Automatically Generated Report

This report was generated automatically by Brain Researcher and edited by Xinhui Li.

Functional Network Connectivity in Schizophrenia: A Multiverse Analysis

Multiverse analysis using NeuroMark 2.2 functional network connectivity from schizophrenia patients and controls

Xinhui Li, Vince D. Calhoun

Date: 2026-05-15

Abstract

We report a full factorial multiverse (specification-curve) analysis of resting-state functional network connectivity (FNC) comparing patients with schizophrenia (SZ; $n = 182$) and healthy controls (HC; $n = 181$), derived from the NeuroMark 2.2 intrinsic connectivity network (ICN) time courses [5, 6]. Five pre-defined forks (connectivity measures, confound regression strategies, dimensionality reduction methods, classifiers, domain labels) yield **320** pipelines ($4 \times 2 \times 5 \times 4 \times 2$); each evaluates three hypotheses [4, 8, 10, 11]. **H2**—greater mean absolute Cohen's d on between-domain than within-domain edges—is highly choice-robust (93.8% favourable specifications; small median effect size). **H3**—greater between-domain loading mass in latent factors—is intermediate (55.5% favourable; ICA-sensitive). **H1**—nested-CV ROC-AUC gain for latent versus raw edges—is fragile at the median (median ΔAUC < 0; 18% favourable) yet strongly fork-dependent (ICA, OLS confounds, partial correlation). Joint binomial count tests reject a stylized independent-α null for all three hypotheses but must be read alongside medians and curves [10].

Contents

1 Introduction **1**
- 1.1 Clinical and methodological background . . . 1
- 1.2 Multiverse analysis . . . 1

2 Hypotheses **3**
- 2.1 H1: Latent versus edge-based classification . . . 3
- 2.2 H2: Between-domain versus within-domain effect sizes . . . 3
- 2.3 H3: Between-domain versus within-domain loading mass in latent factors . . . 4

3 Methods **5**
- 3.1 Dataset . . . 5
- 3.2 NeuroMark template . . . 6
- 3.3 Connectivity estimation methods (D1) . . . 6
- 3.4 Confounding variables (D2) . . . 7
- 3.5 Dimensionality reduction methods (D3) . . . 7
- 3.6 Classifiers (D4) . . . 8
- 3.7 Domain labels (D5) . . . 8
- 3.8 Multiverse grid, implementation, and reproducibility . . . 9

4 Results **12**
- 4.1 H1: Raw edges achieve higher median nested-CV ROC-AUC than latent factors; latent gains concentrate under ICA, OLS confounds, and partial-correlation connectivity. 12
- 4.2 H2: Between-domain edges show larger mean $|d|$ than within-domain edges in most pipelines (93.8%), with a modest median gap. . . . 12
- 4.3 H3: A slim majority of latent pipelines favour excess between-domain loading mass (55.5%), driven especially by ICA and correlation-based connectivity. . . . 13

5 Discussion **20**
- 5.1 H1: Edge-level prediction remains the conservative default; latent models are exploratory and hinge on ICA, nuisance regression, and partial correlations. . . . 20
- 5.2 H2: Domain enrichment supports long-range dysconnectivity narratives statistically but does not localize circuits or prove causal mechanisms. . . . 20
- 5.3 H3: ICA-linked loading structure motivates mechanistic hypotheses while pooled estimation limits biomarker claims. . . . 21
- 5.4 Limitations . . . 21
- 5.5 Future directions . . . 21

6 Brain Researcher MCP review **23**
6.1 Scientific review 23
6.2 Code review 23
6.3 Reviewer questions 23

7 Conclusion **24**

List of Figures

1 **Study design schematic.** Inputs (ICN time courses, labels, confounds), crossed multiverse forks (D1–D5), and parallel H1–H3 branches. The workflow ties high-dimensional edges to three complementary questions—latent versus edge-based classification (H1), between-domain versus within-domain effect sizes (H2), and between-domain versus within-domain loading mass in latent factors (H3). 2

2 **Distributions of confounding variables / covariates.** Several covariates differ across groups or sites; ignoring D2 can blend diagnosis signal with shared nuisance structure. 5

3 **Spatial maps of the 105 ICNs from the NeuroMark 2.2 template.** Reprinted from "Addressing inconsistency in functional neuroimaging: A replicable data-driven multi-scale functional atlas for canonical brain networks" by Kyle M. Jensen, Jessica A. Turner, Lucina Q. Uddin, Vince D. Calhoun, and Armin Iraji, 2024, bioRxiv [6]. . . 10

4 **Group-mean FNC for HC, SZ, and group difference from different connectivity measures (Pearson Fisher-z, Spearman, partial correlation, and mutual information).** 11

5 **H1 specification curve.** One column per latent-evaluable pipeline ($N = 256$), sorted by ΔAUC (latent minus edges). The main trace shows the curve; coloured tiles beneath encode D1–D5 fork choices showing that favourable tails align with ICA, OLS, partial correlation, etc. 14

6 **H1 raincloud plot.** Marginal influence of fork levels on ΔAUC across the factorial mix. ICA and OLS visibly shift mass toward latent-favourable outcomes versus FA/NMF/PCA and `none`. 15

7 **H2 specification curve.** One column per specification ($N = 320$), sorted by the standardized H2 summary statistic (between-domain enrichment of mean $|d|$). Tiles encode forks; flat plateaus at favourable values indicate robustness across connectivity and confound choices. 16

8 **H2 raincloud plot.** Marginal distributions of H2 outcomes by fork level, illustrating stability drivers (correlation-based connectivity, OLS) versus weaker slices. 17

9 **H3 specification curve.** Latent-only specifications ($N = 256$), sorted by the plotted H3 outcome (Wilcoxon p-value). Significant rows are common but decomposition-dependent; ICA-rich columns dominate favourable extremes. 18

10 **H3 raincloud plot.** Robustness of loading-mass contrasts across forks. ICA + correlation-based connectivity lifts favourable fractions relative to PCA or MI paths. . 19

List of Tables

1 **Participant demographics**. Welch's t-test compares mean age (unequal variances); Pearson χ^2 with 1 d.f. tests sex $\times$ diagnosis association. Groups are similar in mean age and sex ratio at $\alpha = 0.05$, which supports interpreting FC contrasts as diagnosis-associated *to the extent* that other covariates are modeled. Balance on age/sex reduces trivial demographic drivers but does not substitute for D2 nuisance regression (age, sex, race, site, motion under `ols`). 5

2 **NeuroMark 2.2 ICN subdomain labels.** Parent domains shown in the second column. H2/H3 "between" edges link ICNs from different rows of this table at the chosen granularity; finer subdomains reclassify some long-range pairs as "within," changing effect summaries. 8

3 **Default factorial grid** (320 specifications). 9

4 **Multiverse robustness summary (H1–H3).** Source: `mv_robustness_summary.csv`. "Median effect" is the hypothesis-specific summary tracked in that table (ΔAUC for H1; Δ mean$|d|$ for H2; median Wilcoxon p for H3). Rank hypotheses by choice-robustness: H2 strongest, H3 intermediate, and H1 most fragile. 12

5 **Joint specification-count (binomial) tests.** Source: `mv_joint_permutation_test.csv`. Expected favourable counts assume independent per-spec Type-I rate $\alpha = 0.05$. Rejecting the independent coin-flip null does not overturn median narratives—especially for H1. 13

6 **Conditional robustness excerpt.** Source: `mv_conditional_robustness.csv`. Full CSV lists all H1–H3 $\times$ fork slices. Fork-wise CSV makes explicit *which* choices drive fragility (H1) vs. stability (H2/H3). 13

Chapter 1

Introduction

1.1 Clinical and methodological background

Schizophrenia is increasingly framed as a disorder of distributed circuits and altered functional integration, not only focal dysfunction [3, 12]. Resting-state fMRI functional connectivity (FC) summarizes temporal coupling between regions or networks; reviews and meta-analyses describe widespread FC alterations whose sign and topology depend on network identity, clinical state, and medication. Case-control FC contrasts are sensitive to motion, scanner site, age, and sex; credible inference therefore requires transparent nuisance handling alongside primary effects.

Functional network connectivity (FNC) denotes pairwise association between intrinsic connectivity network (ICN) time courses (e.g., Pearson correlation with Fisher z), vectorized to one feature vector per subject. With NeuroMark 2.2 [5], this codebase uses 105 ICNs and thus 5,460 undirected edges per subject—a systems-scale, high-dimensional representation that motivates both predictive modelling and summaries stratified by anatomical domain [2].

