A Fixed Coherence-Based Pipeline Repeatedly Exceeds a Matched Reference Within HCP-YA

A Brain Researcher case study

Brain Researcher

Date: 2026-08-13

Contents

Abstract 1

1. Testing connectivity prediction against a matched reference 2

2. Searching for a coherence-based prediction pipeline 4

3. Cognition: fixed pipeline versus matched reference 8

4. The fixed pipeline across four additional behavioural dimensions 10

5. Results that limit the main comparison 13

6. How Brain Researcher made the analysis auditable 15

7. Limits and the next test 16

8. Conclusion 18

Methods and provenance 19

References 20

Abstract

We used Brain Researcher (BR) to pursue a bounded analysis of resting-state functional connectivity in the Human Connectome Project Young Adult cohort (HCP-YA). The analysis began from a published benchmark that maps functional connectivity to five behavioural dimensions and provides a concrete reference procedure. BR evaluated 116 candidate analyses across an initial and an expanded search; 104 completed successfully. The search first established a plateau among precision-connectivity models, then broadened across increasingly different representation families. The strongest observed candidate, reached late in that sweep, pointed to whole-band coherence with ridge regression. After fixed follow-ups, rerunning the 12 failed proposals, and a separate development-only analysis of procedure selection, a related coherence–kernel-ridge pipeline was included in a human-authorised contract and fixed before comparison with a locally matched version of the published procedure on ten repeated, family-grouped partitions of the same 244-person development cohort.

For Cognition, the fixed pipeline had higher pooled out-of-fold correlation, higher R^2, and lower error in all ten repeats. Median correlation was $.332$ for the fixed pipeline and $.235$ for the matched reference, with median paired $\Delta r = .098$. A conditional paired sensitivity gave a nominal one-sided $p = .006$ and a pointwise 95% interval of $[.011, .177]$. The same pipeline was then applied without target-specific retuning to Tobacco Use, Personality and Emotion, Illicit Drug Use, and Mental Health. Median paired Δr remained positive for all four outcomes, with favourable directions in 10/10, 9/10, 10/10, and 8/10 repeats, respectively. Absolute predictive utility was clearest for Cognition and Tobacco Use and was near zero or negative for the other outcomes.

The result is a repeat-stable, same-cohort benchmark difference, not an externally validated discovery. The candidate was selected using the same cohort; repeated partitions overlap; target construction was not wholly fold-local; a separate predeclared calibration-repair decision aid failed its gate; and no cross-outcome comparison survived simultaneous-interval or weak-FWER control. BR's contribution was to preserve the search denominator, fix the evaluated pipeline, separate each evidence tier, and retain the failed checks. It did not autonomously establish scientific confirmation.

What this report establishes

A fixed coherence-based connectivity pipeline repeatedly outperformed a locally matched reference procedure for Cognition within one HCP-YA development cohort, and its directional advantage extended descriptively to four additional behavioural dimensions. This report does *not* establish paper-exact superiority, search-adjusted confirmation, multiplicity-confirmed transfer, calibrated utility across outcomes, independent replication, or external generalisation.

1. Testing connectivity prediction against a matched reference

HCP-YA combines resting-state fMRI with a broad behavioural battery [1]. Liu and colleagues used these data to benchmark methods for mapping functional connectivity to five summary dimensions: Cognition, Tobacco Use, Personality and Emotion, Illicit Drug Use, and Mental Health. Their published analysis and released resources gave this case study a useful starting point: a defined dataset, interpretable targets, and a reference analysis that could be rebuilt under matched local conditions [2].

The scientific question was deliberately narrower than "find the best brain predictor." We asked whether a candidate discovered during bounded model search would still look better after it was fixed and compared with a matched reference procedure on the same people, folds, outcomes, covariate policy, and metrics. We then asked whether the *same fixed pipeline*, without tuning it again for each target, showed a similar directional pattern on the other four behavioural dimensions.

The reference in this report is a local, nested implementation modeled on the published procedure [2]. It searched 74 eligible connectivity representations, four model families, and four regularisation values inside each outer training fold. It is not a paper-exact reproduction of every element of the original benchmark, so the comparison should be read as fixed candidate versus matched local reference, not as a claim that one paper has been superseded.

1.1 What was reconstructed

The local target is a Cognition component reconstructed from the available HCP-YA behavioural resources. It shares the scientific role of the published Cognition dimension but is not guaranteed to reproduce every preprocessing, item-selection, or component-estimation decision in the original study [2]. The HCP-YA source pool used for this case contained 326 people. It was partitioned into a 245-person discovery and development portion from 244 families and an 81-person, 81-family internal holdout that remained sealed for a one-time analysis. Later matched analyses used 244 people from 243 families. These samples all come from HCP-YA; family separation inside that cohort is not external validation.

The candidate space combined multiple connectivity representations with classical, connectome-predictive, tree, graph, and neural prediction models; the connectome-predictive family followed the standard CPM framework [3]. The search agent did not receive subject-level connectivity, targets, or fold predictions. It saw a sanitized catalogue and aggregate scores, then proposed bounded representation–model tests. This design made it possible to inspect what the agent chose next

without exposing participant-level data, but it did not remove adaptive selection from the scientific interpretation.

