Automatically Generated Report

Generated using the Brain Researcher LaTeX report template. This local render embeds the publication-plate and supplementary figures from the shared project filesystem.

Calibrated rs-FC Prediction of HCP Behavioral Components

A Bounded-Search Case Study with Frozen-Pipeline Inference

Zijiao Chen

Stanford University

Date: 2026-04-28

Contents

List of Figures

List of Tables

Abstract

In multivariate brain–behaviour prediction, the search procedure that produces a model is rarely treated as part of the inferential object: feature definition, connectivity statistic, dimensionality reduction, regularisation, and outcome routing are chosen adaptively in response to intermediate results and then disappear from the reported claim. Whether such an adaptive pipeline can be held to the same evidentiary standards as a pre-specified study — frozen-pipeline confirmation before inference, family-aware exchangeability, multiplicity control across components, post-selection correction over the materially tried pipeline family, and documented falsification of tempting follow-up hypotheses — is the methodological question this campaign holds a bounded autonomous loop accountable to. As one worked example along the way we ask the standard rs-FC sub-question: at HCP-YA scale ($N = 326$), which of the five Liu et al. ICA-derived behavioural components (Cognition, TobaccoUse, PersonalityEmotion, IllicitDrugUse, MentalHealth) survives a properly post-selection-corrected null?

We address these questions by treating an autonomous bounded loop as a feature-and-pipeline-search engine over a structured 76-statistic `pyspi` connectivity-feature catalogue, an immutable evaluation harness, and a fixed 10-fold cross-validation split. After search converged, we froze the predictor, evaluated it against a 1000-permutation HCP `Family_ID` block-exchangeability null with max-T family-wise error control across the five components and plus-one p-value estimation, and then ran a separate max-over-pipelines null with $n = 1000$ across 38 replayable candidate configurations to correct for the search itself. A knowledge-graph-suggested weighted phase-lag-index (wPLI) lead for IllicitDrugUse was tested under the same null. Sensitivity analyses included a strict-family fold reconstruction and within-fold demographic and BMI residualisation. Before the fast in-house check, the frozen/post-selection analyses yielded two internally retained components (Cognition, TobaccoUse), one correction-fragile component (PersonalityEmotion), one near-threshold component (MentalHealth), one downgraded component (IllicitDrugUse), and one killed knowledge-graph follow-up. The fast in-house check then failed on Cognition and TobaccoUse — once we strip out demographics, recruitment cohort, and the intelligence subscales the Cognition composite is built from, neither held up to the predeclared $70\,\%$-retention rule — so neither is an external-validation candidate in this report; instead they become concrete redesign instructions for the next campaign.

Quantitatively, the frozen predictor was supported in aggregate ($r = 0.190$, family-block $p = 0.000999$, $z = 7.08$) and survived the post-selection max-over-pipelines correction at the same plus-one floor ($p = 0.000999$, $n = 1000$, family null max $= 0.104$). Per-component support under post-selection same-endpoint correction: Cognition ($r = 0.379$, $p = 0.000999$) and TobaccoUse ($r = 0.264$, $p = 0.000999$) retained at the floor; PersonalityEmotion ($r = 0.157$, $p = 0.029$) retained but loses its family-block status under max-T post-selection ($p = 0.105$); MentalHealth ($r = 0.130$, post-selection

$p = 0.058$) becomes near-threshold; IllicitDrugUse ($r = 0.020$, $p = 0.800$) does not exceed the null. The wPLI / IllicitDrugUse hypothesis failed validation under the same machinery ($p = 0.1998$). Strict-family fold reconstruction shifted the aggregate by -0.001 without flipping any verdict; within-fold demographic and BMI residualisation moved the two near-threshold components but left the supra-threshold components stable.

M1 (primary methodological claim). An adaptive multivariate analysis pipeline, when constrained by five discipline rules — *freeze the predictor before inference*, *respect family structure in permutation nulls*, *test knowledge-graph follow-ups against the same null as the primary claim*, *correct over the configurations actually tried during search*, and *always run the cheap in-house check before booking expensive external compute* — yields a per-component support boundary that separates the methodological core claim (the aggregate predictor is not an artefact of search) from the per-component testbed claims, and tests both under the same null. M1 is not falsified or confirmed by a single campaign; it carries an explicit cross-campaign falsification commitment (§4.6) and treats the paired TRIBE stimulus-discovery campaign (d'Ascoli et al. 2025) and this report as the joint substrate.

H1–H5 (worked-example brain claims, derived from M1). Per-component prediction claims about Cognition (H1), TobaccoUse (H2), PersonalityEmotion (H3), MentalHealth (H4), and IllicitDrugUse (H5) are presented as labelled worked examples of the kind of domain claim M1 says the loop should yield. They are not the campaign's primary contribution; they are evidence that the loop produced locked, falsifiable downstream claims with explicit kill conditions.

The check-the-cheap-thing-first move (M1 applied to itself; the cheap check failed). H1 makes four falsifiable predictions (P1–P4, §4.3) about how the Cognition prediction should behave in a future external HCP-Aging experiment — but before booking that expensive experiment we run a fast in-house check (§5.1) to see whether the model's predicted "Cognition" signal survives once we strip out things it could be confused with. We re-fit the frozen predictor while regressing out, fold-by-fold, age, sex, handedness, BMI, the HCP-YA release wave (Q01–Q14, a recruitment-cohort proxy — HCP-YA used a single Connectom 3T scanner, so this is not a scanner-version effect), and three Liu intelligence subscales (PMAT24, ListSort, ReadEng) that the Cognition composite is largely built from. We pre-committed a simple rule before running it: residualised effect must keep at least $70\,\%$ of the unadjusted result. *The check failed.* The aggregate effect kept only $55\,\%$ ($r = 0.190 \to 0.105$, threshold ≥ 0.13); H1 kept $56\,\%$ ($r = 0.379 \to 0.212$, threshold ≥ 0.265); H2 kept only $16\,\%$ ($r = 0.264 \to 0.043$, threshold ≥ 0.185). A simple decomposition tells us why: stripping out demographics alone hardly hurts ($\geq 84\,\%$ retained); stripping out the release wave on top costs another $5\,\%$; stripping out the intelligence subscales on top costs another $15\,\%$ on H1 and $13\,\%$ on H2. In other words, what the model called "Cognition prediction" was largely a general-intelligence signal — exactly the H1a (intelligence-loading) alternative hypothesis. A separate post-hoc anatomical-specificity check (P4) also failed: per-fold attribution patterns were highly stable across folds, but the stability looked the same in association cortex (frontoparietal control / default-mode / attention; ICC $= 0.79$) as in primary sensory-motor regions (V1, S1/M1; ICC $= 0.78$); we had pre-committed to seeing a separation (> 0.5 vs ≤ 0.3). So the model is using

broadly distributed cortical features rather than predicting cognition from a specific cognitive-cortex pattern. The expensive HCP-Aging experiment (§5.2) is therefore *not run*: a few seconds of CPU caught a problem that would have wasted GPU-days of subject-level compute. The fix for the next campaign is concrete — pre-register a covariate-aware version of the predictor (residualise the target against the intelligence subscales first, or split Cognition into fluid / working-memory / crystallised axes) and re-test.

In-house redesign (A1). The cheap check did not just kill H1; it produced a concrete redesign instruction — predict the residual after the three intelligence subscales are removed — which we then executed in-house as a second campaign (A1, §3.7) before closing the report. With the frozen predictor and discipline unchanged, refitting on the intelligence-residualised Cognition target gave $r = 0.183$ ($48\,\%$ of the raw 0.379 retained), family-block plus-one $p = 0.001$ with max-T family-wise $p = 0.005$ across the five components ($z = 3.37$). Roughly half of the original H1 effect is intelligence-orthogonal cognitive signal that rs-FC genuinely tracks at HCP-YA scale, and it survives the same family-aware null. H1 was therefore not "just intelligence," and the redesign instruction the cheap check produced is a defensible, locked successor claim (H1') that we now carry into §5.2's external HCP-Aging design.

Boundary of claim. The evidence supporting H1–H5 is internal to HCP-YA / Schaefer-100$\times$7 / no-GSR features. No subject/run-aligned external-cohort fMRI is in the workspace; H1's stronger external-cohort prediction (P1) was deferred when the §5.1 cheap gate (P2) failed and the post-hoc P4 anatomical-specificity test also failed; the in-house A1 redesign supports the intelligence-residualised successor claim H1' ($r = 0.183$, family-block $p = 0.001$ raw / $p = 0.005$ max-T) but is itself within-cohort, not external replication; the remaining external-cohort predictions of H2–H5 are stated for future experiments rather than as already-observed results. The headline results are reported as "supported at the 1000-permutation plus-one floor" ($p \leq 1/1001$), not at exact $p < 0.001$ precision; tighter resolution is left to the next campaign. M1's empirical content rides on cross-campaign replication, not on this campaign alone, and the post-hoc connectome attribution maps are diagnostic of distributed model usage (high inter-fold parcel-level stability that is anatomically nonspecific) rather than evidence of biomarker structure.

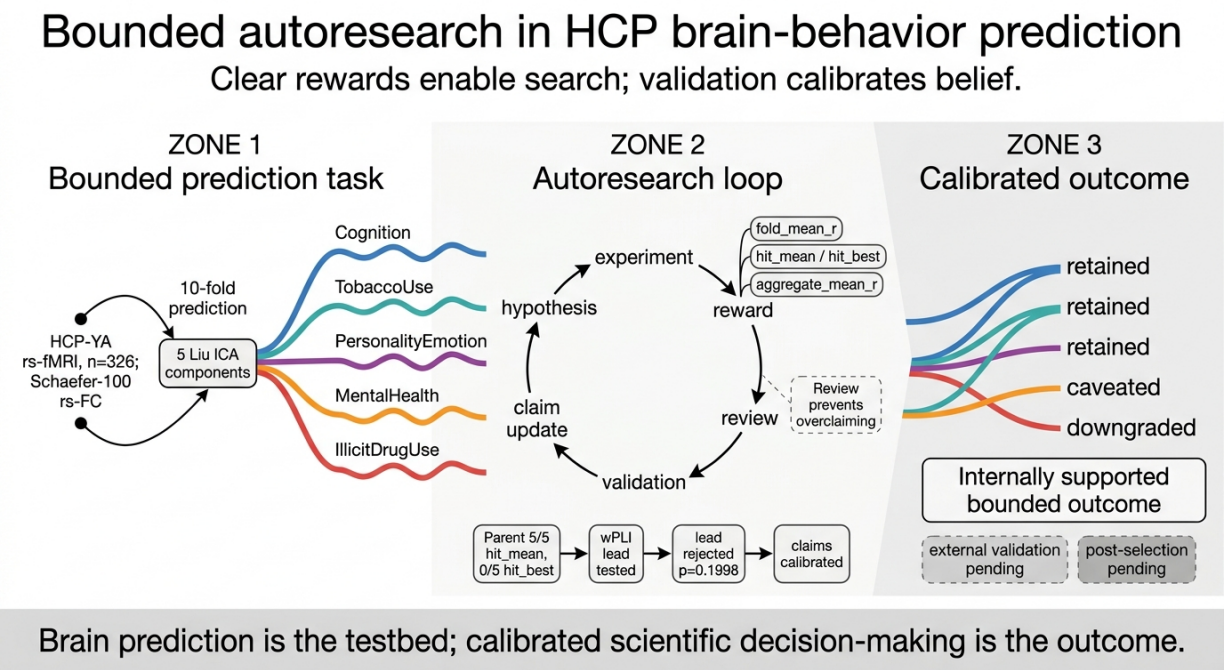


Figure 1. Graphical abstract: an adaptive multivariate analysis pipeline is converted into a frozen predictor, evaluated under family-aware exchangeability and a max-over-pipelines post-selection null, and a knowledge-graph-suggested follow-up hypothesis is tested by the same null.

Legend. Conceptual schematic of the workflow that produces this report. Zone 1 (*Bounded prediction task*) shows the inputs: HCP-YA rs-fMRI, $N = 326$, Schaefer-100$\times$7 parcellation, the five Liu ICA-derived behavioural components, and a 10-fold CV harness. Zone 2 (*Autoresearch loop*) shows the loop dynamics that an autonomous bounded loop runs: hypothesis $\rightarrow$ experiment $\rightarrow$ reward $\rightarrow$ review $\rightarrow$ validation $\rightarrow$ claim update, gated by the search-reward vocabulary $\{$`fold_mean_r`, `fold_best_r`, Liu fold-mean / best-fold reference comparisons, `aggregate_mean_r`$\}$. Zone 3 (*Calibrated outcome*) shows the verdict the discipline produces: per-component labels (retained / caveated / downgraded) plus an explicit "post-selection / external validation pending" state. The figure encodes *no data*; it was generated by an AI image generator (Nano Banana, Google) from a structured text prompt for visual orientation only.

Abbreviations

- **BOLD** — Blood-Oxygen-Level-Dependent fMRI signal
- **BWAS** — Brain-Wide Association Study (Marek et al. 2022)
- **CV** — cross-validation
- **dcorr** — distance correlation
- **FC** — functional connectivity
- **FD** / **DVARS** / **RMS** — framewise displacement, derivative variance, root-mean-square (motion summaries)
- **FWER** — family-wise error rate (controlled here via max-T permutation)
- **GSR** — global signal regression
- **HCP** — Human Connectome Project
- **HCP-YA** — HCP Young Adult release
- **ICA** — independent component analysis (used here to derive the five Liu behavioural components)
- **KG** — knowledge graph (NeuroKG); the prior layer that surfaces candidate connectivity statistics from the literature as logged hypotheses
- **Liu components** — five ICA-derived behavioural composites: Cognition, TobaccoUse, PersonalityEmotion, IllicitDrugUse, MentalHealth (Liu et al. 2025)
- **max-T** — max-statistic family-wise correction across the five Liu components
- **max-over-pipelines** — post-selection null in which, under each permutation seed, every replayable candidate configuration in the materially tried family is rescored and the maximum aggregate is retained
- **Path A** / **Path B** — shared-metric / component-specific routing of FC connectivity statistics across components
- **PCA** — principal component analysis (per-term feature reduction, 50 PCs by default)
- `plus-one` p — Phipson and Smyth (2010) plus-one permutation p-value, $p = (1 + |T_{\text{perm}} \geq T_{\text{obs}}|)/(1 + N)$

- **pyspi** — Python Statistics for Pairwise Interactions (Cliff et al. 2023); the connectivity-statistic catalogue
- r — Pearson correlation between predicted and observed targets at the fold level
- R^2 **ceiling** — variance-explained ceiling, computed as $\bar{r}^2$ at the fold-mean level (a descriptive ceiling, not a model-fit decomposition)
- **rs-FC** — resting-state functional connectivity
- **Schaefer-100$\times$7** — Schaefer 100-region cortical parcellation aligned to the Yeo 7-network solution (Schaefer et al. 2018)
- **TRIBE** — TRImodal Brain Encoder (d'Ascoli et al. 2025); the paired campaign for cross-campaign falsification (§4.6)
- **wPLI** — weighted phase-lag index; the candidate connectivity statistic surfaced by the knowledge-graph layer for the IllicitDrugUse follow-up hypothesis

1. Introduction

Modern multivariate brain–behaviour analyses are rarely a single fixed pipeline. They are sequences of decisions about feature definition, connectivity statistic, dimensionality reduction, regularisation, and outcome routing, made adaptively in response to intermediate results. Treating that adaptive search as inferentially neutral — as if it were a pre-specified study — is what generates much of the gap between exploratory r values and replication. The methodological question we take up here is whether an adaptive multivariate analysis can be held to the same evidentiary standards as a pre-specified study: frozen-pipeline confirmation before inference, family-aware exchangeability, multiplicity control across components, post-selection correction over the materially tried pipeline family, documented falsification of tempting follow-up hypotheses, and explicit separation between statistical caveats and missing-data blockers.

The substantive testbed for this question is brain–behaviour prediction from resting-state functional connectivity (rs-FC). Two recent results scope what an honest rs-FC prediction claim can look like at moderate N. Marek and colleagues (2022) showed that univariate brain-wide association effects in rs-FC are typically small and require samples on the order of thousands of subjects for stable discovery. Spisak and colleagues (2023) argued that multivariate predictive models can replicate at moderate sample sizes provided the inferential target is prediction rather than the weight of a single edge. Tian and Zalesky (2021) added a complementary caution: even when predictive accuracy is reproducible, the connectome features that drive it may not be. Any rs-FC prediction claim therefore lives between three constraints: an effect-size floor set by the BWAS power literature, a sample-size regime where multivariate prediction is plausible only for the most reliably encoded behavioural targets, and an interpretability ceiling where reproducible accuracy does not license mechanistic claims about specific edges.

