BRAINIACS
BRAIN IMAGING AND COMPUTING SCIENCES

2024 Volume 5
Issue 2 Edoc XEE8F298E

Quantifying Similarity between Graph-Theoretic Resting-State fMRI Data Processing Pipelines for Efficient Multiverse Analysis*

Micha Burkhardt, Andrea Hildebrandt, Carsten Gießing, Daniel Kristanto†

Abstract

Multiverse analysis aims to enhance the robustness and replicability of scientific findings by testing research hypotheses through multiple, well-justified analysis pipelines. However, the multiverse of pipelines is often large making exhaustive evaluation computationally infeasible. Thus, a key goal is to approximate the multiverse by sampling a manageable number of pipelines for robustness analysis. For such an approximation, it is necessary to quantify the similarity between analysis pipelines and guide pipeline sampling by these similarities. To this end, we first used meta-analytic data from Kristanto et al. (2024) on fMRI processing pipelines collected from a representative set of papers. Using this meta-analytic data, we propose a Graph Convolutional Network (GCN)-based approach combined with Deep Graph Infomax (DGI) to assess pipeline similarity. Graph-based embeddings were computed using unsupervised learning and subsequently used to derive pipeline features. Pipeline similarity was then quantified via Euclidean distance. Traditional similarity measures, namely Jaccard, Hamming and Levenshtein distances were also computed based on the meta-analytic data for comparison. Clustering analysis revealed consistency across the GCN, Hamming, and Levenshtein measures. Similarity measures based on Hamming and Levenshtein distances treated all processing steps identically, thus biasing them towards pipelines with identical step lengths. In contrast, the GCN-based measure generated distinct features for each step, allowing each to contribute differently to the pipeline similarity measure. Second, we compared the meta-analytically derived pipeline similarity measures with similarity measures obtained from multiverse analysis conducted on empirical data using resting-state fMRI measures from the Human Connectome Project. The comparison showed satisfactory results for the proposed approach, which aims to replace empirical similarity with meta-analytic similarity estimates for computationally efficient multiverse analysis in graph-theoretic fMRI research. These findings will inform future studies aimed at validating meta-analytic pipeline similarity measures based on empirical similarity estimates, providing a solid basis for the development of computationally feasible and valid multiverse analyses.

Keyphrases

Resting-state fMRI, multiverse analysis, data processing pipeline, graph neural network, similarity metric.

Contents

*Presented 2024-10-09 at Guardians 2024 with slides and video.

†Correspondence to micha.burkhardt@uol.de.

 © 2024 BHA

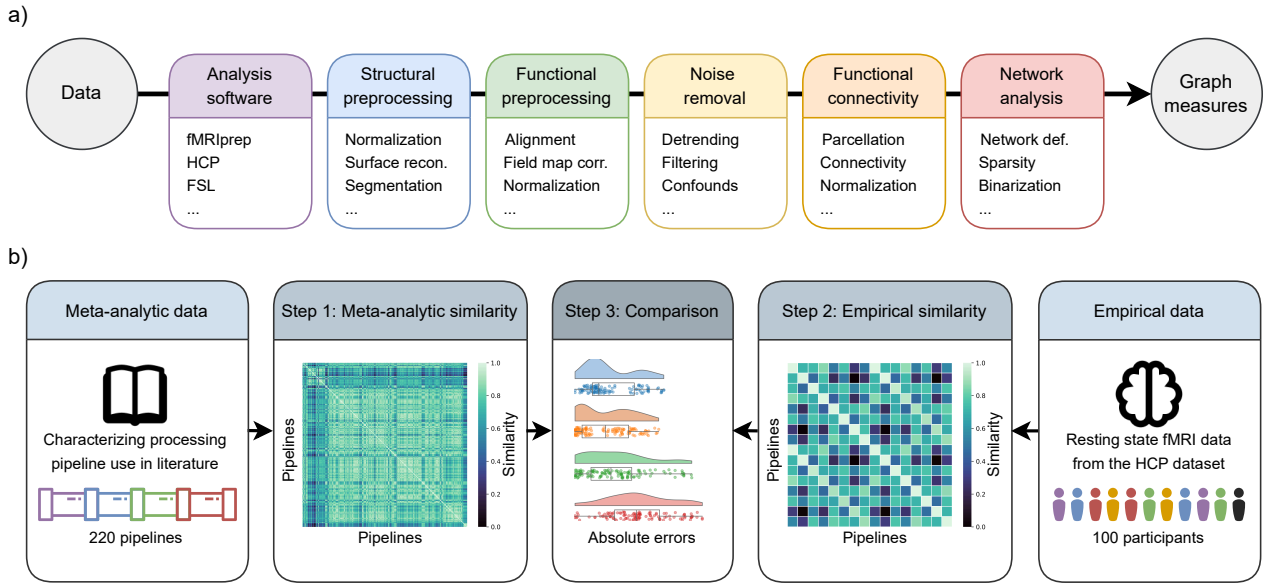

Figure 1: Pipeline Data: 220 graph analytical fMRI preprocessing/analysis pipelines were derived from literature. In total, these pipelines contain 61 distinct analysis steps, which can be grouped into six conceptual categories: Analysis software, structural preprocessing, functional preprocessing, noise removal, functional connectivity estimation, and network analysis.

Introduction

In many fields of computational science, researchers face a plethora of arbitrary yet defensible decisions when designing studies and analysing data. Given this multiplicity of decisions, also known as the many *researcher's degrees of freedom*, particular choices can inadvertently introduce bias and contribute to the ongoing replication crisis in science (Simmons et al. 2011). This issue is particularly pronounced in cognitive neuroscience, where the complexity of neuroimaging data requires extensive preprocessing and analysis pipelines to handle the inherent noise and complexity of the data (Kristanto et al. 2024). Consequently, methodological choices were shown to impact the robustness of results (Botvinik-Nezer et al. 2020), highlighting the need for more thorough data analysis practices.

In light of these challenges, there have been increasing calls to address the robustness of findings in scientific research (Open-Science-Collaboration 2015; Frias-Navarro et al. 2020), and to prioritize statistical replicability over narrative appeal (Huber et al. 2019). Rather than only reporting a subset of the findings in line with a planned story, researchers are urged to prioritize transparency, replicability, and methodological rigor to improve the credibility of findings. This shift in focus is critical for advancing the field, ensuring that results are not only compelling but also robust.

Towards such a shift, multiverse analysis has recently been proposed as an approach to enhance the robustness of research findings (Steegen et al. 2016; Del Giudice and Gangestad 2021). In contrast to just performing (and reporting) a single analysis, multiverse analysis involves running statistical tests over a wide range of specifications. This approach not only reveals whether different specifications lead to similar results but also offers exploratory insights. For instance, hidden structures in the data could include patterns such as latent clusters, nonlinear relationships, or variable interactions that become visible only under certain analytical choices. Additionally, the methods analysed may reveal clusters depending on the characteristics of the data, such as sensitivity to noise or differences in model assumptions. However, implementing multiverse analysis can be computationally expensive, especially in fields like neuroimaging research, where a vast number of analytical decisions are available. To address this challenge, a recent study proposed an active learning approach for multiverse analysis (Dafflon et al. 2022). This approach creates a search space of pipelines by running all analysis pipelines on a subset of the data and quantifies their similarity based on the outputs (e.g., graph measures derived from fMRI data). An active learning algorithm then samples and tests a small subset of pipelines from the search space. Specifically, the algorithm uses these samples to model associations between pipeline features such as analytical decisions, and research outcomes (e.g., predicting cognitive scores from brain data). This allows the algorithm to infer outcomes for the remaining pipelines in the search space without running them.

