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: 2026-06-26

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 **480** pipelines ($4 \times 3 \times 5 \times 4 \times 2$), including ComBat harmonization in the confound fork; each evaluates three hypotheses [4, 8, 10, 11]. **H2**—greater mean absolute Cohen's d on between-domain than within-domain edges—is estimator-split: 12/24 unique connectivity–confound–domain contrasts are favourable (50.0%), with Pearson and Spearman favourable and partial correlation / mutual information unfavourable. **H3**—greater between-domain loading mass in latent factors—has limited direction-aware support (100/384, 26.0% favourable; median Wilcoxon $p = 0.0543$), with factor analysis and PCA more favourable than ICA/NMF under the revised majority-direction criterion. **H1**—nested-CV ROC-AUC gain for latent versus raw edges—is fragile at the median (median ΔAUC < 0; 18.8% favourable) yet strongly fork-dependent (especially ICA, ComBat, and partial correlation). Joint binomial count tests reject a stylized independent-α null for all three hypotheses but must be read alongside medians, direction-aware definitions, 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 . . . 8

4 Results **12**
- 4.1 H1: Raw edges achieve higher median nested-CV ROC-AUC than latent factors; latent gains concentrate under ICA, ComBat, and partial-correlation connectivity. . . . 12
- 4.2 H2: Between-domain effect-size enrichment is estimator-split rather than globally robust. . . . 12
- 4.3 H3: Direction-aware loading-mass evidence is limited overall and strongest for FA/PCA plus correlation-based connectivity. . . . 13

5 Discussion **17**
- 5.1 H1: Edge-level prediction remains the conservative default; latent models are exploratory and hinge on ICA, nuisance regression, and partial correlations. . . . 17
- 5.2 H2: Domain enrichment supports long-range dysconnectivity narratives only under correlation-based FNC. . . . 17
- 5.3 H3: Direction-aware loading structure motivates follow-up but is not robust enough for confirmatory claims. . . . 18
- 5.4 Limitations . . . 18
- 5.5 Future directions . . . 18

6 Brain Researcher MCP review **20**
6.1 Scientific review 20
6.2 Code review 20
6.3 Reviewer questions 20

7 Conclusion **22**

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 ($N = 384$), 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, ComBat, partial correlation, etc. 14

6 **H1 raincloud plot.** Marginal influence of fork levels on ΔAUC across the factorial mix. ICA and ComBat visibly shift mass toward latent-favourable outcomes versus FA/NMF/PCA and `none`. 14

7 **H2 specification curve.** One column per unique H2 contrast ($N = 24$ after collapsing classifier/reduction duplicates), sorted by the standardized H2 summary statistic (between-domain enrichment of mean $|d|$). Tiles encode forks; the favourable and unfavourable halves track connectivity estimator. 15

8 **H2 raincloud plot.** Marginal distributions of H2 outcomes by fork level, illustrating the separation between correlation-based connectivity and partial-correlation / mutual-information paths. 15

9 **H3 specification curve.** Latent-only specifications ($N = 384$), sorted by the plotted H3 outcome (Wilcoxon p-value). Favourable rows require both nominal significance and a strict majority of components with between-domain loading mass greater than within-domain loading mass. 16

10 **H3 raincloud plot.** Robustness of loading-mass contrasts across forks, using the number of components with between-domain loading mass greater than within-domain loading mass. FA/PCA and correlation-based connectivity show the strongest direction-aware support. 16

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. 9

3 **Default factorial grid** (480 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 using the one-sided H2 permutation p for favourable calls; median Wilcoxon p for H3). These results support a cautious reading: H1 is fragile, H2 is estimator-split, and H3 has limited direction-aware support. 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 and direction-aware support. 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 480 specifications. Each specification evaluates H1–H3 where applicable.

