CROSS-SECTIONAL ACCURACY DOES NOT CERTIFY WITHIN-PATIENT TRACKING IN FUNCTIONAL CONNECTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Resting-state functional connectivity supports two applications at once, classifying disease and identifying individuals from their connectivity patterns. A profile carries two kinds of feature, those that stay fixed within a person and differ between people, and those that change within a person from visit to visit. Which kind carries the separable disease information has not been asked, yet the answer decides what a connectivity marker can be used for. We first localise that information to the fixed component, which reads the diagnosis at 0.728 against 0.734 from the whole profile, and the localisation holds under seven alternative removal methods. We then show that centring a profile on its subject mean cancels the class-mean difference exactly whenever the label is constant within a subject, so a linear readout of the residual is at chance by construction and cannot settle the question, while a label that does vary within a subject is retained. The test that remains is whether change within a person carries the label, which we run directly on 57 subjects with repeated Mini-Mental State Examination scores, where the subject mean predicts the level at a Spearman correlation of 0.364 while the within-subject rate of change predicts the rate of cognitive change at -0.186, against a detection limit of 0.33. The same placement further recurs on four of six cohorts spanning five disorders. A cross-sectional accuracy therefore certifies that patients differ from one another, not that connectivity detects change within one patient, and a marker meant to follow one patient has to be validated on within-subject change.

1 INTRODUCTION

Resting-state functional connectivity has become one of the most widely used measurements in computational neuroimaging, because it imposes no radiation burden and can be acquired serially within a single participant (Sperling, 2011). One line of work classifies disease from connectivity, with architectures designed specifically for connectivity matrices and brain graphs (Kawahara et al., 2017; Kan et al., 2022). A second line identifies individuals from the same measurement, a phenomenon known as connectome fingerprinting, in which a person is matched to themselves from a connectivity profile alone, across scanning sessions and between task and rest (Finn et al., 2015), and that identifiability persists from months to years (Horien et al., 2019; Amico & Goñi, 2018). A profile therefore offers two kinds of feature to read from. Inter-individual features differ between people and stay fixed within one person, and are what identification matches on. Intra-individual features change within a person from visit to visit, and are the only ones that can capture a change in that person. One profile serves both applications, yet no study has asked whether the disease information sits in the inter-individual features or the intra-individual ones.

The answer decides what a connectivity marker can be used for. A biomarker is read as a proxy for the underlying disease state, and Alzheimer's disease is now defined by biomarkers that stage a participant along a continuum (Jack et al., 2018), so a marker for this disease is asked to detect movement between stages within one patient. Connectivity is applied to exactly that setting, most prominently to predict which patients with mild cognitive impairment will convert (Hojjati et al., 2017; 2018). Inter-individual variation is among the largest sources in a connectivity profile, and published disease effects are small beside it, so a disease effect can be read only once that component is set aside (Gratton et al., 2018). A marker read from the inter-individual features would return the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

same value at every visit of one patient, leaving deterioration, treatment response and any trial endpoint defined on change undetectable, while the cross-sectional accuracy against which such a marker is validated would be unaffected.

We conducted a series of experiments to localise the disease information within the inter-individual and intra-individual feature spaces. First, we ranked every edge twice, once by the absolute Welch t statistic for the diagnostic contrast and once by the ratio of between-subject to within-subject variance, which is a monotone transform of the intraclass correlation (Shrout & Fleiss, 1979). The top thousand edges by each ranking share 197 edges, against 88 under a permutation null that shuffles the diagnosis across subjects (Figure 1), and the shared edges cross network boundaries rather than falling in any one cortical system of Yeo et al. (2011). The subject mean alone carries almost all of the separable disease information, reading the diagnosis at 0.728 against 0.734 from the whole profile. Centring each scan on that mean, the standard operation for separating within-person from between-person effects (Curran & Bauer, 2011), leaves accuracy at 0.495. For a label that never changes within a subject the operator cancels the class-mean difference exactly, so a linear readout of the residual is at chance whatever the residual contains.

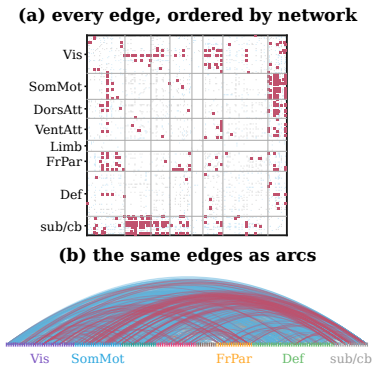


Figure 1: Disease-discriminative and identity-relevant edges. Grey: identity only. Blue: diagnosis only. Red: both, 197 of 1000 against a permutation null of 88.

We next asked whether the localisation depends on centring itself. Across seven removal methods spanning subspace removal, adversarial training and disentanglement, centring among them, each achieves the invariance its own design specifies and none recovers the diagnosis, the highest reaching 0.626. The disease information therefore stays bound to the identity component under every removal method. We then checked what centring leaves readable. Head motion, which varies from scan to scan, is retained after centring at 0.708 ($p = 0.001$), and a simulated signal placed in the residual is recovered and scales with its strength. Centring keeps what varies within a subject and removes the rest, so the test a marker has to pass is whether change within a person carries the label.

We put the test in its direct form, using a clinical score that moves within a person. On the 57 subjects whose visits carry a Mini-Mental State Examination score, the subject mean predicts the level at a Spearman correlation of 0.364, while the within-subject rate of change predicts the rate of cognitive change at -0.186, against a detection limit of 0.33. The within-subject design is not itself the obstacle, because on 217 same-day scans the structural rate of change reads the diagnosis at 0.807 while the connectivity rate reaches 0.515.

We then asked whether the same placement appears outside ADNI. On six cohorts spanning five disorders, the angle between the between-subject and the disease subspaces falls below a label-permutation null on four, covering autism, schizophrenia and attention deficit. We also tested, on data without a disease label, whether centring retains a target varying within a unit and removes a target fixed within it. On the Human Connectome Project (Van Essen et al., 2013), sex is recoverable from the subject mean at 0.910. On CREMA-D (Cao et al., 2014), a corpus of acted speech, one set of acoustic features from 91 actors carries two targets. Emotion varies within an actor and is retained after centring at 0.880. Speaker sex is constant within an actor and falls to 0.499. The target rather than the data decides the outcome, and the disease information in connectivity is carried by the fixed component. A cross-sectional accuracy therefore shows only that patients differ from one another, not that connectivity can follow one patient over time.

Main contributions. **(i)** We localise the separable disease information in resting-state connectivity to the inter-individual features, the same edges that identify the person (§4.1). **(ii)** We show that centring cannot test a label that is fixed within a subject, since the cancellation is exact, and give calibrated readouts of within-subject change that can (§4.2). **(iii)** We test tracking directly and find that change in connectivity does not follow change in cognition within a person, while the same pipeline reads cognition between people, recurring on four of six further cohorts (§4.3). **(iv)** We show that the target rather than the data decides the outcome, as centring on speech features retains emotion, varying within a speaker, and removes speaker sex, fixed within one (§4.4).

2 Related work

Individual identifiability. Connectivity is distinctive enough that a held-out profile can be matched to the person who produced it by maximal correlation, across sessions and across task and rest (Finn et al., 2015), over intervals from months to years (Horien et al., 2019), and more strongly after reconstruction from group-level modes (Amico & Goñi, 2018). These studies characterise the identifying component by how well a person can be recognised, and the nearest to ours measures subject and diagnostic identifiability in one framework on the Alzheimer's spectrum (Svaldi et al., 2018). What remains open is which of the two parts carries a diagnostic label, and we settle it by decomposing the profile and reading the diagnosis from each part.

Between-subject and within-subject decomposition. The operation that answers such a question splits connectivity variance into a part that differs between people and a part that varies within one. Applied across subjects, sessions and tasks, it attributes most of the variance to stable individual features (Gratton et al., 2018). Applied to a single edge, it defines the intraclass correlation (Shrout & Fleiss, 1979), which a meta-analysis across a decade of connectivity studies places in the poor range (Noble et al., 2019). In longitudinal modelling the same operation is centring on the unit mean (Curran & Bauer, 2011), under which a predictor fixed within a unit is absorbed by that mean. Each use describes the measurement. None asks where a disease label falls, although that placement decides whether a marker can follow one patient, and we decompose the profile to report which part carries the separable disease information.

Invariance and disentanglement. A separate body of work removes a chosen attribute from a representation and reports what the removal costs. Invariance to an attribute imposes a lower bound on error whenever the target correlates with that attribute, and the bound rises as the invariance improves (Zhao & Gordon, 2019), while a large-scale evaluation found that well-disentangled models cannot be identified without supervision (Locatello et al., 2019). In neuroimaging the attribute removed has been a site, a scanner or a demographic label, and subject identity has not been evaluated in that role, leaving open the risk of shortcut learning, where a model solves a task through a feature that is predictive but not the one intended (Geirhos et al., 2020). We test it by removing subject identity through seven methods spanning subspace removal, adversarial training and disentanglement, each attaining the invariance its objective specifies without recovering a separable representation. Extended related work is in Appendix A.

3 Method

Our goal is to determine which component of a connectivity profile carries the separable disease information. Instead of asking how accurately a classifier separates patients from controls, we decompose each profile into a part that is constant within a unit and a part that varies, and ask which part the label can be read from. Section 3.1 fixes the notation, 3.2 the benchmarks, 3.3 the localisation procedure, 3.4 the ablation around it, and 3.5 the evaluation protocol. Notation, cohort details and every hyperparameter are tabulated in Appendix B.

3.1 Problem formulation

Decomposition. We work with repeated observations of one unit. Let $\mathbf{x}_{i,t} \in \mathbb{R}^p$ denote the profile of unit i at observation t, with $t = 1, \dots, T_i$, and let $y_{i,t}$ denote the label recorded at that observation. A unit is a subject on the connectivity cohorts and a speaker on CREMA-D. We split every profile into two parts,

$$\mathbf{x}_{i,t} = \underbrace{\bar{\mathbf{x}}_i}_{\text{inter-individual}} + \underbrace{\tilde{\mathbf{x}}_{i,t}}_{\text{intra-individual}}, \qquad \bar{\mathbf{x}}_i = \tfrac{1}{T_i}\textstyle\sum_t \mathbf{x}_{i,t}, \qquad \tilde{\mathbf{x}}_{i,t} = \mathbf{x}_{i,t} - \bar{\mathbf{x}}_i. \tag{1}$$

Here $\bar{\mathbf{x}}_i$ is the unit mean, which is constant within a unit, and $\tilde{\mathbf{x}}_{i,t}$ is the residual, which has zero mean within a unit. We write $\mathcal{R}$ for the centring operator that returns $\tilde{\mathbf{x}}_{i,t}$. Centring on the unit mean is the standard operation for separating within-person from between-person effects in longitudinal modelling (Curran & Bauer, 2011), and we use it as the reference operation throughout.

