2023-08

Advancing the Study of Functional Connectome Development

Graff, Kirk

UNIVERSITY OF CALGARY

Advancing the Study of Functional Connectome Development

by

Kirk Jason Graff

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN NEUROSCIENCE

CALGARY, ALBERTA

AUGUST, 2023

Abstract

      A better understanding of functional changes in the brain across childhood offers the potential to better support neurodevelopmental and learning challenges. However, neuroimaging tools such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) are vulnerable to head motion and other artifacts, and studies have had limited reproducibility. To accomplish research goals, we need to understand the reliability and validity of data collection, processing, and analysis strategies. Neuroimaging datasets contain individually unique information, but identifiability is reduced by noise or lack of signal, suggesting it can be a measure of validity. The goal of this thesis was to use identifiability to benchmark different methodologies, and describe how identifiability associates with age across early childhood.

      I first compared several different fMRI preprocessing pipelines for data collected from young children. Preprocessing techniques are often controversial due to specific drawbacks and have typically been assessed with adult datasets, which have much less head motion. I found benefits to the use of global signal regression and temporal censoring, but overly strict censoring can impact identifiability, suggesting noise removed must be balanced against signal retained.

      I also compared several different EEG measures of functional connectivity (FC). EEG can be vulnerable to volume conduction artifacts that can be mitigated by only considering shared information with a time delay between signals. However, I found that mitigation strategies result in lower identifiability, suggesting that while removing confounding noise they also discard substantial signal of interest.

      Individual experiences may shape development in an individually unique way, which is supported by evidence that adults have more individually identifiable patterns of FC than

children. I found that across 4 to 8 years of age, identifiability increased via increased self-stability, but without changes in similarity-to-others.

In the absence of ground truth, it is difficult to argue for or against analysis decisions based solely on a theoretical framework and need to also be validated. My work highlights the importance of not thinking about techniques in a valid-invalid dichotomy; certain methods may be sub-optimal while still being preferable to alternatives if they better manage the trade off between noise removed and signal retained.

Chapters 2-4 of this thesis consist of primary research articles that are either published (Chapters 2 and 3) or are in preparation (Chapter 4). References to these articles are provided below:

Chapter 2: Graff, K., Tansey, R., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S., 2022a. Benchmarking common preprocessing strategies in early childhood functional connectivity and intersubject correlation fMRI. Developmental Cognitive Neuroscience 54, 101087. https://doi.org/10.1016/j.dcn.2022.101087

Chapter 3: Graff, K., Tansey, R., Rai, S., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S., 2022b. Functional connectomes become more longitudinally self-stable, but not more distinct from others, across early childhood. NeuroImage 258, 119367. https://doi.org/10.1016/j.neuroimage.2022.119367

Chapter 4: Graff, K., Rai, S., Yin, S., Godfrey, K., Merrikh, D., Tansey, R., Vanderwal, T., Protzner, A., Bray, S. 2023. Reliability and validity of the electrophysiological connectome across phase-based connectivity measures. Manuscript in preparation.

All content from these works is included here with permission from co-authors (Appendix E).

Acknowledgements

To properly acknowledge everyone who has had a role in shaping this thesis would be a thesis in and of itself. I am immensely grateful for all the support I've received and I am in the debt of many. Most importantly, this thesis would not have been possible without my supervisor, Dr. Signe Bray. Over the past 5 years of graduate school, often I found myself lost, both as a researcher and as a human, with the pandemic years being especially disorienting. Signe's endless knowledge, support, and flexibility has allowed me to make it to the finish line. Or maybe this is just the starting line. I don't know. Signe, help?

Graduate school would have been unbearably lonely without my lab mates and everyone in the broader CAIR community. A special shoutout to Ryann Tansey, my graduate school twin, second author, and commiseration partner. I hope you will soon be able to finally go home, relax, and have a nice, long nap. Thank you for not getting mad when my brilliant code was found to have rather substantial bugs, meaning all your data needed to be reprocessed.

Thank you to my committee and everyone else who has appraised my knowledge or work, including my thesis examination panel. I hope you enjoy reading my work here.

Thank you to Kat Davenport, Jeptha Davenport, and Tamara Pringsheim for getting me interested in neuroscience.

A big thank you to my now wife, Taylor Graff. I have always appreciated your willingness to listen to my daily stories of graduate school adventures, even when I was completely unable to explain k-space, or censoring thresholds, or volume conduction. Thank you for your endless help with figures; I think every figure in this document was touched up by you in at least a small way.

This thesis would not have occurred if not for the work of John Pemberton, the inventor of Coca-Cola, and James Schlatter, who discovered aspartame. Together, their brilliance combined to create Coke Zero, a much healthier addiction than most addictions. Sorry, teeth.

Table of Contents

List of Tables

Chapter 2

Chapter 4

List of Figures

List of Abbreviations

ANATICOR	Anatomy-Based Image Correction (software)
ANTs	Advanced Normalization Tools (software library)
AFNI	Analysis of Functional Neuroimages (software library)
BG	basal ganglia
BOLD	blood oxygen-level dependent
BRAVO	Axial MRI 3D brain volume (MRI sequence)
coh	coherence
CompCor	component-based noise correction (software)
CSF	cerebrospinal fluid
DMN	default mode network
EEG	electroencephalography
eLORETA	exact Low Resolution Electromagnetic Tomography (algorithm)
EPI	echo-planar imaging
FA	flip angle
FC	functional connectivity
FD	framewise displacement
fMRI	functional magnetic resonance imaging
FSL	FMRIB software library
FSL BET	Brain Extraction Tool (software)
FSL FLIRT	FMRIB's Linear Image Registration Tool (software)
FSL MCFLIRT	Motion Correction FLIRT (software)
GSR	global signal regression

HbO	oxyhemoglobin
HbR	deoxyhemoglobin
HMP	head motion parameter
ICA	independent component analysis
ICA-AROMA	ICA-based Automatic Removal of Motion Artifacts (software)
ICA-FIX	FMRIB's ICA-based Xnoiseifier (software)
ICC	intraclass correlation
imcoh	imaginary coherence
ISC	intersubject correlation
MEG	magnetoencephalography
MIST	Multiresolution Intrinsic Segmentation Template
MNI	Montreal Neurological Institute
pli	phase lag index
plv	phase-locking value
psi	phase slope index
MRI	magnetic resonance imaging
QC-FC	quality-control functional connectivity
T1w	longitudinal relaxing time-weighted
TE	echo time
TR	repetition time
WM	white matter
wpli	weighted phase lag index

Chapter 1: Introduction

Neuroscience has long sought to understand human neurodevelopment. While interesting in and of itself, a more thorough understanding of functional changes during development has potential societal benefits, such as offering the possibility of developing interventions to help struggling students, including those with learning disabilities or neurodiverse populations.

Functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) are two of the primary methodologies to investigate human brain function and its development. Investigations into synchrony between brain regions – known as functional connectivity (FC) - have successfully been used to measure changes over time (Grayson and Fair, 2017), understand differences between neurotypical and neurodiverse individuals (González-Madruga et al., 2022), and better understand behavior (Rosenberg et al., 2020). However, data collected from fMRI and EEG have a relatively low signal-to-noise ratio, and are vulnerable to artifacts such as head motion, leading to concerns about reliability and validity. Sufficient reliability and validity are a major challenge to inference in developmental connectomics, and arguably limited reliability of measurements has led to poor reproducibility in the field (Marek et al., 2022).

Reliability is the extent to which multiple recordings of the same features, such as the same individual or the same group under similar conditions, produce consistent measurements, such as stable estimates of FC. On the other hand, validity is the extent to which measurements accurately capture the feature of interest. While reliability is a precondition for validity, more reliable measurements are not always more valid, as data can be systematically biased. For example, a poorly calibrated scale could consistently give the same inaccurate reading of an object's weight. This can occur in neuroimaging: Earlier fMRI research reliably showed that young children have weaker long-range connectivity, but the validity of these findings was later

reassessed, as head motion that is not properly accounted for causes longer connections to look weaker, and children tend to move more in the scanner (Grayson and Fair, 2017; Power et al., 2012).

It can be difficult to assess validity in human neuroscience, as non-invasive measures such as fMRI and EEG have limitations, and there is typically no ground truth available. Nonetheless, several properties associate with validity. One is the ability to capture individually specific information, allowing for participant identifiability. While previous work often describes individual distinctiveness as "reliability" (Noble et al., 2019), here I use "reliability" in a more limited sense of any repeatable measurement, to emphasize that certain features can be consistently captured while deviating from underlying ground truth, while noting that identifiability is only a proxy of true validity. In this thesis, I used participant identifiability, along with other measures of validity, to assess the reliability and validity of fMRI and EEG FC in developmental samples, and I assessed changes in identifiability with age in an early childhood sample.

1.1.1 Functional magnetic resonance imaging

Magnetic resonance imaging (MRI) uses strong magnetic fields and radio waves to create images of the body's anatomy and physiology. In addition to many clinical uses, it has become well used in neuroscience research to non-invasively image the human brain in terms of structure and function. Functional MRI is sensitive to changes in blood oxygenation, which is an indirect measure of neuronal activity (Ogawa et al., 1990). This is accomplished by utilizing the properties of hemoglobin, the protein responsible for transporting oxygen in red blood cells, which exists in two forms. The first, oxyhemoglobin (HbO), is the form actively carrying

oxygen. By nature of how oxygen binds to hemoglobin, HbO has no unpaired electrons and is therefore diamagnetic, i.e., repelled from external magnetic fields. The second form, deoxyhemoglobin (HbR), is the form not carrying oxygen. It has unpaired electrons, causing it to be paramagnetic, i.e., attracted to external magnetic fields, and thus creates local magnetic field distortions.

The relationship between neural activity and blood flow is known as neurovascular coupling (Hillman, 2014). Through neurovascular coupling, when neurons fire, blood flow increases to the region of neuronal activity to increase oxygen delivery. HbO exceeds what is necessary to account for an increase in oxygen demand, resulting in a drop in the concentration in HbR. fMRI utilizes the magnetic field changes due to HbR to map changes in the blood oxygen-level dependent (BOLD) response (Hillman, 2014). That is, a drop in the HbR concentration leads to magnetic field changes that are recorded as an increased fMRI BOLD signal. Therefore, in BOLD fMRI we interpret higher BOLD signals as more neural activity, although the relationship is complex and not fully understood (Glover, 2011; Kim, 2018; Logothetis, 2002)

In a typical BOLD fMRI acquisition, the whole brain is imaged in a few seconds (or fractions of a second with multi-slice accelerated sequences), generating hundreds of volumes of data over a span of minutes. For example, if a brain volume is recorded in 2 seconds, then in a 1-minute scan 30 volumes of data are collected. For any given voxel (a 3-dimensional pixel) there will then be a 30-point time series. While a lot can be learned about brain activity at this temporal resolution, fMRI is poorly suited for investigating higher frequency brain activity (>0.5 Hz). One of the biggest advantages of fMRI compared to other non-invasive methodologies used to investigate human brain function (such as EEG) is excellent spatial resolution, with whole

brain imaging at the millimetre scale. Despite this comparative advantage, any given fMRI voxel is coalescing data from tens of thousands of neurons and other cells, limiting interpretability at the cellular level. Further, fMRI is susceptible to distortion and signal drop out in specific regions (Rua et al., 2018). Despite these limitations, fMRI is currently the most popular method used to investigate human FC.

1.1.2 fMRI functional connectivity

In addition to the investigation of task-evoked neural responses – such as how flashing lights lead to BOLD signal changes in the occipital cortex (Kwong et al., 1992) – fMRI can also be used to investigate the interactions between different regions of the brain. One way to assess this is with FC, which quantifies the statistical relationship in activity between brain regions (Biswal et al., 1995). In fMRI, FC is typically determined by calculating the Pearson correlation between the temporal signals from two voxels or brain regions, with higher correlations in activity interpreted as a stronger functional connection. However, note that a strong statistical correlation does not necessarily reflect or imply direct physical connections.

FC-fMRI studies have been used to generate maps of 'connections' within the brain, known as functional connectomes. Connectomes are typically created by parcellating the brain into a set of regions, often using a predefined atlas, such as the Schaefer atlas (Schaefer et al., 2018). Afterwards, the average time series is calculated within each region, and then FC is calculated between all pairs of brain regions. These functional connectomes can then be compared between people (Jung et al., 2020), longitudinally within people (Calabro et al., 2020), or when switching between tasks (Harrewijn et al., 2020), revealing both inter-individual differences and intra-individual changes. While connectomes can be generated using data

collected with traditional task-based paradigms, such as memory or motor tasks, they are often used to investigate 'intrinsic' fluctuations in brain activity. To do this, a participant is given either instructions to stare at a fixation cross (a resting state scan), or a prolonged attention task, such as watching a movie (often called passive viewing). These scans can last up to hours (Gordon et al., 2017), allowing detailed investigations into an individual's connectivity patterns.

Connectome studies have shown that the brain is organized into a set of distinct functional networks, where each functional network shows greater intra-network connectivity than inter-network connectivity (Bressler and Menon, 2010). The exact number of networks and their location can vary by dataset, study, and methodology, but network descriptions generally include 7 to 11 networks (Shen et al., 2013; Urchs et al., 2019). These include networks covering sensory and motor regions, such as the somatomotor and visual networks, networks involved in task-positive cognition, such as the dorsal attention and the frontoparietal networks, and the default mode network (DMN). The DMN in particular has been well-studied, as it tends to be negatively correlated with attention networks and is related to self-referential mental activity and recollecting prior experiences (Raichle, 2015). Some brain networks are primarily localized to a single bilateral location, while others consist of spatially distinct regions, such as the DMN which consists of the ventral medial prefrontal cortex, the dorsal medial prefrontal cortex, and the posterior cingulate cortex (Raichle, 2015).

1.1.3 A novel measure of validity: fingerprinting and individualization

Finn et al. (2015) showed that FC-fMRI is both stable and robust enough to successfully match two connectomes from the same person from among a set of connectomes from other people. Specifically, given a dataset where participants are scanned at least twice, by calculating

a functional connectome for each scan and then calculating the correlation between connectomes, >90% of the time a scan's highest correlation is a second scan from the same participant, rather than with someone else. Finn et al. (2015) tested this method using both the entire connectome and only specific networks. By testing 8 networks individually, they found that the medial frontal and frontoparietal networks were best for a successful match, while visual networks tended to be the worst, with the DMN being in the middle. Using the medial frontal and frontoparietal networks together led to higher matching rates than using the whole brain. This suggests that the medial frontal and frontoparietal networks are more individualized from person to person, while visual networks are more similar.

A number of related studies have further validated the fingerprinting approach and expanded our understanding of the individualization of functional connectomes. Using youth participants (7.5-14.6 years old), Miranda-Dominguez et al. (2018) found average correlations between longitudinal scans from the same person of R = 0.58, compared with correlations of R = 0.44 between siblings and R = 0.42 between unrelated individuals, with the last two values being significantly different from each other. This suggests that an individual's functional connectome is relatively stable across up to two years of development, and that siblings have more similar functional connectomes than unrelated individuals. Similar findings were also seen in adults, with higher correlations in the frontoparietal, DMN, and attention networks. Peña-Gómez et al. (2018) found that a small subset of connections, primarily in the frontoparietal network, are sufficient for fingerprinting success, with Byrge and Kennedy (2019) found that as few as 40 connections, when selected on the basis of their distinctiveness, can be used to successfully fingerprint an individual, while also finding that seemingly all edges contain individually specific information.

More valid data should contain more individually specific information, allowing for better identifiability. This has been demonstrated in a variety of studies, which have found that methodological decisions affect fingerprinting success. Finn et al. (2015) found that matching accuracy was affected by choice of task, with match rates ranging from 64% to 99%. The highest match rates were found when matching resting state connectomes to resting state connectomes, compared with rest-to-task, task-to-rest, or task-to-task, where tasks included working memory, motor, language, and emotion tasks. Further, Vanderwal et al. (2017) found that passive viewing led to even higher fingerprinting success than resting state, which was found both when watching a movie or a tranquil video. Some studies have considered the relationship between scan length and fingerprinting success, with longer scans having higher accuracy (Finn et al., 2015; Horien et al., 2018). Other factors have been shown to affect fingerprinting success, such as choice of parcellation (Finn et al, 2015) and quantity of head motion (Horien et al., 2018). Thus, while individually specific information in a connectome has primarily been investigated as a feature of interest, it can also be used as a benchmark of data validity, as low quality data will have reduced identifiability due to connectomes looking more similar across participants (Horien et al., 2018; Waller et al., 2017).

1.2.1 How does fMRI preprocessing impact reliability and validity?

Due to motion-related, physiological, and other noise in fMRI, it has become standard to carry out a set of preprocessing steps as part of fMRI and FC-fMRI studies. Many of these steps have seen almost universal adoption throughout the literature, however their implementation often varies from study to study. Steps typically include rigid body realignment, slice time correction, masking out non-brain regions, registration of images to a template, spatial

smoothing, detrending and other temporal filtering, and white matter (WM) and cerebrospinal fluid (CSF) signal regression (Ciric et al., 2018). However, one of the biggest challenges in fMRI data analysis is the impact of motion-related artifacts. Initially, it was assumed that regressing out estimates of participant motion from voxel time courses was largely sufficient for removing motion-related noise (Friston et al., 1996), but several studies reported that even with regression, head motion leads to immediate and substantial changes in fMRI signal that are systematic (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). These systematic changes are problematic for FC-fMRI studies as they result in long-distance correlations decreasing and short-distance correlations increasing. This appreciation of methodological challenges has led to a number of previous findings being reconsidered, especially in the clinical and developmental literature (Grayson and Fair, 2017), while also leading to a greater emphasis on noise mitigation. Denoising strategies typically come with various strengths and drawbacks, leading to debate on the most appropriate methods; two such debated strategies are volume censoring and global signal regression (GSR).

To remove the effect of specific motion contaminated volumes, it has become common to either remove or interpolate over the temporal signal at specific time points, in a process known as censoring or scrubbing. This is due to sharp scan-to-scan head movements having a larger effect on data than total movement across the entire scan duration (Power et al., 2012). Censoring approaches lead to a loss of data and disrupt temporal autocorrelations, which have the potential to affect FC estimates, though the exact effect has not been well studied. While censoring removes the volumes most damaged by noise, it fails to remove prolonged signal changes due to spin history, and it requires the adoption of a motion threshold, above which volumes are removed and below which any signal changes persist. Exact censoring thresholds

have been debated, along with whether to censor out only contaminated volumes or neighbouring volumes as well (Satterthwaite et al., 2019).

GSR has become a widely used and widely debated preprocessing step (Murphy and Fox, 2017). By regressing out the mean time series for the whole brain, head motion and other artifacts can be removed from the data, based on the assumption that any process captured brain-wide cannot be due to regional neuronal activity. GSR is a simple and effective denoising technique, improving the specificity of positive correlations and showing results that are more consistent with anatomy (Murphy and Fox, 2017). GSR has the drawback of changing the distribution of FC so that it is nearly zero centered, decreasing positive BOLD responses and creating or amplifying anti-correlations (Aguirre et al., 1998). The global signal has also been shown to resemble established networks and to be significantly related to life outcomes and psychological function, suggesting it contains information on cognition and behavior (Li et al., 2019), and that by extension GSR should be used with caution as it removes information of interest in addition to noise.

A large collection of additional preprocessing steps has also been developed, and there has been little consistency in which steps are applied and the order that they are applied in. A study by Carp (2012) showed that of 241 fMRI articles, nearly every one had a slightly different preprocessing pipeline. While there have been attempts in the community to standardize preprocessing, such as through the use of fMRIprep (Esteban et al., 2019), there remains a need to understand the appropriateness of different strategies, especially in high-motion datasets such as those collected from children.

1.2.2 Benchmarking pipelines: how do preprocessing choices impact reliability and validity?

In response to the proliferation of preprocessing approaches, recent studies have compared the effectiveness of different pipelines (Churchill et al., 2017; Ciric et al., 2017; Parkes et al., 2018). Typically, these papers have looked at the amount of noise remaining in a dataset following preprocessing by using a metric known as quality control – functional connectivity (QC-FC). For a set of participants, QC-FC is the correlation between the strength of a given network edge, across participants, and the mean amount of participant motion (as measured by framewise displacement between volumes). Based on the assumption that there should be no correlation between motion and FC, if such a correlation exists then one can argue that there is motion-related noise in the data. QC-FC has two main uses – one is to look at the quantity of network edges related to motion, while the other is to assess whether an edge's QC-FC is related to the inter-node distance – i.e., are short-range connections more vulnerable to head-motion than long-range connections? (Power et al., 2015). While QC-FC is a useful benchmark, it only estimates noise due to motion and not other possible sources of noise, such as physiological effects.

Both Ciric et al. (2017) and Parkes et al. (2018) found that pipelines with GSR have fewer edges significantly related to subject motion, but show distance-dependent effects, though it is unclear whether GSR increases distance-dependence or fails to successfully remove components of noise. Additionally, pipelines containing censoring were particularly effective at removing motion. However, Parkes et al. (2018) also found that preprocessing pipelines with the highest amount of remaining noise had the highest test-retest reliability, suggesting a component of noise is highly reproducible across repeated scans from the same individual. Ciric et al. (2017)

noted that "…certain de-noising methods could conceivably both minimize QC-FC relationships and even enhance reliability by aggressively removing both signal and noise, but in the process diminish sensitivity to meaningful individual differences." This suggests that reliability and measures of validity related to noise contamination are not sufficient to assess the overall validity of preprocessing decisions, and that a wider array of benchmarks are necessary. One useful approach for further pipeline comparison could be to specifically assess how information about individual differences changes as a function of preprocessing steps. Preprocessing pipelines that better retain signal of interest, even if noisier, may allow for better fingerprinting, suggesting higher validity (Phạm et al., 2023).

Preprocessing pipelines have been compared using data from adults, but they have not been systematically compared in young children. There are a number of reasons why studies with children are different than studies of adults, necessitating further validation of preprocessing strategies. Young children are a population with much higher motion than adults (Dosenbach et al., 2017), which systemically alters FC. To reduce movement, FC-fMRI studies in children are increasingly conducted using passive viewing tasks such as movies (Greene et al., 2018; Rohr et al., 2017), and it is unknown how this stimulus might interact with preprocessing steps. Additionally, children have physiological differences from adults. These include faster respiratory rates and heart rates (Fleming et al., 2011), which may affect physiological noise in ways unaccounted for with QC-FC benchmarks. Similarly, smaller head sizes and other fundamental differences in brain structure may impact distance-dependent preprocessing strategies such as GSR.

1.3.1 Connectome changes with age

Functional network development begins prior to birth (Dehaene-Lambertz and Spelke, 2015), with development continuing through childhood, adolescence, and into adulthood (Sydnor et al., 2021). Across the brain, connection strength increases across the first years of development in both connections that are initially strong and initially weak (Lin et al., 2008), and by two years of age the brain strongly resembles adult functional organization, with immature versions of all networks (Grayson and Fair, 2017). While network neurodevelopment is multi-faceted, development roughly progresses from 'lower order' sensory and motor networks to 'higher order' association networks responsible for executive, socioemotional, and mentalizing functions (Sydnor et al., 2021). Newborns already show adult-like topographies in motor, visual, and auditory networks (Gao et al., 2015b; Gilmore et al., 2018). These networks have been shown to display little internetwork change from early childhood to adulthood, staying well-segregated (Gu et al., 2015), though within these networks age correlates with the strength of connections (Rohr et al., 2018). Higher order networks tend to develop adult-like architecture more gradually – i.e., between birth and two years of age – and then continue to slowly change over the next decades (Emerson et al., 2016; Gilmore et al., 2018; Sydnor et al., 2021). For example, the DMN tends to only show adult-like spatial synchrony by six months of age (Gao et al., 2015a), though within-network connections remain weak even at nine years old, with connections strengthening into adulthood (Fair et al., 2008). In adolescents, networks involved in higher level cognition, such as salience, attention, and executive skills, become increasingly segregated from other networks and with a more variable spatial structure (Bassett et al., 2018; Dosenbach et al., 2010; Gu et al., 2015).

It was originally shown that over childhood the brain shifts from a more local to a more global functional organization (Fair et al., 2009), but the validity of these findings have been questioned as the systematic effects of head motion on FC have been investigated in more detail (Grayson and Fair, 2017; Power et al., 2012; Satterthwaite et al., 2012). More recent findings suggest that brain development trajectories from childhood to adulthood involves more specific functional changes, such as more integration between the cingulo-opercular/salience and somatomotor networks (Marek et al., 2015), and more synchronicity with DMN connections (Gu et al., 2015). However, much remains unknown, especially in early childhood due to difficulties with obtaining usable data (Vanderwal et al., 2019). While some broad patterns of typical development trajectories seem clear, there remains a lack of large-scale longitudinal studies and a lack of replication across data sets. As thus, more precise trajectories and how they relate to underlying structural changes remains to be studied. Further, while there is substantial variation between people, individual-specific developmental changes have not been well-investigated.

1.3.2 Age and connectome identifiability

While high identifiability – as demonstrated with connectome fingerprinting – has been shown in adults, the extent to which individualization is a developmental process is not well understood. Broadly speaking, studies seem to suggest that longitudinal identifiability increases with age. Under 25% matching rates were shown in a study matching 1.5-month-old children to themselves at 9 months old (Dufford et al., 2021), up to 38% matching accuracies were shown in older children, between 6 and 15 (Sato et al., 2021), and no difference in longitudinal match rates were found between adolescents and adults (Jalbrzikowski et al., 2020). However, no study has

directly investigated the effect of age on longitudinal identifiability across early childhood, raising questions if these findings can be better explained by differences in data quality.

Further, it remains unclear what might be driving changes in age and identifiability, as identifiability is related to both self-stability and similarity-to-others. Some evidence suggests that adults have higher longitudinal self-stability than adolescents (Liao et al., 2021), though the relationship between age and short-term across-task self-stability has shown inconsistent findings (Kaufmann et al., 2017; Vanderwal et al., 2021). Vanderwal et al. (2021) also found no relationship between age and similarity-to-others, while some evidence suggests that participants with higher self-stability have greater, rather than reduced, similarity-to-others (Vanderwal et al., 2021; Liao et al, 2021).

A related question is whether age-related changes in individualization vary across networks. Since sensorimotor networks tend to show reduced identifiability in adults compared to higher order networks such as the DMN (Finn et al., 2015; Byrge and Kennedy, 2019; Vanderwal et al., 2017), they may also show reduced identifiability in children. On the other hand, given the protracted development of higher order networks and their continued change across early childhood (Gilmore et al., 2018), higher order network identifiability may be impaired relative to the fully developed adult versions of these networks. A better understanding of these differences may illuminate the extent to which the brain develops in individually specific ways and the patterns in which this occurs across early childhood.

1.4.1 Electroencephalography

While much has been learned with fMRI and it will remain a useful tool going forward, the validity of findings can be better assessed if they can be replicated and expanded upon with

other, complementary, modalities. One area of opportunity is with EEG. While BOLD fMRI

only indirectly measures neural activity, EEG directly measures electric fields generated from

neuronal activity. Further, while fMRI makes measurements on the order of 1 Hz, EEG can be

recorded at 1000 Hz or higher. EEG also has the advantage of being far more cost efficient than

fMRI, making research more accessible.

In EEG, electrodes placed on a participant's scalp record activity from underlying brain

regions. Neuronal action potentials are the result of a rapid exchange of ions across the cell

membrane, which generates electromagnetic fields; when several nearby neurons fire in unison,

the resulting field can be detected by the electrodes (Cohen, 2014; St. Louis et al., 2016). Each

electrode measures the difference in electrical potential between itself and a reference electrode

(or electrodes) placed elsewhere on the scalp, often at the vertex of the head. While EEG can be

carried out with just a single electrode (plus reference), typically dozens of electrodes

simultaneously record brain activity. These are arranged in standardized positions, such as the

10-20 system, to capture information across the entire scalp. One convenient approach to EEG is

using a sensor net, where electrodes are imbedded in a cap that can be quickly placed on a

participant's head. These caps often have 32, 64, 128, or 256 electrodes.

While EEG has several advantages compared to fMRI, its major drawback is spatial

resolution. In general, EEG does a poor job of detecting activity from deeper brain structures and

is mostly limited to measuring the activity of the cerebral cortex. A related issue is that of

volume conduction (Nunez et al., 1997). Due to the spatial separation between where electric

fields are generated (the brain) and where they are recorded (the scalp), the signal picked up

from one electrode will reflect the activity of multiple brain regions, and conversely, the activity

at a single brain region will be picked up across multiple electrodes. Together, these limitations

make EEG a poor analysis technique for testing hypotheses involving precise functional localization. One partial workaround is that various source localization algorithms have been developed that attempt to reconstruct the underlying brain activity by combining information about the electrodes' spatial orientation and what they recorded (Pascual-Marqui, 2002). However, no perfect solution exists for this process - nor can exist - and each algorithm requires specific assumptions (Sadaghiani et al., 2022). Further, source localization cannot fully eliminate volume conduction effects (Nolte et al., 2004; Sarvas, 1987). Even after source localization the same underlying activity will show up in multiple estimated sources, an effect known as spatial leakage (Wens et al., 2015).

1.4.2 EEG FC and volume conduction

Inspired by fMRI findings, researchers have begun analyzing the functional connectome with EEG (Sadaghiani et al., 2022). However, answering developmental questions using EEG FC can be difficult due to the dozens of EEG FC measures that have been deployed, making it challenging to choose an analytic approach and to synthesize findings across studies (Cao et al., 2022). One driver in the proliferation of measures is attempts to handle the aforementioned problem of volume conduction (Nolte et al., 2004; Stam et al., 2007). If the same underlying brain activity is recorded in multiple electrodes, estimates of FC will be artificially inflated. One dependable way to assess true synchrony is therefore with a time lag. If two electrodes or brain regions have a consistent pattern of activity with a slight delay between them, it is generally assumed that there is a genuine interaction between them. On the other hand, signal synchrony without a measured time delay could either be true synchrony or the result of the same signal being measured twice. The earliest EEG FC studies did not mitigate volume conduction, utilizing

FC measures such as coherence, but concerns about result validity crept in (Nunez et al., 1997). Thus, connectivity measures were developed that mitigated volume conduction artifacts by ignoring synchrony with zero lag, likely failing to capture some true connectivity in the process (Nolte et al., 2004; Stam et al., 2007). These measures include imaginary coherence, phase lag index, and phase slope index. While the effectiveness of FC measures that mitigate volume conduction has been demonstrated (Nentwich et al., 2020; Nolte et al., 2004; Stam et al., 2007; Wirsich et al., 2021), it is currently unknown if the benefits of reduced noise outweigh the drawback of capturing less signal of interest. A better understanding of the reliability and validity of EEG FC measures will allow for more useful EEG connectome research going forward.

1.5 Overview of thesis

Reliable and valid data are necessary to meaningfully investigate the relationship between brain activity and characteristics such as age and behavior. This is especially important to consider in developmental studies, as data collected from children often contains more noise, while processing and analysis methods are typically validated using adult data. Both excessive noise and insufficient signal of interest can hamper validity. As attempts to remove noise can invertedly remove signal of interest or change the signal in undesirable ways, it is difficult to determine best practices. This makes it necessary to benchmark different methodologies, establishing their overall validity.

Findings from fingerprinting studies suggest that individuals have a unique and stable functional connectome, but that data quality can affect identifiability. Thus, fingerprinting can be used to assess validity, as lower quality data, due to excess noise or lack of signal, will have

lower fingerprinting success. This can be used along with other measures of validity to benchmark the quality of preprocessing pipelines or to compare different measures of connectivity. Using strategies to improve the reliability and validity of neuroimaging data collected from young children can allow for better investigation of developmental questions.

1.5.1 Chapter 2

In Chapter 2 I compared different preprocessing strategies for FC fMRI in young children. While preprocessing pipeline effects have been studied in adults, they have not been systematically examined in young children, and previous measures have focused more on noise remaining after preprocessing rather than signal of interest retained. By comparing fingerprinting success across several preprocessing pipelines, I assessed whether pipelines that are successful at removing noise are inadvertently removing signal-of-interest (i.e., information about individual differences). I used a dataset of 56 children, initially between 4 and 7, who were scanned twice, one year apart. Further, by comparing preprocessing pipeline effects in both low and high motion scans, I also assessed whether high-motion groups warrant more aggressive preprocessing due to increased noise contamination.

1.5.2 Chapter 3

In Chapter 3 I investigated how identifiability changes across early childhood. I considered two aspects related to identifiability: self-stability over time, and (dis)similarity compared to others. While the literature has shown that functional organization is consistent in adults, both over a matter of days and over a one-year period, networks are likely undergoing refinement in early childhood. Self-stability may thus increase across early childhood.

Developmental changes, if individually unique, may also correspond to decreased similarity to others. Using an expanded version of the dataset from Chapter 3, I looked at 73 children, initially between 4 and 7, who were scanned twice, one year apart. While I investigated this in the whole brain, I also considered changes in identifiability within individual networks. Since most sensory and motor network development occurs prior to 4 years of age, these networks may show little change in identifiability between 4 and 7. On the other hand, higher order networks such as the frontoparietal network, which are thought to continue to develop across early childhood, may show more age-related changes.

1.5.3 Chapter 4

In Chapter 4, I investigated the reliability and validity of different EEG connectivity measures. This study utilized a unique dataset where 80 minutes of EEG data per participant was collected from 25 parent-child pairs, with data collected across 4 sessions and 3 passive viewing tasks. This dense sampling allowed me to better investigate reliability across days and allowed for multiple ways to assess validity. FC measures that mitigate volume conduction ignore synchrony with no time delay. This reduces both noise and signal-of-interest, and it is unknown if this improves or impairs reliability and validity. I thus compared 6 different FC measures, where 4 of them utilized strategies to mitigate volume conduction artifacts. I assessed reliability by considering the across-session self-stability of the connectome and I assessed validity with participant identifiability, along with task- and age-sensitivity.

1.6 Statement of contributions

Chapters 2-4 of this thesis cover research articles that were co-authored with collaborators. This section outlines how each co-author contributed to these works, using the Contributor Roles Taxonomy (CRediT) format.

1.6.1 Chapter 2

Article: Graff, K., Tansey, R., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S., 2022a. Benchmarking common preprocessing strategies in early childhood functional connectivity and intersubject correlation fMRI. Developmental Cognitive Neuroscience 54, 101087. https://doi.org/10.1016/j.dcn.2022.101087

Author contributions: **Kirk Graff**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Ryann Tansey**: Investigation, Writing – review & editing. **Amanda Ip:** Resources, Investigation. **Christiane Rohr**: Investigation, Writing – review & editing. **Dennis Dimond**: Investigation, Writing – review & editing. **Deborah Dewey**: Methodology, Writing – review & editing. **Signe Bray**: Conceptualization, Investigation, Methodology, Supervision, Project administration, Funding acquisition, Writing – review & editing.

1.6.2 Chapter 3

Article: Graff, K., Tansey, R., Rai, S., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S., 2022b. Functional connectomes become more longitudinally self-stable, but not more distinct from

others, across early childhood. NeuroImage 258, 119367.

https://doi.org/10.1016/j.neuroimage.2022.119367

Author contributions: **Kirk Graff**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Ryann Tansey**: Investigation, Writing – review & editing. **Shefali Rai**: Investigation, Writing – review & editing. **Amanda Ip:** Resources, Investigation. **Christiane Rohr:** Investigation, Writing – review & editing. **Dennis Dimond**: Investigation, Writing – review & editing. **Deborah Dewey**: Methodology, Writing – review & editing. **Signe Bray**: Conceptualization, Investigation, Methodology, Supervision, Project administration, Funding acquisition, Writing – review & editing.

1.6.3 Chapter 4

Article: Graff, K., Rai, S., Yin, S., Godfrey, K., Merrikh, D., Tansey, R., Vanderwal, T., Protzner, A., Bray, S. 2023. Reliability and validity of the electrophysiological connectome across phase-based connectivity measures. Manuscript in preparation.

Author contributions: **Kirk Graff**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization, Funding acquisition. **Shefali Rai:** Investigation, Writing – review & editing. **Shelly Yin:** Investigation, Writing – review & editing. **Kate Godfrey:** Investigation, Writing – review & editing. **Daria Merrikh:** Investigation, Writing – review & editing. **Ryann Tansey:** Investigation, Writing – review & editing. **Tamara Vanderwal:** Methodology, Writing – review

& editing. **Andrea Protzner:** Methodology, Writing – review & editing. **Signe Bray:**

Conceptualization, Investigation, Methodology, Supervision, Project administration, Funding

acquisition, Writing – review & editing.

Chapter 2: Benchmarking Common Preprocessing Strategies in Early Childhood Functional Connectivity and Intersubject Correlation fMRI

2.0 Abstract

Preprocessing choices present a particular challenge for researchers working with functional magnetic resonance imaging (fMRI) data from young children. Steps which have been shown to be important for mitigating head motion, such as censoring and global signal regression (GSR), remain controversial, and benchmarking studies comparing preprocessing pipelines have been conducted using resting data from older participants who tend to move less than young children. Here, we conducted benchmarking of fMRI preprocessing steps in a population with high head-motion, children aged 4-8 years, leveraging a unique longitudinal, passive viewing fMRI dataset. We systematically investigated combinations of global signal regression (GSR), volume censoring, and ICA-AROMA. Pipelines were compared using previously established metrics of noise removal as well as metrics sensitive to recovery of individual differences (i.e., connectome fingerprinting), and stimulus-evoked responses (i.e., intersubject correlations; ISC). We found that: 1) the most efficacious pipeline for both noise removal and information recovery included censoring, GSR, bandpass filtering, and head motion parameter (HMP) regression, 2) ICA-AROMA performed similarly to HMP regression and did not obviate the need for censoring, 3) GSR had a minimal impact on connectome fingerprinting but improved ISC, and 4) the strictest censoring approaches reduced motion correlated edges but negatively impacted identifiability.

2.1 Introduction

Functional connectivity magnetic resonance imaging (FC-MRI) has become a popular tool for investigating functional brain development (Ball et al., 2014; Grayson and Fair, 2017; Marek et al., 2015) and brain-behavior associations in children (Fair et al., 2012; Rohr et al., 2017; Vanderwal et al., 2021). While FC-MRI identifies consistent FC patterns across individuals (Damoiseaux et al., 2006; Fox et al., 2005), it is highly sensitive to artifacts from physiological sources, such as heart rate and respiration, and head motion (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). This presents a particular challenge in fMRI studies that include young children as they have increased head motion in the scanner (Dosenbach et al., 2017; Greene et al., 2018) leading to systematic artifacts (Fair et al., 2012; Power et al., 2012). For this reason, developmental fMRI studies have struggled to collect usable resting state data from young children, and consequently have often excluded participants younger than seven (Vanderwal et al., 2019), with at least one large study revising their study protocol to omit the resting scan from children younger than six years of age due to difficulty in obtaining high-quality data (Alexander et al., 2017).

To reduce head motion and increase compliance, early childhood FC-MRI studies are increasingly conducted using passive viewing tasks such as movies (Alexander et al., 2017; Moraczewski et al., 2018; Reynolds et al., 2020; Rohr et al., 2017; Vanderwal et al., 2019). Movie watching significantly reduces head motion in participants younger than 10 years of age, with Greene et al. (2018) reporting a more than 70% reduction in mean framewise displacement (FD). Importantly, FC networks derived from passive viewing paradigms have been shown to be globally similar to those derived from resting-state data (Bray et al., 2014; Greene et al., 2018; Vanderwal et al., 2019). Despite these improvements, as in all FC-MRI studies, head motion

noise remains a major concern, and noise mitigating preprocessing steps are essential to prepare data for analysis. Such steps involve tradeoffs, however, and remain widely debated (Murphy and Fox, 2017; Satterthwaite et al., 2019); there is currently no gold-standard preprocessing pipeline, for early childhood studies or otherwise.

In response to the proliferation of preprocessing approaches (Carp, 2012), several studies have compared the effectiveness of different pipelines (Churchill et al., 2017; Ciric et al., 2017; Kassinopoulos and Mitsis, 2021; Parkes et al., 2018; Taymourtash et al., 2020), but no study has specifically considered the impact of preprocessing steps on data from young children, which is important for several reasons. First, children move more than adults, even when implementing strategies such as passive viewing (Dosenbach et al., 2017; Greene et al., 2018), meaning that child datasets may be particularly sensitive to the impact of preprocessing choices. Further, smaller head size could alter the impact of rotational motion, and because GSR shows distance-dependent effects, this step warrants investigation in young children. Children also have faster respiratory and heart rates than adults (Fleming et al., 2011), which may alter the temporal characteristics of physiological noise. Notably, physiological noise is not accounted for with quality control – functional connectivity (QC-FC), a widely used benchmarking metric that is specific to head motion (Power et al., 2012; Satterthwaite et al., 2013). Impacts of physiological noise may be better assessed using metrics sensitive to information recovery.

Common motion-correction strategies include regression of motion estimates, independent component analysis (ICA)-based approaches, regression of the global signal, censoring volumes of high FD, and temporal filtering. Regressing out motion estimates remains one of the most common denoising approaches (Satterthwaite et al., 2019). However, based on QC-FC benchmarks, Ciric et al. (2017) suggest that this approach is insufficient on its own and

leads to concerns about losses in degrees of freedom. An alternative approach is to remove effects of head motion through ICA-based approaches, such as ICA-AROMA (Pruim et al., 2015), which decomposes data into components that reflect either brain activity or structured noise. ICA-AROMA automatically classifies components as noise using both temporal features (high frequency content, correlation with realignment parameters) and spatial features (near CSF or the edge of the brain) (Pruim et al., 2015). These structured noise components can then be regressed from the data (Thomas et al., 2002).

Global signal regression (GSR) is an often used but widely debated preprocessing step (Chai et al., 2012; Gotts et al., 2013; Murphy and Fox, 2017). GSR is a simple and arguably effective (Ciric et al., 2017; Parkes et al., 2018) denoising technique, improving the specificity of positive correlations and showing results that are more consistent with anatomical connectivity (Fox et al., 2009). However, the use of GSR tends to increase the apparent strength of short-range connections while decreasing the apparent strength of long-range connections (Ciric et al., 2017; Parkes et al., 2018; Saad et al., 2012), and creates anti-correlations which may not exist by effectively centering the connectome (Aguirre et al., 1998). The global signal has been shown to temporally resemble established networks and significantly associates with behavioral data (Li et al., 2019), suggesting GSR should be used with caution as it may remove signal of interest in addition to noise.

To mitigate the effect of specific motion contaminated volumes, it has become common to either remove, or interpolate over, specific time points by censoring or 'scrubbing' (Power et al., 2012). Despite efficacy in removing noise (Ciric et al., 2017) there are concerns with censoring, such as disrupting temporal autocorrelations, and leaving variable amounts of scan data between participants. Further, even if the number of censored volumes is matched across

individuals, not all volumes are equally rich in information (Power et al., 2015). Censoring strategies vary widely across early childhood studies (Fair et al., 2012; Miranda-Dominguez et al., 2018; Rohr et al., 2019; Vanderwal et al., 2021) and it is unclear which level of censoring optimizes the recovery of individual-specific FC information.

A further step to consider is temporal filtering, as fMRI signals at both very low frequencies (under 0.01 Hz) and high frequencies (above 0.1 Hz) are variably filtered out to remove noise (Satterthwaite et al., 2013). It has been suggested that filtering above 0.1 Hz may also be removing connectivity information (Niazy et al., 2011), or artificially increasing correlations by introducing sample dependence (Davey et al., 2013), leading to concerns regarding the appropriateness of a bandpass filter compared to a highpass filter (Satterthwaite et al., 2019).

Ciric et al. (2017) speculated that aggressive preprocessing choices could be removing both signal and noise, improving outcomes on metrics sensitive to noise removed, such as QC-FC, but in the process reducing sensitivity to individual differences. Intra-class correlation (ICC; Shrout and Fleiss, 1979) is a commonly used measure of test-retest reliability, and has been used as a preprocessing benchmark of signal retention after preprocessing (Parkes et al., 2018; Kassinopoulos and Mitis, 2021). However, both Parkes et al. (2018) and Kassinopoulos and Mitis (2021) counterintuitively found that pipelines with poor denoising had higher ICC, suggesting higher reliability with less denoising. An alternative metric that provides information about individual-specific information remaining or recovered following preprocessing is functional connectome fingerprinting (Finn et al., 2015; subsequently referred to simply as "fingerprinting"), which aims to match scans from the same participant based on connectome similarity. Analysis of connectome individuality through fingerprinting has become popular in

recent years, though how preprocessing affects FC information necessary for individual identifiability is not yet well understood.

As there is no ground-truth in resting FC-MRI analyses, it has been suggested that preprocessing benchmark studies should ideally consider task-evoked effects to ensure that while noise is mitigated signal variation of interest is preserved (Bijsterbosch et al., 2021). Here we leverage the passive viewing task to assess how preprocessing choices impact functional responses to the video using temporal intersubject correlation (ISC; Hasson et al., 2004). ISC is calculated as the temporal signal correlation between a given brain region across individuals during passive viewing of the same video stimulus.

The aim of the present study is to extend previous pipeline benchmarking studies (Ciric et al., 2017; Parkes et al., 2018) to passive viewing fMRI data from young children, while extending benchmarks beyond metrics sensitive to head motion, namely with the inclusion of ICC, fingerprinting, and ISC. As preprocessing steps likely have a larger impact in noisier samples, we also investigate whether preprocessing strategies are particularly efficacious depending on quantity of head motion by dividing scans into lower- and higher- motion groups. Our findings can support researchers conducting fMRI in early childhood samples to consider the tradeoffs and effectiveness of common preprocessing steps.

2.2 Methods

2.2.1 Participants

Data were collected as part of a longitudinal study of early childhood brain development (Dimond et al., 2020b, 2020a; Rohr et al., 2019, 2017). Participants were recruited from the local community through advertisements and existing databases. All procedures were approved by the

University of Calgary Conjoint Health Research Ethics Board. Parents provided informed consent and children provided assent to participate. Participants were children between 4 and 7 years of age at baseline without any major health concerns. Children were excluded if they had full-scale IQ more than 2 standard deviations below the standardized population mean of 100 on the Wechsler Preschool and Primary Scale of Intelligence – Fourth Edition, a history of neurodevelopmental or psychiatric disorders, or any neurological diagnoses. At the time the analyses for this study were initiated, 168 participants had completed a baseline scan and of these 59 (15 male) had completed a 12-month follow-up scan. From the sample of participants with both baseline and follow-up data, participants were included if after volume-wise censoring of the fMRI data (described in more detail below), both scans retained at least 11 minutes of uncensored data. 56 of the 59 children (14 male) reached this threshold, for a total of 112 scans used in the present study.

2.2.2 Data collection

MRI data were acquired on a 3T GE MR750w MRI (Waukesha, WI) scanner using a 32-channel head coil, at the Alberta Children's Hospital. fMRI was acquired during a passive-viewing task, where participants watched clips from a children's television show (Elmo's World) for 1100 s. Prior to scans, children underwent a practice scan in an MRI simulator during which they watched the same video and practiced staying still. fMRI scans were collected using a gradient-echo echo-planar imaging (EPI) sequence (TR = 2.5 s, TE = 30 ms, FA = 70°, voxel size 3.5x3.5x3.5 mm^3). An anatomical scan was acquired using a T1w 3D BRAVO sequence (TR = 6.764 ms, TE = 2.908 ms, FA = 10°, voxel size 0.8x0.8x0.8 mm^3).

2.2.3 Higher- vs. lower-motion subgroups

Preprocessing steps may impact data differently depending on the amount of contamination from motion or other sources. Low motion samples may see only modest improvements with more aggressive strategies such as censoring or GSR, whereas metrics derived from high motion samples will likely change more if, for example, a greater number of volumes are censored. Therefore, we consider here how preprocessing may differentially impact lower- and higher- motion scans. Towards this goal, after preprocessing pipelines were first compared using the whole sample of 112 scans, the 112 scans were median-split into two groups of 56 scans, based on average FD. The preprocessing pipelines were then separately compared for the higher- and lower-motion subgroups, using the metrics described below.

2.2.4 First-stage preprocessing steps common across pipelines

The following first-stage preprocessing steps were run on all scans, ahead of the specific pipeline variations tested here (described below). All preprocessing was carried out with custom Python scripts integrating Nipype functionality (version 1.1.5; Gorgolewski et al., 2011) using FSL version 6.0.0 (Smith et al., 2004), ANTs version 3.0.0.0 (Avants et al., 2011), and AFNI version 18.3.03 (Cox, 1996). Scans from the same individual were preprocessed separately. Structural (T1w) images were preprocessed using ANTs, including bias field correction, brain extraction, and tissue segmentation. For generation of WM and CSF time courses for regression, tissue masks were eroded using AFNI (CSF eroded twice, WM eroded 7 times).

Basic preprocessing for EPI data was as follows, largely following the procedure described by Ciric et al. (2018): a) FSL *MCFLIRT* to estimate head motion parameters (HMPs) and FD (Jenkinson et al., 2002). B) FSL *slicetimer* for slice time correction. C) FSL *MCFLIRT*

for rigid body realignment. We followed the recommendation in Power et al. (2017) to estimate

HMP on raw data but carry out slice time correction prior to rigid body realignment,

necessitating two FSL *MCFLIRT* steps. D) FSL *BET* to skull-strip EPI images (Smith, 2002). E)

ANTs *Registration* (Avants et al., 2011) to generate a transformation matrix to warp the EPI

image to a study-specific EPI template. This template was produced based on the procedure

described by Huang et al. (2010). Specifically, a 3D EPI reference image was taken from each

fMRI scan, chosen as a volume of low motion approximately in the middle of the scan. These

references were warped to MNI space, then averaged together and smoothed to create the final

study-specific template. F) FSL *FLIRT* boundary-based registration (Jenkinson et al., 2002) was

used to generate a transformation matrix to warp the EPI image to the T1w image, then the

inverse transformation matrix was used to warp tissue segmentations to functional image space.

All preprocessing and confound mitigation steps were carried out in native space. G) A linear

regression to remove the mean, linear trends, and quadratic trends from each voxel was

conducted. For pipelines that include censoring, time points marked for censoring were excluded

from the model to calculate linear and quadratic trends. In pipelines that did not include

censoring, all timepoints were included.

2.2.5 Preprocessing pipelines

After first-stage preprocessing, pipelines varied systematically in whether they used GSR,

censoring, and ICA-AROMA, as shown in Table 2.1. For comparison, we also tested two

minimal pipelines: M1, which included first-stage preprocessing and temporal filtering, but no

nuisance regression, and M2, which included first-stage preprocessing, temporal filtering, and

regression of the WM and the CSF time series.

Pipeline	Primary motion artifact removal	WM + CSF regression	Global signal regression (GSR)	Censoring
M1	None			
M2		Yes		
R1	Regress head motion parameters (HMPs)	Yes		
R2		Yes	Yes	
R3		Yes		Yes
R4		Yes	Yes	Yes
I1	ICA-AROMA	Yes		
I2		Yes	Yes	
I3		Yes		Yes
I4		Yes	Yes	Yes

Table 2.1. Preprocessing pipelines tested to compare effects of ICA-AROMA, global signal regression and censoring.

2.2.6 Additional preprocessing

2.2.6.1 ICA-AROMA

For pipelines I1 through I4, ICA-AROMA was applied immediately following the first-stage preprocessing steps described above. Given that this was a relatively high motion sample, we used ICA-AROMA's aggressive denoising feature, where all variance associated with noise components are removed (Pruim et al., 2015), but otherwise default options. The ICA-AROMA code was modified to warp to MNI space via the ANTs transformation matrix rather than a FNIRT transformation.

2.2.6.2 Temporal filtering

For all pipelines, data were temporal bandpass filtered (0.01 – 0.08 Hz) using a fast Fourier transform. For pipelines I1 through 14, this occurred following ICA-AROMA. For all other pipelines, this occurred immediately after first-stage preprocessing. In follow-up comparisons we tested the effect of a highpass filter; see section *2.2.9.1. Filtering.* To avoid

reintroducing artifacts, HMPs were also filtered, and WM, CSF, and the global signal were calculated following temporal filtering (Lindquist et al., 2019).

2.2.6.3 Nuisance regression

Following filtering, nuisance regression was applied as part of all pipelines, with the exception of M1, which did not undergo further preprocessing. Excepting M1, all pipelines included WM + CSF regression. In pipelines R1 through R4 the six HMPs were regressed. Pipelines R2, R4, I2 and I4 included GSR. For all nuisance parameters, both linear and quadratic terms, along with the first temporal derivative of those terms were regressed (i.e., 4 regressors per parameter). Censoring was carried out as part of the regression step for pipelines R3, R4, I3 and I4. As a default, we censored volumes above a FD threshold of 0.25 mm (based on FSL *MCFLIRT*, i.e., $FD_{Jenkinson}$ as described in Ciric et al., 2018), and censored only the identified frames. Varied censoring schemes were tested separately as described below. Pipelines are summarized in Table 2.1.

2.2.7 Connectome generation

Following preprocessing, each scan was registered to the study specific template using the ANTs transformation matrix. Each voxel was then assigned to one of 325 nodes (regions) within the MIST 325 parcellation (Urchs et al., 2019). The mean time course was calculated for each node by averaging the time courses for all voxels within the node. The Pearson correlation between each pair of nodes was then calculated, generating 52650 edges. Correlation values were Fisher z- transformed to better approximate a normal distribution.

2.2.8 Pipeline comparison metrics

2.2.8.1 Motion correlated edges (QC-FC)

QC-FC was calculated using the approach described in Satterthwaite et al. (2013). Specifically, for each edge the correlation was calculated between average FD (as estimated from the first FSL *MCFLIRT* step) and edge strength, across all 112 scans. Due to the concern that preprocessing steps such as GSR affect edges differently depending on the inter-node distance, we also assessed distance-dependent effects by plotting the edge strength-motion correlation vs. Euclidian inter-node distance for each edge (Satterthwaite et al., 2013). Inter-node distance was calculated as the Euclidian distance between the center of mass of the two nodes linked by a given edge, as defined by the MIST parcellation. Three QC-FC metrics were extracted: (1) the percentage of edges correlated with head motion at a p-value <0.05 uncorrected; (2) the mean of the absolute value of the correlation between head motion and edge strength; absolute value was used to assess the magnitude of the motion-effect, rather than the direction; and (3) QC-FC distance dependence as the correlation between edge strength-motion correlation and Euclidian edge length. In theory, more effective preprocessing pipelines will lead to fewer edges significantly associating with head motion, and ideally this relationship would not depend on distance, such that across these three metrics, smaller values are considered better.

2.2.8.2 Fingerprinting

Pipelines that aggressively remove noise may also have the undesirable side-effect of removing signal of interest (Ciric et al., 2017). We used functional connectome fingerprinting to determine whether information unique to an individual is reduced or amplified by noise-mitigation steps. Following the approach used by Finn et al. (2015) and related work (Byrge and

Kennedy, 2019; Kaufmann et al., 2017; Miranda-Dominguez et al., 2018), connectomes were vectorized, then correlated for each pair of scans. The fingerprinting match rate was calculated by determining how often a scan correlated most strongly with the second scan from the same individual, and dividing by the total number of scans (112). We expressed this as a percentage. For each scan we also calculated 1) group similarity, i.e., the mean correlation to a scan from another individual, 2) stability, i.e., the correlation between scans from the same person, and 3) individualization, the difference between stability and the highest correlation to a scan from another individual (nearest miss). Negative individualization scores reflect scans that did not successfully match. To assess whether pipelines affected associations with head motion, for each pipeline we regressed stability and individualization against head motion, calculating R^2, the slope, and the intercept. For each individual, we used the higher of their scans' two motion values for this analysis.

2.2.8.3 Intra-class correlation (ICC)

For each pipeline, we calculated the intra-class correlation (ICC; Shrout and Fleiss, 1979) of each edge as a measure of test-retest reliability. ICC was calculated in two ways: 1) a between session ICC, where we used the two scans from each individual to calculate the within-subject mean square, and 2) a within-session (split-half) ICC for each participant's first session, where each scan's time series was split in two in order to calculate the within-subject mean square. For pipelines that included censoring, split-half time series were generated based on volumes retained after censoring.

2.2.8.4 Intersubject correlation (ISC)

We used intersubject correlation (ISC; Hasson et al., 2004) to assess whether stimulus-evoked activity is amplified or suppressed by noise-mitigation. Preprocessing pipelines that fail to adequately remove noise, or remove task-evoked signal, will have lower ISC values. ISC values were calculated as the temporal correlation between pairs of individuals for each node, and averaged across pairs. For pipelines that included censoring, we did not include time points that were censored for either scan in a given pair. ISC calculated on passive viewing fMRI are typically highest in visual and auditory regions (Hasson et al., 2004; Kauppi et al., 2010). For pipeline comparisons, we identified the 10 nodes with the highest average ISC across all pipelines and averaged these generate a single ISC score for each scan for each pipeline. In addition to group-level ISC, we calculated the "intra-subject correlation", or the time series correlation between scans from the same participant.

As noted above, scans were median-split into higher- and lower- motion subgroups. For ISC, when both of a participant's scans were in the same subgroup, we only retained the more representative scan, i.e., the lower motion scan in the lower-motion subgroup, or the higher motion scan in the higher-motion subgroup.

2.2.8.5 Intra-scan inter-pipeline correlation

Preprocessing choices may have relatively large or relatively small effects on connectomes, and these may influence downstream analyses and convergence across studies. To assess the impact of preprocessing choices on FC estimates, for each scan we calculated the correlation between edge strengths across pipelines. These were Fisher z-transformed, averaged across the 112 scans, then converted back to correlations for ease of comparison.

2.2.9 Follow-up comparisons

2.2.9.1. Filtering

In addition to the highpass filter applied as part of nearly all fMRI analyses, the use of a lowpass filter has become common as a noise mitigation step in FC-FMRI due to frequencies above approximately 0.1 Hz being more highly associated with noise than signal of interest (Satterthwaite et al., 2019). However, this remains controversial, as connectivity information at higher frequencies will be lost (Niazy et al., 2011). We therefore repeated benchmark comparisons using a highpass (>0.01 Hz) rather than a bandpass filter. All other preprocessing steps were identical.

2.2.9.2 Varying censoring parameters

We conducted an additional comparison focusing specifically on the effect of different censoring strategies. There is variation in thresholds used in the literature, with older studies generally using more lenient censoring relative to more recent work (Satterthwaite et al., 2019). There is also variation in the literature as to whether censoring a single volume per motion artifact (Satterthwaite et al., 2013) or censoring multiple volumes per motion artifact (Power et al., 2012) is preferred. Thresholding decisions are challenging because there is limited guidance in early childhood samples.

We used pipeline R4 above, which performed well on all metrics, and modified the approach to censoring using 11 additional pipelines that varied in the censoring threshold and the volumes censored per motion artifact (Table 2.2). In pipelines that censored 3 volumes per motion artifact, we censored volumes immediately before and after any volume or series of volumes flagged for censoring. Similarly, for pipelines that censored 4 volumes per motion

artifact, 1 volume prior and 2 volumes after the motion-contaminated volume(s) were censored.

All 112 scans were included for these comparisons, even if stricter censoring caused scans to fall

below the threshold of 11 minutes of uncensored data.