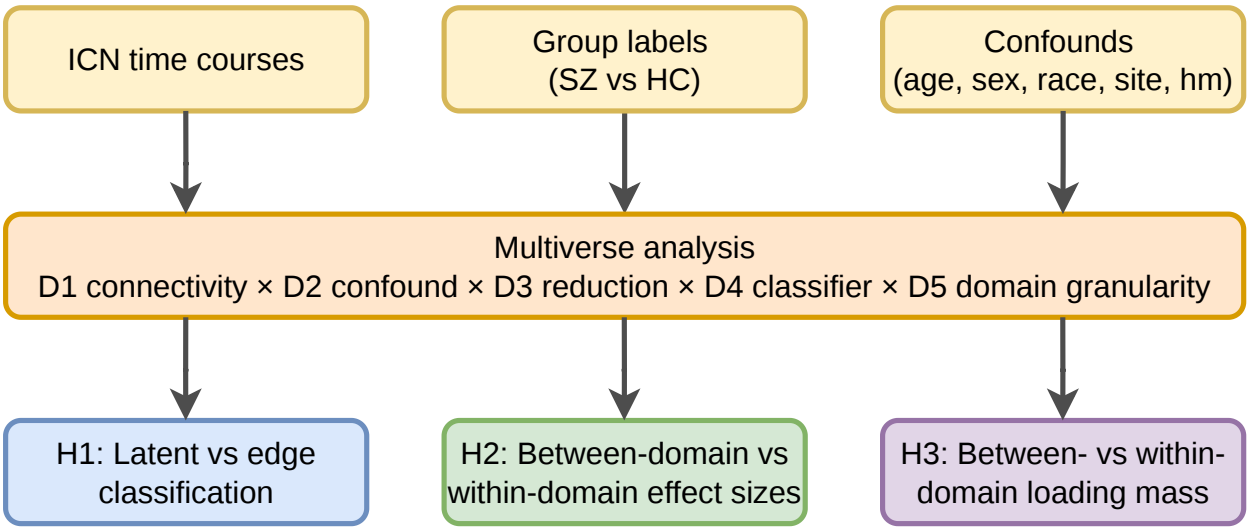

Figure 1: Study design schematic. Inputs (ICN time courses, labels, confounds), crossed multiverse forks (D1–D5), and parallel H1–H3 branches. The workflow ties high-dimensional edges to three complementary questions—latent versus edge-based classification (H1), between-domain versus within-domain effect 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 (384 evaluable latent specifications).

Interpretive lens. Latent superiority is not the default: most specifications favour edges at the median, yet ICA, ComBat harmonization, 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. The H2 helper records both one-sided and two-sided permutation values, but the multiverse code writes `p_value_one_sided` to `h2_p`; therefore the report tables and figures use the one-sided H2

alternative, $\mathrm{mean}(|d|_{\mathrm{between}}) > \mathrm{mean}(|d|_{\mathrm{within}})$. Favourable if one-sided $p < 0.05$ and $\Delta\,\mathrm{mean}|d| > 0$. H2 is independent of classifier and reduction, so the robustness summary reports 24 unique connectivity–confound–domain contrasts, even though the per-pipeline table contains duplicated H2 values across all 480 rows.

Interpretive lens. H2 is estimator-conditional: Pearson and Spearman support the expected between>within ordering, whereas partial correlation and mutual information do not. Even favourable H2 slices support consistent ordering more than large effect gaps, and they do 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 $|\mathrm{loading}|$ is summarized between- vs. within-domain per component; paired differences across components are tested with a Wilcoxon signed-rank test. A specification is favourable only if $p < 0.05$ *and* a strict majority of fitted components have between-domain loading mass greater than within-domain loading mass. This direction-aware rule is a post-review correction of the earlier direction-blind significance count. H3 requires specifications with dimension reduction (384 specifications).

Interpretive lens. Direction-aware support is limited overall. Factor analysis and PCA are more often favourable than ICA/NMF under the majority-component rule, while correlation-based connectivity remains more supportive than mutual information. 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%	—

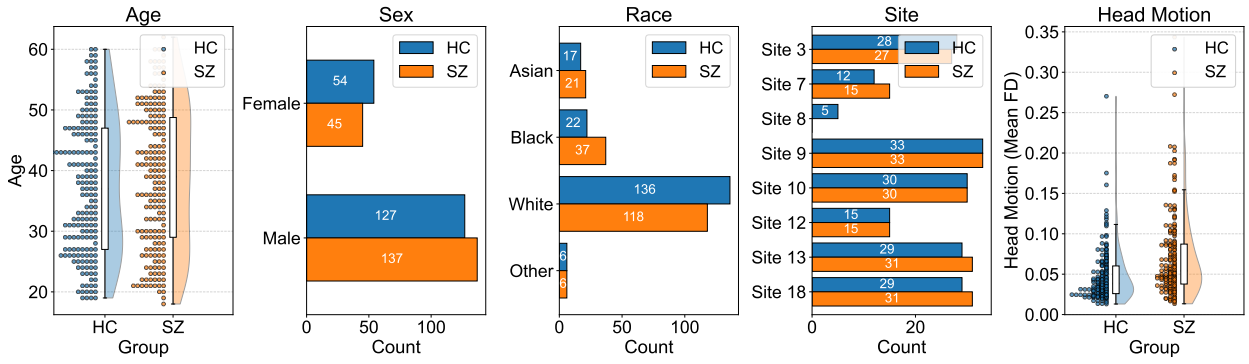

Figure 2: Distributions of confounding variables / covariates. Several covariates differ across groups or sites; ignoring D2 can blend diagnosis signal with shared nuisance structure.

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.

`combat`. ComBat harmonization is included as a third D2 level. This fork targets site-related distributional shifts while retaining diagnosis contrasts and should be interpreted as a harmonization sensitivity analysis rather than a guarantee that all site-by-diagnosis structure has been removed.

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).

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 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

Table 3: Default factorial grid (480 specifications).