1.2 The planned sequence from search to evaluation

The research programme was not "search until a large score appears." It was planned as a sequence of questions with different evidence ceilings:

1. Can a bounded search agent find promising representation–model directions from aggregate feedback?
2. Do fixed controls, alternative partitions, shuffles, and reruns of failed proposals show that the result is executable and not a trivial harness artifact?
3. Does a development-only selector audit nominate a stable procedure, or does it expose instability that requires a researcher-owned choice?
4. Once one implementation is fixed before evaluation, does it repeatedly exceed a locally matched published-style procedure on the same cohort?
5. Does that fixed implementation retain its direction on other outcomes, and does relative improvement coexist with useful absolute prediction?
6. Which result, if any, deserves an external-cohort test?

The later sections retain the intermediate and negative answers because they determine what the final positive comparison can mean.

The report in three sentences

First, bounded search identified a coherence-based prediction direction worth auditing and carrying into a related fixed implementation. Second, once fixed, that pipeline exceeded a matched local reference in every repeated Cognition comparison. Third, the same direction appeared across four additional outcomes, but the multiplicity analysis and absolute R^2 values prevent upgrading that pattern to general predictive superiority.

2. Searching for a coherence-based prediction pipeline

The search used a development cohort of 245 people from 244 families and a catalogue of connectivity representations and prediction models. BR presented the search agent with aggregate results rather than subject-level data. Each proposal had to name a representation–model pair, a rationale, and a falsifier. The denominator contains 20 proposals in the initial search and 96 in the expanded search. Each search also produced four prespecified diagnostic analyses, which are reported separately and are not counted among the 116 candidates.

Recorded search-agent interface

The initial and expanded searches used the same fixed Brain Researcher proposal interface. Each request required strict JSON with exactly two proposal decisions. Every decision named a prediction model and connectivity representation, together with a hypothesis, falsifier, and rationale. The initial search used ten requests and the expanded search used 48. The persisted fields expose the proposed scientific move, but they do not reconstruct hidden chain-of-thought.

Table 1: Representative persisted search directions. These rows show how the search moved; they are not independent confirmatory tests.

Phase	Proposal	Question	Declared falsifier	Why it came next
Coverage	Ledoit–Wolf covariance with ridge versus Spearman connectivity with CPM	Compare a regularized linear candidate with a connectome-predictive candidate.	The expected ordering does not recur in the next aggregate ledger.	Establish model coverage before local exploitation.
Local precision	Shrinkage and empirical precision connectivity with CPM	Probe neighboring precision-connectivity estimators.	Neither neighboring estimator preserves the local signal.	Test a bounded family-level hypothesis rather than declare a general precision advantage.
Threshold test	Unconstrained versus Sakoe–Chiba soft-DTW with ridge	Ask whether related temporal-alignment representations reach $r = .3815$.	Both remain below the predeclared threshold.	Move beyond the current representation family using a declared stopping rule.
Spectral follow-up	Whole-band coherence versus low-band phase locking, followed by low- and high-band coherence	Test whether the new spectral lead is specific to connectivity type or frequency range.	Whole-band coherence fails to exceed $r = .45437$, or the band decomposition does not preserve the pattern.	Distinguish a specific coherence result from a generic spectral advantage.

The initial search covered 13 prediction models across three representation families. After broad model coverage, it concentrated around covariance and precision connectivity. Its strongest

candidate used a covariance representation with support-vector regression: mean-fold $r = .3726$, but $R^2 = -.0359$. This was a useful warning that a high correlation alone did not yet establish useful prediction.

The expanded search restarted with broad coverage rather than simply replaying that winner. It tested 15 classifier keys, 60 representation terms, and 19 metric families. Early proposals found a cross-model signal and then a precision-connectivity plateau: precision with CPM reached $r = .3858$, and precision with ridge reached $r = .4544$. The search agent then kept the leading regularised linear model class approximately fixed while traversing cross-correlation, distance and dependence, temporal-alignment, information-theoretic, and finally spectral representations. Whole-band mean multitaper coherence was first tested as the 109th candidate across the two searches. It raised the observed maximum to $r = .4871$ and $R^2 = .2118$.

That late lead did not automatically become the final pipeline. Fixed follow-ups supported phase locking and low-band mean coherence but rejected the declared thresholds for high-band coherence and maximum-aggregation variants. Four prespecified checks also included an alternate partition, a different-stratum runner-up, a family-block shuffle, and a synthetic positive control. Twelve proposals in the expanded search failed before returning a result. They were rerun without reopening the completed proposals; this established computational equivalence for the recovered configuration, not replication and not completion of the original search record.

Post-search work then separated candidate discovery from pipeline freezing. A same-cohort selector audit examined fixed alternatives and proposed four directions rather than declaring an automatic champion. Only after that audit did a researcher authorise a comparison of seven fixed pipelines that included the final coherence–kernel-ridge pipeline. Thus the adaptive ridge result identified a representation–model direction; the exact cosine-kernel implementation used below was fixed later and was neither the only predeclared arm nor the prespecified primary endpoint.

2.1 From adaptive search to fixed evaluation

Table 2 restores the intermediate stages that are easy to lose when only the final comparison is plotted. The stages are ordered from adaptive observations to fixed procedures; a later result cannot retroactively validate an earlier search.

2.2 Diagnostic checks and model-selection stability

The fixed diagnostic screen completed all ten authorised analyses without adaptation or substitution. The synthetic positive control passed ($r = .9655$). Two single family-block shuffles favored their observed counterparts, but one shuffle per comparison is only a directional diagnostic, not a permutation null or a p-value. Alternate partitions reused the same cohort and therefore measured split sensitivity rather than replication.