Volumes censored per artifact	Censoring threshold (mm FD)	Mean and std of number of volumes censored
1	0.3	31.8 ± 35.0
	0.25	39.3 ± 41.1
	0.2	51.0 ± 49.8
	0.15	72.2 ± 62.2
3	0.3	57.6 ± 59.4
	0.25	69.7 ± 67.6
	0.2	88.6 ± 77.7
	0.15	121.9 ± 91.0
4	0.3	67.6 ± 68.2
	0.25	81.3 ± 76.6
	0.2	102.5 ± 86.5
	0.15	139.3 ± 98.5

Table 2.2. Preprocessing pipelines tested to compare effects of censoring at different thresholds and volumes censored per motion artifact. Note that the pipeline with 1 volume censored at a threshold of 0.25 mm is equivalent to pipeline R4 described in Table 1.

2.2.10 Statistical Comparisons

To statistically assess the impact of specific preprocessing steps, we compared pairs of

pipelines that differed in their use of GSR, censoring, ICA-AROMA, or temporal filtering. For

most analyses, pipelines were compared using paired samples t-tests with uncorrected p-values

and Cohen's d reported. We note there are multiple ways to calculate Cohen's d for a paired test;

here we use mean difference ÷ standard deviation of the difference. We do not report p-values

for edgewise comparisons due to the lack of independence between edges. We note that across

metrics sample sizes vary depending on what specifically is being assessed, so statistical analyses are not directly comparable.

2.3 Results

2.3.1 Participant characteristics

The mean age of participants at baseline was 5.47 years old (standard deviation: 0.76 years), and 6.52 for the second scan. Time from initial to follow up scan ranged from 0.88 to 1.19 years (mean: = 1.05 years, SD = 0.069 years). The average motion of our sample was high, with a mean average FD of 0.126 mm across all 112 scans (median average FD 0.095 mm). For comparison, in a study of adolescents aged 8-23, the whole sample had a mean average FD of 0.062 mm, with lower- and higher- motion subgroups having a mean average FD of 0.029 mm and 0.097 mm respectively (Satterthwaite et al., 2013).

The correlation in average FD between scans from the same participant was r = 0.18 (p = 0.19). While other studies found comparatively large within-subject FD correlation (Zeng et al., 2014), we attribute a smaller correlation to the long period of time between scans in a developing sample, where motion may be less trait-like. We used a linear mixed model to assess the correlation between age and mean FD and found a non-significant negative relationship (p = 0.186; Supplemental Figure 2.1, Appendix A).

For analyses in which scans were divided into two groups based on median head motion, the 56 low motion scans had a mean average FD of 0.062 mm (range: 0.035 to 0.094 mm, std = 0.016 mm). The 56 high motion scans had a mean average FD of 0.191 mm (range: 0.097 to 0.506 mm, std = 0.088 mm). Mean age at scan was similar across groups (lower-motion: 6.01 years, std = 0.920 years; higher-motion: 5.98 years, std = 0.935 years). For the higher- and

lower-motion group ISC analysis, scans were divided into two groups based on median head motion, such that only one scan per individual was included in each group, resulting in a total of 40 scans per group. The lower-motion scans had a mean average FD of 0.059 mm (range: 0.035 to 0.092 mm, std = 0.016 mm). The higher-motion scans had a mean average FD of 0.202 mm (range: 0.097 to 0.506 mm, std = 0.096 mm). Mean age at scan remained similar across groups (lower-motion: 6.13 years, std = 0.888 years; higher-motion: 6.03 years, std = 0.934 years).
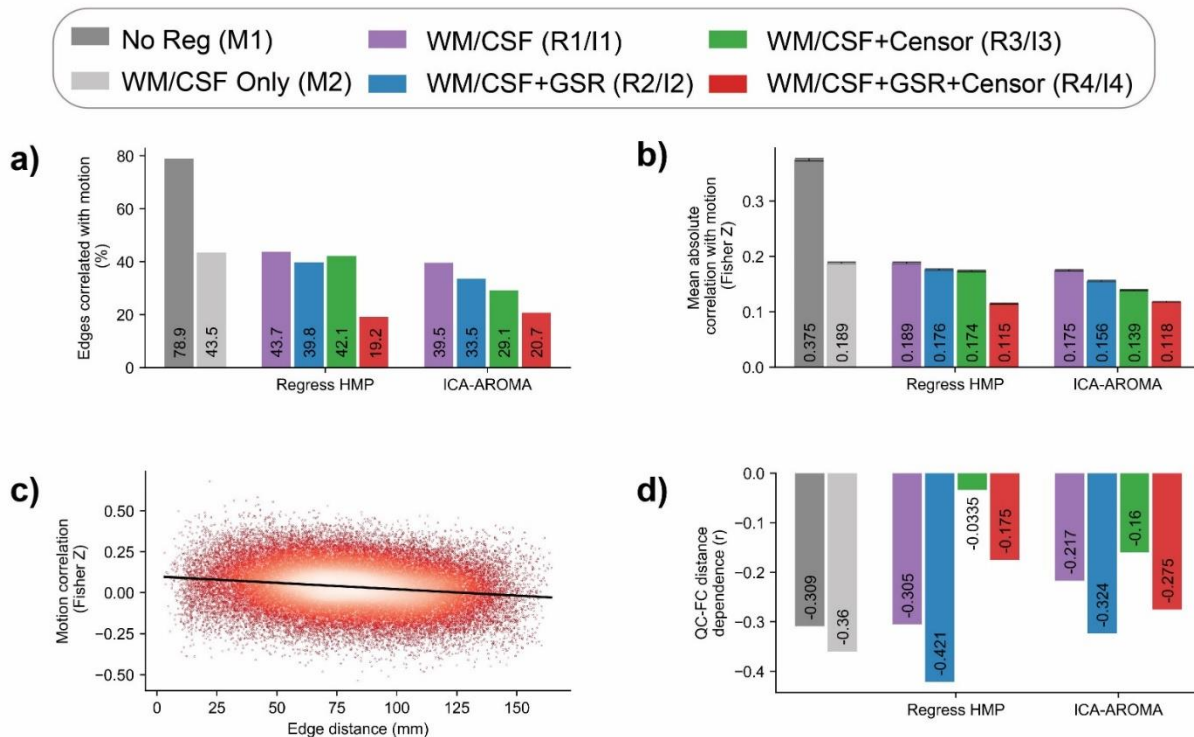
Figure 2.1. **Quality control-functional connectivity (QC-FC) across pipelines. a)** Percentage of edges with a significant correlation between edge value and head motion across all 112 scans (uncorrected $p < 0.05$). **b)** Mean and 99% confidence interval of the absolute correlation between edge strength and subject motion across all 112 scans. **c)** Example QC-FC distance dependence plot, from pipeline R4 (regress HMP + WM/CSF + GSR + censor). Each point is an edge in the connectome, plotted based on the length between its nodes (edge distance) and correlation between edge strength and subject motion. Overlapping points are represented by brighter colors. **d)** QC-FC distance dependence for each pipeline. This is the correlation between edge length and the association between edge strength and subject motion.

2.3.2 QC-FC

Figure 2.1a shows the percentage of edges significantly correlated with motion, ranging from 19% to 79%. In the minimally processed M1 pipeline, the vast majority of edges were correlated with motion (79%). Regressing only WM and CSF (M2) reduced this to 44%. Both GSR and censoring further reduced the number of edges correlated with motion, and pipelines that paired these two steps had the smallest number of motion-associated edges (e.g., R1: 44% vs. R4: 19%). In general, ICA-AROMA pipelines fared better than comparable pipelines that regressed HMP, though the pipeline with the fewest QC-FC associated edges (R4 – regress HMP+GSR+censoring) did not use ICA-AROMA.

Figure 2.1b (Supplemental Table 1, Appendix B) shows the mean absolute correlation with motion for all edges. Here the same trends are observed, with a large motion effect in minimally processed data (M1) greatly reduced with the inclusion of regressing WM and CSF (M1: $z = 0.375$ vs. M2: $z = 0.189$, $d = 0.88$). Again, GSR and censoring further reduced the effect of motion when paired (e.g., R1: $z = 0.189$ vs. R4: $z = 0.115$). Intermediate improvements, with smaller effect sizes, were seen when using only one of GSR or censoring (Supplemental Table 1).

Figure 2.1c shows an example from pipeline R4 of edge correlation with motion vs. edge distance with a linear fit. Figure 2.1d shows the QC-FC distance dependence for each pipeline. Pipelines that include GSR were associated with a more negative correlation (e.g., R1: $r = -0.305$ vs. R2: $r = -0.421$), suggesting that GSR affects shorter edges differently than longer edges, with shorter edges having more remaining motion influence and longer edges more likely to be negatively correlated with motion. However, censoring reduced this correlation, partially compensating for the effect of GSR.
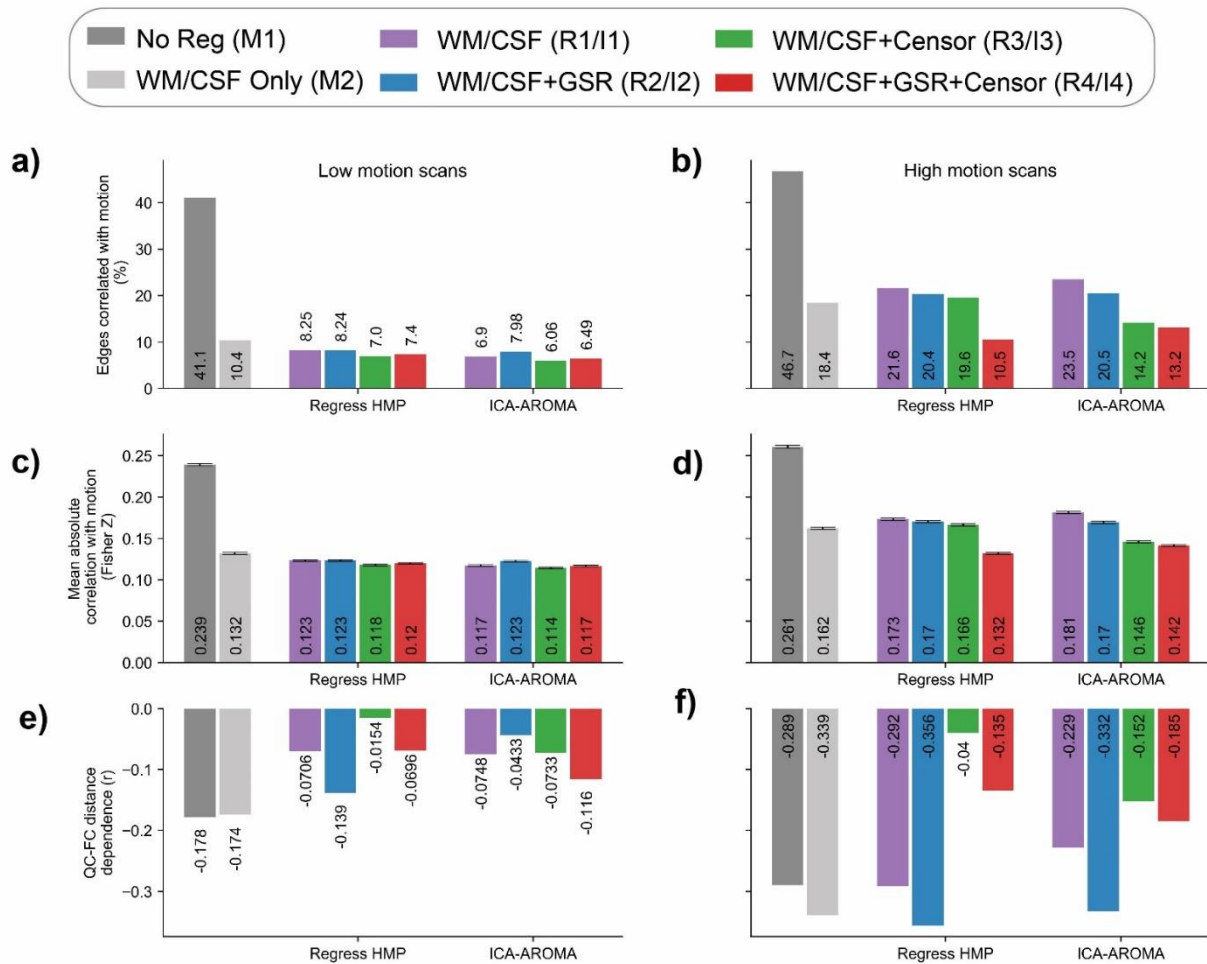
Figure 2.2. **Quality control-functional connectivity (QC-FC) across pipelines, separately for lower or higher motion scans.** Higher and lower motion scans were identified as above or below median average framewise displacement. **a, b)** Percentage of edges that have a significant correlation between edge strength and head motion (uncorrected $p < 0.05$). **c, d)** Mean and 99% confidence interval of absolute correlation between edge strength and subject motion. **e, f)** QC-FC distance dependence. The correlation between edge distance and association between edge strength and head motion.

When scans were split into lower- and higher- motion subgroups, the higher-motion scans fared worse on all metrics (Figure 2.2a vs 2.2b, 2.2c vs 2.2d, 2.2e vs 2.2f). Both the lower- and higher-motion groups had fewer edges significantly correlated with motion than the entire group (Figures 2.2a and 2.2b vs 2.1a), and differences between pipelines on mean absolute correlation with motion showed smaller effect sizes (Supplemental Tables 2 and 3 vs

Supplemental Table 1, Appendix B). We attribute these changes to a smaller sample size and a narrower range of motion values. In the lower- motion group, with the exception of the poor-performing minimal pipelines, choice of pipeline had a minimal effect on the percent of edges correlated with motion (ranging from 6.1% to 8.3% in pipelines R1-R4, I1-I4; Figure 2.2a). We saw a similar lack of impact on the mean absolute correlation with motion, ranging from $z = 0.11$ to $z = 0.12$ in pipelines R1-R4 and I1-I4 (Figure 2.2c), and effect sizes comparing pipelines smaller than 0.15 (Supplemental Table 2). Small improvements were seen with censoring, and pipelines that included GSR fared slightly worse. Most of the effects noted in the entire sample (Figure 2.1) were only seen in the higher-motion group (Figures 2.2b and 2.2d; Supplemental Table 3), where censoring and GSR used together improved the percentage of edges significantly correlated with motion (e.g., R1: 22% vs. R4: 11%) and the mean absolute correlation with motion (e.g., R1: $z = 0.17$ vs. R4: $z = 0.13$, though effect sizes remained small). For inter-pipeline distance dependence, trends were similar to the entire sample in lower- and higher-motion groups (2e and 2f vs 1d), with GSR making the association stronger, but censoring mitigating this effect.

2.3.3 Fingerprinting

Fingerprinting match-rate was assessed for each pipeline (Figure 2.3a). Overall, except for the two minimally processed pipelines (M1 and M2), both with under 50% success, the match rate was high for all pipelines tested, ranging from 71% to 90% (chance <1%). GSR had a minimal effect on the overall match rate (e.g., R1: 75.0% vs. R2: 76.8%), but pipelines that included censoring were more successful, especially when regressing HMP rather than using ICA-AROMA (e.g., R3: 88.4% and R4: 90.2%).
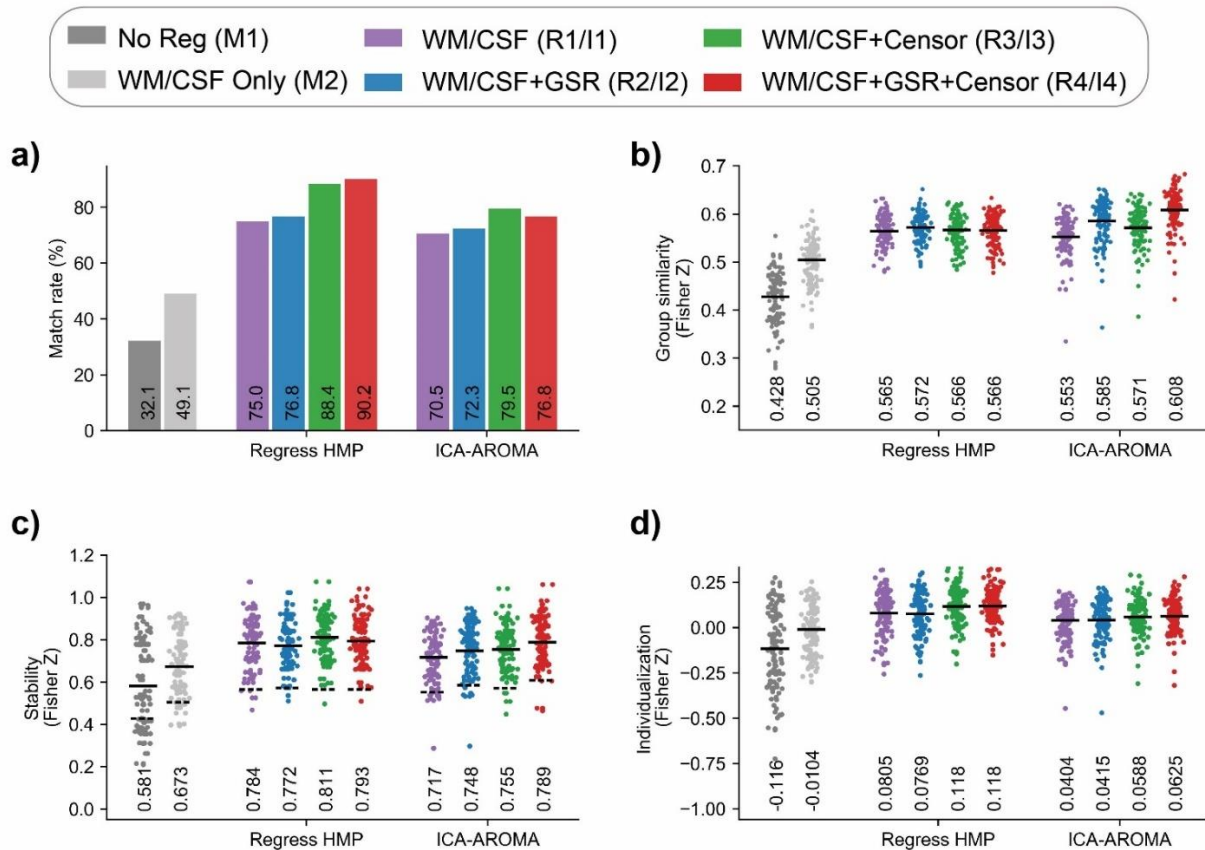
Figure 2.3. Functional connectome fingerprinting across pipelines. a) Match rate across pipelines. A scan matched if its highest correlation was to the other scan from the same individual. **b)** Group similarity across pipelines. Each dot represents an individual scan. Group similarity was assessed as the average correlation to scans from other participants. Lines represent mean values. **c)** Stability across pipelines. Each dot represents one individual. Stability was assessed as the correlation between scans from the same individual. Solid lines represent mean values, dashed lines represent the average group similarity (from 3b) for comparison. **d)** Individualization across pipelines. Each dot represents one scan. Individualization was assessed as the difference between stability and the highest correlation to a scan from another participant. Any point below 0 fails to successfully match. Lines represent mean values.

While group similarity varied minimally in terms of absolute numbers across pipelines

(Figure 2.3b), effect sizes comparing pipelines were large, suggesting consistent effects across

participants (e.g., R1: $z = 0.565$ vs. R2: $z = 0.572$, $p \approx 10^{-31}$, $d = 1.57$; Supplemental Table 4,

Appendix B). The two pipelines with the highest stability were R3 ($z = 0.81$) and R4 ($z = 0.79$;

Figure 2.3c), though interestingly pipeline I4 (ICA-AROMA+GSR+censoring) had a comparably

high stability (z = 0.79), but a smaller gap between stability and group similarity. Effect sizes between pipelines were large for stability (e.g., R1 vs. R3: d = 1.41, p $\approx 10^{-14}$; Supplemental Table 5, Appendix B).

Figure 2.3d (Supplemental Table 6, Appendix B) shows individualization (i.e., the fingerprinting margin) for each of the 112 scans. As expected, pipelines with a higher match-rate (Figure 2.3a) had a higher average individualization, which ranged from z = 0.040 to z = 0.081 for less successful pipelines (R1, R2, I1 through I4), while reaching z = 0.12 for R3 and R4 (d > 1.6 when comparing other pipelines to R3 or R4). There was no significant difference between R3 and R4. The similar distributions and large effect sizes between pipelines suggests that in more successful pipelines, scans that fail to match nonetheless see greater recovery of individual information via a reduced margin of failure.

Figure 2.4 shows individualization plotted against head motion; a higher r^2 value indicates greater sensitivity to motion noise. Censoring was the most impactful option on reducing this metric (e.g., R1: $r^2 = 0.40$ vs. R3: $r^2 = 0.27$), suggesting greater impact of censoring on higher motion scans. The participant with the most motion (mean FD of 0.51 mm for one of their two scans) could be considered an outlier – repeating the analysis with that participant removed reduces the r^2 and makes the slopes less negative, but trends between pipelines remain (Supplemental Figure 2.2, Appendix A). We repeated the analysis using stability rather than individualization (Supplemental Figure 2.3, Appendix A). Here we did not find differences between pipelines differing in GSR or censoring, but ICA-AROMA pipelines tended to have modestly higher r^2 values on this metric (e.g., R2: $r^2 = 0.36$ vs. I2: r^2 0.41).
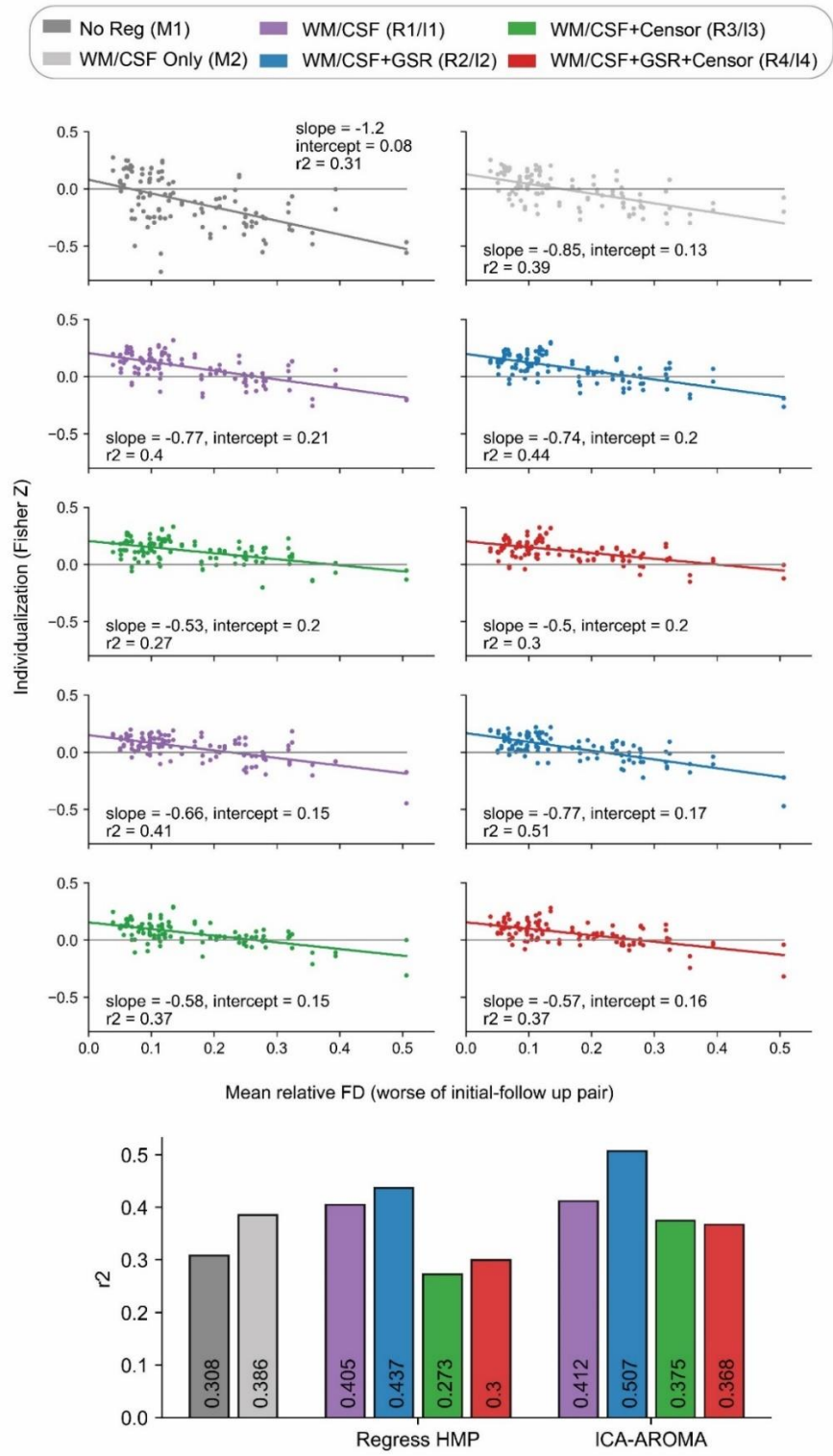
Figure 2.4. Individualization as a function of head motion. The difference between each scan's stability (self-correlation) and the highest correlation to a scan from another participant, plotted by motion. Motion was calculated by taking the worse of each participant's two scans' mean relative framewise displacement; each participant has two points for their two scans. Any point below 0 on the y axis fails to successfully match.
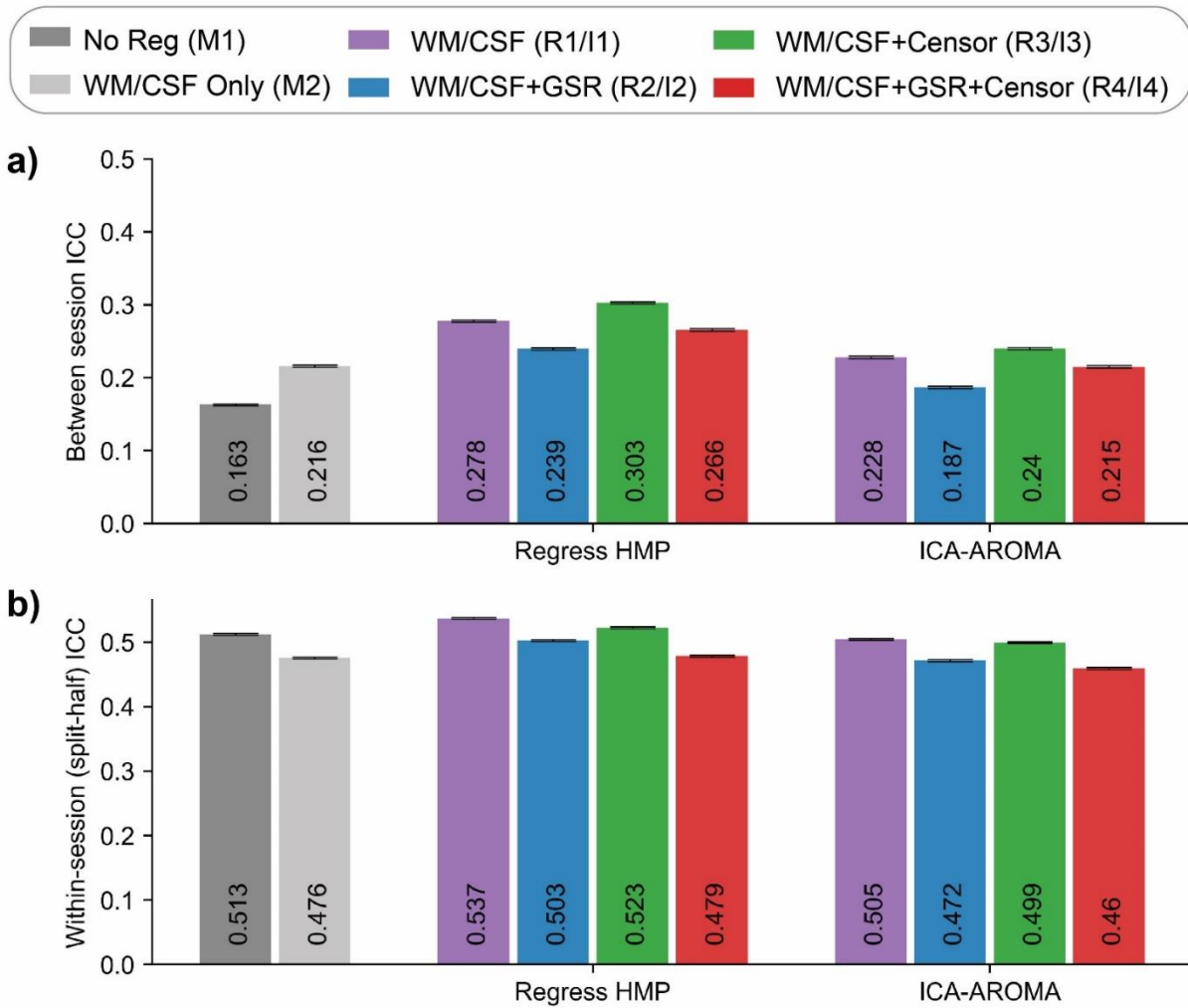
Figure 2.5. Intra-class correlation (ICC). a) Mean and 99% confidence interval for between-session ICC across pipelines. **b)** Mean and 99% confidence interval for the within session (split-half) ICC by pipeline.

2.3.4 Intra-class correlation (ICC)

For between-session ICC (Figure 2.5a; Supplemental Table 7, Appendix B), pipelines

that regressed HMP had higher average ICC than corresponding pipelines using ICA-AROMA

(e.g., R1: 0.278 vs. I1: 0.228, d = 0.32). Across pipelines, the use of GSR lowered ICC

coefficients (e.g., R1: 0.278 vs. R2: 0.239, d = 0.34). Meanwhile, the use of censoring raised ICC

coefficients (e.g., R1: 0.278 vs. R3: 0.303, d = 0.27). The two minimal preprocessing pipelines (M1 and M2) had low average ICC coefficients, of 0.163 and 0.216 respectively.

Within-session ICC showed a much smaller range across pipelines, ranging from 0.46 to 0.537 (Figure 2.5b; Supplemental Table 8, Appendix B). M1, the pipeline with no regressors, had the third highest mean ICC (0.513). Both GSR and censoring resulted in lower ICC coefficients, with the value dropping from 0.537 for R1 (regress HMP+no GSR+no censor) to 0.479 for R4 (regress HMP+GSR+censor). Pipelines differing in censoring and regressing HMP vs ICA-AROMA had small effect sizes (d < 0.27), while pipelines differing in use of GSR had larger effect sizes (d > 0.37).

2.3.5 Intersubject correlation (ISC)

Figure 2.6a (Supplemental Table 9, Appendix B) shows each scan's mean ISC across pipelines, averaged across the 10 nodes with the highest values (See Supplemental Figure 2.4 for location of these nodes, Appendix A). Both GSR and censoring improved ISC, with large effect sizes (d > 2). Figure 2.6b (Supplemental Table 10, Appendix B) shows each participant's "intrasubject correlation" across their two scans. While intrasubject correlations showed a greater range than mean ISC, this is likely related to being a comparison between only two scans, rather than an average across many scans. Average intrasubject correlations were higher than inter-SC, but trends between pipelines were the same, with GSR and censoring raising the correlation ($p < 10^{-7}$, d > 0.89).

When only lower- motion scans were compared using ISC (Figure 2.6c; Supplemental Table 11, Appendix B) the benefits of censoring were reduced, with or without GSR. GSR itself increased ISC (e.g., R1: z = 0.19 vs. R2: z = 0.22, $p \approx 10^{-22}$, d = 3.14). However, when only

higher- motion scans were compared (Figure 2.6d; Supplemental Table 12, Appendix B), the

benefits of censoring were more apparent, especially in combination with GSR (e.g., R2: z =

0.17 vs. R4: z= 0.19, p ≈ 10^{-24}, d = 3.77).
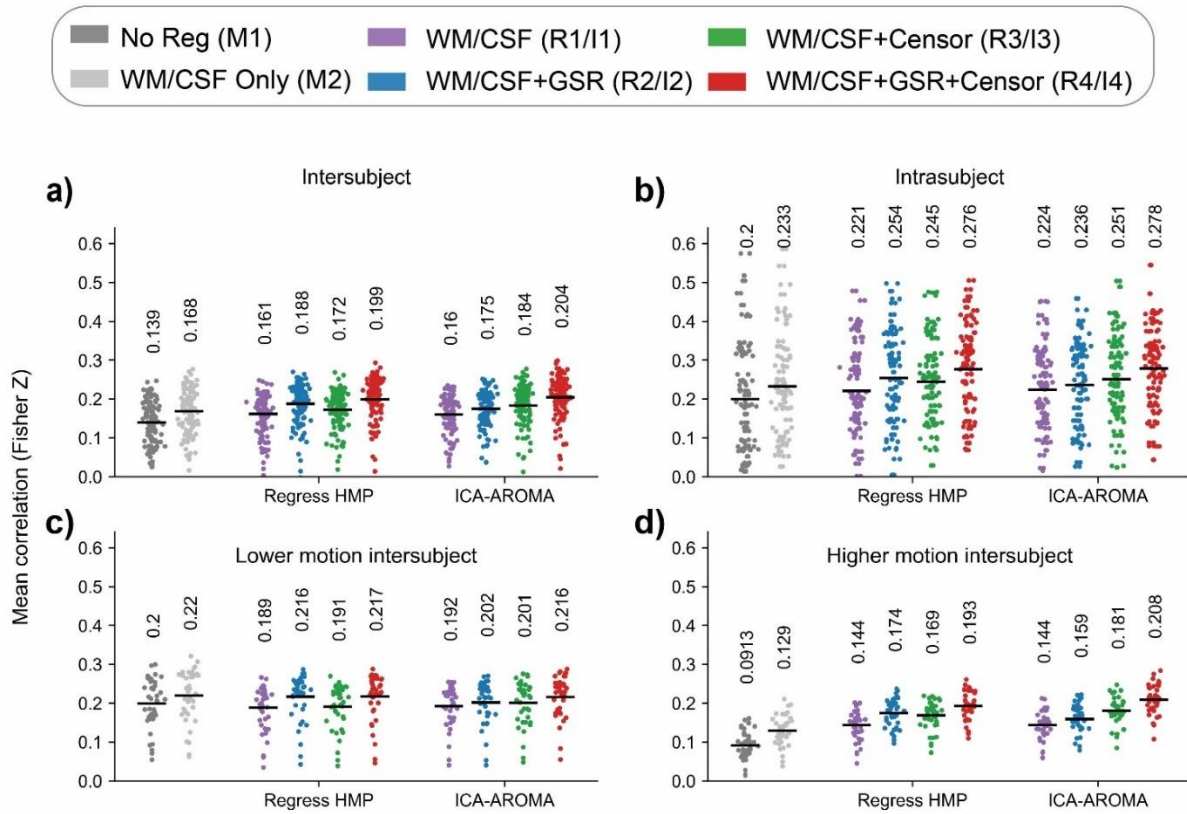
Figure 2.6. Mean intersubject correlation (ISC) values. Each point represents the mean ISC for a scan to all other scans, averaged across the 10 nodes with the highest ISCs. Lines represent mean values. **a)** Mean ISCs for all 112 scans, each compared to all other scans. **b)** Each participant's intrasubject correlation, the correlation between time series from the two scans' time series. **c)** Mean ISCs for each of the 40 below-median-motion scans. **d)** Mean ISCs for each of the 40 above-median-motion scans, compared to the other 39.

2.3.6 Intrascan inter-pipeline correlation

With the exception of the two minimally preprocessed pipelines, correlations between

connectomes from different pipelines were high overall, ranging from 0.67 to 0.97 (Figure 2.7a).

A correlation as low as 0.67 for the same scans preprocessed in two different ways – in this case
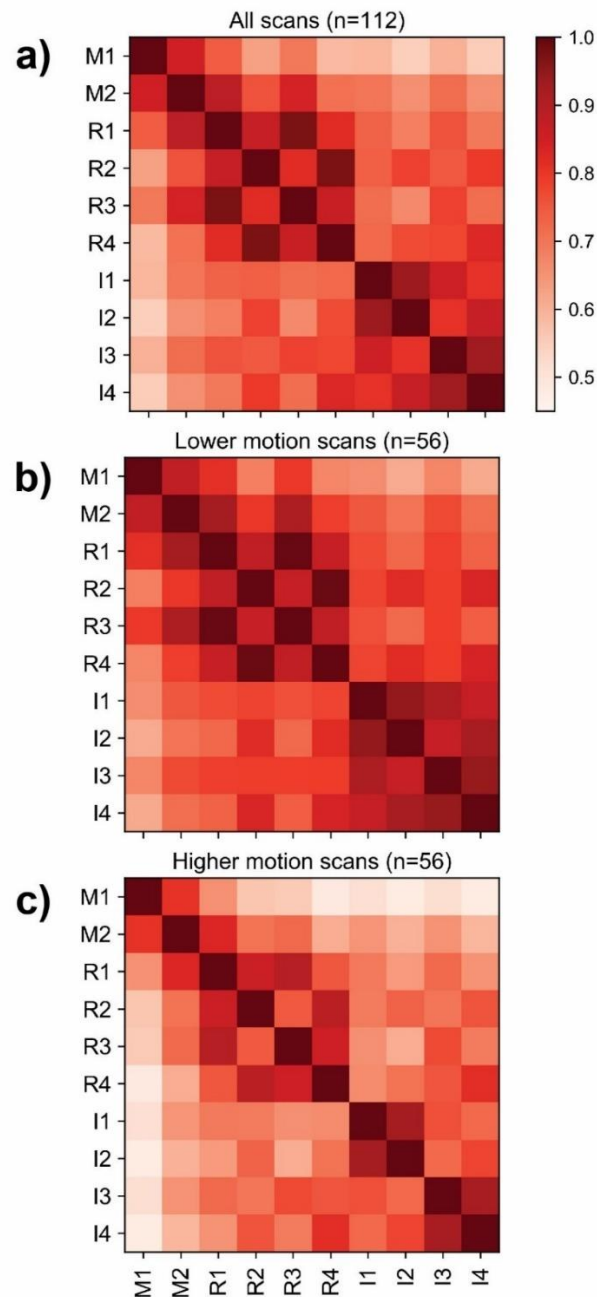
Figure 2.7. Intrascan inter-pipeline correlations. For each pair of pipelines, the correlation between connectomes from the same individual across the two pipelines was calculated, Fisher z-transformed, averaged across scans, then converted back to correlations. Pipelines are listed in Table 1, briefly: M1 – no regression; M2 – regress WM/CSF; R1 – regress HMP + WM/CSF; R2 – regress HMP + WM/CSF + GSR; R3 – regress HMP + WM/CSF + censor; R4 – regress HMP + WM/CSF + GSR + censor; I1 – ICA-AROMA + WM/CSF; I2- ICA-AROMA + WM/CSF + GSR; I3 – ICA-AROMA + WM/CSF + censor; I4 – ICA-AROMA + WM/CSF + GSR + censor. **a)** Average correlation based on all 112 scans**. b)** Average correlation using the 56 scans below the median average framewise displacement**. c)** Average correlation using the 56 scans above the median average framewise displacement.

between pipelines R3 (regress HMP+no GSR+censor) and I2 (ICA-AROMA+GSR+no censor) –

highlights concerns about replicability given the influence of preprocessing choices. Figure 2.7

also highlights how each preprocessing choice progressively alters functional connectivity

estimates. Pipelines that differed only in the use of one of GSR or censoring were similar (e.g.,

R1 vs. R3: r = 0.97) while pipelines that differed in multiple preprocessing steps showed

incremental differences (e.g., R1 vs. R4: r = 0.82). The choice to use ICA-AROMA vs.

regressing HMP had a relatively larger effect, with correlations ranging from 0.73 to 0.83.

Again, these correlations decreased further if the pipelines also differed in use of GSR or

censoring. Unsurprisingly, preprocessing choices had a smaller effect on lower-motion scans

(Figure 2.7b), with correlations between pipelines ranging from 0.72 to 0.99 for pipelines R1-R4,

I1-I4, though were as low as 0.61 between M1 (regress nothing) and I2. For higher-motion scans

(Figure 2.7c), correlations ranged from 0.61 to 0.93 for pipelines R1-R4, I1-I4, dropping as low

as 0.48 between M1 and I4.

2.3.7 Filtering comparison

For highpass relative to bandpass filtering, overall trends between pipelines were quite

similar (see Supplemental Figures 2.5-2.7, Appendix A). With a highpass filter, the number of

edges significantly correlated with motion ranged from 25-50% for pipelines R1-R4, I1-I4

(Supplemental Figure 2.5a). Other than for pipeline I1 (which fared poorly on this metric

regardless), bandpass filtering resulted in fewer edges significantly correlated with motion than

highpass filtering (e.g., R4: 19% for bandpass vs. 26% for highpass). Similarly, pipelines that

used bandpass filtering had a smaller mean absolute correlation with motion (e.g., R4: $z = 0.12$

for bandpass vs. $z = 0.13$ for highpass, $d = 0.24$; Supplemental Figure 2.5b; Supplemental Table

13, Appendix B). Highpass filtering resulted in more negative motion vs. edge distance correlations than bandpass filtering (e.g., R4: r = -0.304 for highpass vs r = -0.175 for bandpass; Supplemental Figure 2.5c), suggesting a bandpass filter leads to reduced distance dependent effects.

For fingerprinting, running the analysis with a highpass filter usually led to a lower fingerprinting match rate (Supplemental Figure 2.6a), such as 67% vs 77% for pipeline R2 (regress HMP+GSR+no censor). However, pipeline R4 (regress HMP+GSR+censor; 90% match rate) was equally successful using either style of filter. Pipelines that used highpass filtering showed increased stability and group similarity (e.g., for mean stability, R4: z = 0.87 for highpass, z = 0.79 for bandpass, $p \approx 10^{-128}$, d = 11.22; Supplemental Figure 2.6b; Supplemental Tables 14 and 15, Appendix B), suggesting a highpass filter makes all scans' FC estimates more similar to each other, relative to bandpass. Filtering choice had a variable effect on individualization (Supplemental Figure 2.6c; Supplemental Table 16, Appendix B). For some pipelines, using a highpass filter led to a lower fingerprinting margin (e.g., R2: z = 0.057 for highpass, z = 0.077 for bandpass, $p \approx 10^{-7}$, d = 0.051), but for pipelines R3 and R4 this margin of fingerprinting success improved with a highpass filter (e.g., R4: z = 0.13 for highpass, z = 0.12 for bandpass, $p \approx 10^{-10}$, d = 0.27). Bandpass filtering increased ISC and intrasubject correlations for all pipelines compared to highpass filtering (e.g., for ISC, R4: z = 0.18 for highpass, z = 0.20 for bandpass, $p \approx 10^{-74}$, d = 3.74; Supplemental Figure 2.7; Supplemental Tables 17 and 18, Appendix B).

2.3.8 Censoring comparison

Starting from the pipeline that performed best on most metrics (R4), we repeated our analyses with variable censoring. When one volume was censored per motion artifact, a lower (more stringent) threshold for censoring resulted in fewer edges correlating with motion, lowering from 20% of edges at a threshold of 0.30 mm, to 17% of edges at a threshold of 0.15 mm (Figure 2.8a). However, the censoring threshold had a minimal effect on the number of edges significantly correlated with motion when 3 or 4 volumes were censored per artifact, ranging from 13% to 15%. Effects were similar in the mean absolute correlation with motion (Figure 2.8b; Supplemental Table 19, Appendix B), with a lower threshold at one volume per artifact reducing the correlation from $z = 0.116$ at 0.30 mm to $z = 0.108$ at 0.15 mm, but having minimal effect at 3 or 4 volumes (ranging from $z = 0.099$ mm to 0.104 mm). However, effect sizes between pipelines were comparatively small ($d < 0.37$) for this analysis.

The fingerprinting match-rate decreased as censoring became more stringent (Figure 2.8c), suggesting that although censoring can help with identifiability, there is an optimal range, above which individual information is reduced. At 1 volume censored per motion artifact, the decrease was from 90% to 88% across thresholds (0.3 mm to 0.15 mm), but at 4 volumes censored per artifact the match-rate fell from 88% to 80%. With stricter censoring, both stability and group similarity decreased (Figure 2.8d; Supplemental Tables 20 and 21, Appendix B), for example at 3 volumes per motion artifact the mean stability dropped from $z = 0.778$ to $z = 0.720$ across thresholds ($p \approx 10^{-20}$, $d = 1.40$). Mean individualization was less affected by censoring (Figure 2.8e; Supplemental Table 22, Appendix B), though effect sizes remained large for most comparisons between pipelines. Censoring had a minimal effect on mean ISC (Figure 2.8f;

Supplemental Table 23, Appendix B), except for the strictest censoring. With ISC, effect sizes

remained large for most comparisons between pipelines.
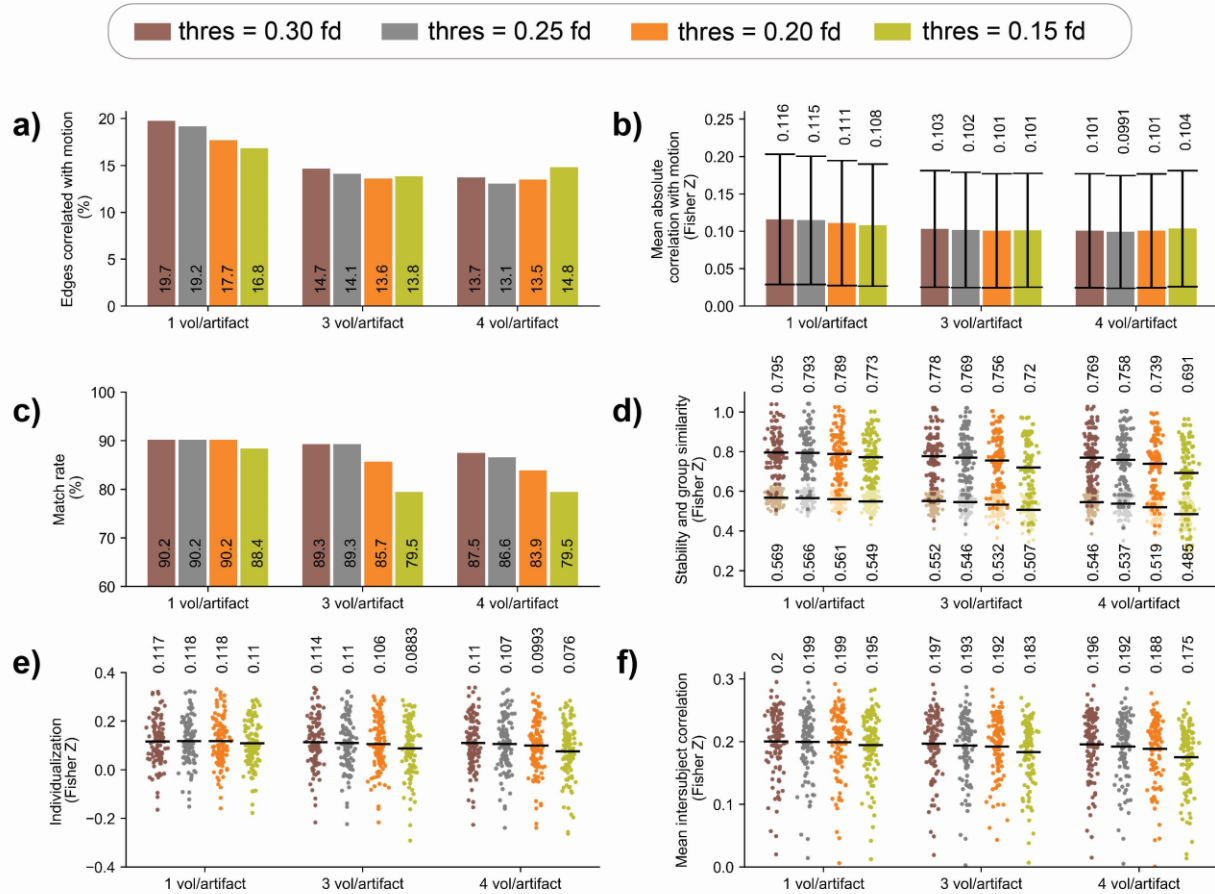
Figure 2.8. Pipeline benchmarks with different censoring thresholds. Colored bars depict four threshold levels and across the x-axis a different number of volumes is censored around the high motion frame. **a)** Percentage of edges with a significant correlation between edge strength and head motion across all 112 scans (uncorrected p < 0.05). **b)** Mean and 99% confidence interval for absolute correlation between edge strength and motion across all 112 scans. **c)** Fingerprinting match rate across pipelines. **d)** Each scan's stability (darker points) and group similarity (lighter points), for each pipeline. Group similarity was assessed as the average correlation to scans from other participants; stability was assessed as the correlation between scans from the same individual. Lines represent mean values. **e)** Individualization across pipelines. Each dot represents one scan. Individualization was assessed as the difference between stability and the highest correlation to a scan from another participant. Lines represent mean values. **f)** Mean ISCs for all 112 scans.

2.4 Discussion

      This study extended preprocessing benchmark comparisons, which have previously been performed in youth and adult resting data, to an early childhood passive viewing sample. Our results support and extend previous findings, suggesting that the highest performing pipelines include both GSR and censoring and that these steps have greater impact in higher-motion data, but remained beneficial in lower-motion data. Further extending previous work, we used connectome fingerprinting to estimate how individual-specific information was retained, or enhanced, by preprocessing steps. In fingerprinting, we found that volume censoring conferred more benefit than GSR, and pipelines with HMP regression outperformed pipelines with ICA-AROMA. When examining ISC, GSR offered the greatest improvement, with censoring showing a specific benefit for high motion scans. Censoring multiple volumes per motion artifact improved QC-FC metrics but had minimal or even a negative effect on other metrics, suggesting an important tradeoff between removal of noise and signal of interest. When examining overall effects of preprocessing choices on connectomes, we found that connectomes differed substantially depending on preprocessing choices, particularly in our higher motion subgroup, which is a concern for cross-study replicability in early childhood fMRI. Overall, our work suggests that for relatively high motion data from early childhood the best preprocessing pipeline includes a bandpass filter, GSR, regressing HMP rather than using ICA-AROMA, and moderate censoring.

      When comparing ICA-AROMA pipelines to pipelines that regressed HMPs, we found ICA-AROMA to be less effective on most metrics. This is somewhat surprising, given that ICA-AROMA primarily works by removing independent components strongly associated with motion estimates (Pruim et al., 2015). Previous benchmarking studies found that ICA-AROMA pipelines

were among the most effective of pipelines that did not include censoring in terms of reducing the quantity of edges correlated with motion – more effective than regressing HMP – though the difference was small between ICA-AROMA+GSR and HMP+GSR (Ciric et al., 2017; Parkes et al., 2018). While our QC-FC results are consistent with previous findings, we found that ICA-AROMA performed relatively poorly on individualization metrics. Even when ICA-AROMA fared comparably well on stability (as in pipeline I4), a necessary component for individualization, it had an elevated similarity to others, decreasing individual identifiability relative to pipelines that regressed HMP. However, ICA-AROMA pipelines showed no obvious difference from regressing HMP on ISC metrics, which measure the evoked response common across participants. One interpretation could be that ICA-AROMA removes both noise and individual-specific information although retaining signal common across participants. Further work is needed to clarify whether ICA-AROMA is less effective in samples of young children or high-motion samples, or whether connectome identifiability is generally poorer in ICA-AROMA relative to HMP pipelines. Other implementations of ICA-based preprocessing exist and are worth systematically investigating. ICA-AROMA is often used in conjunction with ICA-FIX, another ICA-based denoising strategy (Salimi-Khorshidi et al., 2014), and ICA-AROMA can also be run without their aggressive denoising option. Some studies have combined ICA-based denoising with regressing HMPs (Jalbrzikowski et al., 2020; Kaufmann et al., 2017), which may have advantages, though effectively removes aspects of the signal related to motion twice, contrary to the original intent of ICA-AROMA (Pruim et al., 2015).

While GSR remains contentious in the field, our results suggest there are advantages to using GSR in a high-motion dataset such as the early childhood sample used here. GSR reduced the number of edges significantly correlated with motion, especially when applied with

censoring. Similarly, while there are distance-dependent effects when GSR is used, this effect was largely mitigated in pipelines that also included censoring, consistent with QC-FC results from previous studies. Ciric et al. (2017) found that their 6 best pipelines, based on fewest edges correlated with motion, all included GSR, while Parkes et al. (2018) found that all 6 pipeline-pairs that differed in the use of GSR benefited from its inclusion. Both studies found that GSR introduced distance-dependent effects but that censoring diminished the impact. Ciric et al. (2017) suggests that GSR introduces these effects by more effectively denoising long range connections; we speculate that censoring mitigates this by removing the worst motion damage that may be missed by GSR. While these aforementioned studies both considered resting state data from adults, a benchmarking paper on fetal fMRI (Taymourtash et al., 2020) found similar benefits to GSR, suggesting that GSR has benefits regardless of age range or scan protocol.

Notably, however, GSR had no obvious impact on connectome individualization metrics. This suggests that while GSR alters FC estimates, most notably via distance-dependent effects, these changes may have a more global and less regionally specific impact compared to censoring (Power et al., 2015). The finding that GSR had a minimal impact on fingerprinting, our metric of individual information, may alleviate concerns about its use when comparing healthy controls and clinical populations, or in developmental research, even in light of research that suggests the global signal is significantly related to life outcomes and psychological function (Li et al., 2019). We also found benefits to the use of GSR in improving signal-to-noise in ISC metrics, suggesting that the benefits in terms of motion and physiological noise removal in an early childhood sample may outweigh the cost of removing some signal of interest (Behzadi et al., 2007).

We found relatively small effects of a bandpass filter compared to a highpass filter, with the main advantage of bandpass filtering on QC-FC metrics. In connectome fingerprinting, while stability was higher in pipelines that used a highpass filter, so too was group similarity by comparable levels. This suggests that any resulting changes to the functional connectome by including higher frequencies are consistent across individuals and are unlikely to reflect individual-specific information. Bandpass filtering also increased ISC metrics; this may suggest that bandpass filtering is effective for removing noise while retaining task signal, but this improvement may also be explained in part by signals composed of a narrower range of frequencies being inherently more similar across individuals.

We found benefits to volume censoring across metrics. On QC-FC benchmarks, we found that censoring improved both the quantity of edges associated with motion and distance-dependent effects. These effects are in line with Ciric et al. (2017) and Parkes et al. (2018) who both found that censoring pipelines outperformed alternatives on edges associated with motion, with both studies finding that the most effective pipelines included censoring. Our work extends these findings to early childhood passive viewing fMRI as well as to other benchmarks. In connectome fingerprinting, censoring increased accuracy more than any of the other preprocessing steps compared here, with censoring also conferring additional benefits to pipelines that used ICA-AROMA, despite ICA-AROMA being intended partially as an alternative to censoring (Pruim et al., 2015). Censoring had only a small effect on lower-motion scans for ISC comparisons, but censoring increased ISC between higher-motion scans.

When only one volume was censored per motion artifact, the more stringent the censoring threshold, the fewer edges significantly correlated with motion, dropping from 20% to 17% of edges. While this pattern was less appreciable when censoring multiple volumes per

motion artifact, even at the most lenient threshold tested multivolume censoring outperformed single volume censoring on this metric. Stricter censoring tended to decrease both stability and group similarity. This suggests that removing data points removes information that is common across participants. Interestingly, the distribution of individualization values was relatively unaffected by the censoring threshold, except when very strict censoring was used, even though the overall match rate decreased when censoring became stricter. This suggests not all scans are affected in the same way by changes in censoring parameters. While censoring had a minimal effect on mean ISC, except for the strictest censoring, it should be noted that for the purposes of ISC, censoring one scan necessarily censors the scan it is being compared to, since the same time points need to be compared. A stricter censoring protocol will lead to a disproportionate reduction in shared time points, so benefits in the removal of noise may be balanced by the decrease in data being compared. Altogether, while we found censoring to be beneficial, and can advise against overly strict censoring, the exact implementation of censoring seems to involve tradeoffs, making it challenging to advocate a one-size-fits-all approach. Future work may further consider how censoring is implemented, with the possibility of utilizing a scan-specific approach.

Previous work has found that ICC values are generally higher in more minimal pipelines that perform worse on other benchmarks (Parkes et al., 2018; Kassinopoulos and Mitis, 2021), suggesting a trade-off between reliability and validity (Noble et al., 2021) due to correlated noise across sessions. Here, we show that within-session ICC scores are consistent with previous work, but between-session ICC values (with ~12 months between sessions) were generally higher in pipelines with more aggressive denoising. Our interpretation of differences across studies is that greater time between scans in a developmental sample reduces the shared influence of motion

and physiological artifacts. Indeed, in our sample the correlation between initial and follow-up scan head motion was non-significant. This interpretation is supported by analyses in Parkes et al. (2018), who similarly found that while poorer performing pipelines had higher ICC, this effect was reduced when scans were collected on average 90 days apart, rather than within the same session. Given the conceptual difficulty of separating reproducible head motion from its biological confounds (Engelhardt et al., 2017; Hodgson et al., 2017), our data presents an advantage. We further note that ICC or other measures of test-retest reliability may be imperfect benchmarks for addressing head motion, and a broader collection of benchmarks, such as those included here, may be more appropriate to best understand how preprocessing choices affect signal changes (Kassinopoulos and Mitis, 2021).

Our findings on the intrascan interpipeline correlations are concerning towards the goal of replicable results in FC-MRI studies. While preprocessing choices have a smaller impact on lower-motion data, our work suggests that when analyzing high motion data, such as data from young children, each preprocessing choice has a measurable and cumulative effect on FC estimates (Li et al., 2021). While broader methodological differences or group-level participant differences are often cited as reasons for different findings among studies, we suggest that differences in preprocessing are likely to also play a role. Preprocessing choices should be both closely considered and accurately reported. While we compared the correlation between the entire functional connectome, future work should explore regionally-specific effects in more detail.

There are several limitations to our work. Our study was done in young children who were participating in a passive viewing task, and our results may not be directly applicable to other populations or other study protocols. Likewise, the benefits of any step we tested may not

be transferable to a more traditional task-based fMRI study, especially censoring data points during the task. We note that several of the steps implemented here could have been applied in alternative forms, which also restricts generalizability. For example, we used a 24 HMP model rather than 6 or 12 HMP (Friston et al., 1996). Similarly, there are many other thresholds that can be used for temporal filtering or censoring, or conceptually similar strategies such as despiking (Patel et al., 2014). There are also other possible approaches to noise mitigation, such as CompCor (Behzadi et al., 2007) or ANATICOR (Jo et al., 2010), that warrant further investigation.

We also acknowledge that choices such as registration and parcellation may have impacted our findings. Previous studies using functional connectome fingerprinting have used a similar number or fewer nodes; for example, Finn et al (2015) used a 268-node atlas while Miranda-Dominguez et al. (2018) used a 333-node atlas. Finn et al. (2015) also found lower fingerprinting success using a 68-node atlas; it is unknown if a higher number of nodes may offset the difference between preprocessing strategies on fingerprinting metrics, especially if considering a specific subset of edges rather than the full connectome (Byrge and Kennedy, 2019). While we chose metrics that we believe are meaningful to gauge the effect of preprocessing, other metrics could have instead been chosen. For example, we did not consider the effect preprocessing has on the magnitude of FC edges, and we also did not consider the effect preprocessing might have on individual networks within the broader connectome (Kassinopoulos and Mitsis, 2021), or other changes in functional organization.

