Automatically Generated Report

Conducted via Brain Researcher MCP and CodeX

High-Dimensional Connectome Matrices: Hubness, Decoding, and Generative Fidelity

Brain Researcher Canonical Manuscript Report

Ce Ju, Bertrand Thirion

Inria, CEA, and Université Paris-Saclay

Date: 2026-05-06

Contents

List of Figures

List of Tables

Abstract

This report studies a single evaluation question in two linked stages: when brain matrices are embedded in a high-dimensional space, can local-neighborhood methods be trusted, and if not, what parts of generative fidelity remain interpretable? The primary experiments are ADNI fMRI connectome analyses; the EEG covariance result is retained as appendix-level supporting evidence rather than a matched primary endpoint. The unified hypothesis is that evaluation of high-dimensional connectome or SPD-matrix data is layered. First, neighborhood-based decoding is only interpretable if the neighborhood graph is not badly distorted by hubness. Second, even after repairing local geometry, biological fidelity of generated connectomes must be evaluated separately at global and disease-sensitive levels rather than inferred from one local-neighborhood score.

The strongest supported findings are deliberately narrower than a clinical or causal claim. A split audit shows that all archived ADNI train/validation partitions are subject-disjoint for both MSDL-39 and Schaefer-100. Within that audited setting, Local Scaling and NICDM are associated with healthier neighbor graphs and higher validation balanced accuracy across the main neighborhood sizes, while the Schaefer-100 k=50 row remains a bounded computational probe. For generated connectomes, global resting-state structure is preserved strongly for Schaefer-100 and moderately for MSDL-39, but disease-sensitive geometry is unstable under the current pooled non-control label. The resulting conclusion is a review-oriented one: this artifact bundle supports a reproducible exploratory scorecard in which local-neighborhood validity, global fidelity, and disease-shift fidelity are separate inferential layers, and several visually striking or representation-specific patterns should be treated as exploratory rather than headline evidence.

Main Takeaways

- The report now advances one layered claim rather than two loosely linked ones: neighborhood validity must be audited before neighborhood-based decoding is interpreted, and generative fidelity must then be separated into global and disease-sensitive layers.
- In the audited ADNI setting, Local Scaling and NICDM are associated with healthier neighbor graphs and better subject-disjoint kNN balanced accuracy.
- The H1 result is no longer carried by one score alone; it is backed by split auditing, paired intervention tests, label-permutation nulls, non-kNN references, and SPD QC.
- Generated connectomes preserve global structure much better than disease-sensitive structure, especially under the current pooled non-control label.
- The strongest retained output is a scorecard, not a single-model victory claim: neighborhood health, decoding utility, global fidelity, reproducibility, and disease-shift fidelity should be reported separately.
- Atlas superiority, disease-conditioning success, the Schaefer-100 `k=50` probe, and purely visual summaries remain exploratory.

Review-Oriented Hypothesis Frame

The report is strongest when framed as one total hypothesis with two nested tests rather than as two parallel ideas.

- **Total hypothesis**: evaluation of high-dimensional brain matrices is layered. A neighborhood-based result is only scientifically interpretable if the local geometry is not badly distorted by hubness, and generative fidelity should not be inferred from local-neighborhood behavior alone.
- **H1, local-validity test**: if hubness materially distorts the local graph, then methods that normalize local density should improve neighborhood-health diagnostics and should coincide with better subject-disjoint kNN decoding.
- **H2, biological-fidelity test**: if local-neighborhood readouts are only one layer of evaluation, then generated connectomes may preserve global resting-state structure while still failing disease-sensitive fidelity tests; therefore global fidelity and disease-shift fidelity must be reported separately.

This framing makes the logical link explicit. H1 is not a side story about hubness for its own sake; it is the prerequisite validity check for any conclusion that depends on neighborhoods. H2 is not a separate generative paper stapled onto H1; it is the second-stage test of what remains scientifically interpretable once local-neighborhood validity has been audited.

Results to Keep vs Exploratory

The current artifact bundle supports a mixed evidence hierarchy rather than one undifferentiated result list.

0.1 Keep as main evidence

- Subject-disjointness of all archived ADNI train/validation splits for MSDL-39 and Schaefer-100.
- Strong H1 intervention pattern: Local Scaling and NICDM improve hubness diagnostics and are associated with better subject-disjoint validation balanced accuracy.
- Completed H1 uncertainty package: paired intervention tests, label-permutation nulls, non-kNN linear references, and SPD QC.
- H2 global-fidelity result: generated connectomes preserve global mean-structure strongly for Schaefer-100 and moderately for MSDL-39.
- H2 negative result: disease-effect and gradient-shift fidelity are unstable under the current control-vs-pooled-non-control label.
- Scorecard conclusion: neighborhood health, decoding utility, global fidelity, reproducibility, and disease-sensitive fidelity should be reported as separate layers.

0.2 Downgrade to exploratory or supporting status

- Any claim that one atlas is intrinsically better than another. The Schaefer-100 versus MSDL-39 contrast is currently atlas-dependent evidence, not a principled theory of atlas superiority.
- The Schaefer-100 `k=50` row. It is a bounded computational probe, not a full matched replication.
- The EEG covariance result. It is useful supporting evidence for the geometry problem, but it is not a co-equal primary endpoint with the ADNI connectome study.
- Mean heatmaps, PCA common-space plots, and chord diagrams. These are useful descriptive diagnostics, but they should not be treated as stand-alone inferential evidence.
- Any statement implying successful disease conditioning by the generator. Under the pooled label, disease-sensitive fidelity is too unstable for that claim.

- Any mechanistic or causal statement that hubness alone explains downstream performance differences. The current evidence supports a neighborhood-intervention association, not isolated causality.
- The Mutual Proximity failure mode as a scientific result. It is currently best treated as an implementation-sensitive engineering warning.

Introduction

High-dimensional brain matrices are now common in neuroimaging and BCI workflows. In fMRI, a subject can be represented by a connectome: a matrix whose entries summarize functional relationships between brain regions. In EEG and other covariance-based pipelines, a trial or subject can be represented by a symmetric positive definite matrix. These representations are attractive because they preserve pairwise structure, but they also create a statistical evaluation problem: once the matrices are embedded into a high-dimensional feature space, local neighborhoods can become unreliable.

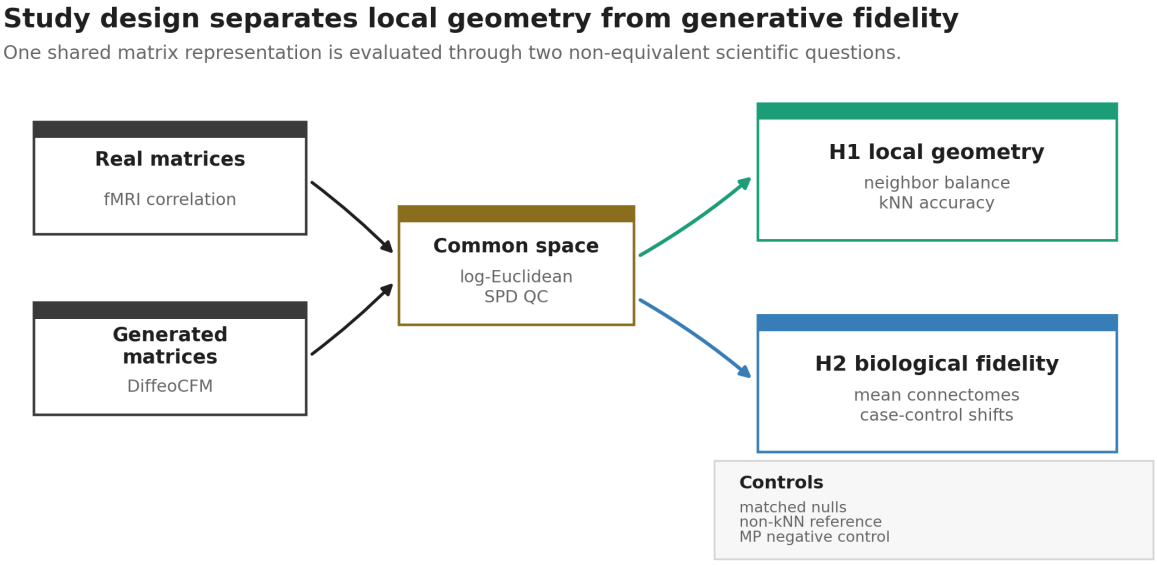

Figure 1: Figure 1. Study design separates local-neighborhood validity from biological fidelity. Real matrices are fMRI correlation connectomes; generated matrices are DiffeoCFM outputs from the generated-matrix setting introduced by Collas et al. (2025). Common space means that real and generated matrices are mapped into the same representation before comparison. Log-Euclidean means that SPD matrices are transformed by the matrix logarithm and then vectorized; SPD QC checks positive eigenvalues and numerical safety before that logarithm is used. H1 asks whether the nearest-neighbor graph is sufficiently well behaved for neighborhood-based decoding to be interpretable. H2 asks whether generated samples preserve global and disease-sensitive structure even when local-neighborhood evaluation is fragile. Controls denote matched null tests, non-kNN classifier references, and Mutual Proximity as a distance-change control.

This problem has been discussed in machine learning under the name **hubness**. In a high-dimensional space, a small number of samples may become nearest neighbors for many other samples. These samples are called hubs. Other samples may almost never appear in anyone's

neighbor list; these are antihubs. If a classifier, retrieval system, or sample-quality metric depends on nearest neighbors, hubness can alter the result even when the original representation looks reasonable.

The terminology and motivation follow the hubness literature. Radovanovic, Nanopoulos, and Ivanovic (2010) showed that high-dimensional distance spaces naturally produce **popular nearest neighbors**, or hubs, and that this can affect learning algorithms based on nearest neighbors. Schnitzer, Flexer, Schedl, and Widmer (2012) later studied hubness-reduction strategies, including **Local Scaling** and **Mutual Proximity**. We use those ideas here because SPD matrices and connectomes become high-dimensional vectors after tangent-space, log-Euclidean, or edge-vector embedding.

The review problem in this report is not simply whether one method beats another on one benchmark. The deeper question is whether a neighborhood-based result means what the reader thinks it means. If the local graph is badly distorted, then a kNN score or nearest-neighbor realism score can be hard to interpret. If the graph is partially repaired, that still does not imply that generated samples preserve biologically meaningful disease structure. This is why the report now adopts a stronger and more explicit total hypothesis: high-dimensional brain-matrix evaluation should be separated into a local-validity layer and a biological-fidelity layer.

H1 is the local-validity test: if hubness is operationally harmful for a neighborhood decoder, then reducing hubness should improve graph diagnostics and should coincide with higher subject-disjoint balanced accuracy. H2 is the biological-fidelity test: if neighborhood overlap is only one layer of validity, then generated connectomes should be assessed separately at global, reproducibility, and disease-sensitive levels. This framing also determines evidence grading. The main argument should rest on the split audit, the Local Scaling/NICDM intervention pattern, the non-kNN references, the completed H1 null package, and the H2 divergence between global and disease-sensitive fidelity. More visually persuasive or representation-specific results, such as the Schaefer-100 `k=50` row, the EEG appendix result, and mean/chord/PCA figures, should remain exploratory.

The metric design follows this logic. H1 uses k-occurrence skewness, antihub fraction, and mutual-neighbor fraction because they directly describe whether the neighbor graph is dominated by hubs, missing many samples, or asymmetric. It uses balanced accuracy because the downstream task is binary and potentially imbalanced. H2 uses global edge correlation because it is a direct map-level fidelity measure, bootstrap reproducibility because stable generated means are useful in population-level modeling, and disease-specific gradient shift because clinical or disease modeling ultimately depends on preserving directional pathology-related structure.

The connection between H1 and H2 is methodological rather than rhetorical. H1 audits whether local-neighborhood readouts are trustworthy in the first place. H2 then asks which conclusions about generated matrices survive once evaluation is decomposed into global structure, reproducibility, and disease-sensitive directions. No single number is enough: a representation can have good global structure but bad local neighborhoods, and a generator can reproduce average connectomes while losing disease-specific shifts.