The procedure-selection audit then replayed the historical 100-workflow selection procedure on the 244-person development set with five outer and three inner family-grouped folds. The broad selector chose partial least squares with shrinkage-precision connectivity overall, whereas each outer fold

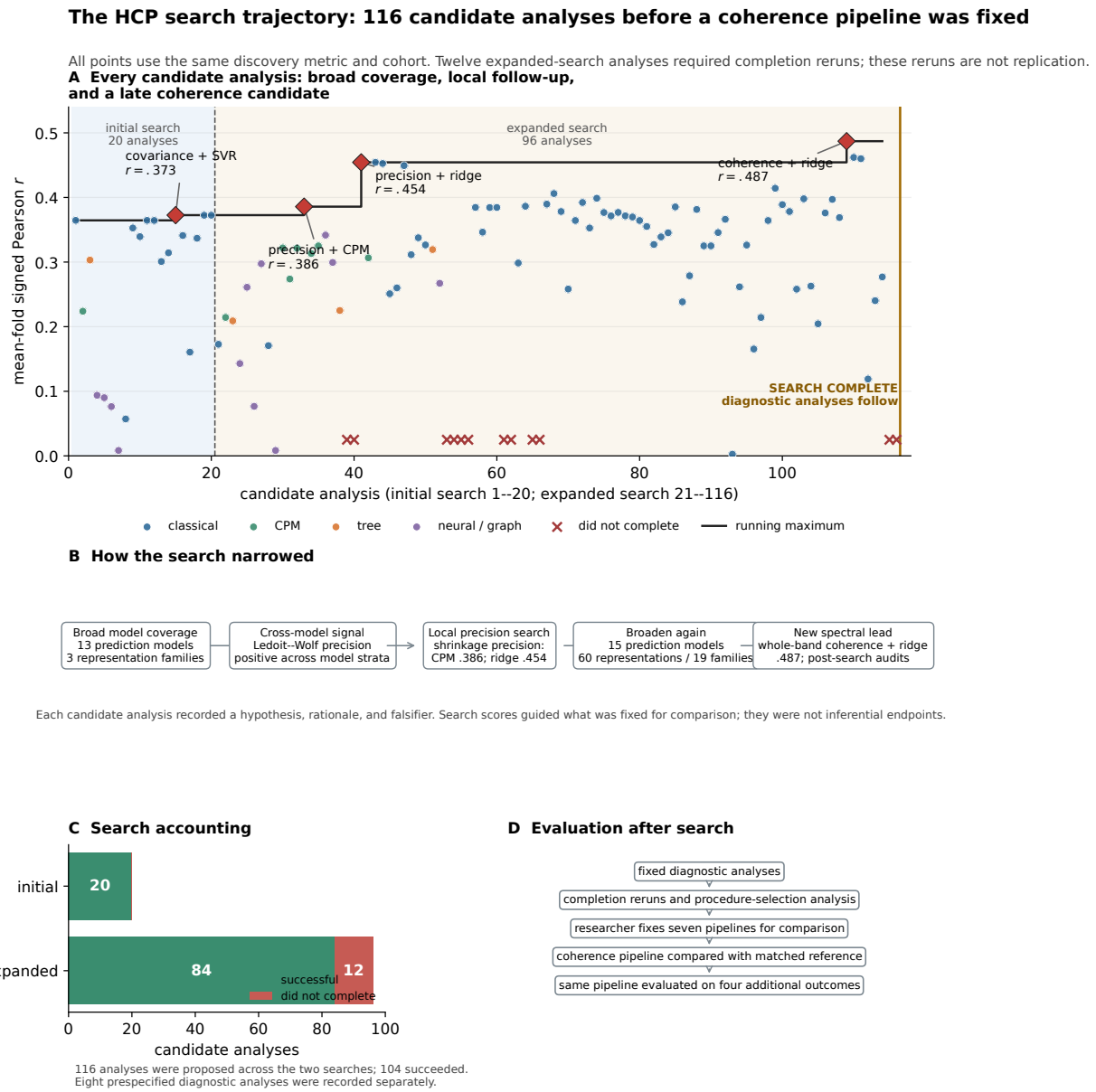

Figure 1: Search trajectory. Panel A shows all 116 candidate proposals rather than only the eventual winner; the running maximum exposes the early precision plateau and the late spectral lead. Panel B summarizes how the search agent narrowed locally and then broadened again. Panel C separates search accounting from later checks, completion reruns, the selector audit, and the human-authorised comparison. The two searches used unequal budgets and the same discovery cohort. Completion reruns are not replication.

Table 2: Execution scale, recorded outcome, and scientific interpretation at each stage.