This study implemented ISC by comparing nodes across scans, rather than a voxel-based approach (Hasson et al., 2004). This has the potential to average out the effect of the task, as not all voxels within a parcel will respond to the passive viewing task in a similar way. We also

chose to focus on the nodes with the highest average ISC values. These choices limit the extent

our findings can be generalized across the whole brain or to voxel-specific changes. Our analysis

has assumed that a higher ISC value reflects greater recovery of true signal, based on the

assumption that unremoved noise will lower the temporal correlation between scans. Parkes et al.

(2018) suggest that reproducible, individual-specific noise can increase FC test-retest reliability.

It is unknown whether noise has a similar effect between subjects in the context of ISC, though

there is the potential for head movements to be more likely at specific times in a video (e.g., due

to laughter). Future work is needed that specifically investigates the effect of preprocessing

choices on ISC values across the brain, especially in high noise scans.

2.5 Conclusions

Due to the impact that both head motion and preprocessing choices have on FC estimates,

benchmarking preprocessing steps in high-motion early childhood samples is critical to support

researchers in making informed choices. While in different datasets the optimal preprocessing

choices may vary, our results suggest that GSR, censoring, a bandpass filter, and HMP

regression are preferable in high motion datasets from early childhood populations engaging in a

passive viewing task. In particular, GSR and censoring showed few disadvantages across our

metrics, and ICA-AROMA showed no major improvement compared to regressing out HMPs,

especially when used without censoring.

All preprocessing choices have unintended effects on data; in light of the major effect that

preprocessing choices has on FC estimates, we urge awareness of the effect of preprocessing

choices within any given research protocol. Ideally, studies should aim to reduce head motion at

the time of scan, for instance through the use of passive viewing, and new preprocessing

strategies should aim to improve the signal-to-noise ratio with fewer tradeoffs.

Chapter 3: Functional Connectomes Become More Longitudinally Self-Stable, but Not More Distinct from Others, Across Early Childhood

3.0 Abstract

Functional connectomes, as measured with functional magnetic resonance imaging (fMRI), are highly individualized, and evidence suggests this individualization may increase across childhood. A connectome can become more individualized either by increasing self-stability or decreasing between-subject-similarity. Here we used a longitudinal early childhood dataset to investigate age associations with connectome self-stability, between-subject-similarity, and developmental individualization, defined as an individual's self-stability across a 12-month interval relative to their between-subject-similarity. fMRI data were collected during an 18-minute passive viewing scan from 73 typically developing children aged 4-7 years, at baseline and 12-month follow-up. We found that young children had highly individualized connectomes, with sufficient self-stability across 12-months for 98% identification accuracy. Linear models showed a significant relationship between age and developmental individualization across the whole brain and in most networks. This association appeared to be largely driven by an increase in self-stability with age, with only weak evidence for relationships between age and similarity across participants. Together our findings suggest that children's connectomes become more individualized across early childhood, and that this effect is driven by increasing self-stability rather than decreasing between-subject-similarity.

3.1 Introduction

Individuality is a fundamental aspect of the human experience. Functional connectomes, measures of inter-regional synchrony across the brain, are highly individualized in adults (Byrge and Kennedy, 2019; Finn et al., 2015; Vanderwal et al., 2021), and reflect individual cognitive and behavioural characteristics (Barch et al., 2013; Dadi et al., 2019; Mevel and Fransson, 2016). It has been suggested that individualization is a developmental process, where unique experiences may push the brain towards increasing individualization as children mature (Freund et al., 2013; Kaufmann et al., 2017; Vanderwal et al., 2021). However, findings on functional connectome individualization in development have been mixed (Kaufmann et al., 2017; Vanderwal et al., 2021). Here we used functional magnetic resonance imaging (fMRI) to examine whether functional connectomes become more individualized across early childhood, a period of rapid brain maturation.

Connectome individualization can be assessed as self-stability (self-similarity) – across two scans, contexts, or timepoints – relative to between-subject-similarity. For example, connectome fingerprinting (Finn et al., 2015) uses correlations to match connectomes from the same participant if the self-correlation exceeds correlation to others. Fingerprinting accuracy depends on factors such as scan duration, number of potential matches, and approach to data cleaning, but generally has high success in adult studies (90%-99%), suggesting that the adult connectome is highly individualized (Byrge and Kennedy, 2019; Finn et al., 2015; Miranda-Dominguez et al., 2014; Peña-Gómez et al., 2018; Vanderwal et al., 2017). Developmental, or longitudinal, individualization can be assessed by collecting pairs of scans from participants at different time points, and comparing self-stability over time to similarity to peers (Dufford et al., 2021; Jalbrzikowski et al., 2020; Sato et al., 2021). Across the first-year of life, longitudinal

identification accuracies are low, with under 25% success matching 1.5-month-old children to themselves at 9 months old, or vice versa (Dufford et al., 2021). In older children (6-15 years old), higher matching accuracies, up to 38%, have been shown across gaps of 20-45 months (Sato et al., 2021), and no significant difference in longitudinal (1.5 year) match rates were found in a study comparing separate groups of adolescents and adults (Jalbrzikowski et al., 2020). While studies have varied in design, and there is a gap in data across early childhood, collectively they suggest that connectomes show greater developmental individualization with increasing age.

Interpreting these findings is challenging, however, as it remains unclear whether age effects on connectome individualization are driven by increased self-stability, decreased between-subject-similarity, or both. Some evidence suggests that changes in developmental individualization are largely driven by changes in self-stability. Liao et al. (2021) found higher longitudinal self-stability in participants as adults compared to when the same individuals were adolescents. Kaufmann et al. (2017) found that age positively associated with short-term across-task self-stability in youth aged 8-23 years, though Vanderwal et al. (2021) found no association between connectome self-stability and age in a similarly designed study. With regards to between-subject-similarity, Vanderwal et al. (2021) also found no association with age, while both Vanderwal et al. (2021) and Liao et al. (2021) found that participants with higher self-stability also had greater, rather than reduced, between-subject-similarity. There also remains the question as to whether increasing self-stability might suggest a slowing down of developmental changes. If older children have more stable connectomes than younger children, the magnitude of change in functional edges may decrease in an asymptotic fashion.

In the present study, we used a longitudinal early childhood sample to separately consider developmental self-stability and between-subject-similarity. Specifically, we asked: 1) Do connectomes become more individualized over time as children get older? 2) Are age-related changes in developmental individualization driven more by changes in self-stability or (dis)similarity to others? And 3) Are changes in developmental individualization related to the magnitude of edge-wise connectome changes?

We further considered network-wise effects of developmental individualization. In adults, frontoparietal association networks are more individualized than sensorimotor networks (Byrge and Kennedy, 2019; Finn et al., 2015; Miranda-Dominguez et al., 2018; Peña-Gómez et al., 2018; Vanderwal et al., 2017), which is believed to reflect greater variance in functional neuroanatomy and evolutionary expansion in association regions (Marek and Dosenbach, 2018; Vendetti and Bunge, 2014). However, brain development is heterochronous, with evidence suggesting association cortices and particularly prefrontal regions follow more protracted development patterns than earlier maturing sensory and motor regions (Alcauter et al., 2015; Gao et al., 2015a; Geng et al., 2017; Gu et al., 2015), which may alter differences between networks in individualization, relative to adults. Early childhood is a period when sensory and sensorimotor functional connectivity and white matter pathways are continuing to mature (Dimond et al., 2020b; Rohr et al., 2018), suggesting individualization in these regions may associate with age, along with later maturing association networks.

3.2 Methods

3.2.1 Participants

173 children between 4 and 7 years of age were recruited as part of a study on early childhood brain maturation (Dimond et al., 2020b, 2020a; Graff et al., 2022a; Rohr et al., 2019, 2017; Tansey et al., 2022). Children were excluded if they had a history of neurodevelopmental or psychiatric disorders, any neurological diagnoses, any major health concerns, any MRI contraindications, or if their full-scale IQ was more than 2 standard deviations below 100 (i.e., <70). For this study, 142 children contributed usable baseline data, and of those 77 (31 male) returned for a follow-up 12 months later. Data were included in this analysis if both an initial and 12-month follow-up scan were available, and if after volume-wise censoring of the fMRI data at 0.25 mm framewise displacement (FD; based on FSL MCFLIRT; Smith et al., 2004) both scans retained at least 11 minutes of data. 73 participants (29 male) reached this threshold, for a total of 146 scans. Participants provided assent to participate, while parents provided informed consent. This study was approved by the University of Calgary Conjoint Health Research Ethics Board.

3.2.2 Data collection

Data was collected using a 3T GE MR750 w (Waukesha, WI) scanner with a 32-channel head coil, located at the Alberta Children's Hospital. A T1w 3D BRAVO sequence (TR = 6.764 ms, TE = 2.908 ms, FA = 10°, voxel size = 0.8x0.8x0.8 mm^3) and a gradient-echo EPI sequence (1100 seconds in duration, TR = 2.5 s, TE = 30 ms, FA = 70°, voxel size = 3.5x3.5x3.5 mm^3) were acquired. During fMRI acquisition, participants watched clips from Elmo's World, a children's television show. During the T1w scan, children watched a video of their choice.

Before each scan, children underwent a practice scan in an MRI simulator, during which they watched the same sequence of video clips from Elmo's World.

3.2.3 Preprocessing

We chose a preprocessing pipeline in accordance with our previous work benchmarking preprocessing pipelines in data from young children (Graff et al., 2022). All preprocessing was carried out using custom Python scripts that integrated Nipype functionality (Gorgolewski et al., 2011) using FSL version 6.0.0 (Smith et al., 2004), ANTs version 3.0.0.0 (Avants et al., 2011), and AFNI version 18.3.03 (Cox, 1996). We preprocessed scans from the same individual separately.

EPI data preprocessing consisted of slice time correction, rigid body realignment, and skull stripping using FSL slicetimer, FSL MCFLIRT, and FSL BET respectively, a linear regression to remove the mean and linear and quadratic trends from each voxel, bandpass temporal filtering (0.01 – 0.08 Hz), and nuisance regression. We regressed out the six head motion parameters, along with the white matter, cerebrospinal fluid, and global signal; both linear and quadratic terms were included, along with their temporal derivatives (4 regressors per parameter). Censoring was carried out as part of the regression step, where we censored volumes above a FD threshold of 0.25 mm (based on FSL MCFLIRT) and censored only the identified frames. After censoring, our shortest scan had 273 volumes; for all other scans volumes 274 onward (of volumes uncensored) were removed to match scan lengths. As not only amount but location of censored volumes influences connectome stability (Power et al., 2015), we carried out a control analysis using an alternate volume-length matching approach where the 18-minute scan was divided into 4.5-minute quartiles, and for each participant their 273 volumes were

chosen sequentially from each quadrant, from uncensored volumes. This allowed a more even

distribution of volumes included across the 18-minute scan. We used ANTs Registration (Avants

et al., 2011) to warp the EPI image to a study-specific EPI template (Huang et al., 2010). This

template was produced by taking a reference volume from each fMRI scan, chosen as a volume

of low motion approximately in the middle of the scan, warping these references to MNI space,

then averaging them together and smoothing with a Gaussian kernel. We smoothed the study-

specific template with a kernel of $\sigma = 3.0$ mm, as we found this kernel best addressed the trade-

off between retaining anatomical information and compensating for imperfect registration of the

reference volumes to MNI space. Functional data used in the analysis were not spatially

smoothed prior to connectome generation.

3.2.4 Connectome generation

Following preprocessing, each voxel was assigned to one of 1095 nodes (parcels) within

the MIST ATOM parcellation (Urchs et al., 2019). We used the MIST parcellation due to its

inclusion of subcortical regions, including the cerebellum. Each node was assigned to one of 12

brain networks, as defined by the MIST 12 parcellation: Basal ganglia and thalamus (for brevity,

hereafter referred to as BG/thalamus; 39 nodes), auditory network and posterior insula (auditory

network; 49 nodes), mesolimbic network (153 nodes), default mode network lateral (lateral

DMN; 48 nodes), visual network (110 nodes), default mode network posteromedial

(posteromedial DMN; 56 nodes), somatomotor network (53 nodes), default mode network

anteromedial and left angular gyrus (anteromedial DMN; 84 nodes), ventral visual stream and

dorsal visual stream (visual stream; 78 nodes), frontoparietal network (157 nodes), ventral

attention network and salience network (ventral attention network; 132 nodes), and cerebellum (136 nodes).

The time courses for voxels in each node were averaged, reducing the brain to 1095 time courses. The Pearson correlation between every pair of nodes was calculated, generating 598 965 edges, which were Fisher z-transformed to approximate a normal distribution.

3.2.5 Self-stability, between-subject-similarity, and individualization metrics

Figure 3.1 summarizes how metrics were generated. Following the approach used by Finn et al. (2015) and related work (Byrge and Kennedy, 2019; Kaufmann et al., 2017; Miranda-Domingeuz et al., 2018), connectomes were vectorized and then each pair of scans was correlated. These correlations were also Fisher z-transformed prior to being included in statistical models. Connectome metrics were then calculated as follows:

1. *Longitudinal self-stability* was assessed as the correlation between a participant's original and follow up scan.

2. *Average-between-subject-similarity* was assessed using the mean of each participant's correlation to all scans from other participants. Both original and follow-up scans were used to generate this average (i.e., for each participant, this value is an average of 2x144=288 correlations). We considered original and follow-up scans separately in a second model (see *2.8.* below).

3. *Developmental individualization* was assessed through two metrics:

    a. The fingerprinting match rate was calculated by comparing each scan against every other scan and dividing the number of times a scan's highest correlation was to the
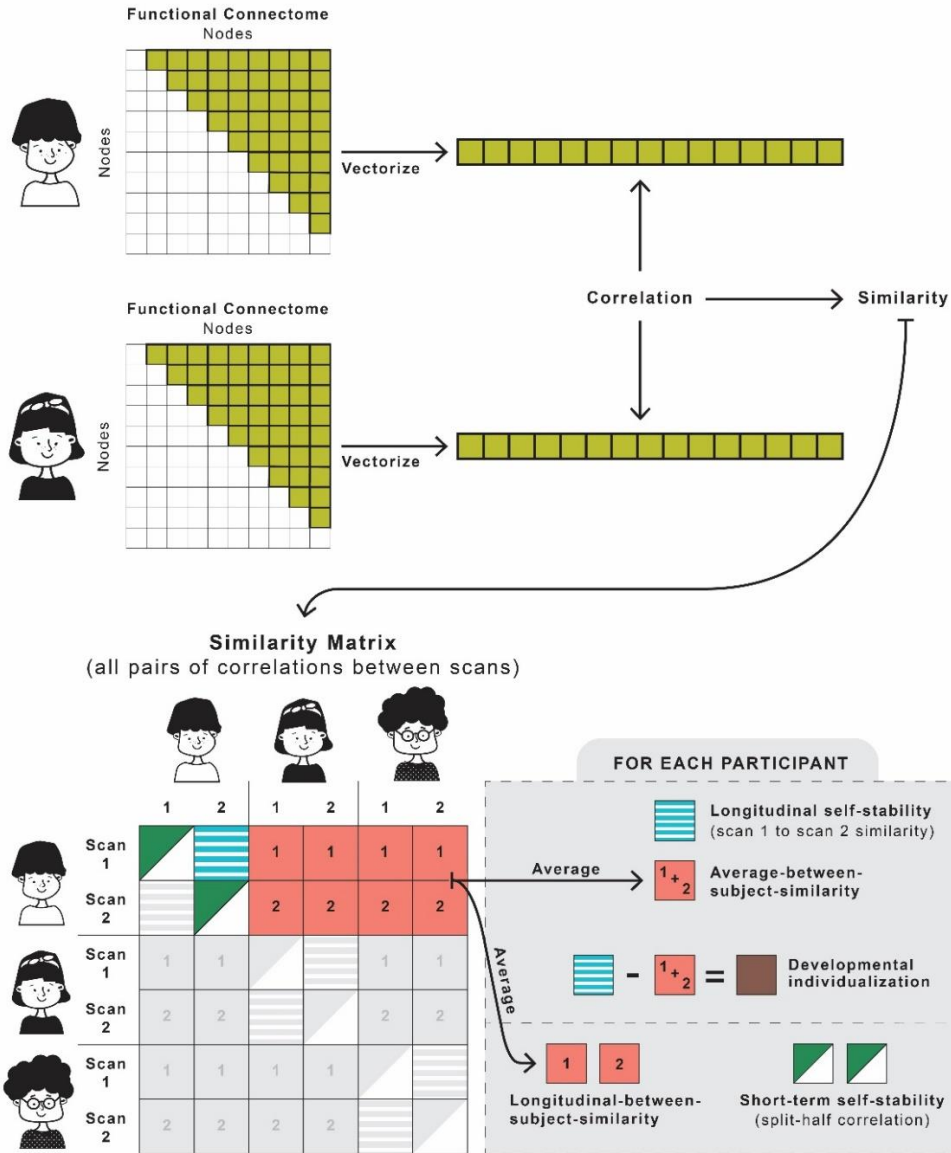
Figure 3.1. Summary of methods. The MIST ATOM parcellation was applied separately to both scans from each participant, then functional connectivity was calculated between all 1095 connectome nodes. The 598 965 edges were then converted into a vector. Between all pairs of vectorized connectomes – both within and between participants – the Pearson correlation was calculated to assess similarity. For each participant, metrics of interest were generated from these similarity values. Longitudinal self-stability was assessed as the correlation between a participant's original and follow up scan (in figure, blue striped box). Short-term self-stability was assessed by dividing scans into halves, generating the functional connectome for the halves separately, then calculating the similarity between these halves (split green-white boxes). Between-subject-similarity was assessed using the mean of each participant's correlation to scans from other participants (pink boxes), either using all correlations from both scan 1 and scan 2 (average-between-subject-similarity) or determining scan 1 and scan 2 mean correlations separately (sessional-between-subject-similarity). Developmental individualization was assessed as the difference between longitudinal self-stability and average-between-subject-similarity (brown box).

other scan from the same participant by the total number of scans, expressed as a

percentage. Here we had 73 individuals, each with two scans, for a denominator of

146. Although this metric has been used widely (Finn et al., 2015), it reduces each

scan to a binary score (match or no match), thus limiting its sensitivity, and is

necessarily affected by sample size, since the larger the sample the more likely a scan

will be compared to a very similar scan (Waller et al., 2017). To address this, we

implemented an additional metric:

b. The difference between longitudinal self-stability and average-between-subject-

similarity. This metric can be thought of as the extent to which a participant is

individualized relative to the group. We used this as our main developmental

individualization metric.

Self-stability, average-between-subject-similarity, and developmental individualization

were calculated for the whole connectome (all 598 965 edges) and for the edges within

individual networks. To account for differences in network size (number of nodes in each

network), we conducted network size-matched control analyses by randomly choosing 36 nodes

(chosen as it was close to but smaller than the 39 nodes in our smallest network) from each

network and only using the (630) edges between those nodes, sampled 1000 times.

3.2.6 Cross-sectional age associations

We used a linear regression to assess associations between age and self-stability, average-

between-subject-similarity, and developmental individualization. For age, we used the

participants' average of the age at the time of their initial scan and age at their follow up scan.

We also included motion (mean relative FD, averaged across the two scans), time between scans,

and sex as covariates. Age associations were calculated for the entire connectome and individual

networks. As described above, we also conducted analyses controlling for network size. For

network-wise analyses, we used Bonferroni correction for multiple comparisons across networks,

and present levels of significance both with and without the correction. Since different networks

may have different distributions in the metric of interest (independently of age), we calculated

standardized beta coefficients. Figures also include 95% confidence intervals around beta

estimates.

3.2.7 Longitudinal-between-subject- similarity analysis

To assess changes in similarity with age, in addition to using average-between-subject-

similarity in a cross-sectional model, which parallels the individualization and self-stability

models (i.e., with one value per participant), we tested a second model. For each participant, we

calculated mean between-subject-similarity for both baseline and follow-up scan separately

(giving two similarity values for each participant, rather than – as in section *2.7* – one similarity

value per participant). We then used a mixed-effects model to test the effect of age on similarity,

including motion and sex as fixed effects and participant identity as a random effect.

3.2.8 Short-term split-half self-stability

To facilitate comparison to prior work (Vanderwal et al., 2021; Kaufmann et al., 2017),

we examined associations between age and short-term self-stability using mixed-effects models.

Short-term self-stability was assessed using split-half correlations: for each scan we calculated

the functional connectome for the first and second half of the scan separately, then correlated the

vectorized edges and Fisher z transformed the result. Models included motion (mean relative FD) and sex as fixed effects and participant identity as a random effect.

To address whether short-term self-stability age effects were driven by temporal autocorrelation between the scan halves, we carried out an additional analysis where we calculated stability between the first third and last third of each scan, discarding the middle third. As before, the functional connectome was calculated for these thirds separately before carrying out the same mixed effects analysis.

3.2.9 Magnitude of longitudinal change

If longitudinal stability increases with age, this may be related to the average change in edge strength (functional connectivity; FC) decreasing, i.e., a slowing of developmental changes. To assess this, we calculated the magnitude of longitudinal change by averaging across edges the absolute edge differences for each participant's scan pairs, both for the whole brain and across networks. We then used a multiple regression to assess the effect of age on magnitude of longitudinal change controlling for motion, time between scans, and sex, identically to other age models.

3.2.10 Within- and between-network effects

In an exploratory analysis, we considered between-network connections for all pairs of networks. We controlled for network size by randomly choosing 18 nodes from each of the two networks and only using the (18x18=324) edges between those nodes, averaged over 1000 random samplings. In this analysis, when assessing within-network we randomly chose two sets of 18 nodes, giving an identical 18x18=324 edges. Reported p-values are derived from using the

entire set of between-network edges, rather than a random subset. While we report p-values with and without multiple comparison correction for transparency, given the large number of network-network pairs we caution against over interpreting significance or lack of significance.

3. Results

3.3.1 Participant characteristics

Figure 3.2 shows the age distribution of participants. The mean age at baseline was 5.54 years old (standard deviation: 0.78 years). Time between scans ranged from 0.87 to 1.33 years (mean = 1.05 years, SD = 0.088 years); mean age at follow-up scan was 6.60 years. Our sample's average motion was high, with a mean average relative FD of 0.127 mm across all scans (median = 0.097 mm). The correlation in average relative FD between scans from the same participant was $r = 0.25$ ($p = 0.035$). Using a linear mixed model, age showed a trend-level negative association with mean FD ($p = 0.081$; Supplemental Figure 3.1, Appendix A), with no significant association between age and median FD ($p = 0.22$) or between age and number of censored volumes ($p = 0.24$). On average, more volumes were censored in the last quartile of the scan than
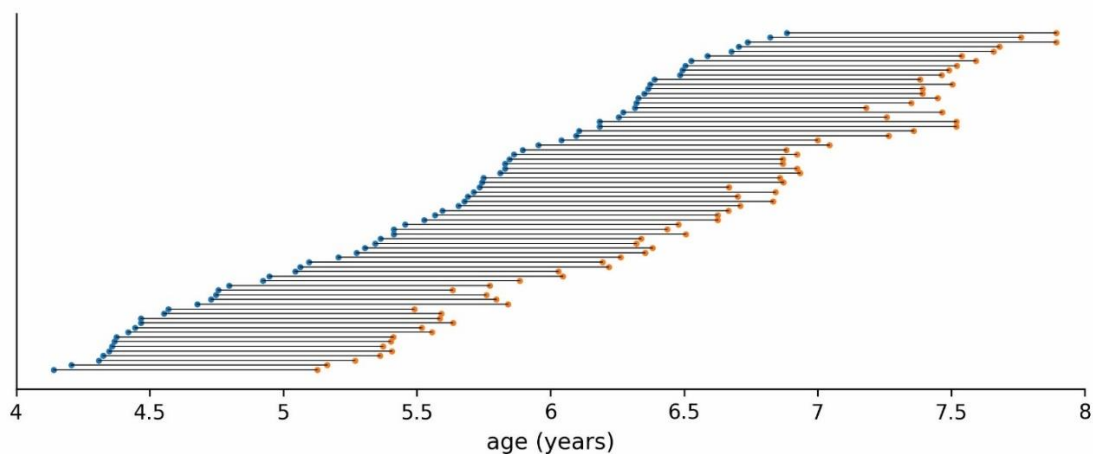
Figure 3.2. Age at scan for all participants. Each dot represents a scan; lines connect scans from the same participant.

other quartiles, but no quartile showed a significant association between age and number of censored volumes (p = 0.97, 0.071, 0.54, 0.18 for the four quartiles respectively: Supplemental Table 3.1).

3.3.2 Connectome individualization

We found a high level of developmental individualization in early childhood with a 98% whole brain fingerprinting match-rate (Supplemental Figure 3.2a, Appendix A). Individual networks varied in individualization, with the frontoparietal network having the highest match rate at 98%, and the BG/thalamus having the lowest at 16%. The posteromedial DMN, visual, visual stream, and ventral attention all had match rates above 80%. When controlling for network size, trends between networks were nearly the same, and the whole brain match rate approached the average across networks (Supplemental Figure 3.2b, Appendix A). All match rates were significantly above chance (e.g., for 16%, the lowest average match rate, $p < 10^{-25}$, based on 24 or more matches out of 146, with each match having a 1/145 chance).

Figure 3.3a shows each participant's mean developmental individualization. The posteromedial DMN showed the highest mean individualization, while the mesolimbic network showed the lowest. The whole brain, which includes internetwork connections, had a comparably low mean individualization. A repeated measures ANOVA showed a significant difference between networks (p = 4.8 x 10^{-82}, $\eta_p^2 = 0.38$). Figures 3.3b and 3.3c show each participant's longitudinal self-stability and average-between-subject-similarity respectively. Differences between networks were more pronounced for these two measures; for self-stability we found an effect size of $\eta_p^2 = 0.68$ and for average-between-subject-similarity we found an effect size of $\eta_p^2 = 0.91$. In general, networks with higher self-stabilities also had higher similarities. The
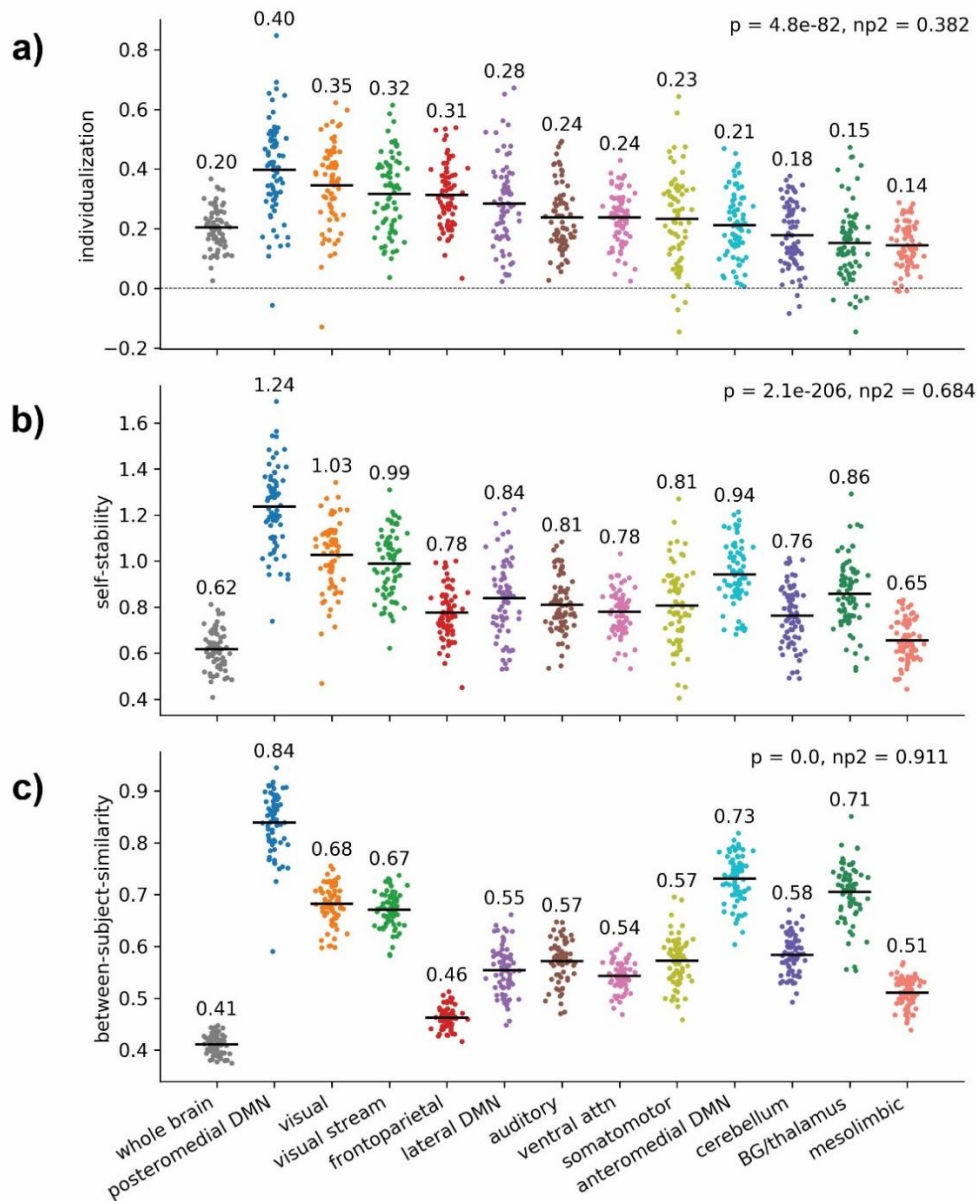
Figure 3.3. Developmental individualization, longitudinal self-stability, and average-between-subject-similarity across networks. Each dot represents one participant. Lines represent mean values across participants. **a)** Developmental individualization across networks, calculated by taking the difference between each participant's self-stability across one year and their average-between-subject-similarity. **b)** Longitudinal self-stability across networks, calculated by correlating FC estimates from the same participant, with scans collected one year apart. Values are Fisher-z transformed. **c)** Average-between-subject-similarity across networks. For each participant, average-between-subject-similarity was calculated as the mean correlation to scans from all other participants, using both their original and follow up scan. Values were Fisher-z transformed prior to averaging. For all three subplots, to test for differences across networks a repeated measures ANOVA was used to determine p and partial-eta squared values; p and np2 values indicated on plots.

posteromedial DMN had both the highest average self-stability ($z = 1.24$) and highest average-between-subject-similarity ($z = 0.84$). Alternatively, the frontoparietal network had among the lowest self-stabilities ($z = 0.78$) and the lowest average-between-subject-similarity between participants ($z = 0.46$), suggesting high individualization can be achieved by either high self-stability or low similarity to others. The whole brain had lower self-stability and similarity than any individual network, which we attribute to its inclusion of both inter- and intra-network connections. No noticeable trend existed between network size and individualization. Small networks showed both high (e.g., posteromedial DMN – 56 nodes) and low (e.g., BG/thalamus – 39 nodes) individualization. When controlling for network size, the relative ranking of each networks' individualization, self-stability, and average-between-subject-similarity scores were essentially unchanged (Supplemental Figure 3.3, Appendix A). In general, larger networks (more nodes) showed larger variance, though this is at least partially an artifact of resampling, as smaller networks have a higher probability of resampling the same nodes. Despite this, the anteromedial DMN, an average-sized network (84 nodes), showed the largest range of self-stability and average-between-subject-similarity scores, suggesting it is a relatively heterogeneous network. When we assessed developmental individualization, longitudinal self-stability, and average-between-subject-similarity using the alternate volume-length matching approach (described in section *2.4.*), measures were mostly unaffected (e.g., the largest mean change was $z = 0.03$) suggesting minimal impact of volume selection decisions (Supplemental Figure 3.4a,c,e, Appendix A).

3.3.3 Effects of age on individualization, self-stability, and average-between-subject-similarity

Figure 3.4a shows associations between developmental individualization and age across networks. We found uncorrected significant associations ($p<0.05$) with age in the whole brain and all networks except auditory, anteromedial DMN, and BG/thalamus. Whole brain, frontoparietal, visual stream, ventral attention, mesolimbic, and cerebellum survived correction for multiple comparisons. The largest standardized effect size was found in the whole brain ($b = 0.47$), followed by the ventral attention network ($b = 0.43$), and visual stream ($b = 0.39$). We note that the 95% confidence intervals around age effect estimates overlapped for most networks, suggesting caution in interpreting age differences between networks.

Age associations with longitudinal self-stability (Figure 3.4b) were nearly identical to individualization (Figure 3.4a; Spearman correlation between age-stability betas and age-individualization betas was $r = 0.99$, $p < 10^{-9}$). On the other hand, no network showed a significant age association for average-between-subject-similarity (Figure 3.4c), although we found a positive association between age and average-between-subject-similarity across the whole brain ($b = 0.35$; as age increased, the similarity to others increased). This suggests that the age effect of developmental individualization is driven by age-related increases in self-stability, rather than a decrease in similarity to others. Although similarity effects were mostly non-significant, age-similarity betas were highly correlated with both age-individualization betas (Spearman's correlation $r = 0.71$, $p = 6.7 \times 10^{-3}$), and age-stability betas (Spearman's correlation $r = 0.79$, $p = 1.5 \times 10^{-3}$).
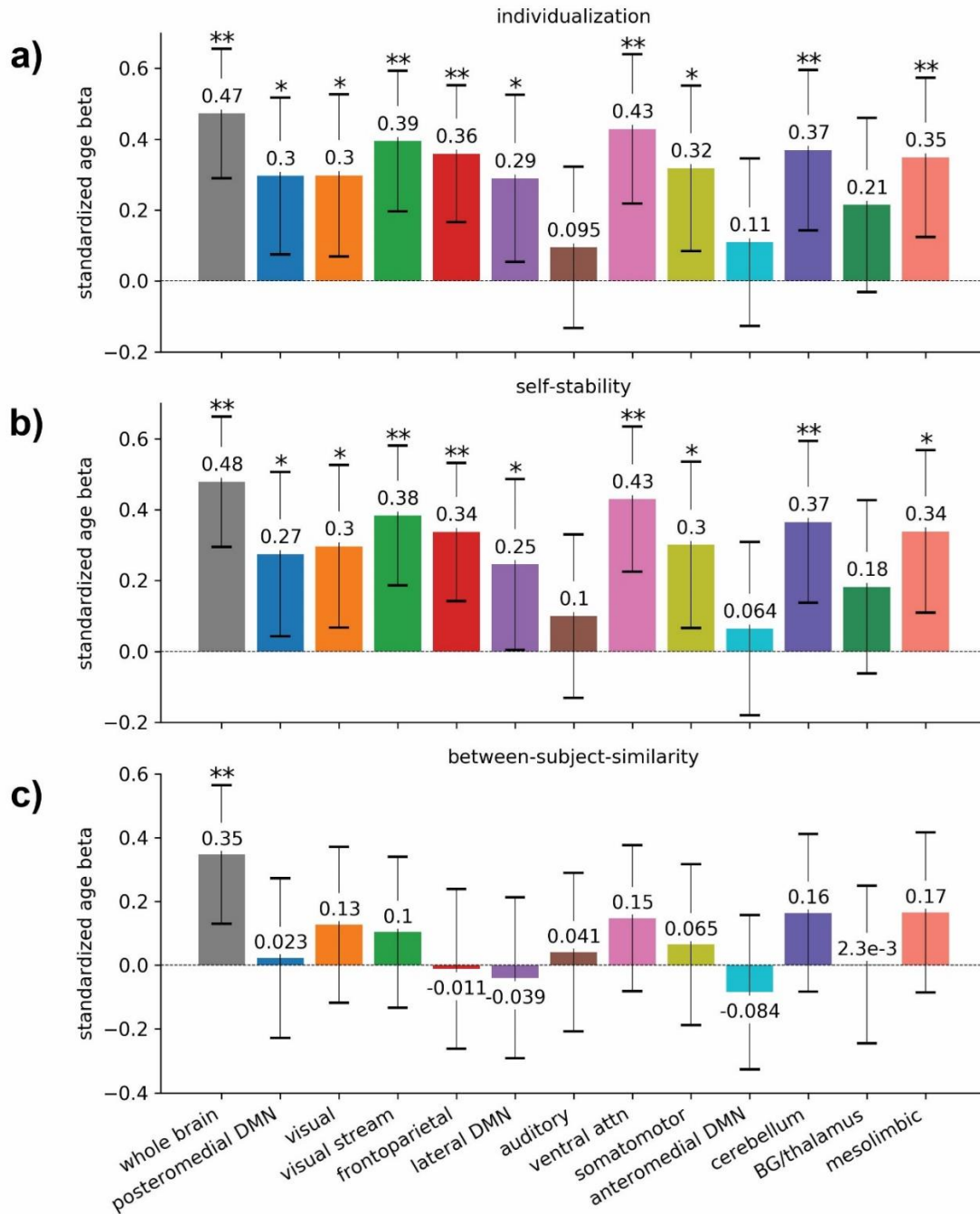
Figure 3.4. Age effects of (a) developmental individualization, (b) longitudinal self-stability, and (c) average-between-subject-similarity by network. A linear regression was used to assess associations between age and the metric of interest, controlling for sex, motion (mean FD), and time between a participant's scans. Error bars depict 95% confidence intervals around standardized beta estimates. * = $p < 0.05$ uncorrected; ** = $p < 0.05$ Bonferroni corrected.

Controlling for network size, we found comparable results (Supplemental Figure 3.5, Appendix A), with the whole brain and the visual stream showing the largest effects for developmental individualization (average b = 0.36 for both), followed by the ventral attention network (average b = 0.35). Individualization and self-stability effect sizes closely matched. As before, the auditory network, anteromedial DMN, and BG/thalamus had the smallest average effect sizes. For each network, mean age effects for average-between-subject-similarity were negligibly small, as with the full network analysis, although the whole brain continued to show a larger, positive effect. When we assessed metrics spreading volumes across quartiles (described in section *2.4.*), age effects were comparable to the main model (Supplemental Figure 3.4b,d,e, Appendix A). The most prominent change was in the mesolimbic network's age effect on developmental individualization, which remained significant at an uncorrected threshold but no longer survived multiple comparisons correction.

In the longitudinal-between-subject-similarity analysis, we calculated mean between-subject similarity for each scan (giving two values per participant – one for their baseline scan, one for their follow up – rather than one overall average), then used a mixed-effects model to assess the effect of age (Supplemental Figure 3.6, Appendix A). Compared to the main, average-between-subject-similarity, model, the whole brain showed a smaller, positive association that approached significance (b = 0.16, p = 0.055). In the whole brain, after controlling for sex and motion, we found that 35 participants had a lower mean between-subject-similarity in their follow up scan, while 38 participants had a higher mean between-subject-similarity in their follow up scan. As previously, no network showed a significant association between age and similarity.

3.3.4 Effects of age on short-term self-stability

For comparison with previous literature, we assessed whether age associated with short term self-stability (i.e., the split-half correlation, rather than correlating separate scans 12-months apart; Supplemental Figure 3.7a, Appendix A). Significant but small effects ($0.19 < b < 0.23$) were found in the whole brain, visual stream, frontoparietal network, and ventral attention network, but did not survive multiple comparisons correction. Most networks showed an effect size of $b = 0.1$ or smaller, suggesting that age has a smaller effect on short-term stability than 12-month longitudinal stability, though we note that split-half analyses use less data and thus provide noisier estimates of connectivity.

In order to minimize temporal autocorrelation effects, we carried out an additional analysis where we assessed short-term self-stability using a "split-third" analysis, i.e., the correlation between the first third and last third of each scan (Supplemental Figure 3.7b, Appendix A). Results were comparable to the split-half analysis, though the effect size was moderately larger in the auditory and frontoparietal networks, the latter of which now survived multiple comparisons correction.

3.3.5 Relationship between age and longitudinal change

As we found that connectomes become more longitudinally self-stable with increasing age, we asked whether there was less overall longitudinal change between scans in older participants. Figure 3.5 shows the association between age and average absolute longitudinal FC change. Throughout the brain, as age increased average absolute longitudinal change in FC decreased. Significant effects were found in the whole brain, visual network, visual stream, cerebellum, and mesolimbic network, though only the whole brain and visual stream survived a
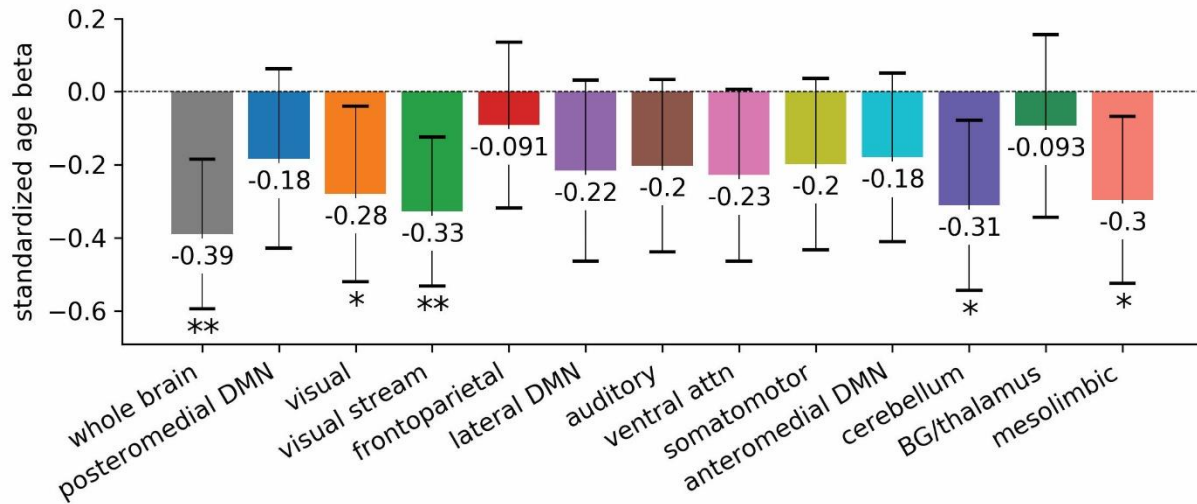
Figure 3.5. Age effects of average change in edge strength. A linear regression was used to assess associations between age and average change in edge strength, controlling for sex, motion (mean FD), and time between scans. For each participant, average change in edge strength was calculated by taking the absolute difference in FC estimates across the participant's two scans, collected one year apart. These differences were then averaged across all edges in the network. Error bars depict 95% confidence intervals around standardized beta estimates. * = $p < 0.05$ uncorrected; ** = $p < 0.05$ Bonferroni corrected.

Multiple comparisons correction. The effect was largest in the whole brain (b = -0.39). The 95% confidence intervals overlapped across most networks, suggesting small differences between networks. Comparing this analysis to the effect of age on longitudinal self-stability (Figure 3.4b) showed broadly similar trends, as all the networks with significant age effects on longitudinal change also showed age effects on stability (Spearman correlation between self-stability age betas and longitudinal change betas: r = -0.69, p = 8.7 x 10^{-3}).

3.3.6 Within-network and between-network effects

Figure 3.6 shows individualization, longitudinal self-stability, and average-between-subject-similarity metrics for both within-network and between-network edges, matched for network size. In general, for most networks developmental individualization was higher within-

network than between-networks, and networks with high within-network individualization tended to also have relatively high between-network individualization (Figure 3.6a). For example, the posteromedial DMN had an average individualization of 0.40, versus 0.34 for the visual network. Posteromedial DMN-to-frontoparietal had an average individualization of 0.31, versus 0.26 for visual-to-frontoparietal. Self-stability (Figure 3.6c) and average-between-subject-similarity (Figure 3.6e) showed comparable effects to each other, where sets of edges with higher self-stability also showed higher similarity. Here we found a clearer gap between intra-network and inter-network connections, with intra-network edges almost universally showing higher average-between-subject-similarity and self-stability than inter-network connections. Fingerprinting match rates (Supplemental Figure 3.2c, Appendix A) were similar to developmental individualization, where networks with high match rates tended to also have internetwork edges with high match rates.

Age associations for individualization, longitudinal self-stability, and average-between-subject-similarity are shown in Figures 3.6b, 3.6d, and 3.6f respectively. As for within-network edges, internetwork edges tended to have comparable age effects for individualization and self-stability, while age effects for average-between-subject-similarity were comparatively small. Our longitudinal-between-subject-similarity model also showed small effect sizes for age (Supplemental Figure 3.6b, Appendix A). Networks with large age effects tended to also have internetwork connections with large age effects. For example, the visual stream showed larger age effects for self-stability than the visual network, and visual stream internetwork connections showed larger age effects than analogous visual internetwork connections.
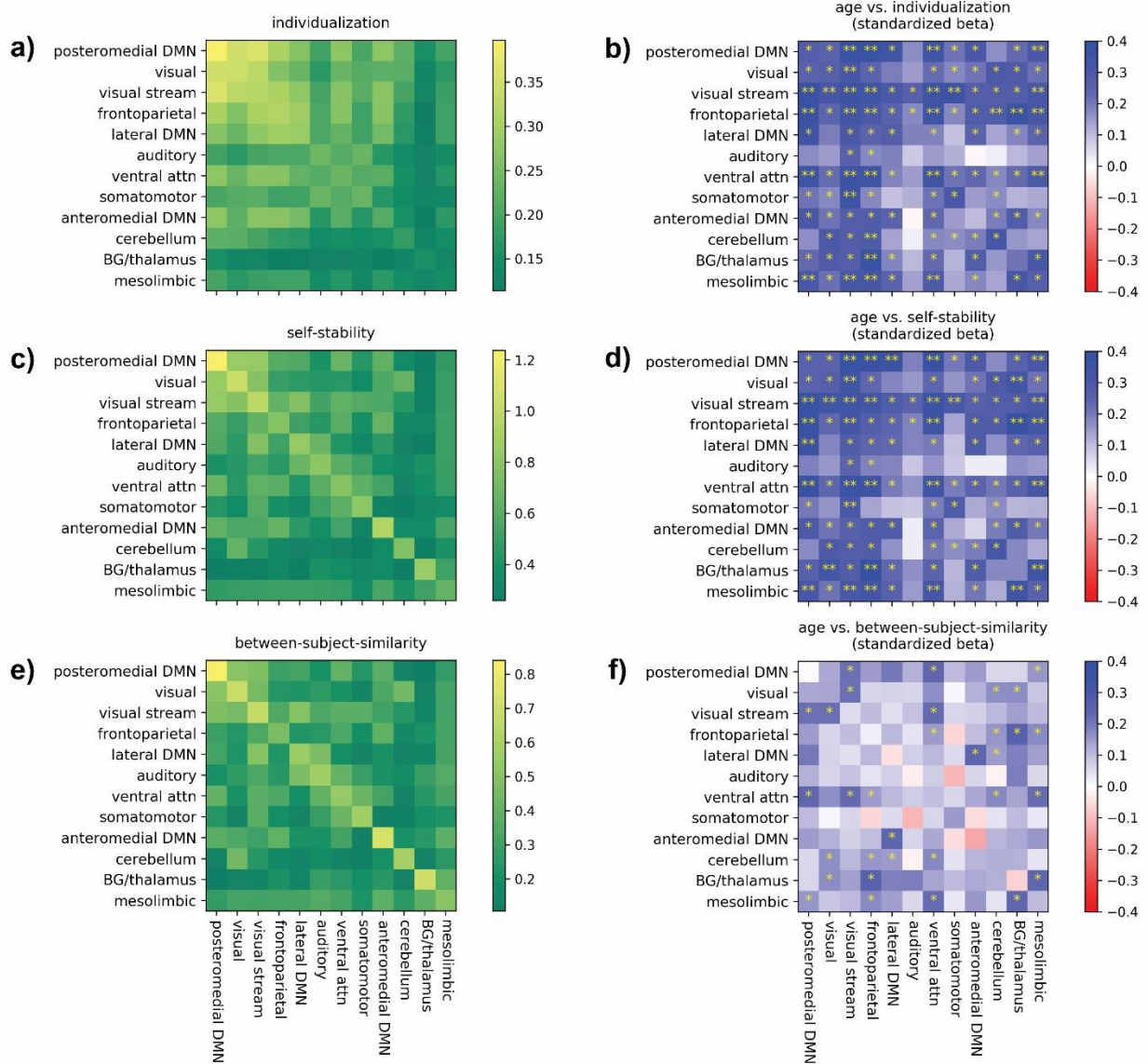
Figure 3.6. Developmental individualization, longitudinal self-stability, and average-between-subject-similarity across networks and their internetwork connections. a, c, e) Each network or internetwork's mean developmental individualization, longitudinal self-stability, and average-between-subject-similarity, averaged across all participants. The network effects, shown along the diagonals, are the same as averages shown in Figure 3.1. **b, d, f)** Age associations. Network size was controlled for by randomly choosing 18 nodes from each of the two networks and only using the (18x18=324) edges between those nodes. To compare the same number of edges, networks were similarly derived using 36 random nodes divided into two sets. Beta values are standardized and are averaged over 1000 random samplings. P-values are derived from using the entire set of edges, rather than a random subset. * = p < 0.05 uncorrected; ** = p < 0.05 Bonferroni corrected.

3.4 Discussion

In this study, we investigated age associations with developmental individualization. We found that most functional networks showed greater developmental individualization with increasing age and that there was a clear parallel with increases in longitudinal self-stability. Further, using two different models we found comparatively minor age associations with similarity to others. Effects persisted when only using random subsets of network nodes or when using an alternate approach to matching scan length between participants. Taken together, these results suggest that across early childhood the functional connectome shows increasing developmental individualization that is driven primarily by increasing self-stability over time rather than a decrease in similarity to others.

To better understand increasing connectome self-stability, we assessed age effects on longitudinal change in edge strength. Age associations for longitudinal change closely mirrored age associations for self-stability, where networks with large positive age effects on self-stability had large negative age effects on longitudinal change, with comparable beta values. For both measures, the largest effect was in the whole brain. This implies that increasing self-stability with age is at least partly due to less absolute change in FC over time, with younger children undergoing more change across a year of development than older children. Further, this suggests that developmental individualization may be sensitive to asymptotic developmental changes in FC. However, not all networks with increasing longitudinal self-stability showed significant age effects for longitudinal change, most notably the frontoparietal network. While part of increasing self-stability with age can be explained by less change in edge strength, more complex relationships may also be at play. For example, already strong edges may strengthen while

weaker edges either change little or weaken. Future work can more closely analyse how individual edges change with age to see what factors explain overall increases in self-stability.

Altogether, our work suggests age only has a minor impact on between-subject-similarity. No network showed a significant relationship between age and between-subject-similarity, both in a cross-sectional model (with one, averaged similarity value per participant) and in a mixed-effects model (with both a baseline and a follow-up mean similarity value for each participant). However, when considering the whole connectome, we found a positive age effect for average-between-subject-similarity to others that persists when controlling for number of nodes. As the whole brain also showed the largest effect between age and individualization, this suggests that increases in self-stability outweigh increases in between-subject-similarity, allowing for an overall increase in individualization. This finding is consistent with previous work that found higher self-stability is associated with higher between-subject-similarity (Vanderwal et al., 2021; Liao et al., 2021). Increasing whole brain between-subject-similarity with age is likely influenced by internetwork effects, for example, internetwork connections associated with the cerebellum, mesolimbic network, and ventral attention network, which our exploratory findings suggest become more similar across children with increasing age (i.e., with effect sizes up to $b = 0.26$). Future work could further explore the factors that determine how similar individuals are to each other, even if age plays a relatively minor role overall, especially within networks.

In adults, higher order networks – such as the frontoparietal and default mode networks – have previously been found to be more individualized than primary sensory networks (Finn et al., 2015; Peña-Gómez et al., 2018; Vanderwal et al., 2017). Our work generally agrees with these results. We found higher individualization scores and match rates in the posteromedial

DMN, lateral DMN, and frontoparietal network compared to the somatomotor, cerebellum, BG/thalamus, and mesolimbic networks. The major exception, relative to previous work, is that we also found high individualization scores and match rates in the visual network and visual stream. While one explanation is that using a passive viewing task emphasizes visual self-stability, previous work in adults (Vanderwal et al., 2017) and older children (Vanderwal et al., 2021) suggests instead that passive viewing conditions, relative to resting state, decreases identifiability in visual networks. We propose that early childhood is a period where visual networks are both sufficiently developed and longitudinally self-stable to allow for comparatively high developmental individualization; however, disentangling task from its effects is not possible with our design.

Network differences in age effects on individualization were relatively minor with a high degree of overlap between networks in 95% confidence intervals around estimates and only the auditory network, anteromedial DMN and BG/thalamus showing non-significant age effects. We note that the degree of individualization and age effects on individualization were not well aligned. We found highly individualized networks with large age effects (e.g., visual stream), highly individualized networks with small age effects (visual network), weakly individualized networks with large age effects (cerebellum), and weakly individualized networks with small age effects (BG/thalamus). This heterogeneity persists when comparing the degree of self-stability and age effects on self-stability, since, in general, more self-stable networks were also more individualized. Inter-network connections were likewise variable, with both strongly and weakly individualized edges having large age effects. Together, this suggests that developmental individualization occurs in a heterogeneous fashion with different brain networks becoming more individualized at different rates, but across early childhood there is an overall trend of

across-network increases in longitudinal individualization. While previous research suggests that sensorimotor networks develop earlier than higher order networks (Alcauter et al., 2015; Gao et al., 2015a, 2015b; Gu et al., 2015), our work suggests they continue to undergo refinement in early childhood along with higher order networks. This is perhaps unsurprising as white matter findings from an overlapping set of participants show pronounced changes in sensorimotor tracts in early childhood (Dimond et al., 2020b).

Previous studies using children have found fingerprinting match rates much lower than in studies using adults, whether matching participants longitudinally (Dufford et al., 2021; Sato et al., 2021) or across short time scales (Vanderwal et al., 2021; Kaufmann et al., 2017). Across studies, there is a trend towards increasing individualization with age, with very low match rates in infants, higher match rates in older childhood, and adult-like match rates in adolescents (Dufford et al., 2021; Sato et al., 2021; Jalbrzikowski et al., 2020). Contrary to these studies, we found near perfect (98%) match rates in the whole brain and in the frontoparietal network, with high (> 50%) match rates in most networks, even when only using 36 random nodes from a given network. Factors that may influence differences in findings include a longer scan time and effective preprocessing that included censoring volumes of high motion. Longer scans have been found to increase test-retest reliability (Noble et al., 2017) and connectome stability (Gordon et al., 2017), and Finn et al. (2015) found that fingerprinting match rate was dependent on scan length. Our comparatively long passive viewing scan, using 11 minutes of post-censoring data per scan, likely yielded a more accurate estimate of the functional connectome, which contributed to a higher match rate. Additionally, our previous work (Graff et al., 2022a) suggests that match rates are affected by choice of preprocessing pipeline. Studies with youth also face the challenge of using parcellations that are typically derived with adult data. This may artificially

lower metrics of individualization, either due to challenges with registration, or by averaging functionally distinct voxels into the same node. Infant data, as in Dufford et al. (2021), has the further challenge of registration due to inverted T1w contrast. This highlights some of the challenges in interpreting individualization findings across studies, especially between adults and youth, particularly as data quality and head motion tend to be worse in younger samples.

In terms of short-term self-stability, as measured with split-half or split-third correlations, we found a significant relationship with age in only the whole brain and 4 of 12 networks, with only the frontoparietal network in the split-third analysis surviving multiple comparisons correction. The effect sizes were smaller than those for longitudinal self-stability, suggesting age associations with longitudinal self-stability are not primarily due to increasing short-term self-stability. Short-term self-stability findings are also interesting in the context of previous work. For example, Kaufmann et al. (2017), in a study of 8-22-year-olds, found that older participants were more highly individualized, when self-stability was assessed across tasks (rest, working memory, and emotion recognition) from the same scan session. In contrast, Vanderwal et al. (2021), considering 6-21-year-olds, found no association between age and self-stability, when self-stability was again assessed across tasks (two rest and two passive viewing tasks) within the same scan session. While due to differences in tasks and age range, it is difficult to directly compare effects reported here with these prior studies, conceptually our short-term self-stability findings align with Kaufmann et al. (2017) albeit with weaker effects. It is important to note that study design and data quality can have a strong impact on findings and split-half analyses necessarily estimate functional connectomes with less data, in this case <6 minutes of post-censoring data, which reduces the reliability of FC estimates (Gordon et al., 2017) and could thereby decrease effect sizes for age associations. Thus, while we only found evidence for

limited age-related changes in short-term self-stability, this may have been due to limitations in our study design.

Our finding of increasing self-stability but more minor changes in similarity to others has implications for understanding how the connectome develops across early childhood. One explanation is that younger children experience more FC change than older children, with divergent (individualized) changes between individuals partially balanced with convergent changes that may reflect typical development trajectories. Another (not mutually exclusive) explanation comes from Kaufmann et al. (2017); they suggest a network tuning effect, in which as aging occurs edges become 'constrained', showing less within-individual variability in their strength while the difference in average strength between individuals may remain constant. While Kaufmann et al. (2017) proposed this across tasks, our findings could support this model in a more generalizable way. Children may exhibit a broader collection of connectome patterns, which becomes more limited with age, thus increasing self-stability, with the average difference between children remaining relatively constant. However, some contrary evidence suggests that older children utilize a larger number of connectome states within a resting state scan (Hutchison and Morton, 2015), though the same study found more complicated dynamic patterns when the brain was engaged in a task. One possible way to shed light on how children engage in different connectome patterns and how this may change with age is with a precision imaging approach, using extended data acquisition of individual subjects (Gratton et al., 2020). This could more thoroughly investigate FC self-stability and between-subject-similarity across a variety of time scales, providing a clearer understanding of how individualization develops.

There are several limitations to our work. We measured the functional connectome during a passive viewing task, and we cannot directly extend our findings to other task

conditions, including rest. While resting state scans are often considered the standard for investigating FC, the pragmatic challenge of getting young children – especially under 6 years old (Alexander et al., 2017) – to remain sufficiently still and calm in a scanner makes this data impractical to collect. This challenge is amplified when collecting longer scans with reduced motion, as is important for reliable estimates of FC (Gordon et al., 2017; Noble et al., 2017; Power et al., 2012). However, there is reason to suggest our findings may extend to resting conditions. Similar network structure is detected in passive viewing paradigms as in rest conditions, suggesting an underlying organization that is well captured by either task (Bray et al., 2015; Vanderwal et al., 2019). Our findings also may not directly line up with other studies due to differences in parcellations. Previous work has found higher match rates in parcellations with more nodes (Finn et al., 2015; Vanderwal et al., 2021), suggesting parcellation choice affects measured identifiability. Additionally, while we used a volume-based approach to parcellation, similar to much of the individualization literature (Finn et al., 2015; Kaufmann et al., 2017; Jalbrzikowski et al., 2020; Vanderwal et al., 2021), accuracy could potentially improve using a cortical surface-based approach, as volume approaches often have poor spatial localization (Coalson et al., 2018).