The generated fMRI and EEG matrix setting in this report follows Collas, Ju, Salvy, and Thirion (2025), **Riemannian flow matching for brain connectivity matrices via pullback geometry**, presented at the Thirty-ninth Annual Conference on Neural Information Processing Systems. The hubness-sensitive generative-evaluation framing is also closely related to later work by Salvy, Talbot, and Thirion (2026) on hubness-aware distance-based evaluation. We treat the present report as an evaluation audit around that generated-matrix setting, not as a claim that one global or neighborhood metric alone is sufficient.

Evidence Chain and Threat Resolution

The main concern with H1 is not whether one table improves after a distance transformation. The concern is whether the improvement can be explained away by a weaker alternative: maybe baseline kNN is simply a poor classifier, maybe the evaluation procedure is broken, maybe any distance change would improve performance, or maybe the effect is not specific to neighborhood methods. The analysis therefore follows a staged resolution chain.

Table 1: Table 1. Five-phase evidence chain used to resolve the main H1 threats.

Phase	Question / Threat	Test Used Here	Resolution
Phase 1	Baseline kNN on log-Euclidean features gives only modest balanced accuracy, around 0.52. Is the method weak, or is the evaluation geometry weak?	Baseline kNN on subject-disjoint ADNI splits.	Baseline performance alone is ambiguous; it motivates graph diagnostics rather than a direct performance claim.
Phase 2	Is the nearest-neighbor graph itself distorted?	k-occurrence skewness, antihub fraction, and mutual-neighbor fraction.	The raw graph is highly asymmetric, with many antihubs and strong hub concentration, so the kNN evaluation is not a neutral readout of representation quality.
Phase 3	Do hubness-reduction methods repair the graph and improve kNN?	Local Scaling and NICDM across k=5/10/20/50 with paired tests and label-permutation nulls.	Both methods improve graph health and subject-disjoint kNN balanced accuracy. Paired improvement is significant in all Local Scaling/NICDM real-to-real rows; label-permutation is formally reported for every available row and is strongest for Schaefer-100 and MSDL-39 k=50.
Phase 4	Is the gain just caused by changing distances or by a generic classifier effect?	Mutual Proximity as a negative-control distance change; Logistic Regression and Linear SVM as non-kNN references.	Not every distance change helps: the current Mutual Proximity implementation worsens graph health and balanced accuracy. Linear models on the same vectors do not use neighbor graphs and do not show the same intervention-style gain, supporting a neighborhood-specific mechanism.

Table 1: Table 1. Five-phase evidence chain used to resolve the main H1 threats.

Phase	Question / Threat	Test Used Here	Resolution
Phase 5	Does local hubness determine whether generated samples are globally good?	H2 global edge fidelity, bootstrap reproducibility, disease-effect vectors, and gradient-shift metrics.	Local hubness and global generative fidelity are separate layers. Generated connectomes can preserve global structure, especially Schaefer-100, while disease-sensitive directions remain unstable under the current pooled disease label.

This chain does not turn the study into a causal proof that one scalar hubness metric alone drives performance. Local Scaling and NICDM alter the full neighborhood geometry. What it does support is a narrower and more useful claim: in these high-dimensional connectome embeddings, the raw nearest-neighbor graph is pathological, graph-aware corrections improve both graph health and kNN decoding, and the effect is not reproduced by either a failed distance-control method or by non-neighborhood linear classifiers.

Study Design

The Alzheimer's Disease Neuroimaging Initiative (ADNI) is a longitudinal, multi-site study designed to support clinical trial development for Alzheimer's disease and related disorders. It collects multimodal biomarkers and neuroimaging data across multiple phases. Data are acquired at more than 60 sites in the United States and Canada, and imaging protocols evolve across phases. For this benchmark, ADNI scans were identified from the available preprocessed subject-session resting-state fMRI files. Atlas-based time series were extracted, matched to phenotype data by `SubjectID` and `Session`, and then filtered using the common post-extraction quality-control procedure. The final benchmark sample comprised 936 participants with 1,997 resting-state fMRI scans, spanning ages 51.0 to 104.0 years with mean age 73.9 years; 427 participants were male.

The available artifacts are organized as train/validation partitions over subject-session scans, and each partition includes a `groups\_*.npy` array containing ADNI subject IDs. A new audit over all 10 splits confirms strict subject-disjointness between train and validation for both ADNI representations. Across 20 dataset-split rows, the total number of overlapping train/validation subjects is 0. The supporting split manifest is listed in Appendix G as `h1\_subject\_disjoint\_manifest.csv`. This removes the strongest repeated-scan leakage concern for H1, although the benchmark remains an archived derived-matrix analysis rather than a full rerun from raw time series.

We use two ADNI-derived fMRI connectome representations:

- ADNI MSDL-39: lower-dimensional atlas, evaluated with the full log-Euclidean vector representation.
- ADNI Schaefer-100: higher-dimensional atlas, evaluated with the full log-Euclidean vector representation for k=5/10/20. For k=50, the full representation was too slow, so we report a bounded raw edge-vector probe using 500 training samples per split.

The log-Euclidean representation requires matrices with positive eigenvalues. A new SPD QC audit checks the stored real and generated ADNI matrices across both atlases and all 10 splits. It found no non-positive eigenvalue counts and no eigendecomposition failures. The smallest observed eigenvalue is positive, and the feature extractor still applies a conservative numerical safety rule by clipping eigenvalues to at least 1e-8 before taking the matrix logarithm.

The original diagnosis labels available for this benchmark contain four clinical groups: CN (837 samples), MCI (847 samples), SMC (133 samples), and AD (180 samples). The primary H1/H2 analyses intentionally collapse these labels into a binary contrast: class 0 is CN, and class 1 is non-CN, pooling MCI, SMC, and AD. We use this CN-vs-non-CN contrast as a broad abnormality screen because the smaller SMC and AD groups are underpowered for stable four-class neighborhood decoding, and because the central question here is whether geometry-aware evaluation

detects any disease-related shift. This choice is clinically coarse: it is not the same as a clean CN-vs-AD, CN-vs-MCI, or severity-stratified disease contrast.

Generated matrices are class-conditional in the archived artifact bundle: the stored generated arrays are paired with binary pooled condition labels, where class 0 is CN and class 1 pools non-control diagnoses. They are also sample-matched to the corresponding real partition in the archived arrays. For example, in split 0, MSDL-39 has 2,108 real train scans and one generated sample per real train scan, with matched class counts 880/1,228; its validation partition has 235 real scans and matched generated class counts 100/135. Schaefer-100 split 0 similarly has 1,797 train scans with class counts 753/1,044 and 200 validation scans with class counts 84/116, again matched by generated samples.

This report therefore evaluates the fidelity of the stored binary condition-labeled generated outputs, not a freshly audited training pipeline. The local bundle does not independently verify the generator's training-time conditioning mechanism, loss, or label-ingestion path. Disease-shift results should consequently be read as a test of whether the archived CN-vs-pooled-non-CN generated matrices preserve the corresponding real-data shift, not as proof that a finely diagnosis-conditioned model has or has not learned AD pathology. This distinction matters because disease-shift fidelity is being tested under a coarse binary conditioning scheme rather than under separate CN-vs-AD, CN-vs-MCI, or severity-aware generation.

H1 tests whether hubness correction improves neighborhood-based decoding. H2 tests whether generated samples preserve brain-like structure on global and disease-sensitive metrics that should not be reduced to local-neighborhood realism alone.

The appendix also reports the earlier EEG covariance-matrix hubness check on `BNCI2014\_001`. That result is kept separate because it was produced with a different protocol: EEG covariance matrices were transformed to tangent-space features and summarized as 5-fold within-CV and bidirectional cross-session averages. The archived EEG table does not contain split-level confidence intervals, so it is treated as supporting exploratory evidence rather than as part of the main inferential fMRI tables.

H1: Hubness Intervention

0.3 Core Question

If a representation has distorted nearest-neighbor structure, can we repair that structure and improve decoding?

Under the current scan-level artifact, Local Scaling and NICDM reduce hubness and improve balanced accuracy. The main evidence is summarized in Figure 2 as an overview aggregation across the available ADNI representations; detailed per-representation values are moved to Appendix F.

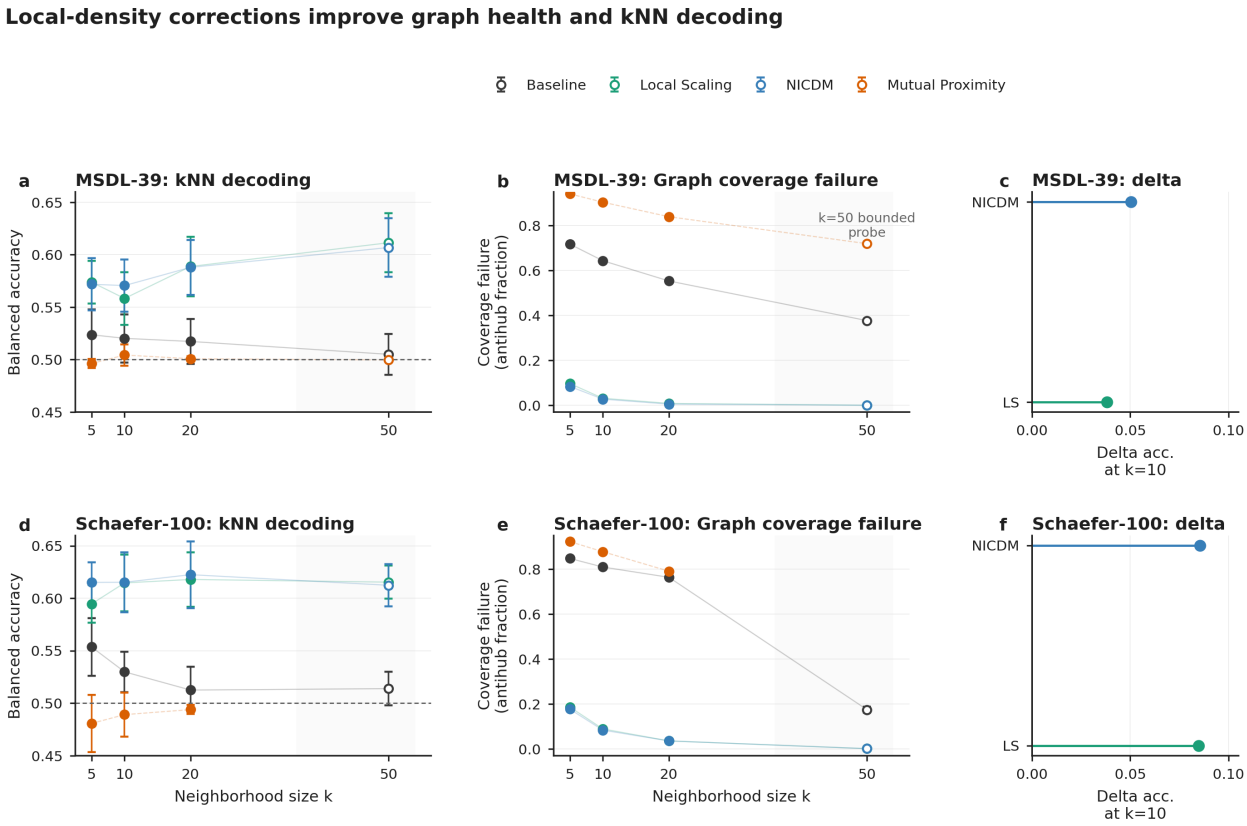

Figure 2: Figure 2. Local-density corrections improve graph health and kNN decoding. MSDL-39 and Schaefer-100 are shown separately; points are split-level means with approximate 95% intervals. The grey k=50 region marks the larger-neighborhood sensitivity probe, with Schaefer-100 using the bounded raw-edge protocol rather than the full log-Euclidean protocol.