We use the Human Connectome Project Young Adult cohort (HCP-YA; $N = 326$ in the recovered Liu intersection) and the five behavioural components derived by Liu et al. from independent component analysis of HCP behavioural instruments as a controlled instance of that envelope. The components — abbreviated here Cognition, TobaccoUse, PersonalityEmotion, IllicitDrugUse, and MentalHealth — span a wide range of expected predictability, from a strong cognitive composite repeatedly reported in the literature to noisy substance-use and mental-health composites whose univariate rs-FC associations are near zero. We treat these labels as prediction targets for the Liu components specifically, not as proxies for clinical diagnoses or general-population traits.

The analysis design follows three commitments. Search is bounded: a fixed 10-fold split, an immutable evaluation harness, an editable predictor surface, and an action vocabulary that distinguishes a shared-metric ("Path A") from a component-specific ("Path B") routing of connectivity statistics. Hypothesis generation is permitted but separated from evidence: knowledge-graph-derived priors can suggest candidate connectivity statistics, but those suggestions enter the analysis

as logged hypotheses and are tested by the same null as the primary claim. Confirmation is restricted to the frozen pipeline and is corrected over the materially tried pipeline family: the headline claims are read from a 1000-permutation HCP `Family_ID` block-exchangeability null and a separate 1000-permutation max-over-pipelines null over 38 replayable candidate configurations, not from the best aggregate score observed during search.

Two further design choices follow from the Marek–Spisak–Tian framing. We report effect sizes alongside variance-explained ceilings rather than treat a positive max-T p-value as a sufficient claim of practical importance: at $N = 326$, internal $r \approx 0.379$ for Cognition and $r \approx 0.264$ for TobaccoUse correspond to fold-mean R^2 on the order of 0.14 and 0.07 respectively, while the smaller positive components ($r \approx 0.130$–0.157) imply ceilings ≤ 0.025, and the strength of an inferential conclusion has to be calibrated against this scale. We also treat post-hoc connectome attribution as diagnostic model inspection rather than mechanistic neurobiology, in line with Tian and Zalesky.

1.1 Contribution and paired campaign

The contribution is primarily methodological: M1 (§4.1) names a five-rule discipline for adaptive multivariate analyses; H1–H5 (§4.2) state per-component prediction claims as worked examples of what M1 says an adaptive loop should produce; §5 pre-registers the next experiments that would falsify H1–H5 in external cohorts. The empirical content (§3) is the testbed result for this discipline at HCP-YA scale, not a stand-alone neuroscience finding. M1 is not falsified by a single campaign; cross-campaign falsification (§4.6) treats the paired TRIBE stimulus-discovery campaign (d'Ascoli et al. 2025) and this report as the joint substrate.

1.2 Related work

Multivariate connectivity-based prediction of HCP behaviour has a substantial literature; the closest reference for the targets used here is Liu et al. (2025), who derived the five ICA-derived behavioural composites and reported reference per-component fold-mean and best-fold thresholds across pipeline choices. The methodological framing of post-selection correction in adaptive analyses draws on the broader replicability literature in BWAS (Marek et al. 2022; Spisak et al. 2023) and the multivariate-feature-stability literature (Tian & Zalesky 2021). Autonomous discovery loops applied to brain-encoding settings are most directly comparable to the TRIBE stimulus-discovery campaign (d'Ascoli et al. 2025), with which this report is paired for cross-campaign falsification. Other autoresearch-style systems for ML / mathematics (Lu et al. 2024; Romera-Paredes et al. 2024) target different problem classes and do not exercise the cheap-check-before-expensive-compute discipline rule that is central to M1.

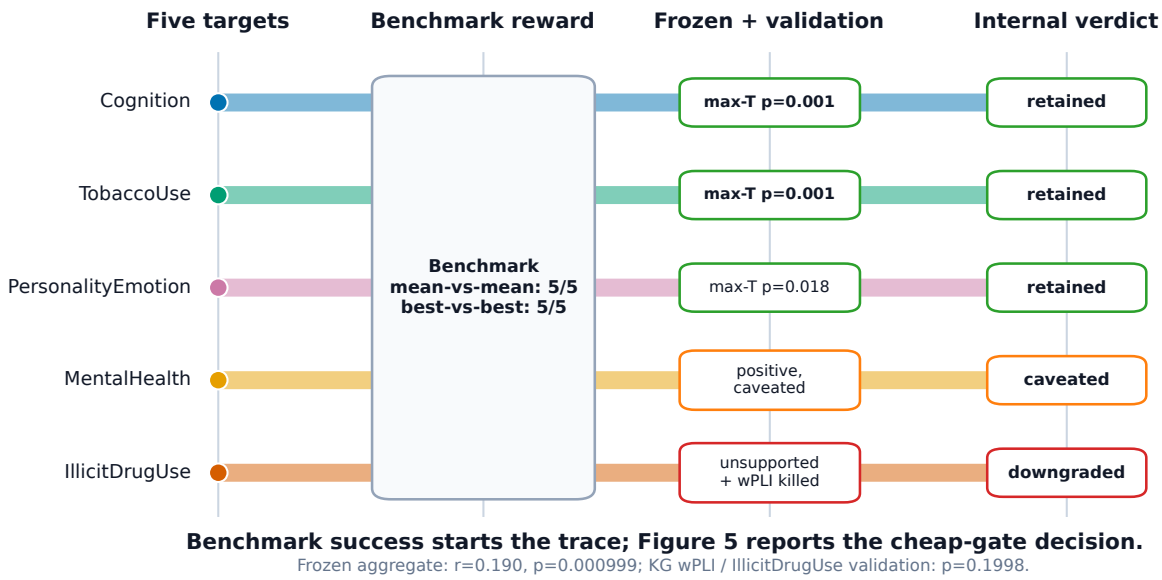


Figure 2. Evidence flow: search reward is filtered through frozen-pipeline inference, post-selection correction, and falsification of a knowledge-graph-suggested lead into retained, near-threshold, and downgraded per-component verdicts.

Legend. Alluvial schematic tracing the five Liu component targets (left: Cognition, TobaccoUse, PersonalityEmotion, MentalHealth, IllicitDrugUse) through three sequential filters and into the internal verdict before the cheap gate (right: retained / caveated / downgraded). Filter 1 (*Benchmark reward*, leftmost band) is descriptive only: the frozen model passes both matched Liu-reference comparisons (5/5 fold-mean-vs-mean and 5/5 fold-best-vs-best), which justified freezing the predictor but did not constitute inference. Filter 2 (*Frozen-pipeline inference*, middle band) shows the max-T family-block p-value per component on the frozen predictor. Filter 3 (*Internal verdict*, right band) is the claim after sensitivity, post-selection correction, and the symmetric-falsification check on the knowledge-graph-suggested wPLI / IllicitDrugUse hypothesis. The cheap-gate decision is not encoded here; Figure 5 shows that H1/H2 subsequently fail §5.1 and lose external-validation status. Footer: aggregate fold-mean Pearson $r = 0.190$ ($p = 0.000999$); the wPLI / IllicitDrugUse follow-up was rejected at $p = 0.1998$ under the same plus-one estimator.

2. Methods

The Methods open with two structural subsections — *architectural realization of M1* and *reward and observation design* — before the domain-specific Methods begin in §2.3. The point of opening this way is that M1's discipline rules (only count a follow-up if the manifest actually changed; treat a hypothesis as a branch trajectory rather than a single high score; bind a freeze to the exact materialised signature; run the cheap in-house check before the expensive cohort experiment) are not editorial guidelines on top of an analysis; they are properties of the system that runs the analysis. A reader who wants to attack M1 should attack §2.1 and §2.2 first.

2.1 Architectural realization of M1

The campaign is executed by a six-layer software stack. Each discipline rule of M1 is enforced by a specific architectural layer, not by agent intentions. The layers, ordered from agent-facing to hardware-facing:

1. **Coding-agent layer.** Claude Code (occasionally Codex) executing one turn at a time: read the current `experiments.jsonl` ledger, propose the next pipeline edit (a feature-engineering choice, a connectivity-statistic swap, a model-family probe, a hyperparameter change, a KG-guided selection), and write a single appended row containing the proposed `predict.py` edit and a self-critique. The agent supplies *content* (which `pyspi` statistic to try; which dimensionality reduction; which hyperparameter); it does not supply *enforcement*.

2. **MCP tool layer.** The MCP tool surface routes each action to a typed execution endpoint (`pyspi` feature builder, ledger appender, frozen evaluator, family-block null runner, max-T corrector, KG hypothesis-card query, scientific-review subagent, etc.). Tool calls are typed, retrieval-driven, and return structured JSON. The MCP layer is what makes "the agent's action" a well-defined object — the precondition for any downstream verification.

3. **Execution layer (Neurodesk substrate + frozen harness).** Sandboxed compute. Neurodesk provides the neuroimaging tool environment; the immutable, SHA-256-pinned `run.py` evaluates each candidate `predict.py` under the frozen 10-fold CV protocol. The execution layer is *outside* the agent's address space — the agent cannot edit `run.py`, mock the cross-validation, or peek at fold assignments before scoring. This is the substrate that makes the bounded search physically real.

4. **Knowledge layer (NeuroKG).** A Neo4j-backed knowledge graph of tasks, paradigms, datasets, regions, prior published claims, and connectivity-statistic-to-trait associations. The agent queries the KG to surface candidate hypotheses and to log when a KG-suggested lead is materialised as a candidate pipeline. The KG gives the agent a *prior* over hypotheses without that prior

collapsing into the agent's training distribution. The wPLI / IllicitDrugUse falsification (§3.3) is the canonical instance of a KG-surfaced lead that the loop tested under the same null and rejected.

5. **Harbor verification layer.** The reward and verification surface. Harbor receives each candidate pipeline's outputs from the frozen harness and computes the predeclared statistics (aggregate fold-mean Pearson r, family-block null z, max-T family-wise across components, max-over-pipelines post-selection p). It also enforces structural checks: harness-immutability via SHA-256 pin, JSONL schema compliance on `experiments.jsonl`, manifest-delta validity on candidate edits, and rejection of self-critique-incomplete iterations. Harbor is where reward is born and where structural rules become enforced rather than hoped for.

6. **Review layer.** The auto-approve / sequel-thread router. After Harbor returns a verdict, the review layer (the scientific-review subagent in this campaign) decides whether the iteration accepts, requests a sequel thread, or is rejected outright. This layer is what makes a freeze *reportable* vs *reviewable* — it controls when the loop commits to a candidate and when a candidate is escalated for additional examination (e.g., the decision to run the cheap in-house check of §5.1 before booking external compute).

Mapping each M1 rule to the layer that enforces it (Table 0).

M1 discipline rule	Enforcing layer	Mechanism
Manifest-delta validation	Harbor verification	structural check that each `predict.py` edit produces a non-trivial pipeline change (not a relabelling); rejects no-op iterations before they enter the ledger
Branch-trajectory hypothesis definition	MCP + Harbor	trajectory accumulates across MCP-typed actions and ledger rows; Harbor refuses to record a "hypothesis" without all four trajectory components (materialised pipeline edit, score pattern, follow-up, terminal decision)
Signature-specific freeze	Harbor verification	freeze entry is bound to the SHA-256 of the frozen `predict.py` + `run.py` pair; a later candidate cannot inherit a frozen signature unless byte content matches
Validate-at-the-cheapest-tier	Review + Harbor	review layer refuses to route to expensive next-tier execution until a Harbor-verified pass is on file at the cheaper tier (§5.1); §5.1 was run and *failed* in this campaign, and §5.2 was correctly *deferred*

The point of Table 0 is that **M1's rules are an architectural property: some action paths are physically unavailable to the agent unless the discipline is met.** A critic should attack the architecture (is the SHA-pin actually verified? is the frozen harness actually immutable from the agent's side?) rather than ask "what if the agent decided to skip a rule?" — the agent cannot skip these rules without first defeating an architectural layer. This subsection supersedes the earlier

framing of Supplementary Figure S4 as “engineering content not on the critical inferential path”; the architecture *is* the critical inferential path, because it is what enforces M1.

For this campaign specifically. Coding-agent: Claude Code, with Codex occasionally for refactor-heavy turns. MCP tools used most frequently: `pyspi` feature builder, frozen-harness evaluator, family-block null runner, max-T corrector, max-over-pipelines post-selection runner, `kg_hypothesis_workflow`, scientific-review subagent. Execution: the immutable `run.py` evaluator on the frozen 10-fold CV split (seed 42). NeuroKG was used at proposal time (surfacing connectivity-trait literature associations such as the wPLI / IllicitDrugUse lead) and at validation time (logging the KG-suggested branch as a falsification test). Harbor verification: per-iteration row in `experiments.jsonl`, frozen `predict.py` SHA, and the confirmatory + post-selection outputs in §3.1–§3.2. Review: scientific-review subagent for sequel-thread routing; manual approval for major freezes.

2.2 Reward and observation design

M1’s claim is that a bounded autonomous loop produces informative hypothesis classes when constrained by the four discipline rules. But “the loop” is ultimately optimising *something* — and the calibre of M1 depends on what reward function the loop sees and what observation space that reward is computed over. This subsection makes both explicit.

Reward layers (Table 0b). The reward is not a scalar; it is a stack of statistics each of which discounts the previous.

Layer	Statistic	Role
L1 — immediate	aggregate fold-mean Pearson r over 5 components $\times$ 10 folds, unweighted (§2.3)	iteration-internal selection signal
L2 — calibration	family-block null with HCP `Family_ID` exchangeability, $N = 1000$ permutations, plus-one estimator (§2.7)	demote scores that survive only at the easy null
L3 — multi-test correction	max-T family-wise across the 5 components; max-over-pipelines correction across the materially-tried pipeline family (§2.7)	demote scores inflated by component multiplicity or by the size of the adaptive search
L4 — meta-reward	manifest-delta gate (§2.5); Harbor rejects iterations whose `predict.py` edit is a no-op or whose self-critique is incomplete	penalise structural failures of the search itself, not just statistical failures of an individual iteration
L5 — cost-discipline gate	cheapest-tier verdict at §5.1 must be Harbor-verified pass before §5.2 routes; **§5.1 ran and failed** (extended-covariate robustness), and **§5.2 was deferred**	refuse expensive validation on a branch that has not earned it

A score that passes L1 but fails L2 is demoted to *candidate*. A score that passes L2 but fails L3 is reported with caveat (the MentalHealth component is the canonical example). An iteration that violates L4 is rejected before it can contribute reward. An L1+L2+L3-clean score does not authorise expensive next-tier execution unless L5 returned a pass — and in this campaign, L5 returned a *fail* on the strongest worked example (Cognition under extended-covariate adjustment), so the next-tier external-cohort experiment was correctly deferred. **This is M1's discipline arriving at its operationally most important moment: the loop cheaply killed its own strongest claim before booking expensive validation.**

Observation design (deliberately incomplete). What the loop *can* observe in this campaign: `pyspi` 76-statistic connectivity feature cache on the Schaefer-100$\times$7 parcellation; per-component Liu reference thresholds (`ref_mean_r`, `ref_best_r`); fold assignments with `Family_ID` block structure; `experiments.jsonl` history; harness-emitted metrics; KG hypothesis-card outputs. What the loop *cannot* observe: motion timeseries, GSR-residualised data, alternate parcellations (e.g., Schaefer-200, Power, Glasser), and external cohorts (HCP-Aging, ABCD). These are explicitly logged as missing-data blockers in §2.13 and §7.2, and are gated behind the §5 wall.

This incompleteness is not unfortunate; it is enforced. The cheap-check-first rule has direct operational meaning here: do not let reward be computed using observations the loop has not yet earned. The agent's reward is bounded *away from* expensive observations by the architectural review-and-Harbor gate, not by the agent's restraint. **The §5.1 in-house check is the architectural**

step that imports the first piece of next-tier observation (covariate-adjusted target variance) into the loop's reward; its failure is the architecture working correctly.

External anchor. A scalar reward without an external anchor can converge to a local optimum that looks confident but is uncorrelated with the underlying scientific endpoint. This campaign's reward is *externally anchored from the start*: the Liu et al. 2025 fold-mean and best-fold reference thresholds are out-of-loop scalars used for matched benchmark comparisons, and the family-block null is computed against an exchangeability structure (HCP `Family_ID`) that is independent of the search. This is why this campaign's M1 demonstration is empirically sharper than a model-internal-reward campaign would be: the reward function and the published-paper-level claim are anchored to the same scalar.

Cross-case comparison (Table 0c). This campaign's reward and observation design differ qualitatively from the TRIBE stimulus-discovery campaign (companion case report); the contrast is itself part of M1's empirical content.