While promising, this method has limitations. For example, constructing the search space requires running all pipelines on a subset of the data, which may be computationally infeasible for large pipeline spaces. In addition, the same data cannot be used to construct the search space and test hypotheses without introducing bias due to circular analysis strategies. Thus, when the sample size is small, loss of data for the main analysis becomes a problem, reducing statistical power. Developing alternative methods to construct the search space without these limitations is therefore a key focus for multiverse analysis research, especially in computationally intensive fields or when large samples are unavailable.

Related research has also sought to address the issue of low robustness and replicability with neuroimaging pipelines. For example, almost two decades ago, Strother (2006) already highlighted that inconsistencies in testing environments and performance metrics hinder the generalisability of findings, and advocated for balanced approaches that not only evaluate isolated analytical steps but also the entire pipeline. More recently, studies such as Bowring et al. (2022) and Luppi et al.

 © 2024 BHA

(2024) have systematically evaluated sources of variability and benchmarked pipeline performance to enhance consistency and robustness in neuroimaging research. However, as the active learning approach, these methods require running all pipelines on the data of a given study, which is computationally intensive. This limitation becomes particularly problematic for multiverse analyses involving large numbers of pipelines, where computational feasibility is a key concern.

To overcome these limitations, we propose replacing the computationally expensive and data-intensive process of constructing a search space of pipelines with a similarity measure based on the configuration of the analysis pipelines as used in the literature. Instead of running pipelines on subsets of data, this approach uses information about the analysis steps and pipeline similarities based on how they are used and reported in the literature for addressing similar research questions. In this context, the 'configuration' of a pipeline refers to the sequence and specific choices of analysis steps that constitute the pipeline, such as preprocessing, feature extraction, and statistical modeling. Step-based pipeline similarity derived from meta-analytic data has garnered attention as a way to streamline multiverse analyses and integrate results efficiently. For instance, it has been suggested that the number of pipelines in a multiverse analysis could be reduced by grouping similar ones, based on the assumption that similar pipelines produce similar outcomes (Cantone and Tomaselli 2024). However, whether this assumption holds true remains an empirical question, as individual analysis steps can decisively alter the data.

Traditional similarity measures used for sequences, such as Jaccard, Hamming, and Levenshtein distances, each have specific strengths and limitations in assessing the similarity of pipelines effectively (Jaccard 1901; Hamming 1950; Levenshtein 1966). For example, Hamming distance detects localised differences by counting mismatches at corresponding positions. Levenshtein distance accounts for edits like substitutions, insertions, and deletions, making it more flexible, but it treats steps as isolated and ignores their relationships. Jaccard similarity measures overlap between sets of elements but disregards the order and structure of sequences. Thus, while these measures are effective for identifying differences or shared components, they might fail to capture the broader, structural relationships that often define processing pipelines as they are not just linear sequences but represent interconnected processes where the order and interdependence of steps carry significant meaning. Traditional similarity measures overlook this global context, making them less effective for accurately comparing pipelines in complex domains like fMRI data processing.

Building on these efforts, we introduce a novel method for assessing pipeline similarity using a Graph Convolutional Neural Network (GCN) combined with Deep Graph Infomax (DGI; Veličković et al. (2018)). Our approach generates graph-based embeddings to capture the relationships between processing steps across entire pipelines by using meta-analytic data indicating how frequently the pipelines are used in the literature. These embeddings are concatenated to form feature representations of pipelines, enabling similarity measurement based on Euclidean distance. Unlike traditional measures, this approach accounts for the structural and contextual relationships between processing steps. We applied this approach to a meta-analytic dataset of 220 fMRI analysis pipelines derived from the literature (Kristanto et al. 2024), estimating their similarity by using features such as the frequency and order of processing steps. To evaluate our approach, we compared the GCN-based similarity measure with traditional measures (Jaccard, Hamming, and Levenshtein distances). We analysed the behavior of these measures and highlighted their differences in the context of comparing fMRI processing pipelines. Additionally, we conducted an empirical multiverse analysis using data from 100 participants of the Human Connectome Project (HCP; Van Essen et al. (2013)). This allowed us to benchmark the GCN approach against empirical results, demonstrating that the GCN-based similarity measure shows promising results by capturing some (but not all) patterns in the data. The present study thus highlights the potential of GCN-based meta-analytic similarity measures for efficient multiverse analysis in computationally intensive fields like neuroimaging. We will discuss how such GCN-based measures can be integrated into frameworks to reduce computational costs, improve methodological rigor, and enhance the robustness of scientific findings.

Methods

The primary aim of this study is to systematically explore algorithms for quantifying the similarity of fMRI processing pipelines based on their analysis steps, as applied in the literature (Kristanto et al. 2024). Specifically, we investigate how different similarity measures, including traditional metrics and a novel Graph Convolutional Network (GCN) based approach, capture patterns of consistency and discrepancy across pipelines. To validate these measures, we compare their estimates to empirical similarity derived through a multiverse analysis using real data.

fMRI Experimental Data

For the empirical multiverse analysis to be compared with the meta-analytic similarity, we used minimally processed data from the publicly accessible Human Connectome Project (HCP) Young Adult dataset (https://www.humanconnectome.org/study/hcp-young-adult). This dataset comprises healthy individuals aged between 22 and 35 years, from which we randomly selected 100 individuals for subsequent analysis. From these, we used the openly available resting-state time-series data, which was cleaned through the HCP minimal processing pipeline (Glasser et al. 2013) and parcellated into 400 cortical regions of interest using the Schaefer et al. (2018) atlas.

fMRI Data Processing Pipelines

The analysis pipelines and associated meta-analytic data characterizing their applications in the literature used in the present study were derived from a systematic literature review, which specifically focused on graph-based methods for fMRI studies (Kristanto et al. 2024). The comprehensive review identified a total of 61 distinct preprocessing and analysis steps commonly employed across studies, with 17 of these steps representing often debated options such as data scrubbing, brain parcellation, or spatial smoothing, which can significantly influence the outcome of fMRI analyses. We grouped the steps based on their functional contribution to processing pipelines to outline the common workflow across pipelines. These groups are: Analysis software, structural preprocessing, functional preprocessing, noise removal, functional connectivity definition, and network analysis (Figure 1a). In total, 220 pipelines were derived, which in the present study serve as the core underlying data for the meta-analytic similarity measures. We emphasize that the list of pipelines used in this study, albeit aiming at different research questions, share a common goal, which is to estimate graph measures from functional connectivity. The meta-analytic data on fMRI processing pipelines contain the following node and edge relevant information: Steps in the pipeline and the frequency of their usage in

 © 2024 BHA

the literature, functional group to which the step belongs (e.g., software selection, structural or functional preprocessing, etc., see also Figure 1a), neighboring steps to which a step is connected, the number of studies in the literature that used a corresponding pair of processing steps consecutively, incoming connections (in-degree), outgoing connections (out-degree).

Traditional Similarity Measures

We first derive similarity measures from three well-established metrics in machine learning and bio-informatics: The Jaccard Index as well as Hamming and Levenshtein distances.