1.2 Multiverse analysis

FC studies require many processing and modelling choices: how connectivity is estimated, how nuisance variance is handled, whether edges are compressed to latent factors, which model is used, and how domains are defined for interpretation. Multiverse analysis runs a pre-specified grid of pipelines [11]; specification-curve displays stress-test the distribution of estimates across those pipelines [10]. Guidance on inference under families of specifications [4] and systematic reviews of FC preprocessing multiverses [8] argue against single-pipeline storytelling. Tooling such as COMET illustrates structured combinatorial workflows in neuroimaging [1].

Interpreting a specification curve requires both where the mass lies (median, spread) and how often conclusions flip under reasonable alternatives (robustness percentages). This report complements aggregated CSV summaries with plots where sorted effects are annotated by fork levels.

The implementation crosses five forks—D1 connectivity, D2 confound strategy, D3 dimensionality reduction, D4 classifier, D5 domain mask granularity—yielding 320 specifications. Each specification evaluates H1–H3.

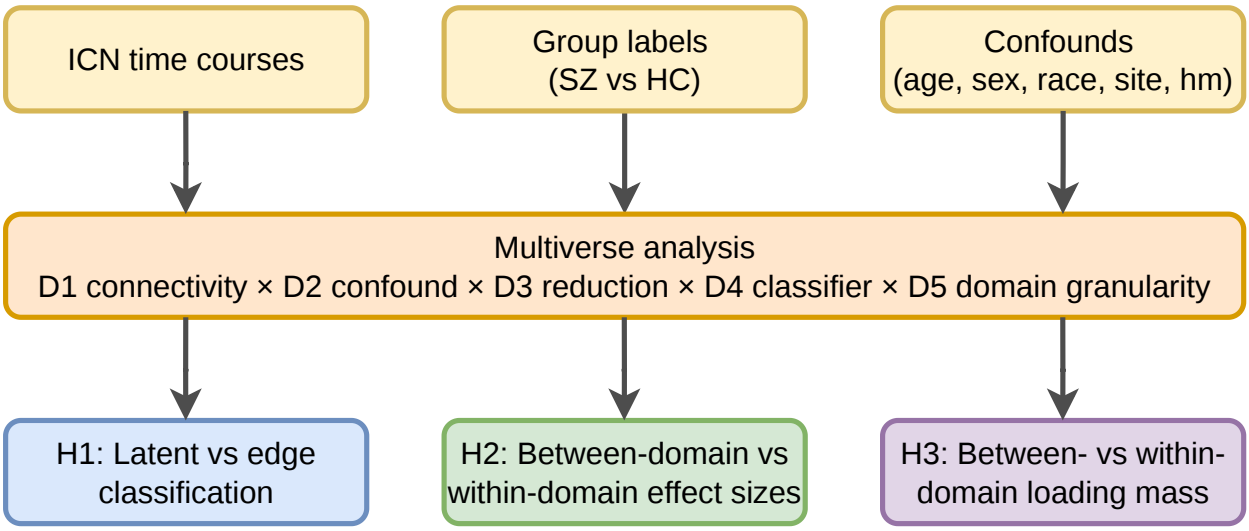

Figure 1: Study design schematic. Inputs (ICN time courses, labels, confounds), crossed multiverse forks (D1–D5), and parallel H1–H3 branches. The workflow ties high-dimensional edges to three complementary questions—latent versus edge-based classification (H1), between-domain versus within-domain effect sizes (H2), and between-domain versus within-domain loading mass in latent factors (H3).

Chapter 2

Hypotheses

2.1 H1: Latent versus edge-based classification

Motivation. All-edge models retain full pairwise information but can overfit with modest sample size and thousands of weak coefficients. Unsupervised compression (FA, ICA, PCA, NMF) may denoise and emphasize shared modes. H1 asks whether latent representations reliably improve SZ vs. HC discrimination versus the same edges classified without compression, under nested cross-validation and leakage-aware confound handling.

Operational definition. For each specification with dimension reduction, two nested-CV arms share folds: edges + classifier vs. reduced edges + same classifier family. ΔAUC $=$ mean outer-fold ROC-AUC(latent) $-$ mean outer-fold ROC-AUC(edges). Favourable if $\Delta\text{AUC} > 0$. Specifications without dimension reduction have no latent arm (256 evaluable latent specifications).

Interpretive lens. Latent superiority is not the default: most specifications favour edges at the median, yet ICA, OLS residualization, and partial-correlation connectivity disproportionately support latent gains—implying strong interactions among decomposition, nuisance control, and connectivity estimator. Classifier and domain forks shift H1 robustness only modestly relative to D2/D3/D1.

Joint count caveat. A binomial test on the *number* of favourable specifications can reject an independent coin-flip null even when the median ΔAUC is negative; read that test as global sensitivity, not as "most pipelines favour latents."

2.2 H2: Between-domain versus within-domain effect sizes

Motivation. Network models of schizophrenia often stress long-range / inter-network dysconnectivity. H2 is a descriptive enrichment question: after per-edge group effects, do between-domain edges show larger absolute Cohen's d than within-domain edges, beyond reassignment of ICNs to domains?

Operational definition. Per edge, Cohen's d (SZ $-$ HC) on the (possibly adjusted) edge matrix. $\Delta\,\text{mean}|d| = \text{mean}(|d|_{\text{between}}) - \text{mean}(|d|_{\text{within}})$. Domain labels are permuted to build a null; two-sided p. Favourable if $p < 0.05$. All 320 specifications yield finite H2 outcomes.

Interpretive lens. Robust H2 supports consistent ordering (between slightly larger on average) more than large effect gaps; it does not identify which circuits drive the pattern.

2.3 H3: Between-domain versus within-domain loading mass in latent factors

Motivation. If dysconnectivity is structured, latent factors fit to edge space might concentrate more absolute loading on between-domain than within-domain edges—an exploratory complement to H2 based on model parameters.

Operational definition. After applying the chosen dimension reduction method on scaled edges, mean $|\text{loading}|$ is summarized between- vs. within-domain per component; paired differences across components are tested with a Wilcoxon signed-rank test. Favourable if $p < 0.05$. Requires specifications with dimension reduction (256 specifications).

Interpretive lens. Agreement between H3 and H2 is strongest under ICA, consistent with seeking extended modes rather than orthogonal variance partitions (PCA). Significant Wilcoxon results support generative hypotheses for follow-up, not validated biomarkers, especially because multiverse H3 uses pooled edge fits (see Limitations).

Chapter 3

Methods

3.1 Dataset

Analyses use 363 participants with ICN time courses and labels (181 HC, 182 SZ) from the Function Biomedical Informatics Research Network (FBIRN) [7]. Table 1 shows the age and sex information of HC and SZ groups. Figure 2 shows distributions of confounding variables (age, sex, race, site, and mean framewise displacement).

Table 1: Participant demographics. Welch's t-test compares mean age (unequal variances); Pearson χ^2 with 1 d.f. tests sex $\times$ diagnosis association. Groups are similar in mean age and sex ratio at $\alpha = 0.05$, which supports interpreting FC contrasts as diagnosis-associated *to the extent* that other covariates are modeled. Balance on age/sex reduces trivial demographic drivers but does not substitute for D2 nuisance regression (age, sex, race, site, motion under `ols`).

Variable	HC ($n = 181$)	SZ ($n = 182$)	Comparison
Age (years), mean $\pm$ SD	37.4 ± 11.2	38.8 ± 11.6	$t = -1.19$, $p = 0.235$
Age (years), median [IQR]	37 [27, 47]	39 [29, 49]	—
Sex, female / male, n	54 / 127	45 / 137	$\chi^2 = 0.95$, $p = 0.330$
Female, %	29.8%	24.7%	—

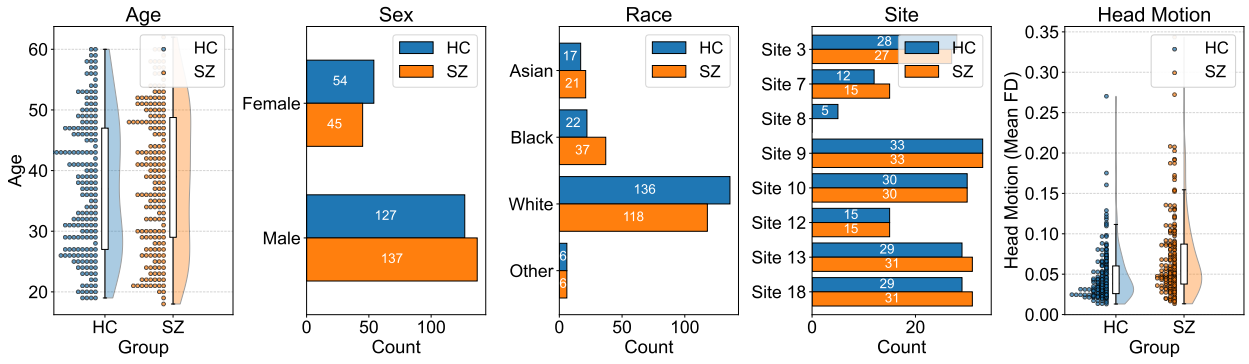

Figure 2: Distributions of confounding variables / covariates. Several covariates differ across groups or sites; ignoring D2 can blend diagnosis signal with shared nuisance structure.