0.4 Neighborhood-Size Sensitivity and Uncertainty

Across this sweep, the baseline graph remains less useful for decoding. Increasing k naturally lowers skewness because each sample can pick more neighbors, but it does not solve the problem by itself. The corrected graphs remain more reciprocal, have fewer antihubs, and produce higher balanced accuracy. The k=50 rows are larger-neighborhood sensitivity checks; Schaefer-100 at k=50 uses a bounded raw edge-vector probe, not the exact full log-Euclidean representation used for k=5/10/20.

At k=10, MSDL-39 balanced accuracy increases from 0.520 at baseline to 0.558 with Local Scaling and 0.571 with NICDM; Schaefer-100 increases from 0.530 to approximately 0.615. A paired split-level check supports the intervention association at k=10: MSDL-39 has mean balanced-accuracy deltas of +0.038 for Local Scaling and +0.050 for NICDM, with paired sign-permutation p-values of 0.004 and 0.002; Schaefer-100 has deltas of +0.085 for both Local Scaling and NICDM, with p-values of 0.004. Mutual Proximity moves in the opposite direction in balanced accuracy. Because the split audit confirms zero subject overlap, these paired checks should be interpreted as subject-disjoint validation results for the archived ADNI matrices.

Table 2: Table 2. Paired intervention-versus-baseline balanced-accuracy deltas at k=10.

Dataset	Intervention vs Baseline	Mean Delta in Balanced Accuracy	Split SD	Paired Sign-Permutation p	Positive Splits
ADNI MSDL-39	Local Scaling	+0.038	0.027	0.004	9/10
ADNI MSDL-39	NICDM	+0.050	0.019	0.002	10/10
ADNI MSDL-39	Mutual Proximity	-0.016	0.035	0.188	2/10
ADNI Schaefer-100	Local Scaling	+0.085	0.055	0.004	9/10
ADNI Schaefer-100	NICDM	+0.085	0.055	0.004	9/10
ADNI Schaefer-100	Mutual Proximity	-0.041	0.039	0.016	1/10

At k=50, MSDL-39 uses the full log-Euclidean representation and Schaefer-100 uses a bounded raw edge-vector probe; both retain the same qualitative pattern.

The strengthened H1 audit resolves the earlier reporting gap. It now attaches split-wise mean and standard deviation, bootstrap confidence intervals, label-permutation null tests, and paired intervention-vs-baseline sign-permutation tests to every available H1 method row, including k=50. Figure 2 remains an overview aggregation; the exact per-atlas inference table is in Appendix F. The k=50 Schaefer-100 rows should still be read as bounded raw-edge probes rather than full log-Euclidean replications.

The strengthened H1 audit now covers all available real-to-real k rows, including the bounded k=50 probe, with split-wise mean +/- SD, bootstrap confidence intervals, paired intervention-vs-baseline sign-permutation tests, and 99-label-permutation null tests per split. Local Scaling and NICDM have paired improvement $p<0.05$ in 16/16 real-to-real rows. Label-permutation evidence is more

conservative: 10/16 Local Scaling/NICDM rows have mean empirical label-permutation $p<0.05$. The remaining MSDL-39 rows are formally tested but borderline rather than uniformly significant.

0.5 Non-kNN Classifier Reference

As a hubness-insensitive reference, Logistic Regression and Linear SVM on the same log-Euclidean/tangent vectors reach balanced accuracy in the 0.556-0.571 range. The best linear reference is logistic regression on ADNI Schaefer-100 at 0.571 +/- 0.037. These values are modest in absolute terms, which is expected in a subject-disjoint ADNI setting with coarse pooled labels, and they should not be read as near-clinical discrimination. These linear models do not use nearest-neighbor graphs, so the Local Scaling/NICDM gain is best interpreted as specific evidence about neighborhood-based decoding rather than a universal classifier improvement.

This matters for interpretation. Local Scaling and NICDM are distance-graph interventions; they are expected to affect kNN and other neighborhood-based procedures. A linear SVM or logistic regression on fixed tangent vectors does not build a nearest-neighbor graph, so it should not benefit directly from hubness correction. The linear-reference result therefore strengthens the causal story in a narrower way: hubness matters for neighborhood decoding, not necessarily for every classifier trained on the same features.

0.6 H1 Interpretation

If a high-dimensional connectome representation is used for nearest-neighbor decoding, retrieval, or generated-sample evaluation, hubness should be checked before trusting the result. In these experiments, the uncorrected graph is highly asymmetric: many samples are never selected, and a few samples dominate the neighbor lists. Local Scaling and NICDM make the graph more balanced and are associated with improved decoding. The word associated is important: these interventions change the full distance structure, so the current evidence does not isolate hubness as the only causal mechanism behind the accuracy gain.

The current Mutual Proximity implementation should be treated cautiously. It sometimes improves raw accuracy, but balanced accuracy stays weak and hubness metrics remain poor. For example, in MSDL-39 at k=50, Mutual Proximity has raw accuracy 0.581 +/- 0.022 but balanced accuracy only 0.500 +/- 0.001. Appendix C shows that part of this failure appears to come from an ordering issue in the current implementation; this is not a general claim that the Mutual Proximity method itself is invalid.

The mathematical explanation and a direct diagnostic for this behavior are provided in Appendix C. The short version is that Local Scaling and NICDM normalize local density, while the implemented Mutual Proximity score behaves like a proximity/tail-probability score whose nearest-neighbor ordering is inverted relative to an ordinary distance.

H2: Generative Fidelity

0.7 Core Question

If generated samples perform poorly under neighborhood-based evaluation, are they necessarily bad samples?

Not necessarily. The generated connectomes can preserve some aspects of global structure and reproducibility while still failing more disease-specific tests. This is the main conceptual distinction in H2, and the strength of the global result depends strongly on the representation.

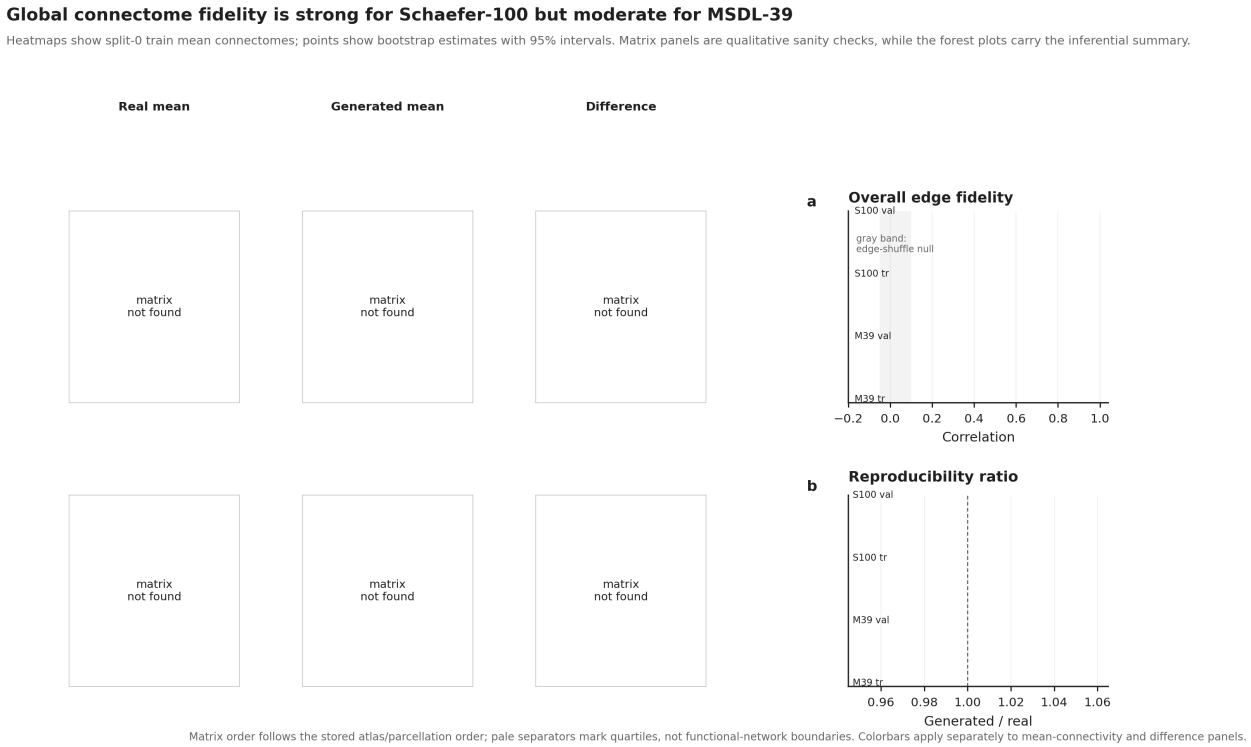

Figure 3: Figure 3. Global connectome fidelity is strong for Schaefer-100 but moderate for MSDL-39. Split-0 mean-connectome heatmaps show real, generated, and difference maps; the difference maps should be read with their explicit color scale rather than by visual contrast alone. Forest plots summarize overall edge correlation and reliability ratio with bootstrap 95% intervals and reference bands. These mean-connectome displays are qualitative sanity checks, not stand-alone inferential evidence; class-wise values are reported in Appendix F.

The H2 uncertainty and null-reference checks are now part of the completed analysis rather than only a planned workflow. The figures show the main pattern; exact bootstrap intervals, matched null intervals, and empirical null p-values are reported in Appendix F to keep the main text focused.

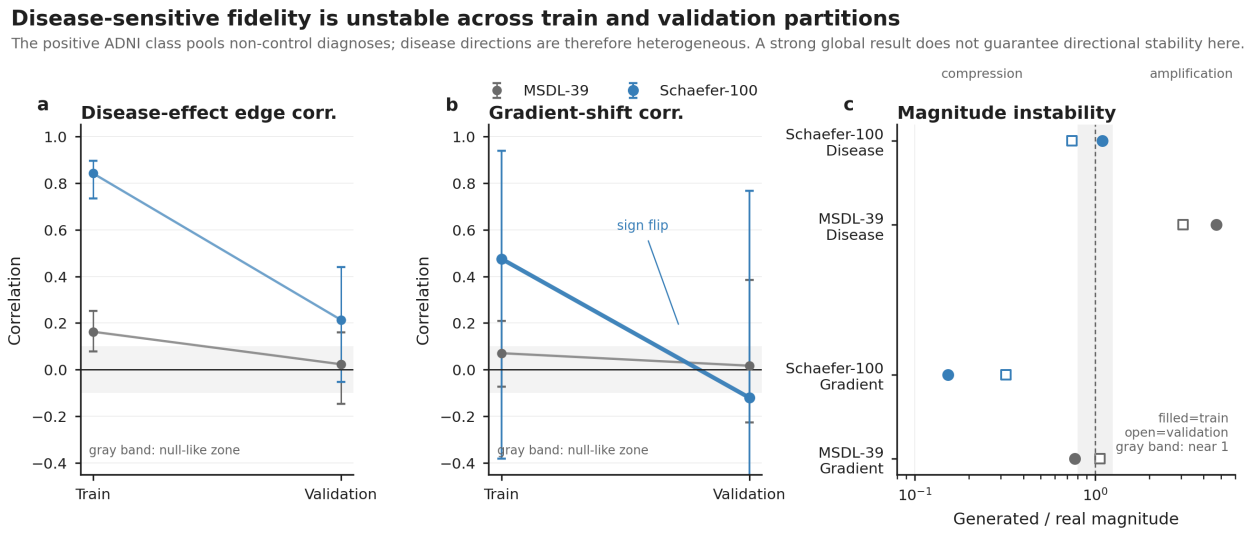

Figure 4: Figure 4. Disease-sensitive fidelity is unstable across train and validation partitions. Paired train-validation displays ask a specific question: does a representation that looks globally plausible also preserve the direction and magnitude of the class-1 minus class-0 disease shift? The disease-effect and gradient-shift panels highlight directional instability, while the log-scale magnitude panel shows compression and amplification relative to the generated/real reference value of 1. Success here requires both directionally consistent and magnitude-stable disease-sensitive structure.