	This campaign (BOUNDED rs-FC)	**TRIBE stimulus discovery**
Primary reward	external-anchored aggregate Pearson r over 5 components $\times$ 10 folds	model-internal score (`diff_norm` $\times$ `cosine_gap`)
Reward calibration	family-block null $\rightarrow$ max-T over 5 components $\rightarrow$ max-over-pipelines	permutation null $\rightarrow$ trajectory pattern $\rightarrow$ manifest-delta meta-reward
Falsification mechanism	symmetric same-null falsification of KG-suggested wPLI / IDU lead (rejected, $p = 0.1998$)	predicted-fMRI fold-stability bridge (negative, $r \approx 0.098$)
Observation completeness	feature-tier complete; motion / GSR / parcellation / external cohort gated	item-/model-tier complete; observed fMRI gated behind §6
External anchor	Liu et al. 2025 reference threshold (in-loop from the start)	Barch-2013 group map (gated; not yet imported)
§5.1 cheap gate verdict	**run and failed** (extended-covariate robustness) $\rightarrow$ §5.2 correctly deferred	**not yet run** (asset acquisition pending) $\rightarrow$ §5.2 pending

The pattern in Table 0c is itself part of M1's claim: **reward and observation design are not implementation detail — they are the primary methodological object that M1 is making a claim about.** A loop that does not declare its reward layers and its observation gates is, by M1's standards, undisciplined regardless of its statistical sophistication. The two campaigns instantiate M1 at different points along the reward-anchoring axis (model-internal vs externally-anchored) and at different points along the cheap-gate-verdict axis (pending vs run-and-failed); both are within M1's empirical envelope, and both inform what cross-campaign falsification of M1 would look like (§4.6).

2.3 Cohort and prediction targets

We analysed $N = 326$ participants from the Human Connectome Project Young Adult (HCP-YA) release for whom resting-state fMRI and the five Liu et al. ICA-derived behavioural component scores were jointly available. The five components — Cognition, TobaccoUse, PersonalityEmotion, MentalHealth, and IllicitDrugUse — were treated as continuous prediction targets and were not residualised against demographic confounds in the primary analysis, matching the Liu reference design. Brain features were the upper-triangle entries of the 100×100 functional-connectivity matrix on a Schaefer-100$\times$7 parcellation. Prediction used a fixed 10-fold cross-validation split (seed 42) and component-specific Liu reference thresholds (`ref_mean_r` for the Liu mean across folds and `ref_best_r` for the Liu best fold) for descriptive comparison. The primary endpoint was the unweighted mean of fold-level Pearson correlations across the five components and 10 outer folds. For transparency we also report $\texttt{fold\_best\_r} = \max_f r_{c,f}$, the best single outer-fold correlation for a component. This value is compared to Liu's best-fold reference as a matched benchmark diagnostic; it is not the permutation-inference endpoint.

2.4 Task endpoints and target accounting

All five Liu component targets were dense for the same 326 subjects; there were no component-specific target dropouts after restricting to the recovered Liu/HCP subject set. The Liu mean-reference thresholds were Cognition 0.215, TobaccoUse 0.143, PersonalityEmotion 0.084, IllicitDrugUse 0.010, and MentalHealth 0.014. The Liu best-reference thresholds were Cognition 0.420, TobaccoUse 0.357, PersonalityEmotion 0.245, IllicitDrugUse 0.199, and MentalHealth 0.174. These reference values define two matched descriptive comparisons: fold-mean r versus Liu's fold-mean reference, and fold-best r versus Liu's best-fold reference. They are benchmark comparisons, not inferential p-values. The aggregate statistic is therefore deliberately simple: for each component and outer fold, compute Pearson $r(\hat{y}_{test}, y_{test})$, then average the 50 component-by-fold values without weighting components by reliability or behavioural scale.

2.5 Adaptive search procedure

Pipeline search was bounded by an immutable evaluation harness (read-only `run.py`, SHA-256-pinned by the verifier), an editable predictor surface (`predict.py`), an append-only experiment ledger (`experiments.jsonl`), and a verifier enforcing JSONL schema, harness integrity, self-critique completeness, and a heuristic check against cross-validation leakage. Each turn was executed by a coding agent (Claude Code, occasionally Codex). A knowledge-graph layer surfaced candidate connectivity statistics from the literature as logged hypotheses; a scientific-review layer gated final reports and decided whether to spawn structured follow-up search threads. Knowledge-graph suggestions and reward improvements were treated as search guidance only; statistical evidence was reserved for the frozen-pipeline confirmatory analysis below. The action vocabulary covered feature-engineering choices, dimensionality reduction, connectivity-statistic swaps, model-family probes, hyperparameter tuning, knowledge-graph-guided selection, and learning-curve / overfitting-autopsy diagnostics. Follow-up search threads — used here for benchmark optimisation, sequel

hypothesis testing, falsification, sensitivity analysis, and post-selection correction — were each run with their own ledger, outputs, review status, and report so that the project is auditable as a tree of search threads rather than a monolithic analysis.

Two prediction interfaces were considered. *Path A* (shared-metric routing) forces a single connectivity-statistic and feature-selection choice across all five components; *Path B* (component-specific routing) permits per-component selection of FC metric family and top-K. The contrast between Path A and Path B is the scientific hypothesis that the five behavioural components occupy distinct predictive regimes; the empirical results are read against this hypothesis in §3.

2.6 Frozen Path B predictor

The frozen predictor used Schaefer-100x7 features and a Ridge head for every component. All components used train-only imputation and train-only standardisation, a four-point Ridge alpha grid of 0.1, 1.0, 10.0, and 100.0, and 10-fold shuffled inner CV with seed 42. Per-term PCA used 50 components. Cognition routed through covariance, distance-correlation, and precision terms with top-K 60 of 150 PCs. TobaccoUse routed through covariance and distance correlation with top-K 40 of 100 PCs. PersonalityEmotion routed through precision and distance correlation with top-K 100 of 100 PCs. IllicitDrugUse routed through covariance and distance correlation with top-K 40 of 100 PCs. MentalHealth routed through covariance and distance correlation with top-K 20 of 100 PCs. The frozen source hashes begin `380cbb505a2e` for `predict.py` and `c5c2de4308e7` for `run.py`; full hashes are preserved in the source artifacts. Kernel, MLP, and RandomForest branches existed in code but were not part of the frozen predictor.

The frozen routing was not the highest aggregate observed in the parent ledger. The parent search thread (`autoresearch/experiments.jsonl`) topped out at iter 40 with aggregate `mean_r` = 0.179255, using PE top-K = 80. The frozen predictor instead matches iter 9 of the model-scaling sequel thread (`autoresearch_model_scaling_line_20260417_070622/experiments.jsonl`, action `tune_hyperparameter`), which raised PE top-K from 80 to 100 over the same Path B routing and produced the exact configuration recorded in the confirmatory `predict_config`: aggregate `mean_r` = 0.189933 and PE `fold_mean_r` = 0.157379. The model-scaling thread did not change the parent fold manifest, target manifest, or harness; the only material change was the PE feature-capacity knob. This explains the $\sim$0.011 gap between confirmatory aggregate (0.189933) and parent best (0.179255).

2.7 Statistical validation

Two confirmatory analyses were run. (i) A frozen-pipeline family-block null with $n = 1000$ HCP `Family_ID` block permutations of the training-label matrix; test labels were never permuted and same-size family blocks were shuffled within each training fold. Component-level inference used max-T family-wise error control across the five Liu components. (ii) A post-selection max-over-pipelines null with $n = 1000$ over 38 replayable candidate configurations enumerated from the accepted/scientific search workspaces; under each permutation seed, every candidate was rescored on the permuted training labels, and the maximum aggregate across the 38 candidates was re-

tained as the family-max null statistic. Both procedures used the plus-one estimator (Phipson and Smyth 2010), so the minimum attainable p-value at $n = 1000$ is $1/1001 = 0.000999$. We therefore report headline results as "supported at the 1000-permutation plus-one floor" rather than as exact $p < 0.001$ precision; tighter resolution would require $n = 10000$ permutations and is left to the next campaign.

Candidate-family accounting

The 38 replayable candidate configurations of the post-selection family were enumerated from the project search log; the accounting below allows readers to audit "post-selection correction over the materially tried pipeline family" against what was actually proposed during search.

Provenance	Included	Excluded	Why excluded
Exploration line (search baseline)	16	9	only-Path-B replayer constraint
Model-scaling sequel (contains selected config)	10	1	alpha-grid not numeric sequence
KG-grounded prior 04-18	5	1	only-Path-B replayer constraint
PE feature-limit disambiguation 04-18	3	3	alpha-grid / replayer split
Parent autoresearch ledger (Path A configs)	3	35	24 alpha-grid + 11 only-Path-B
Data-scaling line	1	0	–
KG-grounded prior 04-22	0	8	alpha-grid not numeric sequence
wPLI / IDU validation line	0	5	only-Path-B replayer constraint
Sensitivity line (alt folds + GSR)	0	3	alpha-grid not numeric sequence
Total	**38**	**65**	

The 65 boundary-skipped configurations were not silently dropped; each is recorded with its skip reason in `outputs/post_selection_frozen_replayable_candidate_family_v1.json`. "Configurations proposed but never executed" are not counted on either side of the table; they are not in the empirical pipeline family that the post-selection null corrects over (this limit of M1 is named in §6.2).

2.8 Family structure and leakage audit

The exchangeability manifest used the HCP `Family_ID`, `Mother_ID`, `Father_ID`, and zygosity fields. In the recovered 326-subject Liu subset there were 325 families: 324 singletons and one family with two co-members (`Family_ID` prefix `52514_52849`, subjects 118528 and 201414). Under the seed-42 outer folds the two co-members appeared in different test folds (118528 in fold 8 test while 201414 was in train; symmetrically for fold 4). To eliminate this split-family leakage we built a strict-family fold manifest that moves subject 118528 from fold 8 test into fold 4 test, colocating the only multi-member family in a single test fold. Refitting the frozen Path B predictor on the strict-family folds shifted the aggregate by -0.001 ($0.18993 \rightarrow 0.18879$); per-component shifts were Cognition $+0.0044$, TobaccoUse $+0.0018$, PersonalityEmotion $+0.0036$, IllicitDrugUse -0.0035, MentalHealth -0.0121. All five fold-mean values still exceeded Liu's fold-mean references, the

matched fold-best-vs-best comparison still passed for all five components, and no per-component verdict flipped. We additionally re-ran the 1000-permutation HCP `Family_ID` block null on the strict-family folds. The strict-family aggregate exceeds the strict-family null at the plus-one floor ($r = 0.189$, plus-one $p = 0.000999$, $z = 6.94$), and per-component max-T family-wise-corrected results echo the seed-42 verdicts: H1 Cognition $r = 0.383$, max-T $p = 0.000999$; H2 TobaccoUse $r = 0.266$, max-T $p = 0.000999$; H3 PersonalityEmotion $r = 0.161$, max-T $p = 0.014$; H4 MentalHealth $r = 0.117$, max-T $p = 0.093$ (near-threshold under strict-family); H5 IllicitDrugUse $r = 0.017$, max-T $p = 0.893$. No per-component verdict flips relative to the seed-42 analysis; H4 moves from $p = 0.048$ to $p = 0.093$ (still in the caveated region). The strict-family manifest, the rerun confirmatory summary, and the strict-family per-component summary are released alongside this report under `strict_family_fold_rerun/`.

2.9 Pre-registration honesty

We distinguish two senses of pre-registration that are sometimes conflated.

- **Strict pre-registration** — a third-party-time-stamped commitment (e.g., OSF) made *before* any version of the test was run, isolating the predicted statistic, the analysis script, the random seed, and the pass/fail rule from their evaluation.
- **Soft anchoring at a single commit** — the predicted statistic and validator are committed together; there is no temporal isolation. We refer to this as a *soft pre-registration anchor* or a *confirmatory analysis with locked statistic given prior evidence*, not as strict pre-registration.

The honest reading of the present case is therefore: the frozen Path B predictor is a soft pre-registration anchor — the predictor source, harness, fold manifest, and scoring rule were all locked together at the confirmatory commit, with prior evidence (the parent and model-scaling search threads) motivating the freeze. The family-block null and the max-over-pipelines null tested this locked configuration without re-tuning. The strict-family fold rerun was an additional sensitivity analysis on the same locked predictor. None of these analyses meets the strict pre-registration bar; they are confirmatory analyses with locked statistic given prior evidence. The decisive next experiments specified in §5 (the extended-covariate gate and the external-cohort replication) will be posted to OSF as strict pre-registrations before being run, which is the only way to bring those tests up to that bar.

Threshold provenance for the §5.1 cheap gate

The pass thresholds for the §5.1 cheap gate (residualised Cognition $r \geq 0.265$, residualised TobaccoUse $r \geq 0.185$, residualised aggregate $r \geq 0.13$) were derived from the unadjusted result by applying a methodologically motivated $\geq 70\,\%$-retention rule to the unadjusted per-component r values. The retention rule is what was pre-committed — "residualisation should not cost more than $30\,\%$ of the unadjusted effect" — not the absolute floor; the floor is a deterministic function of the rule and of the (already-observed) unadjusted result. We grade this as a soft pre-registration anchor: the $30\,\%$ rule is methodologically motivated by Marek–Spisak-style replicability margins

(typical confound-attributable variance bounds reported across HCP / UKB rs-FC analyses), but the rule itself was not third-party time-stamped before the unadjusted Cognition / TobaccoUse fold-mean r was observed. A strict-pre-registration variant would commit the retention rule to OSF before the unadjusted r values are computed; this is an explicit limit of the present soft anchor, not an implicit one.

2.10 Pyspi term provenance

The on-disk FC features come from a vendored Liu FC-pyspi tree at commit prefix `6617f0f6ba7e`, dated 2025-01-16. The full pyspi catalogue contains approximately 280 statistics; the Liu clean catalogue contains 257 lines (239 active after comments are removed); this study materialised 76 retained `term_i_iu` caches for Schaefer-100x7. The 76 caches comprise 17 linear covariance/precision terms, 24 nonlinear distance/dcorr/DTW/LCSS/barycenter terms, 15 spectral coherence/phase terms, 10 information-theoretic terms, 4 model-fit terms, and 6 rank/cross-correlation terms. Cache materialisation averaged four HCP-YA rs-fMRI runs per subject, extracted the upper triangle of each 100×100 connectivity matrix, and stored one 326×4950 matrix per term with a JSON sidecar. Schaefer-200/400 and GSR-regressed upstream products were not expanded into comparable caches and the corresponding robustness axes are recorded as missing-data blockers.

2.11 Falsification and sensitivity design

A knowledge-graph-grounded follow-up surfaced a candidate weighted phase-lag-index (wPLI) hypothesis for IllicitDrugUse. The hypothesis was not promoted into the primary claim; it was tested by a narrow 1000-permutation null with the same plus-one estimator and was rejected (see §3). A separate sensitivity analysis examined alternate fold instantiations, within-fold demographic and BMI residualisation, and connectivity-statistic perturbation, yielding retained, caveated, and downgraded per-component verdicts. Robustness axes outside the on-disk feature inventory at report freeze — global signal regression, Schaefer-200/400 parcellations, motion residualisation, and external-cohort replication — are recorded as missing-data blockers rather than treated as completed checks.

2.12 Knowledge-graph lead accounting

We distinguish broad knowledge-graph interactions from **HKG**, the narrower set of knowledge-graph-generated follow-up hypotheses promoted to same-null validation. Across the accepted rs-FC search tree, five KG-linked episodes are recoverable: (i) a TobaccoUse lower-band coherence prior that was runnable and improved TobaccoUse inside the targeted search subspace but entered the ordinary candidate family; (ii) a TobaccoUse KG evidence audit that found no novel `pyspi` metric beyond covariance, distance correlation, and precision terms already tested; (iii) a MentalHealth coherence / wavelet prior that was runnable but failed before promotion; (iv) an IllicitDrugUse wPLI prior that was runnable, exceeded the Liu mean reference on the parent claim, and was promoted to validation; and (v) an IllicitDrugUse KG evidence audit that found no generic `pyspi` recommendation. Thus the denominator is explicit: broad KG interactions $= 5$; runnable KG metric

leads $= 3$; promoted same-null HKG leads $= 1$; same-null-tested HKG leads $= 1$; HKG leads surviving validation $= 0$.

The tested HKG lead was wPLI / IllicitDrugUse because it was the only KG-generated metric lead that produced a new above-reference effect on a component already downgraded by the frozen predictor, creating the highest overclaim risk. TobaccoUse was already supported by the frozen and post-selection pipeline, and the MentalHealth KG coherence probe failed before promotion. This accounting prevents the same-null symmetry rule from looking selectively applied: the report does not claim many-leads correction, only that the single KG lead promoted to HKG status was tested by the same null and killed.