Fork	Levels
D1 Connectivity	`pearson_z`, `spearman`, `partial_corr`, `mutual_info`
D2 Confound	`none`, `ols`, `combat`
D3 Reduction	`none`, `fa`, `ica`, `pca`, `nmf`
D4 Classifier	`elasticnet`, `logistic_l2`, `svm_linear`, `rf`
D5 Domain	`domain_7`, `subdomain_14`
Count	$4 \times 3 \times 5 \times 4 \times 2 = \mathbf{480}$

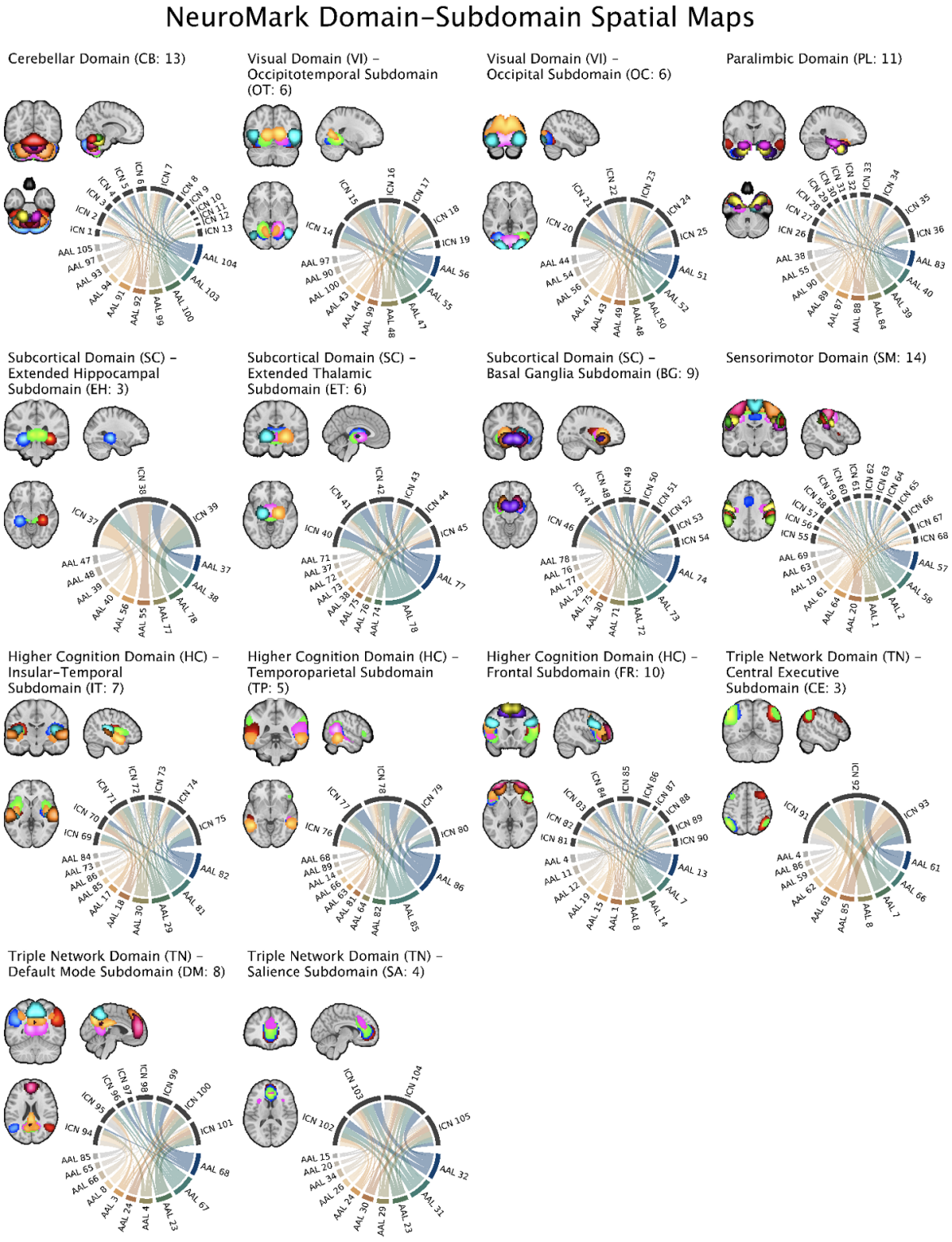

Figure 3: Spatial maps of the 105 ICNs from the NeuroMark 2.2 template. Reprinted from "Addressing inconsistency in functional neuroimaging: A replicable data-driven multi-scale functional atlas for canonical brain networks" by Kyle M. Jensen, Jessica A. Turner, Lucina Q. Uddin, Vince D. Calhoun, and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