The question. We ask which of the two parts carries the separable information about y, meaning information a classifier can recover from one observation at a time. Four measurements answer

it. First, we read the label from $\bar{\mathbf{x}}_i$ alone and compare the accuracy with that of the whole profile. Second, we rank the features by how well they separate the two classes and by how much they differ between units, and ask whether the two rankings agree beyond a permutation null. Third, we measure the angle separating the between-subject subspace from the disease subspace. One observation per unit suffices, so the measurement extends to cohorts scanned once. Fourth, we ask whether within-subject change carries the label, taking the profile's rate of change as the input and either the diagnosis or the rate of change of a cognitive score as the target. The first three locate the information, and the fourth tests what a marker would need to follow one patient.

What centring can and cannot test. Removing $\bar{\mathbf{x}}_i$ discards the half of the decomposition that identification uses. Its effect on a label depends only on whether that label varies within a unit. A label that varies within a unit is retained, as head motion and emotion show. A label that is constant within a unit is cancelled exactly, so a linear readout of the residual is at chance by construction (Appendix D). A simulation confirms the two cases, placing a signal of known strength either in the unit mean or in the residual and reading it from the residual alone. Subject identity is a categorical attribute with one value per person, whereas $\bar{\mathbf{x}}_i$ is part of the measurement, and removing what predicts identity differs from removing the mean.

3.2 Benchmarks

Table 1 lists the eleven cohorts and the analysis each carries. Connectivity is the vector of Fisher-transformed Pearson correlations between parcel time series, $19{,}900$ under CC200 and $79{,}800$ under Schaefer-400, and CREMA-D is summarised as 88 acoustic functionals (Cao et al., 2014).

Table 1: **Cohorts.** Scans / subjects where subjects repeat, subjects otherwise. Visits is the number of observations per unit. The four ADNI subsets are nested (Appendix B.2).

Cohort	Contrast	Scans / subjects	Visits	Analysis
Primary cohort				
ADNI multi-visit	CN vs AD	245 / 61	≥ 2	localisation, removal methods
ADNI matched	CN vs AD	227 / 61	≥ 2	modality comparison
ADNI cognition	MMSE rate	206 / 57	≥ 2	tracking test
ADNI full	CN vs AD	263 / 79	≥ 1	benchmark, leakage, subspace angle
Replication cohorts				
HCP	sex	90 / 45	2	fixed target outside disease
CREMA-D	emotion, sex	7,442 / 91	76–82	two targets on one feature set
Cross-sectional cohorts				
ABIDE-I	CN vs ASD	989	1	benchmark, subspace angle
UCLA-CNP	HC vs SCHZ	172	1	benchmark, subspace angle
ds002748	CN vs MDD	72	1	benchmark, subspace angle
COBRE	HC vs SCHZ	146	1	subspace angle
ADHD-200	HC vs ADHD	567	1	subspace angle

3.3 Localisation procedure

Two feature rankings. We rank the features twice. Diagnostic discriminability of feature e is the absolute Welch statistic on the label contrast, which does not assume equal variance between classes, and inter-individual relevance is the ratio of between-unit to within-unit variance from a one-way random-effects decomposition,

$$|t_e| \;=\; \left| \frac{\bar{x}_{e,1} - \bar{x}_{e,0}}{\sqrt{s^2_{e,1}/n_1 + s^2_{e,0}/n_0}} \right|, \tag{2}$$

$$R_e \;=\; \frac{\sigma^2_{\text{between},e}}{\sigma^2_{\text{within},e}}, \qquad \text{ICC}_e \;=\; \frac{R_e}{1+R_e}. \tag{3}$$

Since $R \mapsto R/(1+R)$ is strictly increasing, ranking by R_e and ranking by the intraclass correlation (Shrout & Fleiss, 1979) produce the same ordering. Let A_k and B_k be the top-k sets of the two orderings. Because $|t_e|$ is computed across scans, edges with high inter-individual relevance carry more variance in that statistic even under a null, so the observed intersection is read against a permutation null rather than against the k^2/p expected under independence. For the anatomical summary,

an edge joins a cortical network only when both endpoints fall in the same network of Yeo et al. (2011), and that count is read against the same null.

Subspace geometry. Let U_{id} be the top-K_{id} principal components of the unit means $\{\bar{\mathbf{x}}_i\}$, which for a cohort with one observation per unit is the between-subject subspace, and let U_{dz} be the top-K_{dz} principal directions of the classifier weight vectors collected across folds. We fix $K_{\mathrm{id}} = 20$ and $K_{\mathrm{dz}} = 10$. Given orthonormal bases Q_{id} and Q_{dz}, the principal angles are the arccosines of the singular values of $Q_{\mathrm{id}}^{\top} Q_{\mathrm{dz}}$, and we report the minimum. The absolute angle is not interpretable on its own, since an ℓ_2-regularised classifier places weight along high-variance directions whether or not the label shares structure with them, so we read it against the permutation null of Section 3.5. Algorithm 1 collects the three localising measures and the centred readout into one procedure, given in full in Appendix C. Its line 5 is what makes the procedure leakage-free, since the rankings, the two subspaces, the selection thresholds and the standardisation are all estimated on training units, so nothing about the held-out units reaches the classifier.

Algorithm 1: Leakage-free localisation

Require: $\{\mathbf{x}_{i,t}\}$, $\{y_{i,t}\}$, folds grouped by unit
Ensure: $\mathrm{AUC}_{\mathrm{raw}}$, $\mathrm{AUC}_{\mathrm{res}}$, $|A_k \cap B_k|$, $\theta_{\min}$
1: **for** each fold **do**
2: $\mathcal{T} \leftarrow$ training units, $\mathcal{E} \leftarrow$ held-out units
3: $\bar{\mathbf{x}}_i \leftarrow T_i^{-1} \sum_t \mathbf{x}_{i,t}$ for every i ▷ Eq. 1
4: $\tilde{\mathbf{x}}_{i,t} \leftarrow \mathbf{x}_{i,t} - \bar{\mathbf{x}}_i$ for every i, t ▷ Eq. 1
5: on $\mathcal{T}$ only: rankings, U_{id}, standardisation ▷ Eqs. 2–3
6: fit on $\mathcal{T}$, test on $\mathcal{E}$: $\mathbf{x} \to \mathrm{AUC}_{\mathrm{raw}}$, $\tilde{\mathbf{x}} \to \mathrm{AUC}_{\mathrm{res}}$
7: collect the fitted weight vector
8: **end for**
9: $U_{\mathrm{dz}} \leftarrow$ top-K_{dz} directions of the weights ▷ §3.3
10: $\theta_{\min} \leftarrow \angle(U_{\mathrm{id}}, U_{\mathrm{dz}})$
11: **return** fold means and $\theta_{\min}$

3.4 Ablation design

The ablation proceeds at four levels, each varying one element of the measurement while the rest are held fixed. The operator is tested first, on head motion, defined as mean displacement above the subject's own median so that the label varies from scan to scan, with its outcome compared against that of the diagnosis under identical folds. We further ablate the modality, forming two summaries in each of the two modalities on the 227 scans that carry a structural session, 217 of them acquired on the same day. The per-visit level admits centring and is reported before and after it. The rate of change is one slope per unit, so it has no counterpart before centring. The third level varies the removal method across seven, of which centring is one. Three remove the identifying directions from a fixed representation, namely centring itself, projection of the top-K principal components (PCA-K), and a domain-adversarial neural network (DANN) whose second head predicts subject identity (Ganin & Lempitsky, 2015). Four let the objective build the representation instead, namely β-TCVAE (Chen et al., 2018) and FactorVAE (Kim & Mnih, 2018), which penalise total correlation among latent factors, a hierarchical variational autoencoder (HVAE) that splits identity from state in the manner of Sø nderby et al. (2016), and a contrastive encoder trained to pull together two visits of one subject (Chen et al., 2020). Invariant risk minimisation (Arjovsky et al., 2019) treats each cohort as an environment and therefore runs leave-one-cohort-out, on a different cohort set. Each method is evaluated against the baseline its own design implies, and hyperparameters are in Table 7. The fourth level fixes the number of removed features rather than the method, deleting k edges in identity-ranked and in random order at matched k over 500 draws. For the rate of change we use the least-squares slope of each feature against age,

$$\hat{\beta}_{i,e} \;=\; \frac{\sum_t \big(\tau_{i,t} - \bar{\tau}_i\big)\big(x_{i,t,e} - \bar{x}_{i,e}\big)}{\sum_t \big(\tau_{i,t} - \bar{\tau}_i\big)^2}, \tag{4}$$

where $\tau_{i,t}$ is age at observation t. The slope is invariant to the unit mean by construction, so no inter-individual information reaches the classifier through the intercept. When the target is itself a rate, the same slope is fitted to the Mini-Mental State Examination score of each subject, a ridge regression with generalised cross-validation maps the connectivity rate to the score's rate, and the out-of-fold prediction is compared with the observed rate.

3.5 Evaluation protocol

Every measurement in this paper is evaluated under one protocol. Folds are grouped by unit rather than by observation, since a split over scans allows a classifier to label a held-out scan by recognising

the person who produced it, and the size of that effect is reported in Appendix C. Throughout we use stratified group five-fold cross-validation, or plain five-fold where each subject contributes one vector and the target is continuous, with ℓ_2-regularised logistic regression as the classifier and the area under the ROC curve as the measure, averaged over folds and then over seeds. To place the reported accuracies against a model designed for connectivity matrices rather than against a linear one alone, we also train BrainNetCNN (Kawahara et al., 2017) on the cross-sectional benchmark. Permutation nulls are formed by shuffling the label and refitting every step that depends on it. Labels are permuted across units when the label is constant within a unit, and within units when it varies inside one, so that the block structure of the data is preserved (Winkler et al., 2015). We draw 1000 permutations and report the observed value, the null mean, its 95th percentile and the permutation p. The same construction gives the null for the subspace angle, where the classifier is refit and the disease subspace re-formed inside each permutation. For a continuous target the measure is the Spearman correlation between the out-of-fold predicted rate and the observed rate, with a bootstrap interval over subjects and a detection limit given by the smallest correlation detectable at 80% power for the available number of subjects.

4 Experiments

We organise our experiments around four questions. **(Q1, §4.1)** Which component of a connectivity profile carries the disease information? **(Q2, §4.2)** What can centring test, and what can it not? **(Q3, §4.3)** Does within-subject change in connectivity track change within a person? **(Q4, §4.4)** Is the outcome decided by the disease or by the target's relation to the unit?

4.1 Q1: cross-sectional accuracy reads the person

Setup. Equation (1) requires more than one observation per unit, so this section uses ADNI multi-visit, the 61 subjects with at least two visits. Under one set of folds we read the diagnosis three times, from the whole profile, from the subject mean alone, and from the residual. We then rank every edge twice, by diagnostic discriminability and by inter-individual relevance (Equations 2 and 3), compare the two rankings at six thresholds against the permutation null, and aggregate both scores to parcels and to the seven cortical networks of Yeo et al. (2011). Three controls follow. We repeat the classification on ADNI full under a scan-level split, regress mean displacement from every edge inside training folds, and classify from the most reliable half of the edges. A staged removal of identity components is in Appendix E.1.