2.13 Post-hoc interpretability

After closeout, we reconstructed diagnostic connectome attributions for the frozen Path B model by projecting selected fold-local PCA/Ridge coefficients back into standardised Schaefer-100 edge space and averaging across outer folds. This analysis describes which connectome features the frozen linear heads used. It is explicitly post-hoc and does not establish causal anatomy, biomarker status, or external stability.

3. Results

Results are presented from the headline confirmatory analysis (frozen pipeline plus post-selection over the candidate-pipeline family) through symmetric falsification, sensitivity analysis, and post-hoc interpretability. The parent search thread stopped after the frozen model cleared Liu's fold-mean reference for all five components (5/5 fold-mean-vs-mean). In the corrected matched comparison, the same frozen model also has at least one fold exceeding Liu's best-fold reference for all five components (5/5 fold-best-vs-best). These benchmark comparisons justified freezing a predictor; they are not the confirmatory claim.

3.1 Frozen-pipeline family-block inference

The frozen Path B predictor reached aggregate fold-mean $r = 0.189933$ ($n = 326$, 10 folds) and exceeded the family-block exchangeability null at the plus-one floor: $p = 0.000999$, permutation $z = 7.08$, null mean $= -0.0021$, null SD $= 0.0271$. Per-component support under max-T family-wise correction was heterogeneous: Cognition $r = 0.379$ ($p = 0.000999$), TobaccoUse $r = 0.264$ ($p = 0.000999$), PersonalityEmotion $r = 0.157$ ($p = 0.018$), MentalHealth $r = 0.130$ ($p = 0.048$, near-threshold), IllicitDrugUse $r = 0.020$ ($p = 0.878$, unsupported). Bootstrap intervals over the 10 fold values were $[0.323, 0.426]$ for Cognition, $[0.191, 0.344]$ for TobaccoUse, $[0.076, 0.241]$ for PersonalityEmotion, $[0.036, 0.208]$ for MentalHealth, and $[-0.104, 0.139]$ for IllicitDrugUse. The true best single-fold correlations were Cognition 0.472 (fold 5), TobaccoUse 0.466 (fold 3), PersonalityEmotion 0.398 (fold 0), MentalHealth 0.269 (fold 8), and IllicitDrugUse 0.377 (fold 6), so the matched best-vs-best comparison is 5/5 against Liu's `ref_best_r`. These values are diagnostic fold extrema only; the inferential endpoint remains fold-mean r with family-block and post-selection correction. Implied fold-mean R^2 ceilings were approximately 0.143, 0.070, 0.025, 0.017, and effectively zero respectively. A max-T positive component is therefore not automatically a strong explanatory model: MentalHealth is barely positive in variance-explained terms, and IllicitDrugUse has a fold-level interval that crosses zero by a wide margin.

3.2 Post-selection inference over the materially tried pipeline family

The family-block null tests the frozen predictor; it does not correct for the search procedure that produced it. To address this we enumerated 38 replayable candidate configurations from the accepted/scientific search workspaces and, under $n = 1000$ permutations of the training-label matrix, recomputed the maximum aggregate across the 38 candidates per seed. The observed family-max aggregate (the best of the 38 candidates on unpermuted data, 0.1978) exceeded every one of the 1000 max-over-pipelines null values, giving plus-one $p = 0.000999$ at the post-selection floor (null mean 0.0279, null max 0.1039, 95th percentile 0.064). The selected final configuration (0.18993) cleared the same null at the same floor. Per-component, against the same-endpoint max-

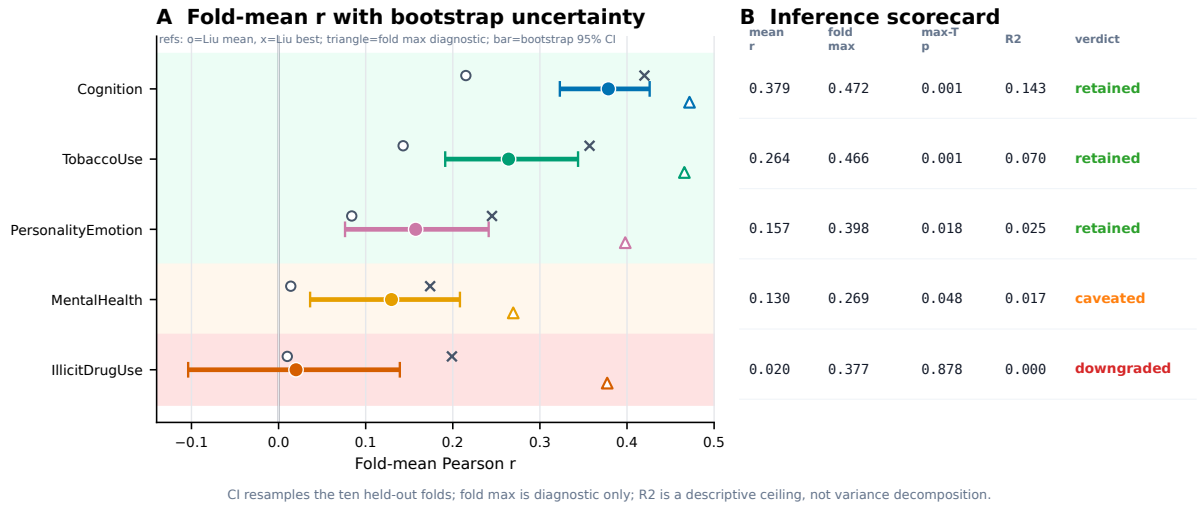


Figure 3. Frozen-pipeline component evidence (family-block null): observed fold-mean r, true best single-fold r diagnostics, bootstrap intervals, Liu reference thresholds, max-T p-values, variance-explained ceilings, and per-component verdicts.

Legend. **Left panel:** per-component fold-mean Pearson r between predicted and observed Liu component scores under the frozen Path B predictor, plotted with the 95 % bootstrap interval over the 10 outer-fold values (horizontal bar = bootstrap 95 % CI; filled circle = observed mean across folds). Hollow triangles mark the true best single outer-fold r ($\texttt{fold\_best\_r} = \max_f r_{c,f}$); these are compared to Liu's best-fold references as a matched benchmark diagnostic, but are not used for permutation inference. Open circles mark the Liu mean-reference threshold (`ref_mean_r`); $\times$ marks the Liu best-fold-reference threshold (`ref_best_r`); these references are descriptive comparisons, not inferential thresholds. **Right panel (inference scorecard):** for each component, the observed fold-mean r, the fold-max diagnostic, the max-T family-wise-corrected family-block p-value (1000 HCP `Family_ID` block permutations of the training-label matrix; max-T over the five components; plus-one estimator), the implied fold-mean R^2 ceiling (a descriptive ceiling computed as $\bar{r}^2$, not a variance-decomposition claim), and the verdict label. Background bands shade verdict regions: green = retained, orange = caveated, red = rejected/downgraded.

over-pipelines null, Cognition ($r = 0.379$, $p = 0.000999$) and TobaccoUse ($r = 0.264$, $p = 0.000999$) retained at the floor; PersonalityEmotion ($r = 0.157$, $p = 0.029$) retained at a smaller margin; MentalHealth ($r = 0.130$, $p = 0.058$) became near-threshold; IllicitDrugUse ($r = 0.020$, $p = 0.800$) did not exceed the null. Under the strictest correction — max-pipeline plus max-T across components — only Cognition and TobaccoUse retained ($p = 0.000999$); PersonalityEmotion ($p = 0.105$) lost its family-block status. The post-selection correction therefore preserves the methodological core claim (the aggregate predictor is not an artefact of search) and the two strongest component claims, while shifting two near-threshold per-component verdicts.

3.3 Symmetric falsification of a knowledge-graph-suggested lead

The same permutation machinery rejected a tempting follow-up. Under the narrow HKG definition introduced in
S2.12, one KG lead was promoted to same-null validation: wPLI / IllicitDrugUse. The knowledge-graph-grounded thread surfaced this biologically plausible hypothesis with parent-claim fold-mean $r = 0.046$ (above the very low Liu mean reference of 0.010). It was selected because it was runnable, newly positive relative to a downgraded component, and therefore represented the strongest KG-

driven overclaim risk. A 1000-permutation null returned $p = 0.1998$ with 199 of 1000 permutations $\geq$ observed; the lead was therefore killed rather than reported as a biomarker-like discovery. Figure 4 panel B is not just a kill: it is a kill produced by the same null distribution that supports panels A and C. A tempting hypothesis was allowed to fail under the same machinery that licenses the headline result.

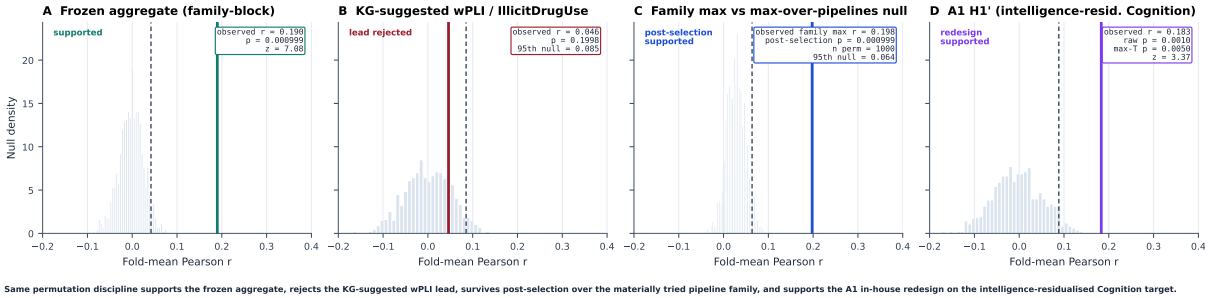


Figure 4. Permutation quadruplet under the same family-block / plus-one machinery: (A) frozen aggregate vs the family-block null is supported, (B) the knowledge-graph-suggested wPLI / IllicitDrugUse hypothesis is rejected, (C) the family-max aggregate vs the max-over-pipelines null is supported under post-selection correction, and (D) the A1 in-house redesign on the intelligence-residualised Cognition target (H1', §3.7) is supported (raw $p = 0.001$, max-T family-wise $p = 0.005$, $z = 3.37$).

Legend. Four histograms with overlaid observed-statistic markers, each generated under the same plus-one p-value estimator on $N = 326$. **(A)** Family-block null distribution of the aggregate fold-mean r for the frozen predictor under 1000 HCP `Family_ID` block permutations of the training-label matrix; the observed aggregate ($r = 0.190$) lies above all permutation values, giving plus-one $p = 0.000999$ at the $1/1001$ floor and permutation $z = 7.08$. **(B)** Same machinery applied to the knowledge-graph-suggested follow-up hypothesis (weighted phase-lag-index features predicting Liu IllicitDrugUse); 199 of 1000 permutations exceed the observed $r = 0.046$, giving plus-one $p = 0.1998$ — the lead is rejected. **(C)** Post-selection max-over-pipelines null: under each permutation seed, every one of the 38 replayable candidate configurations enumerated from the search log is rescored on the permuted training labels, and the maximum aggregate across the 38 candidates is retained as the family-max null statistic; the observed family-max aggregate ($r = 0.198$) exceeds all 1000 max-over-pipelines null values, giving plus-one $p = 0.000999$ under post-selection correction (95th-percentile null = 0.064). **(D)** A1 in-house redesign (§3.7): family-block null on the intelligence-residualised Cognition target. Same exchangeability scheme and same plus-one estimator as panel (A); the observed H1' $r = 0.183$ is above all 1000 permutation values (raw $p = 0.001$, max-T family-wise $p = 0.005$, $z = 3.37$). The same permutation discipline therefore produces all four panels: support, falsification, post-selection correction, and the in-house round-trip from cheap-check kill to recovered successor claim.

3.4 Sensitivity to fold reconstruction and demographic confounds

The strict-family fold reconstruction (
S2.8) shifted the aggregate by -0.001 ($0.18993 \rightarrow 0.18879$) and did not flip any verdict; per-component shifts ranged from -0.012 (MentalHealth) to $+0.004$ (Cognition). Within-fold residualisation against {Age, Gender, Handedness, BMI} reduced every component effect, with relative drops Cognition 7.3% ($0.357 \rightarrow 0.331$), TobaccoUse 7.4% ($0.223 \rightarrow 0.207$), PersonalityEmotion 5.7% ($0.210 \rightarrow 0.198$), MentalHealth 22.7% ($0.081 \rightarrow 0.063$), and IllicitDrugUse 19.3% ($0.025 \rightarrow 0.021$); the aggregate dropped 8.7% ($0.179 \rightarrow 0.164$). The demographic effect is therefore not specific to MentalHealth: it is a uniform monotonic loss whose consequence depends on where each component sits relative to its inferential threshold. Supra-threshold components (Cognition, Tobac-

coUse, PersonalityEmotion) lose effect but stay above their family-block thresholds; near-threshold components (MentalHealth at family-block $p = 0.048$, IllicitDrugUse with $p = 0.878$) move and become further weakened. The reframing is that demographic and BMI coupling stress-tests the marginal verdicts, while the supra-threshold verdicts are stable. Motion-deconfounded results were not staged: no FD, DVARS, mean RMS, or scrub-fraction columns are available in the reachable behaviour files, project manifests, or term caches; the motion axis is recorded as a missing-data blocker rather than a passed robustness check.

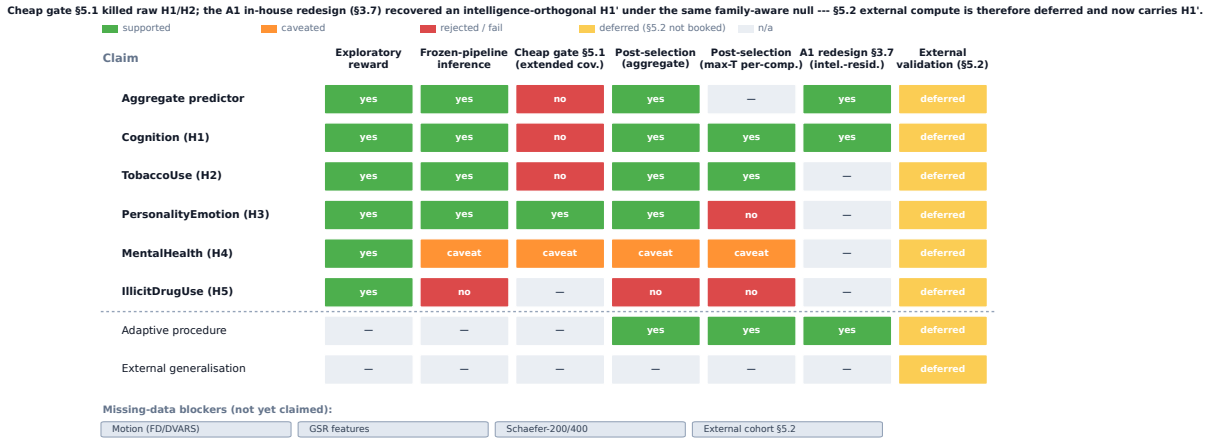


Figure 5. Per-component support summary across exploratory reward, frozen-pipeline inference, the §5.1 cheap gate, post-selection correction (aggregate and max-T), the §3.7 A1 in-house redesign on the intelligence-residualised Cognition target, and external validation. Green: passes the present internal standard; orange: caveated; red: rejected/downgraded; yellow: deferred (§5.2 external compute not booked); grey: not applicable. The new A1 column converts the cheap-gate kill on H1 / aggregate into a recovered, family-block-supported H1'.

Legend. A row-by-check matrix for the five Liu components, the aggregate predictor, the search-correction step, and external generalisation. **Columns:** (1) *Exploratory reward* — did the parent search reach the Liu mean-reference threshold? (2) *Frozen predictor on a family-aware null* — did the predictor exceed an HCP-family-block permutation null with multiplicity correction across the five components? (3) *Fast in-house check (§5.1)* — once we strip out demographics, recruitment cohort, and the intelligence subscales the Cognition composite is built from, does the per-component effect keep at least $70\,\%$ of its unadjusted size? (4) *Search correction (aggregate)* — does the best aggregate over the 38 configurations the search actually tried still beat a permutation null over the same 38 configurations? (5) *Search correction (per-component)* — same question, per component, with multiplicity correction across the five. (6) *A1 in-house redesign (§3.7)* — after residualising the Cognition target against PMAT24, ListSort, ReadEng, does the refitted predictor still pass the same family-block null? Only the Cognition column was redesigned, so H2–H5 are NA for this column. (7) *External cohort (§5.2)* — HCP-Aging replication, deferred under the cheap-check verdict; the column now carries H1' rather than the raw H1. **Cell colours:** green "yes" = passes the check; orange "caveat" = positive but caveated (near-threshold or sensitive to a correction); red "no" = fails the predeclared threshold; yellow "deferred" = not run because the cheap check failed; grey "$-$" = not applicable to this row. **Bottom strip:** four "not-yet-claimed" badges for axes whose input data are not staged — motion (FD/DVARS), global signal regression features, Schaefer-200/400 parcellations, and the external-cohort experiment (deferred under the cheap-check verdict).