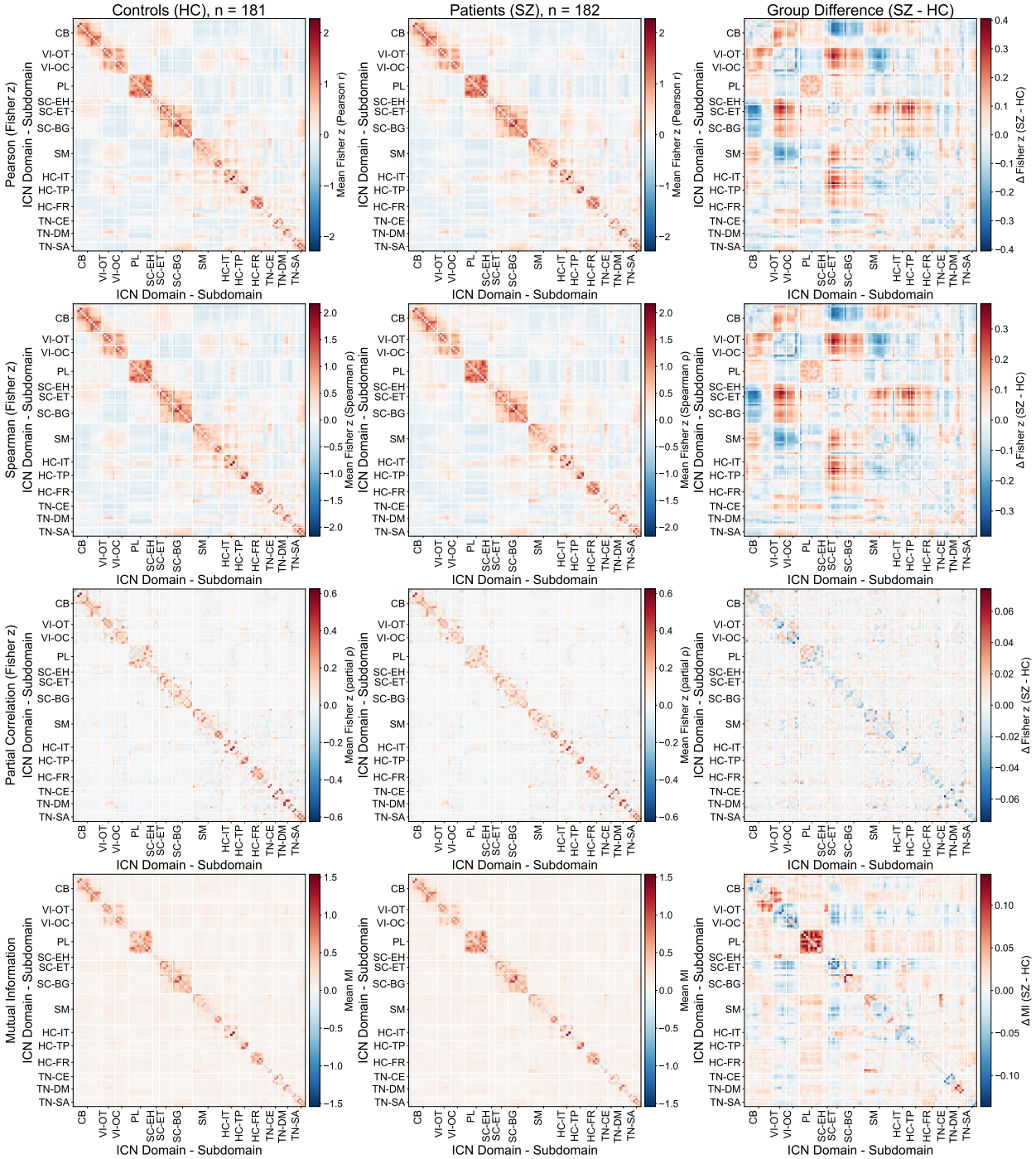

Figure 4: Group-mean FNC for HC, SZ, and group difference from different connectivity measures (Pearson Fisher-z, Spearman, partial correlation, and mutual information).

Chapter 4

Results

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

Across 384 latent-evaluable specifications, 72 (18.8%) favour the latent arm; median ΔAUC $= -0.0319$, so edges win in aggregate. Conditional splits (`mv_conditional_robustness.csv`) reinforce path dependence: ICA (54.2% favourable within ICA rows) vs. FA (8.3%), NMF (10.4%), and PCA (2.1%); ComBat (26.6%) vs. OLS (18.8%) and `none` (10.9%); partial correlation (29.2%) vs. mutual information (8.3%). Classifier shifts are modest by comparison, and domain granularity is neutral (18.8% for both 7-domain and 14-subdomain masks).

The joint binomial test rejects H_0: "independent $\alpha = 0.05$ favourable draws" ($p \approx 4.8 \times 10^{-22}$) because 72 positives still exceeds the expected 19.2 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 using the one-sided H2 permutation p for favourable calls; median Wilcoxon p for H3). These results support a cautious reading: H1 is fragile, H2 is estimator-split, and H3 has limited direction-aware support.

Hypothesis	N	Favourable	%	Median effect
H1: Latent $>$ edges	384	72	18.8	-0.0319
H2: Between $>$ within	24	12	50.0	0.0079
H3: Loading advantage	384	100	26.0	0.0543 median p

4.2 H2: Between-domain effect-size enrichment is estimator-split rather than globally robust.

After deduplicating over classifier and reduction, 12/24 (50.0%) unique H2 contrasts are favourable with median $\Delta\,\mathrm{mean}|d| = 0.0079$. The split is almost entirely driven by connectivity estimator: Pearson (Fisher z) and Spearman are 100% favourable across confound/domain cells, whereas partial correlation and mutual information are 0% favourable. Confound strategy (ComBat, `none`, OLS)

and domain granularity each sit at 50%, indicating they do not rescue the estimator dependence in this 480-specification analysis.