3.2 NeuroMark template

Functional neuroimaging studies often disagree on what is meant by the same network name, and fixed anatomical seeds can misalign with subject- and time-varying functional sources. NeuroMark addresses this by providing a data-driven, whole-brain reference for spatially constrained independent component analysis (ICA), so intrinsic connectivity networks (ICNs) can be compared in a shared functional space across datasets and analyses [5, 6]. The NeuroMark 2.2 atlas builds on large-scale resting-state fMRI (more than 100,000 scans), uses multi-model-order spatial ICA to retain complementary information across spatial scales, and applies stringent stability and distinctness criteria to yield 105 highly replicable ICNs (Figure 3). Each ICN receives an interpretable label, and the set is organized into seven domains and fourteen subdomains (e.g., cerebellar, visual, paralimbic, subcortical, sensorimotor, higher cognition, triple-network groupings) using overlap with standard anatomical atlases, visual inspection on MNI152, and alignment with cognitive neuroscience terminology. In this study, fork D5 (7 domains vs. 14 subdomains) applies the same multi-scale domain–subdomain grouping when defining between- versus within-network masks for H2–H3, using NeuroMark-style labels supplied with the FBIRN network time courses (the multiverse summarizes pre-extracted ICN time courses and does not re-run spatial ICA from raw fMRI volumes).

3.3 Connectivity estimation methods (D1)

For each subject, an ICN×ICN association matrix is computed from the resting-state ICN time courses $(T \times 105)$; upper-triangle edges (5,460 unique pairs) form the FNC feature vector.

`pearson_z`. Pearson product-moment correlation r is computed column-wise across time (`numpy.corrcoef`, variables = ICNs). Values are clipped to $[-1, 1]$ and mapped to Fisher $z = \operatorname{arctanh}(r)$ via `fisher_z`, stabilizing variance for group comparisons.

`spearman`. Each ICN time course is replaced by its per-timepoint rank; Pearson correlation is applied to ranked series, equivalent to Spearman ρ. The result is clipped and Fisher-z transformed like Pearson.

`partial_corr`. Ledoit–Wolf shrinkage estimates a well-conditioned covariance of ICN time series; the precision matrix $\mathbf{P}$ yields pairwise partial correlations $\rho_{ij} = -P_{ij}/\sqrt{P_{ii}P_{jj}}$ (diagonal zeroed, clipped to $[-1, 1]$), then Fisher-z transformed [9]. This removes pairwise linear influence of all other ICNs under multivariate Gaussian assumptions.

`mutual_info`. For each ICN i, `sklearn.feature_selection.mutual_info_regression` estimates mutual information between other ICNs and ICN i using the Kraskov–Stögbauer–Grassberger k-nearest-neighbour estimator (`n_neighbors`=5, fixed `random_state`). Rows are symmetrized; diagonal zeroed. Values remain in sklearn's mutual information (MI) units (not Fisher-z), which affects comparability across estimators but preserves rank/nonlinear coupling sensitivity.

Figure 4 shows group-mean FNC and group difference estimated by different measures. The factorial grid treats all four estimators as equally countable coverage; weaker H2/H3 slices under MI likely reflect noise and scale differences relative to correlation-based edges.

3.4 Confounding variables (D2)

`none`. Edge matrices (and downstream features) enter group contrasts and classifiers without residualizing nuisance regressors at the forked edge stage.

`ols`. When enabled, `fbirn_experiment/confounds.py` builds a design matrix from tabulated covariates (typically age, sex, race, scanner site, head motion) and residualizes edge values by OLS (training-fold fit in CV paths so labels and confounds do not leak across folds). This targets shared demographic and acquisition variance that could otherwise inflate diagnosis effects.

3.5 Dimensionality reduction methods (D3)

All latent methods operate on standardized edge vectors (within training folds). Let $p = 5{,}460$ edges.

`none`. No compression; H1 uses only the edge arm; H3 is undefined.

Factor analysis (FA). `sklearn.decomposition.FactorAnalysis` models edges as linear combinations of latent factors plus diagonal uniqueness. On each inner-CV training split, the number of factors k is chosen from $\{5, 10, \ldots, 50\}$ (clamped to sample/feature limits) by minimizing **BIC** over the grid (`select_n_components_fa`); `FactorAnalysisTransform` then fits FA with that k for prediction.

Independent component analysis (ICA). `FastICATransform` wraps `sklearn.decomposition.FastICA` (`whiten="unit-variance"`, `max_iter`=1000). For FA/PCA, k is selected by BIC; for ICA (and NMF below), the same candidate grid is scanned using **FastICA reconstruction mean squared error** as the selection score (`select_n_components_ica`)—even when the downstream dimension reduction method is NMF, k is chosen via this ICA reconstruction objective before fitting the declared dimension reduction method.

Principal component analysis (PCA). `PCA(n_components=`k`)` with orthogonal axes maximizing variance; k from the FA/BIC path. Components are signed mixtures of edges suitable for linear classification.

Non-negative matrix factorization (NMF). Edges are shifted per feature to be non-negative (`_NMFTransform`: subtract column minima, clip), then `NMF` (`max_iter`=500) learns parts-based factors. k uses the ICA reconstruction grid noted above.

H1 always compares edges vs. latent on identical outer CV folds, with inner grids for k and classifier hyperparameters refit separately per arm.

3.6 Classifiers (D4)

Each classifier is wrapped after scaling (and optional latent block).

Elastic net logistic regression. `SGDClassifier(loss="log_loss", penalty="elasticnet")` with inner-GridSearchCV over α (log-spaced) and ℓ_1 ratio, yielding sparse or dense linear coefficients depending on the draw.

L2 logistic regression. `LogisticRegression(penalty="l2", solver="lbfgs")` with hyperparameter grid on inverse regularization strength C.

Linear support vector machine. `LinearSVC` (`dual="auto"`) provides a hinge-loss margin separator; because raw decision scores are used for ROC-AUC, calibration is implemented via `CalibratedClassifierCV(..., cv=3)` when exposing probabilities is required by the pipeline—here primarily for compatibility with nested tuning.

Random forest. `RandomForestClassifier(n_estimators=500)` captures nonlinear boundaries and interaction structure without explicit edge products.

3.7 Domain labels (D5)

Labels come from NeuroMark 2.2 metadata aligned with each ICN. Between-domain vs. within-domain masks for H2/H3 depend on whether two ICNs share the same aggregated label.

Fourteen subdomains. Fourteen composite codes (Table 2); this is the finer partition.

Seven domains. Subdomains map to seven parent domains CB, VI, PL, SC, SM, HC, TN (Visual subdomains VI-OT and VI-OC $\rightarrow$ VI; Subcortical SC-* $\rightarrow$ SC; Higher-Cognition HC-* $\rightarrow$ HC; Triple-Network TN-* $\rightarrow$ TN; singleton CB, PL, SM unchanged).

Table 2: NeuroMark 2.2 ICN subdomain labels. Parent domains shown in the second column. H2/H3 "between" edges link ICNs from different rows of this table at the chosen granularity; finer subdomains reclassify some long-range pairs as "within," changing effect summaries.

Code	Parent (7)	Description
CB	CB	Cerebellar
VI-OT	VI	Visual, occipitotemporal
VI-OC	VI	Visual, occipital
PL	PL	Paralimbic
SC-EH	SC	Subcortical, extended hippocampal
SC-ET	SC	Subcortical, extended thalamic
SC-BG	SC	Subcortical, basal ganglia
SM	SM	Sensorimotor
HC-IT	HC	Higher cognition, insular–temporal
HC-TP	HC	Higher cognition, temporoparietal
HC-FR	HC	Higher cognition, frontal
TN-CE	TN	Triple network, central executive
TN-DM	TN	Triple network, default mode
TN-SA	TN	Triple network, salience

3.8 Multiverse grid, implementation, and reproducibility

Crossing D1–D5 mimics independent choices a team might make; the value is coverage—claims surviving many cells are more credible than claims hinging on one estimator or confound policy [4].

Table 3: Default factorial grid (320 specifications).