0.8 Visual Interpretation of Generated Connectomes

The next three figures are designed to make the H2 result easier to inspect visually. They do not replace the quantitative scorecard; instead, they show why the scorecard needs separate layers. The same generated samples can look well organized in a global representation space, share recognizable atlas-level connectivity structure, and still fail a more delicate disease-shift test. Throughout this section, the figures are descriptive or qualitative unless explicitly tied back to bootstrap or null-based summaries.

0.9 Global Structure and Reproducibility

Schaefer-100 shows strong global fidelity. The generated and real mean connectomes are highly correlated in both train and validation partitions (0.995 and 0.981). Bootstrap reproducibility is also close to the real-data reference, with reliability ratios near 1. This means the generator is not producing random or unstable samples. It is learning a stable global connectome pattern for this representation.

MSDL-39 is only moderate. Its validation overall edge correlation is 0.418, which is not trivial but is far below Schaefer-100. This shows that generative fidelity depends on representation choice, so the report should not claim uniformly strong global preservation.

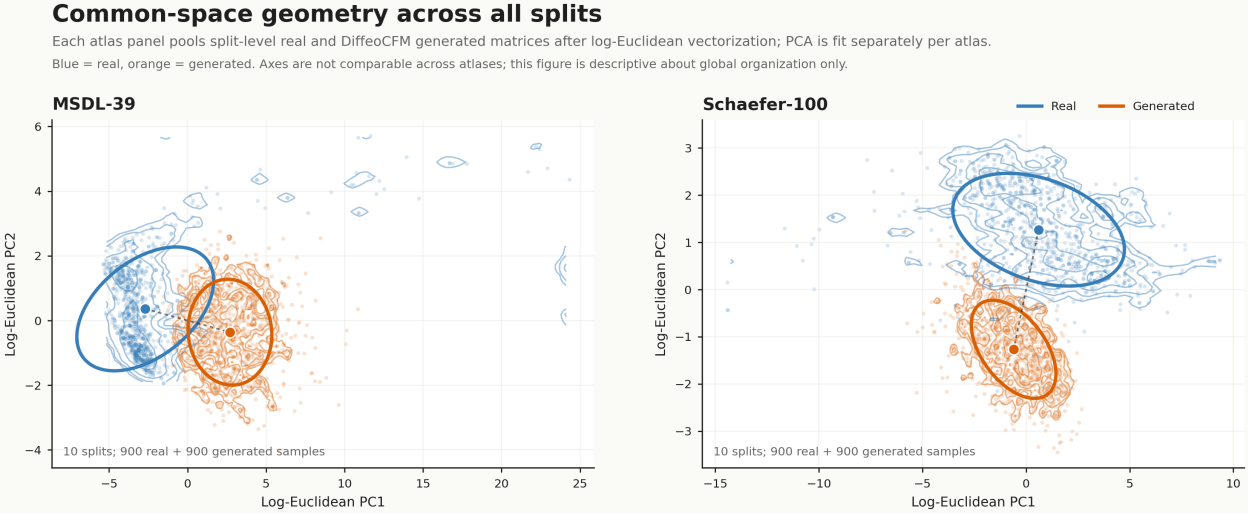

Figure 5: Figure 5. Common-space geometry across all splits. Each point is one real or generated matrix after log-Euclidean vectorization; train and validation samples are pooled within atlas, and PCA is fit separately for MSDL-39 and Schaefer-100, so the axes should be read within, not across, panels. Blue points are real matrices and orange points are generated matrices; density contours and ellipses summarize the two empirical clouds. The separation between real and generated clouds shows that the generator does not perfectly reproduce the full sample distribution, while the compact and structured generated cloud shows that the samples are not random. This visualizes global organization only; it does not establish disease-sensitive fidelity.

The edge-shuffled null makes this distinction clearer. For both atlases, the observed overall edge correlation is far outside the edge-shuffled null distribution, but Schaefer-100 is much farther from null than MSDL-39. This supports the claim that global structure preservation is not a trivial high-dimensional correlation artifact while still showing representation dependence.

Reliability ratios near 1 are useful but not sufficient. They indicate that generated class-level averages have similar bootstrap stability to real averages. They do not rule out oversmoothing or mode collapse; a generator can be stable because it captures reproducible structure, but it can also be stable because it produces overly similar samples. That is why reliability is interpreted together with edge correlations, class-wise metrics, and disease-sensitive shifts.

0.10 Disease-Specific Shift Fidelity

Disease-sensitive structure is much harder to reproduce than global structure. In Schaefer-100, disease-effect edge correlation is high in train (0.913) but much weaker in validation (0.317). The gradient-shift result is more concerning: it is positive in train (0.519) but negative in validation (-0.265). That sign change means the disease-specific gradient direction is not stable across partitions.

The coarse ADNI label likely contributes to this instability. Class 1 is not one disease stage; it pools MCI, SMC, and AD into one positive class, with original sample counts of 847, 133, and 180 respectively. SMC is especially difficult because it represents subjective memory complaint

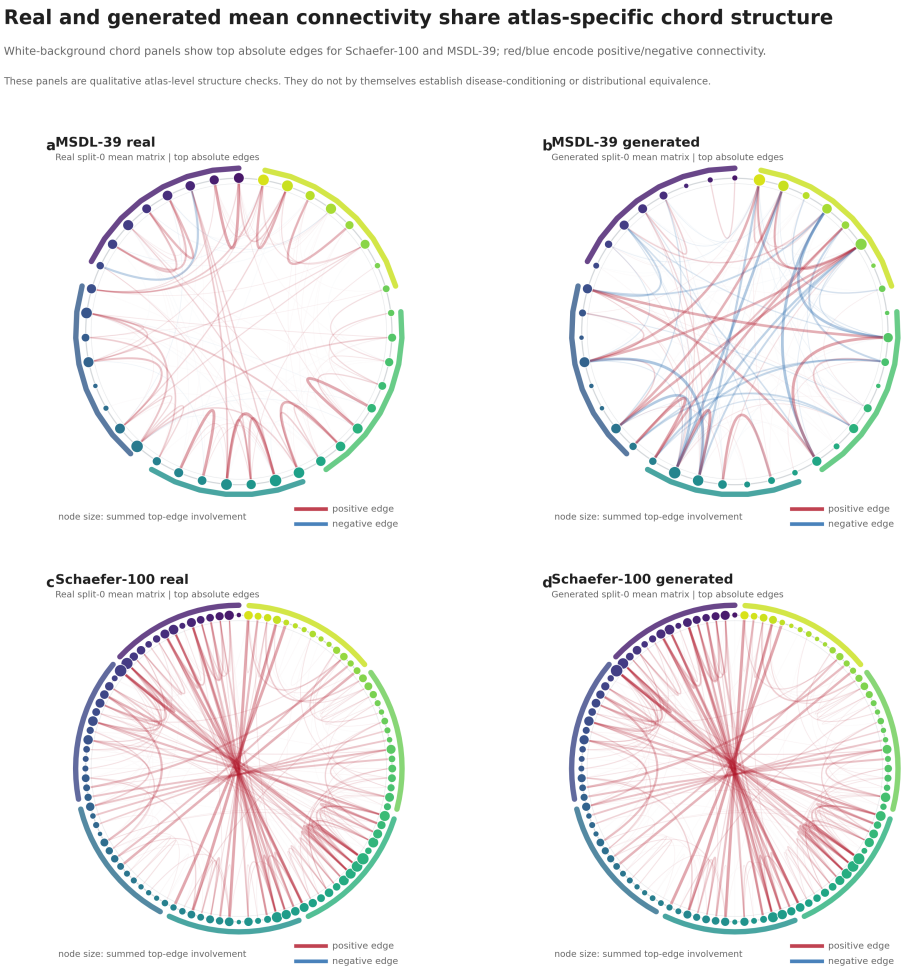

Figure 6: Figure 6. Mean-connectivity chord diagrams for real and generated matrices. The top row shows MSDL-39 and the bottom row shows Schaefer-100; within each row, the left panel is the real split-0 mean matrix and the right panel is the generated split-0 mean matrix. Curved chords show the strongest absolute connectivity edges, red indicates positive connectivity, blue indicates negative connectivity, and node size reflects how strongly a region participates in the displayed top edges. The purpose is explicitly qualitative: these mean-matrix summaries can reveal broad atlas-specific organization, but they cannot by themselves establish conditional fidelity, disease sensitivity, or distributional equivalence.

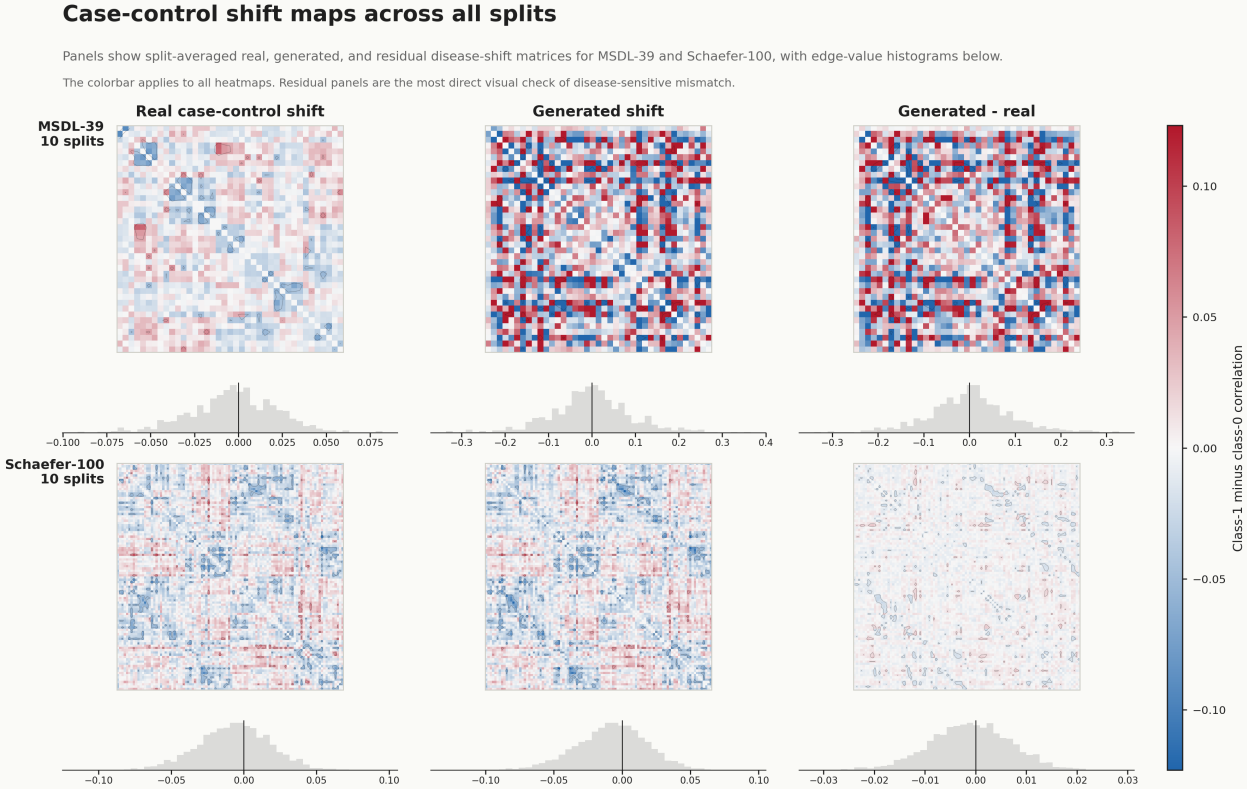

Figure 7: Figure 7. Case-control shift maps across all splits. Rows compare MSDL-39 and Schaefer-100; columns show the split-averaged real disease shift, generated disease shift, and generated-minus-real residual. The heatmaps show class-1 minus class-0 connectivity effects, where class 0 is healthy/control and class 1 is pooled non-control. Histograms below each heatmap show the edge-value distribution. The residual panels explain why disease-sensitive fidelity is harder than global fidelity: generated matrices can reproduce broad mean-connectome structure while still over- or under-shooting class-related edge shifts. Because the positive class pools multiple non-control diagnoses, these disease-shift maps should be interpreted as a coarse abnormality contrast rather than a clean CN-vs-AD disease map.

rather than a sharply separated imaging phenotype, while AD is comparatively small in this local benchmark. If those subgroups differ in severity, progression, or connectome direction, their average class effect can be weak, internally inconsistent, or split-dependent. In that setting, a generated sample can preserve global connectome structure while still failing the disease-effect vector or gradient-shift metric, because the target disease direction itself is a mixture of multiple clinical directions.