Jaccard Index: The Jaccard index quantifies the proportion of data processing steps that are shared between pipelines (Jaccard 1901). By representing each pipeline as a set of steps, the ratio of the size of the intersection (i.e., the common steps between two pipelines) to the size of the union (i.e., all the unique steps across the two pipelines) is calculated. For example, consider two sets of steps: $A = \{1, 2, 3\}$ and $B = \{2, 3, 4, 5\}$. The intersection $A \cap B$ contains the common steps $\{2, 3\}$, and the union $A \cup B$ contains all unique steps $\{1, 2, 3, 4, 5\}$. The Jaccard index is calculated as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{5} = 0.4 \quad (1)$$

It is important to note that the Jaccard index does not account for the order of elements, meaning that it only considers the presence or absence of elements within the sets, regardless of their sequence.

Hamming Distance: The Hamming distance quantifies the dissimilarity between two pipelines by comparing the sequences of processing steps (Hamming 1950). Each pipeline is represented as an ordered sequence, and the Hamming distance is defined as the number of mismatched steps between two pipelines when they are aligned step-by-step. For example, consider two binary strings representing processing steps: '10101' and '10011'. The Hamming distance between these strings is 2, because they differ at the third and fifth positions. Similarly, if two pipelines have identical steps but in different orders, the Hamming distance will be non-zero, reflecting these positional discrepancies.

This metric is particularly useful in scenarios where the order of steps is critical to the outcome. We use the complement of the distance as a measure of similarity. Unlike the Jaccard index, which only considers the presence or absence of steps, the Hamming distance accounts for the order of the steps by calculating the number of positions at which the corresponding steps in two pipelines differ.

Levenshtein Distance: The Levenshtein distance estimates the distance between two pipelines by measuring the minimum number of single-step edits required to transform one pipeline into the other (Levenshtein 1966). These edits can include substitutions, insertions, or deletions of processing steps. For example, the Levenshtein distance between the strings 'kitten' and 'sitting' is 3, as it involves two substitutions ('k' $\rightarrow$'s' and 'e' $\rightarrow$ 'i') and one insertion ('g' at the end). Unlike the Hamming distance, the Levenshtein distance accounts for sequences of different lengths by incorporating these insertion and deletion operations. The Levenshtein distance thus provides a way to assess how similar or different two pipelines are, considering both the order of steps and the specific modifications needed to align one sequence with the other. This metric is particularly useful in scenarios where small differences between pipelines—such as an extra step or a substituted processing method—can have significant implications.

Moreover, we also implemented the Damerau-Levenshtein distance in our analysis, which is an extension of the Levenshtein distance that additionally accounts for adjacent transpositions (i.e., swapping two neighboring elements). This extension is particularly relevant in scenarios where adjacent transpositions are a common source of variation between sequences. However, we found highly similar results as with the traditional Levenshtein distance (a correlation value of 1 between both distance measures). Results for the Damerau-Levenshtein distance are available in the supplementary Python notebooks.

Table 1: Pipeline decisions for multiverse analysis. HCP: Human Connectome Project, WM/CSF: White matter and cerebrospinal fluid regressors.

Pipeline Step	Parameter(s)
Preprocessing	HCP minimal processing pipeline
Cleaning	None
	6-parameter movement, WM/CSF
	Global signal regression
	All combined
Temporal filtering	None
	Band-pass (0.01 - 0.1 Hz)
Parcellation	Schaefer 400
Network construction	Discard negative, 50% density
Graph measure	Global efficiency

Graph Convolutional Network (GCN)

We propose a new way of measuring the similarity between analysis pipelines, which utilizes a Graph Convolutional Network (GCN) combined with Deep Graph Infomax (DGI). Unlike traditional methods, this approach encodes each analysis step in the pipeline as a distinct feature vector, with information provided both by the step itself and by external features (such as its functional group as depicted in Figure 1; see Section for details). The GCN also learns from neighboring steps in the pipeline by aggregating information from adjacent nodes, allowing the model to capture relationships between steps. This enables the model to weigh each step differently based on its role and connections in the pipeline, which in turn influences the similarity scores between pipelines. An advantage of this method is that it reflects both the presence of steps and how they are used in detail, making it more representative of real-world differences in processing pipelines.

Network Construction: The aggregate of the analysis pipelines derived from the literature can be analysed as a weighted and directed graph. Here, the nodes of the graph are the individual processing steps in the pipeline (e.g., spatial normalization, motion regression, parcellation), and the weighted directed edges are the number of studies in the literature that used the corresponding pair of processing steps consecutively. We also included nodal features, namely the frequency of a step (number of studies that applied the step), its incoming connections (in-degree), outgoing connections (out-degree), individual identity, and a group identity (e.g., structural or functional preprocessing, functional preprocessing, noise removal).

The GCN was then combined with the DGI algorithm to learn node representations for the fMRI processing pipeline in an unsupervised manner. The GCN generated initial embeddings by aggregating information from each node's neighbors, capturing local structural and feature information. DGI then refined these embeddings by introducing corrupted versions of the graph and training the model to distinguish between true and corrupted data (Veličković et al. 2018). This pro-

 © 2024 BHA

Table 2: Multiverse Analysis Pipeline Configurations. GSR: Global signal regression, WM/CSF: White matter and cerebrospinal fluid regressors.

Pipeline	Step i+1	Step i+2	Step i+3
Pipeline 1	Parcellation: Schaefer 400	Confounds: none	Band-pass filtering
Pipeline 2	Parcellation: Schaefer 400	Confounds: none	No filtering
Pipeline 3	Parcellation: Schaefer 400	Confounds: GSR	Band-pass filtering
Pipeline 4	Parcellation: Schaefer 400	Confounds: GSR	No filtering
Pipeline 5	Parcellation: Schaefer 400	Confounds: motion + WM/CSF	Band-pass filtering
Pipeline 6	Parcellation: Schaefer 400	Confounds: motion + WM/CSF	No filtering
Pipeline 7	Parcellation: Schaefer 400	Confounds: GSR + motion + WM/CSF	Band-pass filtering
Pipeline 8	Parcellation: Schaefer 400	Confounds: GSR + motion + WM/CSF	No filtering
Pipeline 9	Confounds: none	Parcellation: Schaefer 400	Band-pass filtering
Pipeline 10	Confounds: none	Parcellation: Schaefer 400	No filtering
Pipeline 11	Confounds: GSR	Parcellation: Schaefer 400	Band-pass filtering
Pipeline 12	Confounds: GSR	Parcellation: Schaefer 400	No filtering
Pipeline 13	Confounds: motion + WM/CSF	Parcellation: Schaefer 400	Band-pass filtering
Pipeline 14	Confounds: motion + WM WM/CSF	Parcellation: Schaefer 400	No filtering
Pipeline 15	Confounds: GSR + motion + WM/CSF	Parcellation: Schaefer 400	Band-pass filtering
Pipeline 16	Confounds: GSR + motion + WM/CSF	Parcellation: Schaefer 400	No filtering