Stage	Scale or interface	Recorded outcome	Scientific interpretation
Initial search	20 candidate proposals plus four diagnostic records	SVR with Ledoit–Wolf covariance: $r = .372637$, $R^2 = -.035933$, MAE $= .790670$	Valid descriptive adaptive trajectory.
Expanded search	96 candidate proposals plus four diagnostic records	84 successes, 12 execution failures; ordinary ridge with whole-band coherence: $r = .487109$, $R^2 = .211800$	Selected maximum from a search record containing execution failures.
Completion reruns	12 proposals that had not returned results	12/12 completed; linear-kernel ridge with whole-band coherence reproduced the prediction class	Computational completion and estimator equivalence, not replication.
Fixed diagnostics	Ten prespecified checks	10/10 completed, including alternate split, runner-up, shuffle, and positive-control checks	Partition and harness diagnostics.
Procedure-selection audit	Five-outer by three-inner family-grouped evaluation of 100 prespecified workflows	Partial least squares with shrinkage-precision connectivity selected overall; five outer folds chose five different winners	Development-only model-selection stability evidence.
Repeated Cognition comparison	Ten repeats, seven fixed pipelines, one reference reselected within each outer training fold	400/400 outer evaluations; the fixed coherence pipeline exceeded the reference on r, R^2, and MAE in 10/10 repeats	Retrospective same-cohort matched comparison.
Paired sensitivity	One selected fixed-pipeline comparison; 243 family clusters; 9,999 swaps	Median $\Delta r = .098272$, nominal $p = .006$, pointwise interval $[.011438, .177251]$	Conditional, non-search-adjusted sensitivity.
One-time internal holdout	One 81-person HCP-YA holdout opening	Positive association for a different ordinary-ridge procedure; negative R^2	Narrow internal held-out association.
Cross-outcome comparison	Four targets, five fixed pipelines plus reference	1,200/1,200 outer evaluations	Retrospective same-cohort transfer stress test.
Multiplicity analysis	20 outcome-by-configuration cells	16 positive medians; two positive pointwise intervals; zero simultaneous or weak-FWER positives	Directional pattern without multiplicity-confirmed transfer.

chose a different winner. The fixed whole-band coherence representation gave the stronger development correlation but only near-zero absolute R^2. This instability is why the workflow did not let the selector automatically crown a champion.

Table 3: Fixed controls and development-only selector diagnostics.

Diagnostic	Recorded result	Boundary
Fixed partition checks	10/10 completed; first partition $r = .4045$, $R^2 = .0221$; second partition $r = .3412$	Same-cohort partition sensitivity; no champion or inferential shuffle claim.
Broad workflow selector	Partial least squares with shrinkage-precision connectivity; mean-fold $r = .2632$, $R^2 = -.1084$; pooled $r = .2109$, $R^2 = -.0736$	Development-only procedure audit.
Fixed whole-band coherence	Mean-fold $r = .3337$, $R^2 = -.0040$; pooled $r = .2822$, $R^2 = .0370$	Fixed diagnostic within the same audit.
Selection stability	Five outer folds selected five different template winners	Evidence of selector variability, not a stable winning rule.

3. Cognition: fixed pipeline versus matched reference

The selected scientific pipeline was fixed before the matched evaluation. We call it the *fixed coherence–ridge pipeline*. It uses whole-band mean multitaper coherence-magnitude connectivity [4] and cosine-kernel ridge prediction with $\alpha = 1$. The comparison used ten repeats of five outer and three inner family-grouped folds in a 244-person, 243-family development cohort. Feature filtering, scaling, and reference-model selection were performed within the training folds.

The fixed pipeline exceeded the matched reference in all ten repeats on all three reported metrics. Median pooled out-of-fold correlation was $.3317$ versus $.2352$; median R^2 was $.1073$ versus $.0086$; and median mean absolute error was $.7679$ versus $.8162$. The paired differences were favourable in 10/10 repeats for correlation, 10/10 for R^2, and 10/10 for error.

Table 4: Fixed Cognition comparison on ten repeated partitions of the same development cohort. Paired differences are medians of within-repeat differences.

Metric	Fixed pipeline	Matched reference	Paired difference	Direction
Pooled out-of-fold r	$.3317$	$.2352$	$+.0983$	10/10 higher
R^2	$.1073$	$.0086$	$+.0989$	10/10 higher
MAE	$.7679$	$.8162$	$-.0507$	10/10 lower

The headline contrast was one cell inside a wider comparison of seven fixed pipelines. Restoring the full table matters because the coherence pipeline was neither the only fixed candidate nor the study's primary calibration endpoint. The eighth procedure was the matched Liu-style selector, which searched 74 eligible representations, four model families, and four regularisation values inside each outer training fold.

Table 5: Median pooled results over ten repeated development partitions for all fixed pipelines and the matched selector.

Pipeline	r	R^2	MAE	Slope
Ledoit–Wolf precision, $\alpha = .1$	.298360	.087076	.789460	.907877
Robust precision, $\alpha = .1$	.307036	.093253	.778202	.988982
Robust precision, $\alpha = .1$, recalibrated	.274413	.074274	.784615	.941795
Robust precision, $\alpha = 1$	.310828	.084147	.774751	1.484662
Robust precision, $\alpha = 1$, recalibrated	.283976	.079864	.780131	.944814
OAS precision, $\alpha = .1$	.297578	.086501	.789464	.905381
Whole-band coherence, $\alpha = 1$	.331748	.107297	.767904	1.136761
Matched Liu-style selector	.235209	.008571	.816184	.525529

A separately authorised sensitivity analysis used only the persisted paired predictions. It did not refit either pipeline or reopen the internal holdout. Across 9,999 family-cluster label-swap draws, 59 were at least as extreme as the observed median $\Delta r = .0983$, yielding a nominal one-sided plus-one $p = .006$. A family-cluster bootstrap gave a pointwise 95% interval of $[.0114, .1773]$. The family-blocking principle follows the exchangeability constraints used in multi-level block permutation [5]; the exact statistic and draw count here are specific to the persisted analysis.