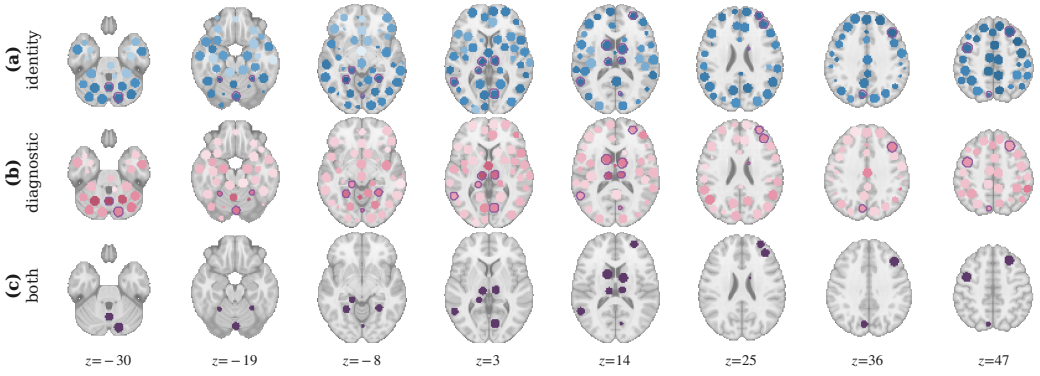


Figure 2: **Diagnosis and identity concentrate in the same parcels.** Each parcel is scored by summing the fold-averaged inter-individual relevance **(a)** and the fold-averaged absolute Welch statistic **(b)** over the 197 shared edges that touch it. The two maps correlate at $\rho = 0.49$, and **(c)** shows the 24 parcels in the top quartile of both against a permutation null of 15.6 ($p = 0.081$), outlined in violet in (a) and (b). Intensity runs from low to high within each map. Slices are axial at the z given.

Results. As shown in Table 2b, the subject mean carries almost all of the separable disease information, reading the diagnosis at 0.728 against 0.734 from the whole profile. Grouping the folds by subject matters, since on ADNI full a scan-level split raises the same measurement to 0.933 while a subject-level split leaves it at 0.742. The features behave the same way. The two rankings share 3.0 times the permutation null at $k = 250$ and 1.3 times at $k = 5000$ (Table 2a). The shared edges

top-k	shared	null	ratio	p
100	3	1.6	1.85	.215
250	24	7.9	3.03	.049
500	79	26.4	2.99	.020
1000	197	87.9	2.24	.015
2000	518	294.3	1.76	.011
5000	1952	1504.7	1.30	.025

(a) Ranking agreement.

Read from	**Cohort**	**AUC**	p
whole profile	multi-visit	.734±.03	.003
subject mean	multi-visit	.728±.05	.020
residual	multi-visit	.495±.01	.669
whole profile, scan split	full	.933	.001
whole profile, subject split	full	.742	.001
whole profile, FD regressed	multi-visit	.724±.02	.004

(b) Where the diagnosis is read from.

Table 2: **The diagnosis is read from the part of the profile that identifies the person.** **(a)** Edges selected by both rankings at each threshold, against a null that permutes the diagnosis across subjects and refits both rankings. **(b)** The diagnosis read from the whole profile, from the subject mean alone, and from the residual, all under folds grouped by subject on ADNI multi-visit. The two splits of ADNI full and the motion control follow. Entries are mean ± s.d. over 20 seeds and five folds, each against a permutation null drawn at the level of its own split.

fall in no single cortical network, since none of the seven networks of Yeo et al. (2011) holds more of them than its permutation null, and at the parcel level the two maps overlap in the thalamus, the basal ganglia and a few parietal and frontal regions, with 24 parcels in the top quartile of both against a null of 15.6 at $p = 0.081$ (Figure 2). Head motion does not account for the overlap, since mean displacement predicts the diagnosis at 0.604, yet regressing it from every edge leaves the accuracy at 0.724 and the overlap at $p = 0.007$. Selecting edges by test-retest reliability raises the accuracy to 0.757 on half of them, and the fitted weights concentrate there at 0.433 against 0.207 (Appendix E.1). Centring leaves the diagnosis at 0.495, which Section 4.2 takes up.

4.2 Q2: WHAT CENTRING CAN AND CANNOT TEST

Setup. We apply the same operator to labels that differ in one property. Head motion enters as mean displacement above the subject's own median, so that its value changes from scan to scan, and runs under the same folds as the diagnosis. We then vary three elements of the measurement in turn, holding the rest fixed. For the modality we form two summaries in each of connectivity and structural imaging, the per-visit level and the within-subject rate, on the 227 scans that carry a structural session, 217 of them acquired on the same day. For the removal method we vary the method across seven, centring among them, spanning subspace removal, adversarial training and disentanglement, each evaluated against the baseline its own design implies. For the feature count we fix how many edges are deleted while varying the order, comparing identity-ranked removal against 500 random draws at matched k. Complete tables are in Appendix E.2.

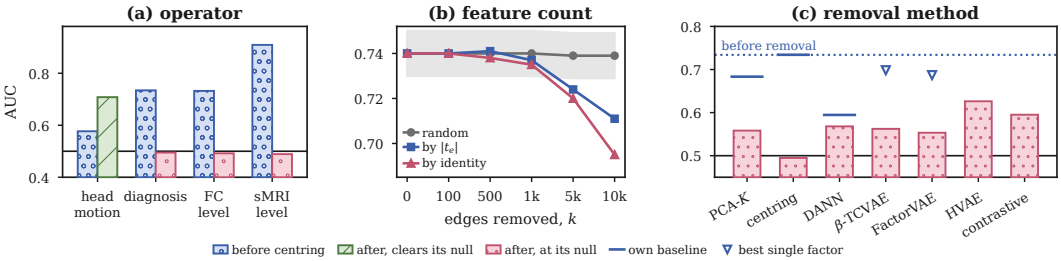


Figure 3: **What the operator does, and what the measurement does not change.** **(a)** AUC before and after centring for four labels, one that varies within a subject and three that do not, each against its own permutation null. **(b)** Disease AUC after removing k edges in random order, by diagnostic discriminability, and by inter-individual relevance, with the band spanning two standard errors over 500 draws. **(c)** Seven removal methods, each against the baseline its own design implies, with the accuracy before any removal marked. Configurations are in Table 7.

Results. As shown in Figure 3a, the operator separates the four labels by one property. Head motion, the only label that varies within a subject, is read weakly before centring at 0.577 ($p = 0.040$) and clearly afterwards at 0.708 ($p = 0.001$), while the three labels fixed within a subject start between 0.732 and 0.909, each with $p \leq 0.003$, and all sit at their nulls after centring, at 0.489 to

0.495 with $p \geq 0.669$. Starting points 0.18 apart converge on one value, the signature of the exact cancellation in Equation (5) rather than a weak signal. Figures 3b and 3c show that the outcome does not depend on how the identity component is removed. Seven removal methods, centring among them, each attain the invariance their objective specifies without recovering the diagnosis, the highest reaching 0.626 against 0.734, and deleting edges in identity-ranked order matches random deletion until $k = 1{,}000$, so the outcome follows which features are removed rather than how many. A simulation calibrates the residual readout. A signal placed only in the subject mean is undetectable in the residual, while one placed in the residual is recovered and scales with the strength injected. Together these results establish that centring keeps what varies within a subject and removes what does not, so the test it supports is one whose target moves within a person (Section 4.3).

4.3 Q3: CONNECTIVITY CHANGE DOES NOT TRACK COGNITIVE CHANGE WITHIN A PERSON

Setup. The tracking test takes a target that itself changes within a person. On the 57 subjects with a Mini-Mental State Examination score at every visit, we fit the least-squares slope against age (Equation 4) to every connectivity feature and to the score, map the connectivity rate to the score's rate by ridge regression inside each fold, and compare the out-of-fold prediction with the observed rate by Spearman correlation. The subject mean read against the score's level gives the cross-sectional reference under the same pipeline, and the smallest correlation detectable at 80% power gives the limit. We then return to the diagnosis, classifying it from the connectivity rate and, on the 217 scans acquired the same day, from the structural rate under one set of folds. Seven further readouts of the residual and their calibration are in Appendix E.3. Finally we compare the between-subject and the disease subspaces on six cohorts spanning five disorders, a measure that needs only one observation per unit.

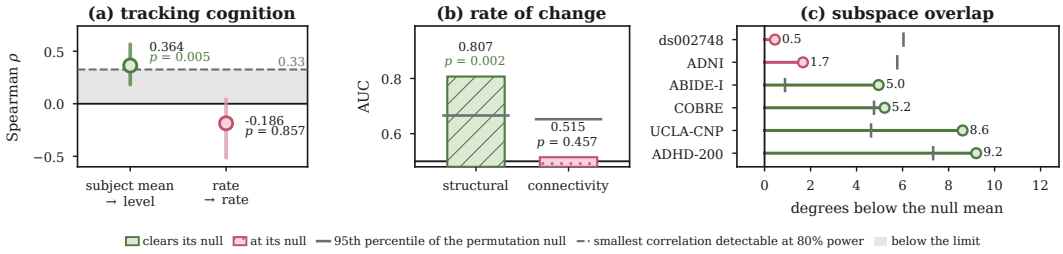


Figure 4: **Connectivity reads cognition between people and not within one. (a)** Spearman correlation between predicted and observed cognition, from the subject mean to the level and from the within-subject rate to the rate, with bootstrap intervals over subjects and the smallest correlation detectable at 80% power, the region below it shaded. **(b)** The diagnosis read from the within-subject rate of change, on the 217 same-day scans in both modalities. **(c)** The minimum principal angle separating the between-subject subspace from the disease subspace, as degrees below the mean of a label-permutation null, with grey marks at the fifth percentile of that null. All nulls use $B = 1000$ with labels permuted across subjects.

Results. As shown in Figure 4a, connectivity reads cognition between people and not within one. The subject mean predicts the level of the score at $\rho = 0.364$ (CI 0.18 to 0.57, $p = 0.005$), while the within-subject rate predicts the rate of cognitive change at -0.186 (CI -0.51 to 0.05, $p = 0.857$). Both come from one pipeline, and the test would have detected a correlation of 0.33, so the null belongs to the target rather than to the sample size. Figure 4b shows the diagnosis read from the rate of change. On scans acquired the same day and under the same folds, the structural rate carries it at 0.807 ($p = 0.002$) while the connectivity rate reaches 0.515 ($p = 0.457$), so a within-subject design can carry the label and connectivity does not. Seven residual readouts covering second-order and temporal structure agree, none exceeding its null with the smallest p at 0.23, and their calibration shows a within-subject variance difference of about fivefold would have been detected. The same geometry appears in other cohorts. Across six cohorts and five disorders, the angle between the between-subject and the disease subspaces falls below its label-permutation null on four, by 5.0 to 9.2 degrees with p from 0.001 to 0.035 (Figure 4c). The absolute angle carries no information on its own, since the null mean ranges from 44.3 to 85.1 degrees, and on ADNI and ds002748 the observed angle does not fall below its null. Together these results place the disease information in the component fixed within a person, whichever way the within-subject component is read.