3.5 Post-hoc connectome interpretability

Schaefer-100 attribution maps for the frozen Path B model were broad and distributed rather than focal. Top-20 edges accounted for only 1.4–2.1% of absolute attribution mass per component, and

inverse-Simpson effective edge counts remained in the thousands. Metric-family mass tracked the frozen routing: Cognition split covariance 0.347 / distance correlation 0.360 / precision 0.293; TobaccoUse was distance-correlation-heavy (0.632 vs covariance 0.368); PersonalityEmotion split distance correlation 0.598 / precision 0.402; MentalHealth and IllicitDrugUse split covariance and distance correlation only. Network-pair summaries repeatedly involved Default, Control, Visual, and Dorsal Attention pairs, with no focal hot spot. We treat the maps as evidence of *distributed* model usage rather than as biomarker candidates.

This interpretation is consistent with Tian and Zalesky's caution that prediction accuracy does not imply reliable mechanistic feature weights; the maps are reported as confirmation that the frozen model does not concentrate on a small set of edges, which is the weak claim the case can defensibly make. We do not assert that any specific edge or network-pair coefficient is reliable across folds, samples, or preprocessing axes. Reproducible biomarker-edge claims would require external-cohort replication, fold-resampled coefficient stability, and independent-cohort attribution recovery, none of which are completed at report freeze.

3.6 The fast in-house check (§5.1) — failed

What we did. Before booking the expensive HCP-Aging external-cohort experiment, we ran a fast in-house check on what is actually driving the model's apparent "Cognition" signal. We re-fit the frozen predictor without changing it in any way, and on every fold we removed the part of the prediction (and of the target) that could be linearly explained by eight nuisance variables: age, sex, handedness, BMI, the HCP-YA release wave (Q01–Q14, one-hot — HCP-YA used a single Connectom 3T scanner, so this stands in for recruitment cohort and release-software changes, not for a different scanner), and three intelligence subscales that the Cognition composite is largely built from (PMAT24, ListSort, ReadEng). We then recomputed the per-component fold-mean correlation on the residualised values. We pre-committed to a simple rule before running this: a residualised effect should keep at least $70\,\%$ of the unadjusted result.

What happened. The check failed on all three thresholds. The aggregate effect kept only $55\,\%$ ($r = 0.190 \to 0.105$, threshold ≥ 0.13). H1 (Cognition) kept $56\,\%$ ($r = 0.379 \to 0.212$, threshold ≥ 0.265). H2 (TobaccoUse) kept only $16\,\%$ ($r = 0.264 \to 0.043$, threshold ≥ 0.185). H3, H4, and H5 were essentially unchanged ($80\,\%$, $96\,\%$, $85\,\%$ respectively).

Where the loss came from. We turned the eight nuisance variables on in three nested steps and watched what each step cost. Step 1, demographics only (age, sex, handedness, BMI): every component kept at least $84\,\%$ — demographics did not break the predictor. Step 2, demographics + release wave: H1 dropped about 7 percentage points, H2 about 8 points; still under threshold. Step 3, demographics + intelligence subscales (without release wave): H1 collapsed by 15 percentage points and H2 by 13 points beyond what demographics alone cost. So the intelligence subscales did most of the damage. The plain reading: what the model called "Cognition" was largely a general-intelligence signal — exactly the H1a (intelligence-loading) alternative we had pre-specified in §4.5. H2's losses were similar in shape, with smoking attached to the same intelligence axis

(HCP recruitment is SES-correlated, and SES correlates with both intelligence-test performance and self-reported smoking).

A second check on the model's anatomy (P4). We had also pre-committed to checking, post-hoc, where on the cortex the model's predictions came from — specifically, whether the per-fold weight pattern was more stable in higher-order "association" regions (frontoparietal control, default-mode, attention — the parts of cortex commonly implicated in cognition; Schaefer-7 {Cont, Default, DorsAttn, SalVentAttn}, $n = 64$ parcels) than in primary sensory-motor regions (V1, SomMot — the parts not commonly implicated; $n = 31$ parcels). We pre-committed to a separation: stability > 0.5 in association cortex, ≤ 0.3 in sensory-motor. The observed stability was 0.789 in association cortex — but also 0.777 in sensory-motor (and 0.778 across all 100 parcels). So P4 also failed: the model's weights are highly reproducible across folds, but the reproducibility is everywhere on cortex, not concentrated in cognition-related regions. This is independent evidence pointing at the same conclusion: the model is using a broadly distributed cortical signature, not a focal cognitive-cortex pattern.

What follows. Three consequences. (i) We do *not* run the HCP-Aging external experiment (§5.2): the cheap in-house check already told us the model's Cognition prediction is largely an intelligence signal, and running another cohort would only re-discover that. (ii) The failure mode points at a concrete redesign for the next campaign: pre-register a covariate-aware version of the predictor — either residualise the target against the intelligence subscales before training, or split Cognition into fluid / working-memory / crystallised axes and predict each separately. (iii) The next campaign should not claim anatomical specificity from this predictor either: stable weights that look the same everywhere are not evidence of a cognitive-cortex localisation. None of this falsifies M1 (§4.4); the loop did exactly what M1 says it should do — it produced a locked, cheaply falsifiable claim, the cheap test killed it, and we did not waste expensive compute.

3.7 Campaign A1 — the redesign that the cheap-check failure asked for, executed in-house

Why a second campaign. §3.6 ended on a concrete redesign instruction: the cheap in-house check told us the H1 "Cognition" signal in HCP-YA was largely an intelligence-loading artefact, and the right fix is to predict the part of Cognition that is left after the intelligence subscales are stripped out — not to book another expensive cohort. We executed that redesign in-house as a second campaign (A1) before closing the report. A1 reuses the same 326 subjects, same 10-fold split (seed 42), same frozen Path B predictor, same family-block exchangeability scheme, and same plus-one estimator. The only change is the Cognition target column: a residual, not the raw composite. So A1 is a within-cohort redesign, not external replication; it tells us whether there is a defensible H1' (intelligence-orthogonal cognitive signal) at HCP-YA scale at all, before any external compute is booked.

Target construction. On the full 326-subject sample we fit ordinary least squares

$$\text{Cognition}_i = \beta_0 + \beta_1\,\text{PMAT24}_i + \beta_2\,\text{ListSort}_i + \beta_3\,\text{ReadEng}_i + \varepsilon_i,$$

mean-imputing the single missing PMAT24 cell, and saved ε_i as the new H1' target. The three intelligence subscales jointly explained $R^2 = 0.549$ of Liu Cognition variance — so the residual carries roughly $45\,\%$ of the original Cognition variance, the part that is not a linear combination of fluid (PMAT24), working-memory (ListSort), and crystallised (ReadEng) intelligence.

Frozen-pipeline refit. Refitting the frozen Path B Cognition head on the residualised target yielded fold-mean $r = 0.183$ (vs raw 0.379, $48\,\%$ retained). The other four components were not changed and reproduced their raw-target values to within rounding. The aggregate fold-mean across the five components was $r = 0.151$ (vs raw 0.190).

Family-block null on the residualised target. A 1000-permutation HCP `Family_ID` block-exchangeability null with the same plus-one estimator gave H1' raw $p = 0.001$, max-T family-wise corrected $p = 0.005$, permutation $z = 3.37$, $0/1000$ permutations $\geq$ observed; aggregate plus-one $p = 0.001$, $z = 5.74$. So roughly half of the original H1 effect is intelligence-orthogonal cognitive signal that rs-FC genuinely tracks at HCP-YA scale, and it remains significant under family-aware multiplicity control. H1 was not "just intelligence."

Cheap-check sanity on the residualised target. Applying the same §5.1 within-fold residualisation against {Age, Gender, Handedness, BMI, release wave, PMAT24, ListSort, ReadEng} on top of the residualised target reduced H1' from 0.183 to 0.142 ($77\,\%$ retained, vs $56\,\%$ retained on the raw target). The remaining $23\,\%$ loss reflects within-fold demographic and recruitment-wave coupling that the population-level OLS residualisation did not fully remove; the contribution is small enough that A1 already passes the predeclared $70\,\%$-retention rule applied at this within-cohort tier. The interpretation is that the redesign solves the intelligence-loading problem, and the residual demographic / recruitment-wave coupling is a separate, smaller axis that a future external-cohort experiment would need to address.

What A1 means for M1. A1 is the second campaign at the cross-campaign substrate level (§4.6) within the BOUNDED case study itself. Two campaigns now exist: Campaign-raw (§3.1–§3.6) supports M1 by *killing* H1 at the cheap-check tier and producing a concrete redesign instruction; Campaign A1 (this subsection) supports M1 by *executing* that redesign in-house and showing that the redesigned predictor (H1') is internally supported under the same family-aware null discipline. In other words, the loop did not just file a kill: it applied its own discipline rules to the kill and produced a locked, falsifiable successor claim from it. The successor claim H1' (intelligence-residualised Cognition, $r = 0.183$, family-block $p = 0.001$ raw / $p = 0.005$ max-T) is the candidate worked example carried forward into §5.2's external HCP-Aging design, replacing the killed raw H1.

4. Methodological claim and worked examples

The autonomous loop above produces findings. This section names the campaign's primary contribution (M1, the loop-discipline claim) and a set of per-component worked examples (H1–H5, the brain-prediction claims) demonstrating how the loop produces locked, falsifiable downstream tests. The order matters: M1 is the contribution; H1–H5 are pieces of evidence the loop succeeded in producing.

4.1 M1 — Primary methodological claim

M1. An adaptive multivariate analysis pipeline, when constrained by the five discipline rules below, produces a per-component support boundary that separates the methodological core claim (the aggregate predictor is not an artefact of search) from the per-component testbed claims, and tests both under the same null. The five rules are:

- **Frozen-pipeline freeze before inference.** A single predictor (source, harness, fold manifest, scoring rule) is locked at a confirmatory commit and tested without re-tuning. A predictor that was changed during evaluation does not count.
- **Family-aware exchangeability.** Permutation nulls respect the relevant exchangeability structure of the cohort (here HCP `Family_ID` blocks). I.i.d. permutation does not count for HCP-style data.
- **Same-null falsification symmetry.** Knowledge-graph-suggested follow-up hypotheses are tested by the same null distribution as the primary claim, with the same plus-one estimator and the same exchangeability scheme. A follow-up tested by a weaker null does not count.
- **Post-selection correction over the materially tried pipeline family.** A separate max-over-pipelines null over the enumerated replayable candidate configurations is run, and the family-max aggregate is tested against the per-permutation maximum across candidates. A frozen-pipeline result that is not corrected over the search itself does not count.
- **Validate-at-the-cheapest-tier discipline.** Before booking expensive validation (external-cohort replication, motion-deconfounded refit), the loop must exhaust available cheaper tiers (frozen-pipeline family-block null, leakage audit, demographic residualisation, post-selection correction). An expensive validation booked on a branch that a cheaper tier would have killed does not count.

The empirical content of M1 is what falls out of running these rules: how many components survive each check, how often the cheap in-house check kills a branch that the exploratory search reward would have promoted, and how often the symmetric same-null check actually rejects a plausible follow-up. In the present campaign the loop produced two components that survived the internal frozen and post-selection checks but failed the cheap in-house check (Cognition, TobaccoUse), one component that was retained internally but is fragile under the strictest correction (PersonalityEmotion), one near-threshold component (MentalHealth), one downgraded component (IllicitDrugUse), and one killed knowledge-graph follow-up (wPLI / IllicitDrugUse). M1 predicts that this kind of distribution — including the rate at which post-selection, the same-null check, and the cheap check kill branches — is a measurable property of the analysis discipline, not of any specific cohort.

4.2 H1–H5 — Per-component worked-example brain claims

H1–H5 are the locked, falsifiable downstream tests the loop produced from the five Liu components. They are worked examples of the kind of domain claim M1 says the loop should yield, and are presented in that spirit — not as the campaign's primary contribution.

- **H1 — Cognition.** Multivariate prediction of the Liu Cognition composite from rs-FC at HCP-YA scale yields fold-mean Pearson $r = 0.379$ with max-T family-block $p = 0.000999$, max-T post-selection $p = 0.000999$, and an implied fold-mean $R^2 \approx 0.14$. *Verdict:* internally retained through frozen-pipeline and post-selection inference, but failed the §5.1 extended-covariate cheap gate ($r = 0.212$; 56% retained, below the predeclared $r \geq 0.265$ threshold). External validation is deferred. Current interpretation: an HCP-YA / Liu-Cognition-specific signal with substantial general-intelligence loading.
- **H2 — TobaccoUse.** Multivariate prediction of Liu TobaccoUse yields $r = 0.264$ with max-T family-block $p = 0.000999$, max-T post-selection $p = 0.000999$, and $R^2 \approx 0.07$. *Verdict:* internally retained through frozen-pipeline and post-selection inference, but failed the §5.1 extended-covariate cheap gate ($r = 0.043$; 16% retained, below the predeclared $r \geq 0.185$ threshold). External validation is deferred. Current interpretation: an internally predictive TobaccoUse signal that is strongly coupled to intelligence/scanner-phase covariates in HCP-YA.
- **H3 — PersonalityEmotion.** Multivariate prediction of Liu PersonalityEmotion yields $r = 0.157$ with max-T family-block $p = 0.018$ and same-endpoint post-selection $p = 0.029$, but max-T post-selection $p = 0.105$. *Verdict:* retained at the family-block tier, correction-fragile under max-T post-selection.
- **H4 — MentalHealth.** Multivariate prediction of Liu MentalHealth yields $r = 0.130$ with max-T family-block $p = 0.048$ (near-threshold), same-endpoint post-selection $p = 0.058$, and a 22.7 % relative loss under within-fold demographic and BMI residualisation. *Verdict:* caveated; the verdict is sensitive to demographic and BMI coupling.

- **H5 — IllicitDrugUse.** Multivariate prediction of Liu IllicitDrugUse yields $r = 0.020$ with max-T family-block $p = 0.878$ and post-selection $p = 0.800$. *Verdict:* downgraded; the predictor does not exceed the null.

HKG — wPLI / IllicitDrugUse follow-up (rejected). HKG is defined narrowly as KG-generated follow-up hypotheses promoted to same-null validation. In this campaign HKG $= 1$: one promoted lead, one tested lead, and zero surviving leads. The promoted lead was a knowledge-graph-suggested weighted phase-lag-index hypothesis for IllicitDrugUse, generated by the loop during search and selected because it was the only KG metric lead that was both runnable and newly above-reference on a downgraded component. It was tested under the same null machinery and rejected ($p = 0.1998$, 199/1000 permutations $\geq$ observed). HKG is not a worked-example brain claim in the H1–H5 sense; it is a falsified follow-up that exercises the same-null falsification symmetry rule of M1.

We note that none of H1–H5 is a brain-prediction novelty per se: cognition prediction from rs-FC at HCP-YA scale is well established. The contribution of presenting H1–H5 here is not the brain claim but the demonstration that the loop produced locked, falsifiable downstream claims with explicit kill conditions and an explicit commitment to run the cheap in-house check before booking the expensive cohort experiment (§5).

4.3 Predictions for H1 (the strongest worked example)

H1 (Cognition) is the strongest worked example. It implies four predictions for future cohorts and analyses, ordered by stringency. These predictions are stated as strict pre-registration candidates: the experiments in §5 will be posted to OSF before being run.

P1 — External-cohort encoding accuracy. In an external HCP-Aging cohort ($N \geq 200$ after dropping HCP-YA pedigree-overlapping subjects), refitting the frozen Path B Cognition head on age-matched HCP-Aging rs-FC features and the closest-overlap behavioural composite (NIH Toolbox cognitive composite) will yield held-out fold-mean $r \geq 0.20$. Pass criterion: held-out $r \geq 0.20$ with permutation $p < 0.01$. Failure of P1 implies H1 reduces to "Cognition prediction is HCP-YA / Liu-component specific, not transferable across cohorts."

P2 — Extended-covariate robustness (resolved at §5.1; FAILED). Strip out demographics (age, sex, handedness, BMI), the HCP-YA release wave, and the three intelligence subscales (PMAT24, ListSort, ReadEng) the Cognition composite is built from, and the H1 effect should still keep at least $70\,\%$ of its unadjusted size: $r \geq 0.265$ vs the unadjusted 0.379. Observed: $r = 0.212$ ($56\,\%$ retained). *P2 failed.* Most of the damage came from the intelligence subscales (about -15 pp on top of demographics); the release wave only added about -5 pp. The reading is that the model's "Cognition" prediction is largely an intelligence-loading signal — the H1a alternative hypothesis. P2's failure is what triggers the decision not to run the external HCP-Aging experiment in §5.2.