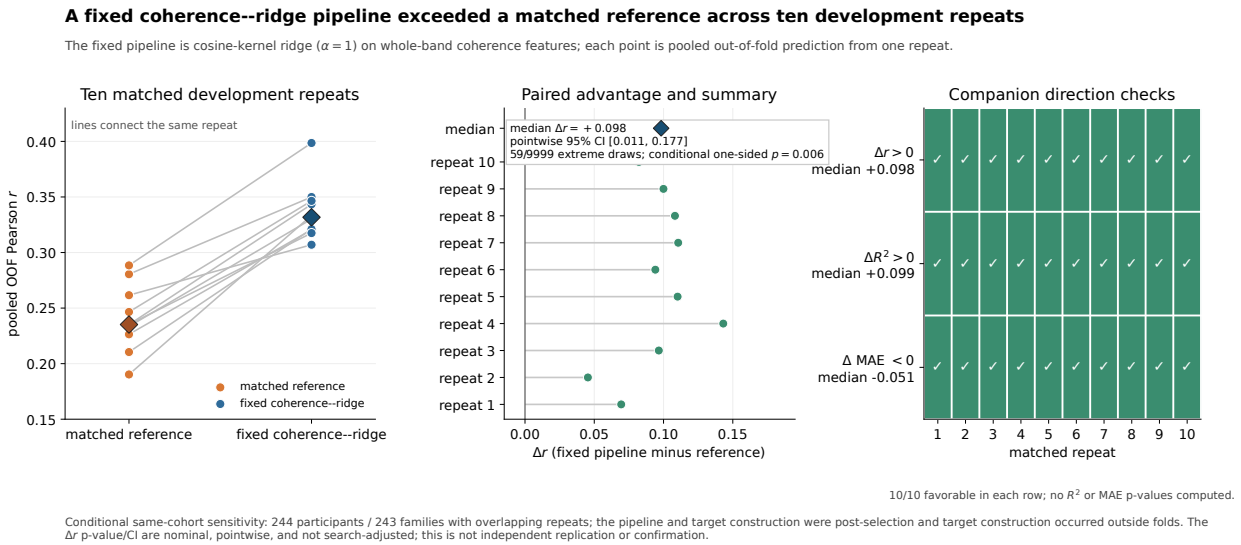

Figure 2: Fixed Cognition comparison. Lines connect the two procedures within the same repeated partition. The conditional p-value and pointwise interval quantify one historically selected comparison; they do not adjust for the preceding model and representation search. Repeats reuse the same participants and are stability checks rather than independent replications.

The clean numerical pattern does not remove the design limits. The selected representation, the Cognition target, and the decision to test this cell were informed by earlier work on the same cohort. The residualised target was constructed before fold assignment rather than rebuilt within each training fold. The algorithm-label test is therefore a conditional sensitivity, not an exact design-based or search-adjusted test. This distinction follows the known optimism of using cross-validation after model selection and the instability of small-sample cross-validation estimates [6, 7].

The shared label-swap vector was applied at the family-cluster level across all ten repeats in each Monte Carlo draw. Of 9,999 draws (seed 20260811), 59 were at least as extreme as the observed median. The family-cluster bootstrap used 9,999 resamples (seed 20260812). Because 242 of the 243 family clusters are singletons, family clustering differs from subject-level resampling for only one two-person cluster. These details make the conditional calculation reproducible; they do not make the historically selected comparison selection-adjusted.

4. The fixed pipeline across four additional behavioural dimensions

The next analysis asked whether the coherence–ridge pipeline was only useful for the Cognition target that nominated it. The pipeline was therefore held fixed and applied, without target-specific retuning, to the other four behavioural dimensions. Each outcome used the same ten-repeat evaluation and the same matched local reference.

The direction was positive for every outcome. Median Δr was $+.100$ for Tobacco Use, $+.068$ for Personality and Emotion, $+.117$ for Illicit Drug Use, and $+.069$ for Mental Health. The fixed pipeline had higher correlation in 10/10, 9/10, 10/10, and 8/10 repeats, respectively. This is the useful part of the transfer result: the originally selected pipeline continued to exceed the matched reference in most repeats when moved to new targets.

The absolute results are equally important. Median R^2 was positive for Cognition ($.107$) and Tobacco Use ($.057$), but negative for Personality and Emotion ($-.019$), Illicit Drug Use ($-.003$), and Mental Health ($-.047$). Thus a positive correlation difference does not imply useful calibrated prediction for every outcome.

Table 6: The same fixed pipeline applied to five behavioural dimensions. Only the Cognition comparison has the conditional $p = .006$ analysis.

Behavioural dimension	Median Δr	Favourable repeats	Absolute median R^2
Cognition	$+.0983$	10/10	$+.1073$
Tobacco Use	$+.1004$	10/10	$+.0568$
Personality and Emotion	$+.0683$	9/10	$-.0190$
Illicit Drug Use	$+.1168$	10/10	$-.0026$
Mental Health	$+.0693$	8/10	$-.0472$

The cross-outcome analysis also carried four other Cognition-nominated fixed pipelines into the same four outcomes. The broader matrix was not used to replace the coherence-pipeline headline with a target-by-target winner. Table 7 instead records the report-side configuration with the largest median Δr for each outcome so that the explored alternatives remain visible. This ordering is retrospective and not selection-adjusted.

The broader multiplicity analysis covered 20 outcome-by-configuration cells. Sixteen had positive median Δr, and two had positive pointwise intervals. None had a positive simultaneous interval, and none had a weak-FWER-adjusted $p < .05$. The transfer result is therefore descriptive and retrospective. It is not multiplicity-confirmed evidence that the pipeline is generally superior across behavioural outcomes.

Table 7: Exploratory view of the wider five-pipeline comparison. The configuration shown for each target is descriptive, not a predeclared winner.