Fork	Levels
D1 Connectivity	`pearson_z`, `spearman`, `partial_corr`, `mutual_info`
D2 Confound	`none`, `ols`
D3 Reduction	`none`, `fa`, `ica`, `pca`, `nmf`
D4 Classifier	`elasticnet`, `logistic_l2`, `svm_linear`, `rf`
D5 Domain	`domain_7`, `subdomain_14`
Count	$4 \times 2 \times 5 \times 4 \times 2 = \mathbf{320}$

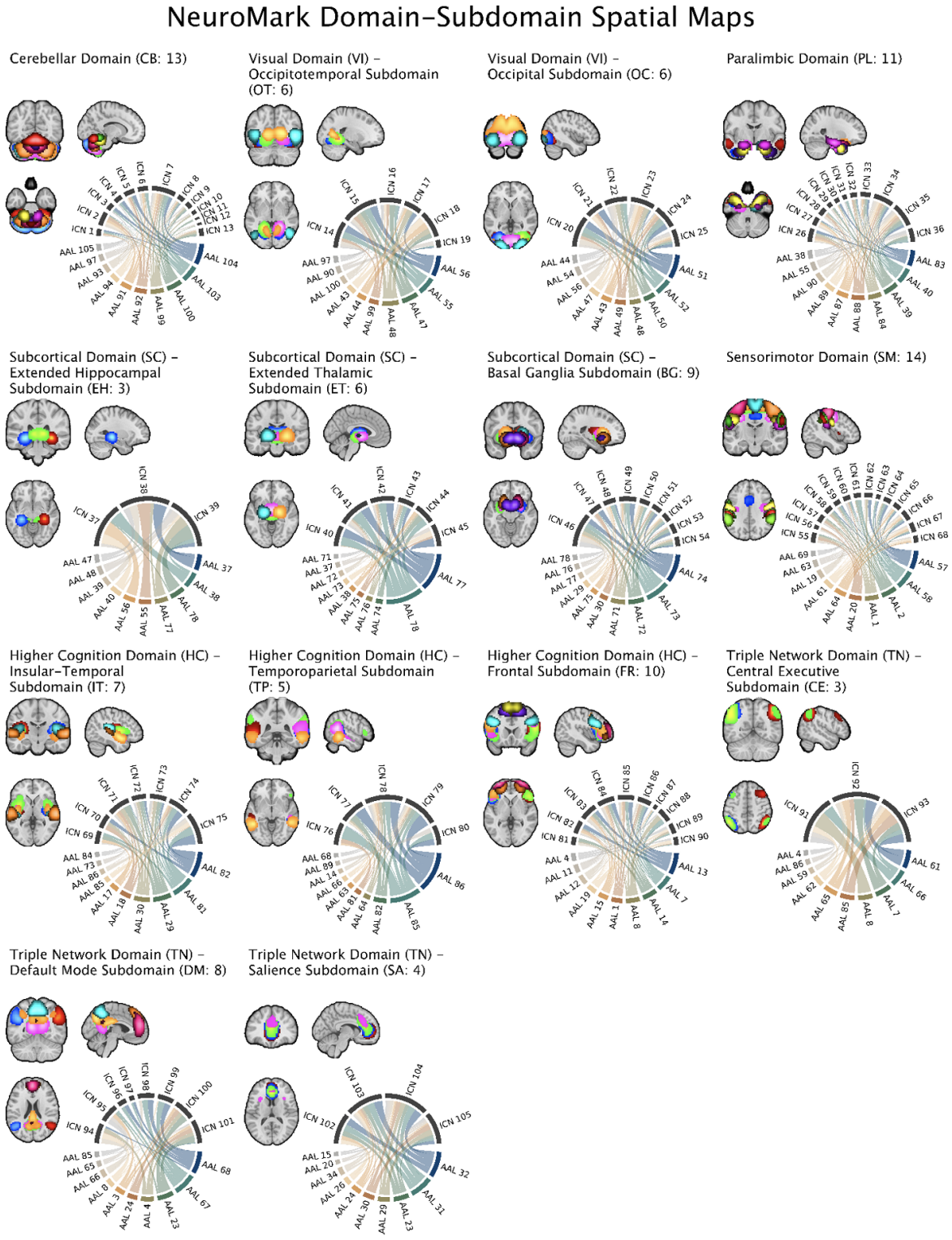

Figure 3: Spatial maps of the 105 ICNs from the NeuroMark 2.2 template. Reprinted from "Addressing inconsistency in functional neuroimaging: A replicable data-driven multi-scale functional atlas for canonical brain networks" by Kyle M. Jensen, Jessica A. Turner, Lucina Q. Uddin, Vince D. Calhoun, and Armin Iraji, 2024, bioRxiv [6].

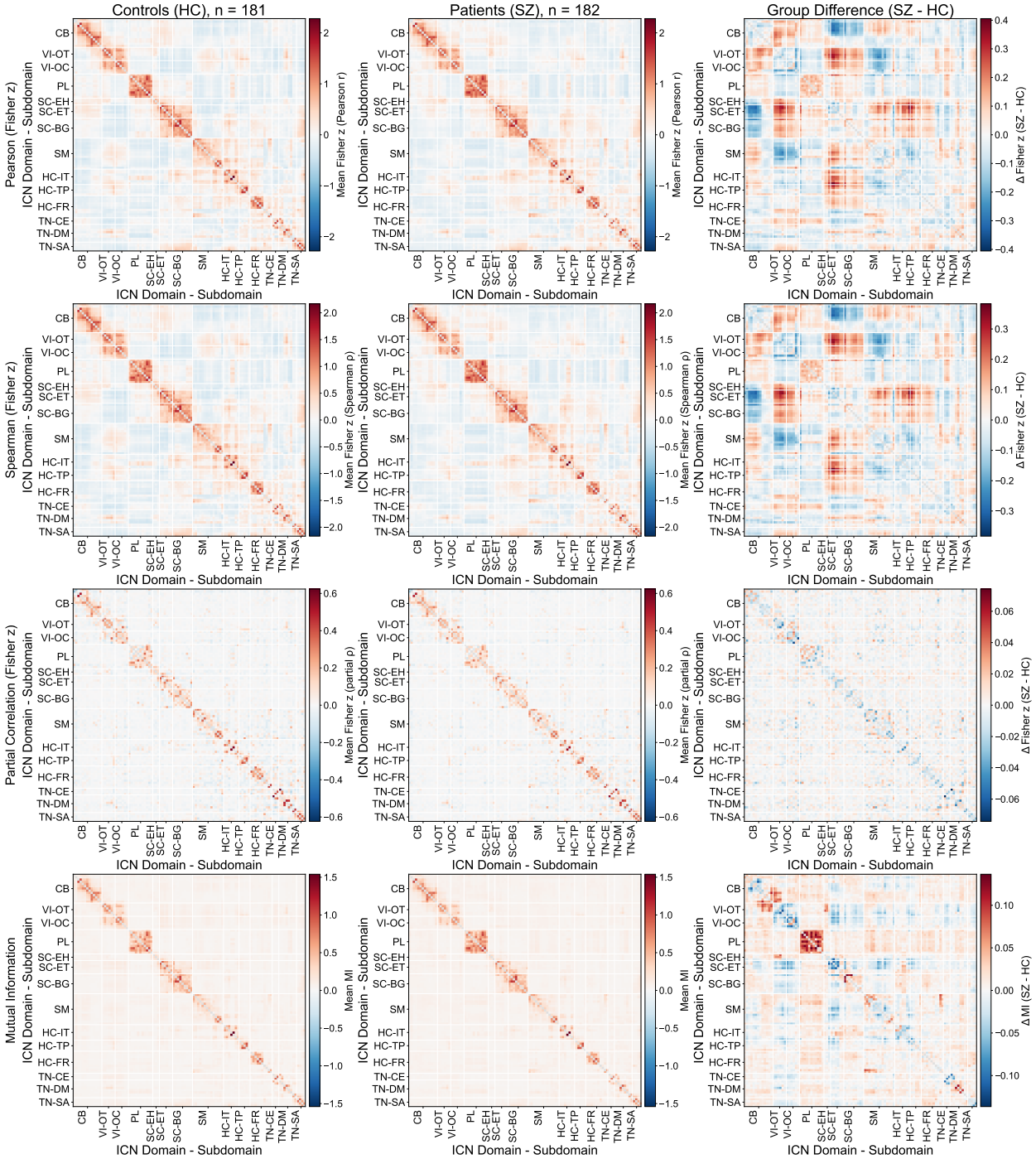

Figure 4: Group-mean FNC for HC, SZ, and group difference from different connectivity measures (Pearson Fisher-z, Spearman, partial correlation, and mutual information).

Chapter 4

Results

4.1 H1: Raw edges achieve higher median nested-CV ROC-AUC than latent factors; latent gains concentrate under ICA, OLS confounds, and partial-correlation connectivity.