We note that our measures of individualization can also be thought of as measures of reliability (Bridgeford et al., 2021; Graff et al., 2022a; Noble et al., 2019), as akin to a metric like interclass correlation we are assessing self-similarity relative to similarity-to-others. While we interpreted findings here to represent developmental changes in the connectome, other explanations are possible. FC estimates are biased by motion (Power et al., 2012), and younger children move more than older children (Greene et al., 2018; Dosenbach et al., 2017). While we found no significant relationship between age and median FD or number of censored volumes,

the relationship between age and mean FD approached significance (p = 0.081). Thus, there is the concern here – or in any other connectome developmental study – whether results could be alternatively explained by differences in motion. While this concern cannot be entirely dismissed, we mitigated motion-related confounds by matching scan lengths post-censoring and including motion as a regressor in our models. The effect size of age on FD was much smaller than age on individualization for most networks, suggesting even if motion could partially explain our findings, it could not entirely explain how age associates with individualization. Similarly, we note that differences between networks in self-stability may be related to differences in how physiological or signal-drop out artifacts may impact BOLD measurements and connectivity metrics. Moreover, older children have slower respiratory and heart rates than their younger peers (Fleming et al., 2011), which could affect age associations. As our study design did not include the measurement of physiological features, this potential confound is hard to address, although physiological noise is captured in the global signal, which we regressed from each voxel's time course. Finally, the present study was specific to early childhood, and findings may not extend to either earlier or later periods of development. Future studies that probe a wider age range of children will be able to assess in more detail when and to what total extent developmental individualization occurs in individual networks, including visual networks.

3.5 Conclusions

In sum, the present study investigated how age was associated with developmental individualization in early childhood. We found that older children have higher individualization, and this was likely due to higher longitudinal self-stability, rather than decreased between-subject similarity. These trends were observed in most networks of the brain and across the entire

94

functional connectome, and our findings suggest that increasing self-stability with age is partially

due to smaller changes in FC edge strength in older children.

Chapter 4: Reliability and Validity of the Electrophysiological Connectome Across Phase-Based Connectivity Measures

4.0 Abstract

EEG connectomics research offers the potential of better understanding human neurodevelopment, disease, and brain-behavior associations. Phase-based functional connectivity (FC) measures are widely used but their reliability and validity are poorly understood. Many of these measures mitigate volume conduction artifacts, but may miss true interactions in the process, which could hamper validity overall. Here, we collected a densely sampled dataset from 25 parent-child pairs, with 80 minutes of passive viewing EEG data collected per participant over 4 sessions. We compared reliability across 6 FC measures by testing the self-stability of the functional connectome across sessions, and we assessed validity with participant identifiability, task-, and age-sensitivity. Measures that do not adjust for volume conduction artifacts - coherence and phase-locking value - had higher reliability, which we interpret as being at least partly artifactual. However, they also had higher identifiability alongside comparable task and age sensitivity. Coherence and phase locking value reached asymptotic reliability and identifiability with only 3 minutes of recording, while measures that mitigate volume conduction were still less reliable and valid with 10 times more data. Of measures that mitigate volume conduction, imaginary coherence had the best performance on metrics used here. Our results emphasize the importance of considering trade-offs between noise removal and signal retention, as missing true brain interactions may outweigh the benefit of removing artifact, and suggest that study designs should consider that mitigating volume conduction comes at a cost in terms of time required for stable measurements. Together our findings can support study design decisions in EEG connectomics research.

4.1 Introduction

Electroencephalography (EEG) connectomics is a fast-growing subfield of EEG neuroscience research (Sadaghiani et al., 2022). While magnetic resonance imaging (MRI)-based modalities have dominated the connectome literature to date, EEG has the advantages of more directly measuring neural activity, millisecond temporal resolution, and greater cost-effectiveness (Sadaghiani et al., 2022). Thus, EEG connectomics research has the potential to contribute to a better understanding of human neurodevelopment, disorders, and disease. EEG's high temporal resolution and inherent confounds have led to dozens of functional connectivity (FC) measures being developed (Cao et al., 2022; Sadaghiani et al., 2022), but the extent to which they contain reliable and valid individual information, necessary for clinical and individual differences research, is not fully known.

From the perspective of brain connectomics, reliability is the extent to which multiple recordings of the same individual produce consistent estimates of FC under similar conditions (Noble et al., 2019). To understand associations between brain activity and relatively stable traits – such as age, sex, cognitive ability, or behavioural tendencies – researchers are generally interested in features of the connectome that are relatively self-consistent across repeated measurement (Colclough et al., 2016; Graff et al., 2022b; Gratton et al., 2018; Levin et al., 2020; Liao et al., 2021; Noble et al., 2017; Wu et al., 2022; Zuo et al., 2013). Nevertheless, more reliable measurements are not always more valid, as connectivity measures are influenced by artifacts such as head motion, meaning that in some cases, higher reliability reflects consistent noise (Parkes et al., 2018). Ideally, connectivity measures should capture only information about brain interactions, but researchers are often faced with a trade-off between retaining signal of interest and removing artifactual noise (Graff et al., 2022a; Nolte et al., 2004; Stam et al., 2007).

Assessing validity in neuroimaging can be difficult in the absence of ground-truth measurements, but several desirable properties associate with validity. One is the presence of individually specific information, i.e., participant identifiability (Finn et al., 2015; Wu et al., 2022). Connectomes are identifiable if self-to-self stability is greater than similarity to others, also known as connectome 'fingerprinting' (Finn et al., 2015; Kaufmann et al., 2017; Graff et al., 2022b; Rajapandian et al., 2020). Likewise, connectomes should be sensitive to the task a participant is performing. While the impact that tasks have on connectomes is multifaceted (Gratton et al., 2018; Mostame and Sadaghiani, 2020), one straightforward marker of task sensitivity is that connectomes should be more similar when collected from the same task relative to a different task (Vanderwal et al., 2017). Connectomes should also be sensitive to age effects (Grayson and Fair, 2017) in that connectomes from participants within an age group should be more similar to one another than to connectomes from other age groups. One further aspect affecting reliability and validity is data quantity. Some FC measures may have high reliability only when sufficient data are collected, as is the case with functional MRI (fMRI) connectomes (Finn et al., 2015; Gordon et al., 2017; Noble et al., 2017; Zuo et al., 2013).

In EEG there are a breadth of measures available with different underlying assumptions and trade-offs, making it challenging to *a priori* choose the 'best' measure. FC measures differ in attributes such as if they are: linear or non-linear, parametric or nonparametric, time-, frequency-, or time-frequency-based, and sensitive to phase or amplitude synchrony (For an overview of EEG FC measures, see Cao et al., 2022). Here we focus on phase-based measures as they are commonly used (Cao et al., 2022; Sadaghiani et al., 2022) and have several advantages including requiring few assumptions (Cohen, 2014), and capturing distinct information from what is captured by fMRI (Engel et al., 2013; Wirsich et al., 2021). One of the main drawbacks to phase-

based FC measures is that they require precise experimental timing (Cohen, 2014) but the exact start- and endpoint of a task is less critical when using fewer, longer task intervals, as in resting state or passive-viewing conditions. Importantly, long stretches of continuous data are not required, allowing noisy epochs to be discarded. This may be especially useful in populations with higher head movement, such as children. Despite their relative simplicity, the category of phase-based FC measures is diverse in their approach to capturing signal of interest or minimizing spurious noise. Some phase-based FC measures only capture phase synchrony, such as phase lag index (pli; Stam et al., 2007) and phase locking value (plv; Lachaux et al., 1999), while others are also sensitive to changes in amplitude, such as coherence (coh; Cohen, 2014) and imaginary coherence (imcoh; Nolte et al., 2004). Further, some phase-based measures, such as phase slope index (psi; Nolte et al., 2008), estimate causal influences in synchrony, making them effective connectivity measures.

Another aspect differentiating phase-based FC measures is whether they mitigate the problem of volume conduction (Nunez et al., 1997). Due to the spatial separation between where electric fields are generated and where they are measured, the same underlying brain activity can be detected across multiple electrodes. This can lead to inflated estimates of FC across these electrodes when calculating FC in sensor space, hampering validity. While source space has become common for EEG analyses, partly as a means to correct for volume conduction (Gross et al., 2001; Schoffelen and Gross, 2009), artifacts remain in the estimated signals (Sarvas, 1987; Sadaghiani et al., 2022). Due to being the same signal recorded twice, spurious connectivity due to volume conduction consists of synchrony with zero lag; as thus, some FC measures, such as imcoh and pli, mitigate artifacts by only being sensitive to time-lagged synchrony (Nolte et al., 2004; Stam et al., 2007). However, these measures likely fail to capture some 'true' connectivity

in the process (Nolte et al., 2004; Stam et al., 2007), and at the same time, they cannot fully remove spurious connectivity (Palva et al., 2018).

Several studies have considered the reliability of various aspects of EEG FC (Büchel et al., 2021; Cannon et al., 2012; Haartsen et al., 2020; Hardmeier et al., 2014; Hatz et al., 2016; Kuntzelman and Miskovic, 2017; Lopez et al., 2023; van der Velde et al., 2019). In general, these studies have each only considered one or two FC measures, have not analyzed the same frequency bands, and have considered a range of non-overlapping graph theory metrics. One tentative conclusion from this work is that measures that do not correct for volume conduction have higher reliability (Lopez et al., 2023). In magnetoencephalography (MEG), a reliability study that directly compared several FC measures came to the same conclusion, but they interpreted the higher reliability of non-corrected measures as likely being due to artifact (Colclough et al., 2016), emphasizing the importance of also assessing validity.

To our knowledge, the validity of measures has not been systematically compared alongside reliability. Reliability and validity are also known to be dependent (to some extent) on the amount of data available, however, whether EEG FC measures differ in their sensitivity to the amount of data collected is not yet known. Here, we compare different phase-based approaches to calculating EEG connectomes by assessing their reliability and validity. We focus on these specific questions: 1) To what extent do different EEG FC measures vary in reliability, as assessed with self-stability, across different measurements? 2) To what extent do different EEG FC measures vary in validity, as assessed with participant identifiability and task- and age-sensitivity? 3) How do FC measures vary in the amount of data required to establish reliability and validity?

4.2 Methods

4.2.1 Participants

We recruited 50 participants consisting of 25 parent-child dyads (5 female-female, 7 female-male, 8 male-female, 5 male-male). Parents were 33.75-47.13 years old (mean: 41.39, std: 3.63 years, 12 female). Children were typically developing and 6.56-8.92 years old at time of first visit (mean: 7.88, std: 0.69 years, 13 female). Our sample included two families where both parents and two children participated. See Supplemental Table 4.1 for more detailed participant information. Participants were excluded if they had psychiatric or neurodevelopmental diagnoses or could not understand English. Parents provided informed consent for both their own and their child's participation, while children provided assent. This study was approved by the University of Calgary Conjoint Health Research Ethics Board.

4.2.2 Data collection

Data were collected at the Alberta Children's Hospital across four sessions per parent-child pair, with each session completed approximately one week apart (see Supplemental Table 4.1). EEG data were collected using a 64-channel Magstim EGI HydroCel Geodesic Sensor Net (Eugene, Oregon), soaked in an electrolytic solution, with data sampled at 1000 Hz. The impedance level was kept below 50 kΩ during recording. MRI data, used for source reconstruction, was collected using a 3T GE MR750 w (Waukesha, WI) scanner with a 32-channel head coil, which consisted of a T1w 3D BRAVO sequence (TR = 6.764 ms, TE = 2.908 ms, FA = 10°, voxel size = 0.8x0.8x0.8 mm^3). Functional and diffusion MRI, and other cognitive and behavioral data were also collected but not analysed for the present study. Order of data collection was pseudo-randomized, i.e., for half the participants, parents underwent EEG then

MRI data collection, while children underwent MRI then EEG data collection. In the other half, the order was reversed.

Each of the four EEG recording sessions included three approximately 7-minute passive video viewing tasks, for a total of 12 unique videos that were drawn from three categories or 'tasks':

a) Relax: a continuous relaxing video with gentle music, slow-moving imagery, and no narration, designed to minimize cognitive load, similar in concept to *Inscapes* (Vanderwal et al., 2015). All participants watched the same Relax videos in the same order across sessions; the videos shown depicted: 1) a Japanese island, 2) a walk through a canyon, 3) scenes from Canadian cities, and 4) the Earth as seen from space.

b) YouTube: a series of popular (i.e., viewed more than 1 million times at the time they were downloaded) and visually engaging YouTube clips (24-65 seconds long, mean = 46.4 seconds), designed to provide constant visual interest with minimal narrative. Clips included dancing, arts and crafts, simple science experiments, scenes from video games, stop-motion animation, and humorous short videos. All participants watched the same YouTube clips in the same order across sessions.

c) Dora: a series of scenes from *Dora and the Lost City of Gold* (2019), a live-action movie. This condition was designed to contain a continuous narrative. Over the course of the study, participants viewed successive clips of this movie sequentially during MR and EEG imaging. Whether participants underwent EEG or MRI first determined if they watched the first sequential video or the last sequential video during EEG recording, i.e., half the participants watched clips 1, 3, 6, and 9 during EEG and half watched clips 3, 6, 9, and 12 during EEG.

The EEG recording was timed to begin and end with video playback. Electrodes were checked in between videos to ensure impedances remained below 50 kΩ. Within each session, video category order was randomized for each participant.

598 EEG recordings were collected in total. Four sessions of data (three recordings each) were excluded due to concerns with correct cap placement and five recordings were excluded due to too many noisy epochs (see *4.2.3 Preprocessing*). This left 581 recordings in the final sample across 50 participants. All participants had at least 8 recordings, with at least two recordings of each video type; 42 participants had all 12 recordings.

4.2.3 Preprocessing

Data were preprocessed using MNE-Python (Gramfort et al., 2013). Data were downsampled to 250 Hz, bandpass filtered from 1-45 Hz, and notch filtered at 60 Hz. Channels were then visually inspected to exclude noisy channels (e.g., excessive artifacts or long stretches of no signal). Electrodes were re-referenced to the average signal of all (remaining) electrodes, data were divided into 2-second (non-overlapping) epochs, and an independent component analysis was carried out to remove eye and other artifacts. Components were classified as brain activity or artifact using mne-icalabel (Li et al., 2022), and classification was confirmed using visual inspection, erring on the side of being conservative with component removal. Following this, each epoch for each channel was checked for noise, where an amplitude exceeding ±70 µV or a change of 100 µV over 100 ms was classified as noisy (Schubert et al., 2015). Channels with noisy epochs were interpolated using data from other channels via spherical spline interpolation. Any epoch with more than 15 unusable channels – either requiring interpolation or having been

excluded during visual inspection – was removed from analysis. Recordings with fewer than 150 non-noisy epochs were excluded.

T1w images were preprocessed using FreeSurfer (Dale et al., 1999). Electrode placements were estimated based on the standard montage for a 64-channel EGI HydroCel Geodesic Sensor Net and by using an automated approach to coregistration with the participant's MRI (Houck and Claus, 2020). Source localization was carried out using the eLORETA algorithm (Pascual-Marqui, 2007). For each of 68 nodes in the Desikan-Killiany atlas (Desikan et al., 2006), the time course was extracted using the first principal component of the vertices via the pca_flip setting.

4.2.4 Connectivity measures and connectome generation

All connectivity measures were calculated in MNE using multitaper spectrum estimation with adaptive weights. We calculated six different connectivity measures, which are summarized in Table 4.1.

1) The magnitude of coherency, often referred to as coherence (coh; Cohen, 2014; Nolte et al., 2004) is one of the most established techniques to estimate electrophysiological synchrony. It measures the linear covariance between two spectra, incorporating both phase and power information. It does not incorporate any strategy to mitigate volume conduction.

2) Phase locking value (plv; Lachaux et al., 1999) measures the degree of phase covariance. It was proposed as an alternative to coh that would be effective for non-stationary sources and would only account for phase information, ignoring amplitude covariance. Like coh, it does not incorporate any strategy to mitigate volume conduction.

3) Imaginary coherence (imcoh; Nolte et al., 2004), involves selecting only the imaginary component from a coherency calculation between two regions. It was proposed as an alternative to coh that mitigates volume conduction artifacts by being insensitive to zero time-lag synchrony between nodes, thereby removing spurious connectivity (while also likely removing true signal of interest). While imcoh and other zero time-lag synchrony measures avoid most of the direct effects of volume conduction, they are still affected by indirect effects such as field spread artifacts (Palva et al., 2018).

4) Phase lag index (pli; Stam et al., 2007) measures the consistency of phase lag between two time series. Like imcoh, it ignores zero time-lag synchrony – spurious or real - but unlike imcoh it is only sensitive to phase covariance, ignoring amplitude effects.

5) Weighted phase lag index (wpli; Vinck et al., 2011) is a modification of pli where phase differences are weighted according to the size of the difference, rather than assessing only the consistency of phase difference. It has been shown to be more robust against noise with increased statistical power (Vinck et al., 2011).

6) Phase slope index (psi; Nolte et al., 2008) is based on the assumption that for an interaction with a given time delay, the phase difference between signals will increase with frequency. It measures the slope of these phase differences, where the sign of the slope indicates which region is responsible for the interaction, making it an effective connectivity measure. Like imcoh, pli, and wpli, it is less affected by volume conduction compared to coh and plv.

FC was calculated across three frequency bands: alpha (8-13 Hz), beta (13-30 Hz), and broadband (2.5-45 Hz). We focused on alpha and beta bands because some evidence suggests reliability is highest in these bands (Lopez et al., 2023), which may allow validity to be better assessed. FC was calculated between all pairs of regions in the Desikan-Killiany atlas, yielding

2278 edges, which were Fisher z-transformed to approximate a normal distribution and vectorized.

Measure	Sensitive to		Volume conduction mitigation	Effective connectivity
	Phase	Amplitude		
Phase slope index	✓		✓	✓
Phase lag index	✓		✓	
Weighted phase lag index	✓		✓	
Imaginary coherence	✓	✓	✓	
Coherence	✓	✓		
Phase locking value	✓			

Table 4.1. Properties of several phase-based FC measures.

4.2.5 Average connectomes

For each connectivity measure and frequency band, we calculated the average connectome across all 581 recordings. These were visualized as connectivity matrices, with regions divided into hemispheres and arranged anatomically from posterior to anterior. Since imcoh and psi have both positive and negative FC values, we visualized the average of both signed and absolute values. To assess the extent to which FC methods were consistent with each other, we calculated the Pearson correlation between each pair of vectorized, averaged connectomes.

4.2.6 Connectivity reliability

While there are multiple ways to assess connectome reliability, a straightforward approach is to vectorize connectomes and calculate the within-subject test-retest correlation (Colclough et al., 2016; Finn et al., 2015; Gordon et al., 2017), hereafter referred to as self-stability. For each connectivity measure and frequency band, pairs of connectomes from the same participant were Pearson correlated to generate a self-stability score. Correlations were

Fisher-z transformed. Unless otherwise stated, analyses refer to between-session-self-stability scores to minimize spuriously higher self-stability due to more similar recording conditions, e.g., cap placement, other electric fields, and electrolytic solution concentration. We also present within-session-self-stability in the supplement for ease of comparison to other studies (Colclough et al., 2016). Participants had between 23 and 54 (mean = 50.6) between-session-self-stability scores, and 5 to 16 (mean = 11.6) within-session-self-stability scores.

4.2.7 Connectivity validity

We considered 3 aspects of connectome validity:

1) Participant identifiability: For each participant, participant identifiability was assessed with individualization, the difference between average self-stability and average similarity-to-others (Graff et al., 2022b). Similarity-to-others was assessed along the same lines as self-stability (Vanderwal et al., 2021), where we calculated the Pearson correlation between pairs of connectomes from different unrelated participants (i.e., not to their parent or their child). For the families where two siblings and both parents participated, siblings were not compared to each other as they are related, but parents were compared to each other (biologically unrelated). Participants had between 4488 (8×561) and 6684 (12×557) similarity-to-other scores (mean = 6445). These were likewise Fisher z-transformed.

We also tested two supplemental metrics of individualization:

-Matching accuracy: for each participant, the fraction of self-stability scores that were higher than all similarity-to-others scores.

-Self-stability percentile rank: for each participant, the average percentile of self-stability scores, relative to similarity-to-others. A score of 50% suggests no difference between self-stability and the median similarity to someone else.

2) Task sensitivity: Task sensitivity was assessed both within each connectivity measure and between connectivity measures. This was done by comparing every participant's average same-task-self-stability (e.g., Dora to Dora) with their average cross-task-self-stability (e.g., Dora to YouTube). Participants had between 8 and 18 same-task-self-stability scores (mean = 16.86) and 15 to 36 cross-task-self-stability scores (mean = 33.7). Within each connectivity measure, we used paired t-tests to assess if same-task-self-stability was significantly higher than cross-task-self-stability. We then calculated the difference between same-task- and cross-task-self-stability for each participant, and used paired t-tests to assess if these differences were significantly different between connectivity measures. Likewise, we also assessed task sensitivity by comparing same-task- and cross-task-similarity-to-others. Because tasks varied in how many epochs were available after the removal of bad epochs (on average, Relax: 201.6 epochs, YouTube: 206.1 epochs, Dora: 198.0 epochs) all connectomes were calculated using only the first 150 non-removed epochs to ensure consistency, where below 150 epochs was the exclusion threshold for a recording.

3) Age sensitivity: Age sensitivity was assessed similarly to task sensitivity, both within each connectivity measure and between connectivity measures. This was done by comparing every participant's average similarity-to-others in their age group (i.e., children's similarity to other children, adults' similarity to adults) with their average similarity-to-others across age groups (i.e., children's similarity to adults, adults' similarity to children). Within each connectivity measure, we used paired t-tests to assess if similarity-to-others was significantly greater within-

relative to across-age-groups. We then calculated the difference between same-age- and cross-age-similarity-to-others for each participant, and used paired t-tests to assess if these differences were significantly different between connectivity measures. Because children and adults varied in how many epochs were available after the removal of bad epochs (on average, children: 200.0 epochs, adults: 203.7 epochs), all connectomes were again calculated using only the first 150 non-removed epochs.

4.2.8 Effect of recording length

fMRI connectome studies have shown that as recording length increases, reliability and validity increase (Finn et al., 2015; Gordon et al., 2017; Noble et al., 2019). Here we similarly investigated the extent to which reliability and validity are sensitive to the amount of data available across FC measures. Of the 42 participants with all 12 recordings, we concatenated the six recordings from their first two sessions, and the six from their last two sessions, generating two sets of data. We then calculated self-stability, similarity-to-others, and individualization using progressively larger quantities of data, starting with 1 epoch from each recording (6 in total per set) and increasing up to 6×150 epochs per set (30 minutes of data per set, 60 minutes in total).

<u>4.3 Results</u>

We found comparable results across frequency bands and therefore describe alpha-band findings. Beta and broadband results are available in Appendices C and D respectively.
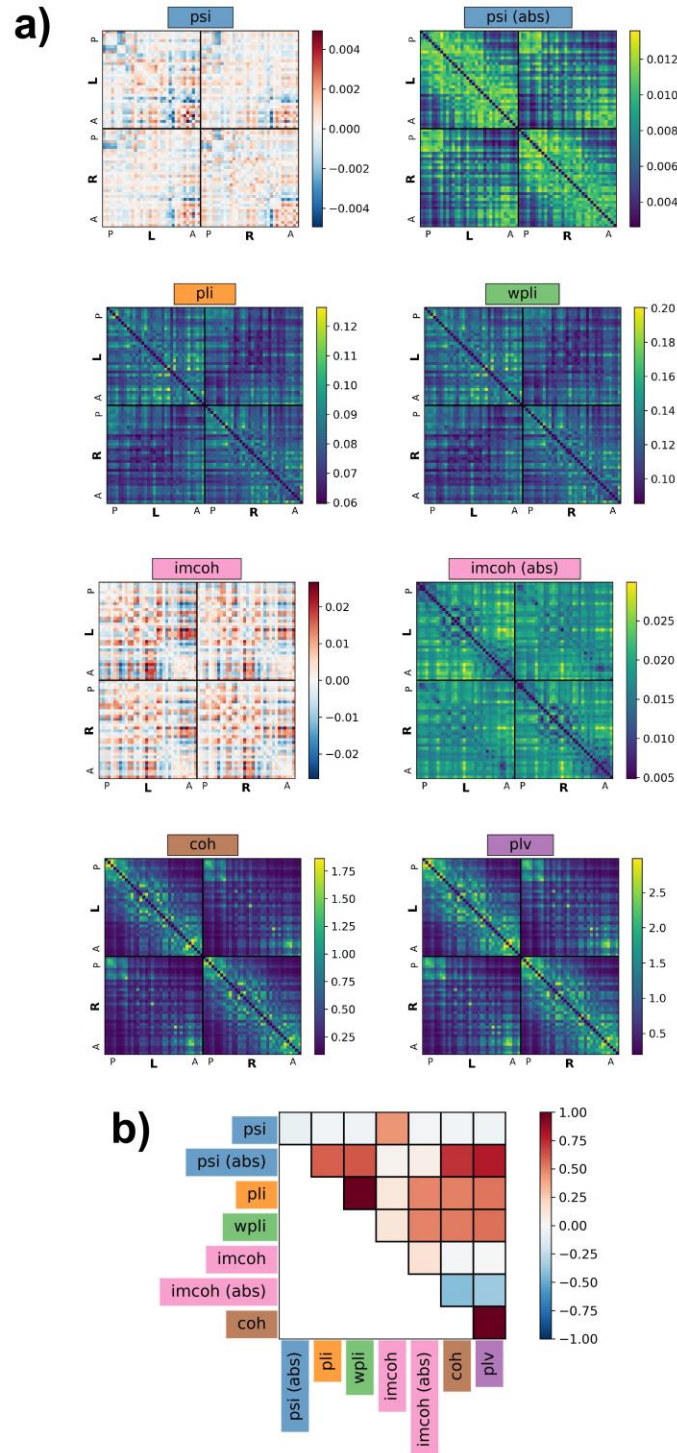
Figure 4.1. Average connectomes for each connectivity measure. a) The average connectome across all 581 recordings. Regions were divided by hemisphere (L and R) and arranged anatomically from posterior (P) to anterior (A). Since imcoh and psi have both positive and negative FC values, both the average of signed values and absolute values are shown. **b)** The Pearson correlation between each pair of vectorized, averaged connectomes. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

4.3.1 Average connectomes

We considered similarities and differences between connectomes calculated using different measures (Figure 4.1). All connectomes show left-right hemisphere symmetry and stronger within-hemisphere than between-hemisphere connectivity (Figure 4.1a). Coh, plv, and psi show notable distance-dependent effects, where short edges are stronger than long edges. On the other hand, for imcoh the weakest connections appear to be between regions spatially close to one another. The correlation between different FC measure connectomes is shown in Figure 4.1b. The signed versions of psi and imcoh, owing to having both positive and negative FC values, showed near zero similarity to other measures, except to each other (r = 0.44), but their absolute values – plv (abs) and imcoh (abs) – were comparable to other FC measures. As suggested visually, wpli and pli showed strong similarity to each other (r = 0.99). They also both showed strong similarity to the other measures (0.49 < r < 0.62 for psi (abs), imcoh (abs), coh, plv). Likewise, coh and plv were highly similar to each other (r = 0.99), with high correlations to psi (abs; r = 0.72 and r = 0.78 respectively), pli, and wpli, but were anti-correlated with imcoh (abs; r = -0.43 and r = -0.37). Overall, this suggests that the 6 FC measures capture overlapping but distinct aspects of brain connectivity, with some measures such as pli and wpli being highly similar, and others such as imcoh and plv showing more distinct FC patterns.

4.3.2 Reliability and participant identifiability

We found the highest reliability, as measured with self-stability, using the two FC measures most sensitive to volume conduction, coh and plv, and comparatively low self-stability when using psi, pli, wpli, and imcoh (Figure 4.2a). Coh and plv also showed relatively high
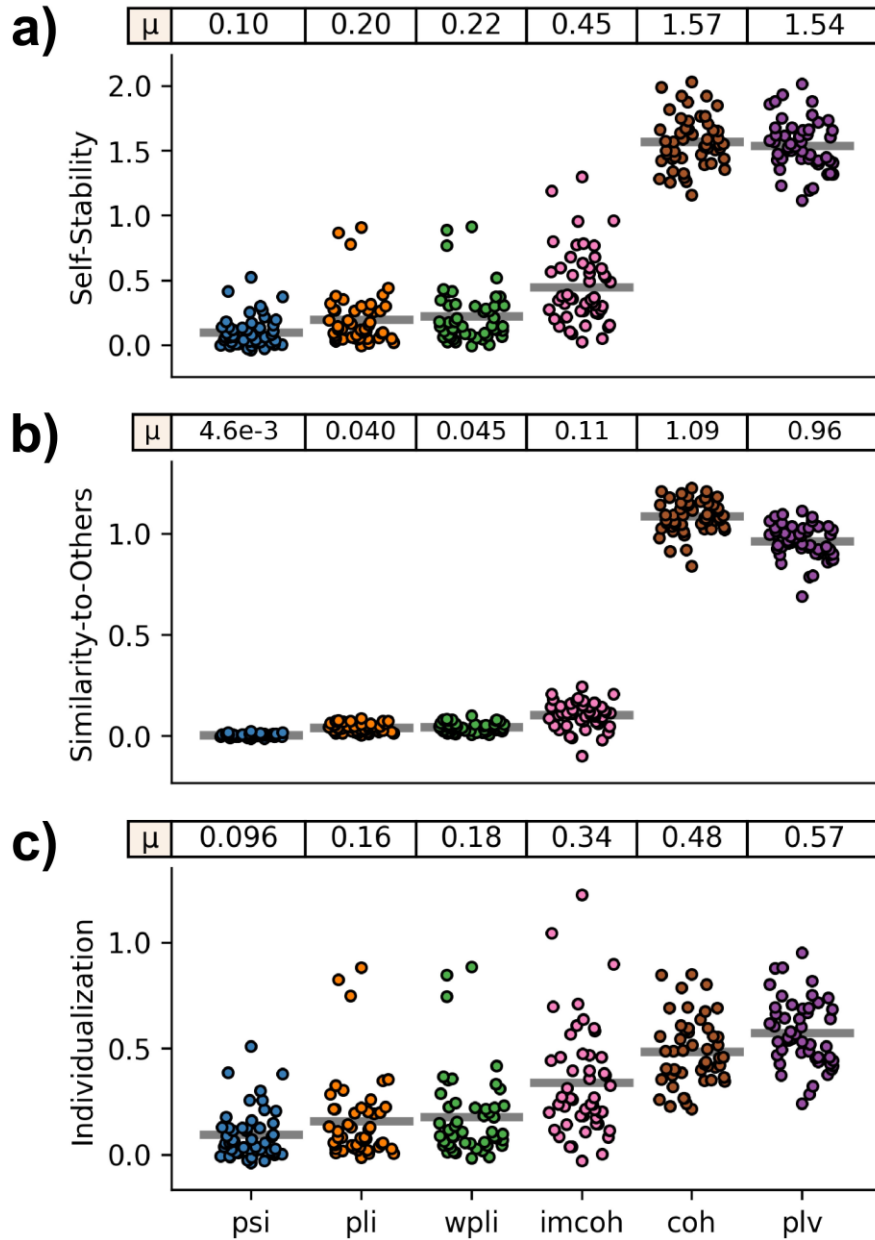
Figure 4.2. Reliability and identifiability across FC measures. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on different days. **b)** Mean similarity-to-others, the average Fisher-z correlation between a participant's connectomes and connectomes from all other participants. **c)** Individualization, the difference between mean self-stability and mean similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

mean similarity-to-others (Figure 4.2b), suggesting that with these FC measures all connectomes are more similar to one another. Nevertheless, coh and plv also had higher validity based on identifiability, as measured with individualization (Figure 4.2c). However, all FC measures had positive average individualization scores, suggesting at least some degree of validity. The ranking of measures using match rate (Supplemental Figure 4.1a, Appendix A) and self-stability percentile (Supplemental Figure 4.1b, Appendix A) were similar to individualization. Of note, average match rates were 51% for plv, compared to under 5% for psi, pli, and wpli. Of FC measures that mitigate volume conduction, imcoh outperformed psi, pli, and wpli in terms of both reliability and validity, while psi had the lowest scores. When repeating the analysis measuring self-stability using scans collected on the same day, we observed the same trend of coh and plv outperforming psi, pli, wpli, and imcoh, with higher identifiability for all FC measures (Supplemental Figure 4.2, Appendix A).

Across reliability and validity metrics, all differences between FC measures were significant (uncorrected p < 0.005; Supplemental Figure 4.3, Appendix A), even if mean differences were small because analyses used paired tests and changes between FC measures were relatively consistent across participants.

4.3.3 Task sensitivity

A valid FC measure should show higher self-stability between connectomes generated from the same task stimuli compared to those obtained from different task stimuli. We found this effect in all FC measures (Figure 4.3a), though psi did not survive Bonferroni correction. When comparing the change in self-stability across measures, the only significant difference was between imcoh and pli (p = 0.04; Supplemental Figure 4.5, Appendix A), suggesting all

measures were near-equally valid on this metric. Likewise, pli, wpli, imcoh, coh, and plv showed

greater similarity-to-others when comparing the same task than between different tasks, though

the magnitude of these differences were small (average $\Delta z < 0.01$; Figure 4.3b). When

comparing the change in similarity-to-others across measures, imcoh was significantly higher

than psi, pli, coh, and plv, but significance did not survive multiple comparisons correction.
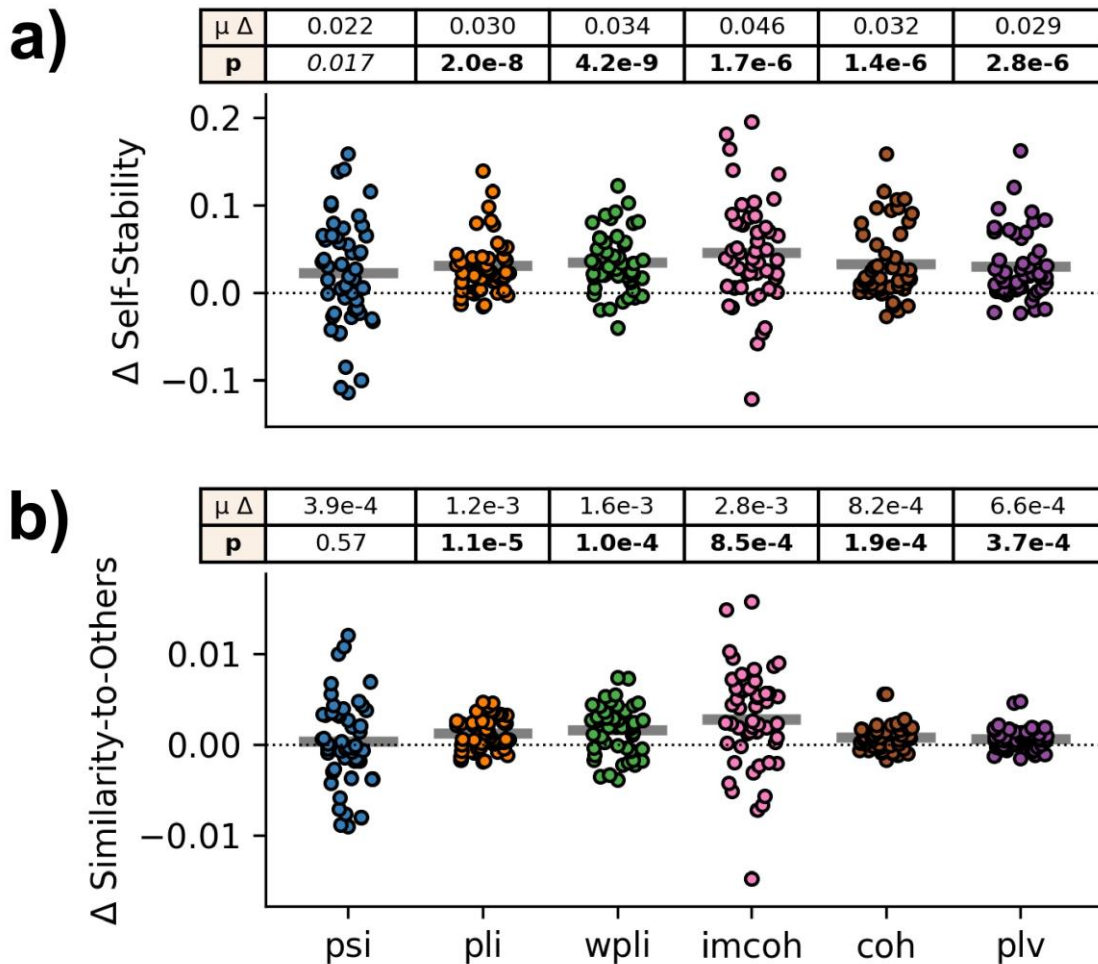
Figure 4.3. Task sensitivity across FC measures. Each dot represents the difference between a participant's average same-task score and their average cross-task score. Lines represent mean differences across participants, which are also displayed at the top of each subplot. For each FC measure, a paired t-test was used to assess if same-task and cross-task scores were significantly different; uncorrected p-values are shown. Bold = p < 0.05 Bonferroni corrected; italics = p < 0.05 uncorrected. **a)** Mean change between same-task- and cross-task-self-stability. **b)** Mean change between same-task- and cross-task-similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

4.3.4 Age sensitivity

Like task sensitivity, a valid FC measure should show greater similarity-to-others when participants are compared to people of similar ages. This was observed in all FC measures (p < 0.005; Figure 4.4). The effect was smallest in psi, but no significant difference was observed between pli, wpli, imcoh, coh, and plv (Supplemental Figure 4.6, Appendix A), suggesting these measures were comparably valid on this metric.

μ Δ	6.0e-3	0.022	0.022	0.026	0.030	0.029
p	**1.6e-3**	**1.5e-6**	**1.5e-5**	**3.4e-3**	**3.6e-4**	**3.8e-4**
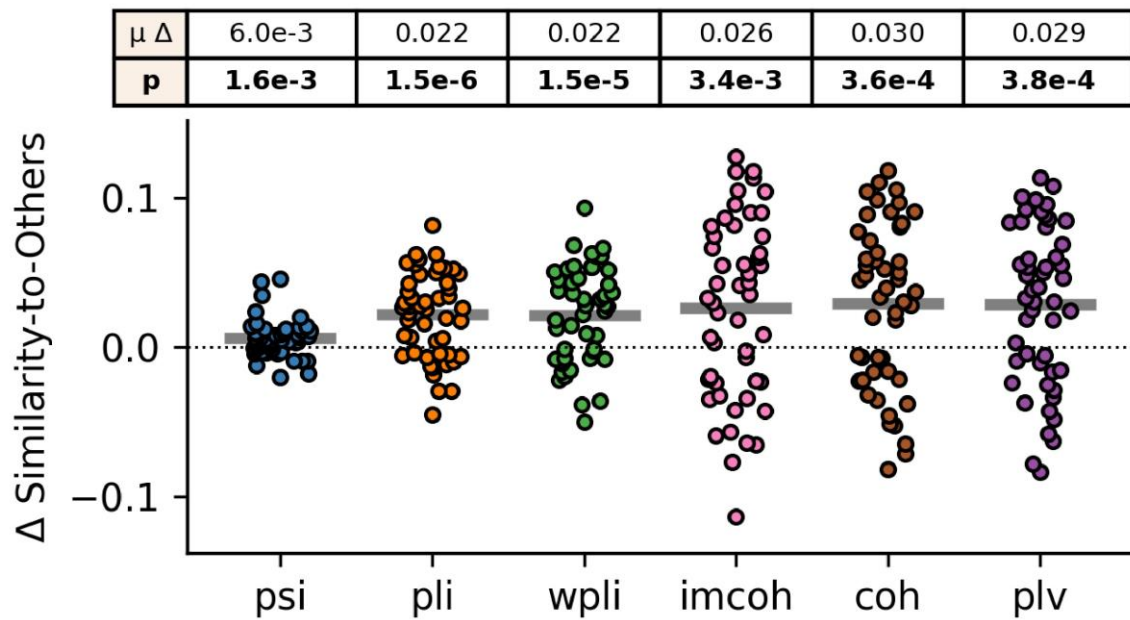
Figure 4.4. Age sensitivity across FC measures. Each dot represents the difference between a participant's average same-age-similarity-others and their average cross-age-similarity-to-others. Lines represent mean differences across participants, which are also displayed at the top of each subplot. For each FC measure, a paired t-test was used to assess if same-age- and cross-age-similarity-to-others scores were significantly different; uncorrected p-values are shown. Bold = p < 0.05 Bonferroni corrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*
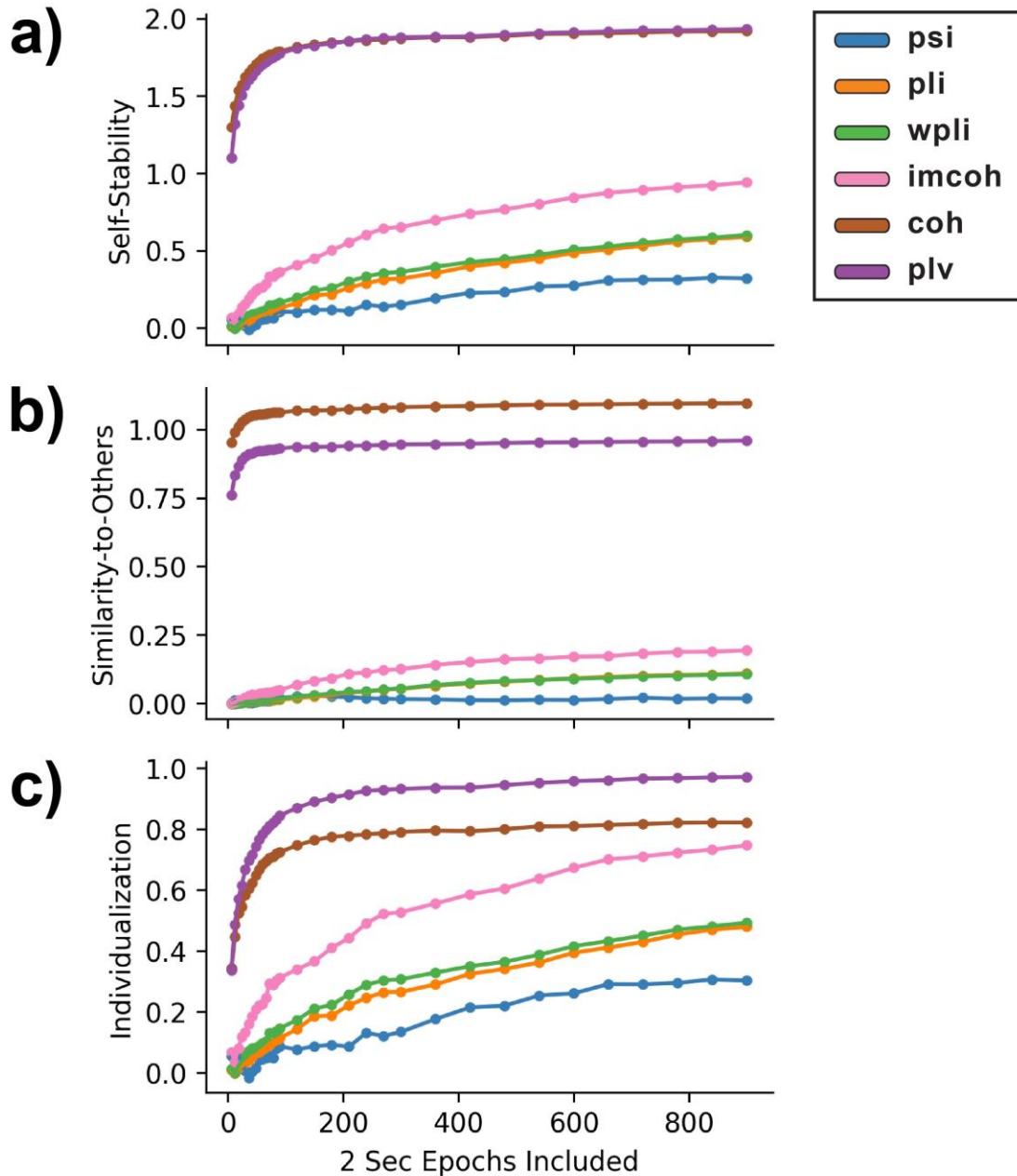
Figure 4.5. Effects of different recording lengths on reliability and identifiability, across FC measures. Points shown are across-participant averages. For each participant, the recordings from their first two sessions and from their last two sessions were concatenated, giving two sets of data per participant. Epochs were added sequentially, e.g., '60 epochs' refers to using the first 10 epochs available from the 6 separate recordings that were combined. **a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on different days, averaged across participants. **b)** Mean similarity-to-others, the average Fisher-z correlation between a participant's connectomes and connectomes from all other participants, averaged across participants. **c)** Individualization, the difference between mean self-stability and mean similarity-to-others, averaged across participants. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

116

4.3.5 Recording length

We assessed the effect of recording length on reliability and validity. Coh and plv reached near-optimum self-stability with relatively short quantities of data (~50 epochs; Figure 4.5a). Though more data was required to reach an asymptote for individualization, there was negligible benefit for more than ~200 epochs (Figure 4.5c). On the other hand, the four measures that are less sensitive to volume conduction continued to show benefit with longer and longer recordings. In particular, imcoh nearly closed the gap in individualization with 900 epochs, though it appeared to also be approaching an asymptote in self-stability, suggesting it may never overtake coh or plv. Regardless of recording length, psi, pli, and wpli tended to show the least reliability and validity, suggesting longer recording lengths are insufficient to improve reliability and validity of these measures.

4.4 Discussion

In this study, we investigated the reliability and validity of six different phase-based EEG FC measures for whole-brain connectomics. We found that measures that do not mitigate volume conduction, namely coh and plv, had much higher reliability, but also showed much higher similarity-to-others, suggesting all connectomes are more similar with these measures. We then assessed validity with participant identifiability, task sensitivity, and age sensitivity. We found that coh and plv had higher validity in terms of identifiability, and comparable validity in task and age sensitivity. Taken together, these results suggest that the higher reliability of coh and plv may be driven by volume conduction artifacts, but despite this susceptibility, coh and plv contain important information about individual differences. Given the proliferation of methods for EEG connectomics, these findings can contribute to choosing appropriate metrics for study goals.

The lack of improved validity when mitigating volume conduction is likely due to the removal of both artifact and true signal of interest. This interpretation is consistent with how measures like imcoh and pli were originally developed. The developers of pli (Stam et al., 2007) note that their approach requires accepting the risk of missing meaningful interactions that have near zero phase coherence, and acknowledging that the remaining connectivity information may be incomplete. Similarly, the developers of imcoh (Nolte et al., 2004), refer to imcoh as an "extreme position" that contains, "at best, only half of the picture". These measures were originally proposed to be entirely artifact-free, making this trade-off more palatable, but newer work suggests they can still be susceptible to false positives due to field spread (Palva et al., 2018).

While the limitations of imcoh and similar methods may make them unappealing, it can be difficult to defend the utility of measures prone to artifact. Colclough et al. (2016) found similar reliability findings in MEG, where FC measures prone to volume conduction had higher within-subject and between-subject consistency, which they suggest is primarily due to volume conduction artifacts. For this reason, they argue against using these measures, suggesting that, "interpretable connectivity estimation is only possible when zero-lag connections are removed or otherwise ignored." While we agree that the interpretation of FC is more difficult for measures susceptible to volume conduction, our work suggests that they still have utility. These aforementioned papers and our work here emphasize the importance of considering trade-offs between maximizing signal and minimizing noise, rather than imposing a false valid/invalid dichotomy. Though volume conduction susceptibility clearly reduces FC validity, coh and plv may still be useful measures if alternatives are unable to retain sufficient signal of interest. While it is common to use only one measure of FC in a study, both due to practical limitations and to

simplify analyses, we echo the recommendation from Nolte et al. (2004) to think of measures like imcoh not as a replacement for coh or plv, but as a necessary additional analysis.

Of the FC measures that correct for volume conduction, imcoh outperformed psi, pli, and wpli on metrics assessed here. It showed both higher reliability, as measured with self-stability, and greater identifiability, both in terms of individualization and fingerprinting match rate. In terms of task sensitivity, it showed the greatest difference of all FC methods between same-task and cross-task similarity-to-others, and it was comparably valid for all other analyses. Of phase-based methods that correct for volume conduction, the use of imcoh has started to become the *de facto* standard for full connectome analyses (Mahjoory et al., 2017; Nentwich et al., 2020; Wirsich et al., 2021). Given our results here, and the need for reproducibility across studies, this convergence seems appropriate. However, we found that functional connectomes generally show modest similarity to each other across FC measures, suggesting that they are capturing different aspects of true interactions. Further, psi is an effective connectivity measure, so while it failed to outperform imcoh on our metrics, imcoh and psi should not be thought of as interchangeable for all analyses. Given these considerations, it may be misleading to think of imcoh as *better* than psi, pli, or wpli, and their different underlying assumptions should always be examined when designing a study.

Considering performance in relation to data quantity, we found that coh and plv required far shorter recordings than other measures to reach an identifiability asymptote (i.e., ~90 epochs, 3 minutes). On the other hand, for standard recording lengths, e.g., 5 minutes (here 150 epochs), imcoh and other measures that correct for volume conduction may have less than half the identifiability of coh or plv, and less than half of the maximum identifiability a longer recording would provide. With sufficient data, imcoh approached the same level of individualization as

coh, but this required each connectome to be generated from 30 minutes of post-processed data collected across six recordings over two sessions. Our results suggest that even greater quantities of data may still prove beneficial to reaching stable imcoh values. Based on this analysis, our findings suggest that for applications where stable volume-conduction adjusted connectomes are needed, relatively long acquisitions are desirable.

While we interpret our main findings as showing that volume conduction mitigation may remove valid signal of interest, other interpretations of our findings are possible. Since volume conduction artifacts are at least partially distance-dependent (Schoffelen and Gross, 2009), subject-specific head geometry will lead to subject-specific artifacts. This could improve identifiability despite worse validity. Similarly, we note that our age sensitivity analysis could be affected by volume conduction, as the impact of volume conduction artifacts is different between children and adults (Grieve et al., 2003). This highlights the importance of considering multiple metrics of validity. These concerns are mitigated by our task sensitivity analysis where participants are only compared to themselves under different conditions, which should result in similar noise profiles. Considering we found coh and plv to have similar validity to other measures in this analysis, it is harder to discount their validity elsewhere. We also note the possibility that methods that mitigate volume conduction may also have distance-dependent biases if nearby regions are more likely to have true near-zero-phase-lag interactions. This could cause longer range connections to have a higher percentage of their true signal captured by the FC measure and cause similar increases in identifiability despite worse validity. To our knowledge this has not been systematically investigated. These concerns necessitate further investigations into the validity of different EEG FC measures. One potential approach could combine conventional EEG with intracranial EEG (iEEG) - which avoids volume conduction –

120

allowing one to more accurately verify the extent to which different FC measures are susceptible to noise relative to captured signal.

There are a number of limitations of our work. While we find that coh and plv are more valid than alternative FC measures, their limitations may be more apparent with other study designs. We estimated FC in source space, which is less affected by volume conduction compared to sensor space (Sadaghiani et al., 2022; Nolte et al., 2004). We do not necessarily expect our findings to hold in sensor space, though exploring that is beyond the scope of the present study. Our choice of parcellation could also impact findings, as the relatively large and functionally heterogeneous parcels in the Desikan-Killiany atlas could avoid the worst aspects of volume conduction, compared to e.g., the Destrieux atlas with smaller parcels (Destrieux et al., 2010). Relatedly, volume conduction artifacts do not impact all edges equally (Schoffelen and Gross, 2009), which make them especially problematic when comparing edges, e.g., when testing which edges are strongest under a given experimental condition or when defining networks. Our analyses did not include any between-edge comparisons or network classifications, instead opting for a whole connectome analysis which may be less affected by these artifacts. Other methodological choices may also limit the scope of findings. Here we split recordings into two-second epochs, though longer epochs may favor connectivity methods that discard more of the signal, such as imcoh. Similarly, we used a 64-channel sensor net, whereas using a 128- or 256-channel device may involve similar benefits for low-signal methods that mitigate noise. We also note that for our task sensitivity analysis, all of our tasks consisted of passive viewing. Using a broader range of tasks could reveal larger differences in validity between FC measures. Similarly, results could vary if we used task-free resting state conditions.

Although we have discussed volume conduction mitigation mainly in terms of choosing connectivity methods like imcoh and pli, various orthogonalization methods are also possible. In these approaches, linear relations between pairs of regions are removed, achieving the same goal of removing zero-phase lag interactions (Colclough et al., 2015; O'Neill et al., 2018). Colclough et al. (2016) found that after orthogonalization, plv had similar reliability as imcoh, pli, and wpli, suggesting that orthogonalization leads to similar outcomes as adopting alternative connectivity measures. We also note that while we focused our analysis on phase-based connectivity methods, many other EEG FC methods exist and warrant investigation. Colclough et al. (2016) argues in favor of amplitude envelope correlations (Hipp et al., 2012), though phase- and amplitude-coupling are thought to reflect different aspects of brain interactions (Mostame and Sadaghiani, 2020; Siems and Siegel, 2020). Many additional EEG FC methods exist (Cao et al., 2022), and other methods may have better trade-offs between retaining signal and removing noise. Future work should compare more connectivity methods, and more directly compare different ways to correct for volume conduction.

4.5 Conclusions

The present study investigated reliability and validity of phase-based EEG FC measures. We found that measures that do not account for volume conduction, namely coh and plv, had higher reliability, potentially due to artifact. However, they also had comparable or higher validity than measures that mitigate volume conduction, i.e., imcoh, pli, wpli, and psi, suggesting that mitigation strategies result in lost signal of interest. We further found that corrected measures require much more data to approach asymptotic reliability and validity. Based on these findings, we recommend careful consideration of which FC measure to use, including in relation

to the amount of data available, and perhaps combining uncorrected with corrected measures for

analyses of brain-behavior relationships.

5.1 Overall summary

To improve our understanding of human functional neurodevelopment, data collected using tools such as fMRI and EEG needs to be both reliable and valid. In this thesis, I used participant identifiability and other markers of validity to assess the reliability and validity of different fMRI preprocessing strategies and EEG FC measures. In general, I found that identifiability increases as noise is removed, but can also drop if insufficient signal is retained. This emphasizes that there are methodological trade-offs involved with noise mitigation. I also assessed how identifiability changes with age as one aspect of development, finding that older children have higher self-stability, which may be related to slower change in FC as children get older, but with no difference in similarity-to-others.

5.1.1 Chapter 2

Researchers have used a variety of fMRI preprocessing strategies, each with advantages and drawbacks. These strategies have not been previously validated in data collected from children – who have higher head motion than adults – which has led to a lack of understanding of which strategies are most appropriate for youth participants. In Chapter 2, I compared several different fMRI preprocessing strategies for high motion early childhood data. I compared ICA-AROMA with the regression of HMP, bandpass filtering with highpass filtering, and tested the use of GSR and censoring. I benchmarked them using QC-FC, which has previously been used to assess noise removed, while adding participant identifiability as a measure of signal retained. The use of passive viewing data also allowed me to use ISC as a further benchmark, as participants watching the same video should demonstrate a similar response, but which may be

obscured by noise or lack of task-evoked signal. I found that across benchmarks, the best

pipeline utilized GSR, censoring, regressing HMP, and bandpass filtering. Further, I found that

ICA-AROMA did not outperform regressing HMP, nor did it obviate the need for censoring, and

that while overly strict censoring improved QC-FC, it negatively impacted identifiability. Taken

together, these results showed that despite the drawbacks of certain preprocessing strategies that

arguably remove some signal of interest along with noise, for high motion data these drawbacks

are largely outweighed by the advantage of recovering true signal. It also demonstrated the

usefulness of identifiability as a measure of validity, while emphasizing the importance of having

multiple benchmarks in the absence of ground truth.

5.1.2 Chapter 3

Individual experiences may shape development in an individually unique way, which is

supported by evidence that adults have more individualized connectomes than children. In

Chapter 3, I investigated how longitudinal identifiability changes across early childhood (i.e., 4-8

years). I found that across both the whole brain and in 9 of 12 networks, identifiability increased

with age. Increases in identifiability corresponded to increases in longitudinal self-stability,

while I found only weak evidence for changes in similarity-to-others across the whole brain. I

also found that older children tended to have less absolute change in FC, partially explaining

increases in self-stability and suggesting non-linear changes even across a relatively narrow age

range. Further, there was only weak evidence for changes in within-scan self-stability, as

assessed with split-half correlations, suggesting increases in longitudinal identifiability are

unlikely an artifact related to data quality. Together, this suggests that increased longitudinal

identifiability with age is primarily driven by increased self-stability, related to progressively

smaller FC changes over time. This work also shows that in addition to being a marker of validity, identifiability may be a feature that increases with age across early childhood.

5.1.3 Chapter 4

A variety of EEG FC measures have been adopted throughout the literature, but the extent to which they capture reliable and valid information is poorly understood. Some measures mitigate volume conduction artifacts, but may fail to capture sufficient signal of interest. In Chapter 4, I compared several different phase-based EEG FC measures on reliability and validity. I found that measures that do not correct for volume conduction had higher reliability, but as they also had higher similarity-to-others, this effect was likely primarily driven by artifact. However, these measures also had higher validity based on identifiability and comparable validity based on both task- and age-sensitivity. Furthermore, measures that do not mitigate volume conduction reached peak reliability and validity with far shorter recording lengths. This suggests that while it is desirable to remove known sources of artifact, such as by utilizing FC measures that are less sensitive to volume conduction, there may be a cost in terms of reduced signal of interest. This work further emphasizes the importance of having multiple benchmarks to assess validity, and like in chapter 2, demonstrates that aggressive artifact removal can overall hamper validity.

5.2 Identifiability as a measure of validity

Validity is an important criterion for scientific measurements but is difficult to establish in practice with neuroimaging as there is often no ground truth available. While work has considered how different aspects of data quality affect identifiability, such as scan length and

126

motion contamination (Horien et al., 2018), to my knowledge identifiability has not previously been used as a benchmark of validity. Here I calculated identifiability by comparing vectorized connectomes both within and across participants, determining if within-subject self-stability is higher than between-subject-similarity. I then compared identifiability across preprocessing pipelines and FC measures, arguing that strategies with greater identifiability were more valid. This approach is based on several assumptions, and has various strengths and drawbacks. It is also worth considering other ways to calculate identifiability as a potential benchmark.

5.2.1 Assumptions

The most basic assumption behind using participant identifiability as a measure of connectome validity is that there is stable and individually unique information captured within functional connectomes. This has been demonstrated in a variety of studies that have included adults (Finn et al., 2015), adolescents (Jalbrzikowski et al., 2020), children (Miranda-Dominguez et al., 2018; Vanderwal et al., 2021), and infants under 1 year of age (Dufford et al., 2021), including longitudinally (Horien et al., 2019; Sato et al., 2021), and across task paradigms (Finn et al., 2015; Gratton et al., 2018; Vanderwal et al., 2017). However, to be a useful measure of validity, identifiability must positively associate with measurement accuracy (i.e., low noise and bias), reliability, and other aspects of data quality. In other words, we assume that individual information within data can be enhanced with data cleaning, and that identifiability will improve if we remove noise while retaining signal. In the absence of ground truth this is difficult to conclusively determine, but several studies suggest that at least certain measures of data quality are related to identifiability. It has been shown that longer scans are necessary to establish connectome stability (Gordon et al., 2017), which matches the finding that connectome

fingerprinting increases as scan length increases (Finn et al., 2015; Horien et al., 2018). Further, considering the systematic bias that head motion can impose on data, it is reassuring that previous work (Horien et al., 2018) and my work in Chapter 2 demonstrates lower identifiability with more motion contamination.