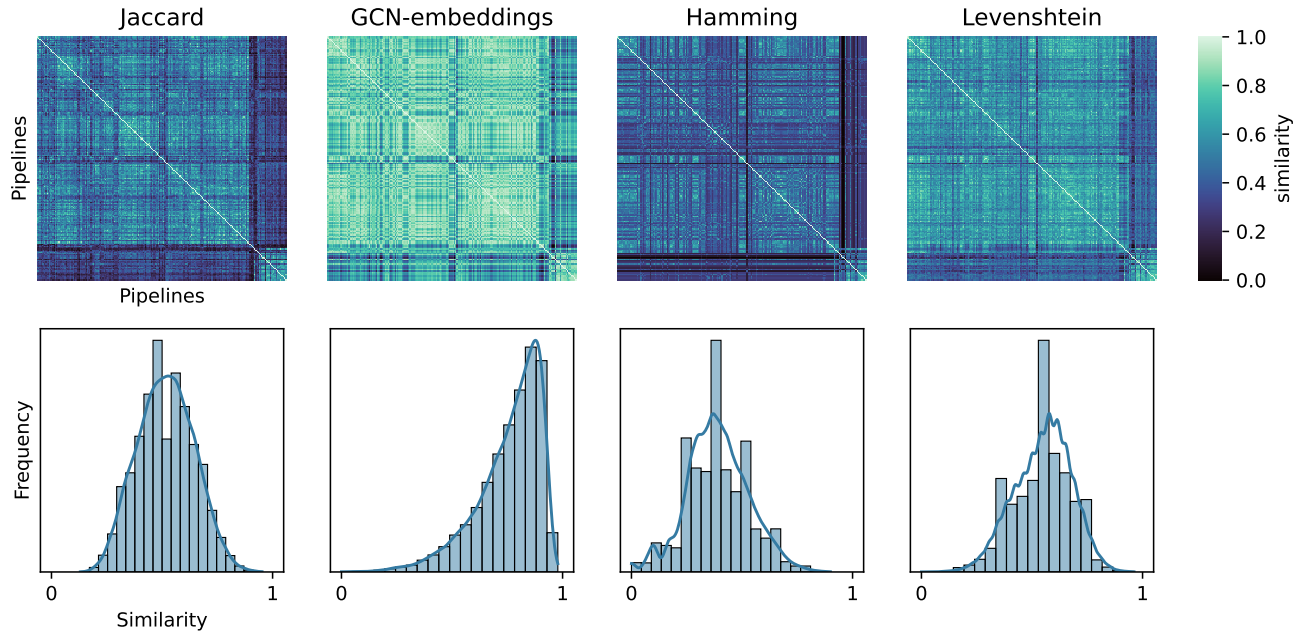

Figure 2: Similarity and distribution of similarity values for the different measures. K-means clustering was performed for the embedding similarity (mean) and all other matrices were ordered accordingly. The figure highlights a considerable overlap between measures.

cess maximizes mutual information between node embeddings and high-level summaries of the graphs, resulting in robust and informative representations for downstream tasks like the proposed estimation of pipeline similarity. As the number of layers in a GCN increases, the network aggregates information from nodes that are increasingly distant in the graph. We chose to use a single layer for our network architecture to focus on local relationships between processing steps only, which helps reduce the influence of distant, potentially less relevant connections. This approach is particularly useful given that not all combinations of processing steps are valid in an analysis pipeline, and focusing on local interactions helps to mitigate the risk of capturing implausible sequences in the embeddings.

Embedding Aggregation: The output embeddings of a GCN correspond to the neurons in its output layer. However, since fMRI processing pipelines vary in length, aggregating these embeddings into a consistent format is a challenge. To preserve the sequential nature of the pipelines, we apply Dynamic Time Warping (DTW), which measures similarity between temporal sequences, allowing for flexible, non-linear alignment of steps and accounting for differences in pipeline length or step ordering (Sakoe and Chiba 1978). Notably, DTW was implemented by treating consecutive steps as being one unit of time apart in sequences. This approach is thus similar to the implementation of Levenshtein distance, with the important distinction that, in DTW, each step is represented by embeddings learned during GCN training, whereas in Levenshtein distance, steps are simply represented by their discrete labels. To distinguish this GCN-based similarity measure from the network itself, we will refer to it as "GCN-embeddings".

In detail, the trained GCN generates embeddings for each node (pro-

 © 2024 BHA

cessing step), where each embedding is a vector corresponding to the network architecture (e.g., a 32-dimensional vector for a single-layer network with 32 neurons). Each pipeline is represented as a list of these embeddings, with the length of the list matching the number of steps in the pipeline. After aggregating the step embeddings into pipeline features, we computed the similarity between pipelines using Euclidean distance.

Normalization Across Measures

To facilitate the comparison between similarity measures, all measures were transformed and scaled into a common scale ranging from 0 (completely dissimilar) to 1 (completely similar):

$$Similarity(D_{ij}) = 1 - \frac{D_{ij} - \min(D)}{\max(D) - \min(D)} \quad (2)$$

with D being the distance matrix for each measure (GCN-embeddings, Jaccard index, Hamming distance, and Levenshtein distance).

Empirical Multiverse Analysis

As a final analysis, we conducted a real-data multiverse analysis to establish an empirical ground truth for pipeline output similarity. This ground truth served as a benchmark to compare the performance of the previously introduced similarity measures. Due to the computational challenges of performing a comprehensive multiverse analysis across all structural and functional preprocessing steps, we focused on the later stages of a standard graph analysis pipeline. Using minimally preprocessed data from the HCP Young Adult dataset, we randomly selected 100 individuals for subsequent analysis. For these individuals, we computed the graph measure, global efficiency, across different analysis pipelines as shown in Table 1. The multiverse analysis was implemented using the Comet toolbox (Burkhardt and Giessing 2024), which provides an integrated framework for functional connectivity, graph analysis, and multiverse analysis.

Analysis pipelines begin with identical preprocessing steps (the HCP minimal processing pipeline; Glasser et al. (2013)) but differ in their noise reduction strategies, which included four confound regression options (none, 6-parameter movement + white matter (WM) + cerebrospinal fluid (CSF), global signal, and both combined) as well as two filtering strategies (none, band-pass filtering between 0.01 and 0.1 Hz). Additionally, we altered the order of data cleaning and parcellation, resulting in two configurations: cleaning performed before parcellation or after. Since the total number of pipelines is the Cartesian product of these decisions, the multiverse comprised 16 pipelines ($2 \times 4 \times 2$).

The remaining parameters were kept consistent across all pipelines. This included parcellation, temporal detrending, calculating functional connectivity using Pearson correlation, constructing graph networks (removing negative correlations and thresholding to 50% density of the network), and computing global efficiency for each participant. To estimate similarity between pipelines, we computed the Pearson correlation of global efficiency values between pairs of pipelines across individuals, resulting in a 16×16 empirical similarity matrix. This matrix was used as a ground truth reference to evaluate the proposed similarity measures, which estimated similarity based solely on the steps in the pipelines without running them on real data.

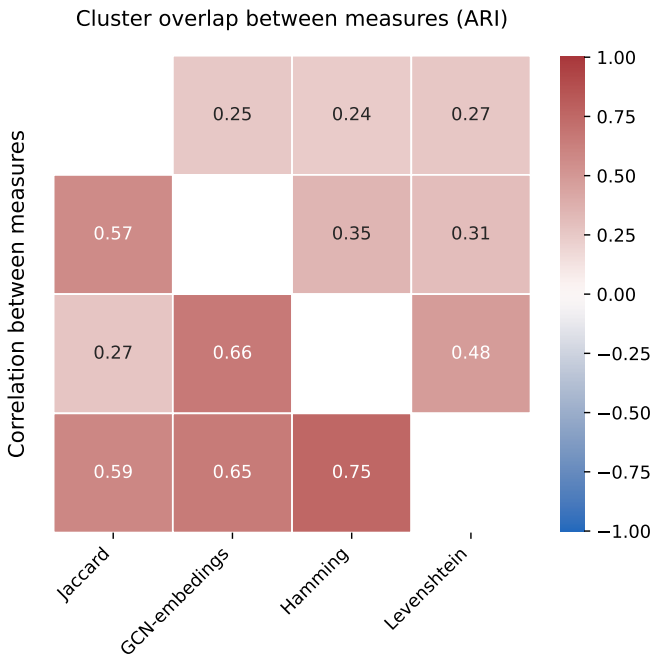

Figure 3: Correlation and Adjusted Rand Index (ARI) matrices between similarity measures. Most measures share a reasonable amount of variance. The adjusted rand index (ARI) is highest for methods which take the order of steps into account (DTW, Hamming, and Levenshtein).