Outcome	Reference r/R^2	Descriptive configuration	Median paired $\Delta r/\Delta R^2/\Delta$MAE	Absolute reading
Tobacco Use	$.1260/-.0292$	Coherence pipeline; 10/10 r wins	$+.100443/+.083372/-.022452$	$r = .2424$, $R^2 = .0568$; clearest additional prediction-level pattern.
Personality and Emotion	$.0059/-.0584$	Robust precision, $\alpha = 1$; 10/10 r wins	$+.081115/+.063346/-.021356$	$r = .1122$, $R^2 = .0102$; small positive absolute fit.
Illicit Drug Use	$.0077/-.0440$	Coherence pipeline; 10/10 r wins	$+.116822/+.041967/-.012804$	$r = .1220$, $R^2 = -.0026$; association-level pattern only.
Mental Health	$-.0620/-.0464$	Robust precision, $\alpha = 1$; 9/10 r wins	$+.084586/+.027304/-.006062$	$r = .0375$, $R^2 = -.0126$; absolute utility unsupported.

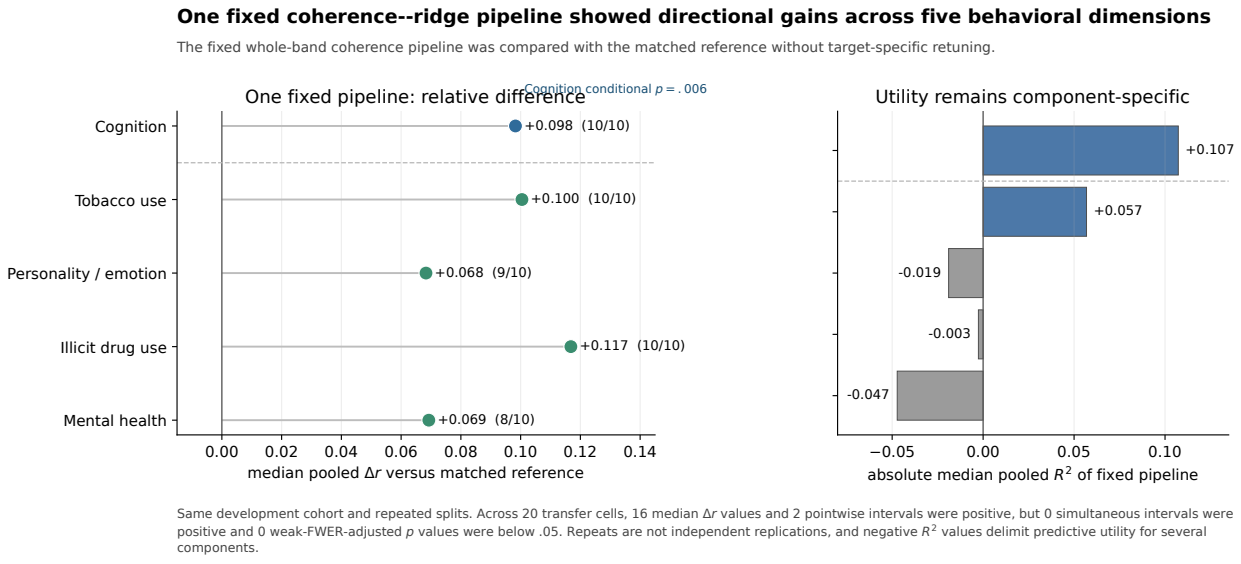

Figure 3: Cross-outcome stress test of one fixed pipeline. The left panel reports relative correlation differences and the number of favourable repeats. The right panel reports absolute R^2. The combination shows broad directional improvement but outcome-specific predictive utility.

The persisted matched-comparison, paired-inference, and transfer records remain under human review: neither scientific acceptance nor confirmation has been recorded. The numerical analyses are complete, but that status prevents the report from silently promoting them to confirmed scientific claims.

5. Results that limit the main comparison

5.1 A separate calibration-repair decision aid failed its own gate

A predeclared calibration-repair decision aid for the robust-precision pipeline with $\alpha = 1$ required improved R^2, non-inferior error, favourable R^2 in at least 8/10 repeats, and a slope between $.8$ and 1.2. Only the slope criterion passed; the joint gate failed three of four criteria. This check did not evaluate the headline coherence–ridge pipeline, so it does not negate the comparison in Section 3. It is retained as a parallel warning that a plausible calibration repair did not work and must not be folded into the headline result.

5.2 A separate internal holdout did not confirm the main comparison

A distinct analysis froze a different ridge version of the coherence representation before opening an 81-person, 81-family HCP-YA holdout once. It found a positive association ($r = .2320$, one-sided family-block $p = .0184$), but its R^2 was negative ($-.0745$) and its calibration slope was $.399$. It also contained no matched reference comparison and did not use the exact pipeline reported in Sections 3–4. This is a narrow internal held-out association for a related procedure, not an independent confirmation of the headline result.

The one-time holdout used ordinary ridge with whole-band coherence magnitude and $\alpha = 100$, rather than the cosine-kernel ridge configuration used in the main comparison. Its primary association interval was $[.0277, .4204]$; MAE was $.8691$. A secondary ordinary-ridge analysis of Kozachenko joint-entropy connectivity had $r = .2724$ and $R^2 = .0226$ but was descriptive. Prediction and target vectors were not persisted, so the one-shot opening cannot be reconstructed from this report alone.

Table 8: Endpoints from the one-time internal HCP-YA holdout.