The small positive median comes from averaging across a bimodal pattern: correlation-based FNC produces positive between-domain enrichment, while partial-correlation and mutual-information FNC produce negative or null enrichment. H2 should therefore be read as estimator-conditional support for long-range dysconnectivity, not a claim that the ordering survives all reasonable connectivity definitions.

4.3 H3: Direction-aware loading-mass evidence is limited overall and strongest for FA/PCA plus correlation-based connectivity.

100/384 (26.0%) latent specifications are favourable under the corrected direction-aware rule; median Wilcoxon $p \approx 0.0543$ (Table 4), just above the nominal 0.05 threshold. Factor analysis (45.8% within-slice favourable) and PCA (33.3%) exceed ICA (12.5%) and NMF (12.5%) when favourability requires both significance and a majority of components with between-domain loading mass greater than within-domain loading mass. Spearman (41.7%) and Pearson (37.5%) outperform partial correlation (25.0%) and mutual information (0.0%). Seven domains (25.0%) and fourteen subdomains (27.1%) are similar, so finer labels do not materially strengthen H3 in this analysis.

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	384	72	18.8	19.2	$\approx 4.8 \times 10^{-22}$
H2	24	12	50.0	1.2	$\approx 3.7 \times 10^{-10}$
H3	384	100	26.0	19.2	$\approx 8.8 \times 10^{-43}$

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 and direction-aware support.

H	Fork level	Total	# Fav.	%
H1	ICA reduction	96	52	54.2
H1	PCA reduction	96	2	2.1
H1	ComBat confound	128	34	26.6
H1	Partial correlation	96	28	29.2
H2	Pearson (Fisher z)	6	6	100.0
H2	Mutual information	6	0	0.0
H3	FA reduction	96	44	45.8
H3	PCA reduction	96	32	33.3
H3	Mutual information	96	0	0.0

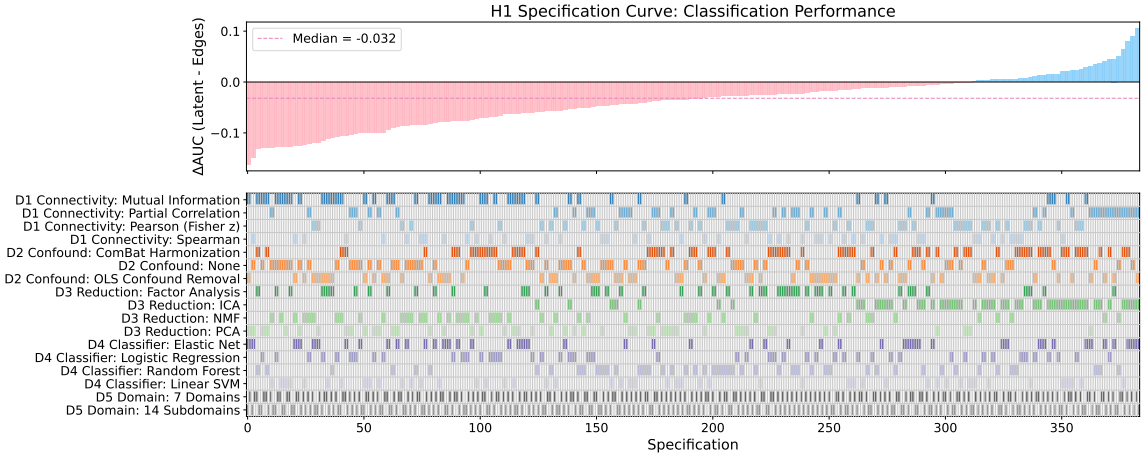

Figure 5: H1 specification curve. One column per latent-evaluable pipeline ($N = 384$), sorted by ΔAUC (latent minus edges). The main trace shows the curve; coloured tiles beneath encode D1–D5 fork choices showing that favourable tails align with ICA, ComBat, partial correlation, etc.