4.4 Q4: THE OUTCOME FOLLOWS THE TARGET, NOT THE DISEASE

Setup. Every disease label so far has been fixed within a subject, so the outcome may follow from the disease or from that property. Separating them calls for two targets from outside the disease setting. The first is not a disease, and we take sex on HCP (Van Essen et al., 2013). The second varies within a unit, and CREMA-D (Cao et al., 2014) supplies it under a control that neuroimaging cannot offer. One set of acoustic features and one set of 91 actors carry two targets, speaker sex fixed within an actor and the emotion portrayed changing from clip to clip, so the two differ in this property and in nothing else. Sizes and protocols are in Appendix E.5.

Results. As shown in Table 3, the two properties come apart, and the outcome is governed by the relation between the target and the unit. The three targets fixed within a unit behave alike although only one is a disease, the diagnosis on ADNI reading from the subject mean at .728 against .734 from the whole profile, sex on HCP at .910 and speaker sex on CREMA-D at .995, each with $p \leq 0.020$, and the two whose residual is defined falling to their nulls once that mean is removed.

Table 3: **What centring leaves behind.** A target is reported only where its design defines a value. HCP has exactly two sessions per subject, so its residual is antisymmetric. All 91 CREMA-D actors portray all six emotions, so an actor mean carries no single emotion label.

Target	Var.	Raw	Mean	Res.
ADNI diagnosis	no	.734	.728	.495
HCP sex	no	.886	.910	—
CREMA-D sp. sex	no	.980	.995	.499
CREMA-D emotion	yes	.840	—	**.880**

The decisive comparison is within CREMA-D, where speaker sex does not change from clip to clip and falls from .980 to .499 after centring, while emotion changes and rises from .840 to .880 with $p = 0.001$. Removing what identifies the speaker improves rather than degrades the reading of the emotion, and the features do carry the speaker, identified among the 91 actors at .712 against chance of .011. Figure 5 collects every target in this paper. The four fixed within a unit start between .732 and .980 and end at their nulls, while three of the four that vary sit on or above the line of no change. The exception is the connectivity rate, which Section 4.3 shows at its null for both the diagnosis and the change in cognition.

Varying within a unit is therefore necessary for a target to be readable after centring, though not sufficient, and the cancellation behind that necessity is exact. Suppose the profile carries the label along a direction $\boldsymbol{d}$, so that $\mathbf{x}_{i,t} = \mu y_{i,t}\boldsymbol{d} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{i,t}$ with $\boldsymbol{\alpha}_i$ the unit offset. Then

$$\mathcal{R}\mathbf{x}_{i,t} \;=\; \mu\big(y_{i,t} - \bar{y}_i\big)\boldsymbol{d} \;+\; \boldsymbol{\varepsilon}_{i,t} - \bar{\boldsymbol{\varepsilon}}_i, \qquad (5)$$

so the offset is removed exactly and the label term is scaled by the within-unit variation of the label, vanishing when $y_{i,t} = y_i$ for every t. Any procedure that removes the directions along which units differ inherits the cancellation, which is why the seven methods of Section 4.2 behave alike. A proof is in Appendix D.

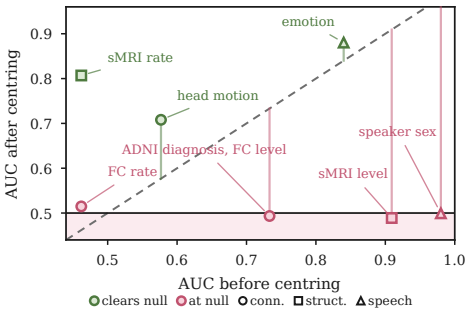


Figure 5: **One rule across two domains.** Each target plotted by its AUC before and after centring. The dashed line marks no change and the shaded band is the null. The two rates have no value before centring and are drawn on the left edge.

5 CONCLUSION

This paper asks which component of a connectivity profile carries the separable disease information. We localise it to the component that stays fixed within a person, which reads the diagnosis almost as well as the whole profile and is built from the edges that identify the person. Centring cancels a subject-constant label exactly, so the usual test cannot settle the question, and what can be tested is whether change within a person carries the label. On the same subjects the subject mean predicts the level of cognition and the within-subject rate does not predict its rate, at -0.186 against a detection limit of 0.33, while on same-day scans the structural rate carries the diagnosis and the connectivity rate does not. A cross-sectional accuracy therefore certifies that patients differ from one another, and a marker meant to follow one of them has to be validated on within-subject change.

THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this work, we used generative AI tools to edit the manuscript for readability. None of the tasks requiring disclosure are applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

ADHD-200 Consortium. The ADHD-200 Consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62, 2012. doi: 10.3389/fnsys.2012.00062.

C. J. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello, J. Liu, A. R. Mayer, N. I. Perrone-Bizzozero, S. Posse, J. M. Stephen, J. A. Turner, V. P. Clark, and V. D. Calhoun. Multimodal neuroimaging in schizophrenia: Description and dissemination. *Neuroinformatics*, 15(4):343–364, 2017. doi: 10.1007/s12021-017-9338-9.

Enrico Amico and Joaquín Goñi. The quest for identifiability in human functional connectomes. *Scientific Reports*, 8(1):8254, 2018. doi: 10.1038/s41598-018-25089-1.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

Dmitry D. Bezmaternykh, Mikhail Ye. Melnikov, Andrey A. Savelov, Lyudmila I. Kozlova, Evgeniy D. Petrovskiy, Kira A. Natarova, and Mark B. Shtark. Brain networks connectivity in mild to moderate depression: Resting state fMRI study with implications to nonpharmacological treatment. *Neural Plasticity*, 2021:8846097, 2021. doi: 10.1155/2021/8846097.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Rastko Ciric, Daniel H. Wolf, Jonathan D. Power, David R. Roalf, Graham L. Baum, Kosha Ruparel, Russell T. Shinohara, Mark A. Elliott, Simon B. Eickhoff, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Danielle S. Bassett, and Theodore D. Satterthwaite. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174–187, 2017. doi: 10.1016/j.neuroimage.2017.03.020.

Patrick J. Curran and Daniel J. Bauer. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology*, 62:583–619, 2011. doi: 10.1146/annurev.psych.093008.100356.

A. Di Martino, C. G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R. A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A.

Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, and M. P. Milham. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014. doi: 10.1038/mp.2013.78.

Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, 2015. doi: 10.1038/nn.4135.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ganin15.html.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Caterina Gratton, Timothy O. Laumann, Ashley N. Nielsen, Deanna J. Greene, Evan M. Gordon, Adrian W. Gilmore, Steven M. Nelson, Rebecca S. Coalson, Abraham Z. Snyder, Bradley L. Schlaggar, Nico U. F. Dosenbach, and Steven E. Petersen. Functional brain networks are dominated by stable group and individual factors, not cognitive or daily variation. *Neuron*, 98(2):439–452.e5, 2018. doi: 10.1016/j.neuron.2018.03.035.

Seyed Hani Hojjati, Ata Ebrahimzadeh, Ali Khazaee, and Abbas Babajani-Feremi. Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM. *Journal of Neuroscience Methods*, 282:69–80, 2017. doi: 10.1016/j.jneumeth.2017.03.006.

Seyed Hani Hojjati, Ata Ebrahimzadeh, Ali Khazaee, and Abbas Babajani-Feremi. Predicting conversion from MCI to AD by integrating rs-fMRI and structural MRI. *Computers in Biology and Medicine*, 102:30–39, 2018. doi: 10.1016/j.compbiomed.2018.09.004.

Corey Horien, Xilin Shen, Dustin Scheinost, and R. Todd Constable. The individual functional connectome is unique and stable over months to years. *NeuroImage*, 189:676–687, 2019. doi: 10.1016/j.neuroimage.2019.02.002.

Clifford R. Jack, Jr., David A. Bennett, Kaj Blennow, Maria C. Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M. Holtzman, William Jagust, Frank Jessen, Jason Karlawish, Enchi Liu, Jose Luis Molinuevo, Thomas Montine, Creighton Phelps, Katherine P. Rankin, Christopher C. Rowe, Philip Scheltens, Eric Siemers, Heather M. Snyder, and Reisa Sperling. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018. doi: 10.1016/j.jalz.2018.02.018.

Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 25586–25599. Curran Associates, Inc., 2022. doi: 10.52202/068431-1855. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a408234a9b80604a9cf6ca518e474550-Paper-Conference.pdf.

Jeremy Kawahara, Colin J. Brown, Steven P. Miller, Brian G. Booth, Vann Chau, Ruth E. Grunau, Jill G. Zwicker, and Ghassan Hamarneh. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017. doi: 10.1016/j.neuroimage.2016.09.046.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Sophia Mueller, Danhong Wang, Michael D. Fox, B. T. Thomas Yeo, Jorge Sepulcre, Mert R. Sabuncu, Rebecca Shafee, Jie Lu, and Hesheng Liu. Individual variability in functional connectivity architecture of the human brain. *Neuron*, 77(3):586–595, 2013. doi: 10.1016/j.neuron.2012.12.028.

Stephanie Noble, Marisa N. Spann, Fuyuze Tokoglu, Xilin Shen, R. Todd Constable, and Dustin Scheinost. Influences on the test-retest reliability of functional connectivity MRI and its relationship with behavioral utility. *Cerebral Cortex*, 27(11):5415–5429, 2017. doi: 10.1093/cercor/bhx230.

Stephanie Noble, Dustin Scheinost, and R. Todd Constable. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, 203:116157, 2019. doi: 10.1016/j.neuroimage.2019.116157.

Ronald C. Petersen, Paul S. Aisen, Laurel A. Beckett, Michael C. Donohue, Anthony C. Gamst, Danielle J. Harvey, Clifford R. Jack, Jr., William J. Jagust, Leslie M. Shaw, Arthur W. Toga, John Q. Trojanowski, and Michael W. Weiner. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology*, 74(3):201–209, 2010. doi: 10.1212/WNL.0b013e3181cb3e25.

R. A. Poldrack, E. Congdon, W. Triplett, K. J. Gorgolewski, K. H. Karlsgodt, J. A. Mumford, F. W. Sabb, N. B. Freimer, E. D. London, T. D. Cannon, and R. M. Bilder. A phenome-wide examination of neural and cognitive function. *Scientific Data*, 3:160110, 2016. doi: 10.1038/sdata.2016.110.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018. doi: 10.1093/cercor/bhx179.

Marc-Andre Schulz, B. T. Thomas Yeo, Joshua T. Vogelstein, Janaina Mourao-Miranada, Jakob N. Kather, Konrad Kording, Blake Richards, and Danilo Bzdok. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, 11(1):4238, 2020. doi: 10.1038/s41467-020-18037-z.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Casper Kaae Sø nderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sø nderby, and Ole Winther. Ladder variational autoencoders. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6ae07dcb33ec3b7c814df797cbda0f87-Paper.pdf.