Results

Comparative Analysis

We first trained the GCN and computed pipeline similarity based on the pipeline features. The training of the model is shown in the supplementary Python notebooks. Next, we computed pipeline similarity using other measures (Jaccard index, Hamming distance, and Levenshtein distance). We then compared the similarity estimates as shown in Figure 2. It becomes clear that there is a considerable overlap between the measures, as indicated by the moderate to high correlation between them (Figure 3). Interestingly, GCN-embeddings shows reasonably high correlation with Hamming ($r = .66$) and Levenshtein ($r = .65$). Further, the distributions of the similarity estimates show considerable differences. Similarity estimates derived from GCN embeddings are left-skewed and thus generally show higher similarity between pipelines. Hamming and Levenshtein distances display less smooth characteristics compared to the other methods, and the Jaccard index based similarity measures appear normally distributed.

Cluster Overlap

A more nuanced understanding of the resulting similarity matrices from different methods can be obtained by using the Adjusted Rand Index (ARI) (Hubert and Arabie 1985). ARI is a measure of the agreement between partitions obtained from a clustering approach. For this comparison, each similarity matrix was clustered into four groups, with the optimal number of clusters determined using the elbow criterion (see accompanying Python notebook for details). As shown in Figure 3, the ARI values were highest for DTW, Hamming, and Levenshtein, meaning that the three measures that account for the step order in the pipelines also show the highest cluster overlap.

 © 2024 BHA

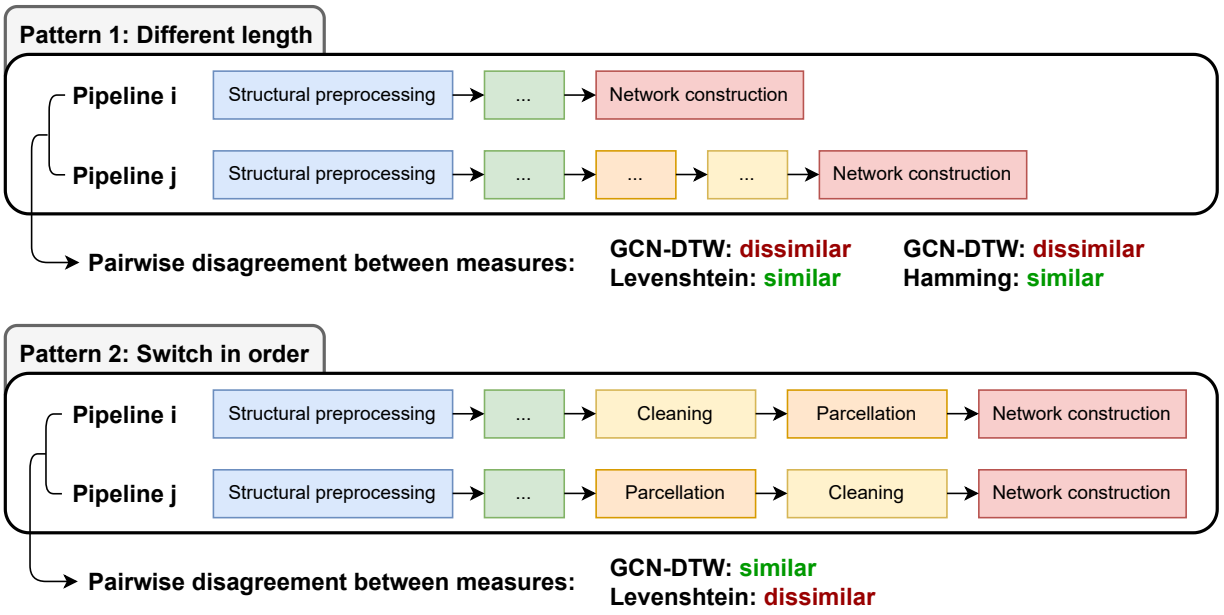

Figure 4: Differences between measures. Measures were evaluated in more detail by observing pairs of pipelines which show highest disagreement between measures. Only pipelines which consider the order of the steps were included. Two distinct patterns emerge. Top: Pattern 1 concerns pipelines of different length. Comparing GCN-embeddings to Levenshtein, GCN-embeddings considers such a pair of pipelines to be more dissimilar, while Levenshtein considers these pipelines to be more similar. The same pattern holds when comparing GCN-embeddings to Hamming. Bottom: Pattern 2 concerns a switch in order between cleaning and parcellation. GCN-embeddings considers these pipelines to be more similar, while Levenshtein considers these pipelines to be more dissimilar.

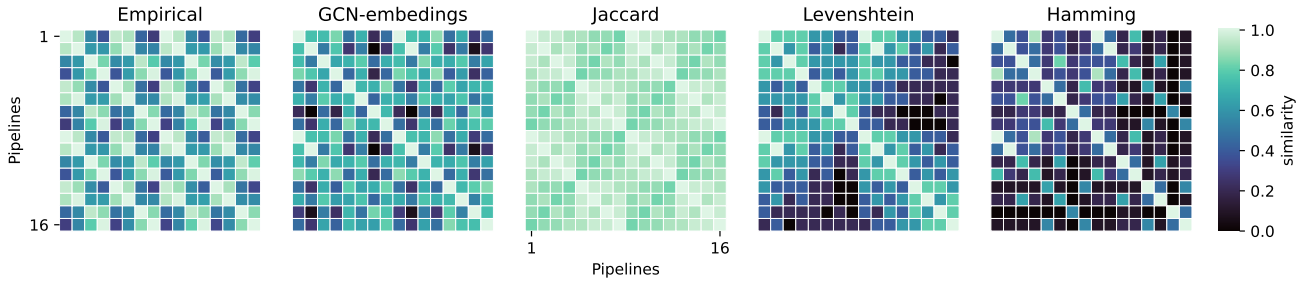

Figure 5: Similarity estimates from pipelines as listed in Table 2. Empirical similarity estimates show a strong grid-like pattern, which can be attributed to global signal regression. The ordering of steps (cleaning before or after parcellation) does not show a noticeable influence. Comparing the empirical similarity to the predicted similarities showed that all measures fail to pick up on this pattern, but mostly group pipelines into blocks of 4 (performing/not performing confound regression for motion, WM, and CSF). Please refer to the multiverse_summary.csv file in the supplementary materials for a detailed description of the pipelines.

Differences between Pipelines

Given the above findings, we further examined the emerging patterns produced by these three measures. We were particularly interested in pairs of pipelines where the similarity values computed by these measures were highly different. We therefore performed a pairwise comparison between the three similarity measures (i.e., GCN-embeddings vs. Hamming, GCN-embeddings vs. Levenshtein, and Hamming vs. Levenshtein) by using their similarity matrices shown in Figure 2. For each pair, we then extracted the 10 items with the highest absolute difference, that is, the 10 pairs of pipelines for which the measures most highly disagree in their similarity estimate. To account for a potential bias in the GCN, we repeated the entire process (including re-training of the GCN) 10 times resulting in 100 pairs of pipelines for each pairwise comparison.