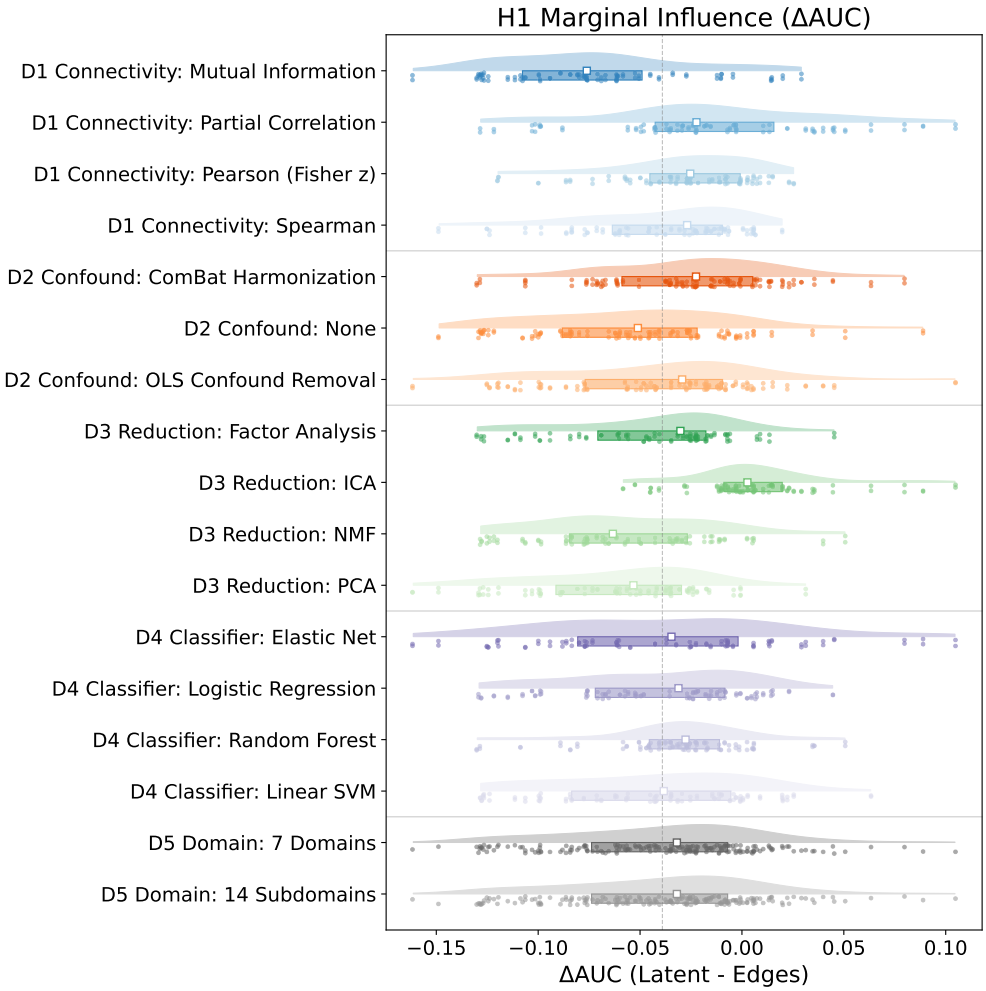

Figure 6: H1 raincloud plot. Marginal influence of fork levels on ΔAUC across the factorial mix. ICA and ComBat visibly shift mass toward latent-favourable outcomes versus FA/NMF/PCA and `none`.

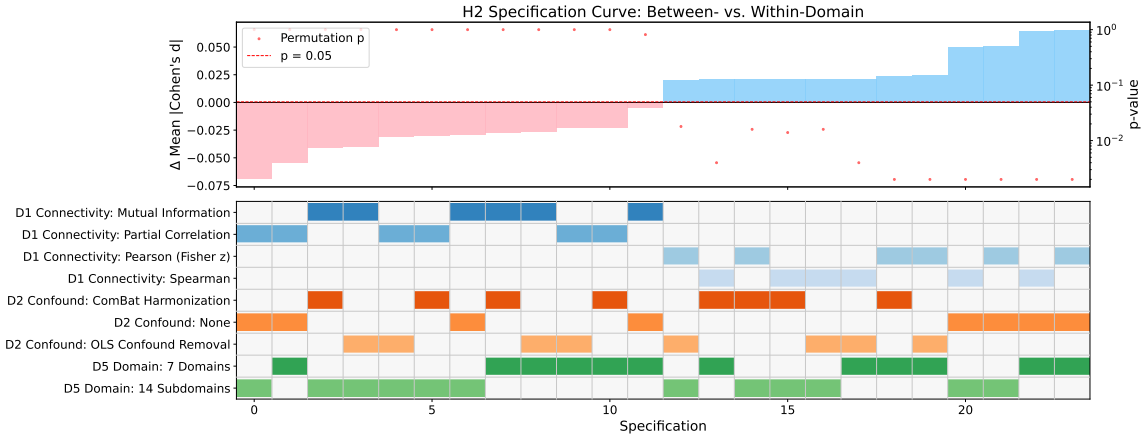

Figure 7: H2 specification curve. One column per unique H2 contrast ($N = 24$ after collapsing classifier/reduction duplicates), sorted by the standardized H2 summary statistic (between-domain enrichment of mean $|d|$). Tiles encode forks; the favourable and unfavourable halves track connectivity estimator.