However, this assumption may not apply in all situations. Some evidence suggests that artifacts are replicable within a participant, leading to higher reliability despite noise contamination (Noble et al., 2021; Parkes et al., 2018). For motion artifacts, this may be more pronounced across short time intervals, e.g., scans from the same session (Parkes et al., 2018), which I similarly found in Chapter 2. This may limit the usefulness of identifiability if a participant is only scanned once. Further, certain methodologies or measures may have low identifiability but remain effective at capturing information about other features of interest. This may be the case with psi, which my work in Chapter 4 suggests captures relatively little information about identifiability, but other studies have used psi to demonstrate properties such as working memory processing (Li et al., 2023) and cognitive decline (González-López et al., 2022). Thus, lack of identifiability (as measured here) does not necessarily imply lack of validity for all possible analyses, such as when considering task effects. Higher identifiability could also be achieved by emphasizing subject-specific features that may not be of interest, such as brain size and shape. If a methodology is more vulnerable to noise, but also more sensitive to these characteristics, identifiability could improve despite worse measurement accuracy. As thus, researchers should always consider if high identifiability can be alternately explained, and should not be overly eager to dismiss methods with low individualization scores.

5.2.2 Advantages

There are several advantages to using identifiability as a measure of validity. While a wide range of potential benchmarks exist for validating different methodological choices (see *5.2.6 Other Benchmarks*), identifiability makes relatively minor and fairly intuitive assumptions about how brain connectomes should look – i.e., that there is more self-stability than similarity-to-others. Other benchmarks make assumptions such as how FC should change during certain tasks (Nolte et al., 2004) or between certain groups (Stam et al., 2007), or the degree to which connectomes show network structure (Ciric et al., 2017), which can be harder to justify *a priori*. Identifiability has the related advantage of transferability, as a wide range of validation studies could use identifiability as a benchmark. A measure such as QC-FC is likely inappropriate for comparing anything not directly related to head motion, while for a benchmark such as network modularity, what is known about fMRI connectomes may not transfer to EEG connectomes. While I here used identifiability for both fMRI and EEG connectomes, identifiability has also been observed with structural data (Sato et al., 2021), suggesting it could also be used to validate, e.g., white matter connectomes. Identifiability also has the advantage of testing the validity of the entire connectome. By comparison, a benchmark that considers a specific task may say relatively little about regions of the connectome unrelated to that task, e.g., non-motor regions when considering a motor task (Nolte et al., 2004). Finally, identifiability (especially fingerprinting) is easy to calculate and easy to interpret; given the importance of reproducibility in science, straightforward measures that anyone can adopt are important towards broader scientific goals.

5.2.3 Drawbacks

The biggest drawback to identifiability is that there are both floor and ceiling effects. With fingerprinting in particular, match rates cannot exceed 100%, and several studies, including my work here in Chapters 2 and 3, have match rates near this boundary (Finn et al., 2015; Vanderwal et al., 2017). A methodology could therefore significantly reduce noise or improve signal of interest without noticeably changing identifiability. Likewise, match rates can only drop as low as ~0%, and some work suggests older datasets may have match rates near this lower bounds when using large sample sizes (Waller et al., 2017). While I was able to mitigate this drawback by using individualization, which quantifies the degree of separation between self-stability and similarity-to-others (rather than a binary match/miss), ceiling effects remain. Fundamentally, people are similar to each other, and this similarity is captured within connectomes (Vanderwal et al., 2021), while at the same time, self-stability seemingly has an upper limit (Gordon et al., 2017). Therefore, while it is likely desirable to remove as much noise as possible from a dataset, beyond a certain threshold, improvements may not be reflected with improved identifiability.

There are also drawbacks in measuring identifiability by utilizing the whole connectome. Edges vary in identifiability (Byrge and Kennedy, 2019), with a lot of variation in networks and subnetworks (Finn et al., 2015). While using more edges offers more power, certain strategies may improve overall identifiability while having negative effects on specific edges, which may be of interest. Thus, there may be benefits to considering how identifiability changes in both the whole connectome and in individual networks (Phạm et al., 2023). This could perhaps be useful to better benchmark measures such as psi, which I found to have low full-connectome

individualization, but may capture individually specific information within specific subsets of the connectome.

Another drawback is that when comparing connectomes using Pearson correlations, as done here, the absolute difference between edges is not accounted for, only how the connectomes covary. In other words, the correlation between two vectorized connectomes is unaffected by either vector's mean or standard deviation. Given that global FC is associated with Alzheimer's (Franzmeier et al., 2017), depressive disorder (Pan et al., 2022), and attention (Chen et al., 2023), and that increases in FC have been shown brain-wide with development (Lin et al., 2008), comparing participants without considering average edge strength likely misses meaningful differences. Similarly, if across development strong edges get stronger, while weak edges stay the same or weaken, such global changes would not be reflected via Pearson correlation. While this has implications for developmental studies or studies with neurodiverse participants, there is also the concern that identifiability might improperly benchmark methodologies that involve global changes in FC. For example, while I found little change in identifiability when comparing pipelines either with or without GSR, this may be due to Pearson correlations failing to detect connectome differences that are lost with GSR.

5.2.4 Other implementations

Identifiability can be measured differently than the approach implemented here, which could improve its sensitivity and validity as a benchmark. One question is whether it is best to use Pearson correlations to compare connectomes. An alternative approach is to calculate the geodesic distance between the correlation matrices that make up the connectome (Abbas et al., 2021; Venkatesh et al., 2020). In addition to improving identifiability in rest and across tasks

(Venkatesh et al., 2020), this may be theoretically preferable, as the vectorized Pearson correlation approach implicitly assumes that all FCs are uncorrelated, which is untrue (Abbas et al., 2021). Pearson correlations are also sensitive to outliers, which could disproportionately affect similarity if certain edges are unusually strong. However, the geodesic approach can require a regularization procedure, which is not always possible (Abbas et al., 2021), limiting its adoptability. Nevertheless, this alternative is worth investigating as both a benchmark of validity and a feature of development.

While using the full connectome is likely preferable in most situations, only using a random subset of the connectome could potentially avoid ceiling effects. For example, Vanderwal et al. (2017) compared match rates across viewing conditions; 14 of 30 conditions had match rates of 90% or higher, which may underemphasize the difference between them. A wider range of identifiability scores could be present if only using, e.g., 100 random edges. This might similarly better distinguish some of the preprocessing pipelines tested in chapter 2. Pipelines that regressed HMP and used censoring all had high (~90%) match rates, despite varying in the use of GSR, the exact censoring parameters, and whether they used bandpass or highpass filtering. However, a random connectome subset approach needs to be validated, as it may instead only shift results down without increasing distinguishability. While random edge selection could decrease the power of analysis, if sufficient sampling occurs (e.g., 10 000 random samples), this strategy could still validate the whole connectome.

5.2.5 Alternative measures of identifiability

It is also possible to assess identifiability using different approaches than the comparison of vectorized connectomes. In MEG, high identifiability has been observed based on how tasks

affect the power spectrum and by using the correlation between timepoints (rather than between spatial nodes; Wu et al., 2022). A potential fMRI analogue is ISC (Hasson et al., 2004). My work in Chapter 2 suggests that while there is a gap between within- and between-subject correlations, it is relatively minor, which may hinder its potential for fingerprinting. Nevertheless, it may remain an appropriate benchmark if higher identifiability can be achieved with noise mitigation, which has not been well studied. Task responses may capture individually unique information that is not present in functional connectomes, for e.g., the speed and duration of response to a stimulus. My work here has established the importance of multiple benchmarks; this could be extended to utilizing both connectome-based and ISC-based identifiability.

5.2.6 Other benchmarks

In addition to identifiability, a wide range of other benchmarks have been previously used to validate both preprocessing pipelines and EEG FC measures. Like identifiability, these metrics have strengths and drawbacks, and future validation work that utilizes identifiability should also consider the metrics briefly described here.

While QC-FC is the most commonly used benchmark to test fMRI-FC preprocessing strategies, a number of further approaches have been used. One approach is to divide participants into low- and high-motion groups and calculate the difference in average FC, where it has been argued that pipelines that better account for noise will show fewer between-group differences (Parkes et al., 2018; Pruim et al., 2015; Satterthwaite et al., 2013). Another strategy is to measure test-retest reliability of FC via ICC (Parkes et al., 2018; Van Dijk et al., 2012), a metric that I also integrated into my work. Several studies have considered various network properties, such as network modularity (Ciric et al., 2017), the reproducibility of networks within a participant

(Pruim et al., 2015), and comparing within- and between-network connections (Kassinopoulos and Mitsis, 2021; Pruim et al., 2015). These strategies assume that better preprocessing strategies should better identify consistent brain networks. Finally, several studies have considered the costliness of preprocessing by considering the degrees of freedom lost in confound regression (Ciric et al., 2017; Parkes et al., 2018; Pruim et al., 2015).

EEG FC measures have not been compared as extensively, limiting the number of benchmarks that have been used. Arguably the most comparable study to my work in Chapter 4 comes from MEG, which also considered the self-stability and similarity-to-others of a variety of FC measures (Colclough et al., 2016). They also randomly divided participants into two groups and calculated group-average connectomes, then measured their consistency. Several more benchmarks come from the papers that first proposed novel FC measures. Nolte et al. (2004) demonstrated the usefulness of imcoh by showing the existence of FC between left and right motor areas during a motor task. Stam et al. (2007) compared FC measures on several metrics: sensitivity to synchronization using a model of oscillation, FC differences during seizure, and FC differences comparing Alzheimer's patients and healthy controls. Psi was tested based on false positives when noise was added to simulation data, and by showing connectivity patterns in an eyes open vs closed experiment (Nolte et al., 2008). While not explicitly for the purposes of benchmarking FC measures, several studies have considered the test-retest reliability of EEG FC measured using ICC (Cannon et al., 2012; Hardmeier et al., 2014; Hatz et al., 2016; Kuntzelman and Miskovic, 2017). In addition to considering the reliability of connectivity estimates themselves, these reliability studies also assessed the reliability of graph theory metrics such as clustering coefficient, path length, and small-world index.

5.3 Methodological trade-offs

My work through this thesis emphasizes the importance of considering trade-offs when debating methods to investigate hypotheses. Many tools I have advocated for here have legitimate drawbacks. For example, the global signal contains information related to trait level cognition and behavior (Li et al., 2019) which is lost with GSR. Likewise, GSR changes the distribution of edges, impacting interpretability (Murphy and Fox, 2017). However, my work in chapter 2 suggests the inclusion of GSR can reduce the number of edges correlated with motion by around 50%, without negatively impacting identifiability, strongly backing its inclusion in preprocessing. Ultimately, there is no single correct decision, and different study designs or hypotheses may favor or disfavor GSR's inclusion. My work on censoring similarly shows its advantages, but for study designs that require the comparison of timepoints across participants, such as with ISC, it may be necessary to avoid censoring, or at least, use less stringent censoring thresholds. While consistent methodologies are preferable towards the goal of reproducibility, some degree of flexibility may be necessary to properly account for study-specific trade-offs.

A related consideration is that methods that better remove noise, despite other unwanted characteristics, may allow the inclusion of participants who are otherwise excluded, or allow the study of groups who would otherwise be impossible to investigate. Given the difficulty of studying young children with fMRI (Vanderwal et al., 2019), developmental studies may need to be more willing to adopt imperfect methods that are effective at removing motion noise, regardless of other negative characteristics, particularly when motion associates with the outcome of interest (age or clinical symptoms). While investigations involving large datasets such as the Healthy Brain Network (Alexander et al., 2017) give researchers the opportunity to exclude a large percentage of their sample, such exclusion rates could bias results in other

undesirable ways. For example, if a dataset has 1000 scans from 5-year-olds available, a researcher could choose to only include 5-year-olds who meet stringent criteria for lack of head motion and still be left with a relatively large sample size of 200 participants. However, this could be a poor representation of a typical 5-year-old, making it preferable to instead adopt looser exclusion criteria (including more participants) and utilize stricter noise-removal strategies even if they have drawbacks.

Although time consuming, one possibility for researchers faced with difficult methodology decisions is to first benchmark their specific dataset. If for a given study design avoiding censoring is preferable, they could first measure QC-FC and identifiability to see if their data is sufficiently noise-free based on certain thresholds, e.g., >85% match rates and <25% of edges significantly correlated with motion. This would require researchers to choose appropriate benchmarks that are sufficiently unrelated to their question of interest in order to avoid circularity. Another possibility is that researchers can adopt multiple methodologies simultaneously, seeing if results are consistent across strategies. For example, data could be preprocessed both with and without GSR (Rohr et al., 2019), or using both imcoh and coh as suggested in Nolte et al. (2004).

5.4 EEG vs fMRI connectomes

EEG connectomics can provide complementary information to fMRI connectomics, but they are not strictly comparable. Notably, they fundamentally consider different aspects of brain activity, with EEG directly measuring neural activity, while fMRI measures blood flow which supports neural activity in a non-linear fashion (Hillman, 2014). Previous work has demonstrated a correlation between the two modalities of r ~ 0.3, which was shown with both phase and

amplitude FC measures (Wirsich et al., 2021), though other work has found lower similarity values (Nentwich et al., 2020). This suggests that while there are likely common features, there is a large amount of unexplained variance between modalities. While some of these differences are no doubt due to fundamental differences in neural and vascular activity, EEG also measures at a much faster timescale, capturing different patterns of activity. Relatedly, while EEG source localization algorithms can be used to measure brain activity with the same parcellations as in fMRI, the inherently decreased spatial resolution results in much noisier estimates of FC between regions. Together, this suggests that certain connectome features or developmental changes may be modality specific, both due to the signal of interest they are able to capture and having different noise vulnerabilities.

My work here demonstrates that individually unique information is available in both fMRI and EEG connectomes. While I found much higher match rates and individualization scores with fMRI, these findings should not be overinterpreted. I used fewer, larger parcels in my EEG study (68 in chapter 4, vs 325 in chapter 2 and 1095 in chapter 3), and previous work suggests coarser parcellations result in lower match rates (Byrge and Kennedy, 2019; Finn et al., 2015). However, while studies have utilized higher resolution parcellations with EEG (Nentwich et al., 2020; Wirsich et al., 2021), at least one study suggests EEG has a resolution limit of around 70 parcels (Farahibozorg et al., 2018), meaning that there may not be identifiability benefits with higher resolution EEG parcellations. Regardless, my EEG work in chapter 4 also considered match rates both within and across tasks, while in chapters 2 and 3 I only considered within task for fMRI; cross-task match rates have been found to be lower (Vanderwal et al., 2017). Relatedly, each recording in chapter 4 had many more potential false matches, potentially lowering fingerprinting rates. There are thus some reasons to suggest my EEG identifiability

scores were deflated, and true identifiability may be comparable in both modalities, at least for

FC measures such as plv with high fingerprinting success. Future work can better establish how

fMRI and EEG connectomes compare in the extent to which they capture individually unique

information.

5.5 Limitations

There are several limitations to my work in this thesis. In each study chapter, I only used

one dataset. As thus, my work may not generalize to other study populations or other

methodologies, and it is worth considering how each design decision may influence findings. For

example, in Chapter 2 I found a censoring threshold of ~0.25 mm FD to be optimal, but this

could change based on who is scanned with what parameters. For participants who move less, a

stricter threshold may be more appropriate if the quantity of data remaining post-censoring is

comparable. Similarly, my censoring analysis was based off of a TR of 2.5 s. In a multislice

sequence with a faster TR, a stricter threshold may be necessary to remove artifacts of the same

absolute size. While it is important in this way to consider the impact that every choice may

have, here I will focus on two broad decisions: video tasks, and parcellation choice.

5.5.1 Video tasks

Through all three body chapters of this thesis, I used passive viewing tasks. While

passive viewing has the advantage of increasing compliance in the scanner, reducing head

motion (Greene et al., 2018), the effects it has on FC should be noted. While passive viewing

connectomes share marked similarities with resting state connectomes, especially with regards to

network structure (Greene et al., 2018), passive viewing still has task-evoked properties (Bray et

al., 2014; Vanderwal et al., 2019). This suggests that resting state datasets should not necessarily be used for replication or even comparison (Vanderwal et al., 2019). This could be worth taking into account when comparing my work to others; for example, Sato et al. (2021) found lower match rates in a similar age range, but using resting state data, potentially limiting comparability. Different passive viewing tasks also result in different FC patterns (Vanderwal et al., 2019), which may mean that results gathered using one specific video may not generalize to other viewing tasks. This leaves open the possibility that the specific passive viewing task used in chapter 3 could be responsible for discrepancies with other research. Future investigations into the effects of age on identifiability could use multiple videos to increase confidence in findings, as I did in chapter 4. One concern with passive viewing is that participants vary in how engaging they find specific videos. Relatedly, participants may have already seen a video, or vary in how familiar they are with the characters in a video, such as Elmo or Dora. This could partially confound my results in chapter 3, as younger children may have found clips from Elmo's World more engaging, or they may have more recently seen Elmo content, resulting in some association between age and FC. Differences in task engagement could also be a driver in my task sensitivity benchmark employed in chapter 4, potentially hampering its effectiveness as a measure of validity. However, it is worth noting that these limitations of passive viewing can also apply to resting state – people's mental state during resting tasks vary, and there may be associations between age and the types of thoughts people have during a scan (including how comfortable they are during the MRI).

5.5.2 Parcellation

Choice of parcellation in particular may limit generalizability. While parcellation is useful as a strategy of dimensionality reduction, it necessarily averages together the activity from separate brain volumes with distinct patterns of activity. Dozens of parcellations have been developed, which vary in both underlying assumptions and in their granularity (Eickhoff et al., 2018), which may limit the transferability of findings. Since parcellations with more nodes seem to better capture individually unique variation in brain activity (Byrge and Kennedy, 2019; Finn et al., 2015), my identifiability findings in Chapter 3 may require similar resolution parcellations to be replicable. For example, I found high individualization in visual networks, contrary to previous works that used lower resolution parcellations, suggesting that the visual network may be relatively homogenous across people at low resolutions but more individually unique information is available at higher resolutions. However, the discrepancy could also be due to the relatively heterogeneous definitions of the canonical networks between parcellations, making it hard to compare, for example, the visual network or visual stream I considered in the 1095 node MIST network with any of the 3 visual networks in the 268 node Shen parcellation (Finn et al., 2015). These reasons, along with other differences in study design, make it difficult to interpret differences in results.

Parcel size could similarly affect my findings in Chapter 4, where I analyzed relatively large parcels in the 68-node Desikan Killiany atlas. A smaller number of larger parcels will mean increased distances between the centers of neighboring parcels. This could reduce the impact of volume conduction artifacts, meaning that it may be more important to mitigate against them when using a higher resolution parcellation. On the other hand, smaller parcels necessitate averaging the signal from fewer voxels, which may mean less signal of interest is captured

within each node. This could disadvantage measures like imcoh that already capture less signal in the first place. As thus, it is not clear how parcellation choice affects the validity of EEG FC measures.

Choice of parcellation could also more broadly affect the extent to which identifiability is an appropriate measure of validity. In order to be a useful benchmark for functional neuroimaging datasets, identifiability should be based on patterns of brain activity, rather than individually specific spatial properties. It is typically necessary to register all brains into a common space; however, brain sizes and positions do not have a 1-to-1 correspondence across participants (Klein et al., 2009). Parcellations with fewer, larger parcels may be less vulnerable to misalignment between participants since each parcel averages data from a larger spatial region that better overlaps between participants. Similarly, since EEG has poor spatial resolution, larger parcels may better ensure correspondence between participants. In either case, if parcels poorly align between participants then higher identifiability could be achieved without increased data validity.

5.6 Future directions

In this thesis I demonstrated the utility of imperfect preprocessing pipelines and FC measures. While it is important to be willing to accept trade-offs, future work should still attempt to find better ways to reduce noise while retaining signal, shifting the balance on the trade-off. This could exist in the form of newer processing strategies, for example a version of GSR with fewer drawbacks (perhaps by decomposing the global signal into noise and signal of interest) or better EEG source localization algorithms. Further, work could be done to reduce artifact at time

of data collection, such as testing a wide range of strategies to reduce in-scanner motion or better

EEG caps that are less sensitive to volume conduction.

The work in this thesis establishes identifiability as a measure of validity. This will

hopefully be adopted for a variety of studies going forward. For example, Sadaghiani et al.

(2022), outlines several areas of future EEG research. These include the development of models

of connectivity that consider "a diversity of signal components across the broad frequency

spectrum" along with better consideration of cross-frequency interactions. These goals may

require developing new measures of connectivity that could be validated with identifiability.

Similarly, machine learning algorithms are currently a popular approach for analyzing resting

state fMRI (Khosla et al., 2019). Considering the diversity of machine learning approaches and

their relative novelty, their validity could be better tested with identifiability, e.g., to compare

clustering algorithms or to test feature selection approaches.

The broader goal of my thesis is to improve the quality of neuroimaging data analysis,

leading to a better understanding of neurodevelopment. In my work I benchmarked different

methodologies, but a big remaining question is why different preprocessing strategies or FC

measures are better than others. For example, while my work suggests that coh is more valid than

imcoh despite volume conduction, I did not determine what features of the connectome coh

managed to capture that disadvantaged imcoh. Determining the exact advantages and

disadvantages of coh and imcoh could better determine when to use one over the other, or if the

data from both could be combined in a meaningful way. Further, given the broad collection of

methods I tested, a deeper understanding of how they are sensitive to noise and how they capture

signal could reveal major blind spots, pointing towards future methodologies that better explain

features of interest. These may require simulations of, e.g., how brain data is detected by EEG

(Barzegaran et al., 2019), or a better understanding of the ground truth of neurovascular coupling (Phillips et al., 2016).

5.7 Significance

The work in this thesis has direct implications for how connectome research with children is carried out. My work in Chapter 2 suggests preferable preprocessing techniques and my work in Chapter 4 has implications for which FC measures to use when assessing EEG connectomes. I have here established identifiability as a benchmark of validity, which can be adopted for future benchmarking studies.

More broadly, the significance of this thesis is in more deeply considering the validity of neuroimaging methods. Neuroimaging studies have suffered from a lack of replicability when studying associations between brain data and more complicated phenotypes, such as neurodiverse conditions or development (Marek et al., 2022). This suggests a lack of reliability or validity, requiring reconsideration of many assumptions.

In the absence of ground truth, it is difficult to argue for or against preprocessing strategies, FC measures, or other analysis decisions based solely on a theoretical framework of how these techniques might affect the data. Techniques need to be validated, and validation needs to occur in a variety of datasets with a range of benchmarks. Techniques established in adults should not be assumed to also work in children. My work here also highlights the importance of not thinking about techniques in a valid-invalid dichotomy; certain methods may be valid for exploring certain hypotheses but not others, or they may have drawbacks while still being better than existing alternatives. Ideally, future neuroimaging methodologies will be able

to capture brain function more reliably with minimal noise, but given current techniques we should be prepared to embrace imperfect strategies.

5.8 Thesis conclusions

In this thesis, I compared fMRI preprocessing strategies and EEG FC measures, using identifiability as a benchmark of validity. For noisy, early childhood datasets, my work here suggests fMRI preprocessing strategies that adopt GSR and censoring are preferable, despite the drawbacks to these strategies. Similarly, EEG FC measures that do not mitigate volume conduction, while vulnerable to artifact, may better capture signal of interest (i.e., stable individual differences). I also considered identifiability as an aspect of neurodevelopment, finding that identifiability increases across early childhood as connectomes become more self-stable. The work in this thesis is important towards the broader goal of improving reproducibility in neuroimaging research, by emphasizing the importance of validating methodologies and considering trade-offs between signal retained and noise removed.

Bibliography

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The inferential impact of global signal covariates in functional neuroimaging analyses. Neuroimage 8, 302–306. https://doi.org/10.1006/nimg.1998.0367

Alcauter, S., Lin, W., Keith Smith, J., Gilmore, J.H., Gao, W., 2015. Consistent anterior-posterior segregation of the insula during the first 2 years of life. Cereb Cortex 25, 1176–1187. https://doi.org/10.1093/cercor/bht312

Alexander, L.M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., Kovacs, M., Litke, S., O'Hagan, B., Andersen, J., Bronstein, B., Bui, A., Bushey, M., Butler, H., Castagna, V., Camacho, N., Chan, E., Citera, D., Clucas, J., Cohen, S., Dufek, S., Eaves, M., Fradera, B., Gardner, J., Grant-Villegas, N., Green, G., Gregory, C., Hart, E., Harris, S., Horton, M., Kahn, D., Kabotyanski, K., Karmel, B., Kelly, S.P., Kleinman, K., Koo, B., Kramer, E., Lennon, E., Lord, C., Mantello, G., Margolis, A., Merikangas, K.R., Milham, J., Minniti, G., Neuhaus, R., Levine, A., Osman, Y., Parra, L.C., Pugh, K.R., Racanello, A., Restrepo, A., Saltzman, T., Septimus, B., Tobe, R., Waltz, R., Williams, A., Yeo, A., Castellanos, F.X., Klein, A., Paus, T., Leventhal, B.L., Craddock, R.C., Koplewicz, H.S., Milham, M.P., 2017. An open resource for transdiagnostic research in pediatric mental health and learning disorders. Sci Data 4, 170181. https://doi.org/10.1038/sdata.2017.181

Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C., 2011. A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage 54, 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025

Ball, G., Aljabar, P., Zebari, S., Tusor, N., Arichi, T., Merchant, N., Robinson, E.C., Ogundipe, E., Rueckert, D., Edwards, A.D., Counsell, S.J., 2014. Rich-club organization of the newborn human brain. Proc Natl Acad Sci U S A 111, 7456–7461. https://doi.org/10.1073/pnas.1324118111

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J.M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A.Z., Van Essen, D.C., WU-Minn HCP Consortium, 2013. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage 80, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Bassett, D.S., Xia, C.H., Satterthwaite, T.D., 2018. Understanding the Emergence of Neuropsychiatric Disorders With Network Neuroscience. Biol Psychiatry Cogn Neurosci Neuroimaging 3, 742–753. https://doi.org/10.1016/j.bpsc.2018.03.015

Behzadi, Y., Restom, K., Liau, J., Liu, T.T., 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage 37, 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042

Bijsterbosch, J.D., Valk, S.L., Wang, D., Glasser, M.F., 2021. Recent developments in representations of the connectome. Neuroimage 243, 118533. https://doi.org/10.1016/j.neuroimage.2021.118533

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magnetic Resonance in Medicine 34, 537–541. https://doi.org/10.1002/mrm.1910340409

Bray, S., Arnold, A.E.G.F., Levy, R.M., Iaria, G., 2014. Spatial and temporal functional connectivity changes between resting and attentive states. Hum Brain Mapp 36, 549–565. https://doi.org/10.1002/hbm.22646

Bressler, S.L., Menon, V., 2010. Large-scale brain networks in cognition: emerging methods and principles. Trends Cogn Sci 14, 277–290. https://doi.org/10.1016/j.tics.2010.04.004

Bridgeford, E.W., Wang, S., Wang, Z., Xu, T., Craddock, C., Dey, J., Kiar, G., Gray-Roncal, W., Colantuoni, C., Douville, C., Noble, S., Priebe, C.E., Caffo, B., Milham, M., Zuo, X.-N., Consortium for Reliability and Reproducibility, Vogelstein, J.T., 2021. Eliminating accidental deviations to minimize generalization error and maximize replicability: Applications in connectomics and genomics. PLoS Comput Biol 17, e1009279. https://doi.org/10.1371/journal.pcbi.1009279

Büchel, D., Lehmann, T., Sandbakk, Ø., Baumeister, J., 2021. EEG-derived brain graphs are reliable measures for exploring exercise-induced changes in brain networks. Sci Rep 11, 20803. https://doi.org/10.1038/s41598-021-00371-x

Byrge, L., Kennedy, D.P., 2019. High-accuracy individual identification using a "thin slice" of the functional connectome. Netw Neurosci 3, 363–383. https://doi.org/10.1162/netn_a_00068

Calabro, F.J., Murty, V.P., Jalbrzikowski, M., Tervo-Clemmens, B., Luna, B., 2020. Development of Hippocampal–Prefrontal Cortex Interactions through Adolescence. Cereb Cortex 30, 1548–1558. https://doi.org/10.1093/cercor/bhz186

Cannon, R.L., Baldwin, D.R., Shaw, T.L., Diloreto, D.J., Phillips, S.M., Scruggs, A.M., Riehl, T.C., 2012. Reliability of quantitative EEG (qEEG) measures and LORETA current source density at 30 days. Neuroscience Letters 518, 27–31. https://doi.org/10.1016/j.neulet.2012.04.035

Cao, J., Zhao, Y., Shan, X., Wei, H.-L., Guo, Y., Chen, L., Erkoyuncu, J.A., Sarrigiannis, P.G., 2022. Brain functional and effective connectivity based on electroencephalography recordings: A review. Hum Brain Mapp 43, 860–879. https://doi.org/10.1002/hbm.25683

Carp, J., 2012. The secret lives of experiments: methods reporting in the fMRI literature. Neuroimage 63, 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004

Chai, X.J., Castañón, A.N., Ongür, D., Whitfield-Gabrieli, S., 2012. Anticorrelations in resting state networks without global signal regression. Neuroimage 59, 1420–1428. https://doi.org/10.1016/j.neuroimage.2011.08.048

Churchill, N.W., Raamana, P., Spring, R., Strother, S.C., 2017. Optimizing fMRI preprocessing pipelines for block-design tasks as a function of age. NeuroImage, Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies 154, 240–254. https://doi.org/10.1016/j.neuroimage.2017.02.028

Ciric, R., Rosen, A.F.G., Erus, G., Cieslak, M., Adebimpe, A., Cook, P.A., Bassett, D.S., Davatzikos, C., Wolf, D.H., Satterthwaite, T.D., 2018. Mitigating head motion artifact in functional connectivity MRI. Nat Protoc 13, 2801–2826. https://doi.org/10.1038/s41596-018-0065-y

Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., Gur, R.C., Gur, R.E., Bassett, D.S., Satterthwaite, T.D., 2017. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. NeuroImage, Cleaning up the fMRI time series: Mitigating noise with advanced

acquisition and correction strategies 154, 174–187.
https://doi.org/10.1016/j.neuroimage.2017.03.020

Coalson, T.S., Van Essen, D.C., Glasser, M.F., 2018. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. Proc Natl Acad Sci U S A 115, E6356–E6365. https://doi.org/10.1073/pnas.1801582115

Cohen, M.X., 2014. Analyzing Neural Time Series Data: Theory and Practice. https://doi.org/10.7551/mitpress/9609.001.0001

Colclough, G.L., Brookes, M.J., Smith, S.M., Woolrich, M.W., 2015. A symmetric multivariate leakage correction for MEG connectomes. NeuroImage 117, 439–448. https://doi.org/10.1016/j.neuroimage.2015.03.071

Colclough, G.L., Woolrich, M.W., Tewarie, P.K., Brookes, M.J., Quinn, A.J., Smith, S.M., 2016. How reliable are MEG resting-state connectivity metrics? Neuroimage 138, 284–293. https://doi.org/10.1016/j.neuroimage.2016.05.070

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29, 162–173. https://doi.org/10.1006/cbmr.1996.0014

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., Alzheimer's Disease Neuroimaging Initiative, 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. Neuroimage 192, 115–134. https://doi.org/10.1016/j.neuroimage.2019.02.062

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9, 179–194. https://doi.org/10.1006/nimg.1998.0395

Damoiseaux, J.S., Rombouts, S. a. R.B., Barkhof, F., Scheltens, P., Stam, C.J., Smith, S.M., Beckmann, C.F., 2006. Consistent resting-state networks across healthy subjects. Proc Natl Acad Sci U S A 103, 13848–13853. https://doi.org/10.1073/pnas.0601417103

Davey, C.E., Grayden, D.B., Egan, G.F., Johnston, L.A., 2013. Filtering induces correlation in fMRI resting state data. Neuroimage 64, 728–740. https://doi.org/10.1016/j.neuroimage.2012.08.022

Dehaene-Lambertz, G., Spelke, E.S., 2015. The Infancy of the Human Brain. Neuron 88, 93–109. https://doi.org/10.1016/j.neuron.2015.09.026

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Destrieux, C., FISCHL, B., DALE, A., HALGREN, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Dimond, D., Heo, S., Ip, A., Rohr, C.S., Tansey, R., Graff, K., Dhollander, T., Smith, R.E., Lebel, C., Dewey, D., Connelly, A., Bray, S., 2020a. Maturation and interhemispheric asymmetry in neurite density and orientation dispersion in early childhood. Neuroimage 221, 117168. https://doi.org/10.1016/j.neuroimage.2020.117168

Dimond, D., Rohr, C.S., Smith, R.E., Dhollander, T., Cho, I., Lebel, C., Dewey, D., Connelly, A., Bray, S., 2020b. Early childhood development of white matter fiber density and morphology. Neuroimage 210, 116552. https://doi.org/10.1016/j.neuroimage.2020.116552

Dosenbach, N.U.F., Koller, J.M., Earl, E.A., Miranda-Dominguez, O., Klein, R.L., Van, A.N., Snyder, A.Z., Nagel, B.J., Nigg, J.T., Nguyen, A.L., Wesevich, V., Greene, D.J., Fair, D.A., 2017. Real-time motion analytics during brain MRI improve data quality and reduce costs. NeuroImage 161, 80–93. https://doi.org/10.1016/j.neuroimage.2017.08.025

Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L., 2010. Prediction of Individual Brain Maturity Using fMRI. Science 329, 1358–1361. https://doi.org/10.1126/science.1194144

Dufford, A.J., Noble, S., Gao, S., Scheinost, D., 2021. The instability of functional connectomes across the first year of life. Developmental Cognitive Neuroscience 51, 101007. https://doi.org/10.1016/j.dcn.2021.101007

Emerson, R.W., Gao, W., Lin, W., 2016. Longitudinal Study of the Emerging Functional Connectivity Asymmetry of Primary Language Regions during Infancy. J Neurosci 36, 10883–10892. https://doi.org/10.1523/JNEUROSCI.3980-15.2016

Engel, A.K., Gerloff, C., Hilgetag, C.C., Nolte, G., 2013. Intrinsic Coupling Modes: Multiscale Interactions in Ongoing Brain Activity. Neuron 80, 867–886. https://doi.org/10.1016/j.neuron.2013.09.038

Engelhardt, L.E., Roe, M.A., Juranek, J., DeMaster, D., Harden, K.P., Tucker-Drob, E.M., Church, J.A., 2017. Children's head motion during fMRI tasks is heritable and stable over time. Dev Cogn Neurosci 25, 58–68. https://doi.org/10.1016/j.dcn.2017.01.011

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S.S., Wright, J., Durnez, J., Poldrack, R.A., Gorgolewski, K.J., 2019. FMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods 16, 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fair, D.A., Cohen, A.L., Dosenbach, N.U.F., Church, J.A., Miezin, F.M., Barch, D.M., Raichle, M.E., Petersen, S.E., Schlaggar, B.L., 2008. The maturing architecture of the brain's default network. Proc Natl Acad Sci U S A 105, 4028–4032. https://doi.org/10.1073/pnas.0800376105

Fair, D.A., Cohen, A.L., Power, J.D., Dosenbach, N.U.F., Church, J.A., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2009. Functional brain networks develop from a "local to distributed" organization. PLoS Comput Biol 5, e1000381. https://doi.org/10.1371/journal.pcbi.1000381

Fair, D.A., Nigg, J.T., Iyer, S., Bathula, D., Mills, K.L., Dosenbach, N.U.F., Schlaggar, B.L., Mennes, M., Gutman, D., Bangaru, S., Buitelaar, J.K., Dickstein, D.P., Di Martino, A., Kennedy, D.N., Kelly, C., Luna, B., Schweitzer, J.B., Velanova, K., Wang, Y.-F., Mostofsky, S., Castellanos, F.X., Milham, M.P., 2012. Distinct neural signatures detected for ADHD subtypes after controlling for micro-movements in resting state functional connectivity MRI data. Front Syst Neurosci 6, 80. https://doi.org/10.3389/fnsys.2012.00080

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci 18, 1664–1671. https://doi.org/10.1038/nn.4135

Fleming, S., Thompson, M., Stevens, R., Heneghan, C., Plüddemann, A., Maconochie, I., Tarassenko, L., Mant, D., 2011. Normal ranges of heart rate and respiratory rate in

children from birth to 18 years of age: a systematic review of observational studies. Lancet 377, 1011–1018. https://doi.org/10.1016/S0140-6736(10)62226-X

Fox, M.D., Snyder, A.Z., Vincent, J.L., Corbetta, M., Van Essen, D.C., Raichle, M.E., 2005. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. Proc Natl Acad Sci U S A 102, 9673–9678. https://doi.org/10.1073/pnas.0504136102

Fox, M.D., Zhang, D., Snyder, A.Z., Raichle, M.E., 2009. The global signal and observed anticorrelated resting state brain networks. J Neurophysiol 101, 3270–3283. https://doi.org/10.1152/jn.90777.2008

Freund, J., Brandmaier, A.M., Lewejohann, L., Kirste, I., Kritzler, M., Krüger, A., Sachser, N., Lindenberger, U., Kempermann, G., 2013. Emergence of individuality in genetically identical mice. Science 340, 756–759. https://doi.org/10.1126/science.1235294

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S., Turner, R., 1996. Movement-related effects in fMRI time-series. Magn Reson Med 35, 346–355. https://doi.org/10.1002/mrm.1910350312

Gao, W., Alcauter, S., Elton, A., Hernandez-Castillo, C.R., Smith, J.K., Ramirez, J., Lin, W., 2015a. Functional Network Development During the First Year: Relative Sequence and Socioeconomic Correlations. Cereb Cortex 25, 2919–2928. https://doi.org/10.1093/cercor/bhu088

Gao, W., Alcauter, S., Smith, J.K., Gilmore, J.H., Lin, W., 2015b. Development of human brain cortical network architecture during infancy. Brain Struct Funct 220, 1173–1186. https://doi.org/10.1007/s00429-014-0710-3

Geng, X., Li, G., Lu, Z., Gao, W., Wang, L., Shen, D., Zhu, H., Gilmore, J.H., 2017. Structural and Maturational Covariance in Early Childhood Brain Development. Cereb Cortex 27, 1795–1807. https://doi.org/10.1093/cercor/bhw022

Gilmore, J.H., Santelli, R.K., Gao, W., 2018. Imaging structural and functional brain development in early childhood. Nat Rev Neurosci 19, 123–137. https://doi.org/10.1038/nrn.2018.1

Glover, G.H., 2011. Overview of functional magnetic resonance imaging. Neurosurg Clin N Am 22, 133–139, vii. https://doi.org/10.1016/j.nec.2010.11.001

González-Madruga, K., Staginnus, M., Fairchild, G., 2022. Alterations in Structural and Functional Connectivity in ADHD: Implications for Theories of ADHD. Curr Top Behav Neurosci 57, 445–481. https://doi.org/10.1007/7854_2022_345

Gordon, E.M., Laumann, T.O., Gilmore, A.W., Newbold, D.J., Greene, D.J., Berg, J.J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J.M., Coalson, R.S., Nguyen, A.L., McDermott, K.B., Shimony, J.S., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., Nelson, S.M., Dosenbach, N.U.F., 2017. Precision Functional Mapping of Individual Human Brains. Neuron 95, 791-807.e7. https://doi.org/10.1016/j.neuron.2017.07.011

Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. Front Neuroinform 5, 13. https://doi.org/10.3389/fninf.2011.00013

Gotts, S.J., Saad, Z.S., Jo, H.J., Wallace, G.L., Cox, R.W., Martin, A., 2013. The perils of global signal regression for group comparisons: a case study of Autism Spectrum Disorders. Front Hum Neurosci 7, 356. https://doi.org/10.3389/fnhum.2013.00356

Graff, K., Tansey, R., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S., 2022a. Benchmarking common preprocessing strategies in early childhood functional connectivity and

intersubject correlation fMRI. Developmental Cognitive Neuroscience 54, 101087. https://doi.org/10.1016/j.dcn.2022.101087

Graff, K., Tansey, R., Rai, S., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S., 2022b. Functional connectomes become more longitudinally self-stable, but not more distinct from others, across early childhood. NeuroImage 258, 119367. https://doi.org/10.1016/j.neuroimage.2022.119367

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M., 2013. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience 7.

Gratton, C., Kraus, B.T., Greene, D.J., Gordon, E.M., Laumann, T.O., Nelson, S.M., Dosenbach, N.U.F., Petersen, S.E., 2020. Defining Individual-Specific Functional Neuroanatomy for Precision Psychiatry. Biol Psychiatry 88, 28–39. https://doi.org/10.1016/j.biopsych.2019.10.026

Gratton, C., Laumann, T.O., Nielsen, A.N., Greene, D.J., Gordon, E.M., Gilmore, A.W., Nelson, S.M., Coalson, R.S., Snyder, A.Z., Schlaggar, B.L., Dosenbach, N.U.F., Petersen, S.E., 2018. Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. Neuron 98, 439-452.e5. https://doi.org/10.1016/j.neuron.2018.03.035

Grayson, D.S., Fair, D.A., 2017. Development of large-scale functional networks from birth to adulthood: A guide to the neuroimaging literature. Neuroimage 160, 15–31. https://doi.org/10.1016/j.neuroimage.2017.01.079

Greene, D.J., Koller, J.M., Hampton, J.M., Wesevich, V., Van, A.N., Nguyen, A.L., Hoyt, C.R., McIntyre, L., Earl, E.A., Klein, R.L., Shimony, J.S., Petersen, S.E., Schlaggar, B.L., Fair, D.A., Dosenbach, N.U.F., 2018. Behavioral interventions for reducing head motion during MRI scans in children. NeuroImage 171, 234–245. https://doi.org/10.1016/j.neuroimage.2018.01.023

Grieve, P.G., Emerson, R.G., Fifer, W.P., Isler, J.R., Stark, R.I., 2003. Spatial correlation of the infant and adult electroencephalogram. Clinical Neurophysiology 114, 1594–1608. https://doi.org/10.1016/S1388-2457(03)00122-6

Gross, J., Kujala, J., Hämäläinen, M., Timmermann, L., Schnitzler, A., Salmelin, R., 2001. Dynamic imaging of coherent sources: Studying neural interactions in the human brain. Proceedings of the National Academy of Sciences 98, 694–699. https://doi.org/10.1073/pnas.98.2.694

Gu, S., Satterthwaite, T.D., Medaglia, J.D., Yang, M., Gur, R.E., Gur, R.C., Bassett, D.S., 2015. Emergence of system roles in normative neurodevelopment. Proceedings of the National Academy of Sciences 112, 13681–13686. https://doi.org/10.1073/pnas.1502829112

Haartsen, R., van der Velde, B., Jones, E.J.H., Johnson, M.H., Kemner, C., 2020. Using multiple short epochs optimises the stability of infant EEG connectivity parameters. Sci Rep 10, 12703. https://doi.org/10.1038/s41598-020-68981-5

Hardmeier, M., Hatz, F., Bousleiman, H., Schindler, C., Stam, C.J., Fuhr, P., 2014. Reproducibility of Functional Connectivity and Graph Measures Based on the Phase Lag Index (PLI) and Weighted Phase Lag Index (wPLI) Derived from High Resolution EEG. PLOS ONE 9, e108648. https://doi.org/10.1371/journal.pone.0108648

Harrewijn, A., Abend, R., Linke, J., Brotman, M.A., Fox, N.A., Leibenluft, E., Winkler, A.M., Pine, D.S., 2020. Combining fMRI during resting state and an attention bias task in children. Neuroimage 205, 116301. https://doi.org/10.1016/j.neuroimage.2019.116301

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., Malach, R., 2004. Intersubject Synchronization of Cortical Activity During Natural Vision. Science 303, 1634–1640. https://doi.org/10.1126/science.1089506

Hatz, F., Hardmeier, M., Bousleiman, H., Rüegg, S., Schindler, C., Fuhr, P., 2016. Reliability of Functional Connectivity of Electroencephalography Applying Microstate-Segmented Versus Classical Calculation of Phase Lag Index. Brain Connectivity 6, 461–469. https://doi.org/10.1089/brain.2015.0368

Hillman, E.M.C., 2014. Coupling mechanism and significance of the BOLD signal: a status report. Annu Rev Neurosci 37, 161–181. https://doi.org/10.1146/annurev-neuro-071013-014111

Hipp, J.F., Hawellek, D.J., Corbetta, M., Siegel, M., Engel, A.K., 2012. Large-scale cortical correlation structure of spontaneous oscillatory activity. Nat Neurosci 15, 10.1038/nn.3101. https://doi.org/10.1038/nn.3101

Hodgson, K., Poldrack, R.A., Curran, J.E., Knowles, E.E., Mathias, S., Göring, H.H.H., Yao, N., Olvera, R.L., Fox, P.T., Almasy, L., Duggirala, R., Barch, D.M., Blangero, J., Glahn, D.C., 2017. Shared Genetic Factors Influence Head Motion During MRI and Body Mass Index. Cereb Cortex 27, 5539–5546. https://doi.org/10.1093/cercor/bhw321

Horien, C., Noble, S., Finn, E.S., Shen, X., Scheinost, D., Constable, R.T., 2018. Considering factors affecting the connectome-based identification process: Comment on Waller et al. Neuroimage 169, 172–175. https://doi.org/10.1016/j.neuroimage.2017.12.045

Houck, J.M., Claus, E.D., 2020. A comparison of automated and manual co-registration for magnetoencephalography. PLOS ONE 15, e0232100. https://doi.org/10.1371/journal.pone.0232100

Huang, C.-M., Lee, S.-H., Hsiao, I.-T., Kuan, W.-C., Wai, Y.-Y., Ko, H.-J., Wan, Y.-L., Hsu, Y.-Y., Liu, H.-L., 2010. Study-specific EPI template improves group analysis in functional MRI of young and older adults. J Neurosci Methods 189, 257–266. https://doi.org/10.1016/j.jneumeth.2010.03.021

Hutchison, R.M., Morton, J.B., 2015. Tracking the Brain's Functional Coupling Dynamics over Development. J Neurosci 35, 6849–6859. https://doi.org/10.1523/JNEUROSCI.4638-14.2015

Jalbrzikowski, M., Liu, F., Foran, W., Klei, L., Calabro, F.J., Roeder, K., Devlin, B., Luna, B., 2020. Functional connectome fingerprinting accuracy in youths and adults is similar when examined on the same day and 1.5-years apart. Human Brain Mapping 41, 4187–4199. https://doi.org/10.1002/hbm.25118

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841. https://doi.org/10.1016/s1053-8119(02)91132-8

Jo, H.J., Saad, Z.S., Simmons, W.K., Milbury, L.A., Cox, R.W., 2010. Mapping sources of correlation in resting state FMRI, with artifact detection and removal. Neuroimage 52, 571–582. https://doi.org/10.1016/j.neuroimage.2010.04.246

Jung, J., Choi, S., Han, K.-M., Kim, A., Kang, W., Paik, J.-W., Lee, H.-W., Ham, B.-J., 2020. Alterations in functional brain networks in depressed patients with a suicide attempt history. Neuropsychopharmacology 45, 964–974. https://doi.org/10.1038/s41386-019-0560-z

Kassinopoulos, M., Mitsis, G.D., 2021. A multi-measure approach for assessing the performance of fMRI preprocessing strategies in resting-state functional connectivity. https://doi.org/10.1101/837609

Kaufmann, T., Alnæs, D., Doan, N.T., Brandt, C.L., Andreassen, O.A., Westlye, L.T., 2017. Delayed stabilization and individualization in connectome development are related to psychiatric disorders. Nat Neurosci 20, 513–515. https://doi.org/10.1038/nn.4511

Kauppi, J.-P., Jääskeläinen, I.P., Sams, M., Tohka, J., 2010. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Front Neuroinform 4, 5. https://doi.org/10.3389/fninf.2010.00005

Kim, S.-G., 2018. Biophysics of BOLD fMRI investigated with animal models. J Magn Reson 292, 82–89. https://doi.org/10.1016/j.jmr.2018.04.006

Kuntzelman, K., Miskovic, V., 2017. Reliability of graph metrics derived from resting-state human EEG. Psychophysiology 54, 51–61. https://doi.org/10.1111/psyp.12600

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., Turner, R., 1992. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proc Natl Acad Sci U S A 89, 5675–5679. https://doi.org/10.1073/pnas.89.12.5675

Lachaux, J.-P., Rodriguez, E., Martinerie, J., Varela, F.J., 1999. Measuring phase synchrony in brain signals. Human Brain Mapping 8, 194–208. https://doi.org/10.1002/(SICI)1097-0193(1999)8:4<194::AID-HBM4>3.0.CO;2-C

Levin, A.R., Naples, A.J., Scheffler, A.W., Webb, S.J., Shic, F., Sugar, C.A., Murias, M., Bernier, R.A., Chawarska, K., Dawson, G., Faja, S., Jeste, S., Nelson, C.A., McPartland, J.C., Şentürk, D., 2020. Day-to-Day Test-Retest Reliability of EEG Profiles in Children With Autism Spectrum Disorder and Typical Development. Front Integr Neurosci 14, 21. https://doi.org/10.3389/fnint.2020.00021

Li, A., Feitelberg, J., Saini, A.P., Höchenberger, R., Scheltienne, M., 2022. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. Journal of Open Source Software 7, 4484. https://doi.org/10.21105/joss.04484

Li, J., Bolt, T., Bzdok, D., Nomi, J.S., Yeo, B.T.T., Spreng, R.N., Uddin, L.Q., 2019. Topography and behavioral relevance of the global signal in the human brain. Sci Rep 9, 14286. https://doi.org/10.1038/s41598-019-50750-8

Li, X., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., Clucas, J., Franco, A., Heinsfeld, A.S., Adebimpe, A., Vogelstein, J.T., Yan, C.-G., Esteban, O., Poldrack, R.A., Craddock, C., Fair, D., Satterthwaite, T., Kiar, G., Milham, M.P., 2021. Moving Beyond Processing and Analysis-Related Variation in Neuroscience. https://doi.org/10.1101/2021.12.01.470790

Liao, Z., Banaschewski, T., Bokde, A.L.W., Desrivières, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Martinot, J.-L., Martinot, M.-L.P., Artiges, E., Nees, F., Orfanos, D.P., Poustka, L., Hohmann, S., Millenet, S., Fröhner, J.H., Smolka, M.N., Walter, H., Whelan, R., Schumann, G., Paus, T., 2021. Similarity and stability of face network across populations and throughout adolescence and adulthood. NeuroImage 244, 118587. https://doi.org/10.1016/j.neuroimage.2021.118587

Lin, W., Zhu, Q., Gao, W., Chen, Y., Toh, C.-H., Styner, M., Gerig, G., Smith, J.K., Biswal, B., Gilmore, J.H., 2008. Functional connectivity MR imaging reveals cortical functional connectivity in the developing brain. AJNR Am J Neuroradiol 29, 1883–1889. https://doi.org/10.3174/ajnr.A1256

Lindquist, M.A., Geuter, S., Wager, T.D., Caffo, B.S., 2019. Modular preprocessing pipelines can reintroduce artifacts into fMRI data. Hum Brain Mapp 40, 2358–2376. https://doi.org/10.1002/hbm.24528

Logothetis, N.K., 2002. The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. Philos Trans R Soc Lond B Biol Sci 357, 1003–1037. https://doi.org/10.1098/rstb.2002.1114

Lopez, K.L., Monachino, A.D., Vincent, K.M., Peck, F.C., Gabard-Durnam, L.J., 2023. Stability, change, and reliable individual differences in electroencephalography measures: A lifespan perspective on progress and opportunities. NeuroImage 275, 120116. https://doi.org/10.1016/j.neuroimage.2023.120116

Mahjoory, K., Nikulin, V.V., Botrel, L., Linkenkaer-Hansen, K., Fato, M.M., Haufe, S., 2017. Consistency of EEG source localization and connectivity estimates. NeuroImage 152, 590–601. https://doi.org/10.1016/j.neuroimage.2017.02.076

Marek, S., Dosenbach, N.U.F., 2018. The frontoparietal network: function, electrophysiology, and importance of individual precision mapping. Dialogues Clin Neurosci 20, 133–140. https://doi.org/10.31887/DCNS.2018.20.2/smarek

Marek, S., Hwang, K., Foran, W., Hallquist, M.N., Luna, B., 2015. The Contribution of Network Organization and Integration to the Development of Cognitive Control. PLoS Biol 13, e1002328. https://doi.org/10.1371/journal.pbio.1002328

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., Malone, S.M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A.M., Earl, E.A., Perrone, A.J., Cordova, M., Doyle, O., Moore, L.A., Conan, G.M., Uriarte, J., Snider, K., Lynch, B.J., Wilgenbusch, J.C., Pengo, T., Tam, A., Chen, J., Newbold, D.J., Zheng, A., Seider, N.A., Van, A.N., Metoki, A., Chauvin, R.J., Laumann, T.O., Greene, D.J., Petersen, S.E., Garavan, H., Thompson, W.K., Nichols, T.E., Yeo, B.T.T., Barch, D.M., Luna, B., Fair, D.A., Dosenbach, N.U.F., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660. https://doi.org/10.1038/s41586-022-04492-9

Mevel, K., Fransson, P., 2016. The functional brain connectome of the child and autism spectrum disorders. Acta Paediatr 105, 1024–1035. https://doi.org/10.1111/apa.13484

Miranda-Dominguez, O., Feczko, E., Grayson, D.S., Walum, H., Nigg, J.T., Fair, D.A., 2018. Heritability of the human connectome: A connectotyping study. Netw Neurosci 2, 175–199. https://doi.org/10.1162/netn_a_00029

Miranda-Dominguez, O., Mills, B.D., Carpenter, S.D., Grant, K.A., Kroenke, C.D., Nigg, J.T., Fair, D.A., 2014. Connectotyping: model based fingerprinting of the functional connectome. PLoS One 9, e111048. https://doi.org/10.1371/journal.pone.0111048

Moraczewski, D., Chen, G., Redcay, E., 2018. Inter-subject synchrony as an index of functional specialization in early childhood. Sci Rep 8, 2252. https://doi.org/10.1038/s41598-018-20600-0

Mostame, P., Sadaghiani, S., 2020. Phase- and amplitude-coupling are tied by an intrinsic spatial organization but show divergent stimulus-related changes. NeuroImage 219, 117051. https://doi.org/10.1016/j.neuroimage.2020.117051

Murphy, K., Fox, M.D., 2017. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. Neuroimage 154, 169–173. https://doi.org/10.1016/j.neuroimage.2016.11.052

Nentwich, M., Ai, L., Madsen, J., Telesford, Q.K., Haufe, S., Milham, M.P., Parra, L.C., 2020. Functional connectivity of EEG is subject-specific, associated with phenotype, and different from fMRI. NeuroImage 218, 117001. https://doi.org/10.1016/j.neuroimage.2020.117001

Niazy, R.K., Xie, J., Miller, K., Beckmann, C.F., Smith, S.M., 2011. Spectral characteristics of resting state networks. Prog Brain Res 193, 259–276. https://doi.org/10.1016/B978-0-444-53839-0.00017-X

Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: a systematic review and meta-analysis. Neuroimage 203, 116157. https://doi.org/10.1016/j.neuroimage.2019.116157

Noble, S., Spann, M.N., Tokoglu, F., Shen, X., Constable, R.T., Scheinost, D., 2017. Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. Cereb Cortex 27, 5415–5429. https://doi.org/10.1093/cercor/bhx230

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M., 2004. Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical Neurophysiology 115, 2292–2307. https://doi.org/10.1016/j.clinph.2004.04.029

Nolte, G., Ziehe, A., Nikulin, V.V., Schlögl, A., Krämer, N., Brismar, T., Müller, K.-R., 2008. Robustly estimating the flow direction of information in complex physical systems. Phys Rev Lett 100, 234101. https://doi.org/10.1103/PhysRevLett.100.234101

Nunez, P.L., Srinivasan, R., Westdorp, A.F., Wijesinghe, R.S., Tucker, D.M., Silberstein, R.B., Cadusch, P.J., 1997. EEG coherency: I: statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. Electroencephalography and Clinical Neurophysiology 103, 499–515. https://doi.org/10.1016/S0013-4694(97)00066-7

Ogawa, S., Lee, T.M., Kay, A.R., Tank, D.W., 1990. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc Natl Acad Sci U S A 87, 9868–9872.

O'Neill, G.C., Tewarie, P., Vidaurre, D., Liuzzi, L., Woolrich, M.W., Brookes, M.J., 2018. Dynamics of large-scale electrophysiological networks: A technical review. NeuroImage, Brain Connectivity Dynamics 180, 559–576. https://doi.org/10.1016/j.neuroimage.2017.10.003

Palva, J.M., Wang, S.H., Palva, S., Zhigalov, A., Monto, S., Brookes, M.J., Schoffelen, J.-M., Jerbi, K., 2018. Ghost interactions in MEG/EEG source space: A note of caution on inter-areal coupling measures. NeuroImage 173, 632–643. https://doi.org/10.1016/j.neuroimage.2018.02.032

Parkes, L., Fulcher, B., Yücel, M., Fornito, A., 2018. An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. Neuroimage 171, 415–436. https://doi.org/10.1016/j.neuroimage.2017.12.073

Pascual-Marqui, R.D., 2007. Discrete, 3D distributed, linear imaging methods of electric neuronal activity. Part 1: exact, zero error localization. https://doi.org/10.48550/arXiv.0710.3341

Pascual-Marqui, R.D., 2002. Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods Find Exp Clin Pharmacol 24 Suppl D, 5–12.