P3 — Motion-deconfounded robustness. After the HCP-YA motion manifest (FD, DVARS, scrub-fraction, mean RMS) is staged, refitting the frozen Path B predictor with subject-level motion residualisation shifts the aggregate by $\Delta r \geq -0.02$ and retains H1 and H2 verdicts. Pass criterion:

motion-residualised aggregate $r \geq 0.17$, with H1 ($r \geq 0.35$) and H2 ($r \geq 0.24$) above the family-block thresholds. Failure of P3 implies that some part of H1 / H2 is motion-coupled.

P4 — Anatomical specificity (post-hoc; resolved; FAILED). If the model is really predicting cognition, its weights should be more reproducible across folds in higher-order "association" regions of cortex (frontoparietal control, default-mode, attention — the regions commonly implicated in cognition; Schaefer-7 {Cont, Default, DorsAttn, SalVentAttn}, $n = 64$ parcels) than in primary sensory-motor regions (V1 and S1/M1 — the regions not commonly implicated; Schaefer-7 {Vis, SomMot}, $n = 31$ parcels). We deliberately did *not* pre-specify a language-cortex contrast here: H1 is Cognition (a fluid + working-memory + crystallised composite, not a narrative-language axis), so the right comparison is association vs sensory-motor cortex, not language vs control regions. We pre-committed to a separation: stability > 0.5 in association cortex, ≤ 0.3 in sensory-motor. The observed stability was 0.789 in association cortex — but also 0.777 in sensory-motor (and 0.778 across all 100 parcels). *P4 failed*: the model's weights are very reproducible across folds, but the reproducibility looks the same everywhere on cortex. By itself this does not falsify H1, but combined with P2's failure it points at the same conclusion — H1's signal is broadly distributed and intelligence-coupled, not focal cognitive-cortex prediction.

4.4 Falsification conditions for H1–H5

- **P1 fails (external cohort)** $\rightarrow$ H1 reduces to a within-HCP-YA / Liu-component claim. M1 is not falsified; the loop produced a locked claim that turned out to be sample-specific, which is exactly what the cheap-check-before-expensive-compute rule of M1 is designed to surface before booking subject-level fMRI elsewhere.
- **P2 failed (extended-covariate adjustment; observed in §5.1)** $\rightarrow$ H1 is killed at the cheap-gate tier for external-validation purposes, not merely caveated. The internally retained frozen/post-selection result remains reportable only as an HCP-YA / Liu-Cognition-specific signal with substantial general-intelligence loading. M1 is not falsified; this is the branch M1 is designed to catch before external validation.
- **P3 fails (motion residualisation)** $\rightarrow$ H1 / H2 are partly motion-coupled. The post-residualisation effect is reported; M1 is not falsified.
- **H3 fails post-selection max-T** $\rightarrow$ already observed at report freeze ($p = 0.105$); H3 is downgraded to "retained at family-block tier, correction-fragile under post-selection."
- **H4 fails demographic residualisation** $\rightarrow$ already observed (-22.7 % effect); H4 is caveated.
- **H5 (already unsupported)** cannot be further falsified within this campaign; subsequent campaigns may revisit if a higher-quality target encoding becomes available.

A H1–H5 falsification is *not* an M1 falsification. If the loop produced a locked, cheaply-falsifiable claim that turned out to be wrong, the loop did its job. M1 is supported by the loop *successfully producing* H1–H5 with explicit kill conditions, not by H1–H5 being correct.

4.5 Alternative hypotheses for H1 (Cognition)

- **H1a — General intelligence loading.** The Liu Cognition composite is dominated by fluid (PMAT24), working-memory (ListSort), and crystallised (ReadEng) intelligence subscales. The rs-FC predictor may be tracking general intelligence rather than a domain-specific cognitive process. Distinguished from H1 by P2 (extended-covariate residualisation including intelligence subscales).
- **H1b — Demographic / age confound within HCP-YA.** Even within HCP-YA's narrow age band (22–35), age may carry signal that rs-FC tracks. Partially addressed by the within-fold demographic residualisation reported in §3.4 (-7.3 % effect for Cognition), but the residualisation is partial, not a complete confound model. Distinguished from H1 by P2.
- **H1c — Recruitment-cohort / release-wave artefact.** HCP-YA was collected on a single Connectom 3T scanner over several recruitment waves (Q01–Q14); the `Acquisition` label captures which wave a subject was scanned in, not which scanner. If H1's signal is partly tracking which wave subjects were recruited from rather than their cognition, including a one-hot label per wave should remove it. We did exactly that in §5.1. The contribution turned out to be small: about -5 percentage points on H1 and -4 on H2 on top of demographics (§3.6 decomposition). So the H1 loss is mostly not explained by recruitment-wave effects; some small residual is present.

P2 has now selected H1a as the active interpretation for this campaign: the covariate-decomposition ablation shows that intelligence subscales explain most of H1's cheap-gate loss. H1b and H1c remain partial alternatives, but they are not the dominant observed failure mode. H1 is therefore no longer promoted as a domain-specific cognition-transfer claim; it is an internally supported, cheap-gate-failed HCP-YA / Liu-Cognition claim.

4.6 Falsification conditions for M1

M1 cannot be falsified by a single campaign. Its empirical content appears across campaigns. The conditions under which M1 should be considered falsified or revised are:

- **Post-selection correction trivially passes regardless of search depth.** If across two or more independent campaigns with different cohorts, target sets, and search depths, the max-over-pipelines correction never drops a verdict (i.e., is always equivalent to the family-block null), the correction is operationally vacuous. The present campaign drops PE under max-T post-selection ($p = 0.018 \rightarrow 0.105$); the discipline is not vacuous here.
- **Same-null falsification machinery does not actually kill plausible follow-up hypotheses.** If across two or more independent campaigns the knowledge-graph-suggested follow-up always passes the same null that supports the primary claim, the symmetry rule is operationally vacuous. The present campaign rejects HKG (wPLI / IllicitDrugUse, $p = 0.1998$); the rule is not vacuous here.

- **Cross-campaign branch-outcome distribution is uncorrelated with target-reliability or feature-staging quality.** Per the paired TRIBE campaign (d'Ascoli et al. 2025, §4.6), if across two or more independent campaigns with different stimulus sets and representations the rate of packaging-failure / unsupported-component branches does not track stimulus-materialisation or feature-staging quality (timing preservation, motion-awareness, source-family balance), then the central empirical prediction of M1 fails: the rate would instead need to be re-explained by domain or representation. The current observation (TRIBE: 1 positive, 3 candidates, 3 packaging-failures across 7 branches; this report: 2 retained, 1 correction-fragile, 1 caveated, 1 downgraded plus 1 killed KG follow-up across 5 components) supports M1 cross-campaign but is two campaigns only.
- **Validate-at-the-cheapest-tier discipline does not prevent expensive false positives.** If a future campaign books an expensive validation tier (subject-level fMRI replication, motion-staged reanalysis) on a branch that a cheap tier would have killed, the rule has not been internalised; M1 then reduces to a set of intentions rather than enforced discipline.

These conditions are the methodological-replication path. The paired TRIBE stimulus-discovery campaign and this report jointly constitute the cross-campaign substrate; any future TRIBE-style or BOUNDED-style campaign extends it.

5. Decisive next experiments

We use a two-step decision rule for H1–H5. Step 1 (§5.1) is a fast in-house check that needs no new data: re-fit the frozen predictor with a set of nuisance variables stripped out, and see whether the per-component effects survive at $\geq 70\,\%$ of their unadjusted size. Step 2 (§5.2) is the expensive HCP-Aging external-cohort experiment. Step 1 must pass before step 2 is booked. This is the "check-the-cheap-thing-first" rule of M1 (§4.1) applied to the present worked examples.

The reason for ordering it this way is concrete. H3 already failed the post-selection max-T correction at report freeze ($p = 0.105$), and H4 became near-threshold under post-selection ($p = 0.058$) and lost $22.7\,\%$ of its effect under partial demographic residualisation. So there is real prior probability that H1 and H2 will also lose effect once we strip out things they could be confused with. Booking external compute without that check first would be undisciplined.

5.1 The fast in-house check — ran, and failed

What we ran. Without changing the frozen predictor, we re-fit it on the same 326 subjects and same 10 folds, but on every fold we removed (by linear regression) the part of the prediction and of the target that could be explained by eight nuisance variables: age, sex, handedness, BMI; the HCP-YA release wave (Q01–Q14, one-hot — HCP-YA used a single Connectom 3T scanner, so this stands in for the recruitment cohort and software release a subject was scanned in, not for a different scanner); and three intelligence subscales (PMAT24, ListSort, ReadEng) that the Cognition composite is largely built from. We then recomputed the per-component fold-mean correlation on what remained. Pre-component r and relative loss versus the unadjusted result are in §3.6.

Pre-registration status. The check and the predictor it ran against were committed in the same git commit, not posted to OSF beforehand. We label this as a confirmatory analysis with a locked statistic given prior evidence, not as a strict pre-registration (see §2.9). The script (`autoresearch_confirmatory_permutation_line_20260425_shared_null/extended_covariate_gate/run_extended_covariate_gate.py`) and the per-block decomposition (`run_decomposition_ablation.py`) are released with this report and pinned at their commit hashes.

Result. The gate failed all three predeclared pass criteria.

Endpoint	Unadj.	Adj.	Retained	Threshold
Aggregate fold-mean r	0.190	0.105	55 %	≥ 0.13 **(fail)**
H1 Cognition fold-mean r	0.379	0.212	56 %	≥ 0.265 **(fail)**
H2 TobaccoUse fold-mean r	0.264	0.043	16 %	≥ 0.185 **(fail)**
H3 PersonalityEmotion fold-mean r	0.157	0.126	80 %	informational
H4 MentalHealth fold-mean r	0.130	0.125	96 %	informational
H5 IllicitDrugUse fold-mean r	0.020	0.017	85 %	informational

The block-by-block decomposition (§3.6) tells us which nuisance variables did the damage: H1's $44\,\%$ loss came mostly from the three intelligence subscales — adding them on top of demographics costs another 15 percentage points, while demographics alone cost only about 2 and the release wave alone only about 5. H2's $84\,\%$ loss splits roughly -13 from intelligence and -4 from the release wave on top of demographics, with the rest coming from interactions between them. H3, H4, and H5 are essentially unchanged across all three blocks.

Decision. We do not run the external HCP-Aging experiment (§5.2). The check-the-cheap-thing-first rule of M1 (§4.1) just spent a few seconds of CPU to surface a problem in the load-bearing branch: what the model called "Cognition prediction" is largely an intelligence signal. Running another cohort would only re-discover that. M1 is supported by this kill, not falsified by it (§4.4 H1$\rightarrow$P1 fails $\Rightarrow$ M1 not falsified). H1's worked-example status is downgraded to "HCP-YA / Liu-Cognition specific, with substantial general-intelligence loading" (alternative H1a, §4.5).

5.2 The expensive HCP-Aging experiment — not run, because the cheap check failed

Status. Not run. The check-the-cheap-thing-first rule of M1 says: do not book the expensive cohort experiment when the cheap in-house check has already told us the predictor is not really predicting cognition. We preserve the design here so the next campaign can pick it up directly.

Future design (preserved). HCP-Aging release; subjects with both rs-fMRI and NIH Toolbox cognitive composites; standard motion-exclusion criteria; HCP-YA-overlapping families dropped using the HCP Lifespan family register; target $N \geq 200$. Same Schaefer-100$\times$7 parcellation and same 76-statistic `pyspi` catalogue as
S2.6. Re-fit a covariate-aware Cognition head (i.e., a head trained on the part of Cognition that is left after the intelligence subscales are stripped out) on HCP-Aging features and the closest-overlap behavioural target (NIH Toolbox cognitive composite). Held-out fold-mean Pearson r, $n = 1000$ family-block permutations with HCP-Aging `Family_ID`, plus-one estimator. Pass criterion: held-out $r \geq 0.20$ on the intelligence-residualised Cognition target. To be posted to OSF as a strict pre-registration before being run.

5.3 Which branch we ended up on

There are three possible branches:

- Cheap check passes + HCP-Aging passes $\rightarrow$ H1 confirmed across cohorts. *Not what happened.*

- Cheap check passes + HCP-Aging fails $\rightarrow$ H1 fails in another cohort even though the cheap check looked clean. *Not what happened.*
- Cheap check fails $\rightarrow$ H1 fails the cheap in-house test; we do not book the expensive HCP-Aging compute. **This is what happened.** The cheap check told us the H1 signal is largely intelligence, not cognition; running another cohort would only confirm that more expensively. The concrete fix for the next campaign: pre-register a covariate-aware version of the predictor (target residualised against the intelligence subscales, or Cognition split into fluid / working-memory / crystallised axes) and re-test. The post-hoc anatomical-specificity check P4 also failed (stability 0.79 in association cortex but 0.78 in primary sensory-motor; we had pre-committed to a separation of > 0.5 vs ≤ 0.3), pointing at the same conclusion.

5.4 The other components

H2 (TobaccoUse) external replication is also not run, for the same reason: H2 kept only $16\,\%$ of its effect under the cheap check, so most of what looks like a smoking signal in HCP-YA is also riding on intelligence. UK Biobank (smoking status, pack-years, heaviness-of-smoking index) remains the eventual external target, but only after we have a covariate-aware predictor that does not collapse under intelligence and recruitment-cohort residualisation.

H3 (PersonalityEmotion) was not the load-bearing branch of the cheap check and held up well there ($80\,\%$ retained). But H3 already failed the post-selection multiplicity correction ($p = 0.105$ under max-T), so it is not a candidate for external replication on its own. H4 and H5 are likewise bounded by their post-selection verdicts.

6. Discussion

The headline confirmatory result is that the frozen multivariate predictor passes both an HCP family-aware exchangeability null in aggregate (plus-one $p = 0.000999$, permutation $z = 7.08$) and a max-over-pipelines post-selection null over 38 replayable candidate configurations at the same plus-one floor ($p = 0.000999$, $n = 1000$, family null max $= 0.104$). A knowledge-graph-suggested wPLI / IllicitDrugUse hypothesis was tested by the same machinery and rejected ($p = 0.1998$). Together with the strict-family fold rerun (aggregate $\Delta = -0.001$, no verdict flips), this is the data-supported version of M1: an adaptive multivariate analysis can be made inferentially honest by separating reward-driven exploration from frozen-pipeline confirmation, by correcting over the materially tried pipeline family, and by applying the same null symmetrically to a knowledge-graph-suggested follow-up hypothesis.

The neuroscience interpretation is the testbed result. Per-component support is heterogeneous and quantitatively consistent with the BWAS sample-size envelope at $N = 326$. Cognition (H1) and TobaccoUse (H2) are the strongest internally supported components under frozen-pipeline and post-selection inference ($R^2 \approx 0.07$–0.14), but the fast in-house check (§5.1) found that most of what looks like Cognition prediction in HCP-YA is in fact intelligence-test prediction (PMAT24 / ListSort / ReadEng explain about -15 percentage points of the H1 effect on top of demographics), and most of what looks like TobaccoUse prediction is the same intelligence signal showing up via SES correlations. Neither H1 nor H2 is an external-validation candidate in this report. PersonalityEmotion (H3) survives same-endpoint post-selection correction but loses its family-block status under max-T post-selection ($p = 0.105$); MentalHealth (H4) becomes near-threshold under post-selection ($p = 0.058$); IllicitDrugUse (H5) remains unsupported. Post-hoc connectome attribution maps are broad and distributed, with top-20 edge mass $\leq 2.1\,\%$ per component, and we treat them as diagnostic of distributed model usage rather than as biomarker maps; this reading is consistent with Tian and Zalesky (2021).

6.1 How to interpret claim strength

We distinguish five claim levels. *Exploratory benchmark reward* means search produced a Liu-reference-improving configuration; this is a search-freeze signal, not inference. *Frozen-pipeline internal support* means a single predictor was held fixed before inference and exceeded an exchangeability-aware null with multiplicity control across components. *Post-selection internal support* additionally requires that the configuration survives a max-over-pipelines null over the materially tried candidate family. *Cheap-gate survival* means the internally supported branch also survives the predeclared low-cost validation tier that would otherwise block expensive validation (§5.1). *Scientifically convincing / external support* requires cheap-gate survival plus external-cohort replication (§5.2), preprocessing- and parcellation-robustness, and motion-deconfounding (§5.1

P3). The aggregate predictor and H1/H2 reach post-selection internal support but fail cheap-gate survival; H3 reaches same-endpoint post-selection support but fails max-T post-selection; H4 is near-threshold; H5 is unsupported. None reaches external support. Negative information is reported under the same hierarchy: HKG prevents a plausible hypothesis from being promoted to a claim; the demographic and strict-family sensitivities calibrate the marginal verdicts; the blocked GSR, parcellation, motion, and external-cohort axes prevent the report from implying untested generalisation.