We then investigated the origin of the differences in similarity estimates and found two distinct patterns (Figure 4). The first pattern emerged from the pairs of pipelines with different length. For the 100 pairs of pipelines for which GCN-embeddings and Levenshtein most highly disagree, the average difference in length was 10.01 processing steps, with GCN-embeddings judging the pair of pipelines to be less similar in all 100 cases. The same pattern can also be observed in the pairwise comparison between GCN-embeddings and Hamming. There, the average difference in pipeline length was 6.85, with GCN-embeddings considering such pipelines to be less similar in 62 out of 100 times. These results indicated that the Hamming-based measure was less

 © 2024 BHA

sensitive to these pipeline pairs compared to the GCN-embeddings measure. It is important to note that while these pipeline pairs differed in length, a significant portion of their steps were identical. Specifically, they shared minimal preprocessing pipelines from the Human Connectome Project, including identical steps for structural and functional preprocessing. The primary differences lay in subsequent noise reduction steps. This similarity in the early, substantial portion of the pipelines may explain why the Hamming distance identified these pairs as more similar than GCN-embeddings. In GCN-embeddings, each step is represented by its own embedding. Some steps may have larger embeddings than others, potentially leading to the identification of these pipeline pairs as less similar.

The second pattern highlighted a switch in order between brain parcellation and steps related to noise removal such as temporal filtering and motion regression. More specifically, we evaluated this order by assessing whether cleaning was performed in a high dimensional brain space (voxel level or high-resolution surface mesh), or on parcellated brain signals (groups of voxels/vertices clustered together into functionally distinct brain regions). This pattern emerged in 14% of comparisons between GCN-embeddings and Hamming, in 46% of comparisons between GCN-embeddings and Levenshtein, but in 0% of comparisons between Hamming and Levenshtein, indicating that the GCN-embeddings measure is robust to this pattern. Further, the comparison between GCN-embeddings and Levenshtein revealed that pairs of pipelines with this pattern were seen as more similar by GCN-embeddings and less similar by Levenshtein.

Empirical Multiverse Analysis

To evaluate the effectiveness of the meta-analytic pipeline similarity measures, we compared them with empirical similarity obtained by running pipelines on real MRI data as described in Methods). A total of 16 analysis pipelines were applied to the HCP dataset, and their empirical similarities were computed.

Figure 5 shows the empirical as well as the predicted meta-analytic similarity matrices of the analysis pipelines listed in Table 2. For the empirical similarity, a grid-like pattern becomes apparent. This can be attributed to global signal regression (GSR). Pipelines with GSR (Pipelines 3, 4, 7, 8, 11, 12, 15, 16) demonstrated high similarity to one another but low similarity to pipelines without GSR (pipelines 1, 2, 5, 6, 9, 10, 13, 14), and vice versa. Although not a particular focus of the present study, this finding once again outlined the significant impact of GSR on analysis pipeline results. The meta-analytic similarity estimates failed to pick up on this pattern, but more closely picked up the pattern of performing/not performing confound regression for motion, WM, and CSF signals. This lead to blocks of 4 being more pronounced in their estimates, which are most visible for Hamming, but also for GCN-embeddings and Levenshtein. Notably, the order of performed steps (first 8 vs. second 8 pipelines) did not play a major role for differences in similarity. For example, pipelines performing cleaning after parcellation (Pipelines 1, 2, 5, 6) were highly similar to pipelines performing cleaning before parcellation (Pipelines 9, 10, 13, 14).

Finally, we computed the absolute errors between the empirical and meta-analytic pipeline similarities. Figure 6 displays the distribution of absolute errors for each measure. GCN-embeddings showed the lowest median absolute error (MAE) of 0.18, with Jaccard (MAE = 0.23), and Levenshtein (MAE = 0.26) trailing closely behind. Hamming showed a substantially higher MAE of 0.45. Please refer to the supplementary Python notebooks for element-wise error matrices. Despite their reasonable performance, all measures failed to capture the influence of critical processing steps (in this case GSR) on pipeline similarity. We will later discuss potential extensions to GCN-embeddings method, informed by empirical evidence, to better account for the substantial influence of certain processing decisions on pipeline performance.

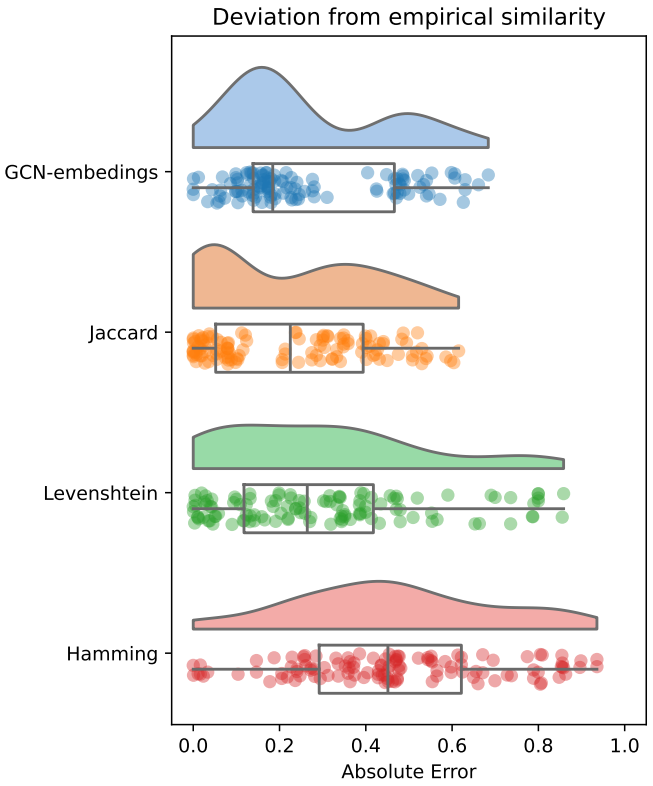

Figure 6: Comparison to ground truth. Comparing the estimated similarity between analysis pipelines to the ground truth from multiverse analysis, GCN-embeddings displays the lowest median absolute error (MAE) of 0.18. Jaccard (MAE = 0.23) and Levenshtein (MAE = 0.26) also produce similar estimates, while Hamming (MAE = 0.45) shows a considerably worse performance.

Discussion

In the present study, we elaborated on measuring similarity between processing pipelines in the context of multiverse analysis for fMRI studies, where pipelines involve complex sequences of steps. We introduced a novel meta-analytic pipeline similarity measure based on a Graph Convolutional Network (GCN) and compared it to traditional measures like Jaccard index, Hamming distance, and Levenshtein distance. Unlike these traditional measures, our GCN-based approach (GCN-embeddings) allows for varying contributions from different steps when computing pipeline similarities. Put simply, while traditional measures treat all steps equally, GCN-based measures assign individual weights to different steps. We expected that this would enable GCN-based measures to capture both consistent and distinct similarity/dissimilarity patterns in pipelines compared to the traditional measures.

We examined the similarity between GCN-embeddings and other traditional measures. Figures 2 and 3 show that the pipeline similarity estimates of GCN-embeddings overlap with those of other measures, suggesting that a GCN-based approach can also capture relevant pat-

 © 2024 BHA

terns. Moreover, a detailed analysis comparing pipeline partitions obtained by these measures revealed that GCN-embeddings was more consistent with Hamming and Levenshtein similarity measures but not with the Jaccard index. This finding was expected, as GCN-embeddings, Hamming, and Levenshtein consider the order of steps when calculating pipeline similarity.