The norm ratios add another warning, but not a one-directional one. Schaefer-100 shows strong gradient-shift compression (0.153 in train and 0.321 in validation). MSDL-39 is different: some disease-related ratios are inflated rather than compressed. The safer interpretation is magnitude instability: generated disease-related effects can be too small or too large depending on representation and partition.

The label-shuffled gradient null is also informative. The observed validation gradient-shift correlation for Schaefer-100 is negative and should not be interpreted as stable preservation; it is a

sign-instability warning. In contrast, high global edge fidelity can coexist with weak or unstable disease-shift fidelity.

0.11 H2 Interpretation

Generated connectomes should not be judged with a single score. They can be useful for one purpose and weak for another. In these experiments, the generator is strong for global Schaefer-100 structure, moderate for MSDL-39 global structure, and unreliable for preserving disease-specific gradient shifts under the current control-vs-pooled-non-control label. The Schaefer-100 versus MSDL discrepancy should not be oversold as a principled victory for one atlas. It is consistent with representation dependence, but it may also reflect unmodeled atlas-specific preprocessing, dimensionality, or generator-behavior differences that this artifact bundle cannot fully disentangle. The safest reading is therefore atlas-dependent fidelity, not a universal conclusion about connectome generation quality.

For engineers, this means generated samples may still be useful for pretraining, stress testing, or population-level simulation. For statisticians, this means global correlation is not enough: disease-effect vectors, gradient-shift correlations, and norm ratios need to be reported separately.

Joint Interpretation

H1 and H2 answer different parts of the same review question. H1 asks whether neighborhood-based readouts are interpretable at all. H2 asks what remains defensible about generated matrices once evaluation is separated into global and disease-sensitive layers. The practical synthesis is straightforward:

- good global fidelity does not guarantee good disease-shift fidelity;
- good hubness correction does not imply that the generator learned pathology;
- weak baseline kNN alone is not enough, but the five-phase evidence chain makes the H1 pattern interpretable;
- once that validity check is in place, H2 shows that global plausibility and disease-sensitive fidelity diverge.

The main lesson is therefore a scorecard one: high-dimensional SPD/connectome evaluation should separate neighborhood health, prediction utility, global structure, and disease-sensitive shift fidelity rather than collapsing them into one headline score.

Reviewer-Facing Summary

The strongest response to the main critique is not that every result is now strong. It is that the report now makes the logic, scope, and evidence grading explicit. The kept claims are narrow and auditable: H1 is a subject-disjoint neighborhood-validity result with completed null and reference checks, and H2 is a layered fidelity result showing that global structure can be preserved while disease-sensitive structure remains unstable. The downgraded claims are also explicit: atlas superiority, disease-conditioning success, visually compelling but non-inferential figures, and the Schaefer-100 `k=50` probe are no longer treated as headline evidence. That shift makes the report weaker in rhetoric but stronger in review defensibility.

Engineering Recommendations

- If evaluation depends on kNN, retrieval, or nearest-neighbor realism, report hubness diagnostics before interpreting downstream scores.
- If the raw neighbor graph has high k-occurrence skewness, many antihubs, or low reciprocity, compare raw neighborhoods with Local Scaling or NICDM before making representation-quality claims.
- If Mutual Proximity is retained, treat the current result as an implementation-sensitive diagnostic and report the ordering check; do not use it as evidence that the method family fails.
- If generated connectomes are evaluated, report global edge correlation, bootstrap reproducibility, disease-effect vectors, gradient-shift metrics, and matched nulls as separate endpoints.
- If a claim requires atlas or disease-label generality, rerun the benchmark with additional atlases, connectivity estimators, and cleaner diagnosis contrasts before promoting the conclusion beyond this scorecard.

Summary and Discussion

The main manuscript-level conclusion is narrow but defensible. In the audited ADNI setting, Local Scaling and NICDM are associated with healthier neighborhood geometry and better subject-disjoint kNN decoding, which makes H1 interpretable as a neighborhood-validity result rather than as a weak benchmark anecdote. In H2, generated connectomes can preserve global structure while failing to preserve disease-sensitive directions, so global plausibility and disease-sensitive fidelity should not be collapsed into one headline metric.

That conclusion is intentionally bounded. The retained claims concern split-audited neighborhood evaluation, non-trivial global fidelity, and the divergence between global and disease-sensitive structure. Atlas superiority, the Schaefer-100 `k=50` probe, and visually compelling diagnostics remain exploratory. The practical product of the current study is therefore a review-ready scorecard rather than a strong disease-modeling claim.

Scientific Review Integration

The report now clears the main H1 review blocker. Possible train/validation leakage has been addressed by the subject-ID split audit, and the neighborhood-intervention result is now backed by paired intervention tests, label-permutation nulls, non-kNN references, and SPD QC. The strengthened SPD QC audit found minimum eigenvalue 0.000042 across the stored ADNI matrices, with 0 non-positive eigenvalue counts, 0 values below 1e-8, and 0 matrix-load/eigendecomposition failures. Feature extraction uses an explicit numerical safety rule: eigenvalues are clipped to at least 1e-8 before the matrix logarithm.

H2 remains narrower by design. The retained claim is that generated connectomes can preserve non-trivial global structure without reliably preserving disease-sensitive directions under the current pooled label. The missing ingredient for a stronger disease-specific manuscript is still phenotype resolution, documented in Appendix G as `h2\_cleaner\_label\_rerun\_status.md`.

The Brain Researcher prod MCP report-generation run is useful as a handoff check, but not as an independent prod-side review of this local filesystem. A direct prod `scientific\_report\_generate` call against the local `results/brain\_researcher\_autoresearch` directory failed because the deployed service could not read `/home/zijiaochen`; the successful prod run rendered an analysis-only handoff from explicit local summaries. The maintained source of truth for this revision is therefore the local report wrapper and its copied evidence artifacts.

Forward-Looking Implications

The most valuable next step is a phenotype-clean rerun of H2 with CN-vs-AD, CN-vs-MCI, and severity-aware contrasts. If disease-shift fidelity improves there, the current negative result is largely a label-heterogeneity warning; if it does not, the limitation is more likely in the generator or conditioning scheme.

The broader benchmark lesson is methodological. Future releases should extend the current audit package to more atlases, more connectivity estimators, and the same split/uncertainty checks used here. Hubness correction should also be treated as part of the evaluation stack, because any method that depends on neighborhoods is implicitly depending on the statistical quality of the neighbor graph.

Limitations

The workspace contains only `ADNI MSDL-39` and `ADNI Schaefer-100`, so the report shows representation dependence but not broad atlas invariance.

The disease label is a pooled control-vs-non-control contrast. That is acceptable for a coarse abnormality screen but weak for disease-shift fidelity, because pooled MCI, SMC, and AD can dilute or rotate the target effect direction.

The archives contain derived matrices rather than the original subject-level time series. That means the split audit and SPD QC can be verified directly, but upstream connectome-construction choices cannot be rerun inside this bundle.

The Schaefer-100 `k=50` row is a bounded raw-edge probe, and the EEG appendix is supporting evidence only. Neither should be read as a matched primary replication of the main ADNI result.

Conclusion

In the audited ADNI setting, hubness-aware local-density corrections are associated with healthier neighborhood geometry and better subject-disjoint kNN decoding, while generated connectomes preserve global structure more reliably than disease-sensitive structure. The most defensible output is therefore a review-ready scorecard separating neighborhood validity, decoding utility, global fidelity, reproducibility, and disease-shift fidelity rather than a stronger disease-modeling claim.

References

- Collas, A., Ju, C., Salvy, N., & Thirion, B. (2025). Riemannian flow matching for brain connectivity matrices via pullback geometry. In The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025).
- Salvy, N., Talbot, H., & Thirion, B. (2026). GICDM: Mitigating hubness for reliable distance-based generative model evaluation. arXiv:2602.16449.
- Radovanovic, M., Nanopoulos, A., & Ivanovic, M. (2010). Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research, 11, 2487-2531.
- Schnitzer, D., Flexer, A., Schedl, M., & Widmer, G. (2012). Local and global scaling reduce hubs in space. Journal of Machine Learning Research, 13, 2871-2902.
- Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage, 180(Part A), 68-77.
- Varoquaux, G., Raamana, P. R., Engemann, D. A., Hoyos-Idrobo, A., Schwartz, Y., & Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage, 145(Part B), 166-179.
- Eitel, F., Schulz, M.-A., Seiler, M., Walter, H., & Ritter, K. (2021). Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. Experimental Neurology, 339, 113608.

Appendix A: Terminology and Method Details

0.12 Matrix and representation terms

- **Connectome**: a matrix summarizing functional relationships between brain regions. In this report, the stored matrices are fMRI connectivity-like matrices.
- **SPD matrix**: a symmetric positive definite matrix. Covariance and regularized correlation matrices are commonly treated as SPD objects because their geometry is better handled on the SPD manifold than by plain matrix flattening.
- **Atlas** / **parcellation**: a definition of brain regions. `MSDL-39` has 39 regions; `Schaefer-100` has 100 regions. More regions usually mean a higher-dimensional representation.
- **Log-Euclidean vector representation**: a geometry-aware way to convert SPD matrices into vectors. Each matrix is mapped through a matrix logarithm and represented in a vector space where standard distances and classifiers can be applied. This representation assumes positive eigenvalues; the current report relies on the stored matrices satisfying that prerequisite.
- **Raw edge-vector representation**: a simpler vectorization of a connectivity matrix. Because the matrix is symmetric, only the upper-triangle entries are kept. This avoids duplicated edges and gives one feature vector per subject or sample.
- **Raw upper-triangle vector probe**: the bounded Schaefer-100 `k=50` experiment. It uses raw upper-triangle edge vectors instead of the full log-Euclidean vector representation because the full Schaefer-100 log-Euclidean `k=50` run was too slow to checkpoint.
- **@@PLACEHOLDER0@@**: a computational cap used only in the bounded Schaefer-100 `k=50` probe. For each split, at most 500 training samples were used to build the neighbor graph. This keeps the large-neighborhood experiment tractable. It should not be interpreted as the full-data Schaefer-100 result.

0.13 Hubness terms

- **Nearest-neighbor graph**: the graph formed by connecting each sample to its `k` closest training samples.
- **Hub**: a sample that appears in many other samples' nearest-neighbor lists.

- **Antihub**: a sample that almost never appears in any nearest-neighbor list.
- **Hubness**: the overall imbalance where a few samples dominate neighbor lists and many samples are rarely selected.
- **k-occurrence**: for each sample, the number of times it appears in other samples' `k` nearest-neighbor lists.
- **k-occurrence skewness**: the skewness of the k-occurrence distribution. Larger values indicate stronger hubness.
- **Antihub fraction**: the fraction of samples that never appear as a neighbor.
- **Mutual-neighbor fraction**: the fraction of neighbor relationships that are reciprocal. Higher values indicate a healthier and less one-sided neighbor graph.
- **Neighbor label purity**: the fraction of neighbors that share the same class label as the query sample.