Across 256 latent-evaluable specifications, 46 (18.0%) favour the latent arm; median $\Delta\text{AUC} = -0.0377$, so edges win slightly more often in aggregate. Conditional splits (`mv_conditional_robustness.csv`) reinforce path dependence: ICA (47.2% favourable within ICA rows) vs. FA (2.8%) and PCA (3.6%); OLS (23.6%) vs. `none` (10.7%); partial correlation (28.1%) vs. mutual information (6.2%). Classifier and domain forks move these percentages only modestly.

The joint binomial test rejects H_0: "independent $\alpha = 0.05$ favourable draws" ($p \approx 5.73 \times 10^{-14}$) because 46 positives still exceeds the expected 12.8 under that caricature—even though the median ΔAUC is negative. That tension is central: count-level sensitivity $\neq$ median superiority.

Table 4: Multiverse robustness summary (H1–H3). Source: `mv_robustness_summary.csv`. "Median effect" is the hypothesis-specific summary tracked in that table (ΔAUC for H1; Δ mean$|d|$ for H2; median Wilcoxon p for H3). Rank hypotheses by choice-robustness: H2 strongest, H3 intermediate, and H1 most fragile.

Hypothesis	N	Favourable	%	Median effect
H1: Latent $>$ edges	256	46	18.0	-0.0377
H2: Between $>$ within	320	300	93.8	0.0079
H3: Loading advantage	256	142	55.5	0.0330

4.2 H2: Between-domain edges show larger mean $|d|$ than within-domain edges in most pipelines (93.8%), with a modest median gap.

300/320 (93.8%) specifications are favourable with median $\Delta\,\text{mean}|d| = 0.0079$—a small absolute increment implying consistent ordering more than large separation. Pearson (Fisher z), Spearman, partial correlation, OLS, and 14-subdomains reach 100% favourable counts; mutual information

(75%), `none` (86.1%), and 7 domains (87.5%) remain majority favourable but weaker, consistent with noisier estimators and coarser masks blurring between/within contrasts.

OLS exceeding `none` suggests part of the raw gap can track shared nuisance structure; residual confounding remains possible after regression.

4.3 H3: A slim majority of latent pipelines favour excess between-domain loading mass (55.5%), driven especially by ICA and correlation-based connectivity.

142/256 (55.5%) latent specifications are favourable; median Wilcoxon $p \approx 0.033$ (Table 4). ICA (77.8% within-slice favourable) dominates PCA (35.7%) and NMF (50%), highlighting sensitivity to identifiability constraints. Pearson (66.7%) and Spearman (67.5%) outperform partial correlation (37.5%) and mutual information (50%). Seven domains (60.9%) exceed fourteen subdomains (50.0%), so finer labels need not strengthen H3—possibly via repartitioning long-range edges or adding variance to paired factor summaries.

Table 5: Joint specification-count (binomial) tests. Source: `mv_joint_permutation_test.csv`. Expected favourable counts assume independent per-spec Type-I rate $\alpha = 0.05$. Rejecting the independent coin-flip null does not overturn median narratives—especially for H1.

Hypothesis	N	Fav.	%	$E[\text{Fav}]$	Binomial p
H1	256	46	18.0	12.8	$\approx 5.7 \times 10^{-14}$
H2	320	300	93.8	16.0	≈ 0
H3	256	142	55.5	12.8	$\approx 6.8 \times 10^{-113}$

Table 6: Conditional robustness excerpt. Source: `mv_conditional_robustness.csv`. Full CSV lists all H1–H3 $\times$ fork slices. Fork-wise CSV makes explicit *which* choices drive fragility (H1) vs. stability (H2/H3).

H	Fork level	Total	# Sig.	%
H1	ICA reduction	72	34	47.2
H1	PCA reduction	56	2	3.6
H1	Partial correlation	64	18	28.1
H2	Pearson (Fisher z)	56	56	100.0
H2	Mutual information	80	60	75.0
H2	OLS confound	176	176	100.0
H3	ICA reduction	72	56	77.8
H3	PCA reduction	56	20	35.7

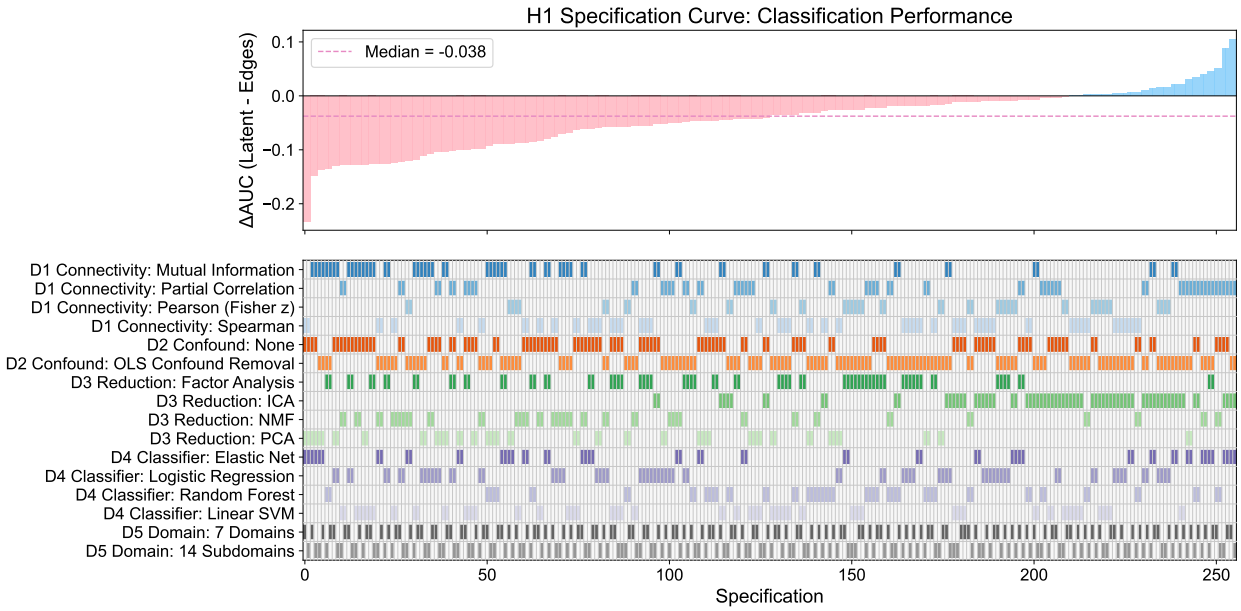

Figure 5: H1 specification curve. One column per latent-evaluable pipeline ($N = 256$), sorted by ΔAUC (latent minus edges). The main trace shows the curve; coloured tiles beneath encode D1–D5 fork choices showing that favourable tails align with ICA, OLS, partial correlation, etc.