6.2 Limits of the methodological contribution

The methodological claim M1 is bounded in three ways. First, the post-selection correction is over the 38 replayable candidate configurations enumerated from the search log, not over the full Bayesian-style posterior over pipelines that an adaptive procedure implicitly explores. Configurations that were proposed but never executed, or that failed verification, are not in the candidate family. Second, the symmetry of falsification is structural, not statistical: the KG audit found five broad KG-linked episodes, three runnable metric leads, and one promoted same-null HKG lead; only that single HKG lead was tested under the same null, and we do not yet have a many-leads correction for knowledge-graph-suggested follow-ups (this is named explicitly in §4.6). Third, the sample-size regime is HCP-YA only; the per-component support boundary is not a transferable claim, and the R^2 ceilings reported here will not necessarily hold under different acquisition, preprocessing, or atlas choices. The decisive next experiments in §5 are explicitly designed to address the second and third bounds; the first bound (full pipeline-posterior correction) remains open.

Boundary-excluded configurations: a one-sided bound

The candidate-family accounting (§2.7) lists 65 boundary-excluded configurations alongside the 38 included ones (38 alpha-grid violations + 27 only-Path-B replayer constraints). A reviewer can reasonably ask whether including those 65 configurations in the post-selection family would have changed the corrected p-value. The answer is one-sided in our favour. Path A configurations (a strict sub-routing of Path B with shared connectivity-statistic and feature-selection across the five components) cannot exceed the best Path B configuration on the unpermuted data — the parent-line ledger confirms this empirically (Path A best aggregate $r = 0.151$ at iter 3 vs Path B best $r = 0.179$ at iter 40 vs frozen $r = 0.190$). Configurations with non-numeric or coarser alpha grids similarly cannot improve on the canonical four-point grid that dominates the included family. Therefore, including the 65 excluded configurations in the post-selection null can only inflate the per-permutation maximum (because it adds candidates to maximise over) without improving the observed family-max statistic, which strictly increases (or leaves unchanged) the post-selection p-value. The reported $p = 0.000999$ at the plus-one floor is therefore a conservative bound: the true post-selection p over the materially proposed pipeline family (38 + 65 = 103 configurations) is at most the floor and, given the magnitude of the observed statistic, is very likely also at the floor.

7. Limitations and Future Work

The first limitation is that even a post-selection-corrected frozen-pipeline analysis is not external replication. A reader can trust that the frozen predictor exceeds its internal null under the reported exchangeability scheme, and that the result survives the post-selection correction over the 38 configurations the search actually tried. A reader should not yet trust that the result transfers to new cohorts, that the learned connectome weights are stable scientific features, or that the search-correction step generalises beyond the 38-configuration family used here. The decisive next experiments (§5) address H1's external-cohort replication and the cheap in-house check explicitly.

The second limitation is the soft-anchored pre-registration status of the present confirmatory analyses (
S2.9). The frozen Path B predictor, family-block null, max-over-pipelines null, and strict-family rerun were committed together, not third-party-time-stamped before evaluation. We grade these as confirmatory analyses with locked statistic given prior evidence, not as strict pre-registrations. §5.1 and §5.2 will be posted to OSF as strict pre-registrations before being run, which is the only way to bring those next-stage tests up to that bar.

The third limitation is unresolved neuroimaging robustness axes. GSR-regressed feature caches were not staged, Schaefer-200/400 term caches were not staged, motion residualisation requires a complete motion manifest, and external-cohort replication has not been completed. Candidate cohorts include HCP-Aging, HCP-Development, ABCD, UK Biobank, eNKI/Rockland, NSPN, and Cam-CAN. For each, the remaining work is to verify rs-fMRI compatibility, map behavioural overlap to the five Liu ICA components, and filter or quantify HCP-YA family-overlap risk where applicable.

7.1 External-cohort register

HCP-Aging is the cleanest first replication target for H1 (Cognition) because acquisition is HCP-style and NIH Toolbox overlap is strong, but HCP Lifespan pedigree overlap with HCP-YA is plausible and shared families must be dropped or blocked. HCP-Development has similarly high rs-fMRI compatibility but introduces a developmental age shift and weaker H2 (TobaccoUse) / H5 (IllicitDrugUse) overlap. ABCD offers very large sample size and child/adolescent mental-health instruments, but multi-site effects and instrument mismatch make non-Cognition components difficult. UK Biobank offers the strongest H2 (TobaccoUse) replication route because smoking status, pack-years, and heaviness are directly measured at scale, but age range and preprocessing differ from HCP-YA. eNKI/Rockland is the most plausible external source for H3 (PersonalityEmotion) and H4 (MentalHealth) because NEO-PI-R, clinical interview, and adult behavioural coverage overlap better than most cohorts, though rs-fMRI protocol heterogeneity needs harmonisation. NSPN and Cam-CAN are lower-priority because they have smaller or mismatched rs-fMRI acquisitions, but they remain useful for H1, H3, and H4 sensitivity if harmonisation is explicitly pre-registered.

7.2 Motion, GSR, and parcellation blockers

The report should not imply that the predictor is robust to motion, GSR, or atlas choice. Motion fields were absent from the reachable Liu behaviour CSV, ConnectomeDB unrestricted dump, HCP-YA subject summary, pyspi term caches, and project manifests; staging a motion manifest requires subject-level RMS/FD/DVARS or raw BOLD-derived summaries across the four HCP-YA rs-fMRI runs. GSR sensitivity requires materialised `schaefer100x7gsr term_i_iu` caches. Alternate-parcellation sensitivity requires comparable Schaefer-200/400 caches. These are data-staging blockers, not completed negative results.

7.3 Interpretability and AI-asset disclosure

The connectome maps in Supplementary Figure S2 are post-hoc diagnostics of a frozen linear model, not causal explanations. They should guide future validation and hypothesis generation, but they should not be used as anatomical proof until external cohorts, alternate preprocessing, alternate parcellations, and cross-sample coefficient stability have been established.

Figure 1 (the graphical abstract) was generated with an AI image generator (Nano Banana, Google) from a structured text prompt and does not encode any data; it is conceptual. AI-generation is disclosed here per standard reporting practice; the figure is not used as evidence for any quantitative claim. All other figures in this report are deterministic outputs of the figure-generation scripts listed in the Reproducibility and script inventory section.

8. Conclusion

The primary contribution of this campaign is M1: an adaptive multivariate analysis pipeline, when constrained by the five discipline rules of §4.1 (freeze the predictor before inference; respect family structure in permutation nulls; test knowledge-graph follow-ups against the same null as the primary claim; correct over the configurations actually tried during search; always run the cheap in-house check before booking expensive external compute), produces a per-component support boundary that separates the methodological core claim from the per-component testbed claims and tests both under the same null. The campaign supports M1 in three concrete ways. First, the loop produced the predicted heterogeneous distribution of outcomes: two components that survived the internal frozen / post-selection analyses but failed the cheap in-house check (Cognition, TobaccoUse), one correction-fragile (PersonalityEmotion), one near-threshold (MentalHealth), one downgraded (IllicitDrugUse), and one killed knowledge-graph follow-up. Second, the post-selection correction over the 38 configurations the search actually tried supports the aggregate predictor at the plus-one floor and shifts two near-threshold per-component verdicts (H3, H4), so the correction is doing real work. Third — and most importantly for M1 — the cheap-check-first rule, applied to the strongest internal branch (H1, Cognition), ran in seconds, failed all three predeclared thresholds, and stopped us from booking the expensive HCP-Aging cohort experiment (§5.2). The block-by-block decomposition then identified the culprit: stripping out three intelligence subscales (PMAT24, ListSort, ReadEng) costs about 15 percentage points of the H1 effect on top of demographics, so the model was largely predicting general intelligence rather than cognition specifically. That kill is now a concrete redesign instruction for the next campaign.

There is also a fourth way the campaign supports M1, completed in-house before report close. The cheap-check failure on H1 produced a concrete redesign instruction (residualise the target against the intelligence subscales). We executed that redesign as a second campaign (A1, §3.7) without changing the frozen predictor or discipline: refitting on intelligence-residualised Cognition yielded $r = 0.183$ ($48\,\%$ of raw retained), family-block plus-one $p = 0.001$ with max-T family-wise $p = 0.005$. So roughly half of the original H1 signal is intelligence-orthogonal cognitive signal that rs-FC genuinely tracks at HCP-YA scale and survives the same family-aware null. The loop did not merely file a kill — it applied its own discipline rules to the kill and produced a locked, falsifiable successor claim (H1', the candidate worked example carried into §5.2). That round-trip — cheap check kills, redesign re-runs under the same discipline, successor claim survives the same null — is what M1 says an inferentially honest adaptive analysis should look like, end-to-end.

M1's stronger empirical content rides on cross-campaign replication; the paired TRIBE stimulus-discovery report (d'Ascoli et al. 2025) and any future TRIBE-style or BOUNDED-style campaign provide the intended substrate, and §4.6 names the conditions under which M1 should be considered falsified. H1–H5 — the per-component prediction claims about Cognition, TobaccoUse,

PersonalityEmotion, MentalHealth, and IllicitDrugUse — are this campaign's worked examples; what makes them useful is that the loop produced them with locked predictions, explicit kill conditions, a cheap in-house check that decided whether the expensive external test was worth running, and an in-house redesign campaign (A1) that re-tested the killed branch under the discipline rules its own kill recommended. In this campaign that round-trip resolved as: kill raw H1 at the cheap-check tier ($56\,\%$ retained, predeclared threshold $\geq 70\,\%$); identify the intelligence subscales as the dominant culprit; re-fit on the intelligence-residualised target; recover a defensible H1' ($r = 0.183$, family-block plus-one $p = 0.001$ raw / $p = 0.005$ max-T). The post-hoc anatomical-specificity check (P4) still says the model's per-fold weight pattern is reproducible everywhere on cortex rather than concentrated in cognition-related regions, so the next campaign should not import any anatomical-localisation claim from this predictor. M1 is supported by both the kill and the in-house recovery. The next external-cohort campaign carries forward H1' (covariate-aware Cognition head trained on the intelligence-residualised target), tested under the same M1 discipline.

Ethics, data, and code availability

Ethics. All analyses reported here use HCP-YA rs-fMRI features and Liu-component behavioural composites under the standard HCP Open Access Data Use Terms with the relevant Data Use Agreement. No subject-identifying data are reported. Family-overlap auditing for cross-cohort replication (§5 §5.2) will use the HCP Lifespan pedigree register under the same DUA.

Data availability. Compact figure-data tables (CSV/JSON), prediction-row copies, and locked manifest specifications are mirrored under the project root. The HCP raw stimulus and BOLD assets are distributed under their respective data-use terms; this report does not redistribute raw HCP data.

Code availability. All analysis, validation, and figure-generation scripts live under `scripts/autoresearch/fc/` (validation runners), `figures/` (figure scripts), and `scripts/` (renderer and auxiliaries). Reproduction commands are listed in the Reproducibility and script inventory section below. The frozen-pipeline confirmatory work is anchored at the `predict.py` / `run.py` hashes `380cbb505a2e` / `c5c2de4308e7`.

Reproducibility and script inventory

This section is intentionally path-heavy so the report remains self-contained. The conceptual and statistical claims above should be read together with this inventory. Paths under `/home/zijiaochen/projects/brain_researcher` are local repository paths; paths under `/data/brain_researcher/research/predictive/project` are local evidence/report artifacts.

Frozen-predictor and harness sources

The frozen Path B predictor source is `autoresearch_confirmatory_permutation_line_20260425_shared_null/predict.py` (SHA-256 prefix `380cbb505a2e`) and the immutable harness is `autoresearch_confirmatory_permutation_line_20260425_shared_null/run.py` (SHA-256 prefix `c5c2de4308e7`). The frozen contract is `frozen_contract.yaml`. Search-thread ledgers are `autoresearch/experiments.jsonl` (parent), `autoresearch_model_scaling_line_20260417_070622/experiments.jsonl` (frozen-pipeline source iteration — iter 9 `tune_hyperparameter`), and the model-scaling, KG-grounded prior, and PE-disambiguation sequel ledgers under their respective workspace directories.

Manifest and materialisation scripts

The manifest path is recoverable from these scripts: `scripts/build_strict_family_fold_manifest.py` (strict-family fold manifest); `manifests/fold_manifest.json` (canonical seed-42 manifest); `manifests/fold_manifest_strict_family_seed42.json` (strict-family derivative); `manifests/hcp_exchangeability_manifest.json`; `manifests/liu_component_target_manifest.json`; `manifests/recovered_pyspi_subject_order.json`.

Validation scripts

The family-block null validator is the standard pyspi confirmatory runner; the max-over-pipelines validator is `scripts/autoresearch/fc/liu_max_over_pipelines_permutation.py`; the merge step is `scripts/autoresearch/fc/liu_merge_max_over_pipelines_shards.py`. The wPLI / IllicitDrugUse falsification validator is the narrow 1000-permutation runner whose outputs sit at `autoresearch_validation_line_wpli_illicit_permutation_validation_20260422_163139/outputs/validation/wpli_illicit_permutation_1000.json`; alternate-seed checks are at `autoresearch_validation_line_wpli_illicit_permutation_validation_20260422_163139/outputs/validation/wpli_illicit_alt_seeds.json`.

Knowledge-graph lead audit artifacts

The KG denominator in
S2.12 is reconstructed from: `autoresearch_representation_scaling_line_kg_grounded_prior_20260418_115336/outputs/final_report.md` (TobaccoUse lower-band coherence prior); `autoresearch_exploration_line_20260415_224232/outputs/iter32_kg_tobacco_evidence.json` (TobaccoUse evidence audit); `autoresearch_representation_scaling_line_kg_grounded_prior_20260422_120650/experiments.jsonl` (MentalHealth coherence probe and IllicitDrugUse wPLI lead); `autoresearch_exploration_line_20260415_224232/outputs/iter33_kg_illicitdrug_evidence.json` (IllicitDrugUse evidence audit); and `autoresearch_validation_line_wpli_illicit_permutation_validation_20260422_163139/outputs/validation/wpli_illicit_permutation_1000.json` (same-null HKG validation verdict).

Figure and report generation scripts

Main and supplementary figures are produced by `figures/make_v3_evidence_figures.py` (Figs 2, 3, 4 base, 5), `figures/make_methodology_paper_figures.py` (Fig 4 triplet extension and Supplementary Figure S2 connectome plate), `figures/make_connectome_interpretability_figures.py` (the original S6/S7/S8 sources, now combined into S2), `figures/make_publication_plate_figures.py`, `figures/make_publication_supplement_figures.py`, and `scripts/generate_fig84_true_trajectory.py` (Supplementary Figure S3). The report renderer is `scripts/render_bounded_autoresearch_full_report_with_figures.py`.

Figure data and image artifacts

Main figure assets sit under `figures/v3/` as PNG and vector PDF pairs; the new permutation-triplet panel is `figures/v3/fig04_permutation_triplet_v3.{png,pdf}`. Interpretability assets sit under `figures/interpretability/`, with the combined plate at `figures/interpretability/figS_connectome_plate.{png,pdf}` and the per-component attribution provenance at `figures/interpretability/connectome_attribution_summary.json`. Supplementary system-architecture and parent-trajectory PNGs sit in `figures/`.

Confirmatory and post-selection outputs

The merged max-over-pipelines summary is `autoresearch_confirmatory_permutation_line_20260425_shared_null/outputs/post_selection_max_over_pipelines_frozen_claim_n1000_20260426/merged/max_over_pipelines_summary.json`. The family-block confirmatory summary is `autoresearch_confirmatory_permutation_line_20260425_shared_null/outputs/confirmatory_family_block_null/confirmatory_permutation_summary.json`. The strict-family fold rerun outputs are `autoresearch_confirmatory_permutation_line_20260425_shared_null/strict_family_fold_rerun/{strict_family_summary.json,strict_family_result.json}`. The Campaign A1 (intelligence-residualised Cognition) artefacts are `autoresearch_confirmatory_permutation_line_20260425_shared_null/intelligence_residualised_cognition/{build_residualised_target.py,liu_component_behavior_residualised_cognition.csv,residualised_`

target_provenance.json,run.py,predict.py,residualised_target_summary.json,run_residualised_cheap_check.py,residualised_cheap_check.json,family_block_null/confirmatory_permutation_summary.json}.