Endpoint	r	Association detail	R^2 / MAE	Interpretation
Ordinary ridge on whole-band coherence, $\alpha = 100$	.232005	One-sided family-block $p = .0184$; interval $[.0277, .4204]$	$-.07445/.86911$	Association gate passed; calibration slope $.39881$ and predictive utility did not.
Ordinary ridge on Kozachenko joint entropy	.27238	Secondary descriptive endpoint	$.02260/.84106$	Positive absolute fit in this one descriptive endpoint; not a confirmatory result.

5.3 The goal-only baseline was a valid null, not a failed run

In a separately initiated goal-only baseline without direct BR use, a computational agent built a substantial HCP analysis from the supplied goal and inputs. It evaluated 239 usable connectivity estimators, deduplicated them to 201 representations, locked one cognition pipeline, and opened a

family-disjoint 81-person holdout once. The locked result did not meet its prespecified validation criterion (adjusted partial $r = .141$, permutation $p = .261$; overall $R^2 = -.013$). Because the holdout had been spent, the agent correctly stopped rather than test a runner-up on the same participants.

This baseline should not be described as stopping before analysis. It completed a careful single-shot experiment and ended with a valid internal null. The BR line followed a different history: it preserved a sequence of bounded follow-up analyses after review. The runs were not matched in prompt, budget, human interaction, or stopping rules, so their different outcomes do not estimate a causal effect of BR on scientific success.

6. How Brain Researcher made the analysis auditable

BR's role in this case was procedural and inspectable:

- It retained the full search denominator: 116 candidate analyses, 104 successful evaluations, and 12 execution failures.
- It kept the expanded search with execution failures distinct from the later completion reruns and from fixed evaluations.
- It fixed the coherence–ridge pipeline before the matched comparison and recorded which later analyses reused that pipeline without retuning.
- It preserved the failed calibration-repair decision aid, negative absolute R^2 values, multiplicity results, and the distinction between same-cohort stability and external replication.
- It left scientific authority with the researcher. Search output could propose a candidate; it could not promote the candidate to confirmation.

The contrast is not auditable versus unauditable, nor an agent that runs versus one that stops. The comparison analysis produced an auditable single-shot analysis. BR additionally recorded the progression from exploratory proposals to fixed evaluations across a sequence of bounded follow-up analyses, including failed gates and fixed pipeline reuse.

7. Limits and the next test

Why the conclusion remains bounded

- The discovery maximum came from adaptive, unequal-budget search on the same cohort used to nominate the later fixed comparison.
- The matched reference is locally reconstructed, not paper-exact.
- Repeated partitions overlap and do not constitute ten replications.
- Target and covariate construction was precomputed rather than wholly fold-local.
- The conditional $p = .006$ and pointwise interval do not adjust for candidate selection.
- A separate predeclared calibration-repair decision aid failed three of four criteria; that gate did not evaluate the headline coherence–ridge pipeline.
- Cross-outcome directional gains did not survive simultaneous-interval or weak-FWER control, and several outcomes had negative absolute R^2.
- The 81-person holdout used a different pipeline and remained inside HCP-YA.

The decisive next experiment is not another replay of the same cohort. It is a predeclared external-cohort evaluation of one fixed pipeline and one fixed target-construction procedure, with target construction performed inside the training data, family overlap excluded or blocked, and both association and absolute predictive utility specified in advance. Until that experiment is run, the highest defensible state is *repeat-stable same-cohort candidate*.

7.1 Candidate external cohorts

The previous research record considered several concrete replication routes rather than leaving "external validation" abstract. HCP-Aging and HCP-Development retain HCP-style acquisition conventions while changing the age range [8]; ABCD offers a large multisite developmental sample [9]; UK Biobank offers population-scale multimodal imaging [10]; and NKI-Rockland combines lifespan imaging with broad phenotyping [11]. HCP-Aging is a plausible first Cognition candidate, and UK Biobank is a plausible Tobacco Use candidate, but those priorities are staging hypotheses rather than demonstrated measurement equivalence. Target items, pedigree overlap, acquisition, preprocessing, and covariates require an item-level audit before registration.

For any cohort, the experiment must pre-specify one target definition, one coherence–ridge implementation, one comparator, and one inference rule before opening outcomes. Behavioural overlap should be documented item by item rather than inferred from similarly named constructs.

7.2 Robustness analyses that remain data-staging tasks

Three neuroimaging sensitivity axes were discussed but not executed. Motion robustness needs subject-level RMS, framewise displacement, DVARS, or run-specific quality summaries aligned across the four HCP-YA resting-state runs. Global-signal-regression sensitivity needs materialized GSR versions of the same connectivity representations. Parcellation sensitivity needs comparable Schaefer-200/400 feature caches. Their absence is a staging blocker, not evidence that the pipeline is robust or non-robust on those axes [12–14].

7.3 A stricter confirmatory design

A decisive follow-up should reconstruct target residualisation entirely inside each training fold, preserve family or pedigree grouping, and evaluate both association and absolute predictive utility. The primary endpoint should require a positive out-of-sample R^2 or another prespecified calibration criterion in addition to correlation. Selection-adjusted inference on the old HCP-YA search would answer a different question; it cannot substitute for a new-cohort test of a genuinely pre-specified procedure.