Reisa Sperling. The potential of functional MRI as a biomarker in early Alzheimer's disease. *Neurobiology of Aging*, 32(Suppl 1):S37–S43, 2011. doi: 10.1016/j.neurobiolaging.2011.09.009.

Rolf Steyer, Manfred Schmitt, and Michael Eid. Latent state–trait theory and research in personality and individual differences. *European Journal of Personality*, 13(5):389–408, 1999. doi: 10.1002/(SICI)1099-0984(199909/10)13:5⟨389::AID-PER361⟩3.0.CO;2-A.

Diana O. Svaldi, Joaquín Goñi, Apoorva Bharthur Sanjay, Enrico Amico, Shannon L. Risacher, John D. West, Mario Dzemidzic, Andrew Saykin, and Liana Apostolova. Towards subject and diagnostic identifiability in the Alzheimer's disease spectrum based on functional connectomes. In *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, volume 11044 of *Lecture Notes in Computer Science*, pp. 74–82. Springer, 2018. doi: 10.1007/978-3-030-00689-1_8.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80: 62–79, 2013. doi: 10.1016/j.neuroimage.2013.05.041.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digital Medicine*, 5(1):48, 2022. doi: 10.1038/s41746-022-00592-y.

Junhao Wen, Elina Thibeau-Sutre, Mauricio Diaz-Melo, Jorge Samper-González, Alexandre Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Ninon Burgos, and Olivier Colliot. Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation. *Medical Image Analysis*, 63:101694, 2020. doi: 10.1016/j.media.2020.101694.

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

B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011. doi: 10.1152/jn.00338.2011.

Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b4189d9de0fb2b9cce090bd1a15e3420-Paper.pdf.

Appendix

Table of Contents

A Extended related work **15**

A.1 Identifiability 15

A.2 Decomposition and reliability 15

A.3 Invariance and evaluation 15

B Datasets and implementation **16**

B.1 Glossary and notation 16

B.2 Cohorts 17

B.3 Preprocessing pipeline 18

B.4 Implementation details 19

C The localisation procedure **21**

D Formal statements **22**

D.1 Exact cancellation of a label that is constant within a unit 22

D.2 What a collapse to chance implies 22

E Additional results **23**

E.1 Cross-sectional benchmark, reliability selection and staged removal 23

E.2 The ablation in full 23

E.3 Within-subject change in full 24

E.4 Conversion within a subject 25

E.5 Cross-domain replication 26

A Extended related work

A.1 Identifiability

Connectome fingerprinting matches a held-out profile to the same person's profile from another session by maximal correlation (Finn et al., 2015; Amico & Goñi, 2018; Horien et al., 2019). Identification succeeds across rest and task, persists from months to years, and is strongest in frontoparietal and medial frontal networks, the same networks that best predict fluid intelligence (Finn et al., 2015). The identifying signal is not fixed by the biology alone. It can be concentrated or diluted by preprocessing (Amico & Goñi, 2018), and individual variability is itself concentrated in heteromodal association cortex (Mueller et al., 2013). These studies establish that a stable, person-specific component exists and is readable, and they supply the component we remove. They treat it as the target of identification, whereas we ask which label it carries.

A.2 Decomposition and reliability

Centring an observation on its unit mean is the standard operation for separating within-person from between-person effects (Curran & Bauer, 2011; Shrout & Fleiss, 1979), and it is the core of latent state–trait theory, which splits a measurement into a stable trait and an occasion-specific state (Steyer et al., 1999). Applied to connectivity, the decomposition shows that between-subject variance dominates and that published disease effects are small beside it (Gratton et al., 2018). Reliability studies quantify the same split. The mean edge-level intraclass correlation across a decade of connectivity studies is 0.29, between-subject variability accounts for 80% of the variance in that coefficient, and the most reliable edges are the stronger, within-network, cortical ones (Noble et al., 2019; 2017). The same operation is standard in longitudinal modelling, where a predictor that does not change within a unit is absorbed by that unit's mean. What these studies establish is that the two components differ in magnitude and in stability. None of them asks which component a disease label occupies.

A.3 Invariance and evaluation

A separate body of work removes a chosen attribute from a representation and reports what the removal costs. Adversarial approaches train a second head to predict the attribute and reverse its gradient (Ganin & Lempitsky, 2015). Latent-variable approaches penalise total correlation among factors (Chen et al., 2018; Kim & Mnih, 2018; Sø nderby et al., 2016), contrastive approaches build the invariance into the training signal (Chen et al., 2020), and invariant risk minimisation seeks a predictor that is optimal in every environment (Arjovsky et al., 2019).

The cost of that removal is known from two directions. Theory gives a lower bound on error whenever the target correlates with the attribute a representation is constrained to be invariant to (Zhao & Gordon, 2019). Empirically, twelve thousand disentanglement models across seven datasets show that well-disentangled models seemingly cannot be identified without supervision (Locatello et al., 2019). Evaluation protocols bear on the same question, since more than half of surveyed convolutional pipelines for Alzheimer's classification may have suffered from data leakage, and below roughly 10^4 subjects linear models perform on par with more heavily parameterised ones on brain images (Wen et al., 2020; Schulz et al., 2020; Varoquaux & Cheplygina, 2022).

The attribute in that literature has been a site, a scanner or a protected demographic label, and subject identity has not been evaluated in that role. That omission matters because a model can solve a task through a feature that is predictive but not the one intended, a failure mode known as shortcut learning (Geirhos et al., 2020). This work addresses the gap common to the three lines. We take the decomposition the second supplies, use it to ask which component a label falls in, and use the identifiability the first establishes to certify that what is removed is what identifies the person. The seven removal methods, their hyperparameters and their baselines are listed in Table 7.

B DATASETS AND IMPLEMENTATION

B.1 GLOSSARY AND NOTATION

Table 4: **Glossary and notation.**

Symbol	Description
Measurement and indices	
p	Feature dimension, 19,900 for CC200, 79,800 for Schaefer-400, 88 for eGeMAPSv02
$i,\ t$	Unit index and observation index
T_i	Number of observations of unit i
$\mathbf{x}_{i,t} \in \mathbb{R}^p$	Profile of unit i at observation t
$y_{i,t}$	Target label recorded at observation t
$\bar{y}_i$	Unit mean of the label, $T_i^{-1} \sum_t y_{i,t}$
s	Unit identity, a categorical attribute with one value per unit
Decomposition	
$\bar{\mathbf{x}}_i = T_i^{-1} \sum_t \mathbf{x}_{i,t}$	Inter-individual component, constant within a unit
$\tilde{\mathbf{x}}_{i,t} = \mathbf{x}_{i,t} - \bar{\mathbf{x}}_i$	Intra-individual component, zero mean within a unit
$\mathcal{R}$	Centring operator, $\mathcal{R}\mathbf{x}_{i,t} = \tilde{\mathbf{x}}_{i,t}$
Feature-level scores	
t_e	Welch statistic for edge e, computed across scans. Diagnostic discriminability is $\lvert t_e \rvert$
$R_e = \sigma^2_{\text{between},e} / \sigma^2_{\text{within},e}$	Inter-individual relevance of edge e
$\text{ICC}_e = R_e/(1+R_e)$	Intraclass correlation, increasing in R_e, so the two induce one ordering (Shrout & Fleiss, 1979)
$A_k,\ B_k$	Top-k sets of the two rankings, whose overlap is read against a subject-level label-permutation null
Subspace geometry	
U_{id}	Top-K_{id} principal components of the unit means, which with one observation per unit is the between-subject subspace
U_{dz}	Top-K_{dz} principal directions of the fitted classifier weights
$K_{\text{id}},\ K_{\text{dz}}$	Subspace dimensions, 20 and 10
$\theta_{\min}$	Minimum principal angle between U_{id} and U_{dz}
Generative form used in the propositions	
$\boldsymbol{\alpha}_i$	Unit-specific offset
$\boldsymbol{d}$	Unit-norm direction along which the label separates
μ	Effect size along $\boldsymbol{d}$
$\boldsymbol{\varepsilon}_{i,t}$	Per-observation noise
Evaluation	
BA, AUC	Balanced accuracy and area under the ROC curve
$I(\mathbf{x}; y)$	Information a profile carries about the label
$I(\mathbf{x}; y \mid s)$	Information a per-observation profile carries about the label once the unit is known
d_{TV}	Total variation distance between the two class-conditional laws

B.2 COHORTS

Table 5: **Source of datasets used in this research.** Size is observations / units where units repeat, units otherwise. Each ADNI subset is defined by the requirement its experiment imposes. The full cross-sectional subset carries the leakage comparison, where scan-level and subject-level splits are drawn from the same data. The multi-visit subset keeps the subjects with at least two visits, for which the decomposition is defined, and has $\{2{:}12,\ 3{:}11,\ 4{:}19,\ 5{:}11,\ 6{:}3,\ 7{:}3,\ 9{:}2\}$ visits per subject. The matched subset keeps only the visits that also carry a structural session. The cognition subset keeps the multi-visit subjects with a Mini-Mental State Examination score at every visit.

Dataset	Size	Details
Primary cohort, repeated observations per subject		
ADNI, full cross-sectional	263 / 79	CN versus AD, 162 CN and 101 AD. Carries the leakage comparison, the deep baseline, and the subspace geometry with one scan per subject (Petersen et al., 2010).
ADNI, multi-visit	245 / 61	CN versus AD, 150 CN and 95 AD. Carries the localisation, the removal curves, and the feature-level analyses. A wider CN-versus-impaired contrast on the same visits gives 252 scans from 77 subjects, in which 5 change label within the observation window, and carries the analysis of Appendix E.4.
ADNI, matched structural	227 / 61	CN versus AD, 142 CN and 85 AD. Additionally carries a structural session, 217 of the 227 acquired on the same day and the remaining ten within 686 days. Modality comparison only.
ADNI, cognition	206 / 57	CN versus AD subjects with a Mini-Mental State Examination score at each of at least two visits. Carries the tracking test of Section 4.3.
Replication cohorts, repeated observations per unit		
HCP	90 / 45	Sex in healthy young adults, two sessions each. Schaefer-400 parcellation, 79,800 edges (Van Essen et al., 2013; Schaefer et al., 2018).
CREMA-D	7,442 / 91	Emotion and speaker sex. Twelve fixed sentences in six emotional states, no exclusions. eGeMAPSv02 functionals, 88 dimensions (Cao et al., 2014).
Single-session cohorts, cross-sectional comparison only		
ABIDE-I	989	ASD versus CN, 480 ASD and 509 CN. CC200 time series from the Preprocessed Connectomes Project, `cpac/filt_noglobal` (Di Martino et al., 2014).
UCLA-CNP	172	SCHZ versus HC, 50 SCHZ and 122 HC, drawn from the 261 subjects released with SCHZ, BIP and ADHD labels. fMRIPrep derivatives released on OpenNeuro (Poldrack et al., 2016).
ds002748	72	MDD versus CN. OpenNeuro accession, preprocessed with the ADNI configuration (Bezmaternykh et al., 2021).
COBRE	146	SCHZ versus HC. Subspace geometry only (Aine et al., 2017).
ADHD-200	567	ADHD versus HC, 193 ADHD and 374 HC. Subspace geometry only (ADHD-200 Consortium, 2012).