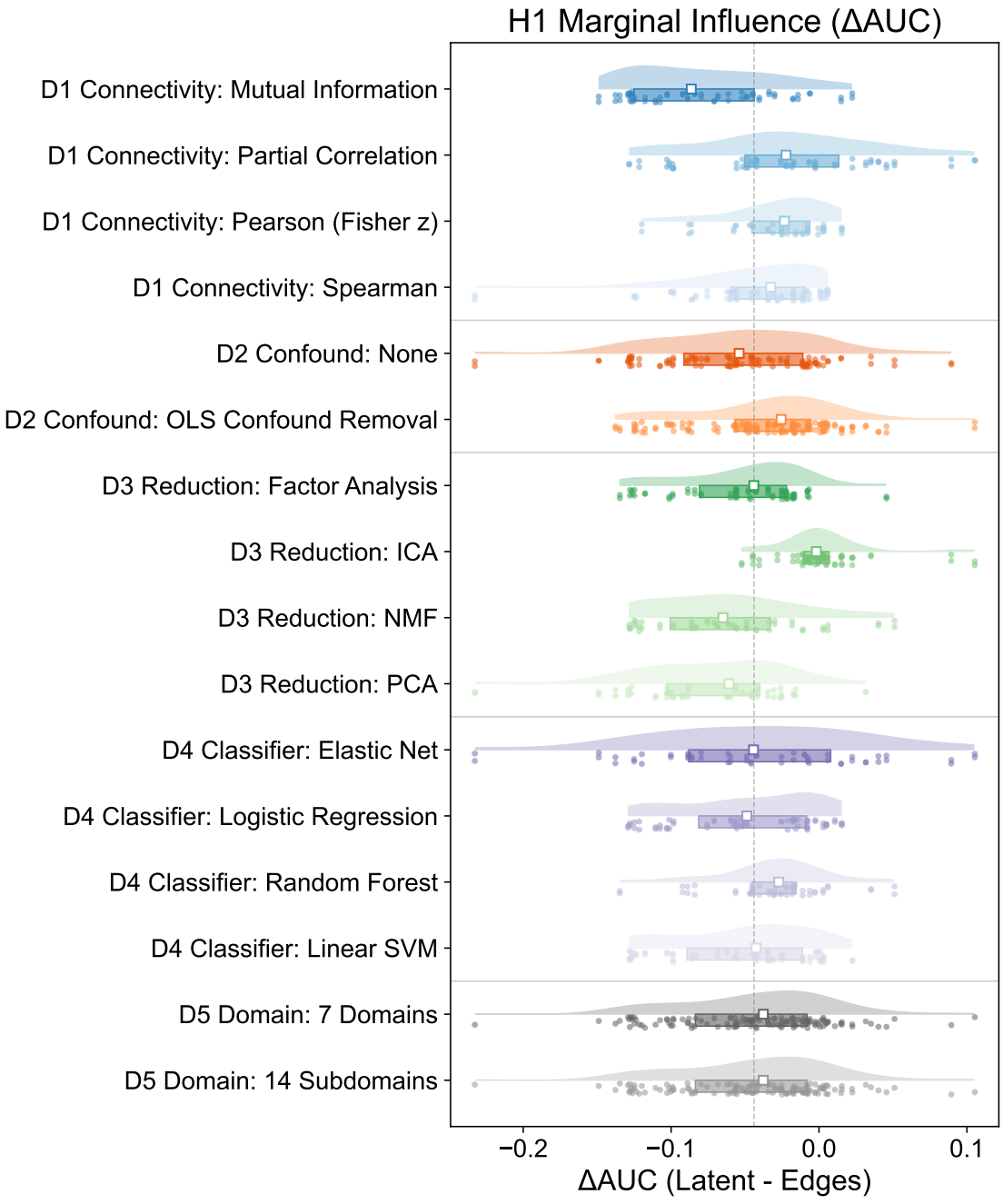

Figure 6: H1 raincloud plot. Marginal influence of fork levels on ΔAUC across the factorial mix. ICA and OLS visibly shift mass toward latent-favourable outcomes versus FA/NMF/PCA and `none`.

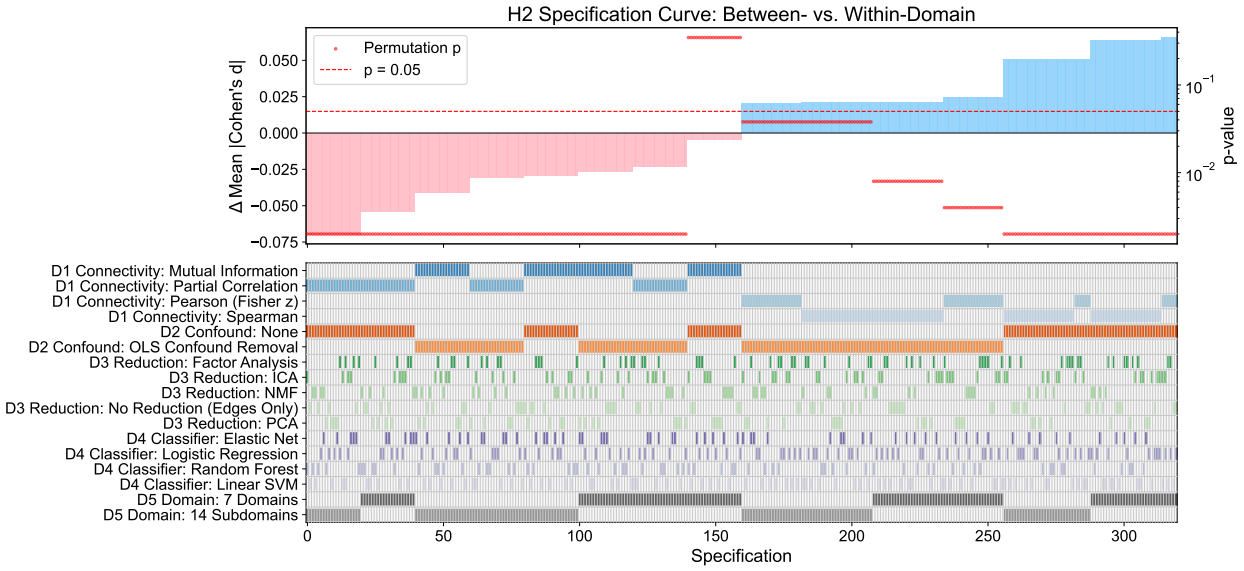

Figure 7: H2 specification curve. One column per specification ($N = 320$), sorted by the standardized H2 summary statistic (between-domain enrichment of mean $|d|$). Tiles encode forks; flat plateaus at favourable values indicate robustness across connectivity and confound choices.

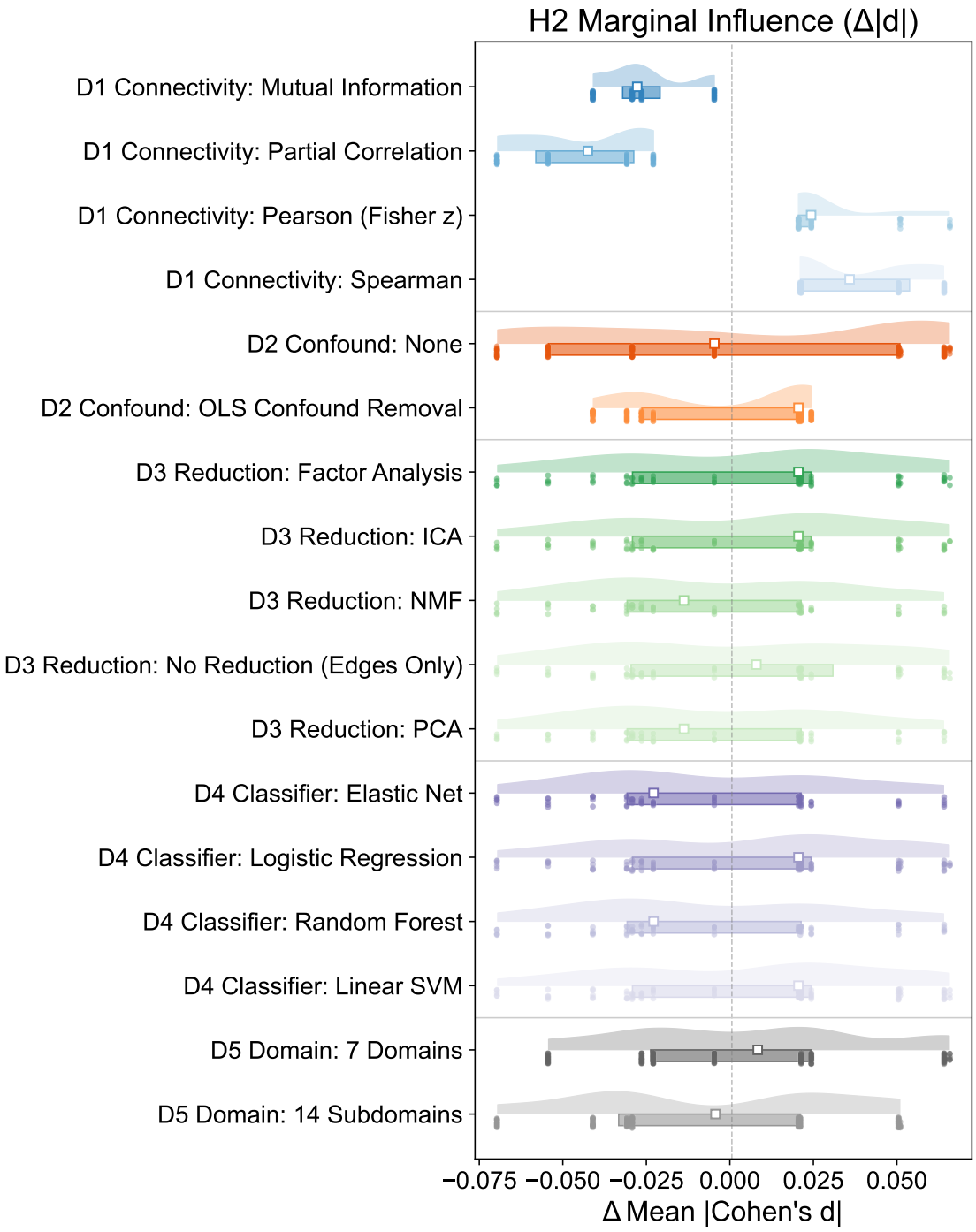

Figure 8: H2 raincloud plot. Marginal distributions of H2 outcomes by fork level, illustrating stability drivers (correlation-based connectivity, OLS) versus weaker slices.