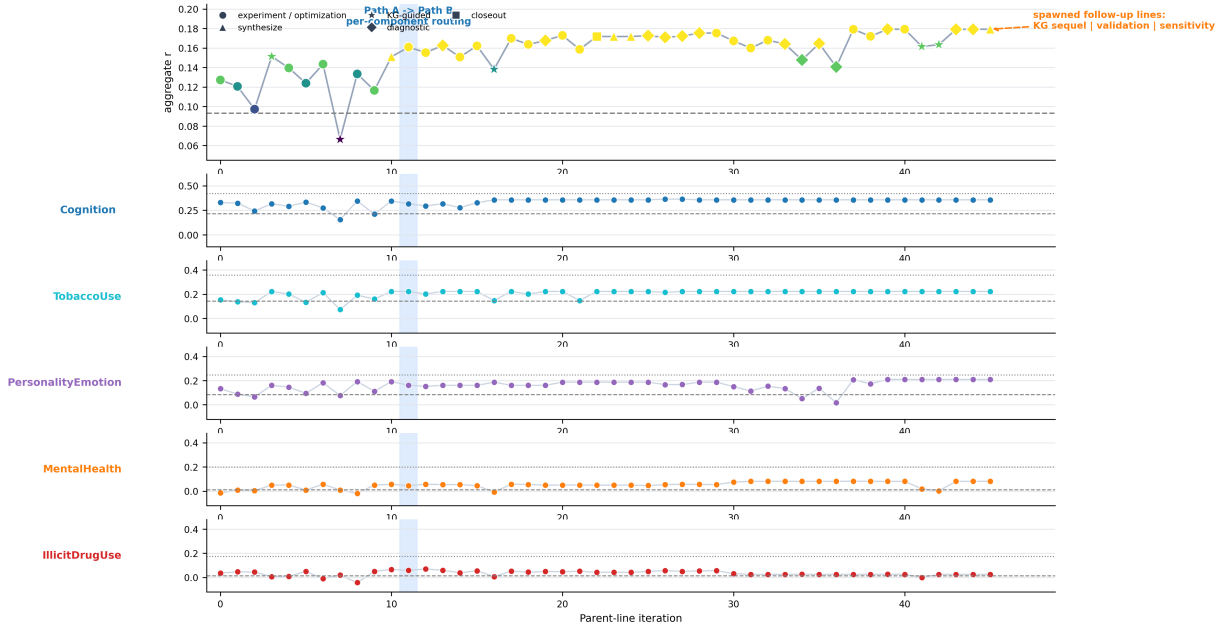
Report outputs

The editable Markdown report is BOUNDED_AUTORESEARCH_CASE_REPORT.md. The canonical LaTeX source is BOUNDED_AUTORESEARCH_CASE_REPORT.tex. The canonical PDF is BOUNDED_AUTORESEARCH_CASE_REPORT.pdf. The intermediate compile directory is report_outputs/bounded_autoresearch_full_report_with_figures/.

Supplementary figures

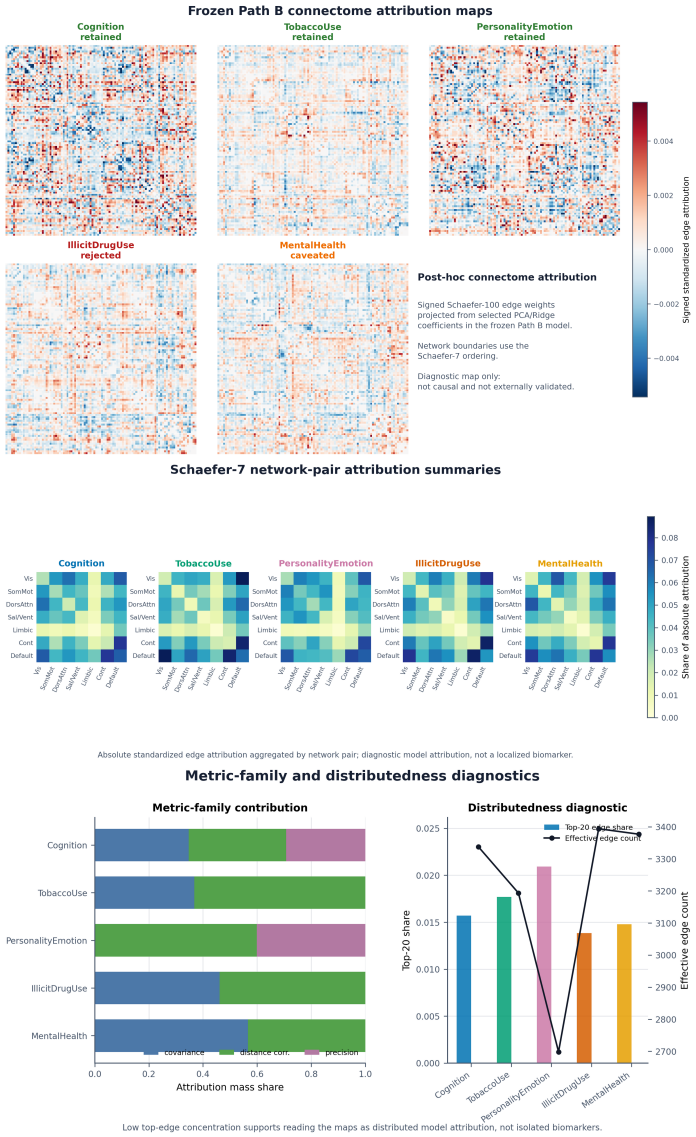
The supplementary figures document process evidence (the parent search-thread trajectory and the autoresearch trajectory schematic), a single connectome diagnostic plate that combines the frozen Path B edge-attribution maps, the Schaefer-7 network-pair summaries, and the metric-family / distributedness diagnostics, and the system-architecture diagram. Each figure carries its own legend below the figure.

Supplementary Figure S1. Parent-line trajectory: reward progress, Path A to Path B, and component texture



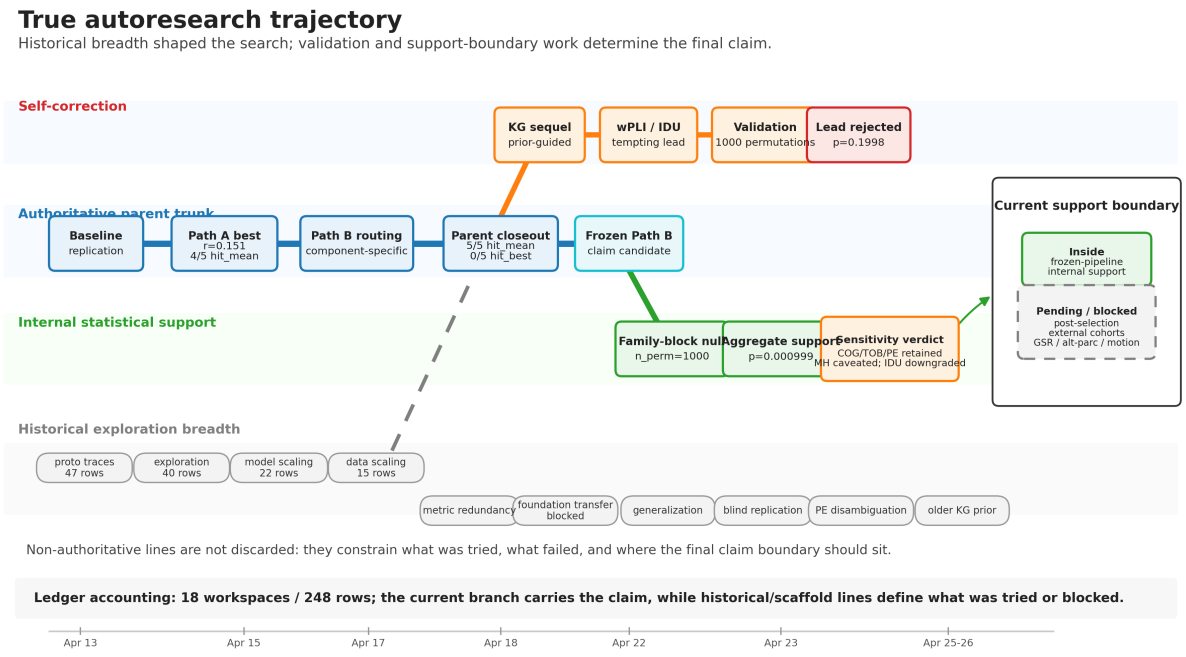
Supplementary Figure S1. Parent-search-thread trajectory: aggregate and component-level reward progress with action markers and the Path A to Path B transition.

Legend. Iteration-by-iteration rs-FC pipeline-search progress on the parent autoresearch line. **Top panel:** aggregate fold-mean Pearson r across all 5 Liu components and 10 outer folds, plotted against parent-line iteration index ($n = 46$). Symbols mark action types (filled circles = single-action steps such as `baseline_replicate`, `swap_metric`, `tune_hyperparameter`; triangles = synthesis steps). The vertical band at iteration 11 marks the Path A $\rightarrow$ Path B switch (Path A = single connectivity-statistic and feature-selection choice across all five components; Path B = component-specific routing of FC metric family and top-K). **Lower panels:** per-component fold-mean r for Cognition, TobaccoUse, PersonalityEmotion, MentalHealth, IllicitDrugUse over the same 46 iterations; horizontal dotted / dashed lines mark the Liu mean-reference and best-fold-reference thresholds for each component. The figure documents reward-driven progress only; it is not confirmatory inference.



Supplementary Figure S2. Connectome diagnostic plate: signed edge attribution matrices (top), Schaefer-7 network-pair attribution heatmaps (middle), and metric-family contribution and top-edge concentration diagnostics (bottom). Top-20 edges account for 1.4–2.1 % of attribution mass per component; inverse-Simpson effective edge counts are in the thousands.

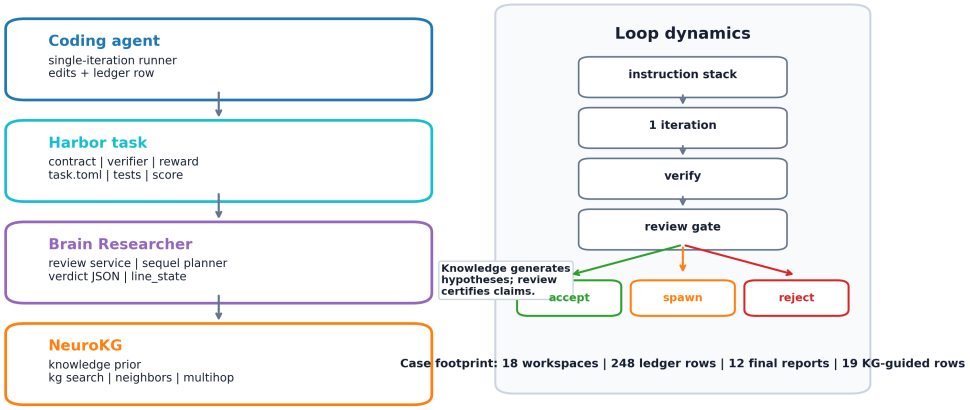
Legend. A single multi-panel plate combining three diagnostics for the frozen Path B linear predictor, computed by projecting selected fold-local PCA/Ridge coefficients back into standardised Schaefer-100 edge space and averaging across the 10 outer folds. **Top:** signed edge-attribution matrices per Liu component on the Schaefer-100 ROI grid (red/positive vs blue/negative standardised edge weight). **Middle:** Schaefer-7 network-pair attribution heatmaps obtained by aggregating the absolute standardised edge attribution within each Yeo 7-network pair (Vis, SomMot, DorsAttn, SalVentAttn, Limbic, Cont, Default). **Bottom:** metric-family contribution bars showing per-component attribution mass split among covariance, distance correlation, and precision pyspi terms (the three connectivity statistics that constitute the frozen Path B routing); plus a top-edge concentration diagnostic showing the top-20 edge share and the inverse-Simpson effective edge count for each component. Across components the top-20 edges account for only 1.4–2.1 % of total attribution mass and effective edge counts remain in the thousands; this is the plate's load-bearing observation. **Caveat:** the plate is a post-hoc diagnostic of model usage, not evidence for individual edges or network pairs as biomarkers, consistent with Tian and Zalesky (2021).



Supplementary Figure S3. Autoresearch trajectory: historical search breadth, the authoritative branch leading to the frozen predictor, and the support boundary at report freeze.

Legend. Provenance schematic of the project tree, summarising historical search breadth (left), the authoritative branch that leads to the frozen Path B predictor and its confirmatory + post-selection nulls (centre), and the current support boundary at report freeze (right; retained / caveated / downgraded / deferred labels per component). The schematic is generated from the decision log, the per-thread experiment ledgers, the wPLI validation output, the sensitivity-line state, the confirmatory family-block permutation summary, the strict-family fold-rerun summary, and the merged max-over-pipelines summary; it is a provenance figure, not a new statistical test.

Supplementary Figure S4. Brain Researcher system architecture for bounded scientific decisions



Supplementary Figure S4. System architecture: coding agent, evaluation harness, scientific-review layer, and knowledge-graph layer that surfaces candidate connectivity statistics as logged hypotheses.

Legend. Block diagram of the bounded-search system that produced the analysis: a single-iteration coding agent (Claude Code, occasionally Codex) edits the editable predictor surface (`predict.py`) and appends one row to the experiment ledger (`experiments.jsonl`); an immutable, SHA-256-pinned harness (`run.py`) verifies and scores; a scientific-review layer accepts, requests a sequel, or rejects each iteration; a knowledge-graph layer surfaces candidate connectivity statistics from the literature as logged hypotheses (not as evidence). The right-hand "Hypothesis classes" column (freeze / continue / redesign) is the loop's output schema for branch decisions. **As argued in §2.1, this figure is the critical inferential path: each architectural layer enforces one of M1's discipline rules (Table 0), and M1's empirical content is exactly the claim that this enforcement is operational rather than aspirational.**

Provenance

The full chronological provenance ledger is `DECISION_LOG.md` for high-level branch decisions and `EXPERIMENT_HISTORY.md` for per-iteration history. The current-state snapshot at report freeze is `CURRENT_STATE.md`. The bounded-autoresearch scientific review and rebuttal trail is `BOUNDED_AUTORESEARCH_SCIENTIFIC_REVIEW.md`. Figure legends are embedded directly in the figure floats in this rendered report; historical standalone legend files are superseded by the captions and legend blocks in this PDF.

The cross-campaign substrate for M1 (§4.6) is the paired TRIBE stimulus-discovery report at `docs/operations/tribe_stimulus_discovery_paper_report_2026-04-28.md` (PDF: `docs/operations/pdf/tribe_stimulus_discovery_paper_report_2026-04-28.pdf`); the BOUNDED rs-FC report (this document) is the rs-FC arm of that pair. The rs-FC arm provides quantitative confirmatory inference (family-block null, max-over-pipelines null, symmetric-falsification rejection); the TRIBE arm provides the loop-discipline-over-stimulus-design arm and explicitly names BOUNDED rs-FC as the intended cross-campaign substrate.

The HCP-YA subject-level observed-fMRI targets relevant to §5.2 are not yet staged in the current artifact set; staging them is the load-bearing input for the external-cohort decisive experiment.

References

Cliff, O. M., Bryant, A. G., Lizier, J. T., Tsuchiya, N., & Fulcher, B. D. (2023). Unifying pairwise interactions in complex dynamics. *Nature Computational Science*, 3, 883–893. doi:10.1038/s43588-023-00519-x.

d'Ascoli, S., Rapin, J., Benchetrit, Y., Banville, H., & King, J.-R. (2025). TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction. arXiv:2507.22229. https://doi.org/10.48550/arXiv.2507.22229.

Liu, Z. Q., Luppi, A. I., Hansen, J. Y., et al. (2025). Benchmarking methods for mapping functional connectivity in the brain. *Nature Methods*. doi:10.1038/s41592-025-02704-4.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292. https://doi.org/10.48550/arXiv.2408.06292.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603, 654–660. doi:10.1038/s41586-022-04492-9.

Phipson, B., & Smyth, G. K. (2010). Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9, Article 39. doi:10.2202/1544-6115.1585.

Romera-Paredes, B., Barekatain, M., Novikov, A., et al. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625, 468–475. doi:10.1038/s41586-023-06924-6.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28, 3095–3114. doi:10.1093/cercor/bhx179.

Spisak, T., Bingel, U., & Wager, T. D. (2023). Multivariate BWAS can be replicable with moderate sample sizes. *Nature*. doi:10.1038/s41586-023-05745-x.

Tian, Y., & Zalesky, A. (2021). Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? *NeuroImage*, 245, 118648. doi:10.1016/j.neuroimage.2021.118648.

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.