B.3 PREPROCESSING PIPELINE

Table 6: **Acquisition, software and preprocessing.** The stage at which each cohort is taken up differs. ADNI and ds002748 are processed from raw data with fMRIPrep and no global signal regression. ABIDE-I is taken as released by the Preprocessed Connectomes Project and UCLA-CNP as fMRIPrep derivatives from OpenNeuro, both already parcellated under CC200, so only the confound and filtering steps are applied. COBRE and ADHD-200 share the parcellation and the connectivity measure but not the preprocessing, and are used for the subspace geometry alone.

Setting	**Value**	**Scope**
ADNI acquisition	3T Philips, TR 3.0 s, TE 30 ms, flip angle 80°, $64 \times 64 \times 48$ voxels, 140 volumes	ADNI
Software	Python 3.10, fMRIPrep 24.1.1, dcm2niix v1.0.20250505, nilearn 0.10.3, scikit-learn 1.3.2, PyTorch 2.1, NumPy 1.26, SciPy 1.11	All
Functional pipeline	fMRIPrep 24.1.1, MNI152NLin2009cAsym 2 mm, ANTs extraction, BBR coregistration, MCFLIRT, no surface reconstruction	ADNI, ds002748
Confounds	6 motion parameters, mean WM and CSF, no derivatives, quadratics or GSR	ADNI, ds002748, UCLA-CNP
Filtering	Band-pass 0.01–0.1 Hz, linear detrend, z-score, first 5 volumes dropped, 6 mm FWHM smoothing during ROI extraction	ADNI, ds002748, UCLA-CNP
Parcellation	Craddock CC200, whose atlas file carries 202 regions of which the first 200 are retained, giving 19,900 Fisher-transformed Pearson correlations	All connectivity except HCP
Quality control	Mean FD ≤ 0.5 mm, fewer than 10% of regions with dropout, no censoring	All connectivity
Motion control	Mean FD regressed from every edge inside training folds. Across the 61 subjects, the QC-FC correlation has median $\lvert r\rvert = 0.118$ before regression, with 13.5% of the 19,900 edges at $p < 0.05$, against 5% expected by chance (Ciric et al., 2017)	ADNI
Acoustic features	openSMILE eGeMAPSv02, 88 dimensions	CREMA-D

B.4 IMPLEMENTATION DETAILS

Table 7: **Evaluation protocol, permutation nulls, deep baseline and removal methods.** Hyperparameters are fixed a priori. Seven removal methods run on the ADNI multi-visit subset under the protocol above, three of them against the same representation without the removal step and two against the best single latent factor. Invariant risk minimisation treats each cohort as an environment and runs leave-one-cohort-out, on a different cohort set.

Setting	Value	Scope
Evaluation protocol		
Cross-validation	Stratified group five-fold, grouped by subject or actor; plain five-fold for the tracking test, where each subject is one row	All accuracies
Aggregation	AUC averaged over folds, then over seeds	All accuracies
Classifier	LR-ℓ_2, `C=1.0`, `max_iter=3000`; `liblinear` for connectivity, `lbfgs` for acoustic features and for every subspace-angle estimate	Linear models
Standardisation	Per fold, train-fold statistics only	All methods
Fold-internal	Subject means, relevance scores, rankings, thresholds and the two subspaces	All analyses
Seeds	20 linear (fold assignment); 5 for CREMA-D; 3 for tuned BrainNetCNN; 10 for DANN, VAE and contrastive (also init and batch order)	Per experiment
Permutation nulls		
Blocking	Labels permuted across units for a label that is constant within a unit, and within units for a label that varies inside one, with every label-dependent step refit in each permutation	All nulls
Counts	$B = 1000$	All nulls
Reported	Observed value, null mean, null 95th percentile, and $p = (1 + \#\{\text{null at least as extreme}\})/(1 + B)$	All nulls
Subspace null	Labels permuted across subjects, the classifier refit in every fold and seed, U_{dz} re-formed and the angle recomputed. Lower-tailed, so the 5th percentile is reported	Subspace geometry
Uncertainty	Subject-level bootstrap, 1000 resamples, reported alongside the across-seed standard deviation	Where reported
Tracking test		
Rates	Least-squares slope against age, Equation (4), for every connectivity feature and for the cognitive score	ADNI cognition
Regressor	Ridge, penalty chosen by generalised cross-validation inside each fold, refit in every permutation	ADNI cognition
Statistic	Spearman correlation between out-of-fold predicted and observed rate, averaged over 20 fold seeds	ADNI cognition
Uncertainty	Bootstrap over subjects, 1000 draws; target permuted across subjects, $B = 1000$	ADNI cognition
Detection limit	Smallest correlation detectable at 80% power, one-sided $\alpha = 0.05$, Fisher z, for the available n	ADNI cognition
Deep baseline		
Optimiser	Adam, lr 10^{-4}, (β_1, β_2)=$(0.9, 0.999)$, batch 16	All deep models
Stopping	Early stopping on validation AUC, patience 20, max 200 epochs; validation is a 10% subject-stratified split of the train fold	All deep models
BrainNetCNN	E2E, E2N, N2G layers; channels $(32, 64, 128)$; dropout 0.5; two-layer head	(Kawahara et al., 2017)
Removal applied to a fixed representation		
PCA-K removal	PCA on train fold, top-K projected out of both, $K \in \{0, 5, 10, 20, 50, 100\}$, then LR-$\ell_2$	K=0
Centring	Classify on $\mathbf{x}_{i,t} - \bar{\mathbf{x}}_i$, subject-level protocol	Raw FC

Table 7: **Evaluation protocol, permutation nulls, deep baseline and removal methods** (continued).

Setting	Value	Scope
DANN	PCA(50); encoder $50 \to 16$; linear disease head (BCE); identity head over $\lvert s\rvert$ subjects (CE); λ_{GRL}=0.3 warmed from 0; Adam 5×10^{-4}, wd 10^{-4}, 100 epochs	PCA-LR(50)
Representation created by the objective		
β-TCVAE	PCA(200); 32 latents; $200 \to 128 \to 64 \to (\mu, \log \sigma^2)$, mirrored decoder; α=1, β=10, γ=1; Adam 10^{-3}, batch 64, 100 epochs	Best single factor
FactorVAE	Same encoder and decoder; γ=10; discriminator Adam 10^{-4}, (β_1, β_2)=(0.5, 0.9), alternating	Best single factor
Hierarchical VAE	PCA(200); $z_{\mathrm{id}}, z_{\mathrm{state}} \in \mathbb{R}^{16}$, width 128; recon $+ \mathrm{KL} + \mathcal{L}_{\mathrm{contr}}(z_{\mathrm{id}}) + 10\,\mathrm{CE}(z_{\mathrm{state}}, y)$; Adam 10^{-3}, 300 epochs	—
Contrastive	PCA(200); backbone $200 \to 128 \to 64 \to 32$; positives are two visits of one subject; InfoNCE τ=0.1, 200 epochs, Adam 10^{-3}; linear probe on the backbone	—
Different cohort set and protocol		
IRM	PCA(50); encoder $50 \to 64 \to 32$; one environment per cohort; λ_{IRM}=1 after 100-epoch warmup; Adam 10^{-3}, 500 epochs, test cohort held out	—

C THE LOCALISATION PROCEDURE

Algorithm 2 Leakage-free localisation with its permutation null, full form

Require: profiles $\{\mathbf{x}_{i,t}\}$, labels $\{y_{i,t}\}$, unit ids, threshold k, dimensions K_{id} and K_{dz}, seeds S, permutations B

Ensure: $\mathrm{AUC}_{\mathrm{raw}}$, $\mathrm{AUC}_{\mathrm{mean}}$, $\mathrm{AUC}_{\mathrm{res}}$, $|A_k \cap B_k|$, $\theta_{\min}$, each with null mean, 95th percentile and p

1: **procedure** MEASURE($\{\mathbf{x}_{i,t}\}, \{y_{i,t}\}$)
2: **for** each seed $s \in S$ **do**
3: draw five folds grouped by unit, stratified by label, seeded by s
4: **for** each fold **do**
5: $\mathcal{T} \leftarrow$ training units, $\mathcal{E} \leftarrow$ held-out units
6: $\bar{\mathbf{x}}_i \leftarrow T_i^{-1} \sum_t \mathbf{x}_{i,t}$ for every unit i
7: $\tilde{\mathbf{x}}_{i,t} \leftarrow \mathbf{x}_{i,t} - \bar{\mathbf{x}}_i$ for every i, t
8: on $\mathcal{T}$ only: $|t_e|$ and R_e for every e ▷ Eqs. 2–3
9: on $\mathcal{T}$ only: $A_k, B_k \leftarrow$ top-k by $|t_e|$ and by R_e
10: on $\mathcal{T}$ only: $U_{\mathrm{id}} \leftarrow$ top-K_{id} PCs of $\{\bar{\mathbf{x}}_i\}_{i \in \mathcal{T}}$
11: on $\mathcal{T}$ only: $\mu_e, \sigma_e \leftarrow$ mean and s.d. of every feature
12: standardise $\mathbf{x}$, $\bar{\mathbf{x}}$, $\tilde{\mathbf{x}}$ on $\mathcal{T}$ and $\mathcal{E}$ with μ_e, σ_e
13: fit ℓ_2 logistic regression on $\mathcal{T}$, score $\mathcal{E}$: $\mathbf{x} \to \mathrm{AUC}_{\mathrm{raw}}$, $\bar{\mathbf{x}} \to \mathrm{AUC}_{\mathrm{mean}}$, $\tilde{\mathbf{x}} \to \mathrm{AUC}_{\mathrm{res}}$
14: record $|A_k \cap B_k|$ and the fitted weight vector $\boldsymbol{w}$
15: **end for**
16: **end for**
17: $U_{\mathrm{dz}} \leftarrow$ top-K_{dz} principal directions of all recorded $\boldsymbol{w}$
18: $\theta_{\min} \leftarrow$ minimum principal angle between U_{id} and U_{dz}
19: **return** each AUC and $|A_k \cap B_k|$ averaged over folds then seeds, and $\theta_{\min}$
20: **end procedure**
21: **procedure** NULL($\{\mathbf{x}_{i,t}\}, \{y_{i,t}\}$)
22: **for** $b = 1, \ldots, B$ **do**
23: permute the label across units, keeping it constant within each unit
24: $\mathrm{stat}_b \leftarrow$ MEASURE($\{\mathbf{x}_{i,t}\}$, permuted labels)
25: **end for**
26: **return** null mean, 95th percentile (5th for $\theta_{\min}$), and $p = (1 + \#\{\mathrm{stat}_b \text{ at least as extreme}\})/(1 + B)$
27: **end procedure**

Lines 5 to 11 use only the training units of the current fold, so no held-out unit influences the feature selection, the two subspaces, or the standardisation. The folds are grouped by unit (line 3) because a scan-level split lets a classifier label a held-out scan by recognising the person who produced it. A subject with T_i scans then has on average $T_i(1 - 1/5)$ of them in the training partition of any fold, so a held-out scan lies close to a mean the classifier has already learned. On ADNI full this raises the AUC from 0.742 under a subject-grouped split to 0.933 under a scan-level split. Line 13 uses ℓ_2 logistic regression. BrainNetCNN (Kawahara et al., 2017), tuned as in Table 7 and trained on the same folds of ADNI full over three seeds, reaches 0.691 ± 0.008 against 0.742 for the linear model, and with its default configuration does not exceed the linear model on any of the four cross-sectional benchmark cohorts (Appendix E), so the linear accuracy is not a capacity limit.