0.14 Hubness intervention methods

- **Baseline**: the uncorrected neighbor graph, using ordinary distances in the chosen representation.
- **Local Scaling**: a distance correction that rescales distances by the local neighborhood radius around each point. Intuitively, it asks whether two points are close relative to their local density, not just close in absolute distance. This reduces the tendency for dense central points to become hubs.
- **NICDM**: a local-density normalization method. It divides pairwise distances by local neighborhood radii. Like Local Scaling, it aims to make distance comparisons fairer across dense and sparse regions of the representation.
- **Mutual Proximity**: a probabilistic distance correction. Instead of using raw distance directly, it estimates how surprising a distance is relative to each point's distance distribution. In this pipeline, the current Mutual Proximity implementation did not reliably improve balanced accuracy or hubness metrics.
- **Negative control**: a method included to test whether any correction helps automatically. The current Mutual Proximity implementation is useful here because it shows that improvement is specific to Local Scaling and NICDM in this implementation, not just caused by changing distances arbitrarily.

0.15 Prediction and evaluation terms

- **kNN decoding**: classification by majority vote among the `k` nearest training samples.
- **Balanced accuracy**: average recall across classes. It is preferred over raw accuracy when classes may be imbalanced.
- **@@PLACEHOLDER0@@**: train/reference samples are real data and test/query samples are also real data.
- **@@PLACEHOLDER0@@**: train/reference samples are generated data and test/query samples are real data. This is a train-on-synthetic, test-on-real style check.
- **Split**: one train/validation partition of the dataset. Reporting across splits gives a more stable estimate than one partition alone.

0.16 Generative fidelity terms

- **Global map fidelity**: whether the generated average connectome resembles the real average connectome.
- **Overall edge correlation**: correlation between real and generated mean connectome edge vectors.
- **Bootstrap reproducibility**: stability of class-level mean connectomes when the data are resampled.
- **Bootstrap Mean [95% CI]**: the mean of a metric over bootstrap resamples, followed by the 2.5th and 97.5th percentile bootstrap interval.
- **Matched Null 95%**: the 95% interval from a null experiment matched to the metric. In this report, edge-shuffled nulls are used for overall edge correlation and label-shuffled nulls are used for gradient-shift correlation.
- **Null p**: an empirical two-sided p-value from the matched null distribution. It asks how often the null produces a value at least as extreme as the observed value.
- **Reliability ratio**: generated bootstrap stability divided by real bootstrap stability. Values near 1 mean generated averages are about as stable as real averages, but this must be interpreted with fidelity metrics because excessive stability can also reflect oversmoothing.
- **Disease-effect vector**: the class-1 mean connectome minus the class-0 mean connectome. In the current ADNI benchmark, class 0 is healthy control and class 1 is pooled non-control, so this vector is a broad abnormality contrast rather than a diagnosis-specific AD vector.

- **Gradient shift**: a low-dimensional summary of how class-related or disease-related organization moves along a connectome gradient.
- **Gradient-shift correlation**: agreement between real and generated disease-related gradient directions.
- **Gradient-shift norm ratio**: generated gradient-shift magnitude divided by real gradient-shift magnitude. Values below 1 indicate compression; values above 1 indicate amplification.

Appendix B: EEG Covariance-Matrix Hubness Check

This appendix adds back the EEG result that was used during the exploratory phase. The dataset is `BNCI2014\_001`, a motor-imagery EEG benchmark with 9 subjects and 4 classes. Each trial was represented as an OAS-regularized covariance matrix and then mapped to tangent-space features before hubness analysis.

The purpose of this appendix is narrow: it checks whether the same local-neighborhood issue appears in EEG covariance representations. It does not change the main fMRI conclusions because the EEG artifact was archived as averaged summary rows, not as the full split-level statistical package used for the ADNI tables.

Table 3: Table 3. EEG covariance-matrix hubness check on BNCI2014_001.

Dataset	Scenario	Method	k	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Neighbor Label Purity
BNCI2014_001	Bidirectional cross-session	Baseline	10	0.513	2.332	0.084	0.323	0.423
BNCI2014_001	Bidirectional cross-session	Local Scaling	10	0.530	1.129	0.006	0.598	0.437
BNCI2014_001	Bidirectional cross-session	Mutual Proximity	10	0.217	12.366	0.827	0.028	0.227
BNCI2014_001	Bidirectional cross-session	NICDM	10	0.535	1.178	0.005	0.598	0.437
BNCI2014_001	Bidirectional cross-session	Baseline	20	0.534	1.886	0.029	0.408	0.403
BNCI2014_001	Bidirectional cross-session	Local Scaling	20	0.547	1.038	0.002	0.669	0.418
BNCI2014_001	Bidirectional cross-session	Mutual Proximity	20	0.199	8.512	0.718	0.045	0.223
BNCI2014_001	Bidirectional cross-session	NICDM	20	0.547	1.082	0.001	0.663	0.417
BNCI2014_001	Within-session pooled 5-fold CV	Baseline	10	0.563	2.281	0.085	0.323	0.419
BNCI2014_001	Within-session pooled 5-fold CV	Local Scaling	10	0.580	1.078	0.005	0.601	0.433
BNCI2014_001	Within-session pooled 5-fold CV	Mutual Proximity	10	0.220	16.286	0.871	0.019	0.234
BNCI2014_001	Within-session pooled 5-fold CV	NICDM	10	0.579	1.156	0.005	0.600	0.431
BNCI2014_001	Within-session pooled 5-fold CV	Baseline	20	0.577	1.908	0.029	0.408	0.402
BNCI2014_001	Within-session pooled 5-fold CV	Local Scaling	20	0.603	1.014	0.002	0.673	0.415
BNCI2014_001	Within-session pooled 5-fold CV	Mutual Proximity	20	0.197	10.921	0.783	0.032	0.226
BNCI2014_001	Within-session pooled 5-fold CV	NICDM	20	0.601	1.074	0.001	0.667	0.414

The EEG pattern is consistent with H1 but weaker than the fMRI connectome pattern. At `k=10`, the within-CV baseline balanced accuracy is `0.5631`; Local Scaling improves it to `0.5804`, and NICDM reaches `0.5785`. The same interventions also reduce k-occurrence skewness from `2.2811` to

roughly `1.1` and almost remove antihubs. Cross-session decoding shows a similar but smaller gain: baseline balanced accuracy is `0.5129`, Local Scaling reaches `0.5295`, and NICDM reaches `0.5345`.

This supports the broader claim that hubness is not only an fMRI-connectome artifact. It can also appear in EEG covariance/tangent-space representations. However, the EEG result should remain in the appendix because it lacks the newer uncertainty reporting requirements: no bootstrap confidence intervals, no label-permutation significance test, and no paired test comparing NICDM or Local Scaling against the baseline.

The current Mutual Proximity implementation again behaves as a useful negative control. In the EEG table, it sharply worsens balanced accuracy and increases hubness, so the result reinforces a methodological point: changing distances is not automatically beneficial. This is a statement about the current implementation and data geometry, not a general rejection of Mutual Proximity as a method.

Appendix C: Why Local Scaling and NICDM Work Here, but Mutual Proximity Does Not

The three interventions all try to repair high-dimensional nearest-neighbor geometry, but they make different mathematical assumptions.

Let `d(i,j)` be the distance between samples `i` and `j`. In a raw high-dimensional space, samples in dense central regions tend to have smaller distances to many other samples, so they become hubs. Local Scaling and NICDM directly correct for this local density effect.

Local Scaling replaces the raw distance with a local-density-aware score:

```
d\_LS(i,j)=1-exp(-d(i,j)\textasciicircum{}2/(sigma\_isigma\_j))
```

where `sigma\_i` is a local radius around sample `i`, typically the distance to a nearby neighbor. A pair is considered close only if it is close relative to both samples' local neighborhoods. This is well matched to hubness caused by density imbalance.

NICDM uses a similar idea:

```
d\_NICDM(i,j)=d(i,j)/sqrt(r\_ir\_j)
```

where `r\_i` is the average distance from sample `i` to its local neighbors. Again, the correction divides out local scale. If a point is a hub simply because it sits in a dense central area, its distances are no longer unfairly small after normalization.

Mutual Proximity is different. It converts a distance into a probability relative to each sample's full distance distribution. A common form asks whether the observed distance is surprisingly small compared with the distances usually seen from both endpoints:

```
MP(i,j)=P(D\_i>d(i,j))P(D\_j>d(i,j))
```

This can work when each sample's distance distribution is well modeled and when the resulting proximity score is converted into the correct ordering for nearest-neighbor search. In this experiment, those conditions are weak.

0.17 Direct Diagnostic

We directly checked the three proposed failure modes on `ADNI MSDL-39`, split 0, `real\_to\_real`, `k=10`. The diagnostic artifacts are stored as `mutual\_proximity\_diagnostic\_adni\_msdl\`

`_split0.csv`, `mutual\_proximity\_rank\_diagnostic\_adni\_msdl\_split0.csv`, and `distance\_distribution\_shape\_adni\_msdl\_split0.csv`.

Table 4: Table 4. Mutual Proximity diagnostic summary for ADNI MSDL-39 split 0.

Method	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Count-Radius Corr
Baseline	0.5343	10.4500	0.5901	0.0602	-0.2742
Local Scaling	0.5617	3.6618	0.0332	0.4676	-0.2103
NICDM	0.5652	3.1428	0.0256	0.4850	-0.1488
Mutual Proximity	0.5267	14.2635	0.9094	0.0129	0.4532
Mutual Proximity, inverted order	0.5091	7.7065	0.0773	0.2707	-0.3917

The first result supports the local-density interpretation. Local Scaling and NICDM reduce hubness and improve balanced accuracy. The correlation between k-occurrence count and local raw-distance radius also moves toward zero, especially for NICDM, which means neighbor popularity is less tied to local density after correction.

The second result directly verifies the ordering problem for Mutual Proximity:

Table 5: Table 5. Rank correlation between corrected scores and raw distances.

Method	Mean Rank Correlation with Raw Distance
Local Scaling	0.5781
NICDM	0.5427
Mutual Proximity	-0.9325
Mutual Proximity, inverted order	0.9325

The current Mutual Proximity ordering has rank correlation `-0.9325` with the raw distance ranking. That is strong evidence that the score is being consumed in the wrong direction for nearest-neighbor search. Reversing the order changes the rank correlation to `0.9325`, reduces antihubs from `0.9094` to `0.0773`, and improves the graph relative to the current Mutual Proximity implementation.

However, the inverted-order Mutual Proximity still does not match Local Scaling or NICDM in balanced accuracy. This means the failure is not only a coding convention. Local density normalization remains better matched to this representation.

Finally, the row-wise distance distributions are visibly non-Gaussian:

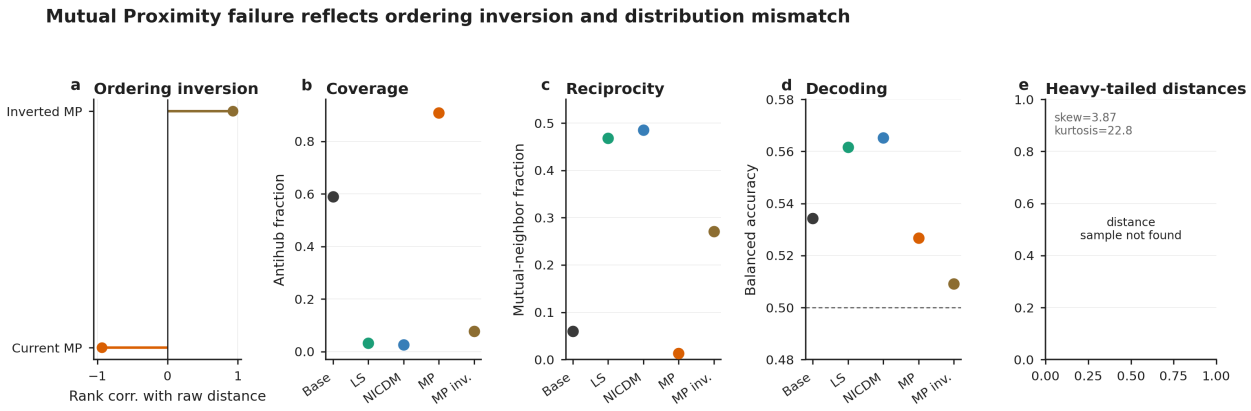

Figure 8: Figure 8. Mutual Proximity failure reflects ordering inversion and distribution mismatch. The signed lollipop plot shows the current MP score is ordered opposite to raw distance; graph-health and decoding panels show that inversion improves coverage but still does not match Local Scaling or NICDM, and the heavy-tailed distance summary explains why the Gaussian-tail approximation is fragile here.