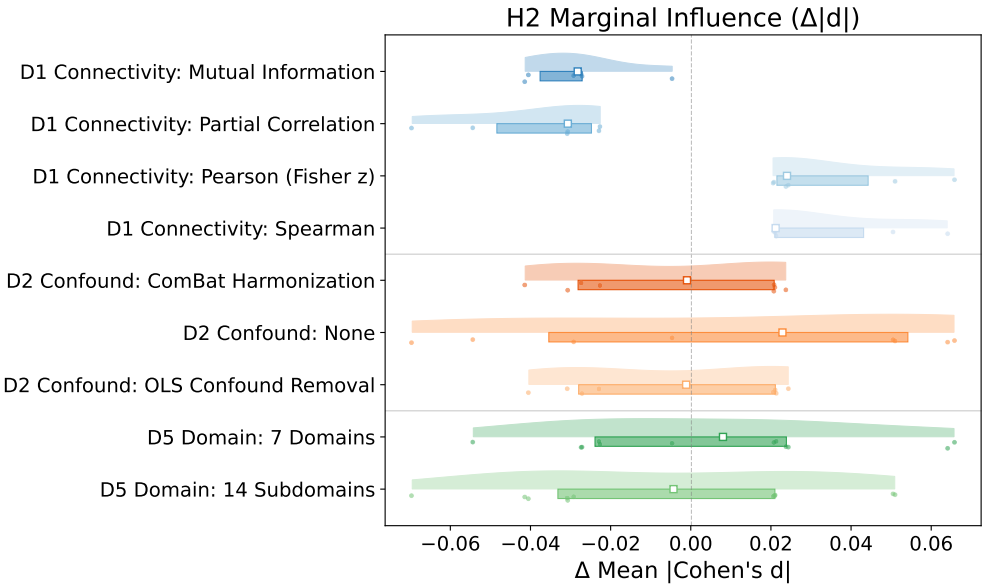

Figure 8: H2 raincloud plot. Marginal distributions of H2 outcomes by fork level, illustrating the separation between correlation-based connectivity and partial-correlation / mutual-information paths.

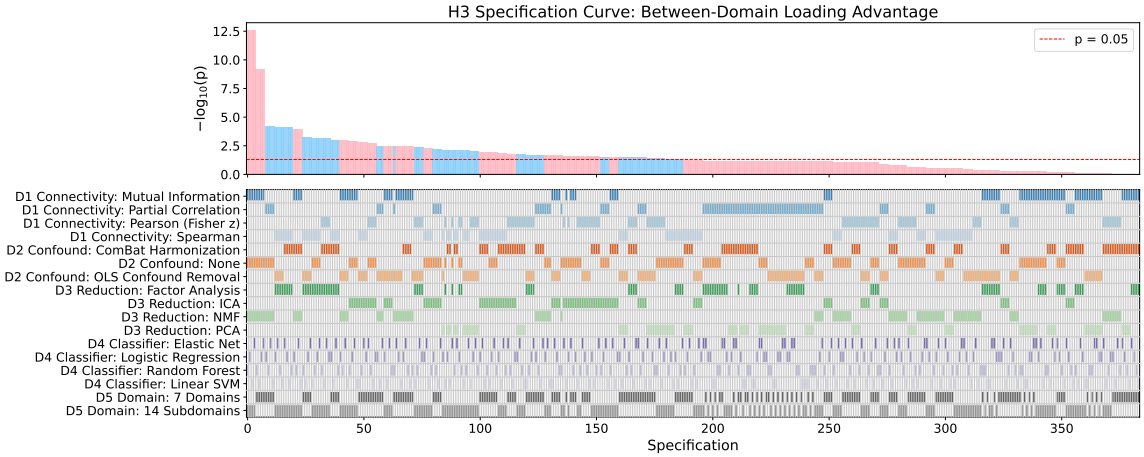

Figure 9: H3 specification curve. Latent-only specifications ($N = 384$), sorted by the plotted H3 outcome (Wilcoxon p-value). Favourable rows require both nominal significance and a strict majority of components with between-domain loading mass greater than within-domain loading mass.