Lines 10 and 18 estimate the two subspaces. The between-subject subspace is estimated from one scan per unit so that all geometry cohorts are treated alike. On ADNI the scan is drawn at random from each of the 79 subjects of the full cross-sectional cohort, giving 47.70°, while using every scan of the same subjects gives 60.74°. Because NULL refits the classifier and re-forms U_{dz} inside every permutation, the covariance structure of the profile and the tendency of an ℓ_2 classifier to load on high-variance directions are both preserved, which is why the null mean ranges from 44.3° on ds002748 to 85.1° on ABIDE-I while two random subspaces of the same dimensions would meet at nearly a right angle on every cohort. Results for all six cohorts are in Appendix E.

D FORMAL STATEMENTS

Two statements support the argument. The first concerns what the centring operator does to a label, and is the reason the removal methods of Section 4.2 behave alike. It also states what remains when the label varies within a unit, namely the label term scaled by its within-unit deviation. The second bounds what a classifier can achieve on the residual. The model is a random-intercept model and the operator is the within transformation of longitudinal analysis (Curran & Bauer, 2011).

D.1 EXACT CANCELLATION OF A LABEL THAT IS CONSTANT WITHIN A UNIT

Proposition 1. *Let $\mathbf{x}_{i,t} = \mu y_{i,t}\boldsymbol{d} + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{i,t}$, where $\boldsymbol{d} \in \mathbb{R}^p$ is a fixed direction, $\mu \in \mathbb{R}$ a scalar, $\boldsymbol{\alpha}_i$ a unit offset, and $\boldsymbol{\varepsilon}_{i,t}$ a residual term. Let $\mathcal{R}$ be the centring operator of Equation (1). Then for every unit i and every observation t,*

$$\mathcal{R}\mathbf{x}_{i,t} \;=\; \mu\big(y_{i,t} - \bar{y}_i\big)\boldsymbol{d} \;+\; \boldsymbol{\varepsilon}_{i,t} - \bar{\boldsymbol{\varepsilon}}_i, \qquad \bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{i,t}. \tag{6}$$

If $y_{i,t} = y_i$ for all t, the first term is identically zero, for every direction $\boldsymbol{d}$, every μ, and every $T_i \geq 1$.

Proof. The unit mean of the profile is

$$\bar{\mathbf{x}}_i \;=\; \frac{1}{T_i}\sum_{t=1}^{T_i}\mathbf{x}_{i,t} \;=\; \mu\bar{y}_i\boldsymbol{d} + \boldsymbol{\alpha}_i + \bar{\boldsymbol{\varepsilon}}_i, \tag{7}$$

since $\boldsymbol{\alpha}_i$ does not depend on t. Subtracting it from $\mathbf{x}_{i,t}$ removes the offset and leaves Equation (6). When the label is constant within the unit, $\bar{y}_i = y_i = y_{i,t}$, so $y_{i,t} - \bar{y}_i = 0$ and the label term vanishes. □

Remark 2. Equation (6) also states what remains when the label does vary within a unit. The label term is retained in proportion to $y_{i,t} - \bar{y}_i$, the deviation of the label at observation t from its unit mean. A label that varies widely within a unit is largely retained, one that barely varies is largely cancelled. This is why head motion is readable after centring where the diagnosis is not, and why emotion on CREMA-D is retained where speaker sex is lost.

Remark 3. The cancellation is a property of the operator and the label, not of the classifier or the feature space. Any procedure whose output is a function of $\mathcal{R}\mathbf{x}_{i,t}$ inherits it. This covers centring directly, and covers projection of the identity subspace, adversarial training and the latent-variable methods to the extent that each removes the directions along which units differ. It is the reason the seven methods of Section 4.2 behave alike despite differing in objective, in representation and in what they treat as the attribute.

D.2 WHAT A COLLAPSE TO CHANCE IMPLIES

Proposition 4. *Let s denote the unit identity and y the label. Suppose the conditional mutual information satisfies $I(\mathbf{x}; y \mid s) \leq \epsilon$, and that the residual $\tilde{\mathbf{x}}$ retains at most δ nats of information about s, so that $I(\tilde{\mathbf{x}}; s) \leq \delta$. Then for balanced classes, any classifier acting on $\tilde{\mathbf{x}}$ has balanced accuracy at most*

$$\mathrm{BA} \;\leq\; \tfrac{1}{2} + \tfrac{1}{2}\sqrt{2(\epsilon + \delta)}. \tag{8}$$

Proof. Adding a variable cannot decrease mutual information, so $I(\tilde{\mathbf{x}}; y) \leq I(\tilde{\mathbf{x}}; (y, s)) = I(\tilde{\mathbf{x}}; s) + I(\tilde{\mathbf{x}}; y \mid s)$ by the chain rule. The first term is at most δ by assumption. The second is at most $I(\mathbf{x}; y \mid s) \leq \epsilon$ by the data processing inequality, since $\tilde{\mathbf{x}}_{i,t}$ is a function of the observations of unit i. For balanced classes, write P_0 and P_1 for the two class-conditional laws of $\tilde{\mathbf{x}}$ and M for their equal mixture, so that $I(\tilde{\mathbf{x}}; y) = \frac{1}{2}D(P_0\|M) + \frac{1}{2}D(P_1\|M)$. Pinsker's inequality gives $d_{\mathrm{TV}}(P_c, M) \leq \sqrt{\frac{1}{2}D(P_c\|M)}$ for each class, and the triangle inequality together with $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ gives $d_{\mathrm{TV}}(P_0, P_1) \leq \sqrt{2\, I(\tilde{\mathbf{x}}; y)} \leq \sqrt{2(\epsilon + \delta)}$. The balanced accuracy of the optimal classifier is $\frac{1}{2} + \frac{1}{2}\, d_{\mathrm{TV}}(P_0, P_1)$, which yields Equation (8). □

Remark 5. The bound is stated for balanced accuracy and does not transfer to the area under the ROC curve. At total variation distance d the sharp bound on AUC is $\frac{1}{2} + d - \frac{1}{2}d^2$, attained by a uniform shift of one class-conditional distribution, which exceeds the balanced-accuracy bound by $\frac{1}{2}d(1-d)$. We therefore report permutation nulls alongside every AUC.

E ADDITIONAL RESULTS

This appendix follows the order of Section 4. Every value is a mean over folds and then over seeds under the protocol of Section 3.5, and every null is a permutation null with $B = 1000$ unless the caption states otherwise.

E.1 CROSS-SECTIONAL BENCHMARK, RELIABILITY SELECTION AND STAGED REMOVAL

Deep baseline. Table 8 places the linear classifier against BrainNetCNN (Kawahara et al., 2017) on the four cross-sectional benchmark cohorts, under the same subject-grouped folds. With its default configuration the network does not exceed the linear model on any cohort, and tuned as in Table 7 on ADNI full over three seeds it reaches 0.691 ± 0.008 against 0.742.

Table 8: **Linear classifier against BrainNetCNN.** AUC under subject-grouped five-fold cross-validation. The BrainNetCNN column uses the default configuration on all four cohorts. The tuned run was completed on ADNI full only.

Cohort	Contrast	LR-ℓ_2	BrainNetCNN
ADNI full	CN vs AD	0.742	0.726
ABIDE-I	CN vs ASD	0.754	0.728
UCLA-CNP	HC vs SCHZ	0.784	0.702
ds002748	CN vs MDD	0.610	0.570

Reliability selection. Marker studies often restrict an analysis to the edges that are most stable across sessions. We rank the edges by intraclass correlation inside each training fold and classify from the most reliable half. Accuracy rises from 0.734 on all edges to 0.757, which clears both a label permutation ($p = 0.002$) and a null that draws a random subset of the same size ($p = 0.022$). The fitted weights also concentrate on those edges, with a loading fraction of 0.433 against 0.207 under a subject-level label permutation, $p = 0.027$. Selecting for stability therefore selects for the part of the profile that distinguishes people.

Staged removal. A second reading removes the identity components in stages, projecting out the top k of them before classifying, at fifteen values of k. The components removed are the principal components of the training scans ranked by their intraclass correlation, so their number is bounded by the training scans rather than by the subjects, and at $k = 200$ the projection has taken out the entire representation. Between $k = 0$ and $k = 150$ the two quantities move in opposite directions. The disease AUC falls to 0.543 by $k = 30$ and does not recover, while identification rises from 0.780 to 0.880 at $k = 50$ before declining. The rise matches the gain in identifiability reported when a connectome is reconstructed from a truncated set of group-level components (Amico & Goñi, 2018), so the components taken out first carry structure shared across subjects rather than structure specific to one. The two quantities are never low together until the representation itself is exhausted (Figure 6, Table 9).

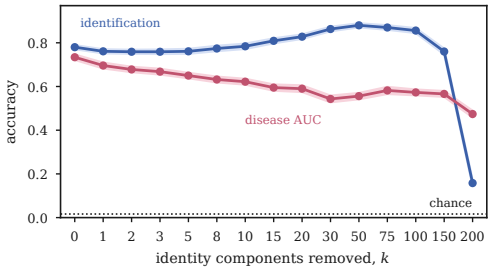


Figure 6: **Disease accuracy and identification move in opposite directions.** Each point is a mean over 20 seeds and five folds, with the shaded ribbon giving one standard error. The dotted line is identification chance, $1/61$.

Table 9: **Staged removal of the identity components, all fifteen values of k.** Mean $\pm$ s.d. over 20 seeds and five folds on ADNI multi-visit. Identification chance is $1/61$.

k	Disease	Ident.	k	Disease	Ident.
0	.734 $\pm$.12	.780 $\pm$.09	20	.590 $\pm$.15	.828 $\pm$.10
1	.696 $\pm$.13	.761 $\pm$.09	30	.543 $\pm$.15	.863 $\pm$.10
2	.678 $\pm$.12	.759 $\pm$.09	50	.556 $\pm$.14	**.880** $\pm$.09
3	.668 $\pm$.13	.759 $\pm$.10	75	.582 $\pm$.16	.870 $\pm$.10
5	.650 $\pm$.12	.761 $\pm$.11	100	.573 $\pm$.13	.856 $\pm$.11
8	.632 $\pm$.13	.774 $\pm$.12	150	.566 $\pm$.13	.760 $\pm$.13
10	.622 $\pm$.13	.784 $\pm$.12	200	.474 $\pm$.15	.158 $\pm$.14
15	.595 $\pm$.14	.809 $\pm$.11			

E.2 THE ABLATION IN FULL

Figure 3 of the main text shows three panels for four levels of the ablation, and this section gives the underlying values. Table 10 reports the operator on three labels. Head motion varies within a subject and is retained. The diagnosis does not vary and is cancelled. Age is continuous and enters as the Spearman correlation between

true age and the age predicted from the profile, for which a permutation of a binary label is not defined, so it is reported as a point estimate whose sign reverses. Table 11 reports the two modalities under two summaries, on the 227 scans that carry a structural session, 217 of them acquired on the same day, matched scan for scan under one set of folds. Two controls accompany it. The modalities differ in dimension, so we reduced the connectivity rate to six dimensions by principal components and repeated it, giving 0.476 against a null 95th percentile of 0.652, while the structural rate at the same dimension reaches 0.804 against 0.664. They also differ before the operator, which is why the comparison is between what remains after the operator in each rather than between the two starting points. Table 12 reports the feature-count control, in which identity-ranked removal is compared against 500 random draws at each k under the same folds and the same classifier. The seven removal methods appear in Figure 3c and their configurations in Table 7. Invariant risk minimisation treats each cohort as an environment and runs leave-one-cohort-out on a different cohort set, reaching 0.471 there. The simulation that calibrates what the residual readouts can detect is in Appendix E.3.