Patel, A.X., Kundu, P., Rubinov, M., Jones, P.S., Vértes, P.E., Ersche, K.D., Suckling, J., Bullmore, E.T., 2014. A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. Neuroimage 95, 287–304. https://doi.org/10.1016/j.neuroimage.2014.03.012

Peña-Gómez, C., Avena-Koenigsberger, A., Sepulcre, J., Sporns, O., 2018. Spatiotemporal Network Markers of Individual Variability in the Human Functional Connectome. Cerebral Cortex 28, 2922–2934. https://doi.org/10.1093/cercor/bhx170

Phạm, D.Đ., McDonald, D.J., Ding, L., Nebel, M.B., Mejia, A.F., 2023. Less is more: balancing noise reduction and data retention in fMRI with data-driven scrubbing. NeuroImage 270, 119972. https://doi.org/10.1016/j.neuroimage.2023.119972

Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E., 2012. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage 59, 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Power, J.D., Plitt, M., Kundu, P., Bandettini, P.A., Martin, A., 2017. Temporal interpolation alters motion in fMRI scans: Magnitudes and consequences for artifact detection. PLoS One 12, e0182939. https://doi.org/10.1371/journal.pone.0182939

Power, J.D., Schlaggar, B.L., Petersen, S.E., 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage 105, 536–551. https://doi.org/10.1016/j.neuroimage.2014.10.044

Pruim, R.H.R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. Neuroimage 112, 267–277. https://doi.org/10.1016/j.neuroimage.2015.02.064

Raichle, M.E., 2015. The brain's default mode network. Annu Rev Neurosci 38, 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030

Rajapandian, M., Amico, E., Abbas, K., Ventresca, M., Goñi, J., 2020. Uncovering differential identifiability in network properties of human brain functional connectomes. Netw Neurosci 4, 698–713. https://doi.org/10.1162/netn_a_00140

Reynolds, J.E., Long, X., Paniukov, D., Bagshawe, M., Lebel, C., 2020. Calgary Preschool magnetic resonance imaging (MRI) dataset. Data Brief 29, 105224. https://doi.org/10.1016/j.dib.2020.105224

Rohr, C.S., Arora, A., Cho, I.Y.K., Katlariwala, P., Dimond, D., Dewey, D., Bray, S., 2018. Functional network integration and attention skills in young children. Dev Cogn Neurosci 30, 200–211. https://doi.org/10.1016/j.dcn.2018.03.007

Rohr, C.S., Dimond, D., Schuetze, M., Cho, I.Y.K., Lichtenstein-Vidne, L., Okon-Singer, H., Dewey, D., Bray, S., 2019. Girls' attentive traits associate with cerebellar to dorsal attention and default mode network connectivity. Neuropsychologia 127, 84–92. https://doi.org/10.1016/j.neuropsychologia.2019.02.011

Rohr, C.S., Vinette, S.A., Parsons, K.A.L., Cho, I.Y.K., Dimond, D., Benischek, A., Lebel, C., Dewey, D., Bray, S., 2017. Functional Connectivity of the Dorsal Attention Network Predicts Selective Attention in 4-7 year-old Girls. Cereb Cortex 27, 4350–4360. https://doi.org/10.1093/cercor/bhw236

Rosenberg, M.D., Scheinost, D., Greene, A.S., Avery, E.W., Kwon, Y.H., Finn, E.S., Ramani, R., Qiu, M., Constable, R.T., Chun, M.M., 2020. Functional connectivity predicts changes in attention observed across minutes, days, and months. Proceedings of the National Academy of Sciences 117, 3797–3807. https://doi.org/10.1073/pnas.1912226117

Rua, C., Wastling, S.J., Costagli, M., Symms, M.R., Biagi, L., Cosottini, M., Del Guerra, A., Tosetti, M., Barker, G.J., 2018. Improving fMRI in signal drop-out regions at 7 T by using tailored radio-frequency pulses: application to the ventral occipito-temporal cortex. MAGMA 31, 257–267. https://doi.org/10.1007/s10334-017-0652-x

Saad, Z.S., Gotts, S.J., Murphy, K., Chen, G., Jo, H.J., Martin, A., Cox, R.W., 2012. Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. Brain Connect 2, 25–32. https://doi.org/10.1089/brain.2012.0080

Sadaghiani, S., Brookes, M.J., Baillet, S., 2022. Connectomics of human electrophysiology. NeuroImage 247, 118788. https://doi.org/10.1016/j.neuroimage.2021.118788

Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449–468. https://doi.org/10.1016/j.neuroimage.2013.11.046

Sarvas, J., 1987. Basic mathematical and electromagnetic concepts of the biomagnetic inverse problem. Phys. Med. Biol. 32, 11. https://doi.org/10.1088/0031-9155/32/1/004

Sato, J.R., Biazoli Jr, C.E., Zugman, A., Pan, P.M., Bueno, A.P.A., Moura, L.M., Gadelha, A., Picon, F.A., Amaro Jr, E., Salum, G.A., Miguel, E.C., Rohde, L.A., Bressan, R.A., Jackowski, A.P., 2021. Long-term stability of the cortical volumetric profile and the functional human connectome throughout childhood and adolescence. European Journal of Neuroscience 54, 6187–6201. https://doi.org/10.1111/ejn.15435

Satterthwaite, T.D., Ciric, R., Roalf, D.R., Davatzikos, C., Bassett, D.S., Wolf, D.H., 2019. Motion artifact in studies of functional connectivity: Characteristics and mitigation strategies. Hum Brain Mapp 40, 2033–2051. https://doi.org/10.1002/hbm.23665

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., Wolf, D.H., 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage 64, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Satterthwaite, T.D., Wolf, D.H., Loughead, J., Ruparel, K., Elliott, M.A., Hakonarson, H., Gur, R.C., Gur, R.E., 2012. Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. Neuroimage 60, 623–632. https://doi.org/10.1016/j.neuroimage.2011.12.063

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex 28, 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schoffelen, J.-M., Gross, J., 2009. Source connectivity analysis with MEG and EEG. Human Brain Mapping 30, 1857–1865. https://doi.org/10.1002/hbm.20745

Schubert, A.-L., Hagemann, D., Voss, A., Schankin, A., Bergmann, K., 2015. Decomposing the relationship between mental speed and mental abilities. Intelligence 51, 28–46. https://doi.org/10.1016/j.intell.2015.05.002

Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage 82, 403–415. https://doi.org/10.1016/j.neuroimage.2013.05.081

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86, 420–428. https://doi.org/10.1037//0033-2909.86.2.420

Siems, M., Siegel, M., 2020. Dissociated neuronal phase- and amplitude-coupling patterns in the human brain. NeuroImage 209, 116538. https://doi.org/10.1016/j.neuroimage.2020.116538

Smith, S.M., 2002. Fast robust automated brain extraction. Hum Brain Mapp 17, 143–155. https://doi.org/10.1002/hbm.10062

Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., Niazy, R.K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J.M., Matthews, P.M., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1, S208-219. https://doi.org/10.1016/j.neuroimage.2004.07.051

St. Louis, E., Frey, L., University of Colorado, Denver, Colorado, Britton, J., Mayo Clinic College of Medicine, Rochester, Minnesota, Hopp, J., University of Maryland, Baltimore, Maryland, Korb, P., University of Colorado, Denver, Colorado, Koubeissi, M., George Washington University, Washington, District of Columbia, Lievens, W., University of Alabama, Birmingham, Alabama, Pestana-Knight, E., Cleveland Clinic Foundation, Cleveland, Ohio, 2016. Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants. American Epilepsy Society. https://doi.org/10.5698/978-0-9979756-0-4

Stam, C.J., Nolte, G., Daffertshofer, A., 2007. Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. Human Brain Mapping 28, 1178–1193. https://doi.org/10.1002/hbm.20346

Sydnor, V.J., Larsen, B., Bassett, D.S., Alexander-Bloch, A., Fair, D.A., Liston, C., Mackey, A.P., Milham, M.P., Pines, A., Roalf, D.R., Seidlitz, J., Xu, T., Raznahan, A., Satterthwaite, T.D., 2021. Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology. Neuron 109, 2820–2846. https://doi.org/10.1016/j.neuron.2021.06.016

Tansey, R., Graff, K., Rohr, C.S., Dimond, D., Ip, A., Dewey, D., Bray, S., 2022. Inattentive and hyperactive traits differentially associate with interindividual functional synchrony during video viewing in young children without ADHD. Cereb Cortex Commun 3, tgac011. https://doi.org/10.1093/texcom/tgac011

Taymourtash, A., Schwartz, E., Nenning, K.-H., Sobotka, D., Diogo, M., Kasprian, G., Prayer, D., Langs, G., 2020. Quantifying Residual Motion Artifacts in Fetal fMRI Data. https://doi.org/10.48550/arXiv.2001.03739

Thomas, C.G., Harshman, R.A., Menon, R.S., 2002. Noise reduction in BOLD-based fMRI using component analysis. Neuroimage 17, 1521–1537. https://doi.org/10.1006/nimg.2002.1200

Urchs, S., Armoza, J., Moreau, C., Benhajali, Y., St-Aubin, J., Orban, P., Bellec, P., 2019. MIST: A multi-resolution parcellation of functional brain networks. MNI Open Res 1, 3. https://doi.org/10.12688/mniopenres.12767.2

van der Velde, B., Haartsen, R., Kemner, C., 2019. Test-retest reliability of EEG network characteristics in infants. Brain and Behavior 9, e01269. https://doi.org/10.1002/brb3.1269

Van Dijk, K.R.A., Sabuncu, M.R., Buckner, R.L., 2012. The influence of head motion on intrinsic functional connectivity MRI. Neuroimage 59, 431–438. https://doi.org/10.1016/j.neuroimage.2011.07.044

Vanderwal, T., Eilbott, J., Castellanos, F.X., 2019. Movies in the magnet: Naturalistic paradigms in developmental functional neuroimaging. Developmental Cognitive Neuroscience 36, 100600. https://doi.org/10.1016/j.dcn.2018.10.004

Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., Castellanos, F.X., 2017. Individual differences in functional connectivity during naturalistic viewing conditions. Neuroimage 157, 521–530. https://doi.org/10.1016/j.neuroimage.2017.06.027

Vanderwal, T., Eilbott, J., Kelly, C., Frew, S.R., Woodward, T.S., Milham, M.P., Castellanos, F.X., 2021. Stability and similarity of the pediatric connectome as developmental measures. NeuroImage 226, 117537. https://doi.org/10.1016/j.neuroimage.2020.117537

Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L.C., Castellanos, F.X., 2015. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. Neuroimage 122, 222–232. https://doi.org/10.1016/j.neuroimage.2015.07.069

Vendetti, M.S., Bunge, S.A., 2014. Evolutionary and developmental changes in the lateral frontoparietal network: a little goes a long way for higher-level cognition. Neuron 84, 906–917. https://doi.org/10.1016/j.neuron.2014.09.035

Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., Pennartz, C.M.A., 2011. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. NeuroImage 55, 1548–1565. https://doi.org/10.1016/j.neuroimage.2011.01.055

Waller, L., Walter, H., Kruschwitz, J.D., Reuter, L., Müller, S., Erk, S., Veer, I.M., 2017. Evaluating the replicability, specificity, and generalizability of connectome fingerprints. NeuroImage 158, 371–377. https://doi.org/10.1016/j.neuroimage.2017.07.016

Wens, V., Marty, B., Mary, A., Bourguignon, M., Op de Beeck, M., Goldman, S., Van Bogaert, P., Peigneux, P., De Tiège, X., 2015. A geometric correction scheme for spatial leakage effects in MEG/EEG seed-based functional connectivity mapping. Hum Brain Mapp 36, 4604–4621. https://doi.org/10.1002/hbm.22943

Wirsich, J., Jorge, J., Iannotti, G.R., Shamshiri, E.A., Grouiller, F., Abreu, R., Lazeyras, F., Giraud, A.-L., Gruetter, R., Sadaghiani, S., Vulliémoz, S., 2021. The relationship between EEG and fMRI connectomes is reproducible across simultaneous EEG-fMRI studies from 1.5T to 7T. NeuroImage 231, 117864. https://doi.org/10.1016/j.neuroimage.2021.117864

Wu, S., Ramdas, A., Wehbe, L., 2022. Brainprints: identifying individuals from magnetoencephalograms. Commun Biol 5, 1–12. https://doi.org/10.1038/s42003-022-03727-9

Zeng, L.-L., Wang, D., Fox, M.D., Sabuncu, M., Hu, D., Ge, M., Buckner, R.L., Liu, H., 2014. Neurobiological basis of head motion in brain imaging. Proc Natl Acad Sci U S A 111, 6058–6062. https://doi.org/10.1073/pnas.1317424111

Zuo, X.-N., Xu, T., Jiang, L., Yang, Z., Cao, X.-Y., He, Y., Zang, Y.-F., Castellanos, F.X., Milham, M.P., 2013. Toward reliable characterization of functional homogeneity in the human brain: preprocessing, scan duration, imaging resolution and computational space. Neuroimage 65, 374–386. https://doi.org/10.1016/j.neuroimage.2012.10.017
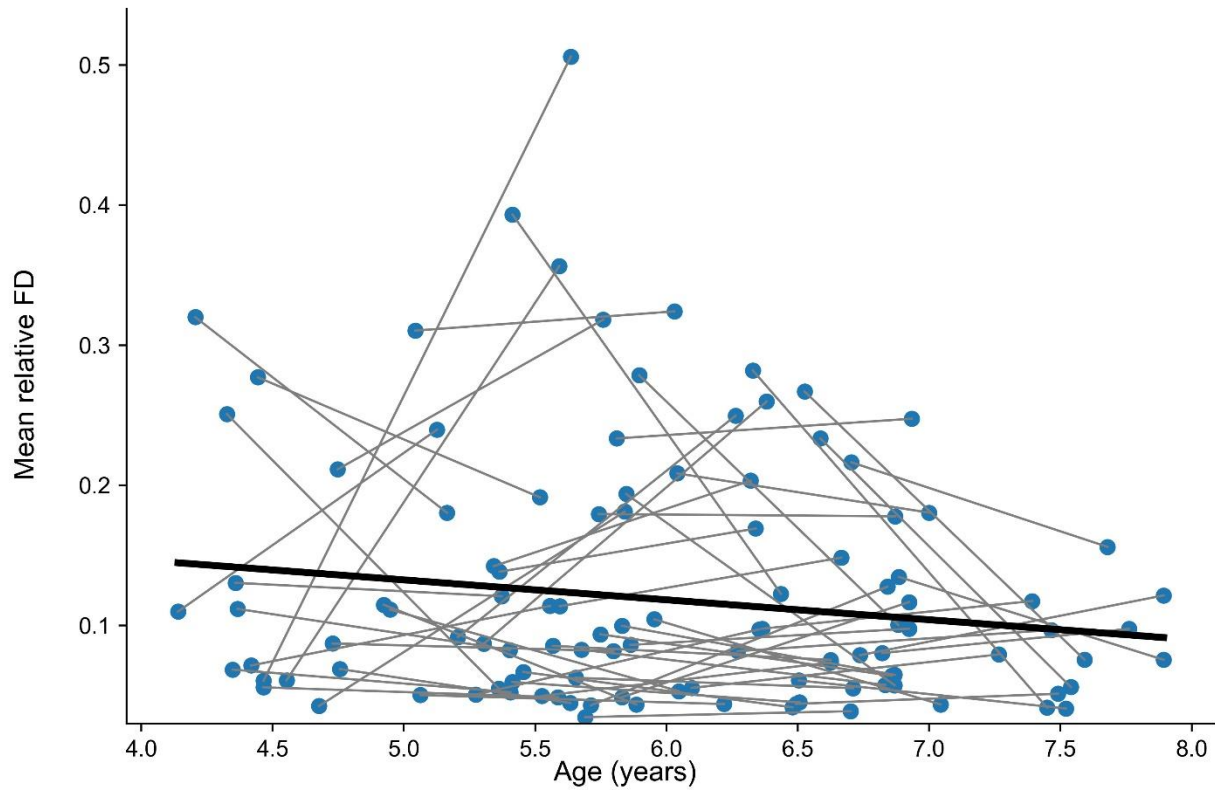
Supplemental Figure 2.1. The mean relative framewise displacement (FD) for each scan, plotted by age of participant at time of scan. Each point represents one scan, with lines between points linking scans from the same participant. Line of best fit derived from a linear mixed model used to predict mean relative FD by age.
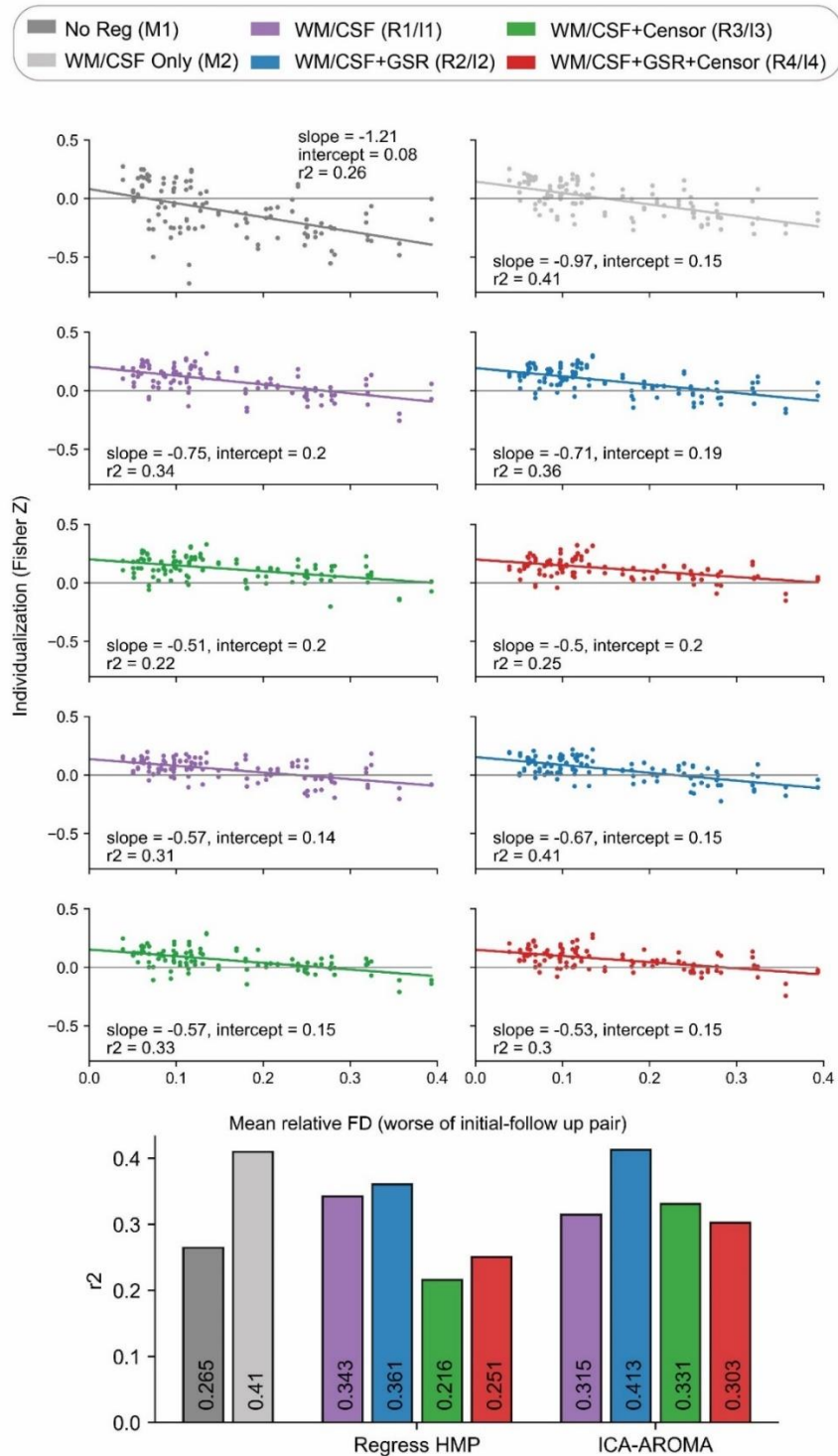
Supplemental Figure 2.2. Individualization as a function of head motion. This is similar to Figure 2.4 with one outlier removed (participant with 0.51 mm FD). The difference between each scan's stability (self-correlation) and the highest correlation to a scan from another participant, plotted by motion. Motion was calculated by taking the worse of each participant's two scans' mean relative framewise displacement; each participant has two points for their two scans. Any point below 0 on the y axis fails to successfully match.
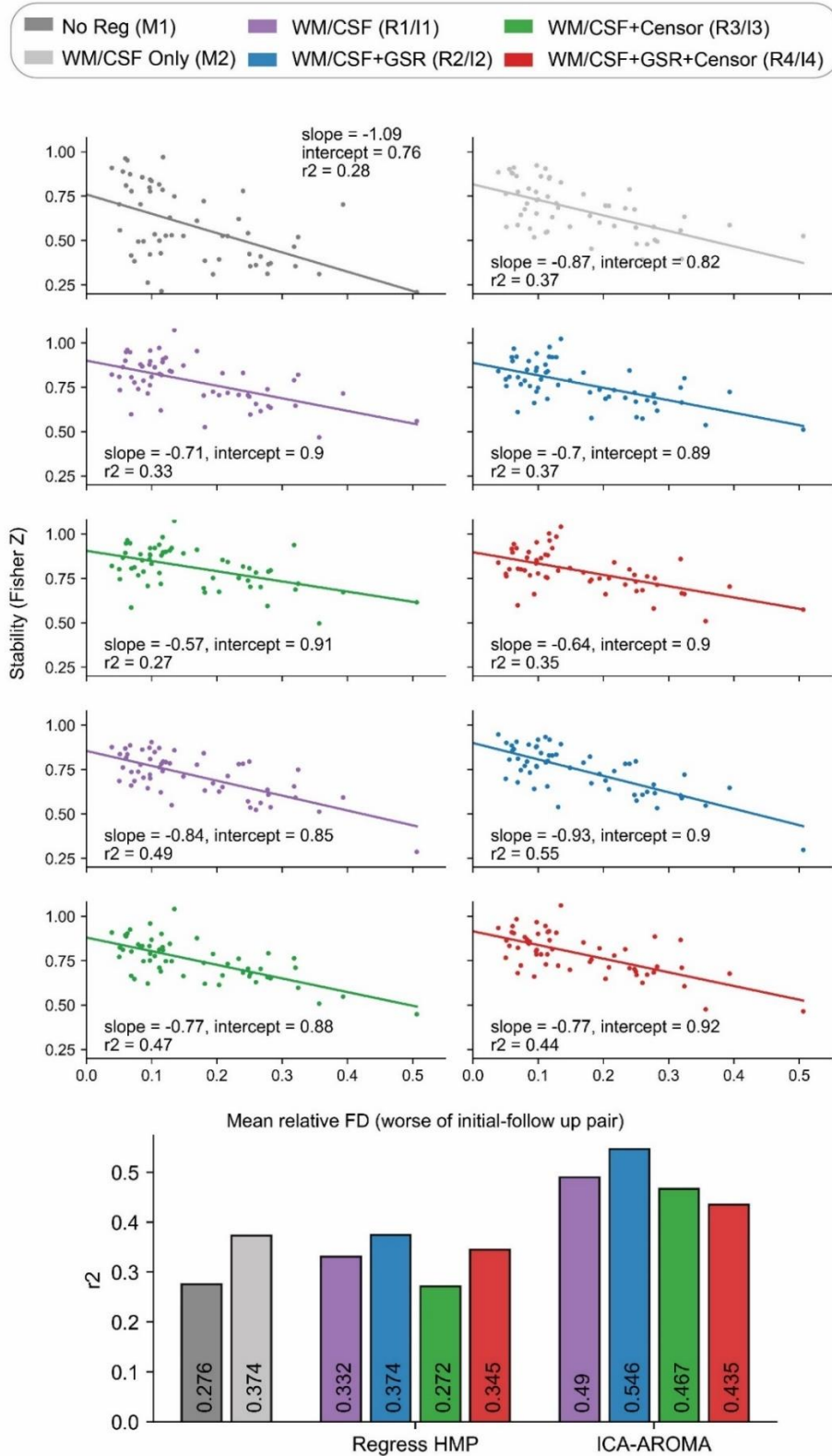
Supplemental Figure 2.3. Stability as a function of head motion. The correlation between each participant's two scans' FC estimates, plotted as a function of motion. Motion was calculated by taking the worse of each participant's two scans' mean relative FD.
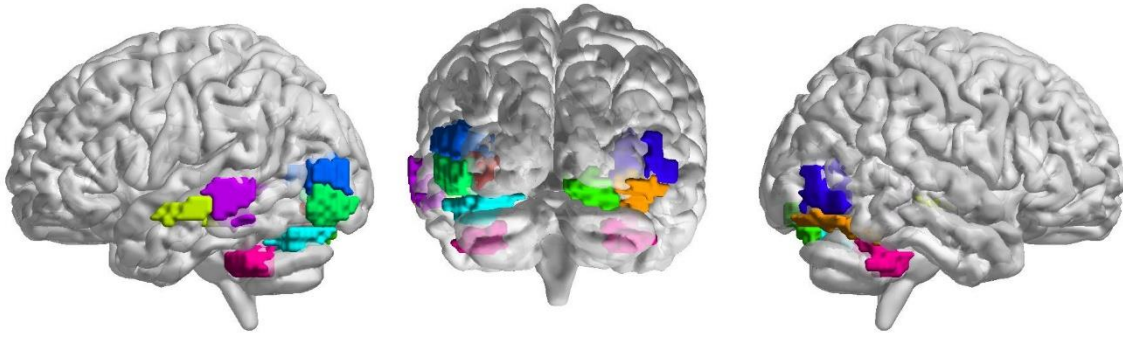
Supplemental Figure 2.4. Location of the 10 nodes with the highest average ISC. For each node defined by the MIST 325 parcellation, the average ISC was calculated across participants for each pipeline. Shown are the 10 nodes with the highest ISC values, averaged across all pipelines tested. Visualization was created with BrainNet Viewer (Xia et al., 2013, http://www.nitrc.org/projects/bnv/).
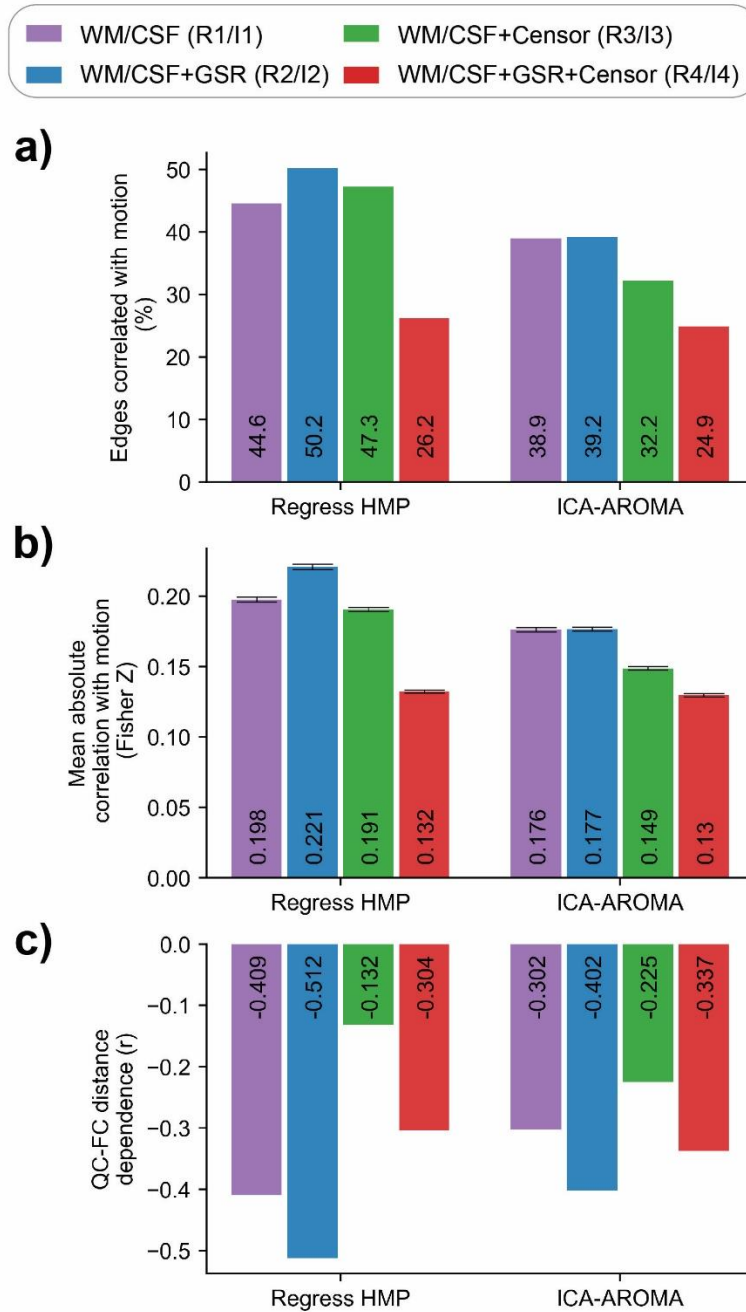
Supplemental Figure 2.5. **Quality control-functional connectivity (QC-FC) across pipelines with a highpass filter. a)** Percentage of edges with a significant correlation between edge strength and motion across all 112 scans (uncorrected p < 0.05). **b)** Mean and 99% confidence interval of the absolute correlation between edge strength and motion across all 112 scans. **c)** The correlation between edge distance and correlation between edge strength and motion.
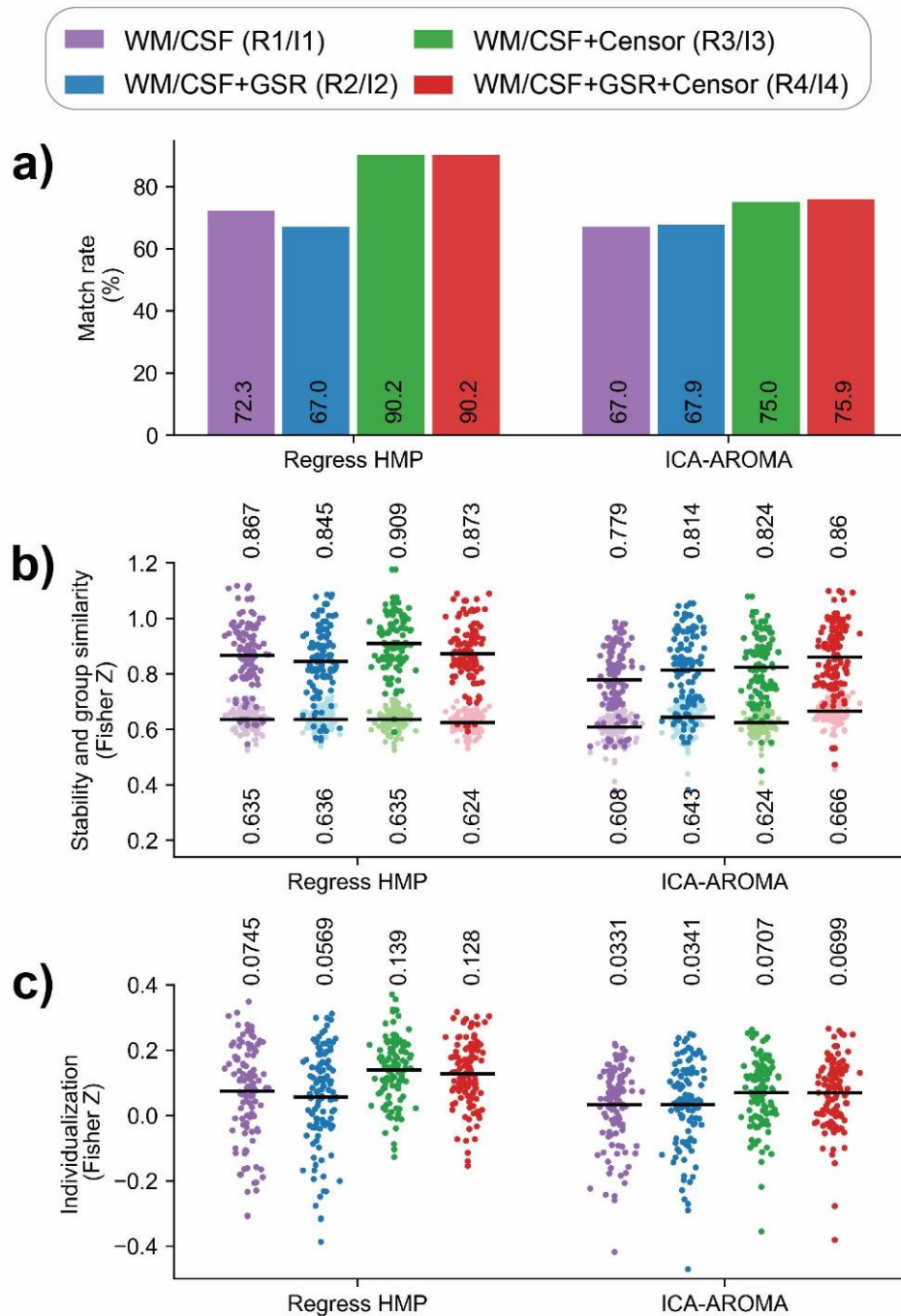
Supplemental Figure 2.6. Functional connectome fingerprinting across pipelines with a highpass filter. a) Match rate across pipelines. A scan matched if its highest correlation was to the other scan from the same individual. **b)** Each scan's stability (darker points) and group similarity (lighter points), for each pipeline. Group similarity was assessed as the average correlation to scans from other participants; stability was assessed as the correlation between scans from the same individual. Lines represent mean values. **c)** Individualization across pipelines. Each dot represents one scan. Individualization was assessed as the difference between stability and the highest correlation to a scan from another participant. Lines represent mean values.
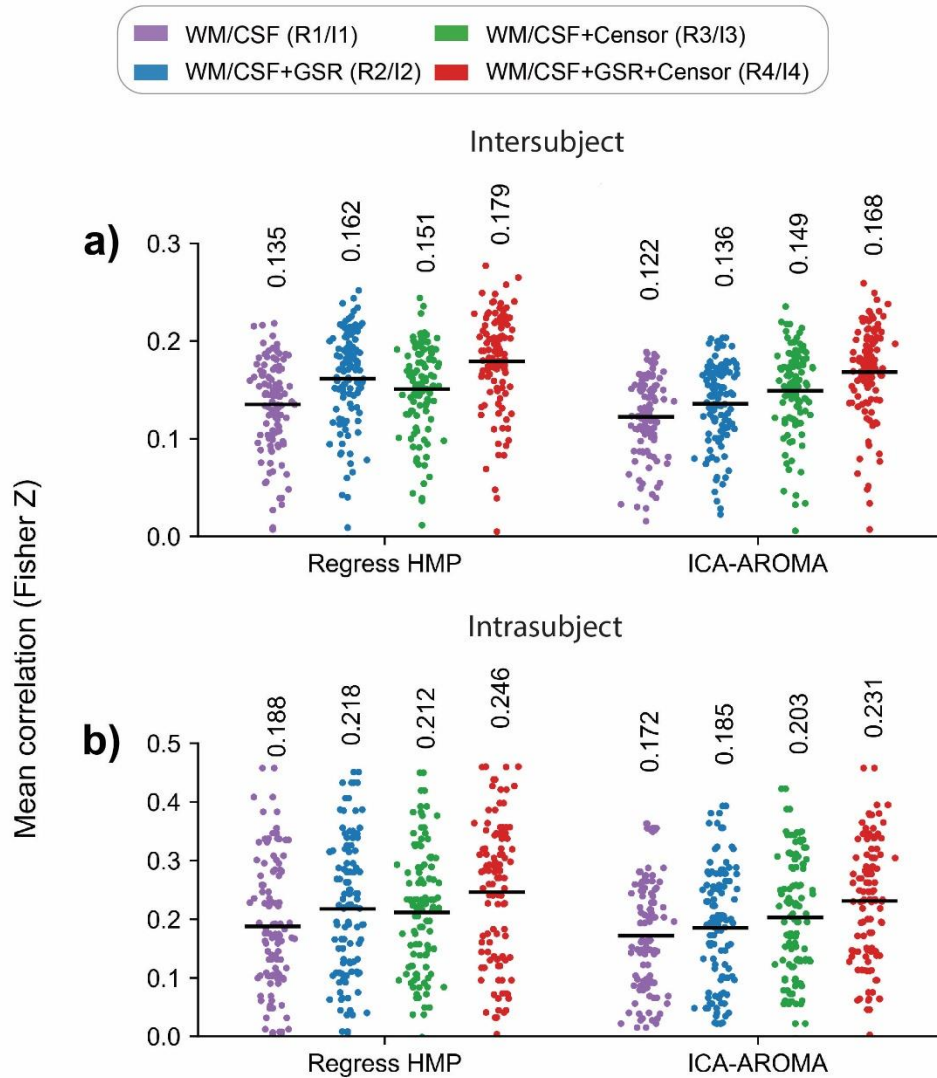
Supplemental Figure 2.7. Mean intersubject correlation (ISC) values with a highpass filter.
Each point represents the mean ISC for a scan to all other scans, averaged across the 10 nodes
with the highest ISC. Lines represent mean values. **a)** Mean ISCs for all 112 scans, each
compared to all other scans from other participants. **b)** Each participant's intrasubject correlation,
the correlation between their two scans' time series.
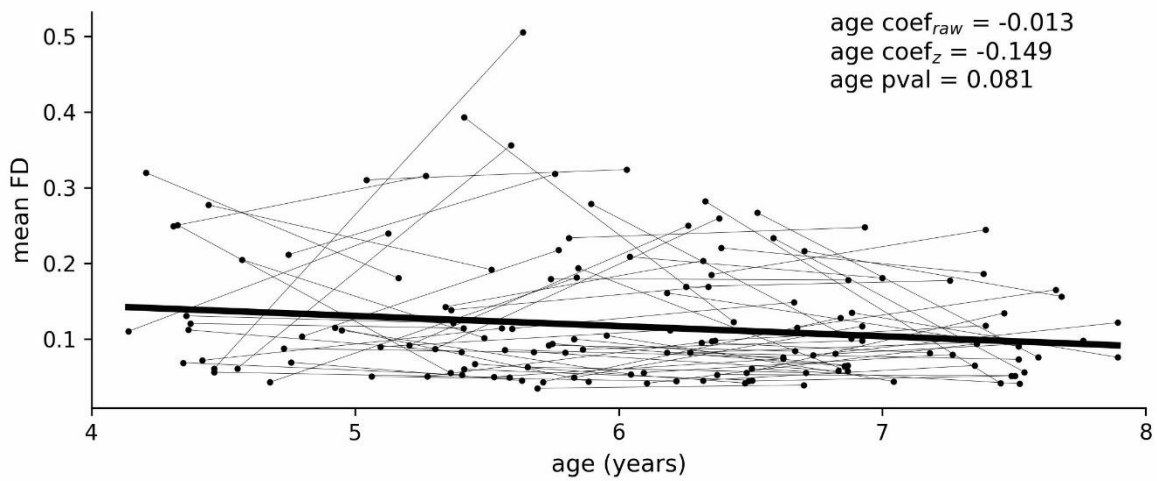
Supplemental Figure 3.1. Mean relative framewise displacement (FD) for each scan by age of participant at time of scan. Each dot represents one scan, with lines between dots showing scans from the same participant. A linear mixed model to predict mean relative FD by age was used to derive the age coefficient, age p value, and line of best fit. Raw data is shown, with both the raw and standardized age coefficient presented.

	Mean (± std) number volumes censored	Median number volumes censored	Number volumes censored vs age (p value)
First quartile	8.59 ± 10.69	4	0.97
Second quartile	8.51 ± 11.51	3.5	0.071
Third quartile	9.38 ± 12.04	5	0.54
Fourth quartile	12.43 ± 14.07	7	0.18
Whole scan	38.91 ± 39.04	26	0.24

Supplemental Table 3.1. Temporal censoring summary, by scan quartile. Censoring was carried out as part of the regression step, where we censored volumes above a FD threshold of 0.25 mm. A linear mixed model was used to predict number of volumes censored by age for the whole scan and each scan quartile separately.

166
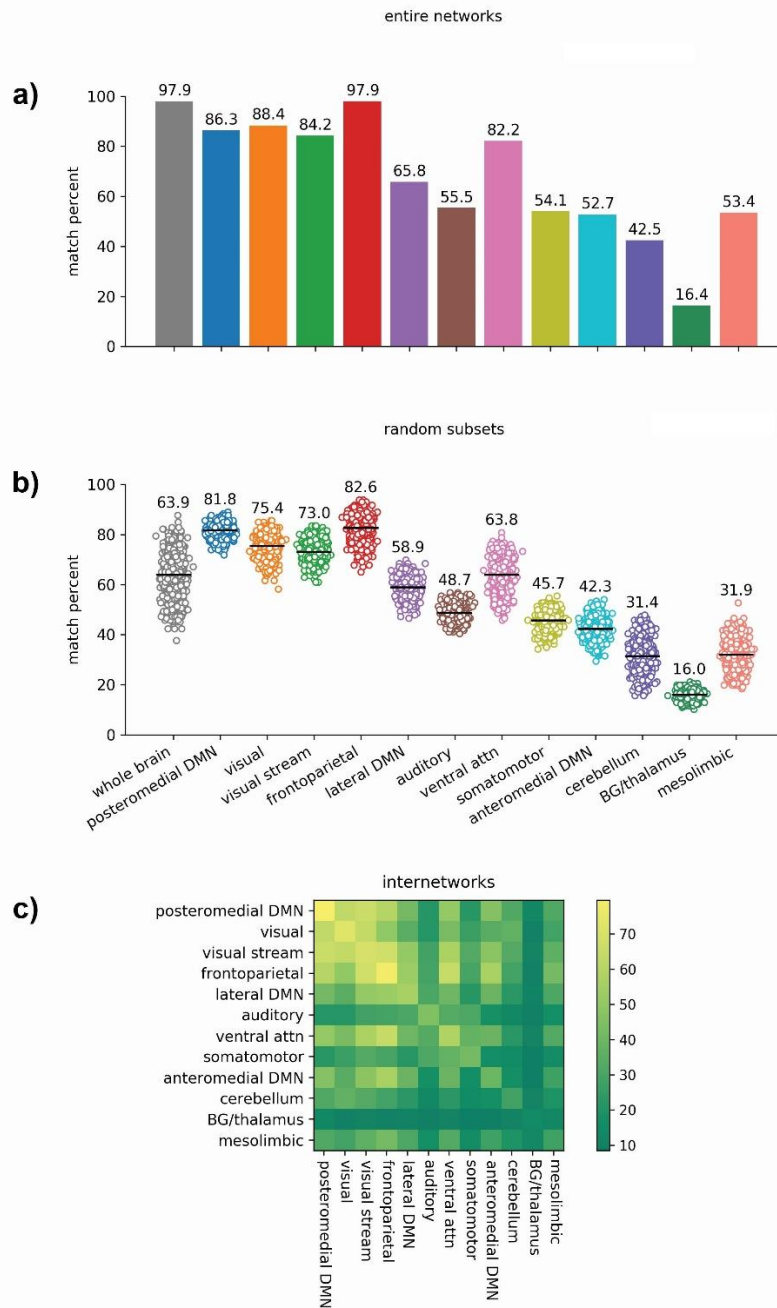
Supplemental Figure 3.2. Functional connectome fingerprinting match rate across networks. A scan matched if its self-stability was higher than the correlation to all other scans; successful match percentages are shown. **a)** Match rate when using all network edges. **b)** Match rate when controlling for network size using a subset of 36 randomly selected nodes, sampled 1000 times. Each dot represents one sampling; lines represent averages across samplings. **c)** Match rates in both networks and internetworks, controlling for network size. For each pair of networks, internetwork connectomes were derived from 18 random nodes from each network, and the (18x18=324) edges between those nodes. To compare the same number of edges, networks were similarly derived using 36 random nodes divided into two sets. Match rates are averaged over 1000 random samplings.
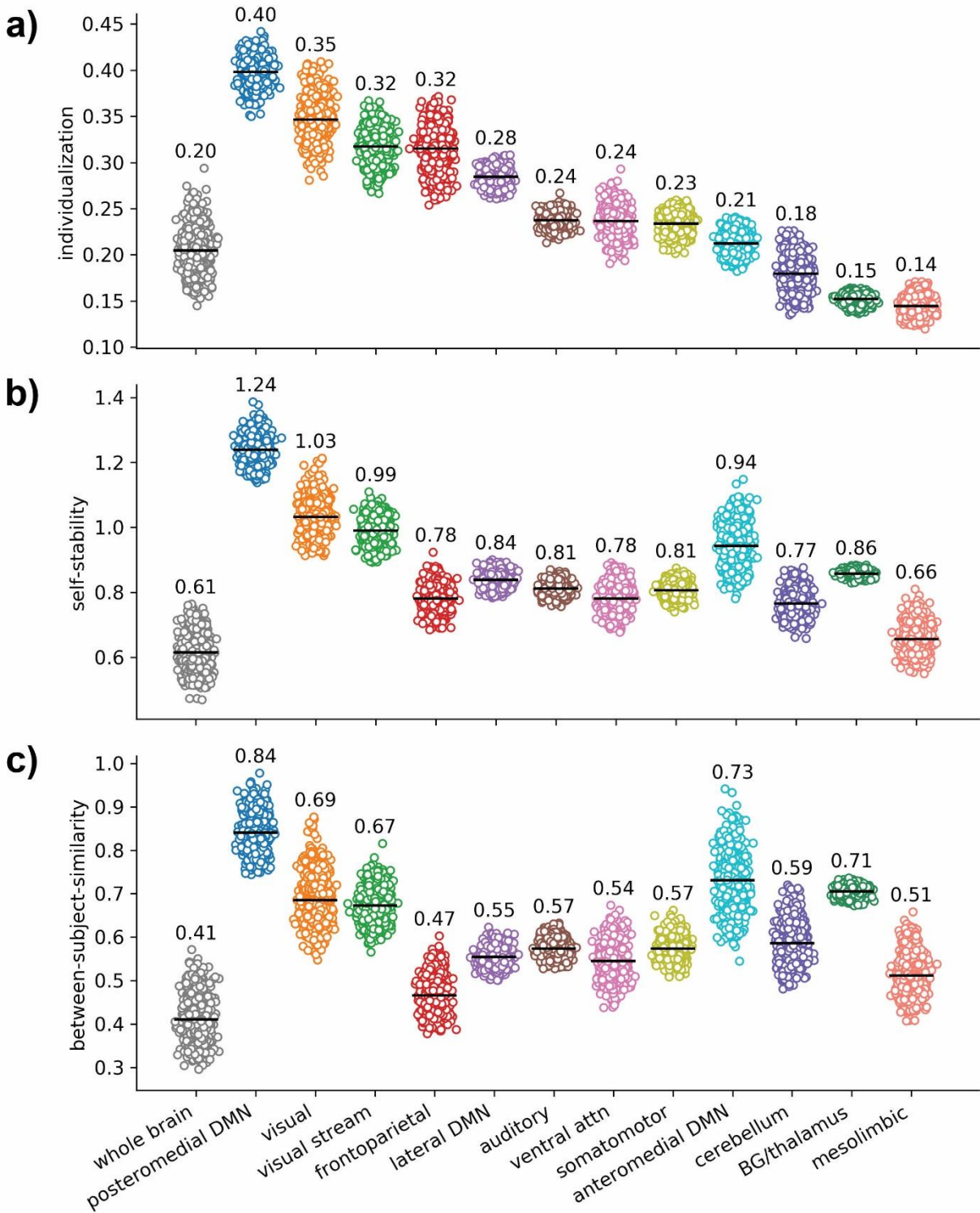
Supplemental Figure 3.3. (a) Developmental individualization, (b) longitudinal self-stability, and (c) average-between-subject-similarity, controlling for network size.
Connectomes were derived from 36 random nodes, sampled 1000 times. Each dot represents the mean value for the metric of interest for one sampling, averaged across all participants. Lines represent averages across samplings.
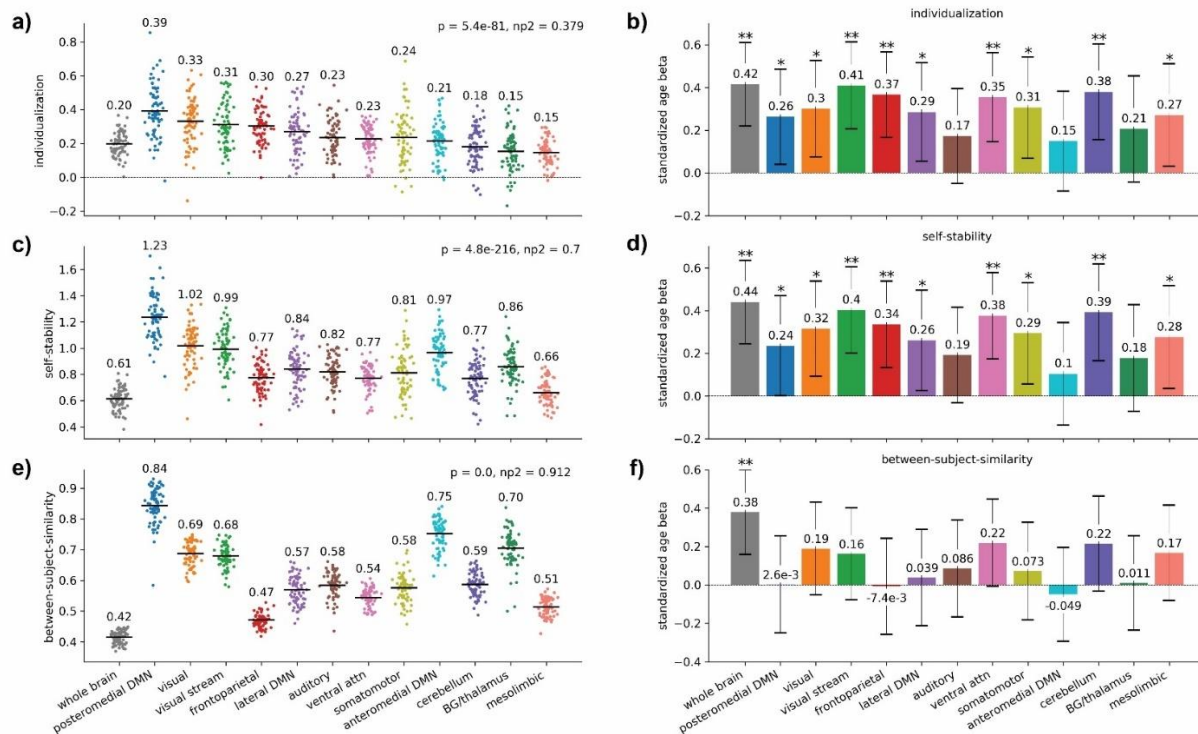
Supplemental Figure 3.4. Alternate volume-length matching analysis of developmental individualization, longitudinal self-stability, and average-between-subject-similarity across networks. This is similar to figures 3.3 and 3.4, using the alternate volume-length matching approach, described in 3.*2.3. Preprocessing*. **a, c, e)** Each participant's developmental individualization, longitudinal self-stability, and average-between-subject-similarity. Lines represent mean values across participants. **b, d, f)** Age associations. A linear regression was used to assess associations between age and the metric of interest, controlling for sex, motion (mean FD), and time between a participant's scans. Error bars depict 95% confidence intervals around standardized beta estimates. * = p < 0.05 uncorrected; ** = p < 0.05 Bonferroni corrected.
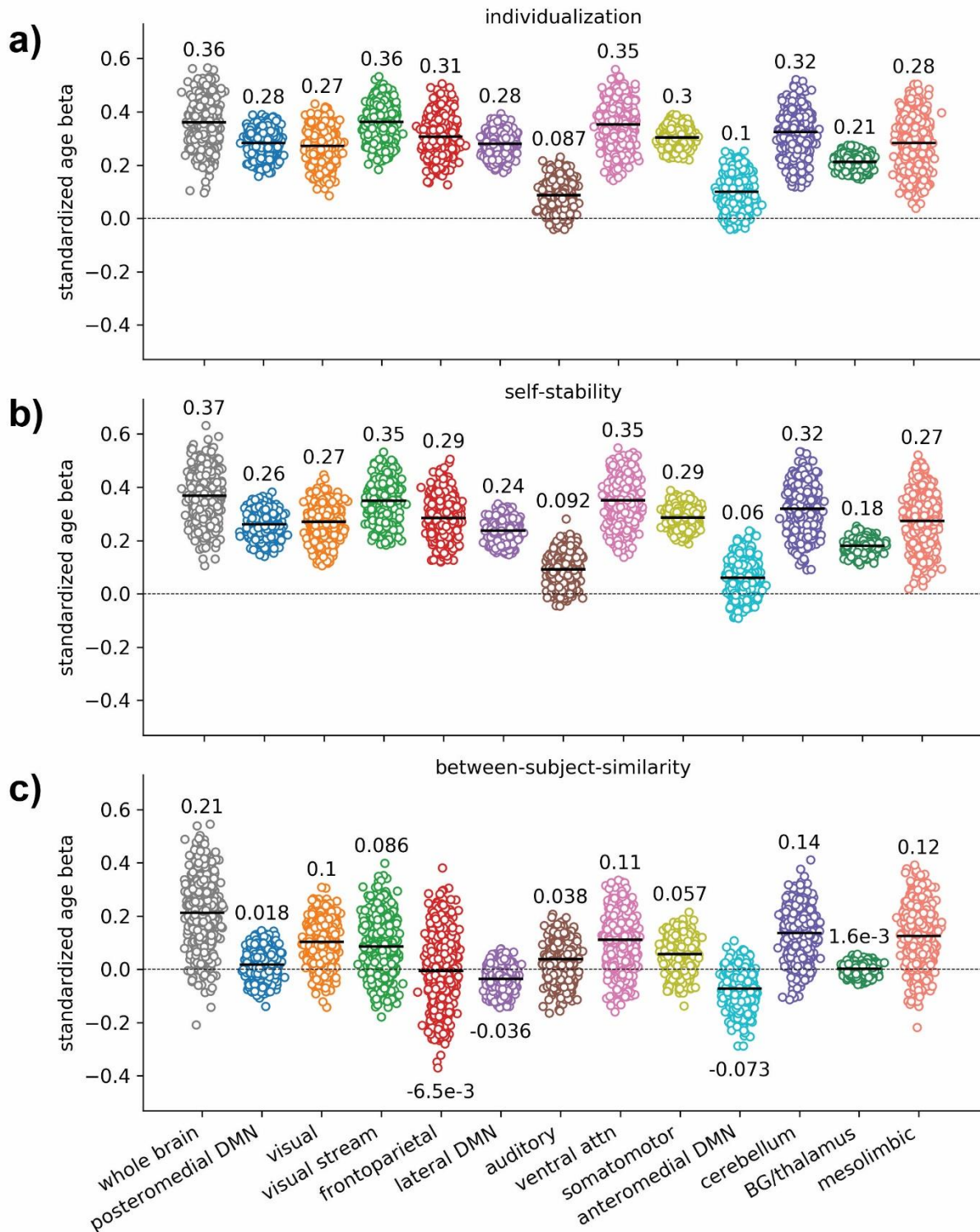
Supplemental Figure 3.5. Age effects of (a) developmental individualization, (b) longitudinal self-stability, and (c) average-between-subject-similarity, controlling for network size. Connectomes were derived from 36 random nodes, sampled 1000 times. Each dot represents the standardized age beta value for the metric of interest for one sampling. Lines represent averages across samplings.
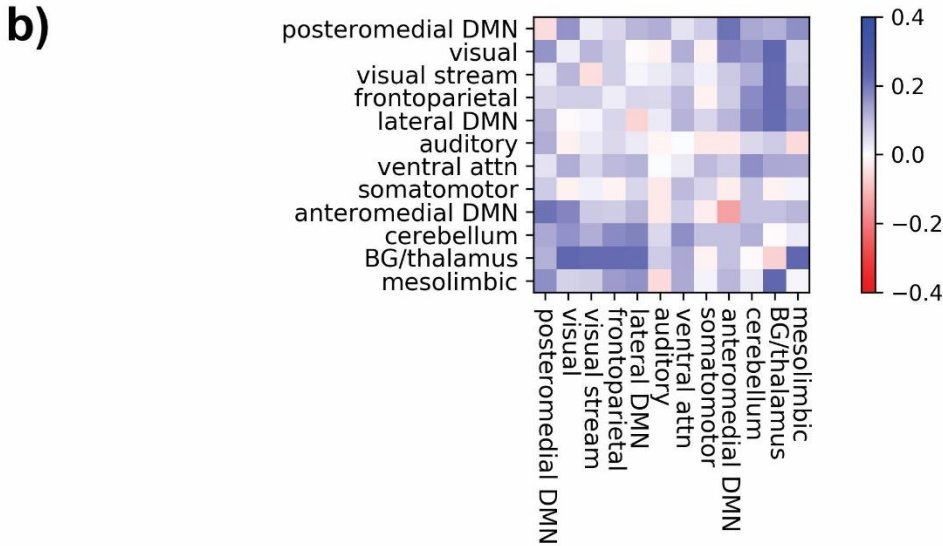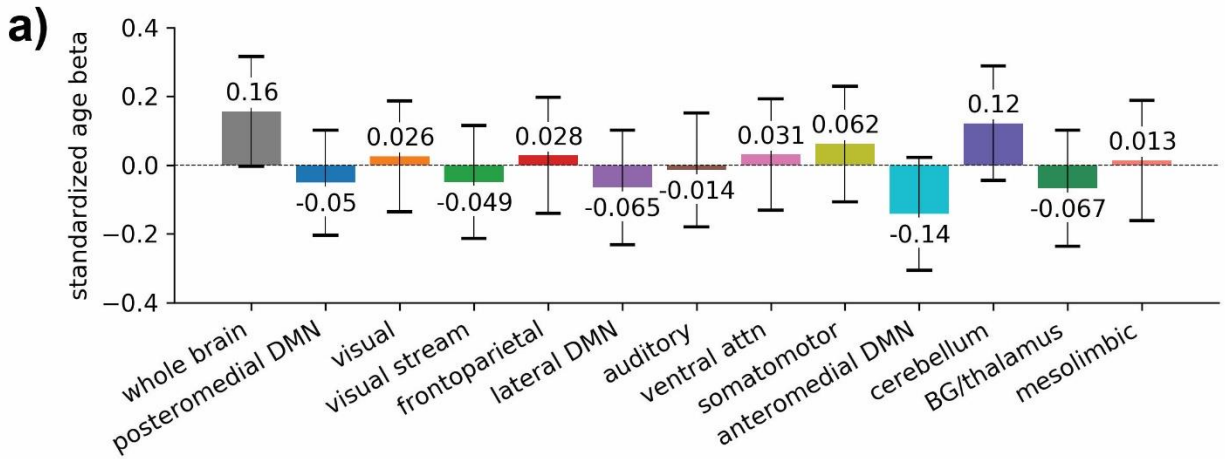
Supplemental Figure 3.6. Age effect of longitudinal-between-subject-similarity. Network effects are shown in (a), while both network and internetwork effects are shown in (b). For each participant, mean between-subject-similarity was calculated for their baseline and follow-up scan separately. A linear mixed effects model was used to test for the effect of age, controlling for motion and sex as fixed effects and participant identity as a random effect. Error bars in plot (a) represent 95% confidence intervals around standardized beta estimates. No effect was significant ($p < 0.05$ uncorrected). In plot (b), the network effects, shown along the diagonals, are the same values shown in figure (a).
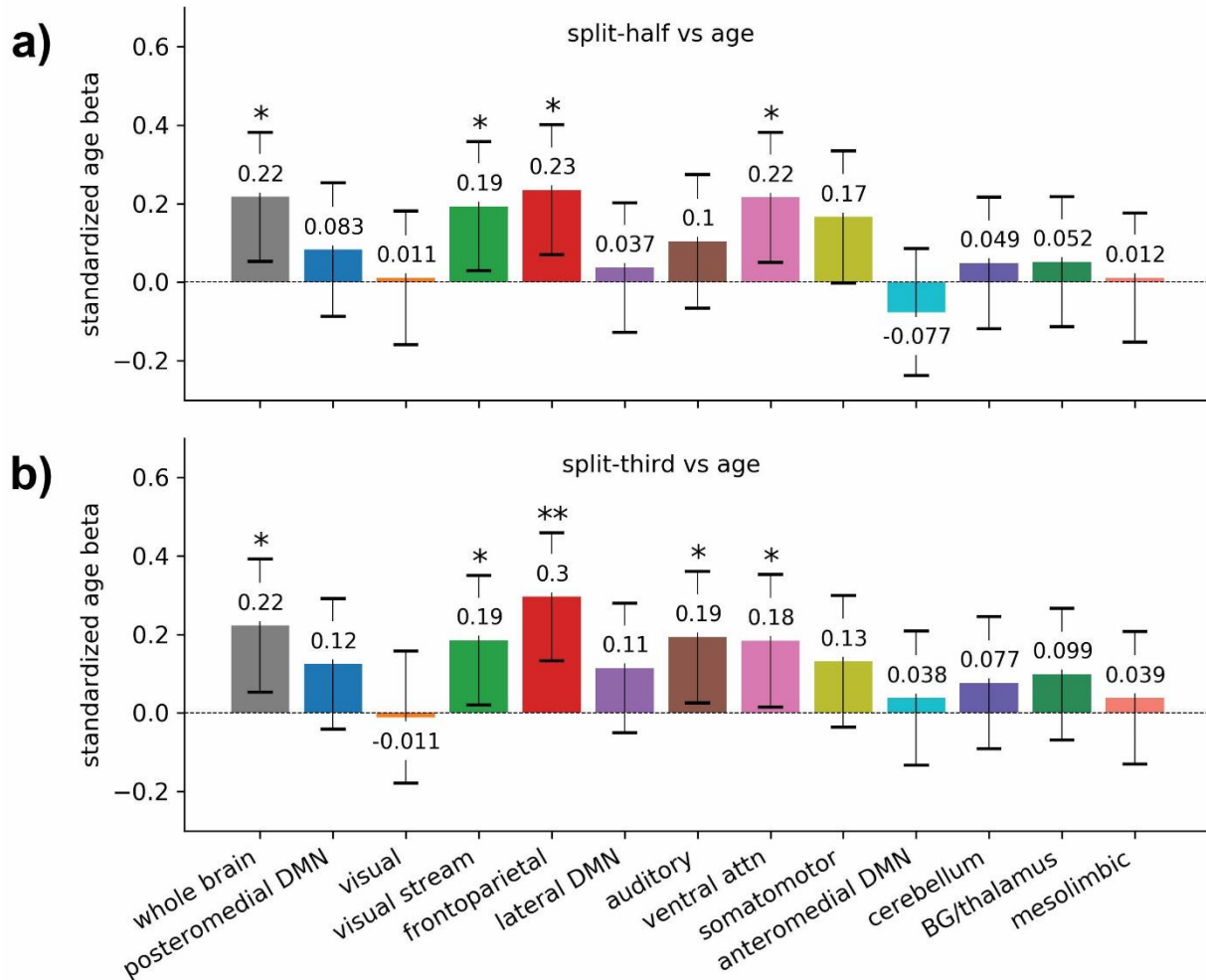
Supplemental Figure 3.7. Age effect of short-term self-stability, using either (a) a split-half correlation or (b) a 'split-third' correlation, the correlation between the first and last third of the scan. A linear mixed model was used to assess associations between age and each scan's split-half or 'split-third' correlation, controlling for sex and motion (average FD) as fixed effects, and participant identity as a random effect. Error bars represent 95% confidence intervals around standardized beta estimates. * = $p < 0.05$ uncorrected; ** = $p < 0.05$ Bonferroni corrected.