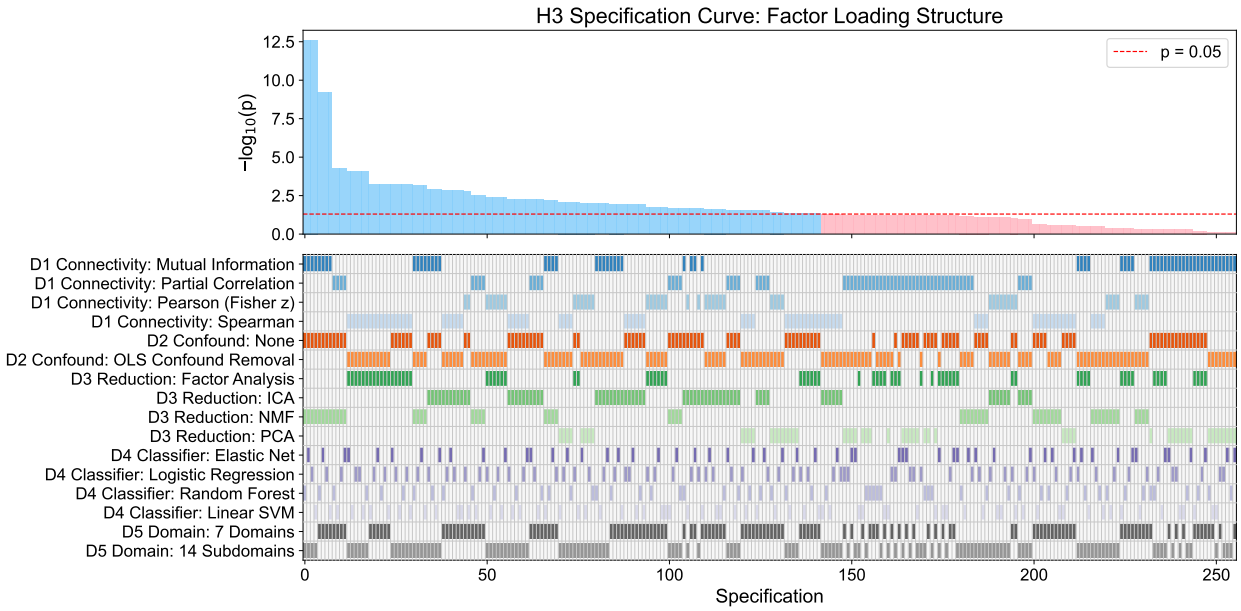

Figure 9: H3 specification curve. Latent-only specifications ($N = 256$), sorted by the plotted H3 outcome (Wilcoxon p-value). Significant rows are common but decomposition-dependent; ICA-rich columns dominate favourable extremes.

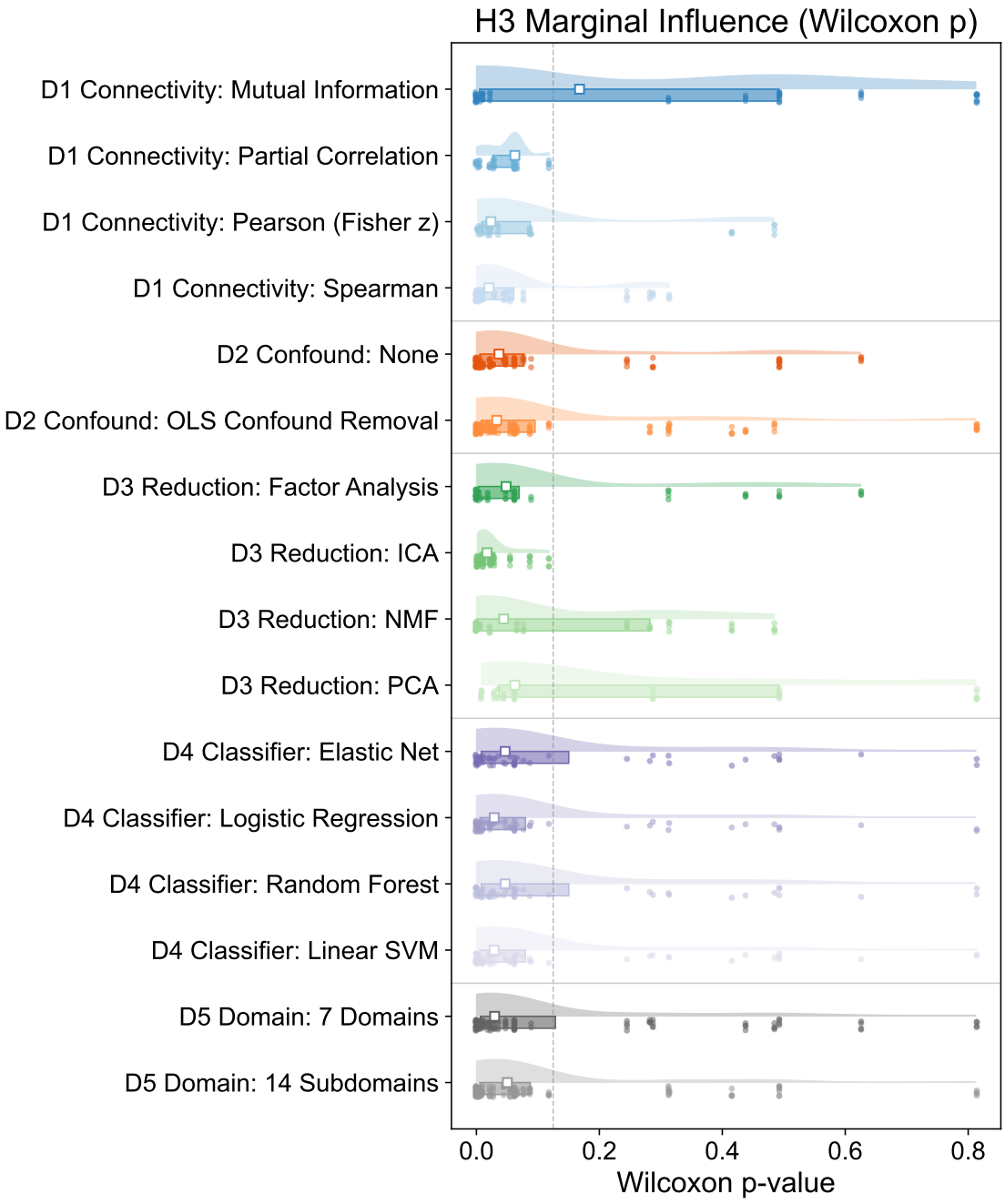

Figure 10: H3 raincloud plot. Robustness of loading-mass contrasts across forks. ICA + correlation-based connectivity lifts favourable fractions relative to PCA or MI paths.

Chapter 5

Discussion

5.1 H1: Edge-level prediction remains the conservative default; latent models are exploratory and hinge on ICA, nuisance regression, and partial correlations.

Dense edges remain a strong predictive default in this sample once confounds are modeled; latent models function better as hypothesis generators tied to ICA/partial-correlation/OLS combinations than as uniformly superior biomarkers.

That aggregate edge advantage need not dismiss latent factor analysis (including ICA): compressing thousands of pairwise couplings into a small number of modes trades a modest amount of discriminative accuracy (here, median ΔAUC slightly favouring edges) for **interpretability**—each component's loading profile can be read as a structured coupling pattern across ICNs rather than a diffuse weight vector over all edges—and for **stability** in the sense of a lower-dimensional representation that emphasises shared variance and can be less sensitive to idiosyncratic edge noise when nuisance structure is handled (consistent with ICA-rich slices showing the largest latent-favourable fractions in this multiverse). Prediction and explanation therefore answer different questions; ICA-aligned latent arms remain valuable for mechanistic follow-up and cross-study alignment even when raw edges win on ROC-AUC.

Nested CV with inner tuning inflates optimism relative to a fully external sample; comparing two tuned arms on shared folds induces correlated AUC noise, so small ΔAUC values demand confirmation with dedicated uncertainty quantification. The significant joint binomial test for H1 underscores that global count tests and median effects tell different stories.

5.2 H2: Domain enrichment supports long-range dysconnectivity narratives statistically but does not localize circuits or prove causal mechanisms.

The enrichment pattern is compatible with long-range dysintegration hypotheses while remaining a massive edge-wise summary: it does not localize circuits. Medication, symptom severity, and illness duration, which are not modeled here, could still covary with FC.

5.3 H3: ICA-linked loading structure motivates mechanistic hypotheses while pooled estimation limits biomarker claims.

H3 bridges H2's group-level enrichment and latent geometry; ICA alignment suggests physically interpretable extended modes may better capture between-domain loading structure than orthogonal PCA axes. Because multiverse H3 fits reducers on pooled scaled edges and treats factor summaries as exchangeable pairs for Wilcoxon tests, significant results support mechanistic follow-up rather than standalone clinical validation.