Table 6: Table 6. Distance-distribution shape diagnostic for ADNI MSDL-39 split 0.

Dataset	Split	Rows Checked	Mean Distance Skewness	Mean Excess Kurtosis
ADNI MSDL-39	0	200	3.8704	22.7972

The average distance distribution has high positive skewness and very high excess kurtosis. That weakens the simple distributional approximation behind this Mutual Proximity implementation and explains why the method is less stable here than Local Scaling or NICDM.

Overall, the direct diagnostic supports a two-part explanation: the current Mutual Proximity implementation has a proximity/distance ordering problem, and even after reversing the order, global distribution-tail normalization is less effective than local-density normalization for these SPD/connectome embeddings.

Appendix D: Metric Definitions

0.18 H1 metrics

- **Accuracy**: fraction of correctly predicted samples.
- **Balanced Accuracy**: average recall across classes. This is more informative than raw accuracy when classes are imbalanced.
- **k-Occurrence**: how often each sample appears in other samples' `k` nearest-neighbor lists.
- **k-Occurrence Skewness**: how unequal the k-occurrence distribution is. Larger values mean stronger hubness.
- **Robinhood Index**: another inequality measure for neighbor occurrence. Larger values mean a more uneven neighbor graph.
- **Antihub Fraction**: fraction of samples that never appear in any neighbor list.
- **Top 1% Occurrence Share**: how much of the neighbor graph is controlled by the most frequent hubs.
- **Mutual Neighbor Fraction**: fraction of neighbor links that are reciprocal.
- **Neighbor Label Purity**: fraction of neighbors that share the query sample's class label.

0.19 H2 metrics

- **Overall Edge Correlation**: correlation between real and generated mean connectomes.
- **Leading Eigenvector Cosine**: similarity between dominant spatial modes of the real and generated mean matrices.
- **Class-wise Edge Correlation**: same as edge correlation, but computed separately within each class.
- **Disease Effect Vector**: class-1 mean connectome minus class-0 mean connectome.
- **Disease Effect Edge Correlation**: correlation between real and generated disease-effect vectors.
- **Disease Effect Norm Ratio**: generated disease-effect magnitude divided by real disease-effect magnitude.

- **Bootstrap Mean Correlation**: stability of class-level mean connectomes under resampling.
- **Bootstrap Reliability Ratio**: generated bootstrap stability divided by real bootstrap stability.
- **Gradient Shift Correlation**: agreement between real and generated disease-related gradient displacement.
- **Gradient Shift Norm Ratio**: generated gradient-shift magnitude divided by real gradient-shift magnitude.

Appendix E: Statistical Reporting Framework

Table 7: Table 7. Recommended uncertainty and null-model reporting by metric family.

Metric Family	Preferred Uncertainty Summary	Recommended Null / Reference	Interpretive Role
Classification metrics	Split-wise mean +/- SD; bootstrap CI for Balanced Accuracy	Label permutation within split	Distinguishes genuine predictive structure from chance decoding
Hubness metrics	Split-wise mean +/- SD; bootstrap CI over samples	Distance-preserving randomization or isotropic null cloud	Tests whether hub concentration exceeds geometric baseline
Edge-correlation metrics	Bootstrap CI over subjects or edges	Edge-shuffled null and class-label null	Separates structural agreement from accidental high-dimensional correlation
Norm-ratio metrics	Median and percentile interval; denominator magnitude reported explicitly	Scale-matched synthetic null	Prevents over-interpretation of unstable ratios
Gradient metrics	Split-wise CI and sign-stability frequency	Gradient-axis permutation / manifold coordinate null	Assesses whether directional disease effects are reproducible

For formal manuscript use, each main result should be reported with a point estimate, split-wise standard deviation, confidence interval, and matched null model. For decoding, the natural null is label permutation. For global structure, edge-shuffled and class-label-shuffled nulls are useful. For gradient shifts, sign stability across splits should be reported because unstable signs are scientifically meaningful.

The current release implements this framework for all available H1 k rows and for H2 global/gradient-shift uncertainty. The EEG appendix remains a supporting analysis rather than a fully audited primary endpoint.

Appendix F: Audit Tables

These tables provide the exact numeric values behind the main-text figures. They are kept in the appendix to avoid making the main report read like exported markdown tables.

0.20 H1 k=10: ADNI MSDL-39

Table 8: Table 8. H1 k=10 audit values for ADNI MSDL-39.

Method	Accuracy	Balanced Accuracy	k-Occurrence Skewness	Robinhood Index	Antihub Fraction	Top 1% Occurrence Share	Mutual Neighbor Fraction	Neighbor Label Purity
Baseline	0.497	0.520	9.369	0.846	0.643	0.591	0.034	0.514
Local Scaling	0.556	0.558	2.816	0.321	0.031	0.063	0.396	0.558
Mutual Proximity	0.546	0.504	14.229	0.972	0.904	0.983	0.011	0.508
NICDM	0.571	0.571	2.733	0.315	0.027	0.059	0.414	0.562

0.21 H1 k=10: ADNI Schaefer-100

Table 9: Table 9. H1 k=10 audit values for ADNI Schaefer-100.

Method	Accuracy	Balanced Accuracy	k-Occurrence Skewness	Robinhood Index	Antihub Fraction	Top 1% Occurrence Share	Mutual Neighbor Fraction	Neighbor Label Purity
Baseline	0.487	0.530	9.725	0.924	0.809	0.627	0.016	0.511
Local Scaling	0.611	0.615	3.936	0.420	0.089	0.094	0.273	0.581
Mutual Proximity	0.538	0.489	9.706	0.940	0.877	0.709	0.025	0.526
NICDM	0.610	0.615	4.234	0.417	0.083	0.093	0.286	0.583

0.22 H1 neighborhood-size sweep

Table 10: Table 10. H1 neighborhood-size sweep across ADNI representations.

Dataset	Method	k	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Protocol Note
ADNI MSDL-39	Baseline	5	0.524	11.925	0.717	0.032	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Local Scaling	5	0.574	3.337	0.097	0.335	Full log-Euclidean vector representation; 10 splits

Table 10: Table 10. H1 neighborhood-size sweep across ADNI representations.

Dataset	Method	k	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Protocol Note
ADNI MSDL-39	NICDM	5	0.572	3.250	0.084	0.354	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Baseline	10	0.520	9.369	0.643	0.034	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Local Scaling	10	0.558	2.816	0.031	0.396	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	NICDM	10	0.571	2.733	0.027	0.414	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Baseline	20	0.517	7.520	0.553	0.042	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Local Scaling	20	0.589	2.127	0.008	0.462	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	NICDM	20	0.588	2.167	0.005	0.477	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Baseline	50	0.505 +/- 0.032	5.634 +/- 0.078	0.377	0.069	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	Local Scaling	50	0.611 +/- 0.045	1.547 +/- 0.056	0.000	0.555	Full log-Euclidean vector representation; 10 splits
ADNI MSDL-39	NICDM	50	0.607 +/- 0.045	1.723 +/- 0.085	0.000	0.564	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	Baseline	5	0.554	10.576	0.847	0.018	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	Local Scaling	5	0.595	4.558	0.186	0.228	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	NICDM	5	0.615	5.067	0.177	0.242	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	Baseline	10	0.530	9.725	0.809	0.016	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	Local Scaling	10	0.615	3.936	0.089	0.273	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	NICDM	10	0.615	4.234	0.083	0.286	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	Baseline	20	0.513	8.086	0.763	0.021	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	Local Scaling	20	0.618	3.321	0.036	0.318	Full log-Euclidean vector representation; 10 splits
ADNI Schaefer-100	NICDM	20	0.623	3.566	0.035	0.335	Full log-Euclidean vector representation; 10 splits

Table 10: Table 10. H1 neighborhood-size sweep across ADNI representations.

Dataset	Method	k	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Protocol Note
ADNI Schaefer-100	Baseline	50	0.514 +/- 0.026	2.681 +/- 0.092	0.175	0.214	Bounded raw edge-vector probe; 500 training samples per split; 10 splits
ADNI Schaefer-100	Local Scaling	50	0.615 +/- 0.025	0.991 +/- 0.147	0.001	0.600	Bounded raw edge-vector probe; 500 training samples per split; 10 splits
ADNI Schaefer-100	NICDM	50	0.612 +/- 0.032	0.977 +/- 0.155	0.001	0.620	Bounded raw edge-vector probe; 500 training samples per split; 10 splits

0.23 H1 strengthened inference audit

Table 11: Table 11. H1 strengthened inference audit.

Dataset	k	Method	Balanced Accuracy Mean +/- SD	Bootstrap 95% CI	Delta vs Baseline	Paired p	Positive Splits	Mean / Worst Label-Permutation p
ADNI MSDL-39	5	Baseline	0.524 +/- 0.039	[0.500, 0.546]	+0.000	Baseline reference	10/10	0.342 / 0.920
ADNI MSDL-39	5	Local Scaling	0.574 +/- 0.033	[0.555, 0.593]	+0.050	0.002	10/10	0.065 / 0.270
ADNI MSDL-39	5	Mutual Proximity	0.496 +/- 0.007	[0.492, 0.500]	-0.027	0.064	2/10	0.638 / 0.930
ADNI MSDL-39	5	NICDM	0.572 +/- 0.040	[0.547, 0.594]	+0.048	0.004	9/10	0.106 / 0.450
ADNI MSDL-39	10	Baseline	0.520 +/- 0.037	[0.498, 0.542]	+0.000	Baseline reference	10/10	0.372 / 0.860
ADNI MSDL-39	10	Local Scaling	0.558 +/- 0.040	[0.534, 0.583]	+0.038	0.004	9/10	0.134 / 0.500
ADNI MSDL-39	10	Mutual Proximity	0.504 +/- 0.017	[0.495, 0.515]	-0.016	0.188	2/10	0.501 / 0.920
ADNI MSDL-39	10	NICDM	0.571 +/- 0.040	[0.549, 0.595]	+0.050	0.002	10/10	0.093 / 0.350
ADNI MSDL-39	20	Baseline	0.517 +/- 0.035	[0.498, 0.538]	+0.000	Baseline reference	10/10	0.377 / 0.840
ADNI MSDL-39	20	Local Scaling	0.589 +/- 0.046	[0.561, 0.616]	+0.071	0.002	10/10	0.054 / 0.260
ADNI MSDL-39	20	Mutual Proximity	0.501 +/- 0.004	[0.498, 0.503]	-0.016	0.180	3/10	0.452 / 0.830
ADNI MSDL-39	20	NICDM	0.588 +/- 0.042	[0.563, 0.613]	+0.071	0.002	10/10	0.055 / 0.300
ADNI MSDL-39	50	Baseline	0.505 +/- 0.032	[0.486, 0.522]	+0.000	Baseline reference	10/10	0.417 / 0.970
ADNI MSDL-39	50	Local Scaling	0.611 +/- 0.045	[0.585, 0.637]	+0.106	0.002	10/10	0.018 / 0.060
ADNI MSDL-39	50	Mutual Proximity	0.500 +/- 0.001	[0.499, 0.500]	-0.005	0.604	3/10	0.663 / 0.780
ADNI MSDL-39	50	NICDM	0.607 +/- 0.045	[0.580, 0.634]	+0.102	0.002	10/10	0.028 / 0.180
ADNI Schaefer-100	5	Baseline	0.554 +/- 0.044	[0.529, 0.580]	+0.000	Baseline reference	10/10	0.191 / 0.650
ADNI Schaefer-100	5	Local Scaling	0.595 +/- 0.029	[0.578, 0.612]	+0.041	0.031	7/10	0.029 / 0.090
ADNI Schaefer-100	5	Mutual Proximity	0.481 +/- 0.044	[0.455, 0.506]	-0.073	0.018	1/10	0.549 / 0.980
ADNI Schaefer-100	5	NICDM	0.615 +/- 0.031	[0.597, 0.633]	+0.062	0.012	8/10	0.021 / 0.090
ADNI Schaefer-100	10	Baseline	0.530 +/- 0.031	[0.510, 0.548]	+0.000	Baseline reference	10/10	0.287 / 0.790
ADNI Schaefer-100	10	Local Scaling	0.615 +/- 0.044	[0.590, 0.640]	+0.085	0.004	9/10	0.031 / 0.110
ADNI Schaefer-100	10	Mutual Proximity	0.489 +/- 0.034	[0.468, 0.507]	-0.041	0.016	1/10	0.547 / 0.980
ADNI Schaefer-100	10	NICDM	0.615 +/- 0.046	[0.587, 0.640]	+0.085	0.004	9/10	0.044 / 0.280
ADNI Schaefer-100	20	Baseline	0.513 +/- 0.036	[0.492, 0.534]	+0.000	Baseline reference	10/10	0.426 / 0.910
ADNI Schaefer-100	20	Local Scaling	0.618 +/- 0.042	[0.594, 0.642]	+0.105	0.004	9/10	0.027 / 0.130
ADNI Schaefer-100	20	Mutual Proximity	0.494 +/- 0.007	[0.489, 0.498]	-0.019	0.082	3/10	0.593 / 0.850

Table 11: Table 11. H1 strengthened inference audit.