Table 10: **The operator on three labels.** The first two entries are AUC against a permutation null, with the p value before centring and after. Age is a Spearman correlation between true and predicted age, reported as a point estimate.

Label	Before	After	p, before $\rightarrow$ after
head motion	0.577	**0.708**	$0.040 \rightarrow 0.001$
diagnosis	0.734	0.495	$0.003 \rightarrow 0.669$
age, ρ	0.394	-0.194	—

Table 11: **Two modalities under two summaries.** All entries come from the 227 matched scans under one set of folds, 217 of them acquired on the same day. Sixty subjects contribute a slope. A rate of change is one slope per subject and has no value before centring.

Modality	Summary	Before	After	p after
connectivity	per-visit level	0.732	0.492 ± 0.012	0.736
structural	per-visit level	0.909	0.489 ± 0.015	0.743
connectivity	rate of change	—	0.515	0.457
structural	rate of change	—	**0.807**	0.002

Table 12: **Disease AUC after removing k edges under three orderings.** The random ordering is a mean over 500 draws. This run uses 10 seeds, so its $k = 0$ baseline differs from the 20-seed 0.734 by seed set alone.

k	Random order	By $\lvert t_e \rvert$	By identity relevance
0	0.740	0.740	0.740
100	0.740	0.740	0.740
500	0.740	0.741	0.738
1000	0.740	0.737	0.735
5000	0.739	0.724	0.720
10,000	0.739	0.711	0.695

E.3 Within-subject change in full

Tracking cognitive change. Section 4.3 asks whether the within-subject rate of change in connectivity follows the within-subject rate of change in cognition. The cohort is the 57 multi-visit subjects, 206 scans, with a Mini-Mental State Examination score at every visit. For each subject the least-squares slope against age is fitted to every connectivity feature and to the score, a ridge regression with generalised cross-validation maps the connectivity rate to the score's rate inside each fold, and the out-of-fold prediction is compared with the observed rate by Spearman correlation, averaged over 20 fold seeds. The same construction with the subject mean as input and the score's level as target gives the cross-sectional reference. Table 13 gives both, with a bootstrap interval over subjects and a null that permutes the target across subjects. The detection limit is the smallest correlation detectable at 80% power for the available number of subjects. On the 55 subjects who also carry a structural session the structural rate is fitted the same way, and neither modality's rate follows the score

Table 13: **Cross-sectional level against within-subject rate.** Spearman correlation between out-of-fold prediction and observation, bootstrap interval over subjects with 1000 draws, and a subject-level permutation null with $B = 1000$. The limit is the smallest correlation detectable at 80% power, 0.33 at both sample sizes.

Input	Target	n	ρ	95% CI	p
subject mean	MMSE level	57	0.364	[0.182, 0.566]	0.005
connectivity rate	MMSE rate	57	−0.186	[−0.514, 0.046]	0.857
connectivity rate	MMSE rate	55	−0.222	[−0.511, 0.028]	0.903
structural rate	MMSE rate	55	−0.084	[−0.252, 0.305]	0.159

at that sample size. The Clinical Dementia Rating sum of boxes was recorded for the same subjects and gives the same pattern, but 26 of the 57 have no within-subject change in that score, so it is not reported as a rate.

Readouts of the residual. Eight readouts of the ADNI residual, one linear and seven that can see structure a linear classifier cannot, are summarised in Section 4.3. Figure 7 and Table 14 give the values. A simulation calibrates them. We inject a signal into the real feature structure in one of three forms, a difference in the subject mean, a difference in residual variance along one direction, and a difference in the residual trend over time, and record how often each readout detects it at $p < 0.05$ over ten replicates. A mean-only signal is never detected in the residual, which is what the cancellation requires. A variance difference is detected by the kernel classifier, the across-visit variance and the mean residual magnitude once one class varies about five times as widely as the other, at which strength the same injection gives a raw AUC of 0.678. A trend difference is not detected at the strengths tried, so the slope and lag-one readouts are reported without a calibrated limit. The variance readouts therefore resolve a within-subject difference of that size, and the reported values rule one out. The residual is not empty of individual structure, since held-out subjects are matched to themselves from the residual alone at 0.245 against a chance level of $1/61$. What the residual carries is person-specific without being diagnostic.

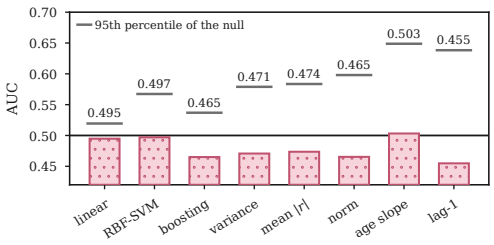


Figure 7: **Eight readouts of the ADNI residual.** Each bar against the 95th percentile of its own subject-level permutation null.

Table 14: **Readouts of the ADNI residual.** Each against a subject-level permutation null with $B = 1000$. The first three are scan-level readouts and the last five are computed per subject. The lag-one readout uses the 49 subjects with at least three visits.

Readout	AUC	Null p95	p
linear	0.495	0.519	0.669
kernel	0.497	0.567	0.234
gradient boosting	0.465	0.537	0.572
across-visit variance	0.471	0.579	0.746
mean residual magnitude	0.474	0.584	0.692
residual norm	0.465	0.598	0.761
slope against age	0.503	0.649	0.443
lag-one correlation	0.455	0.638	0.692

Subspace geometry on six cohorts. Table 15 gives the minimum principal angle between the between-subject and the disease subspaces on every cohort, with the null of Appendix C. Four of the six fall below the fifth percentile of their null. On ADNI, one scan per subject gives 47.70° and every scan of the same 79 subjects gives 60.74°, neither below its null.

E.4 Conversion within a subject

ADNI is the only cohort in this paper that carries a disease label capable of changing within a subject. The analysis uses the multi-visit subset defined by a control-versus-impaired contrast, 252 scans from 77 subjects, in which 5 change label within the observation window and 72 do not. Three of the five move from normal cognition to impairment and two move in the opposite direction, and each carries exactly one scan in the changed state, so five differences cannot support a within-subject test. Table 16 reports what the cohort supports cross-sectionally. Separating the five from the 72 reaches 0.789, which does not clear a subject-level permutation null, and restricting the features to the baseline visit leaves it at 0.572. Splitting the cohort by label change reproduces the result of Section 4.1 among the 72, while the five, whose 17 scans cannot support a five-fold split, return an identical value on every measurement.

Table 15: **Minimum principal angle against a label-permutation null.** $K_{\mathrm{id}} = 20$, $K_{\mathrm{dz}} = 10$, $B = 1000$. The test is lower-tailed, so the fifth percentile of the null is reported. Shaded rows fall below it.

Cohort	Contrast	Observed	Null mean	Null 5th pct	p
ABIDE-I	CN vs ASD	80.17	85.13	84.24	0.001
UCLA-CNP	HC vs SCHZ	58.20	66.81	62.18	0.007
ADHD-200	HC vs ADHD	69.65	78.85	71.52	0.007
COBRE	HC vs SCHZ	57.02	62.24	57.48	0.035
ds002748	CN vs MDD	43.87	44.32	38.28	0.442
ADNI, one scan per subject	CN vs AD	47.70	49.37	43.60	0.312

Table 16: **What the converter cohort supports.** The cross-sectional rows separate the 5 subjects whose label changes from the 72 whose label does not. The subgroup rows report the control-versus-impaired contrast within each group.

Measurement	Group	Value	p
Cross-sectional, from the subject mean			
who changes label, all visits	5 versus 72	0.789	0.067
who changes label, baseline only	5 versus 72	0.572	0.377
present diagnosis	61 subjects	0.728	0.020
Subgroups, raw profile and residual			
raw profile	72 stable	0.713 ± 0.030	—
residual	72 stable	0.494 ± 0.006	—
all measurements	5 changed	0.600	—

E.5 Cross-domain replication

Table 17 gives the values behind Table 3 of the main text. HCP contributes 45 subjects, 31 female and 14 male, each with exactly two resting-state sessions under Schaefer-400 (Schaefer et al., 2018). The residual is not reported, because with two sessions the two residuals are antisymmetric, so a classifier trained on one is trained on the negation of the other. Subjects are matched to themselves from the subject mean alone at 0.667 against a chance level of $1/45$. CREMA-D contributes 7,442 clips from 91 actors, 48 male and 43 female, summarised as 88 eGeMAPSv02 acoustic functionals, with folds grouped by actor. Emotion has no actor-mean value, since all 91 actors portray all six emotions with 76 to 82 clips each, so an actor mean carries no single emotion label. On ADNI the diagnosis is constant within a subject, so the subject mean is well defined over the 61 multi-visit subjects, 24 with Alzheimer's disease and 37 controls, and reaches within 0.01 of the 0.734 available from the whole profile. Table 18 gives the accuracy for each emotion against the remaining five, with anger the most separable and disgust the least.

Table 17: **The three unit-mean estimates.** All three use ℓ_2-regularised logistic regression at $C = 1$, stratified group five-fold cross-validation, 20 seeds, and standardisation fitted on training folds alone. Nulls permute the label across units with $B = 1000$.

Data	Target	AUC	Seed s.d.	Null p95	p
HCP	sex	0.910	0.033	0.749	0.002
CREMA-D	speaker sex	0.995	0.008	0.634	0.001
ADNI	diagnosis	0.728	0.046	0.686	0.020

Table 18: **CREMA-D, one emotion against the rest.** Macro one-versus-rest accuracy on the raw features, each against a within-actor permutation null at $p = 0.001$. Emotion rises from 0.840 on the raw features to 0.880 on the residual, exceeding the raw value in all five seeds. Speaker sex falls from 0.980 to 0.499, and the speaker is identified among the 91 actors at 0.712 against a chance level of 0.011.

Emotion	anger	neutral	sadness	happiness	fear	disgust
AUC	0.925	0.867	0.864	0.825	0.799	0.762