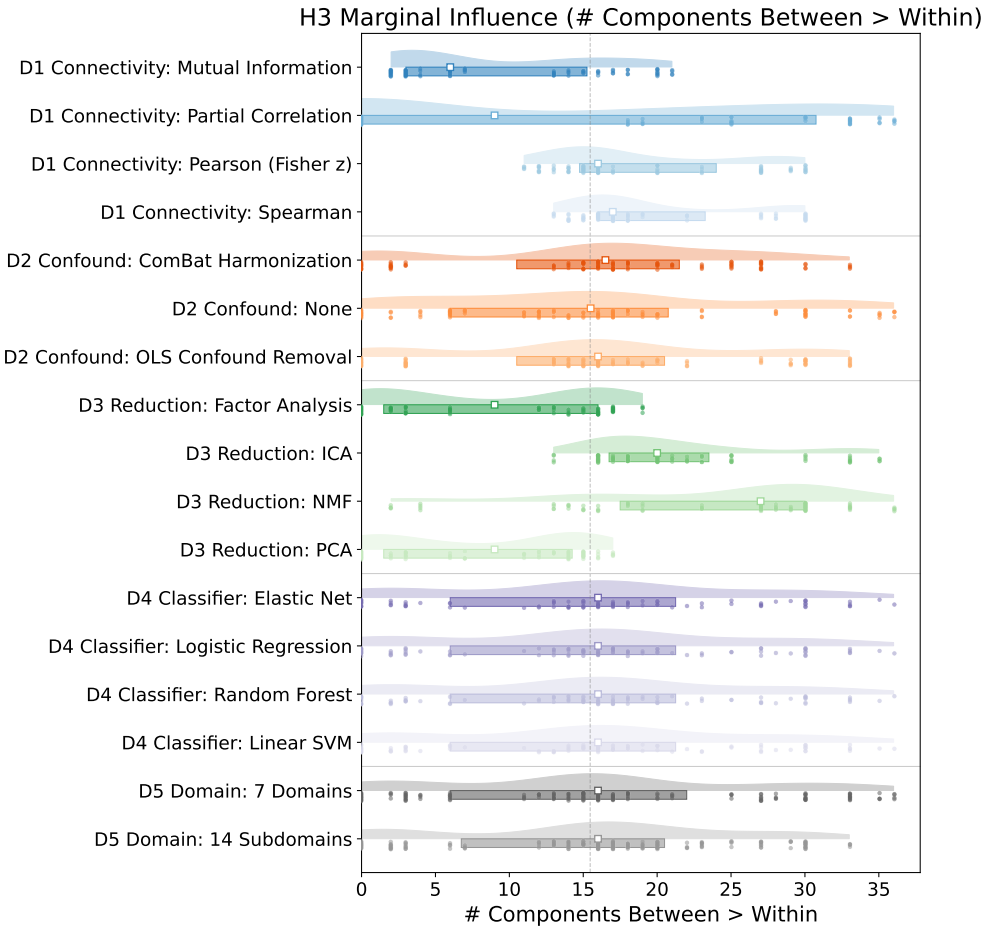

Figure 10: H3 raincloud plot. Robustness of loading-mass contrasts across forks, using the number of components with between-domain loading mass greater than within-domain loading mass. FA/PCA and correlation-based connectivity show the strongest direction-aware support.

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 only under correlation-based FNC.

The Pearson and Spearman enrichment pattern is compatible with long-range dysintegration hypotheses, but the partial-correlation and mutual-information slices argue against treating H2 as estimator-invariant. Even where favourable, H2 remains 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: Direction-aware loading structure motivates follow-up but is not robust enough for confirmatory claims.

H3 bridges H2's group-level enrichment and latent geometry, but the corrected direction-aware summary lowers the favourable fraction to 26.0%. The median Wilcoxon p is slightly above 0.05, and support is concentrated in FA/PCA plus correlation-based connectivity rather than across all latent methods. Because multiverse H3 fits reducers on pooled scaled edges and treats factor summaries as exchangeable pairs for Wilcoxon tests, favourable slices 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].

Revised H3 decision rule. H3 favourability now requires both a significant Wilcoxon test and a majority of components in the hypothesized between$>$within direction. This corrects an earlier direction-blind count of significant Wilcoxon results and should be treated as a revised exploratory decision rule unless prospectively preregistered.

Harmonization sensitivity. Including ComBat expands D2 and improves coverage of site-harmonization choices. ComBat itself is not a substitute for external replication, and site-by-diagnosis structure can remain difficult to separate in a case-control cohort.

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 additional 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, predeclare whether H3 uses majority-component count, signed median/mean loading difference, or one-sided Wilcoxon as its primary estimand, 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 an independent **Brain Researcher** MCP scientific-review and code-review pass invoked on 2026-06-26 for the report source, the `fbirn_experiment` code, and the 480-specification artifacts in `results/multiverse_full/`.

6.1 Scientific review

Overall decision: `explore_more`. **Correctness:** `flag` (non-blocking findings). **Completeness:** `incomplete`. **Report action:** `revise_report`. The review verified that the report's primary multiverse summaries are internally consistent with the current result table: H1 72/384, H2 12/24, and H3 100/384 favourable specifications.

Remaining findings. The reviewed `results/multiverse_full/` directory lacks a `run_manifest.json`; joint binomial specification-count tests are dependence-naive and should be treated as descriptive; ComBat H2/H3 branches require site-by-diagnosis and residual site-predictability diagnostics; and the H3 majority-direction rule should be treated as an exploratory decision rule unless preregistered.

6.2 Code review

Decision: non-blocking revision. Splitting CSV regeneration from plotting, or making `matplotlib` a hard CLI dependency with a clear error, would improve reproducibility.

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 (direction-aware loading mass); see Hypotheses and Results.
- **Preprocessing** / **confounds:** Connectivity estimators D1; D2 confound fork (`none`, OLS residualization, ComBat harmonization); 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 with one-sided `h2_p` in the multiverse summaries; 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; result provenance should be recorded in `run_manifest.json`.
- **Software:** Python; NumPy/SciPy; scikit-learn (see Methods).

Chapter 7

Conclusion

The **480**-specification multiverse yields a cautious ordering of evidence: **H1** shows median edge superiority with minority latent-favourable pipelines concentrated under ICA, ComBat, and partial correlation; **H2** shows between-domain enrichment only for correlation-based FNC estimators, not partial correlation or mutual information; and **H3** shows limited direction-aware loading-structure support (26.0% favourable) concentrated under FA/PCA and correlation-based connectivity. 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.