8. Conclusion

BR began with a published HCP connectivity benchmark, explored 116 bounded candidate analyses, and identified a coherence-based direction worth carrying through fixed checks and a human-authorised comparison. In the cleanest subsequent comparison, that fixed pipeline exceeded a matched local reference for Cognition in every one of ten repeated partitions, with median $\Delta r = .098$, positive companion changes in R^2 and error, and a nominal conditional sensitivity of $p = .006$. Without retuning, the same pipeline also showed positive median correlation differences on four additional behavioural dimensions.

The stress tests determine the interpretation. Absolute median R^2 was positive only for Cognition and Tobacco Use; multiplicity-adjusted transfer was unsupported; a separate calibration-repair decision aid failed its gate; and no external cohort was tested. The result is therefore scientifically interesting but still bounded: a fixed, repeat-stable same-cohort connectivity candidate whose directional behaviour extends beyond its original target, not a confirmed or generally superior predictor.

Methods and provenance

Design summary

The 326-person HCP-YA source pool was partitioned into a 245-person discovery and development portion from 244 families and an 81-participant, 81-family internal holdout. The matched evaluation used 244 participants from 243 families with ten repeated five-outer by three-inner family-grouped partitions. The fixed pipeline used whole-band mean multitaper coherence-magnitude connectivity with cosine-kernel ridge prediction at $\alpha = 1$. The matched reference searched 74 eligible representations, four model families, and four regularisation values within each outer fold. The cross-outcome stress test applied the fixed pipeline without target-specific retuning. The separate one-time analysis opened the sealed internal holdout once using a different ridge procedure.

Public evidence and reproducibility

This public report preserves the logical evidence path rather than redistributing controlled HCP inputs or internal execution records: an open 20-plus-96-candidate screen, fixed diagnostics and a development-only selector audit, a researcher-authorised fixed comparison, repeated matched Cognition evaluation, cross-outcome stress tests, and separate calibration-repair and internal-holdout checks. The report source bundle contains the LaTeX text, bibliography, and the figures required to regenerate this PDF. The accompanying paper is available at arXiv:2608.19902; public software and portable reproducibility-pack documentation are available at the Brain Researcher public repository and its reproducibility directory. Those public materials document the software and portable verification format, but they are not an immutable replay package for this controlled-data episode. The numerical claim boundaries in this report therefore rest on the stated design and evidence tiers, not on access to private receipts. In particular, the 116-candidate screen is descriptive selection evidence, the ten repeated matched Cognition comparisons are same-cohort stability evidence, and the cross-outcome matrix is a transfer stress test rather than a collection of newly selected winners. The calibration-repair decision aid and the internal holdout remain limiting analyses because neither re-evaluated the exact headline comparison under an independent cohort design.

The self-contained source directory supports transparent document regeneration: it carries the report text, citation database, report style, and only the figure PDFs cited here. It deliberately does not carry person-level HCP data, target vectors, folds, prediction vectors, or private execution receipts. A reader can therefore inspect the scientific narrative, plots, equations, captions, and references without mistaking document reproducibility for a replay of the controlled analysis. Re-running a future external-cohort evaluation would require separately authorised data access and a new predeclared analysis plan.

References

[1] David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn human connectome project: An overview. *NeuroImage*, 80: 62–79, 2013. doi: 10.1016/j.neuroimage.2013.05.041.

[2] Zhen-Qi Liu, Andrea I. Luppi, Justine Y. Hansen, Ye Ella Tian, Andrew Zalesky, B. T. Thomas Yeo, Ben D. Fulcher, and Bratislav Misic. Benchmarking methods for mapping functional connectivity in the brain. *Nature Methods*, 22:1593–1602, 2025. doi: 10.1038/s41592-025-02704-4.

[3] Xilin Shen, Emily S. Finn, Dustin Scheinost, Monica D. Rosenberg, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nature Protocols*, 12:506–518, 2017. doi: 10.1038/nprot.2016.178.

[4] David J. Thomson. Spectrum estimation and harmonic analysis. *Proceedings of the IEEE*, 70 (9):1055–1096, 1982. doi: 10.1109/PROC.1982.12433.

[5] Anderson M. Winkler, Matthew A. Webster, Diego Vidaurre, Thomas E. Nichols, and Stephen M. Smith. Multi-level block permutation. *NeuroImage*, 123:253–268, 2015. doi: 10.1016/j.neuroimage.2015.05.092.

[6] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

[7] Gaël Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180:68–77, 2018. doi: 10.1016/j.neuroimage.2017.06.061.

[8] Michael P. Harms, Leah H. Somerville, Beau M. Ances, et al. Extending the human connectome project across ages: Imaging protocols for the lifespan development and aging projects. *NeuroImage*, 183:972–984, 2018. doi: 10.1016/j.neuroimage.2018.09.060.

[9] B. J. Casey, Tariq Cannonier, May I. Conley, et al. The adolescent brain cognitive development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018. doi: 10.1016/j.dcn.2018.03.001.

[10] Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19: 1523–1536, 2016. doi: 10.1038/nn.4393.

[11] Kate Brody Nooner, Stanley J. Colcombe, Russell H. Tobe, et al. The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, 6:152, 2012. doi: 10.3389/fnins.2012.00152.

[12] Rastko Ciric, Daniel H. Wolf, Jonathan D. Power, et al. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174–187, 2017. doi: 10.1016/j.neuroimage.2017.03.020.

[13] Kevin Murphy and Michael D. Fox. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *NeuroImage*, 154:169–173, 2017. doi: 10.1016/j.neuroimage.2016.11.052.

[14] Alexander Schaefer, Ru Kong, Evan M. Gordon, et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28:3095–3114, 2018. doi: 10.1093/cercor/bhx179.