172

Family	Parent		Child		Days after visit 1		
	Sex	Age (Visit 1; years)	Sex	Age (Visit 1; years)	Visit 2	Visit 3	Visit 4
1	F	40.44	F	7.46	11	14	22
2	F	36.77	M	7.26	51	86	125
3	F	36.78	M	7.21	53	111	467
4	F	39.56	M	8.91	7	28	35
4B	M	42.05	M	7.34	7	28	147
5	M	44.57	M	8.84	28	35	42
5B	F	44.04	F	7.22	7	49	63
6	F	45.03	M	7.26	70	77	84
7	M	38.03	M	7.62	78	118	126
8	F	44.25	M	8.90	17	103	129
9	F	42.00	M	8.70	77	92	108
10	F	44.92	M	8.07	81	116	145
11	M	36.77	M	7.57	7	14	42
12	F	40.22	F	7.94	56	109	182
13	M	43.60	M	8.02	7	21	34
14	M	46.94	F	7.60	28	35	98
15	M	42.63	F	8.55	7	14	35
16	M	33.75	F	8.26	23	26	50
17	F	47.13	F	8.82	7	36	41
18	M	41.73	F	8.48	28	35	42
19	M	42.12	F	7.98	7	13	20
20	F	44.46	F	8.49	7	45	59
21	M	38.09	F	7.03	6	42	49
22	M	34.88	F	6.90	7	14	21
23	M	43.58	F	6.56	21	43	50

Supplemental Table 4.1. Age, sex, and visit schedule of participants.
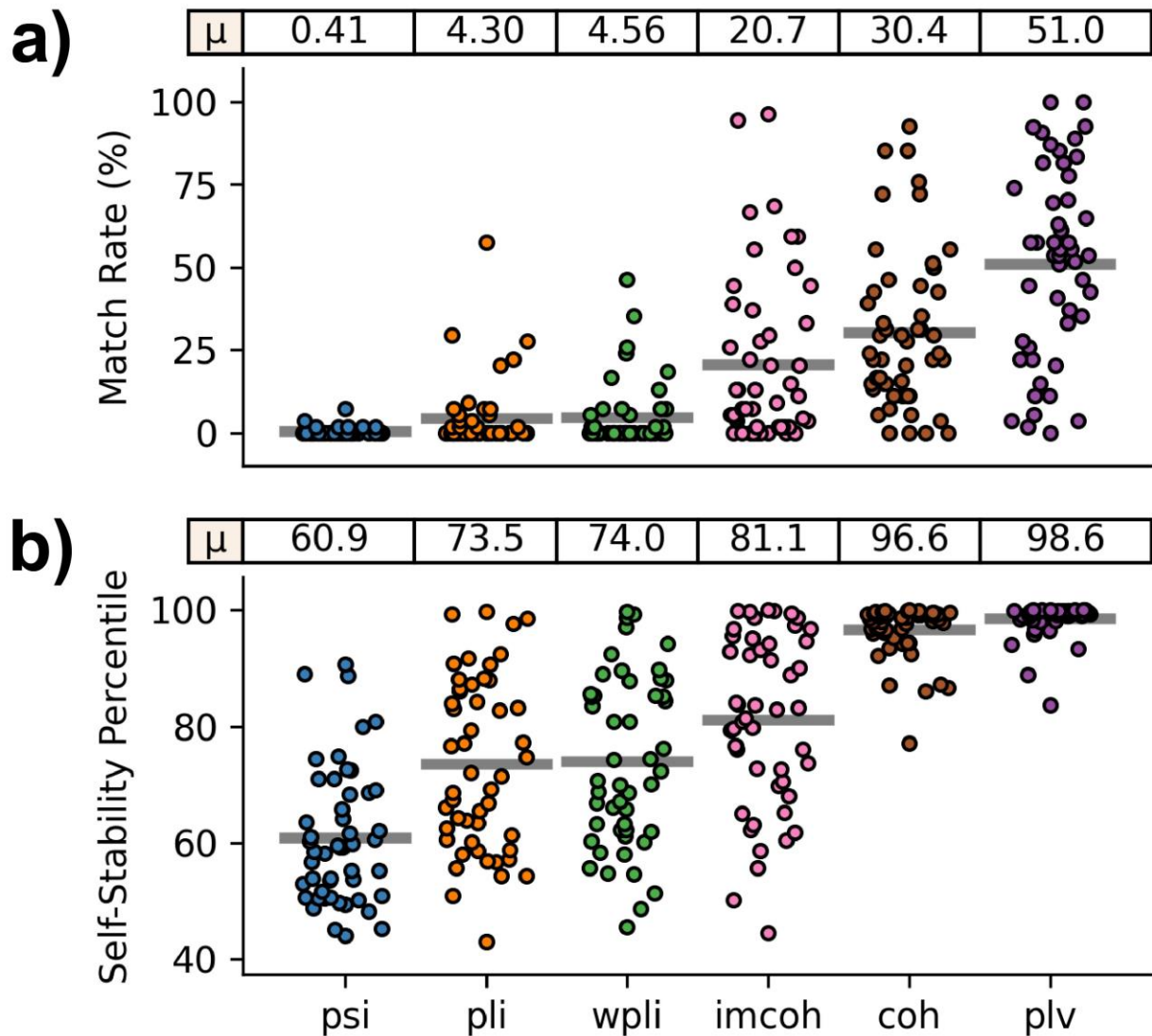
Supplemental Figure 4.1. Additional identifiability metrics across FC measures. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **a)** Match rate, the percentage of self-stability scores that were higher than all similarity-to-others scores. **b)** Self-stability percentile, the average percentile of self-stability scores, relative to similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*
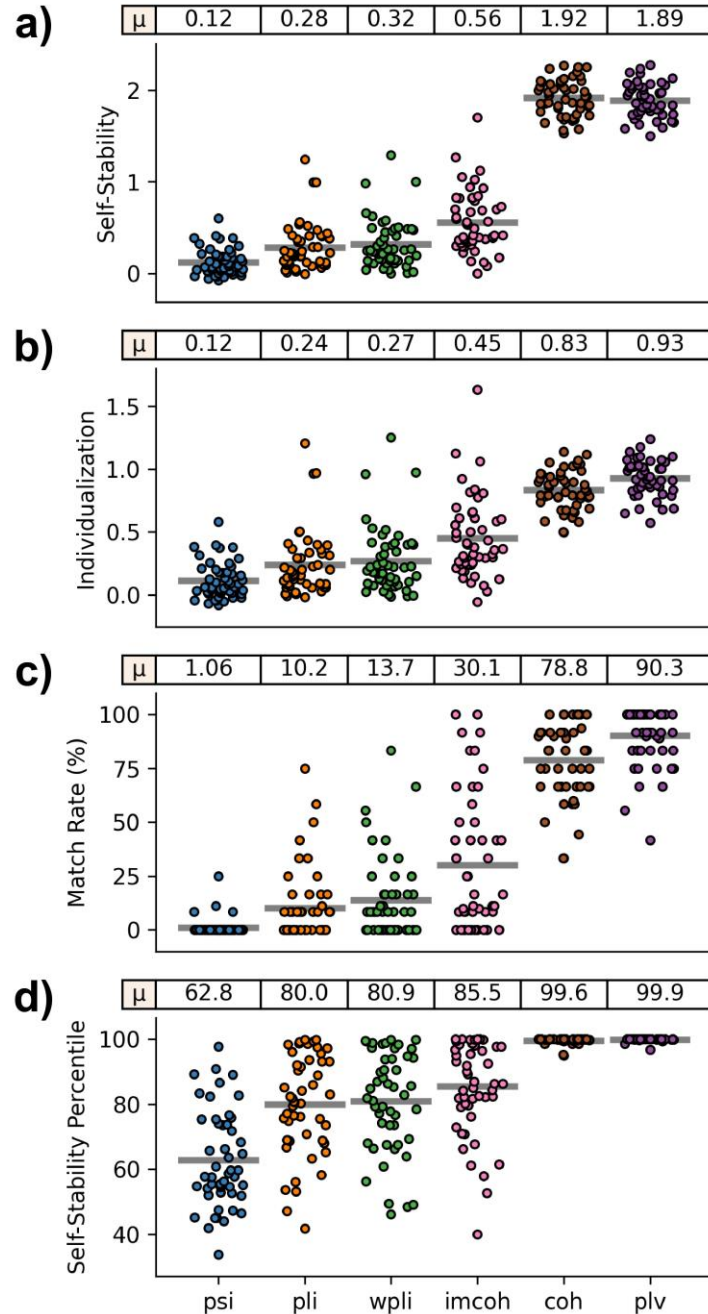
Supplemental Figure 4.2. Within-session reliability and identifiability across FC measures.
This is similar to Figure 4.2 and Supplemental Figure 4.1 but calculating self-stability by
comparing connectomes from the same session, rather than different sessions. Each dot
represents one participant. Lines represent mean values across participants, which are also
displayed at the top of each subplot. **a)** Mean self-stability, the average Fisher-z correlation
between connectomes of the same participant, collected on the same day. **b)** Individualization,
the difference between mean self-stability and mean similarity-to-others. **c)** Match rate, the
percentage of self-stability scores that were higher than all similarity-to-others scores. **d)** Self-
stability percentile, the average percentile of self-stability scores, relative to similarity-to-others.
*Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary
coherence; coh: coherence; plv: phase locking value.*

Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 2.7e-5	p: 1.3e-6	p: 5.5e-14	p: 1.7e-43	p: 3.3e-42
pli		p: 1.1e-7	p: 2.1e-18	p: 1.5e-38	p: 6.2e-37
wpli			p: 1.1e-18	p: 1.5e-38	p: 7.4e-37
imcoh				p: 2.2e-30	p: 1.2e-28
coh					p: 4.0e-5

Similarity-to-Others

	pli	wpli	imcoh	coh	plv
psi	p: 1.4e-15	p: 1.7e-15	p: 6.8e-15	p: 9.0e-57	p: 3.0e-53
pli		p: 9.8e-7	p: 1.3e-12	p: 5.6e-57	p: 4.0e-53
wpli			p: 4.9e-12	p: 4.2e-57	p: 3.8e-53
imcoh				p: 1.7e-52	p: 2.1e-48
coh					p: 4.2e-43

Individualization

	pli	wpli	imcoh	coh	plv
psi	p: 3.3e-3	p: 3.2e-4	p: 5.2e-11	p: 2.2e-22	p: 1.7e-25
pli		p: 7.2e-7	p: 1.6e-15	p: 9.0e-15	p: 7.3e-18
wpli			p: 1.4e-15	p: 5.7e-14	p: 3.6e-17
imcoh				p: 3.5e-4	p: 3.1e-7
coh					p: 8.4e-21

Match Rate (%)

	pli	wpli	imcoh	coh	plv
psi	p: 0.011	p: 3.9e-3	p: 6.5e-7	p: 1.4e-11	p: 1.3e-16
pli		p: 0.58	p: 2.6e-7	p: 2.0e-9	p: 7.6e-15
wpli			p: 1.3e-7	p: 1.8e-9	p: 5.1e-15
imcoh				p: 0.038	p: 2.4e-7
coh					p: 1.5e-12

Self-Stability Percentile

	pli	wpli	imcoh	coh	plv
psi	p: 1.0e-8	p: 1.5e-9	p: 3.6e-14	p: 6.8e-28	p: 1.2e-28
pli		p: 0.24	p: 1.1e-8	p: 3.3e-15	p: 1.3e-16
wpli			p: 4.5e-9	p: 2.9e-15	p: 1.1e-16
imcoh				p: 1.6e-9	p: 3.4e-11
coh					p: 9.8e-8

Supplemental Figure 4.3. Statistical comparisons between reliability and identifiability measures, as presented in Figure 4.2 and Supplemental Figure 4.1. Between each FC measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 6.4e-7	p: 1.5e-8	p: 2.6e-15	p: 3.2e-45	p: 7.1e-45
pli		p: 7.3e-9	p: 1.7e-19	p: 3.8e-39	p: 8.5e-38
wpli			p: 1.3e-19	p: 1.1e-38	p: 2.6e-37
imcoh				p: 4.6e-32	p: 1.2e-30
coh					p: 7.5e-5

Individualization

	pli	wpli	imcoh	coh	plv
psi	p: 3.2e-5	p: 9.8e-7	p: 8.4e-13	p: 1.8e-30	p: 2.6e-33
pli		p: 5.7e-8	p: 1.3e-16	p: 6.6e-22	p: 8.4e-24
wpli			p: 1.9e-16	p: 1.1e-20	p: 1.2e-22
imcoh				p: 5.7e-12	p: 2.1e-14
coh					p: 3.1e-19

Match Rate (%)

	pli	wpli	imcoh	coh	plv
psi	p: 2.0e-4	p: 7.2e-6	p: 2.2e-8	p: 4.1e-33	p: 3.8e-41
pli		p: 2.6e-3	p: 5.2e-8	p: 1.3e-25	p: 7.4e-30
wpli			p: 4.8e-6	p: 6.7e-24	p: 9.9e-28
imcoh				p: 1.7e-14	p: 7.5e-18
coh					p: 1.7e-8

Self-Stability Percentile

	pli	wpli	imcoh	coh	plv
psi	p: 4.5e-10	p: 6.2e-12	p: 4.5e-15	p: 6.2e-23	p: 5.5e-23
pli		p: 0.26	p: 2.7e-5	p: 3.3e-12	p: 2.2e-12
wpli			p: 4.1e-4	p: 1.8e-11	p: 1.2e-11
imcoh				p: 7.0e-9	p: 4.5e-9
coh					p: 5.8e-3

Supplemental Figure 4.4. Statistical comparisons between within-session reliability and identifiability measures, as presented in Supplemental Figure 4.2. Between each FC measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = $p < 0.05$ Bonferroni corrected; green shading = $p < 0.05$ uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Δ Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 0.40	p: 0.22	p: 0.056	p: 0.31	p: 0.47
pli		p: 0.13	p: 0.037	p: 0.76	p: 0.84
wpli			p: 0.13	p: 0.73	p: 0.42
imcoh				p: 0.12	p: 0.068
coh					p: 0.060

Δ Similarity-to-Others

	pli	wpli	imcoh	coh	plv
psi	p: 0.26	p: 0.14	p: 0.023	p: 0.53	p: 0.69
pli		p: 0.12	p: 0.040	p: 0.23	p: 0.081
wpli			p: 0.11	p: 0.074	p: 0.030
imcoh				p: 0.030	p: 0.017
coh					p: 5.7e-3

Supplemental Figure 4.5. Statistical comparisons between FC measures of task sensitivity, as presented in Figure 4.3. Between each measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = $p < 0.05$ Bonferroni corrected; green shading = $p < 0.05$ uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Δ Similarity-to-Others	pli	wpli	imcoh	coh	plv
psi	p: 8.8e-4	p: 3.3e-3	p: 0.026	p: 4.5e-3	p: 4.6e-3
pli		p: 0.83	p: 0.45	p: 0.19	p: 0.22
wpli			p: 0.39	p: 0.16	p: 0.18
imcoh				p: 0.66	p: 0.72
coh					p: 0.38

Supplemental Figure 4.6. Statistical comparisons between FC measures of age sensitivity, as presented in Figure 4.4. Between each measure, a paired t-test was used to assess the statistical difference between participants' similarity-to-others difference. Yellow shading = $p < 0.05$ Bonferroni corrected; green shading = $p < 0.05$ uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

No Reg (M1)	WM/CSF (R1/I1)	WM/CSF+Censor (R3/I3)
WM/CSF Only (M2)	WM/CSF+GSR (R2/I2)	WM/CSF+GSR+Censor (R4/I4)

Absolute correlation with motion

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	Cohen's d
Minimal - M1	Minimal - M2	0.881
Minimal - M2	Regress HMP - R1	0.975
Minimal - M2	ICA-AROMA - I1	0.845

Effect of GSR

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R2	0.116
Regress HMP - R3	Regress HMP - R4	0.334
ICA-AROMA - I1	ICA-AROMA - I2	0.034
ICA-AROMA - I3	ICA-AROMA - I4	0.238

Effect of Censoring

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R3	0.107
Regress HMP - R2	Regress HMP - R4	0.33
ICA-AROMA - I1	ICA-AROMA - I3	0.158
ICA-AROMA - I2	ICA-AROMA - I4	0.373

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.09
Regress HMP - R2	ICA-AROMA - I2	0.323
Regress HMP - R3	ICA-AROMA - I3	0.014
Regress HMP - R4	ICA-AROMA - I4	0.222

Supplemental Table 1. Absolute correlation with motion, statistical comparisons between pipelines. See Figure 2.1b for main results. Cohen's d was used to compare pipelines across all 52650 edges.

Absolute correlation with motion
(low motion)

| No Reg (M1) | WM/CSF (R1/I1) | WM/CSF+Censor (R3/I3) |
| WM/CSF Only (M2) | WM/CSF+GSR (R2/I2) | WM/CSF+GSR+Censor (R4/I4) |

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	Cohen's d
Minimal - M1	Minimal - M2	0.781
Minimal - M2	Regress HMP - R1	0.816
Minimal - M2	ICA-AROMA - I1	0.773

Effect of GSR

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R2	0.127
Regress HMP - R3	Regress HMP - R4	0.034
ICA-AROMA - I1	ICA-AROMA - I2	0.144
ICA-AROMA - I3	ICA-AROMA - I4	0.034

Effect of Censoring

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R3	0.054
Regress HMP - R2	Regress HMP - R4	0.036
ICA-AROMA - I1	ICA-AROMA - I3	0.055
ICA-AROMA - I2	ICA-AROMA - I4	0.028

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.061
Regress HMP - R2	ICA-AROMA - I2	0.053
Regress HMP - R3	ICA-AROMA - I3	0.077
Regress HMP - R4	ICA-AROMA - I4	0.084

Supplemental Table 2. Absolute correlation with motion for low motion scans, statistical comparisons between pipelines. See Figure 2.2c for main results. Cohen's d was used to compare pipelines across all 52650 edges.

Absolute correlation with motion
(high motion)

Legend:
- ■ No Reg (M1)
- ■ WM/CSF Only (M2)
- ■ WM/CSF (R1/I1)
- ■ WM/CSF+GSR (R2/I2)
- ■ WM/CSF+Censor (R3/I3)
- ■ WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	Cohen's d
Minimal - M1	Minimal - M2	0.574
Minimal - M2	Regress HMP - R1	0.512
Minimal - M2	ICA-AROMA - I1	0.414

Effect of GSR

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R2	0.049
Regress HMP - R3	Regress HMP - R4	0.265
ICA-AROMA - I1	ICA-AROMA - I2	0.139
ICA-AROMA - I3	ICA-AROMA - I4	0.176

Effect of Censoring

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R3	0.05
Regress HMP - R2	Regress HMP - R4	0.159
ICA-AROMA - I1	ICA-AROMA - I3	0.058
ICA-AROMA - I2	ICA-AROMA - I4	0.085

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.092
Regress HMP - R2	ICA-AROMA - I2	0.094
Regress HMP - R3	ICA-AROMA - I3	0.1
Regress HMP - R4	ICA-AROMA - I4	0.067

Supplemental Table 3. Absolute correlation with motion for high motion scans, statistical comparisons between pipelines. See Figure 2.2d for main results. Cohen's d was used to compare pipelines across all 52650 edges.

Similarity

Legend:
- No Reg (M1)
- WM/CSF Only (M2)
- WM/CSF (R1/I1)
- WM/CSF+GSR (R2/I2)
- WM/CSF+Censor (R3/I3)
- WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	1.62e-104	8.354
Minimal - M2	Regress HMP - R1	1.45e-76	6.796
Minimal - M2	ICA-AROMA - I1	2.83e-74	7.856

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	1.09e-31	1.568
Regress HMP - R3	Regress HMP - R4	0.074	0.171
ICA-AROMA - I1	ICA-AROMA - I2	1.74e-94	6.759
ICA-AROMA - I3	ICA-AROMA - I4	1.42e-118	11.217

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	7.87e-4	0.328
Regress HMP - R2	Regress HMP - R4	5.15e-25	1.275
ICA-AROMA - I1	ICA-AROMA - I3	9.05e-50	2.512
ICA-AROMA - I2	ICA-AROMA - I4	2.19e-57	3.004

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	2.29e-13	0.791
Regress HMP - R2	ICA-AROMA - I2	2.49e-12	0.747
Regress HMP - R3	ICA-AROMA - I3	5.41e-6	0.454
Regress HMP - R4	ICA-AROMA - I4	1.06e-62	3.391

Supplemental Table 4. Group similarity, statistical comparisons between pipelines. See Figure 2.3b for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

Stability

Legend: No Reg (M1) | WM/CSF Only (M2) | WM/CSF (R1/I1) | WM/CSF+GSR (R2/I2) | WM/CSF+Censor (R3/I3) | WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	1.39e-13	1.314
Minimal - M2	Regress HMP - R1	1.45e-32	2.146
Minimal - M2	ICA-AROMA - I1	3.23e-10	1.321

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	1.42e-5	0.643
Regress HMP - R3	Regress HMP - R4	3.30e-10	1.031
ICA-AROMA - I1	ICA-AROMA - I2	7.35e-30	3.097
ICA-AROMA - I3	ICA-AROMA - I4	7.70e-26	2.563

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	1.16e-14	1.408
Regress HMP - R2	Regress HMP - R4	1.62e-13	1.308
ICA-AROMA - I1	ICA-AROMA - I3	2.94e-14	1.373
ICA-AROMA - I2	ICA-AROMA - I4	2.58e-15	1.466

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	5.25e-27	2.71
Regress HMP - R2	ICA-AROMA - I2	3.10e-7	0.785
Regress HMP - R3	ICA-AROMA - I3	1.03e-37	4.407
Regress HMP - R4	ICA-AROMA - I4	0.232	0.163

Supplemental Table 5. Stability, statistical comparisons between pipelines. See Figure 2.3c for main results. A paired t-test was used to compare pipelines across 56 participants. P-values are uncorrected.

Individualization

Legend

- No Reg (M1)
- WM/CSF Only (M2)
- WM/CSF (R1/I1)
- WM/CSF+GSR (R2/I2)
- WM/CSF+Censor (R3/I3)
- WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	5.03e-25	1.275
Minimal - M2	Regress HMP - R1	2.33e-65	1.979
Minimal - M2	ICA-AROMA - I1	4.79e-19	1.294

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	1.88e-3	0.302
Regress HMP - R3	Regress HMP - R4	0.593	0.051
ICA-AROMA - I1	ICA-AROMA - I2	0.269	0.105
ICA-AROMA - I3	ICA-AROMA - I4	5.57e-3	0.268

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	1.45e-32	1.608
Regress HMP - R2	Regress HMP - R4	1.21e-32	1.612
ICA-AROMA - I1	ICA-AROMA - I3	1.58e-12	0.756
ICA-AROMA - I2	ICA-AROMA - I4	7.28e-15	0.854

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	3.21e-26	1.327
Regress HMP - R2	ICA-AROMA - I2	1.10e-32	1.614
Regress HMP - R3	ICA-AROMA - I3	2.53e-68	3.844
Regress HMP - R4	ICA-AROMA - I4	7.16e-63	3.404

Supplemental Table 6. Individualization, statistical comparisons between pipelines. See Figure 2.3d for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected

185

Between session ICC

No Reg (M1)	WM/CSF (R1/I1)	WM/CSF+Censor (R3/I3)
WM/CSF Only (M2)	WM/CSF+GSR (R2/I2)	WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	Cohen's d
Minimal - M1	Minimal - M2	0.394
Minimal - M2	Regress HMP - R1	0.751
Minimal - M2	ICA-AROMA - I1	0.368

Effect of GSR

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R2	0.336
Regress HMP - R3	Regress HMP - R4	0.313
ICA-AROMA - I1	ICA-AROMA - I2	0.393
ICA-AROMA - I3	ICA-AROMA - I4	0.24

Effect of Censoring

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R3	0.269
Regress HMP - R2	Regress HMP - R4	0.263
ICA-AROMA - I1	ICA-AROMA - I3	0.084
ICA-AROMA - I2	ICA-AROMA - I4	0.198

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.315
Regress HMP - R2	ICA-AROMA - I2	0.339
Regress HMP - R3	ICA-AROMA - I3	0.437
Regress HMP - R4	ICA-AROMA - I4	0.354

Supplemental Table 7. Between session ICC, statistical comparisons between pipelines. See Figure 2.5a for main results. Cohen's d was used to compare pipelines across all 52650 edges.

Within-session ICC

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	Cohen's d
Minimal - M1	Minimal - M2	0.33
Minimal - M2	Regress HMP - R1	0.188
Minimal - M2	ICA-AROMA - I1	0.053

Effect of GSR

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R2	0.37
Regress HMP - R3	Regress HMP - R4	0.456
ICA-AROMA - I1	ICA-AROMA - I2	0.373
ICA-AROMA - I3	ICA-AROMA - I4	0.429

Effect of Censoring

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	Regress HMP - R3	0.178
Regress HMP - R2	Regress HMP - R4	0.269
ICA-AROMA - I1	ICA-AROMA - I3	0.046
ICA-AROMA - I2	ICA-AROMA - I4	0.1

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.256
Regress HMP - R2	ICA-AROMA - I2	0.229
Regress HMP - R3	ICA-AROMA - I3	0.19
Regress HMP - R4	ICA-AROMA - I4	0.147

Supplemental Table 8. Within session ICC, statistical comparisons between pipelines. See Figure 2.5b for main results. Cohen's d was used to compare pipelines across all 52650 edges.

187

Intersubject

Legend: No Reg (M1) | WM/CSF (R1/I1) | WM/CSF+Censor (R3/I3) | WM/CSF Only (M2) | WM/CSF+GSR (R2/I2) | WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	1.75e-78	4.8
Minimal - M2	Regress HMP - R1	1.60e-9	1.647
Minimal - M2	ICA-AROMA - I1	1.42e-10	1.375

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	2.54e-73	4.291
Regress HMP - R3	Regress HMP - R4	1.91e-71	4.118
ICA-AROMA - I1	ICA-AROMA - I2	8.98e-70	3.969
ICA-AROMA - I3	ICA-AROMA - I4	4.72e-67	3.737

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	1.16e-47	2.388
Regress HMP - R2	Regress HMP - R4	9.15e-41	2.009
ICA-AROMA - I1	ICA-AROMA - I3	3.89e-62	3.348
ICA-AROMA - I2	ICA-AROMA - I4	3.83e-63	3.425

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	2.20e-3	0.298
Regress HMP - R2	ICA-AROMA - I2	3.34e-55	2.856
Regress HMP - R3	ICA-AROMA - I3	2.30e-35	1.74
Regress HMP - R4	ICA-AROMA - I4	6.05e-16	0.899

Supplemental Table 9. Intersubject correlation, statistical comparisons between pipelines.
See Figure 2.6a for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

Intrasubject

Legend:
- No Reg (M1)
- WM/CSF Only (M2)
- WM/CSF (R1/I1)
- WM/CSF+GSR (R2/I2)
- WM/CSF+Censor (R3/I3)
- WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	3.82e-14	1.363
Minimal - M2	Regress HMP - R1	4.76e-3	0.535
Minimal - M2	ICA-AROMA - I1	0.083	0.557

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	2.76e-20	1.939
Regress HMP - R3	Regress HMP - R4	8.50e-19	1.791
ICA-AROMA - I1	ICA-AROMA - I2	1.57e-8	0.893
ICA-AROMA - I3	ICA-AROMA - I4	1.17e-19	1.876

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	6.17e-14	1.345
Regress HMP - R2	Regress HMP - R4	1.20e-14	1.407
ICA-AROMA - I1	ICA-AROMA - I3	3.35e-23	2.252
ICA-AROMA - I2	ICA-AROMA - I4	5.30e-30	3.117

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.195	0.177
Regress HMP - R2	ICA-AROMA - I2	1.14e-9	0.987
Regress HMP - R3	ICA-AROMA - I3	3.88e-3	0.407
Regress HMP - R4	ICA-AROMA - I4	0.252	0.156

Supplemental Table 10. Intrasubject correlation, statistical comparisons between pipelines. See Figure 2.6b for main results. A paired t-test was used to compare pipelines across 56 participants. P-values are uncorrected.

Intersubject
(low motion)

Legend:
- No Reg (M1)
- WM/CSF Only (M2)
- WM/CSF (R1/I1)
- WM/CSF+GSR (R2/I2)
- WM/CSF+Censor (R3/I3)
- WM/CSF+GSR+Censor (R4/I4)

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	4.45e-17	2.287
Minimal - M2	Regress HMP - R1	5.89e-18	0.745
Minimal - M2	ICA-AROMA - I1	7.41e-13	0.412

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	8.54e-22	3.139
Regress HMP - R3	Regress HMP - R4	7.08e-21	2.956
ICA-AROMA - I1	ICA-AROMA - I2	8.77e-16	2.086
ICA-AROMA - I3	ICA-AROMA - I4	8.32e-17	2.244

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	1.85e-4	0.662
Regress HMP - R2	Regress HMP - R4	0.072	0.296
ICA-AROMA - I1	ICA-AROMA - I3	3.34e-10	1.336
ICA-AROMA - I2	ICA-AROMA - I4	3.20e-15	2.002

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	4.06e-3	0.489
Regress HMP - R2	ICA-AROMA - I2	1.22e-12	1.643
Regress HMP - R3	ICA-AROMA - I3	1.72e-9	1.25
Regress HMP - R4	ICA-AROMA - I4	0.326	0.159

Supplemental Table 11. Intersubject correlation for low motion scans, statistical comparisons between pipelines. See Figure 2.6c for main results. A paired t-test was used to compare pipelines across 40 scans. P-values are uncorrected.

Intersubject
(high motion)

Legend:
- No Reg (M1) — dark gray
- WM/CSF Only (M2) — light gray
- WM/CSF (R1/I1) — purple
- WM/CSF+GSR (R2/I2) — blue
- WM/CSF+Censor (R3/I3) — green
- WM/CSF+GSR+Censor (R4/I4) — red

Effect of Basic Preprocessing

Pipeline 1	Pipeline 2	p	Cohen's d
Minimal - M1	Minimal - M2	8.20e-37	7.924
Minimal - M2	Regress HMP - R1	6.88e-14	5.346
Minimal - M2	ICA-AROMA - I1	3.59e-17	6.992

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R2	3.92e-29	4.978
Regress HMP - R3	Regress HMP - R4	1.77e-22	3.28
ICA-AROMA - I1	ICA-AROMA - I2	1.92e-23	3.49
ICA-AROMA - I3	ICA-AROMA - I4	8.14e-27	4.314

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	Regress HMP - R3	5.10e-29	4.943
Regress HMP - R2	Regress HMP - R4	1.15e-24	3.771
ICA-AROMA - I1	ICA-AROMA - I3	7.26e-29	4.896
ICA-AROMA - I2	ICA-AROMA - I4	3.05e-37	8.131

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
Regress HMP - R1	ICA-AROMA - I1	0.713	0.059
Regress HMP - R2	ICA-AROMA - I2	2.65e-18	2.489
Regress HMP - R3	ICA-AROMA - I3	3.25e-14	1.858
Regress HMP - R4	ICA-AROMA - I4	4.11e-21	3.002

Supplemental Table 12. Intersubject correlation for high motion scan, statistical comparisons between pipelines. See Figure 2.6d for main results. A paired t-test was used to compare pipelines across 40 scans. P-values are uncorrected.

Absolute correlation with motion
(highpass)

Effect of Highpass vs Bandpass

Pipeline 1	Pipeline 2	Cohen's d
HP - Regress HMP - R1	BP - Regress HMP - R1	0.11
HP - Regress HMP - R2	BP - Regress HMP - R2	0.54
HP - Regress HMP - R3	BP - Regress HMP - R3	0.166
HP - Regress HMP - R4	BP - Regress HMP - R4	0.239
HP - ICA-AROMA - I1	BP - ICA-AROMA - I1	0.204
HP - ICA-AROMA - I2	BP - ICA-AROMA - I2	0.493
HP - ICA-AROMA - I3	BP - ICA-AROMA - I3	0.005
HP - ICA-AROMA - I4	BP - ICA-AROMA - I4	0.206

Effect of GSR

Pipeline 1	Pipeline 2	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R2	0.531
HP - Regress HMP - R3	HP - Regress HMP - R4	0.336
HP - ICA-AROMA - I1	HP - ICA-AROMA - I2	0.641
HP - ICA-AROMA - I3	HP - ICA-AROMA - I4	0.385

Effect of Censoring

Pipeline 1	Pipeline 2	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R3	0.15
HP - Regress HMP - R2	HP - Regress HMP - R4	0.035
HP - ICA-AROMA - I1	HP - ICA-AROMA - I3	0.289
HP - ICA-AROMA - I2	HP - ICA-AROMA - I4	0.023

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	Cohen's d
HP - Regress HMP - R1	HP - ICA-AROMA - I1	0.508
HP - Regress HMP - R2	HP - ICA-AROMA - I2	0.3
HP - Regress HMP - R3	HP - ICA-AROMA - I3	0.355
HP - Regress HMP - R4	HP - ICA-AROMA - I4	0.259

Supplemental Table 13. Absolute correlation with motion when using a highpass filter, statistical comparisons between pipelines. See Supplemental Figure 2.5b for main results. Cohen's d was used to compare pipelines across all 52650 edges.

192

Legend: WM/CSF (R1/I1) · WM/CSF+GSR (R2/I2) · WM/CSF+Censor (R3/I3) · WM/CSF+GSR+Censor (R4/I4)

Stability
(highpass)

Effect of Highpass vs Bandpass

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	BP - Regress HMP - R1	9.08e-36	0.643
HP - Regress HMP - R2	BP - Regress HMP - R2	1.65e-33	1.031
HP - Regress HMP - R3	BP - Regress HMP - R3	3.30e-43	3.097
HP - Regress HMP - R4	BP - Regress HMP - R4	1.67e-41	2.563
HP - ICA-AROMA - I1	BP - ICA-AROMA - I1	2.27e-33	1.454
HP - ICA-AROMA - I2	BP - ICA-AROMA - I2	4.23e-31	1.926
HP - ICA-AROMA - I3	BP - ICA-AROMA - I3	3.94e-40	2.079
HP - ICA-AROMA - I4	BP - ICA-AROMA - I4	3.01e-40	2.688

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R2	3.49e-15	1.308
HP - Regress HMP - R3	HP - Regress HMP - R4	3.73e-20	1.466
HP - ICA-AROMA - I1	HP - ICA-AROMA - I2	1.27e-21	0.993
HP - ICA-AROMA - I3	HP - ICA-AROMA - I4	7.77e-27	1.492

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R3	4.82e-11	0.785
HP - Regress HMP - R2	HP - Regress HMP - R4	9.41e-10	0.163
HP - ICA-AROMA - I1	HP - ICA-AROMA - I3	3.62e-21	1.088
HP - ICA-AROMA - I2	HP - ICA-AROMA - I4	1.31e-15	0.409

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - ICA-AROMA - I1	1.09e-39	3.656
HP - Regress HMP - R2	HP - ICA-AROMA - I2	6.68e-11	5.2
HP - Regress HMP - R3	HP - ICA-AROMA - I3	1.21e-35	3.278
HP - Regress HMP - R4	HP - ICA-AROMA - I4	3.68e-3	4.924

Supplemental Table 14. Stability when using a highpass filter, statistical comparisons between pipelines. See Supplemental Figure 2.6b for main results. A paired t-test was used to compare pipelines across 56 participants. P-values are uncorrected.

Legend:
- WM/CSF (R1/I1) — purple
- WM/CSF+GSR (R2/I2) — blue
- WM/CSF+Censor (R3/I3) — green
- WM/CSF+GSR+Censor (R4/I4) — red

Similarity
(highpass)

Effect of Highpass vs Bandpass

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	BP - Regress HMP - R1	1.92e-112	1.568
HP - Regress HMP - R2	BP - Regress HMP - R2	3.40e-107	0.171
HP - Regress HMP - R3	BP - Regress HMP - R3	5.84e-123	6.759
HP - Regress HMP - R4	BP - Regress HMP - R4	4.40e-128	11.217
HP - ICA-AROMA - I1	BP - ICA-AROMA - I1	3.62e-122	0.073
HP - ICA-AROMA - I2	BP - ICA-AROMA - I2	1.72e-115	1.552
HP - ICA-AROMA - I3	BP - ICA-AROMA - I3	3.85e-123	6.881
HP - ICA-AROMA - I4	BP - ICA-AROMA - I4	4.04e-127	9.468

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R2	0.445	1.275
HP - Regress HMP - R3	HP - Regress HMP - R4	2.52e-31	3.004
HP - ICA-AROMA - I1	HP - ICA-AROMA - I2	2.53e-95	2.249
HP - ICA-AROMA - I3	HP - ICA-AROMA - I4	1.77e-110	3.352

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R3	0.478	0.747
HP - Regress HMP - R2	HP - Regress HMP - R4	3.14e-45	3.391
HP - ICA-AROMA - I1	HP - ICA-AROMA - I3	3.93e-41	0.509
HP - ICA-AROMA - I2	HP - ICA-AROMA - I4	3.39e-62	3.023

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - ICA-AROMA - I1	5.34e-49	8.838
HP - Regress HMP - R2	HP - ICA-AROMA - I2	4.44e-7	13.681
HP - Regress HMP - R3	HP - ICA-AROMA - I3	1.29e-20	10.517
HP - Regress HMP - R4	HP - ICA-AROMA - I4	1.18e-57	13.409

Supplemental Table 15. Group similarity when using a highpass filter, statistical comparisons between pipelines. See Supplemental Figure 2.6b for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

194

Individualization
(highpass)

Effect of Highpass vs Bandpass

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	BP - Regress HMP - R1	0.017	0.302
HP - Regress HMP - R2	BP - Regress HMP - R2	1.30e-7	0.051
HP - Regress HMP - R3	BP - Regress HMP - R3	8.35e-44	0.105
HP - Regress HMP - R4	BP - Regress HMP - R4	2.62e-10	0.268
HP - ICA-AROMA - I1	BP - ICA-AROMA - I1	2.34e-4	1.27
HP - ICA-AROMA - I2	BP - ICA-AROMA - I2	2.87e-3	0.881
HP - ICA-AROMA - I3	BP - ICA-AROMA - I3	6.26e-13	0.058
HP - ICA-AROMA - I4	BP - ICA-AROMA - I4	6.02e-7	0.078

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R2	6.62e-25	1.612
HP - Regress HMP - R3	HP - Regress HMP - R4	1.64e-15	0.854
HP - ICA-AROMA - I1	HP - ICA-AROMA - I2	0.541	1.385
HP - ICA-AROMA - I3	HP - ICA-AROMA - I4	0.414	1.056

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R3	9.46e-26	1.614
HP - Regress HMP - R2	HP - Regress HMP - R4	1.43e-27	3.404
HP - ICA-AROMA - I1	HP - ICA-AROMA - I3	4.53e-29	0.867
HP - ICA-AROMA - I2	HP - ICA-AROMA - I4	8.89e-20	2.654

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - ICA-AROMA - I1	1.34e-26	0.536
HP - Regress HMP - R2	HP - ICA-AROMA - I2	3.43e-15	0.66
HP - Regress HMP - R3	HP - ICA-AROMA - I3	5.81e-65	0.289
HP - Regress HMP - R4	HP - ICA-AROMA - I4	4.58e-52	0.503

Supplemental Table 16. Individualization when using a highpass filter, statistical comparisons between pipelines. See Supplemental Figure 2.6c for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

Intersubject
(highpass)

Effect of Highpass vs Bandpass

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	BP - Regress HMP - R1	1.08e-85	4.291
HP - Regress HMP - R2	BP - Regress HMP - R2	7.09e-78	4.118
HP - Regress HMP - R3	BP - Regress HMP - R3	4.23e-74	3.969
HP - Regress HMP - R4	BP - Regress HMP - R4	1.86e-74	3.737
HP - ICA-AROMA - I1	BP - ICA-AROMA - I1	1.89e-73	4.698
HP - ICA-AROMA - I2	BP - ICA-AROMA - I2	3.47e-77	4.386
HP - ICA-AROMA - I3	BP - ICA-AROMA - I3	6.22e-74	3.732
HP - ICA-AROMA - I4	BP - ICA-AROMA - I4	2.27e-75	3.975

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R2	1.72e-77	2.009
HP - Regress HMP - R3	HP - Regress HMP - R4	2.49e-74	3.425
HP - ICA-AROMA - I1	HP - ICA-AROMA - I2	5.42e-67	3.0
HP - ICA-AROMA - I3	HP - ICA-AROMA - I4	7.75e-70	3.516

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R3	6.76e-55	2.856
HP - Regress HMP - R2	HP - Regress HMP - R4	2.49e-57	0.899
HP - ICA-AROMA - I1	HP - ICA-AROMA - I3	3.33e-66	3.165
HP - ICA-AROMA - I2	HP - ICA-AROMA - I4	2.59e-64	2.133

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - ICA-AROMA - I1	4.92e-41	4.738
HP - Regress HMP - R2	HP - ICA-AROMA - I2	1.15e-59	4.398
HP - Regress HMP - R3	HP - ICA-AROMA - I3	6.06e-4	4.667
HP - Regress HMP - R4	HP - ICA-AROMA - I4	4.05e-43	4.487

Supplemental Table 17. Intersubject correlation when using a highpass filter, statistical comparisons between pipelines. See Supplemental Figure 2.7a for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

WM/CSF (R1/I1)   WM/CSF+Censor (R3/I3)
WM/CSF+GSR (R2/I2)   WM/CSF+GSR+Censor (R4/I4)

Intrasubject
(highpass)

Effect of Highpass vs Bandpass

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	BP - Regress HMP - R1	4.10e-26	1.939
HP - Regress HMP - R2	BP - Regress HMP - R2	4.57e-26	1.791
HP - Regress HMP - R3	BP - Regress HMP - R3	4.59e-26	0.893
HP - Regress HMP - R4	BP - Regress HMP - R4	3.24e-23	1.876
HP - ICA-AROMA - I1	BP - ICA-AROMA - I1	8.32e-26	1.823
HP - ICA-AROMA - I2	BP - ICA-AROMA - I2	3.20e-29	1.836
HP - ICA-AROMA - I3	BP - ICA-AROMA - I3	1.37e-23	0.964
HP - ICA-AROMA - I4	BP - ICA-AROMA - I4	1.69e-27	2.268

Effect of GSR

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R2	3.99e-19	1.407
HP - Regress HMP - R3	HP - Regress HMP - R4	2.91e-19	3.117
HP - ICA-AROMA - I1	HP - ICA-AROMA - I2	2.16e-9	1.696
HP - ICA-AROMA - I3	HP - ICA-AROMA - I4	2.45e-23	3.697

Effect of Censoring

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - Regress HMP - R3	1.72e-14	0.987
HP - Regress HMP - R2	HP - Regress HMP - R4	8.04e-18	0.156
HP - ICA-AROMA - I1	HP - ICA-AROMA - I3	1.41e-21	1.722
HP - ICA-AROMA - I2	HP - ICA-AROMA - I4	9.28e-34	0.738

Effect of Regress HMP vs ICA-AROMA

Pipeline 1	Pipeline 2	p	Cohen's d
HP - Regress HMP - R1	HP - ICA-AROMA - I1	5.38e-7	2.591
HP - Regress HMP - R2	HP - ICA-AROMA - I2	4.38e-18	2.254
HP - Regress HMP - R3	HP - ICA-AROMA - I3	9.20e-6	3.007
HP - Regress HMP - R4	HP - ICA-AROMA - I4	1.11e-6	2.774

Supplemental Table 18. Intrasubject correlation when using a highpass filter, statistical comparisons between pipelines. See Supplemental Figure 2.7b for main results. A paired t-test was used to compare pipelines across 56 participants. P-values are uncorrected.

thres = 0.30 fd thres = 0.20 fd
thres = 0.25 fd thres = 0.15 fd

Absolute correlation with motion
(censor comparison)

Effect of Threshold

Pipeline 1	Pipeline 2	Cohen's d
1 vol at 0.30 mm	1 vol at 0.25 mm	0.341
1 vol at 0.30 mm	1 vol at 0.20 mm	0.183
1 vol at 0.30 mm	1 vol at 0.15 mm	0.027
1 vol at 0.25 mm	1 vol at 0.20 mm	0.252
1 vol at 0.25 mm	1 vol at 0.15 mm	0.337
1 vol at 0.20 mm	1 vol at 0.15 mm	0.189
3 vol at 0.30 mm	3 vol at 0.25 mm	0.037
3 vol at 0.30 mm	3 vol at 0.20 mm	0.109
3 vol at 0.30 mm	3 vol at 0.15 mm	0.014
3 vol at 0.25 mm	3 vol at 0.20 mm	0.155
3 vol at 0.25 mm	3 vol at 0.15 mm	0.03
3 vol at 0.20 mm	3 vol at 0.15 mm	0.113
4 vol at 0.30 mm	4 vol at 0.25 mm	0.236
4 vol at 0.30 mm	4 vol at 0.20 mm	0.086
4 vol at 0.30 mm	4 vol at 0.15 mm	0.038
4 vol at 0.25 mm	4 vol at 0.20 mm	0.236
4 vol at 0.25 mm	4 vol at 0.15 mm	0.131
4 vol at 0.20 mm	4 vol at 0.15 mm	0.004

Effect of 1 vol vs 3 vol

Pipeline 1	Pipeline 2	Cohen's d
1 vol at 0.30 mm	3 vol at 0.30 mm	0.012
1 vol at 0.25 mm	3 vol at 0.25 mm	0.36
1 vol at 0.20 mm	3 vol at 0.20 mm	0.076
1 vol at 0.15 mm	3 vol at 0.15 mm	0.022

Effect of 1 vol vs 4 vol

Pipeline 1	Pipeline 2	Cohen's d
1 vol at 0.30 mm	4 vol at 0.30 mm	0.0
1 vol at 0.25 mm	4 vol at 0.25 mm	0.211
1 vol at 0.20 mm	4 vol at 0.20 mm	0.169
1 vol at 0.15 mm	4 vol at 0.15 mm	0.022

Effect of 3 vol vs 4 vol

Pipeline 1	Pipeline 2	Cohen's d
3 vol at 0.30 mm	4 vol at 0.30 mm	0.01
3 vol at 0.25 mm	4 vol at 0.25 mm	0.239
3 vol at 0.20 mm	4 vol at 0.20 mm	0.109
3 vol at 0.15 mm	4 vol at 0.15 mm	0.048

Supplemental Table 19. Absolute correlation with motion when varying censoring, statistical comparisons between pipelines. See Figure 2.8b for main results. Cohen's d was used to compare pipelines across all 52650 edges.

Stability
(censor comparison)

Legend:
- thres = 0.30 fd
- thres = 0.25 fd
- thres = 0.20 fd
- thres = 0.15 fd

Effect of Threshold

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	1 vol at 0.25 mm	0.023	0.042
1 vol at 0.30 mm	1 vol at 0.20 mm	3.83e-4	1.427
1 vol at 0.30 mm	1 vol at 0.15 mm	2.65e-16	0.999
1 vol at 0.25 mm	1 vol at 0.20 mm	0.011	2.046
1 vol at 0.25 mm	1 vol at 0.15 mm	3.54e-15	0.81
1 vol at 0.20 mm	1 vol at 0.15 mm	6.50e-15	1.177
3 vol at 0.30 mm	3 vol at 0.25 mm	2.22e-7	1.915
3 vol at 0.30 mm	3 vol at 0.20 mm	1.31e-12	1.526
3 vol at 0.30 mm	3 vol at 0.15 mm	5.58e-20	1.403
3 vol at 0.25 mm	3 vol at 0.20 mm	8.01e-10	0.66
3 vol at 0.25 mm	3 vol at 0.15 mm	8.21e-19	1.116
3 vol at 0.20 mm	3 vol at 0.15 mm	6.51e-18	0.975
4 vol at 0.30 mm	4 vol at 0.25 mm	1.08e-10	1.101
4 vol at 0.30 mm	4 vol at 0.20 mm	4.48e-14	0.813
4 vol at 0.30 mm	4 vol at 0.15 mm	2.23e-18	1.75
4 vol at 0.25 mm	4 vol at 0.20 mm	3.10e-11	0.99
4 vol at 0.25 mm	4 vol at 0.15 mm	3.37e-17	1.62
4 vol at 0.20 mm	4 vol at 0.15 mm	3.85e-17	1.707

Effect of 1 vol vs 3 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	3 vol at 0.30 mm	1.05e-9	1.705
1 vol at 0.25 mm	3 vol at 0.25 mm	1.55e-11	0.13
1 vol at 0.20 mm	3 vol at 0.20 mm	7.12e-15	1.43
1 vol at 0.15 mm	3 vol at 0.15 mm	5.46e-16	1.209

Effect of 1 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	4 vol at 0.30 mm	4.70e-11	1.148
1 vol at 0.25 mm	4 vol at 0.25 mm	7.59e-13	2.685
1 vol at 0.20 mm	4 vol at 0.20 mm	1.54e-14	0.851
1 vol at 0.15 mm	4 vol at 0.15 mm	1.96e-15	1.647

Effect of 3 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
3 vol at 0.30 mm	4 vol at 0.30 mm	1.43e-7	2.137
3 vol at 0.25 mm	4 vol at 0.25 mm	6.83e-7	1.254
3 vol at 0.20 mm	4 vol at 0.20 mm	3.89e-9	0.429
3 vol at 0.15 mm	4 vol at 0.15 mm	8.00e-11	1.632

Supplemental Table 20. Stability when varying censoring, statistical comparisons between pipelines. See Figure 2.8d for main results. A paired t-test was used to compare pipelines across 56 participants. P-values are uncorrected.

Similarity
(censor comparison)

Effect of Threshold

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	1 vol at 0.25 mm	2.17e-34	4.861
1 vol at 0.30 mm	1 vol at 0.20 mm	7.81e-41	2.676
1 vol at 0.30 mm	1 vol at 0.15 mm	1.56e-62	2.342
1 vol at 0.25 mm	1 vol at 0.20 mm	1.37e-30	1.039
1 vol at 0.25 mm	1 vol at 0.15 mm	4.10e-60	4.387
1 vol at 0.20 mm	1 vol at 0.15 mm	1.95e-67	2.165
3 vol at 0.30 mm	3 vol at 0.25 mm	5.16e-41	2.389
3 vol at 0.30 mm	3 vol at 0.20 mm	9.20e-53	2.154
3 vol at 0.30 mm	3 vol at 0.15 mm	9.31e-50	1.99
3 vol at 0.25 mm	3 vol at 0.20 mm	7.12e-47	1.162
3 vol at 0.25 mm	3 vol at 0.15 mm	7.45e-47	1.999
3 vol at 0.20 mm	3 vol at 0.15 mm	5.65e-42	1.788
4 vol at 0.30 mm	4 vol at 0.25 mm	8.38e-38	2.234
4 vol at 0.30 mm	4 vol at 0.20 mm	5.71e-46	1.491
4 vol at 0.30 mm	4 vol at 0.15 mm	6.32e-45	2.232
4 vol at 0.25 mm	4 vol at 0.20 mm	1.44e-40	2.341
4 vol at 0.25 mm	4 vol at 0.15 mm	1.88e-43	2.306
4 vol at 0.20 mm	4 vol at 0.15 mm	2.23e-40	2.178

Effect of 1 vol vs 3 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	3 vol at 0.30 mm	7.40e-47	2.072
1 vol at 0.25 mm	3 vol at 0.25 mm	6.66e-46	5.478
1 vol at 0.20 mm	3 vol at 0.20 mm	2.03e-52	3.769
1 vol at 0.15 mm	3 vol at 0.15 mm	1.70e-43	2.015

Effect of 1 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	4 vol at 0.30 mm	5.85e-45	2.602
1 vol at 0.25 mm	4 vol at 0.25 mm	2.83e-44	0.207
1 vol at 0.20 mm	4 vol at 0.20 mm	3.17e-46	1.843
1 vol at 0.15 mm	4 vol at 0.15 mm	1.64e-39	2.103

Effect of 3 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
3 vol at 0.30 mm	4 vol at 0.30 mm	5.55e-30	2.642
3 vol at 0.25 mm	4 vol at 0.25 mm	2.37e-30	2.288
3 vol at 0.20 mm	4 vol at 0.20 mm	1.03e-30	0.9
3 vol at 0.15 mm	4 vol at 0.15 mm	3.93e-29	1.989

Supplemental Table 21. Group similarity when varying censoring, statistical comparisons between pipelines. See Figure 2.8d for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

Individualization
(censor comparison)

Legend:
- thres = 0.30 fd
- thres = 0.25 fd
- thres = 0.20 fd
- thres = 0.15 fd

Effect of Threshold

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	1 vol at 0.25 mm	2.20e-3	2.872
1 vol at 0.30 mm	1 vol at 0.20 mm	0.016	0.983
1 vol at 0.30 mm	1 vol at 0.15 mm	2.08e-7	0.631
1 vol at 0.25 mm	1 vol at 0.20 mm	0.897	4.218
1 vol at 0.25 mm	1 vol at 0.15 mm	6.82e-13	3.773
1 vol at 0.20 mm	1 vol at 0.15 mm	1.81e-15	0.724
3 vol at 0.30 mm	3 vol at 0.25 mm	1.20e-8	1.475
3 vol at 0.30 mm	3 vol at 0.20 mm	1.15e-13	1.137
3 vol at 0.30 mm	3 vol at 0.15 mm	3.10e-33	1.071
3 vol at 0.25 mm	3 vol at 0.20 mm	1.18e-9	0.214
3 vol at 0.25 mm	3 vol at 0.15 mm	1.37e-31	0.683
3 vol at 0.20 mm	3 vol at 0.15 mm	2.24e-28	0.551
4 vol at 0.30 mm	4 vol at 0.25 mm	7.79e-8	0.375
4 vol at 0.30 mm	4 vol at 0.20 mm	5.50e-14	0.456
4 vol at 0.30 mm	4 vol at 0.15 mm	4.51e-25	1.277
4 vol at 0.25 mm	4 vol at 0.20 mm	7.91e-11	0.215
4 vol at 0.25 mm	4 vol at 0.15 mm	3.26e-24	1.048
4 vol at 0.20 mm	4 vol at 0.15 mm	7.46e-21	1.265

Effect of 1 vol vs 3 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	3 vol at 0.30 mm	0.026	1.42
1 vol at 0.25 mm	3 vol at 0.25 mm	1.21e-10	3.018
1 vol at 0.20 mm	3 vol at 0.20 mm	5.47e-18	0.879
1 vol at 0.15 mm	3 vol at 0.15 mm	1.00e-21	0.775

Effect of 1 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	4 vol at 0.30 mm	1.35e-4	0.484
1 vol at 0.25 mm	4 vol at 0.25 mm	2.24e-12	3.929
1 vol at 0.20 mm	4 vol at 0.20 mm	3.12e-16	0.471
1 vol at 0.15 mm	4 vol at 0.15 mm	5.19e-18	1.211

Effect of 3 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
3 vol at 0.30 mm	4 vol at 0.30 mm	4.91e-6	1.519
3 vol at 0.25 mm	4 vol at 0.25 mm	1.03e-6	0.552
3 vol at 0.20 mm	4 vol at 0.20 mm	1.02e-6	0.364
3 vol at 0.15 mm	4 vol at 0.15 mm	3.64e-10	1.101

Supplemental Table 22. Individualization when varying censoring, statistical comparisons between pipelines. See Figure 2.8e for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

Intersubject
(censor comparison)

Effect of Threshold

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	1 vol at 0.25 mm	6.73e-3	2.065
1 vol at 0.30 mm	1 vol at 0.20 mm	5.97e-4	1.486
1 vol at 0.30 mm	1 vol at 0.15 mm	1.26e-40	2.002
1 vol at 0.25 mm	1 vol at 0.20 mm	0.173	2.537
1 vol at 0.25 mm	1 vol at 0.15 mm	1.04e-28	3.164
1 vol at 0.20 mm	1 vol at 0.15 mm	7.32e-30	0.335
3 vol at 0.30 mm	3 vol at 0.25 mm	7.55e-26	2.563
3 vol at 0.30 mm	3 vol at 0.20 mm	5.79e-21	2.301
3 vol at 0.30 mm	3 vol at 0.15 mm	2.98e-66	3.671
3 vol at 0.25 mm	3 vol at 0.20 mm	1.96e-3	0.301
3 vol at 0.25 mm	3 vol at 0.15 mm	1.33e-50	1.311
3 vol at 0.20 mm	3 vol at 0.15 mm	3.74e-46	1.105
4 vol at 0.30 mm	4 vol at 0.25 mm	9.59e-25	3.568
4 vol at 0.30 mm	4 vol at 0.20 mm	4.96e-41	2.605
4 vol at 0.30 mm	4 vol at 0.15 mm	1.40e-76	4.606
4 vol at 0.25 mm	4 vol at 0.20 mm	1.09e-18	1.012
4 vol at 0.25 mm	4 vol at 0.15 mm	5.63e-65	1.264
4 vol at 0.20 mm	4 vol at 0.15 mm	2.72e-51	2.023

Effect of 1 vol vs 3 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	3 vol at 0.30 mm	1.38e-21	2.973
1 vol at 0.25 mm	3 vol at 0.25 mm	2.42e-39	1.483
1 vol at 0.20 mm	3 vol at 0.20 mm	1.67e-34	1.699
1 vol at 0.15 mm	3 vol at 0.15 mm	6.27e-57	1.131

Effect of 1 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
1 vol at 0.30 mm	4 vol at 0.30 mm	5.75e-25	3.658
1 vol at 0.25 mm	4 vol at 0.25 mm	1.15e-44	1.377
1 vol at 0.20 mm	4 vol at 0.20 mm	2.92e-50	2.542
1 vol at 0.15 mm	4 vol at 0.15 mm	4.37e-66	1.273

Effect of 3 vol vs 4 vol

Pipeline 1	Pipeline 2	p	Cohen's d
3 vol at 0.30 mm	4 vol at 0.30 mm	5.12e-10	1.988
3 vol at 0.25 mm	4 vol at 0.25 mm	2.42e-7	0.522
3 vol at 0.20 mm	4 vol at 0.20 mm	2.64e-22	1.161
3 vol at 0.15 mm	4 vol at 0.15 mm	2.28e-40	0.647

Supplemental Table 23. Intersubject correlation when varying censoring, statistical comparisons between pipelines. See Figure 2.8f for main results. A paired t-test was used to compare pipelines across 112 scans. P-values are uncorrected.