It is important to clarify that the focus of the present work was not to validate pipelines for specific cognitive neuroscience hypothesis testing efforts. Instead, we relied on a meta-analytic dataset consisting of pipelines designed for different research purposes, unified by the common aim of estimating graph measures from functional connectivity. The primary contribution of this study was the development of a computationally efficient framework for quantifying pipeline similarity, which is critical for subsampling the multiverse of analytical decisions in a manageable and representative way, allowing researchers to explore variability across pipelines without requiring the exhaustive evaluation of all possible combinations.

Therefore, the validity of specific pipelines to test a particular hypothesis in cognitive neuroscience and its interpretations is outside the scope of this work. Instead, our focus was to assess whether meta-analytic data on the use of pipelines across multiple individual studies in graph-theoretic fMRI analyses can be used to effectively estimate pipeline similarity that approximates well empirical similarity measures, and which could be used to design multiverse analyses and efficient sampling from the multiverse in situations where the multiverse cannot be computed exhaustively but only approximated. While the present work advances methods for multiverse analysis, future studies could expand upon this framework by integrating hypothesis-specific considerations and further validating the approach in the context of specific cognitive neuroscience experimental paradigms. For now, the approach is aimed to serve as a methodological tool to facilitate efficient subsampling and variability assessment within the multiverse, independent of the specific experimental context.

Patterns in Similarity Discrepancies

Through a more in-depth analysis focusing only on the similarity measures that take the order of processing steps into account, we identified patterns in how these measures distinguish similar and dissimilar analysis pipelines. We focused on pairs of pipelines that exhibited the greatest discrepancies in similarity values computed by these measures. The first pattern was found in pipeline pairs that have different length. For example, two pipelines might use the Human Connectome Project (HCP) minimal preprocessing pipeline in earlier steps, but differ in length in the later part of the pipeline for cleaning or network construction. Hamming and Levenshtein score such pairs as highly similar due to the large number of common steps, while GCN-embeddings assigns a lower similarity score. This can be explained by examining processing step embeddings computed by the GCN, where steps related to network reconstruction have higher weights (mean embedding values, see supplementary Python notebook) compared to other earlier steps in the pipeline. Thus, pipelines with different network reconstruction steps would be less similar even if they share many other earlier steps. Second, comparing GCN-embeddings and Levenshtein, discrepancies were also found in pipeline pairs that differed in when cleaning steps (e.g., temporal filtering, motion regression) were employed. One pipeline might perform cleaning after brain parcellation, while others might do so before. Levenshtein considered these less similar due to the difference in order, and because it treated all steps equally. However, GCN-embeddings assigned them higher similarity because the weights it computed for brain parcellation and cleaning steps (e.g., temporal filtering, motion regression) were similar (mean embedding values, see supplementary Python notebook). Thus, these pipelines were considered more similar, despite their different sequence of steps, based on their embeddings. Importantly, the embeddings of a step also capture information about its neighbours, suggesting that similarity in embeddings implies that these steps may have an overlap in common neighbours.

Empirical Comparison

Comparing the meta-analytic similarity measures (using features characterizing their use in the literature) with empirical measures in a small multiverse of 16 pipelines revealed that GCN-embeddings performed, in terms of absolute error, comparably to traditional measures such as Levenshtein and Hamming distances. However, none of the methods — including GCN-embeddings — were able to adequately capture the substantial influence of global signal regression (GSR) on empirical similarity, underscoring a key limitation in current approaches: the inability to fully account for individual analysis steps with disproportionate effects on the outcome. In contrast, GCN-embeddings (as well as Jaccard and Levenshtein distances) was more sensitive to differences in pipeline lengths caused by variations in the number of individual steps within specific categories (in this case cleaning). These differences are amplified by the current coding schemes in which certain pipeline categories, like cleaning, may include a varying number of steps.

Implications and Future Directions

The findings of the present study suggest that GCN-based meta-analytic similarity measures may serve as a simple foundational tool for incorporating prior knowledge from an extensive literature into multiverse analysis frameworks. While the proposed method does not yet fully capture the effects of influential individual analysis steps, it already generates valuable information with relatively low computational effort. Future work is required to validate GCN-embeddings (or other GCN-based approaches) with larger and more comprehensive multiverse analyses and examine its consistency with empirically derived similarity measures. Establishing robust empirical ground truths will enable the refinement of the GCN, such as exploring deeper architectures to better capture global features across pipelines. Incorporating contextual information on the level of individual analysis steps, informed by expert knowledge about disproportionately influential steps, could also enhance the ability of the algorithm to distinguish meaningful differences between pipelines. Finally, automating the extraction of pipelines from literature would expand the meta-analytic dataset used here significantly, facilitating more robust training and testing of the model.

Broader Impact

While our study was primarily focused on quantifying pipeline similarity within the context of fMRI, its broader implications extend to addressing potential "fallacies and pitfalls" in other life science research domains (Hecker et al. 2023) that rely on complex data preprocessing, such as Positron Emission Tomography (PET) (Naseri et al. 2024), Electroencephalography (EEG) (Jacobsen et al. 2024), or genome-wide association studies (GWAS) (Hecker et al. 2023). This approach, by enabling a deeper understanding of how different processing and analysis choices can subtly affect results, promotes greater transparency

 © 2024 BHA

and reproducibility in such computational disciplines. Such enhanced understanding can help mitigate the risk of drawing misleading conclusions due to pipeline variability, a common pitfall in data-intensive research. Our work contributes to the broader goal of improving the validity and integrity of scientific findings in these research areas that rely on complex, multidimensional data.

Study Limitations

First, the meta-analytic dataset used in this analysis was limited to 220 pipelines, constraining the scope of the analysis. Automating the extraction of pipelines from the literature would address this limitation. Second, parameter-level differences in steps (e.g., specific software package, type of brain parcellation, number of motion regressors, or filtering options) were not considered, despite their known influence on results (Parkes et al. 2018; Luppi et al. 2024). Including these factors in future analyses is essential, though this was infeasible in the current study as this would lead to a sparse network, which would hinder the training of the GCN. Third, the empirical comparison was limited to a small multiverse of 16 pipelines with significant overlap in steps due to the HCP preprocessing pipeline. Expanding the analysis to a more diverse and larger set of pipelines would provide deeper insights into the behavior of the proposed measures.

Another avenue for improvement involves ensuring a uniform number of steps across pipelines, as this could enable fairer comparisons between methods and potentially reduce absolute error. Standardizing step representations, such as collapsing all noise reduction strategies under a single step with specific options (e.g., "Noise Reduction: GSR, None, or Both"), could improve performance. Such an approach would also require data sets from the literature to follow a consistent coding scheme, ensuring that all steps are comparable across pipelines.

Conclusion

The present study highlights the importance of quantifying pipeline similarities as a step toward improving the efficiency of multiverse analysis and developing tools for enhanced reproducibility in computationally intensive research workflows. By integrating step embeddings and sequential characteristics, GCN-based methods provide a simple framework to inform future algorithm development. While the current GCN-based similarity measure does not yet fully address variability (e.g. due to particularly influential analysis steps), its ability to generate valuable prior knowledge with low computational effort makes it a promising foundation for future advancements in this area.

Acknowledgments

This work was supported by a grant from the German Research Foundation (DFG) awarded to Andrea Hildebrandt (HI 1780/7-1) and Carsten Gießing (GI 682/5-1) as part of the DFG priority program "META-REP: A Meta-scientific Programme to Analyse and Optimise Replicability in the Behavioural, Social, and Cognitive Sciences" (SPP 2317).