Dataset	k	Method	Balanced Accuracy Mean +/- SD	Bootstrap 95% CI	Delta vs Baseline	Paired p	Positive Splits	Mean / Worst Label-Permutation p
ADNI Schaefer-100	20	NICDM	0.623 +/- 0.051	[0.591, 0.652]	+0.110	0.004	9/10	0.041 / 0.220
ADNI Schaefer-100	50	Baseline	0.514 +/- 0.026	[0.500, 0.530]	+0.000	Baseline reference	10/10	0.425 / 0.780
ADNI Schaefer-100	50	Local Scaling	0.615 +/- 0.025	[0.601, 0.630]	+0.102	0.002	10/10	0.029 / 0.100
ADNI Schaefer-100	50	NICDM	0.612 +/- 0.032	[0.592, 0.630]	+0.099	0.002	10/10	0.036 / 0.180

0.24 H1 non-kNN classifier reference

Table 12: Table 12. H1 non-kNN classifier reference.

Dataset	Classifier	Representation	Accuracy Mean +/- SD	Balanced Accuracy Mean +/- SD
ADNI MSDL-39	Linear SVM	logeuclidean tangent vectors	0.566 +/- 0.029	0.560 +/- 0.028
ADNI MSDL-39	Logistic Regression	logeuclidean tangent vectors	0.564 +/- 0.025	0.558 +/- 0.025
ADNI Schaefer-100	Linear SVM	logeuclidean tangent vectors	0.564 +/- 0.037	0.556 +/- 0.038
ADNI Schaefer-100	Logistic Regression	logeuclidean tangent vectors	0.577 +/- 0.036	0.571 +/- 0.037

0.25 H1 k=50: ADNI MSDL-39

Table 13: Table 13. H1 k=50 audit values for ADNI MSDL-39.

Method	Accuracy	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Neighbor Label Purity	Protocol Note
Baseline	0.455 +/- 0.029	0.505 +/- 0.032	5.634 +/- 0.078	0.377	0.069	0.504	Full log-Euclidean vector representation; 10 splits
Local Scaling	0.624 +/- 0.043	0.611 +/- 0.045	1.547 +/- 0.056	0.000	0.555	0.549	Full log-Euclidean vector representation; 10 splits
Mutual Proximity	0.581 +/- 0.022	0.500 +/- 0.001	6.269 +/- 0.012	0.719	0.045	0.518	Full log-Euclidean vector representation; 10 splits
NICDM	0.623 +/- 0.042	0.607 +/- 0.045	1.723 +/- 0.085	0.000	0.564	0.551	Full log-Euclidean vector representation; 10 splits

0.26 H1 k=50: ADNI Schaefer-100

Table 14: Table 14. H1 k=50 audit values for ADNI Schaefer-100.

Method	Accuracy	Balanced Accuracy	k-Occurrence Skewness	Antihub Fraction	Mutual Neighbor Fraction	Neighbor Label Purity	Protocol Note
Baseline	0.444 +/- 0.054	0.514 +/- 0.026	2.681 +/- 0.092	0.175	0.214	0.518	Bounded raw edge-vector probe; 500 training samples per split; 10 splits
Local Scaling	0.604 +/- 0.037	0.615 +/- 0.025	0.991 +/- 0.147	0.001	0.600	0.536	Bounded raw edge-vector probe; 500 training samples per split; 10 splits
Mutual Proximity	Not available	Not available	Not available	Not available	Not available	Not available	Bounded raw edge-vector probe; 500 training samples per split; 10 splits; method row not present in copied artifact
NICDM	0.604 +/- 0.042	0.612 +/- 0.032	0.977 +/- 0.155	0.001	0.620	0.537	Bounded raw edge-vector probe; 500 training samples per split; 10 splits

0.27 H2 global fidelity

Table 15: Table 15. H2 global fidelity summary.

Dataset	Partition	Overall Edge Corr	Leading Eigvec Cosine	Class 0 Edge Corr	Class 1 Edge Corr	Real Bootstrap Corr	Generated Bootstrap Corr	Reliability Ratio
ADNI MSDL-39	TRAIN	0.423	0.289	0.408	0.411	0.990	0.995	1.005
ADNI MSDL-39	VAL	0.418	0.286	0.394	0.397	0.921	0.960	1.042
ADNI Schaefer-100	TRAIN	0.995	0.949	0.994	0.995	0.995	0.997	1.002
ADNI Schaefer-100	VAL	0.981	0.891	0.962	0.972	0.956	0.973	1.018

0.28 H2 disease-sensitive fidelity

Table 16: Table 16. H2 disease-sensitive fidelity summary.

Dataset	Partition	Disease Effect Edge Corr	Disease Effect Norm Ratio	Class 0 Gradient Corr	Class 1 Gradient Corr	Gradient Shift Corr	Gradient Shift Cosine	Gradient Shift Norm Ratio
ADNI MSDL-39	TRAIN	0.175	4.703	0.028	0.402	0.071	0.083	0.770
ADNI MSDL-39	VAL	0.028	3.071	0.045	0.388	0.007	0.100	1.063
ADNI Schaefer-100	TRAIN	0.913	1.097	0.344	0.998	0.519	0.536	0.153
ADNI Schaefer-100	VAL	0.317	0.745	0.356	0.911	-0.265	0.472	0.321

0.29 H2 bootstrap and matched-null audit

Table 17: Table 17. H2 bootstrap and matched-null audit.

Dataset	Partition	Metric	Bootstrap Mean [95% CI]	Matched Null 95%	Null p
ADNI MSDL-39	TRAIN	Overall edge corr	0.422 [0.359, 0.494]	[-0.070, 0.062]	0.004
ADNI MSDL-39	TRAIN	Gradient-shift corr	0.071 [-0.073, 0.210]	[-0.323, 0.326]	0.777
ADNI MSDL-39	TRAIN	Reliability ratio	1.005 [1.002, 1.009]	Not applicable	Not applicable
ADNI MSDL-39	TRAIN	Gradient norm ratio	0.792 [0.366, 1.341]	Not applicable	Not applicable
ADNI MSDL-39	VAL	Overall edge corr	0.412 [0.325, 0.518]	[-0.074, 0.069]	0.004
ADNI MSDL-39	VAL	Gradient-shift corr	0.017 [-0.227, 0.384]	[-0.313, 0.322]	0.984
ADNI MSDL-39	VAL	Reliability ratio	1.041 [0.999, 1.097]	Not applicable	Not applicable
ADNI MSDL-39	VAL	Gradient norm ratio	1.026 [0.360, 2.755]	Not applicable	Not applicable
ADNI Schaefer-100	TRAIN	Overall edge corr	0.994 [0.991, 0.996]	[-0.026, 0.030]	0.004
ADNI Schaefer-100	TRAIN	Gradient-shift corr	0.476 [-0.382, 0.938]	[-0.681, 0.708]	0.335
ADNI Schaefer-100	TRAIN	Reliability ratio	1.002 [1.001, 1.004]	Not applicable	Not applicable
ADNI Schaefer-100	TRAIN	Gradient norm ratio	0.181 [0.065, 0.595]	Not applicable	Not applicable
ADNI Schaefer-100	VAL	Overall edge corr	0.972 [0.943, 0.982]	[-0.028, 0.029]	0.004
ADNI Schaefer-100	VAL	Gradient-shift corr	-0.120 [-0.789, 0.767]	[-0.758, 0.700]	0.622
ADNI Schaefer-100	VAL	Reliability ratio	1.017 [0.998, 1.041]	Not applicable	Not applicable
ADNI Schaefer-100	VAL	Gradient norm ratio	0.565 [0.096, 3.529]	Not applicable	Not applicable

Bootstrap Mean [95% CI] means the metric was recomputed over bootstrap resamples and summarized by its bootstrap mean and 2.5th/97.5th percentiles. **Matched Null 95%** is the corresponding null interval: edge-shuffled nulls for overall edge correlation and label-shuffled nulls for gradient-shift correlation. **Null p** is the empirical two-sided null probability of seeing a value at least as extreme as the observed metric; small values mean the observed structure is unlikely under that matched null. Ratio metrics do not use these shuffle nulls in the table because their natural reference is 1, not 0.

0.30 SPD QC audit

Table 18: Table 18. SPD QC audit summary.

Dataset	Matrices Audited	Matrix Dim	Minimum Eigenvalue	Non-Positive Count	< 1e-8 Count	Failure Count	Log-Safety Rule
ADNI MSDL-39	46860	39	0.000042	0	0	0	clip eigenvalues to >=1e-8 before matrix log
ADNI Schaefer-100	39940	100	0.000709	0	0	0	clip eigenvalues to >=1e-8 before matrix log

Appendix G: Deliverables

- Reusable experiment code under `high\_dim\_spd\_experiment/`
- Exploratory k-sweep code in `explore\_representation\_hubness\_grid.py`
- Canonical report directory: `results/brain\_researcher\_report/`
- Canonical artifacts directory: `results/brain\_researcher\_report/artifacts/`
- Final markdown report: `report.md`
- Final LaTeX wrapper: `report.tex`
- Local report style file: `scientific\_report.sty`
- Final PDF report: `report.pdf`
- Autoresearch wrapper mirror: `results/brain\_researcher\_autoresearch/final\_report.tex`
- Prod MCP report-generation handoff run: `br\_20260506\_071319\_5048410fc3` (analysis-only render; not a prod-side filesystem review)
- Mutual Proximity diagnostic summary: `mutual\_proximity\_diagnostic\_adni\_msdl\_split0.csv`
- Mutual Proximity rank diagnostic: `mutual\_proximity\_rank\_diagnostic\_adni\_msdl\_split0.csv`
- Distance-shape diagnostic: `distance\_distribution\_shape\_adni\_msdl\_split0.csv`
- H1 split audit: `h1\_subject\_disjoint\_manifest.csv`
- H1 strengthened inference summary: `h1\_inference\_summary.csv`
- H1 label-permutation audit: `h1\_label\_permutation\_full.csv`
- H1 non-kNN linear-reference summary: `h1\_linear\_reference\_summary.csv`
- H1 non-kNN linear-reference split metrics: `h1\_linear\_reference\_split\_metrics.csv`
- SPD QC summary: `spd\_qc\_summary.csv`
- H2 cleaner-label rerun status: `h2\_cleaner\_label\_rerun\_status.md`
- Report figures directory: `figures/`