Appendix C: Beta Band Results for Chapter 4

In Chapter 4, results were presented for alpha band connectivity. Here, results are presented for **beta band connectivity (13.0 – 30.0 Hz)**. For ease of comparison, figures are labelled as either 'Figure' or 'Supplemental Figure' to match their designation in Chapter 4 or Appendix A.
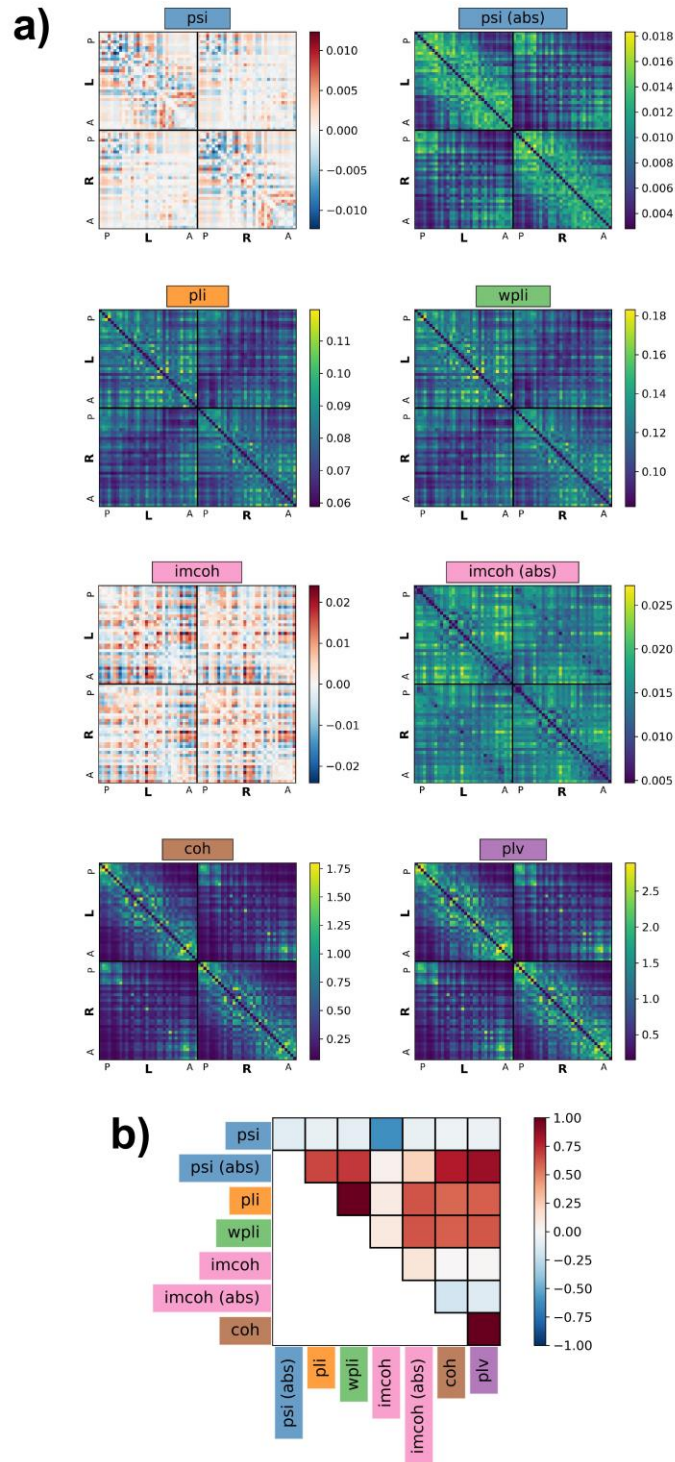
Figure C.1. Average connectomes for each connectivity measure. (a) The average connectome across all 581 recordings. Regions were divided by hemisphere (L and R) and arranged anatomically from posterior (P) to anterior (A). Since imcoh and psi have both positive and negative FC values, both the average of signed values and absolute values are shown. **(b)** The Pearson correlation between each pair of vectorized, averaged connectomes. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*
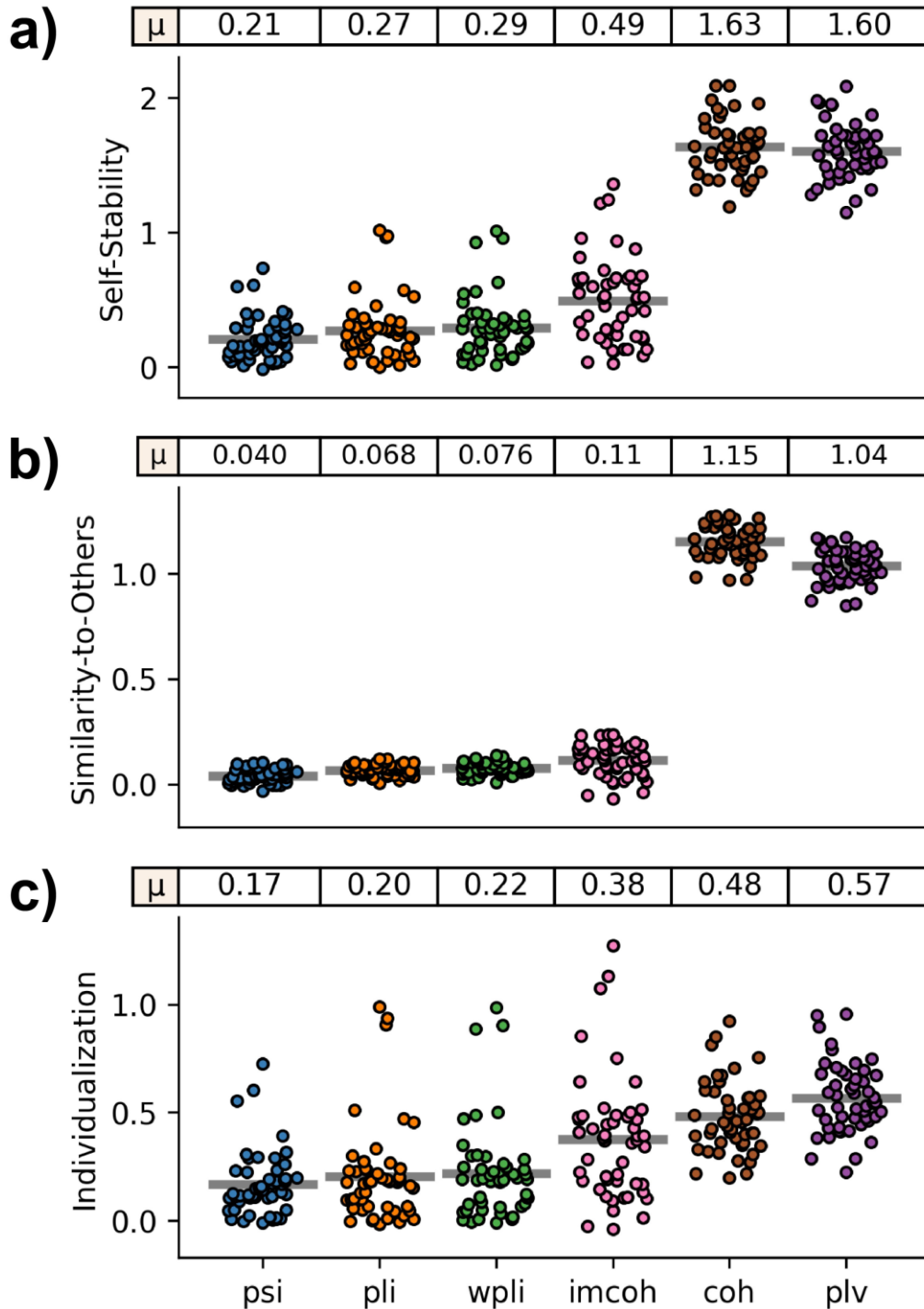
Figure C.2. Reliability and identifiability across FC measures. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **(a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on different days. **(b)** Mean similarity-to-others, the average Fisher-z correlation between a participant's connectomes and connectomes from all other participants. **(c)** Individualization, the difference between mean self-stability and mean similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Figure C.3. Task sensitivity across FC measures. Each dot represents the difference between a participant's average same-task score and their average cross-task score. Lines represent mean differences across participants, which are also displayed at the top of each subplot. For each FC measure, a paired t-test was used to assess if same-task and cross-task scores were significantly different; uncorrected p-values are shown. Bold = p < 0.05 Bonferroni corrected; italics = p < 0.05 uncorrected. **(a)** Mean change between same-task- and cross-task-self-stability. **(b)** Mean change between same-task- and cross-task-similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

μ Δ	0.023	0.017	0.017	0.020	0.025	0.025
p	3.0e-5	3.9e-6	1.5e-5	*0.015*	1.0e-3	9.6e-4

Figure C.4. Age sensitivity across FC measures. Each dot represents the difference between a participant's average similarity-to-same-age-others and their average similarity-to-different-age-others. Lines represent mean differences across participants, which are also displayed at the top of each subplot. For each FC measure, a paired t-test was used to assess if same-age- and cross-age-similarity-to-others scores were significantly different; uncorrected p-values are shown. Bold = p < 0.05 Bonferroni corrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

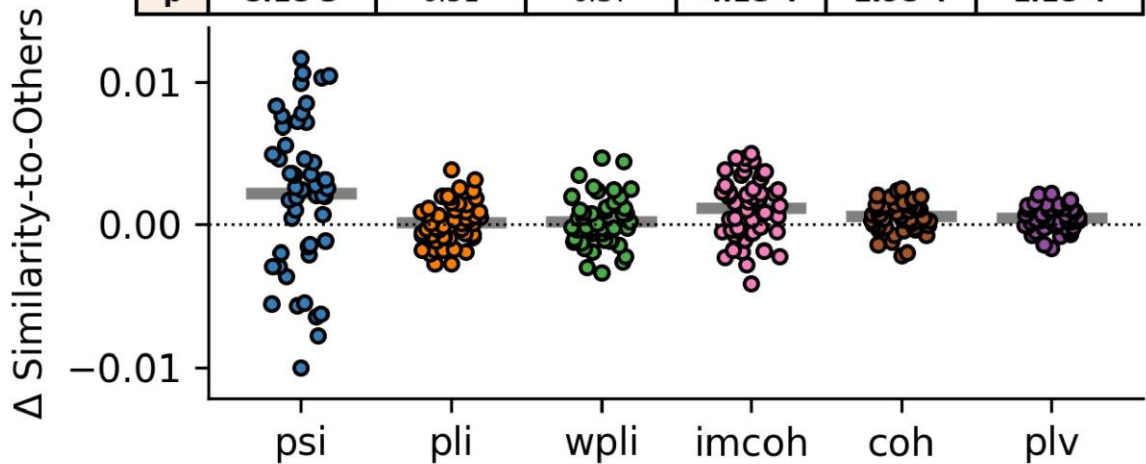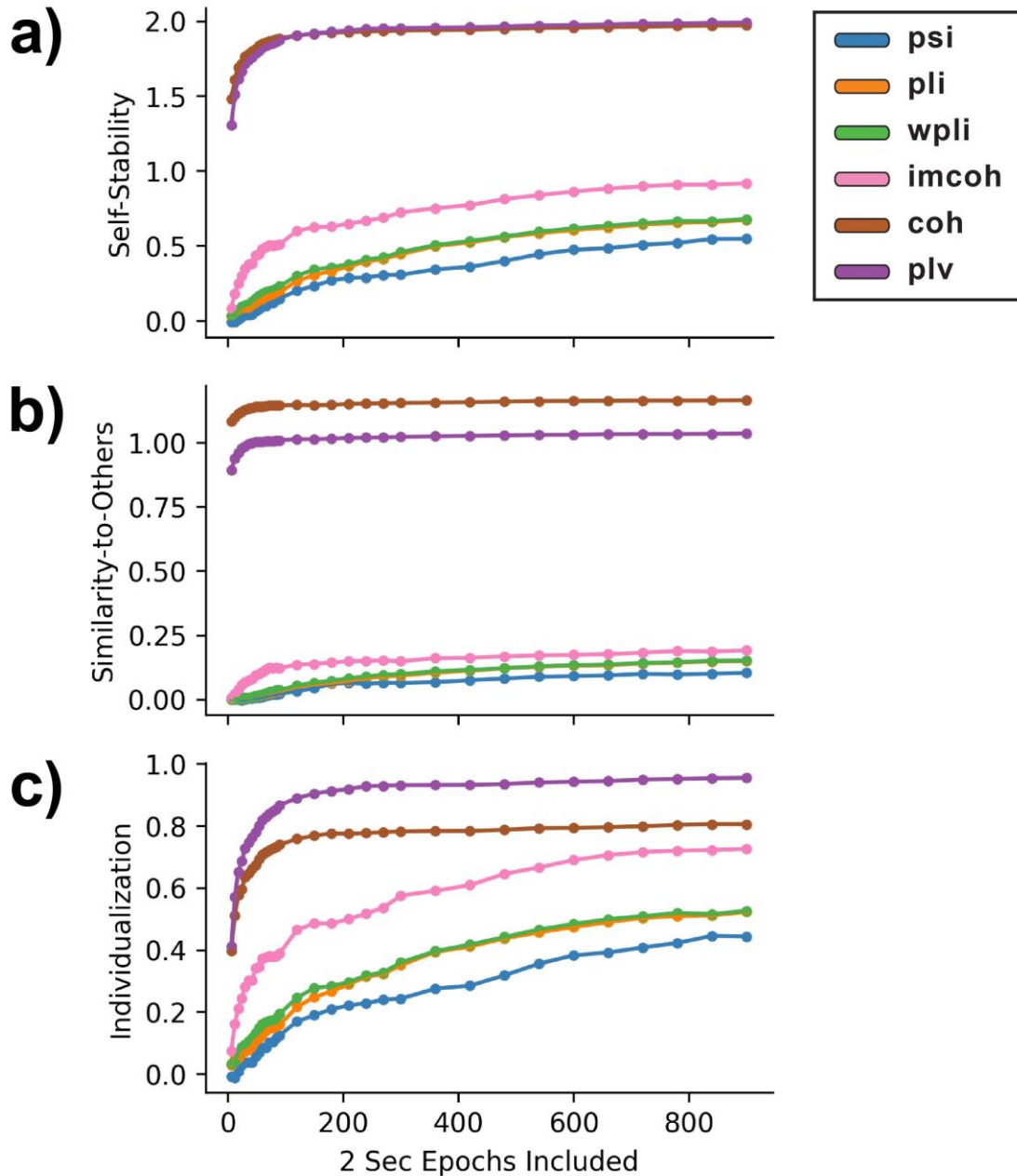**Figure C.5. Effects of different recording lengths on reliability and identifiability, across FC measures.** Points shown are across-participant averages. For each participant, the recordings from their first two sessions and from their last two sessions were concatenated, giving two sets of data per participant. Epochs were added sequentially, e.g., '60 epochs' refers to using the first 10 epochs available from the 6 separate recordings that were combined. **(a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on different days, averaged across participants. **(b)** Mean similarity-to-others, the average Fisher-z correlation between a participant's connectomes and connectomes from all other participants, averaged across participants. **(c)** Individualization, the difference between mean self-stability and mean similarity-to-others, averaged across participants. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Supplemental Figure C.1. Additional identifiability metrics across FC measures. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **(a)** Match rate, the percentage of self-stability scores that were higher than all similarity-to-others scores. **(b)** Self-stability percentile, the average percentile of self-stability scores, relative to similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*
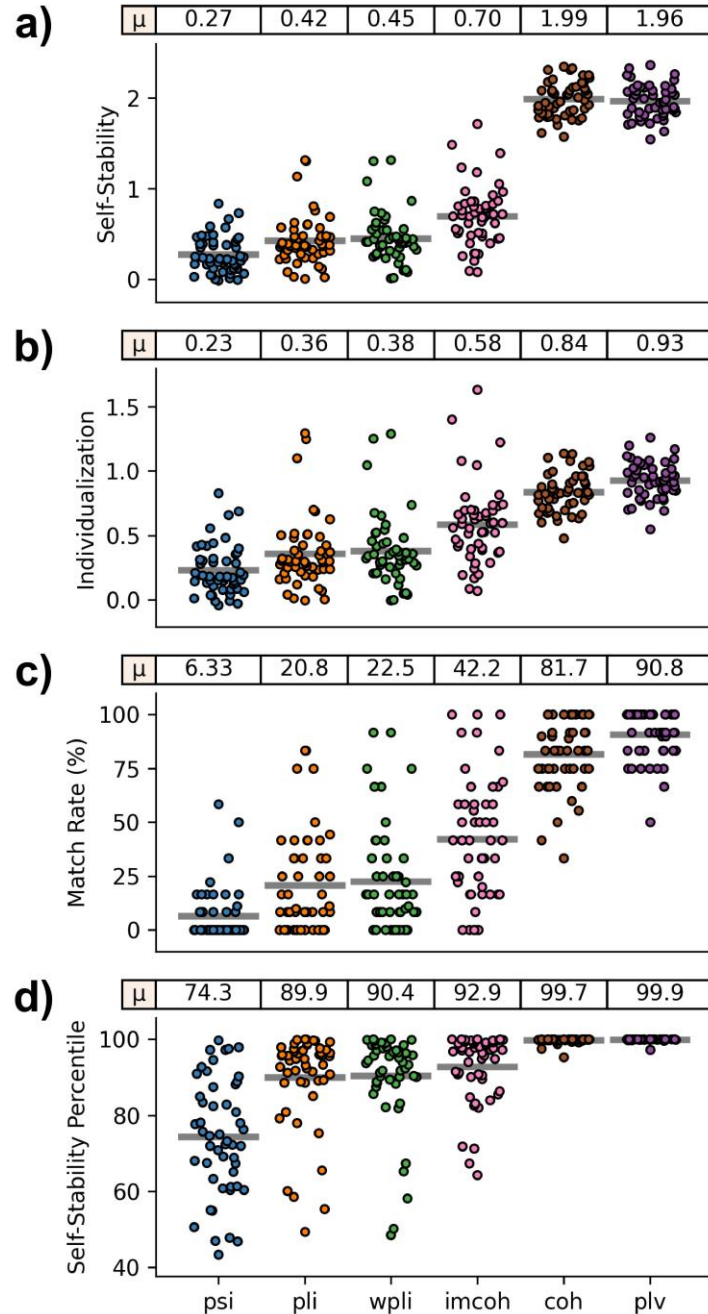
Supplemental Figure C.2. Within-session reliability and identifiability across FC measures.
This is similar to Figure C.2 and Supplemental Figure C.1 but calculating self-stability by comparing connectomes from the same session, rather than different sessions. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **(a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on the same day. **(b)** Individualization, the difference between mean self-stability and mean similarity-to-others. **(c)** Match rate, the percentage of self-stability scores that were higher than all similarity-to-others scores. **(d)** Self-stability percentile, the average percentile of self-stability scores, relative to similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 0.017	p: 2.1e-3	p: 2.7e-9	p: 1.5e-41	p: 3.1e-40
pli		p: 5.3e-5	p: 3.5e-14	p: 1.1e-37	p: 2.0e-36
wpli			p: 1.2e-12	p: 3.3e-38	p: 7.3e-37
imcoh				p: 1.9e-29	p: 2.7e-28
coh					p: 2.2e-7

Similarity-to-Others

	pli	wpli	imcoh	coh	plv
psi	p: 3.0e-5	p: 5.2e-7	p: 4.2e-7	p: 2.8e-58	p: 2.1e-55
pli		p: 1.1e-12	p: 2.1e-6	p: 4.6e-58	p: 7.3e-55
wpli			p: 5.1e-5	p: 1.8e-58	p: 3.5e-55
imcoh				p: 1.1e-50	p: 1.1e-47
coh					p: 1.1e-48

Individualization

	pli	wpli	imcoh	coh	plv
psi	p: 0.13	p: 0.045	p: 1.6e-8	p: 2.3e-16	p: 2.0e-19
pli		p: 6.3e-3	p: 1.9e-14	p: 1.6e-11	p: 7.4e-15
wpli			p: 3.9e-13	p: 2.3e-11	p: 8.7e-15
imcoh				p: 8.1e-3	p: 1.1e-5
coh					p: 7.3e-24

Match Rate (%)

	pli	wpli	imcoh	coh	plv
psi	p: 0.020	p: 4.0e-3	p: 2.1e-6	p: 7.5e-11	p: 5.1e-16
pli		p: 0.015	p: 1.5e-6	p: 2.0e-8	p: 1.0e-13
wpli			p: 8.0e-6	p: 8.2e-8	p: 4.6e-13
imcoh				p: 0.029	p: 1.2e-6
coh					p: 8.0e-14

Self-Stability Percentile

	pli	wpli	imcoh	coh	plv
psi	p: 2.2e-4	p: 7.6e-5	p: 1.9e-8	p: 3.8e-18	p: 1.1e-19
pli		p: 0.61	p: 3.0e-5	p: 9.5e-12	p: 5.9e-13
wpli			p: 6.0e-5	p: 5.8e-12	p: 3.5e-13
imcoh				p: 4.2e-8	p: 1.3e-9
coh					p: 3.2e-9

Supplemental Figure C.3. Statistical comparisons between reliability and identifiability measures, as presented in Figure C.2 and Supplemental Figure C.1. Between each FC measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 3.4e-5	p: 1.6e-6	p: 2.3e-13	p: 3.6e-42	p: 1.7e-41
pli		p: 6.9e-6	p: 2.9e-20	p: 2.5e-36	p: 1.6e-35
wpli			p: 7.2e-19	p: 7.2e-37	p: 5.8e-36
imcoh				p: 1.7e-29	p: 1.0e-28
coh					p: 3.0e-5

Individualization

	pli	wpli	imcoh	coh	plv
psi	p: 1.5e-4	p: 1.4e-5	p: 1.6e-13	p: 1.6e-24	p: 1.0e-26
pli		p: 2.8e-4	p: 3.1e-21	p: 3.5e-16	p: 1.0e-18
wpli			p: 4.4e-20	p: 4.9e-16	p: 1.3e-18
imcoh				p: 6.5e-7	p: 7.8e-10
coh					p: 3.6e-21

Match Rate (%)

	pli	wpli	imcoh	coh	plv
psi	p: 3.6e-6	p: 2.2e-6	p: 1.5e-13	p: 1.2e-29	p: 8.3e-35
pli		p: 0.19	p: 1.0e-10	p: 6.3e-20	p: 3.9e-24
wpli			p: 9.7e-9	p: 7.6e-19	p: 2.2e-23
imcoh				p: 1.2e-11	p: 3.5e-15
coh					p: 1.3e-7

Self-Stability Percentile

	pli	wpli	imcoh	coh	plv
psi	p: 5.3e-10	p: 4.2e-10	p: 5.6e-12	p: 6.7e-16	p: 4.4e-16
pli		p: 0.39	p: 2.8e-3	p: 1.0e-6	p: 6.8e-7
wpli			p: 0.019	p: 1.7e-6	p: 1.2e-6
imcoh				p: 2.8e-6	p: 1.4e-6
coh					p: 3.8e-3

Supplemental Figure C.4. Statistical comparisons between within-session reliability and identifiability measures, as presented in Supplemental Figure C.2. Between each FC measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = $p < 0.05$ Bonferroni corrected; green shading = $p < 0.05$ uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Δ Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 0.20	p: 0.093	p: 0.047	p: 0.026	p: 0.062
pli		p: 0.030	p: 0.066	p: 0.14	p: 0.36
wpli			p: 0.32	p: 0.43	p: 0.82
imcoh				p: 0.95	p: 0.69
coh					p: 0.010

Δ Similarity-to-Others

|	pli	wpli	imcoh	coh	plv
psi	p: 0.010	p: 0.013	p: 0.21	p: 0.043	p: 0.031
pli		p: 0.59	p: 8.2e-3	p: 0.091	p: 0.14
wpli			p: 0.016	p: 0.20	p: 0.29
imcoh				p: 0.11	p: 0.053
coh					p: 0.11

Supplemental Figure C.5. Statistical comparisons between FC measures of task sensitivity, as presented in Figure C.3. Between each measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = $p < 0.05$ Bonferroni corrected; green shading = $p < 0.05$ uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

|	pli	wpli	imcoh	coh	plv
psi	p: 0.18	p: 0.16	p: 0.81	p: 0.69	p: 0.79
pli		p: 0.90	p: 0.70	p: 0.16	p: 0.18
wpli			p: 0.73	p: 0.14	p: 0.16
imcoh				p: 0.72	p: 0.76
coh					p: 0.30

Δ Similarity-to-Others

Supplemental Figure C.6. Statistical comparisons between FC measures of age sensitivity, as presented in Figure C.4. Between each measure, a paired t-test was used to assess the statistical difference between participants' similarity-to-others difference. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Appendix D: Broadband Results for Chapter 4

In Chapter 4, results were presented for alpha band connectivity. Here, results are presented for **broadband connectivity (2.5 – 45.0 Hz)**. For ease of comparison, figures are labelled as either 'Figure' or 'Supplemental Figure' to match their designation in Chapter 4 or Appendix A.
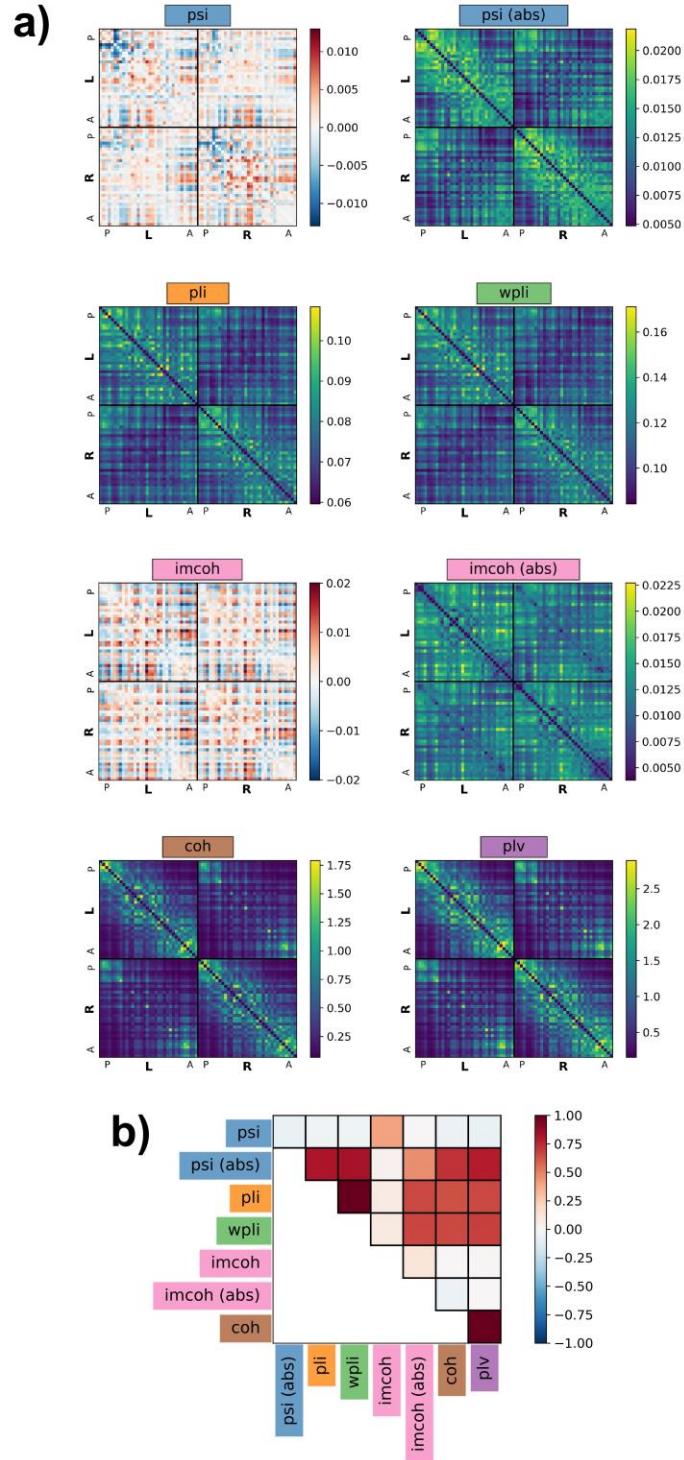
Figure D.1. Average connectomes for each connectivity measure. (a) The average connectome across all 581 recordings. Regions were divided by hemisphere (L and R) and arranged anatomically from posterior (P) to anterior (A). Since imcoh and psi have both positive and negative FC values, both the average of signed values and absolute values are shown. **(b)** The Pearson correlation between each pair of vectorized, averaged connectomes. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*
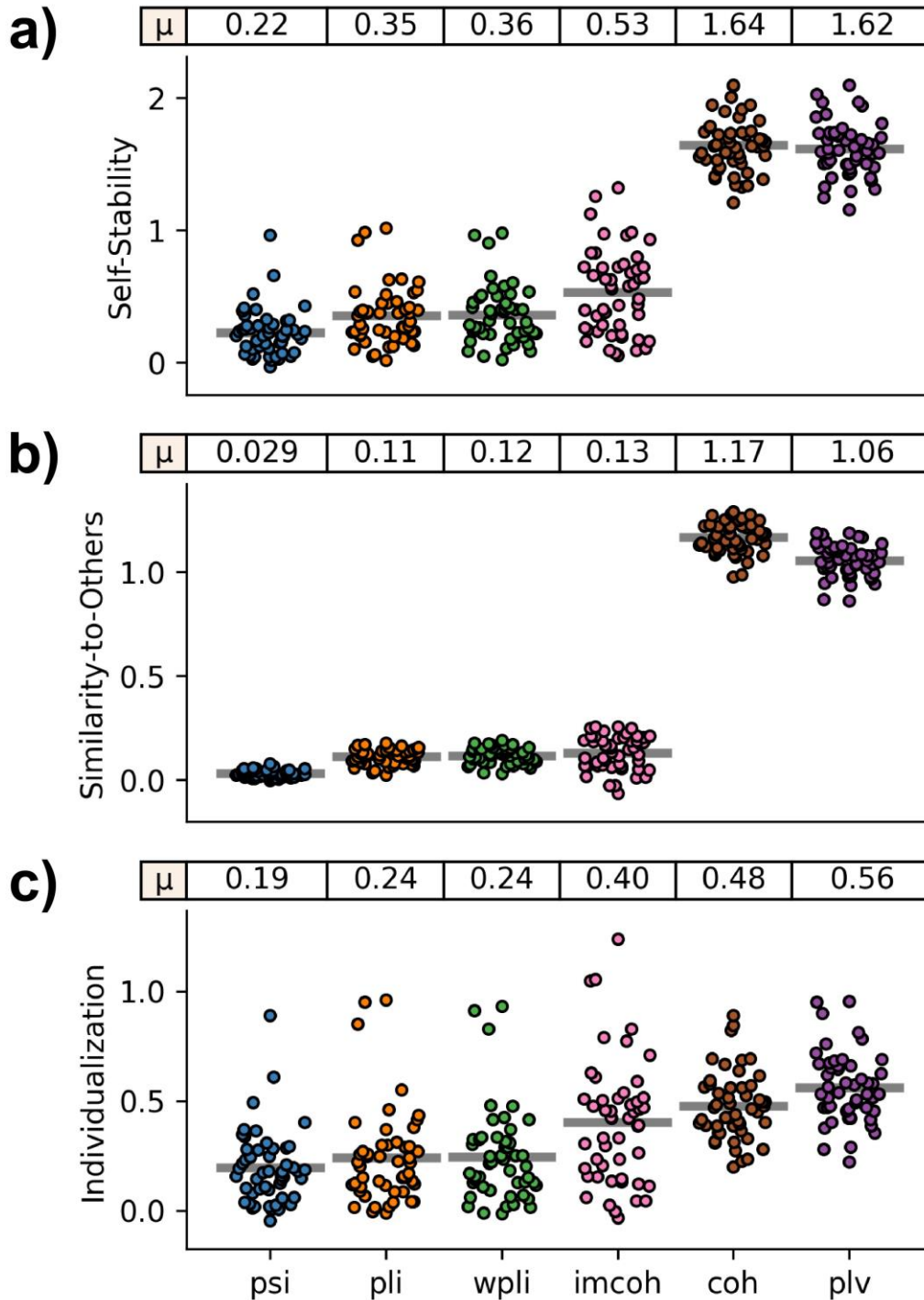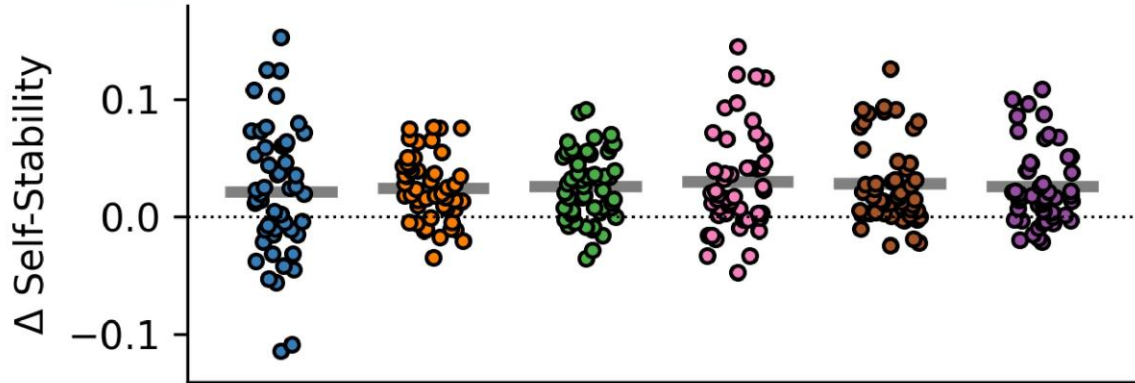
Figure D.2. Reliability and identifiability across FC measures. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **(a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on different days. **(b)** Mean similarity-to-others, the average Fisher-z correlation between a participant's connectomes and connectomes from all other participants. **(c)** Individualization, the difference between mean self-stability and mean similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

a)

μ Δ	0.021	0.025	0.026	0.030	0.028	0.026
p	*0.010*	**8.6e-8**	**9.8e-8**	**1.0e-5**	**8.0e-7**	**1.3e-6**

b)

μ Δ	2.6e-3	2.0e-4	4.9e-4	9.5e-4	5.5e-4	5.1e-4
p	**1.4e-3**	0.38	0.099	**1.9e-3**	**3.3e-4**	**1.8e-4**
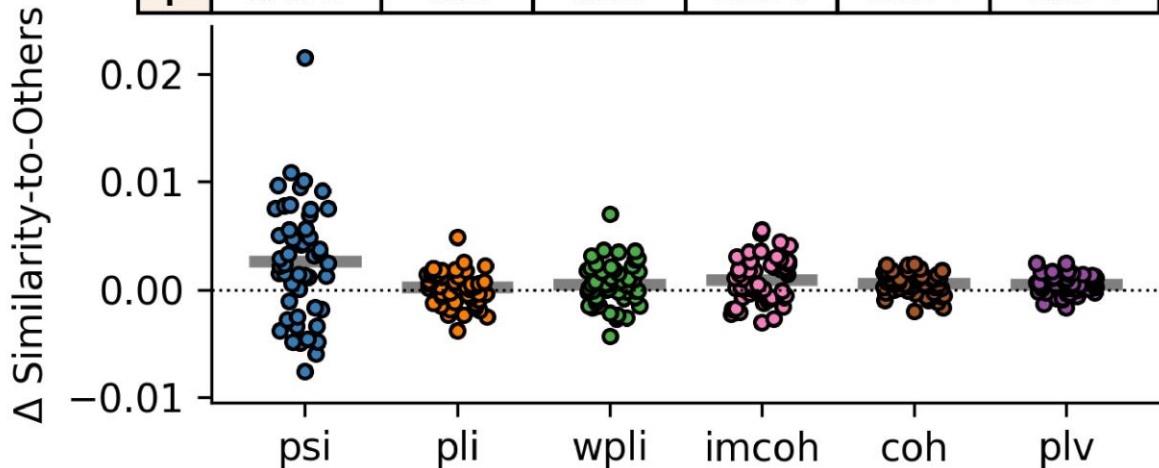
Figure D.3. Task sensitivity across FC measures. Each dot represents the difference between a participant's average same-task score and their average cross-task score. Lines represent mean differences across participants, which are also displayed at the top of each subplot. For each FC measure, a paired t-test was used to assess if same-task and cross-task scores were significantly different; uncorrected p-values are shown. Bold = p < 0.05 Bonferroni corrected; italics = p < 0.05 uncorrected. **(a)** Mean change between same-task- and cross-task-self-stability. **(b)** Mean change between same-task- and cross-task-similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

| μ Δ	0.061	0.030	0.028	0.021	0.023	0.022
p	**5.1e-10**	**8.7e-7**	**7.9e-6**	*9.8e-3*	**1.1e-4**	**8.3e-5**

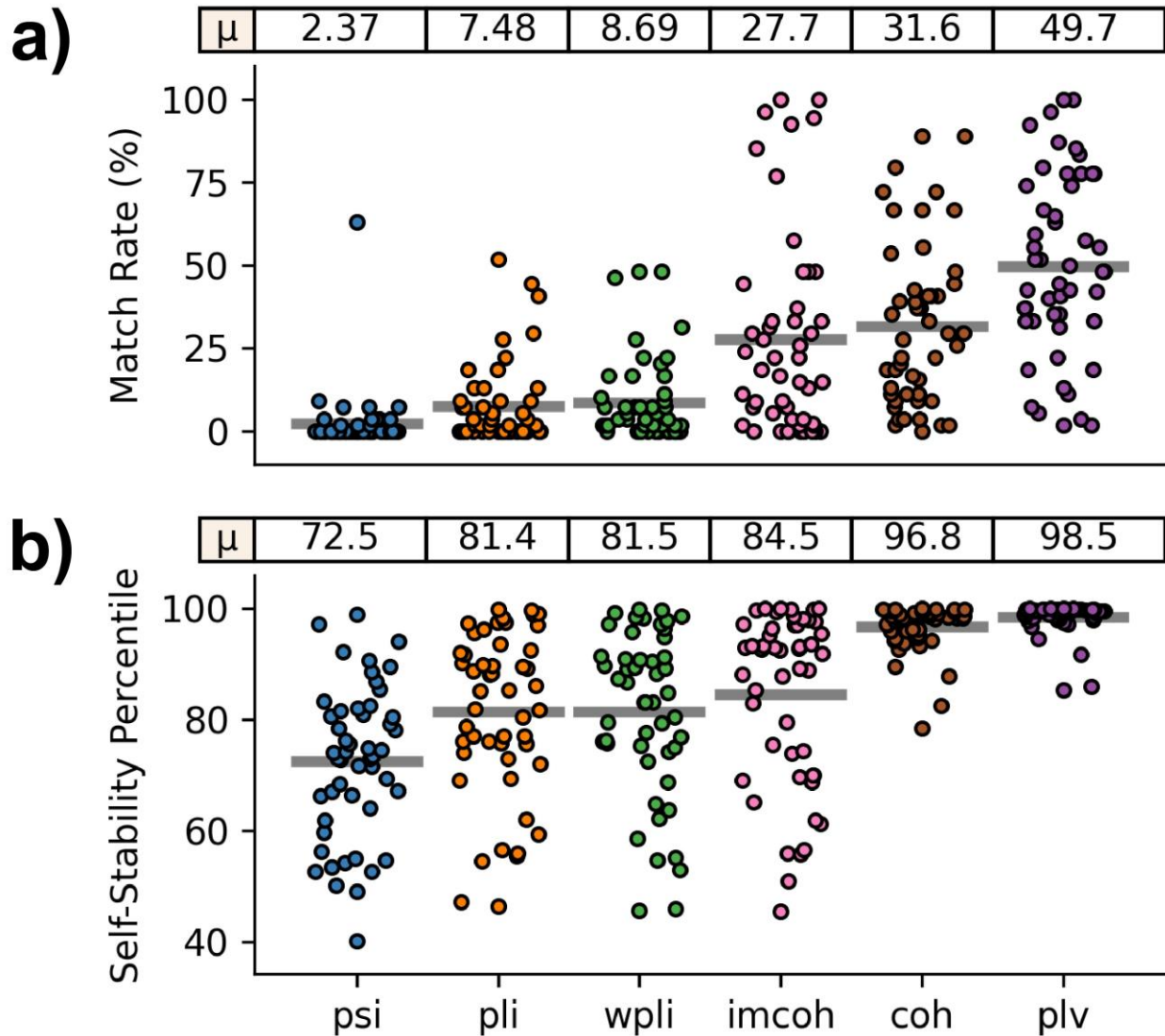**Figure D.4. Age sensitivity across FC measures.** Each dot represents the difference between a participant's average similarity-to-same-age-others and their average similarity-to-different-age-others. Lines represent mean differences across participants, which are also displayed at the top of each subplot. For each FC measure, a paired t-test was used to assess if same-age- and cross-age-similarity-to-others scores were significantly different; uncorrected p-values are shown. Bold = p < 0.05 Bonferroni corrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*
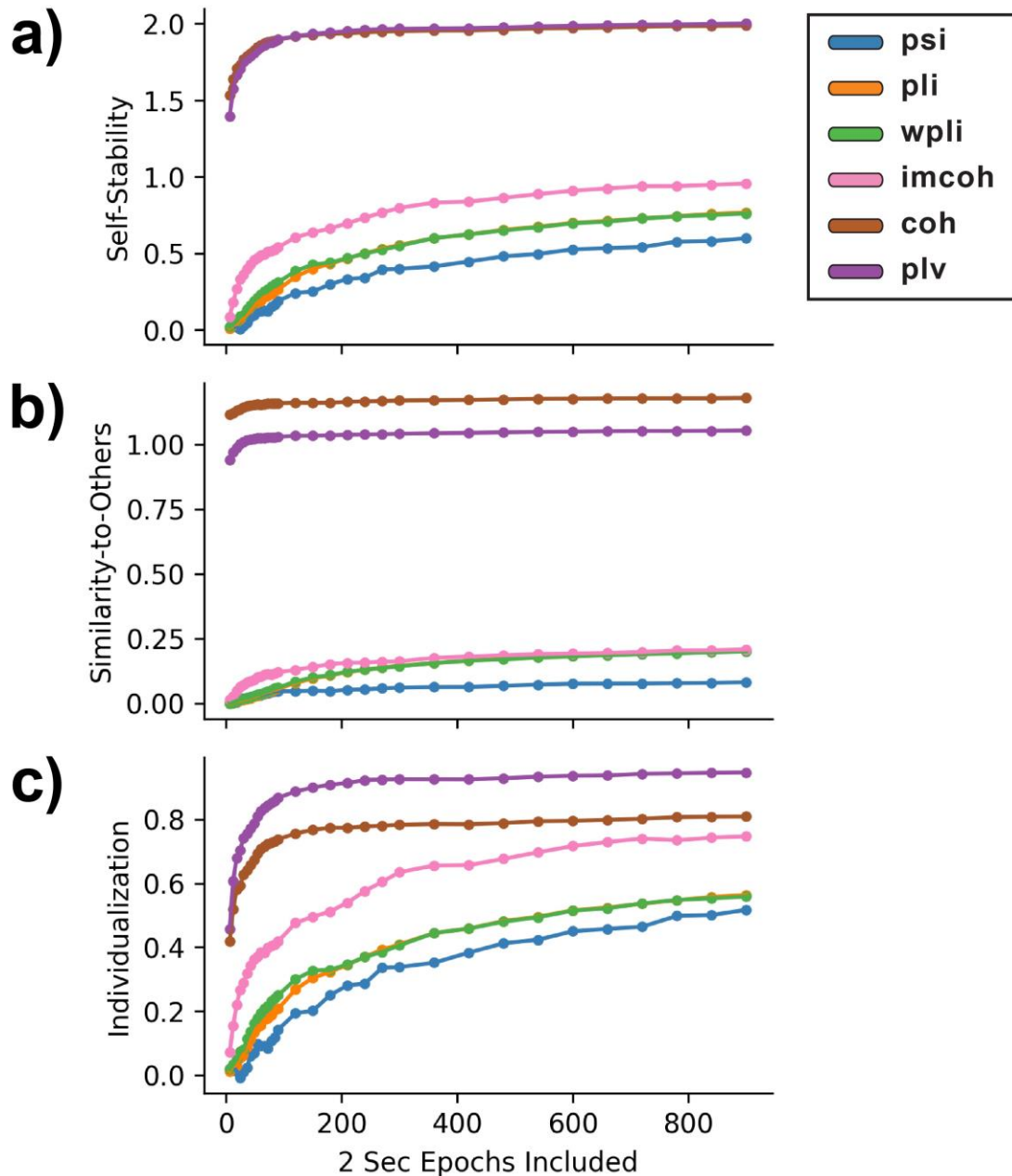
Figure D.5. Effects of different recording lengths on reliability and identifiability, across FC measures. Points shown are across-participant averages. For each participant, the recordings from their first two sessions and from their last two sessions were concatenated, giving two sets of data per participant. Epochs were added sequentially, e.g., '60 epochs' refers to using the first 10 epochs available from the 6 separate recordings that were combined. **(a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on different days, averaged across participants. **(b)** Mean similarity-to-others, the average Fisher-z correlation between a participant's connectomes and connectomes from all other participants, averaged across participants. **(c)** Individualization, the difference between mean self-stability and mean similarity-to-others, averaged across participants. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

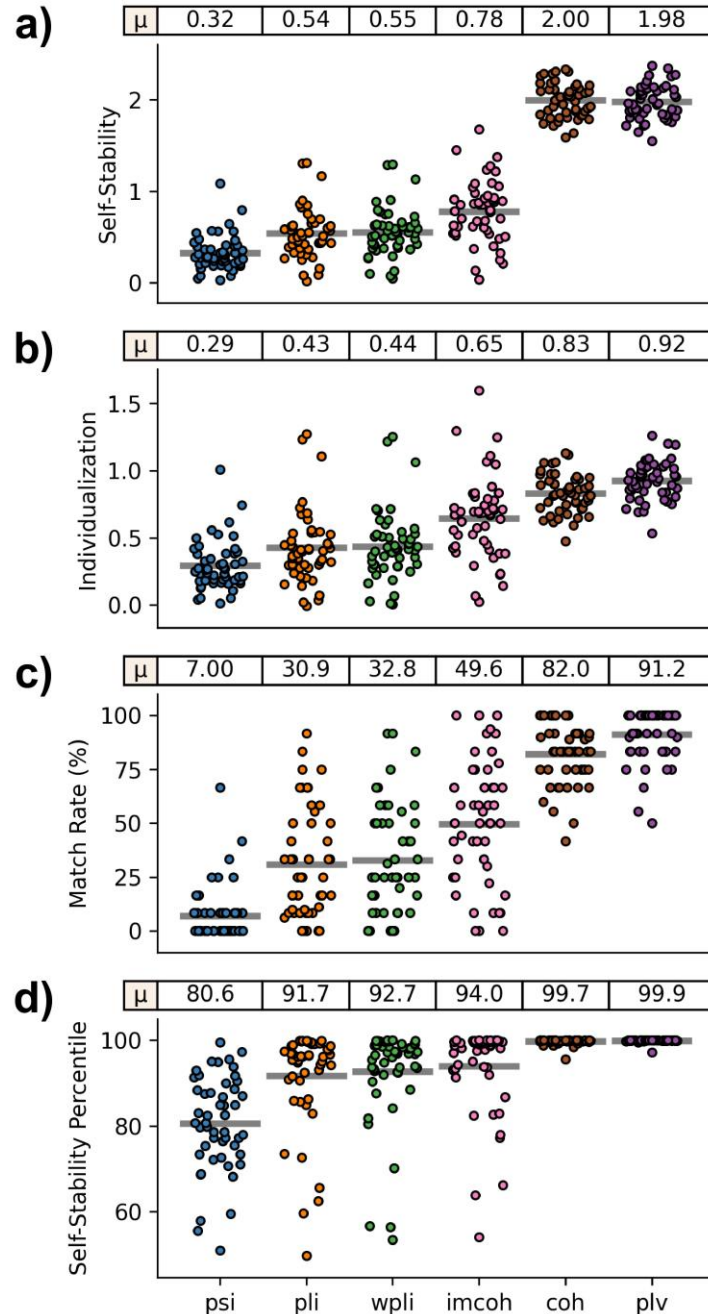**Supplemental Figure D.1. Additional identifiability metrics across FC measures.** Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **(a)** Match rate, the percentage of self-stability scores that were higher than all similarity-to-others scores. **(b)** Self-stability percentile, the average percentile of self-stability scores, relative to similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Supplemental Figure D.2. Within-session reliability and identifiability across FC measures. This is similar to Figure D.2 and Supplemental Figure D.1 but calculating self-stability by comparing connectomes from the same session, rather than different sessions. Each dot represents one participant. Lines represent mean values across participants, which are also displayed at the top of each subplot. **(a)** Mean self-stability, the average Fisher-z correlation between connectomes of the same participant, collected on the same day. **(b)** Individualization, the difference between mean self-stability and mean similarity-to-others. **(c)** Match rate, the percentage of self-stability scores that were higher than all similarity-to-others scores. **(d)** Self-stability percentile, the average percentile of self-stability scores, relative to similarity-to-others. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 6.6e-9	p: 7.2e-10	p: 5.7e-12	p: 1.4e-41	p: 4.5e-40
pli		p: 0.048	p: 1.8e-9	p: 5.1e-38	p: 1.1e-36
wpli			p: 1.2e-8	p: 1.4e-38	p: 3.1e-37
imcoh				p: 5.6e-29	p: 8.1e-28
coh					p: 9.5e-8

Similarity-to-Others

	pli	wpli	imcoh	coh	plv
psi	p: 1.9e-20	p: 8.8e-21	p: 1.3e-11	p: 5.9e-60	p: 6.2e-57
pli		p: 2.2e-6	p: 0.054	p: 5.1e-58	p: 1.5e-54
wpli			p: 0.17	p: 1.4e-58	p: 5.5e-55
imcoh				p: 7.7e-51	p: 1.2e-47
coh					p: 5.9e-45

Individualization

	pli	wpli	imcoh	coh	plv
psi	p: 0.015	p: 7.7e-3	p: 1.2e-9	p: 5.3e-14	p: 8.4e-17
pli		p: 0.44	p: 2.7e-11	p: 4.1e-10	p: 2.2e-13
wpli			p: 1.2e-10	p: 2.4e-10	p: 1.1e-13
imcoh				p: 0.055	p: 2.4e-4
coh					p: 1.2e-23

Match Rate (%)

	pli	wpli	imcoh	coh	plv
psi	p: 9.2e-4	p: 3.6e-4	p: 1.1e-7	p: 1.3e-10	p: 1.9e-15
pli		p: 0.093	p: 2.0e-7	p: 2.2e-8	p: 1.6e-13
wpli			p: 2.9e-7	p: 2.6e-8	p: 1.4e-13
imcoh				p: 0.40	p: 1.2e-4
coh					p: 1.8e-14

Self-Stability Percentile

	pli	wpli	imcoh	coh	plv
psi	p: 4.7e-6	p: 4.9e-6	p: 3.1e-7	p: 7.0e-17	p: 5.0e-18
pli		p: 0.82	p: 0.045	p: 3.7e-10	p: 2.5e-11
wpli			p: 0.052	p: 6.6e-10	p: 4.1e-11
imcoh				p: 7.0e-7	p: 3.9e-8
coh					p: 1.9e-9

Supplemental Figure D.3. Statistical comparisons between reliability and identifiability measures, as presented in Figure D.2 and Supplemental Figure D.1. Between each FC measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 9.5e-11	p: 4.8e-12	p: 1.7e-16	p: 8.9e-43	p: 6.8e-42
pli		p: 3.4e-3	p: 4.2e-15	p: 1.1e-36	p: 9.2e-36
wpli			p: 4.0e-14	p: 1.7e-37	p: 1.5e-36
imcoh				p: 3.2e-29	p: 1.9e-28
coh					p: 1.0e-3

Individualization

	pli	wpli	imcoh	coh	plv
psi	p: 7.5e-6	p: 1.1e-6	p: 8.8e-15	p: 2.8e-22	p: 1.9e-24
pli		p: 0.038	p: 1.2e-17	p: 2.7e-14	p: 5.3e-17
wpli			p: 3.1e-17	p: 1.3e-14	p: 2.1e-17
imcoh				p: 9.3e-5	p: 1.0e-7
coh					p: 8.3e-24

Match Rate (%)

	pli	wpli	imcoh	coh	plv
psi	p: 1.0e-8	p: 1.7e-9	p: 5.7e-15	p: 5.4e-29	p: 2.9e-32
pli		p: 0.17	p: 7.2e-8	p: 1.9e-17	p: 5.1e-20
wpli			p: 3.8e-8	p: 3.4e-17	p: 1.1e-19
imcoh				p: 2.3e-9	p: 1.3e-12
coh					p: 9.5e-8

Self-Stability Percentile

	pli	wpli	imcoh	coh	plv
psi	p: 6.5e-9	p: 3.3e-10	p: 3.1e-11	p: 2.6e-16	p: 1.6e-16
pli		p: 0.041	p: 0.011	p: 1.1e-5	p: 7.9e-6
wpli			p: 0.089	p: 4.8e-5	p: 3.3e-5
imcoh				p: 2.8e-4	p: 1.9e-4
coh					p: 8.8e-4

Supplemental Figure D.4. Statistical comparisons between within-session reliability and identifiability measures, as presented in Supplemental Figure D.2. Between each FC measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

Δ Self-Stability

	pli	wpli	imcoh	coh	plv
psi	p: 0.62	p: 0.53	p: 0.27	p: 0.40	p: 0.57
pli		p: 0.48	p: 0.26	p: 0.52	p: 0.82
wpli			p: 0.38	p: 0.68	p: 1.0
imcoh				p: 0.83	p: 0.58
coh					p: 0.049

Δ Similarity-to-Others

	pli	wpli	imcoh	coh	plv
psi	p: 3.4e-3	p: 0.010	p: 0.045	p: 0.011	p: 0.011
pli		p: 0.021	p: 0.036	p: 0.17	p: 0.21
wpli			p: 0.21	p: 0.84	p: 0.93
imcoh				p: 0.27	p: 0.21
coh					p: 0.41

Supplemental Figure D.5. Statistical comparisons between FC measures of task sensitivity, as presented in Figure D.3. Between each measure, a paired t-test was used to assess the statistical difference between participants' scores. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

225

Δ Similarity-to-Others	pli	wpli	imcoh	coh	plv
psi	p: 4.3e-4	p: 3.3e-4	p: 1.0e-4	p: 7.9e-5	p: 4.9e-5
pli		p: 0.14	p: 0.39	p: 0.062	p: 0.055
wpli			p: 0.51	p: 0.16	p: 0.14
imcoh				p: 0.86	p: 0.89
coh					p: 0.46

Supplemental Figure D.6. Statistical comparisons between FC measures of age sensitivity, as presented in Figure D.4. Between each measure, a paired t-test was used to assess the statistical difference between participants' similarity-to-others difference. Yellow shading = p < 0.05 Bonferroni corrected; green shading = p < 0.05 uncorrected. *Psi: phase slope index; pli: phase lag index; wpli: weighted phase lag index; imcoh: imaginary coherence; coh: coherence; plv: phase locking value.*

226

Appendix E: Co-Author Permissions

From: Kirk Graff
To: Ryann Tansey, Amanda Ip, Christiane Rohr, Dennis Dimond, Deborah Dewey, Signe Bray, Shefali Rai, Shelly Yin, Kate Godfrey, Daria Merrikh, Tamara Vanderwal, Andrea Protzner
Subject: Request for permission to include manuscript(s) in thesis
Date: 2023-07-17 7:44 PM

Hello co-authors,

I would like to request your permission to include our co-authored papers as chapters in my doctoral thesis. Please respond to this email to indicate your approval. Copies of your responses will be appended to my thesis.

Below are the manuscripts I would like to include:

Benchmarking common preprocessing strategies in early childhood functional connectivity and intersubject correlation fMRI. Developmental Cognitive Neuroscience.
Graff, K., Tansey, R., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S.
https://doi.org/10.1016/j.dcn.2022.101087

Functional connectomes become more longitudinally self-stable, but not more distinct from others, across early childhood. Neuroimage.
Graff, K., Tansey, R., Rai, S., Ip, A., Rohr, C., Dimond, D., Dewey, D., Bray, S.
https://doi.org/10.1016/j.neuroimage.2022.119367

Reliability and validity of the electrophysiological connectome across phase-based connectivity measures.
Graff, K., Rai, S., Yin, S., Godfrey, K., Merrikh, D., Tansey, R., Vanderwal, T., Protzner, A., Bray, S.
Manuscript in preparation.

Thank you,

Kirk

From: Ryann Tansey
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-17 7:50 PM

Hi Kirk,

I give permission for the papers to appear in your thesis.

Ryann

From: Andrea Protzner
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-17 9:57 PM

You have my permission.
Cheers,
Andrea

From: Shelly Yin
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 9:29 AM

Hi Kirk,

You have my permission to include the paper that I am a co-author of in your thesis. Good luck!

Shelly

From: Tamara Vanderwal
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 9:31 AM

Hi Kirk,
I am happy to give permission to use the manuscript on which I am a co-author in your thesis.
Excellent work!
Best,
Tammy

From: Kate Godfrey
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 9:54 AM

I am happy to approve inclusion of our co-authored manuscript as one of your thesis chapters.

From: Deborah Dewey
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 9:55 AM

Dear Kirk,

You have my permission to include the papers on which I am listed as a co- author in your doctoral thesis.

From: Shefali Rai
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 1:47 PM

I approve!

From: Dennis Dimond
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 3:23 PM

Hi Kirk,

You have my unreserved approval! Good luck finishing up your thesis and congratulations!

Best,
Dennis

From: Amanda Ip
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-18 5:13 PM

Go for it! Are you defending soon?!

From: Christiane Rohr
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-21 8:59 AM

Approve of course Kirk!

Great work for your doctoral thesis and happy to have been a small part of it.

Hope you are well and wishing you Best of luck 🤞
Christiane

From: Daria Merrikh
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-21 2:58 PM

Hi Kirk,

You have my permission to include the co-authored paper as a chapter in your manuscript. Thank you so much!

Kind Regards,

Daria

From: Signe Bray
To: Kirk Graff
Subject: Re: Request for permission to include manuscript(s) in thesis
Date: 2023-07-21 6:26 PM

I approve.