Experimental data were provided by the Human Connectome Project, MGH-USC Consortium (Principal Investigators: Bruce R. Rosen, Arthur W. Toga and Van Wedeen; U01MH093765) funded by the NIH Blueprint Initiative for Neuroscience Research grant; the National Institutes of Health grant P41EB015896; and the Instrumentation Grants S10RR023043, 1S10RR023401, 1S10RR019307.

Conflicts of Interest

The authors have no personal, professional, or financial conflicts of interest to declare.

Availability of Code and Data

Code and data to run the analyses included in this study are available at github.com/metascience-uol/GCN-pipelines. The meta-analytic data on pipelines was openly published by Kristanto et al. (2024). The empirical data used to validate the meta-analytic pipeline similarity measures are part of the Human Connectome Project which can be requested for scientific use at humanconnectome.org/study/hcp-young-adult.

Citation

Brainiacs 2024 Volume 5 Issue 2 Edoc XEE8F298E
Title: "Quantifying Similarity between Graph-Theoretic Resting-State fMRI Data Processing Pipelines for Efficient Multiverse Analysis"
Authors: Micha Burkhardt, Andrea Hildebrandt, Carsten Gießing, Daniel Kristanto
Dates: created 2024-09-09, presented 2024-10-09, updated 2024-12-22, published 2024-12-23, endorsed 2025-03-26
Copyright: © 2024 Brain Health Alliance
Contact: micha.burkhardt@uol.de
URL: BrainiacsJournal.org/arc/pub/Burkhardt2024QSMPPEMA
PDP: /Nexus/Brainiacs/Burkhardt2024QSMPPEMA
DOI: /10.48085/XEE8F298E

References

[1] R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, et al. "Variability in the analysis of a single neuroimaging dataset by many teams." *Nature* 582 (2020), pp. 84–88. DOI: https://doi.org/10.1038/s41586-020-2314-9 (cited p. 2).

[2] A. Bowring, T. E. Nichols, and C. Maumet. "Isolating the sources of pipeline-variability in group-level task-fMRI results." *Human brain mapping* 43.3 (2022), pp. 1112–1128 (cited p. 2).

[3] M. Burkhardt and C. Giessing. "A dynamic functional connectivity toolbox for multiverse analysis." *bioRxiv* (2024), pp. 2024–01 (cited p. 6).

[4] G. G. Cantone and V. Tomaselli. "Theory and methods of the multiverse: an application for panel-based models." *Quality & Quantity* 58 (2024), pp. 1447–1480. DOI: https://doi.org/10.1007/s11135-023-01698-5 (cited p. 3).

[5] J. Dafflon, P. F. Da Costa, F. Váša, R. P. Monti, et al. "A guided multiverse study of neuroimaging analyses." *Nature Communications* 13.1 (2022), p. 3758. ISSN: 2041-1723. DOI: https://doi.org/10.1038/s41467-022-31347-8 (cited p. 2).

[6] M. Del Giudice and S. W. Gangestad. "A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions." *Advances in Methods and Practices in Psychological Science* 4.1 (2021). DOI: https://doi.org/10.1177/2515245920954925 (cited p. 2).

[7] D. Frias-Navarro, J. Pascual-Llobell, M. Pascual-Soler, J. Perezgonzalez, and J. Berrios-Riquelme. "Replication crisis or an opportunity to improve scientific production?" *European Journal of Education* 55.4 (2020), pp. 618–631. DOI: https://doi.org/10.1111/ejed.12417 (cited p. 2).

[8] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, et al. "The minimal preprocessing pipelines for the Human Connectome Project." *Neuroimage* 80 (2013), pp. 105–124 (cited pp. 3, 6).

[9] R. W. Hamming. "Error detecting and error correcting codes." *The Bell System Technical Journal* 29.2 (1950), pp. 147–160 (cited pp. 3, 4).

[10] J. Hecker, A. Craig, A. Hughes, J. Neidich, C. Taswell, and N. Laird. "Fallacies and Pitfalls in Genome-Wide Association Studies." *2023 Guardians Workshop (Guardians)* (2023) (cited p. 9).

[11] D. E. Huber, K. W. Potter, and L. D. Huszar. "Less "story" and more "reliability" in cognitive neuroscience." *Cortex* 113 (2019), pp. 347–349. DOI: https://doi.org/10.1016/j.cortex.2018.10.030 (cited p. 2).

[12] L. Hubert and P. Arabie. "Comparing partitions." *Journal of classification* 2.1 (1985), pp. 193–218 (cited p. 6).

[13] P. Jaccard. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura." *Bulletin de la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 547–579 (cited pp. 3, 4).

[14] N. S. Jacobsen, D. Kristanto, S. Welp, Y. C. Inceler, and S. Debener. "Preprocessing Choices for P3 Analyses with Mobile EEG: A Systematic Literature Review and Interactive Exploration." *bioRxiv* (2024), pp. 2024–04 (cited p. 9).

[15] D. Kristanto, M. Burkhardt, C. Thiel, S. Debener, C. Gießing, and A. Hildebrandt. "The multiverse of data preprocessing and analysis in graph-based fMRI: A systematic literature review of analytical choices fed into a decision support tool for informed analysis." *Neuroscience & Biobehavioral Reviews* 165 (2024), p. 105846. DOI: https://doi.org/10.1016/j.neubiorev.2024.105846 (cited pp. 1–3, 10).

[16] V. I. Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics Doklady* 10.8 (1966), pp. 707–710 (cited pp. 3, 4).

[17] A. I. Luppi, H. M. Gellersen, Z.-Q. Liu, A. R. Peattie, et al. "Systematic evaluation of fMRI data-processing pipelines for consistent functional connectomics." *Nature Communications* 15.1 (2024), p. 4745 (cited pp. 2, 10).

[18] M. Naseri, S. Ramakrishnapillai, and O. T. Carmichael. "Reproducible brain PET data analysis: easier said than done." *Frontiers in Neuroinformatics* 18 (2024), p. 1420315 (cited p. 9).

[19] Open-Science-Collaboration. "Estimating the reproducibility of psychological science." *Science* 349 (2015), p. 943 (cited p. 2).

[20] L. Parkes, B. Fulcher, M. Yücel, and A. Fornito. "An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI." *Neuroimage* 171 (2018), pp. 415–436 (cited p. 10).

[21] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition." *IEEE transactions on acoustics, speech, and signal processing* 26.1 (1978), pp. 43–49 (cited p. 5).

[22] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo. "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI." *Cerebral cortex* 28.9 (2018), pp. 3095–3114 (cited p. 3).

[23] J. P. Simmons, L. D. Nelson, and U. Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22.11 (2011), pp. 1359–1366. DOI: https://doi.org/10.1177/0956797611417632 (cited p. 2).

[24] S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. "Increasing Transparency Through a Multiverse Analysis." *Perspectives on Psychological Science* 11.5 (2016), pp. 702–712. DOI: https://doi.org/10.1177/1745691616658637 (cited p. 2).

[25] S. Strother. "Evaluating fMRI preprocessing pipelines." *IEEE Engineering in Medicine and Biology Magazine* 25.2 (2006), pp. 27–41 (cited p. 2).

[26] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. "The WU-Minn human connectome project: an overview." *Neuroimage* 80 (2013), pp. 62–79 (cited p. 3).

[27] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm. "Deep graph infomax." *arXiv preprint arXiv:1809.10341* (2018) (cited pp. 3, 4).

 © 2024 BHA