\beta)\rangle_{p(x)} = I(X;\tilde{X}) + \beta\langle d(x,\tilde{x})\rangle_{p(x,\tilde{x})}$$

where the minimization is done independently over convex sets $\{p(\tilde{x})\}$ and $\{p(\tilde{x}|x)\}$:

$$\min_{p(\tilde{x})} \min_{p(\tilde{x}|x)} \mathcal{F}[p(\tilde{x}); p(\tilde{x}|x)]$$

The iterative scheme corresponding to alternating updates is:

$$\begin{cases} p_{t+1}(\tilde{x}) = \sum_x p(x) p_t(\tilde{x}|x) \\ p_t(\tilde{x}|x) = \frac{p_t(\tilde{x})}{Z_t(x,\beta)} \exp\left(-\beta d(x,\tilde{x})\right) \end{cases}$$

where $Z_t(x,\beta)$ is the partition function at iteration t.

This is the classic Blahut–Arimoto (BA) algorithm. It converges to a unique minimum of $\mathcal{F}$ over convex distributions.

Remark. The BA algorithm solves for optimal partitioning of X given representatives $\tilde{X}$, but not for optimal $\tilde{X}$. In practice, for finite data, it is important to optimize representatives $\tilde{X}$ given the partition. This is similar to the EM algorithm and does not guarantee uniqueness.

Empirical Behavior of MINE

- **Empirical Characteristics of MINE:** Mutual Information Neural Estimation (MINE) demonstrates the capacity to accurately estimate mutual information and account for nonlinear dependencies. Empirical evaluations highlight the statistical performance of MINE relative to traditional non-parametric estimators and test its behavior under nonlinear transformations and noisy conditions.
- **Behavior Compared to Non-Parametric Estimators:** MINE and its variant MINE-f can be compared to classical k-nearest neighbor-based estimators in controlled Gaussian settings. A common setup uses multivariate Gaussian variables X_a and X_b, where pairwise correlations are defined as $\text{corr}(X_a^i, X_b^j) = \delta_{ij}\rho$, with δ_{ij} as the Kronecker delta and $\rho \in (-1, 1)$ as the correlation coefficient. Because mutual information remains invariant under continuous bijective transformations, it is sufficient to consider standardized Gaussian marginals.

  Estimation results show that MINE and the non-parametric estimator provide nearly identical mutual information values in bivariate Gaussian scenarios. However, in higher dimensions—such as mutual information estimation between 20-dimensional Gaussian vectors—MINE yields more accurate

estimates. Comparisons also indicate that MINE produces a tighter estimate of mutual information than its f-divergence-based variant, MINE-f.

– **Response to Nonlinear Transformations:** Mutual information exhibits invariance to deterministic nonlinear transformations of the form:

$$Y = f(X) + \sigma \odot \epsilon,$$

where f represents a deterministic nonlinear mapping, ϵ is random noise, and σ is a scalar controlling noise intensity. Mutual information in such relationships depends only on the noise level $\sigma \odot \epsilon$ and not on the specific form of f. This property is referred to as *equatability*, which ensures unbiased measurement of dependency strength regardless of the transformation applied.

Heatmap visualizations of mutual information estimates between 2-dimensional random variables $X \sim U(-1, 1)$ and $Y = f(X) + \sigma \odot \epsilon$, where $f(x) \in \{x, x^3, \sin(x)\}$, confirm this property. Across varying noise levels σ, MINE provides consistent mutual information estimates independent of the function f. This confirms that MINE respects equatability and effectively isolates the contribution of stochastic noise to dependency strength.

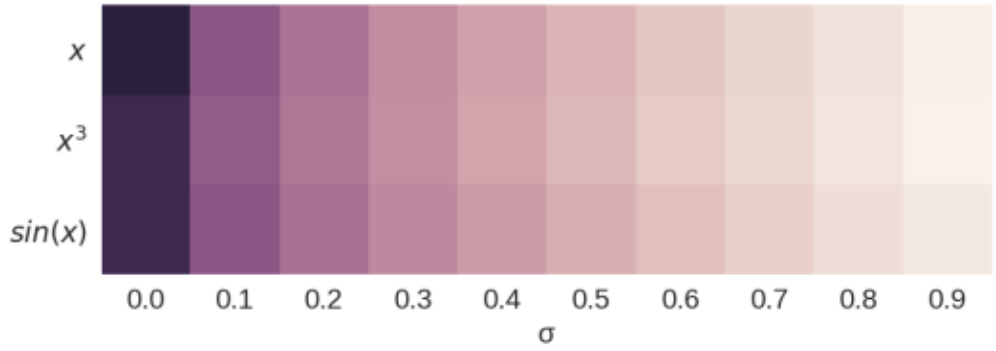


Figure 1: The heatmap illustrates MINE's invariance to deterministic nonlinear transformations. It shows mutual information estimates computed by MINE between 2-dimensional random variables $X \sim \mathcal{U}(-1, 1)$ and $Y = f(X) + \sigma \odot \epsilon$, where $f(x) \in \{x, x^3, \sin(x)\}$ and $\epsilon \sim \mathcal{N}(0, I)$. The estimates remain consistent across different functions f, highlighting MINE's equatability property [24].