5.4 Limitations

Single cohort and single parcellation. Results may not generalise beyond the FBIRN dataset and the NeuroMark 2.2 template.

Unmeasured confounding. Age, sex, race, site, and motion enter via D2 OLS, but due to the complexity of psychopathology, potential confounders such as medication history and symptom heterogeneity should be considered.

Dependence and multiplicity. Although connectivity edges are estimated separately for each subject, the resulting edge-wise statistical tests would be dependent because they are computed from the same participants and overlapping ICN time courses. Joint binomial checks are stylized sensitivity analyses, not substitutes for preregistered primary pipelines [10].

Clinical translation. Robust statistical structure does not establish causality, symptom mechanisms, or treatment targets.

5.5 Future directions

Replication and preregistration. The highest priority is held-out or multi-site validation with one or two pre-registered primary pipelines chosen *a priori* (connectivity + confound + reduction), reporting uncertainty for ΔAUC (H1) and raw effect sizes (H2/H3), rather than relying solely on within-sample multiverse counts.

Expanded analytic forks. Natural extensions include ComBat or other harmonisation strategies [8], alternative ICN atlases or ICA-derived networks, dynamic or multi-band FC, and severity-, medication-, or cognitive-stratified subsets to test whether H2/H3 patterns track clinical heterogeneity.

Inferential refinements. Dependence across edges and correlated specifications motivate structured nulls, empirical Bayes shrinkage across the curve, or simulation-based calibration rather than only stylised binomial Joint-count summaries [4, 10].

H1/H3 methodology. Future work could nest latent decomposition fully inside outer CV for H3-equivalent questions, add bootstrap confidence intervals for fork-wise robustness percentages, and benchmark latent pipelines against regularised edge models matched on degrees of freedom.

Mechanistic translation. Combining FC multiverses with multimodal data, perturbational paradigms, or genetics could narrow which between-domain enrichments are reproducible *beyond* statistical screens—still short of clinical biomarkers without prospective designs.

Chapter 6

Brain Researcher MCP review

This chapter records **Brain Researcher** MCP outputs from `run_find_latest_reviewable`, `run_scientific_review`, and `run_code_review` (invoked 2026-05-04).

6.1 Scientific review

Overall decision: `proceed`. **Correctness:** `pass` (no deterministic findings). **Completeness:** `complete` — random seed pinned; atlas version pinned; ordering rule declared.

6.2 Code review

Decision: `approve`. **Risk level:** low. **Findings:** none.

6.3 Reviewer questions

- **Modality** / **space:** Resting-state fMRI; group-level ICN time courses and FNC edges (105 ICNs; see Methods).
- **Scientific question** / **estimand:** Multiverse robustness of H1 (nested-CV ΔAUC), H2 (between- vs. within-domain $|d|$), H3 (loading mass); see Hypotheses and Results.
- **Preprocessing** / **confounds:** Connectivity estimators D1; OLS residualization fork D2; standardization within folds for latent arms (Methods).
- **Inference** / **multiple comparisons:** Per-specification tests; joint binomial sensitivity (Results); limitations on dependence (Discussion).
- **Null** / **permutation:** H2 domain-label permutation; H1/H3 as described in code and Limitations.
- **Predictive split** / **leakage:** Nested CV for H1 with fold-wise confound fit; H3 pooled fit caveat (Discussion).
- **Key artifacts:** `multiverse_results.csv`, `mv_*.csv`, specification-curve figures.
- **Software:** Python; NumPy/SciPy; scikit-learn (see Methods).

Chapter 7

Conclusion

The **320**-specification multiverse yields a clear ordering of claim strength: **H2** shows highly robust between-domain enrichment in $|d|$ with small magnitude; **H3** shows partial, ICA-sensitive loading-structure evidence; **H1** shows median edge superiority with minority latent-favourable pipelines concentrated under ICA, OLS, and partial correlation. Readers should combine medians, percentages, specification curves, fork-wise CSVs, and joint binomial summaries rather than any single scalar. External replication with preregistered primary models remains the appropriate next step [4, 8].

Bibliography

[1] Micha Burkhardt and Carsten Gießing. The comet toolbox: Improving robustness in network neuroscience through multiverse analysis. *Imaging Neuroscience*, 4:IMAG.a.1122, 02 2026. ISSN 2837-6056. doi: 10.1162/IMAG.a.1122. URL https://doi.org/10.1162/IMAG.a.1122.

[2] Vince D. Calhoun and Jing Sui. Multimodal fusion of brain imaging data: A key to finding the missing link(s) in complex mental illness. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(3):230–244, 2016. ISSN 2451-9022. doi: https://doi.org/10.1016/j.bpsc.2015.12.005. URL https://www.sciencedirect.com/science/article/pii/S2451902216000598. Brain Connectivity in Psychopathology.

[3] Alex Fornito, Andrew Zalesky, Christos Pantelis, and Edward T. Bullmore. Schizophrenia, neuroimaging and connectomics. *NeuroImage*, 62(4):2296–2314, 2012. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2011.12.090. URL https://www.sciencedirect.com/science/article/pii/S1053811912002133. Connectivity.

[4] Marco Del Giudice and Steven W. Gangestad. A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920954925, 2021. doi: 10.1177/2515245920954925. URL https://doi.org/10.1177/2515245920954925.

[5] A. Iraji, Z. Fu, A. Faghiri, M. Duda, J. Chen, S. Rachakonda, T. DeRamus, P. Kochunov, B. M. Adhikari, A. Belger, J. M. Ford, D. H. Mathalon, G. D. Pearlson, S. G. Potkin, A. Preda, J. A. Turner, T. G. M. van Erp, J. R. Bustillo, K. Yang, K. Ishizuka, A. Faria, A. Sawa, K. Hutchison, E. A. Osuch, J. Theberge, C. Abbott, B. A. Mueller, D. Zhi, C. Zhuo, S. Liu, Y. Xu, M. Salman, J. Liu, Y. Du, J. Sui, T. Adali, and V. D. Calhoun. Identifying canonical and replicable multi-scale intrinsic connectivity networks in 100k+ resting-state fmri datasets. *Human Brain Mapping*, 44(17):5729–5748, 2023. doi: https://doi.org/10.1002/hbm.26472. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.26472.

[6] Kyle M. Jensen, Jessica A. Turner, Lucina Q. Uddin, Vince D. Calhoun, and Armin Iraji. Addressing inconsistency in functional neuroimaging: A replicable data-driven multi-scale functional atlas for canonical brain networks. *bioRxiv*, 2024. doi: 10.1101/2024.09.09.612129. URL https://www.biorxiv.org/content/early/2024/12/03/2024.09.09.612129.

[7] David B. Keator, Theo G.M. van Erp, Jessica A. Turner, Gary H. Glover, Bryon A. Mueller, Thomas T. Liu, James T. Voyvodic, Jerod Rasmussen, Vince D. Calhoun, Hyo Jong Lee, Arthur W. Toga, Sarah McEwen, Judith M. Ford, Daniel H. Mathalon, Michele Diaz, Daniel S. O'Leary, H. Jeremy Bockholt, Syam Gadde, Adrian Preda, Cynthia G. Wible, Hal S. Stern, Aysenil Belger, Gregory McCarthy, Burak Ozyurt, and Steven G. Potkin. The function

biomedical informatics research network data repository. *NeuroImage*, 124:1074–1079, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2015.09.003. URL https://www.sciencedirect.com/science/article/pii/S1053811915007995.

[8] Daniel Kristanto, Micha Burkhardt, Christiane Thiel, Stefan Debener, Carsten Gießing, and Andrea Hildebrandt. The multiverse of data preprocessing and analysis in graph-based fmri: A systematic literature review of analytical choices fed into a decision support tool for informed analysis. *Neuroscience & Biobehavioral Reviews*, 165:105846, 2024. ISSN 0149-7634. doi: https://doi.org/10.1016/j.neubiorev.2024.105846. URL https://www.sciencedirect.com/science/article/pii/S0149763424003154.

[9] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. ISSN 0047-259X. doi: https://doi.org/10.1016/S0047-259X(03)00096-4. URL https://www.sciencedirect.com/science/article/pii/S0047259X03000964.

[10] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 2020. doi: https://doi.org/10.1038/s41562-020-0912-z. URL https://doi.org/10.1038/s41562-020-0912-z.

[11] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016. doi: 10.1177/1745691616658637. URL https://doi.org/10.1177/1745691616658637. PMID: 27694465.

[12] Peter J. Uhlhaas and Wolf Singer. Abnormal neural oscillations and synchrony in schizophrenia. *Nature Reviews Neuroscience*, 11(2):100–113, 2010. doi: https://doi.org/10.1038/nrn2774. URL https://doi.org/10.1038/nrn2774.