Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Conditioned connectomes outperform the standard correlation kernel for predicting cognition in an independent cohort

Chulmin Pyo

pyo@ai3.kr

AI3 Discovery https://orcid.org/0009-0003-1439-0740

Research Article

Keywords: functional connectivity, kernel ridge regression, partial correlation, transductive learning, cognition prediction, benchmark, Human Connectome Project, Aging Adult Brain Connectome, external validation, foundation models

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11794795/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The author is the head of research of AI3 Discovery, a for-profit company; the author's affiliation, AI3 Discovery, is part of AI3. The research used AI3 Discovery's scientific-discovery harness, described in the manuscript under AI use. The work received no specific funding, and no patent application has been filed in relation to it. The code and results are openly released (code under the MIT licence, result files under CC BY 4.0).

Conditioned connectomes outperform the standard correlation kernel for predicting cognition in an independent cohort

Chulmin Pyo
AI3 Discovery
Corresponding author: discovery@ai3.kr

Abstract

Connectome-based cognition predictions are scored differently across papers. We re-scored a public Human Connectome Project benchmark (753 subjects, six fMRI modalities, 60 family-aware splits) as its authors do, with a public age band in place of exact age, and evaluated a kernel ridge pipeline that conditions each subject's connectome and filters the between-subject kernel's spectrum. With fixed settings it reached r = 0.7215, against 0.6779 for the benchmark paper's best model and 0.6798 for our re-run of its kernel (corrected t = 3.23). This margin is transductive: the filter uses all subjects' unlabelled features, helps only after conditioning, and gives no significant gain fitted on training subjects alone; without it, the difference was not significant. On 263 HCP subjects outside the benchmark, most used in development, a pre-registered comparison favoured the filtered pipeline, computed with them present, over the benchmark's kernel by 0.064 (95% interval 0.030 to 0.099). In an independent cohort never used in development (1,290 adults aged 36 to 89, resting state), the conditioned connectome, fitted on training subjects only, beat the benchmark's kernel as pre-registered (r = 0.5861 against 0.5257; corrected t = 5.47); the filter added nothing there. Against a frozen fMRI foundation model on identical splits, the unfiltered pipeline was ahead in 7 of 8 settings, significantly on resting state (corrected t = 3.88), held-out subjects and ABIDE autism diagnosis, and nominally in AOMIC. Of 18 pre-registered single-decision contrasts, 9 were met, 5 by counting repeats. Two scoring choices each reversed a method ordering.

Keywords: functional connectivity; kernel ridge regression; partial correlation; transductive learning; cognition prediction; benchmark; Human Connectome Project; Aging Adult Brain Connectome; external validation; foundation models

1 Introduction

Kernel ridge regression (KRR) on the edges of functional-connectivity (FC) matrices is a widely used baseline for predicting individual cognition, and on the same data it predicts as well as deep networks [1]. It is the estimator of a public multimodal benchmark on the Human Connectome Project (HCP) [2,3]. With the estimator and the evaluation fixed, the representation of each subject's connectome is the main free choice. Studies of that choice disagree. Benchmarks of resting-state pipelines favoured tangent-space and regularised representations [4,5]. A recent screen, which motivated a spectral pre-training target for fMRI encoders, ranked unregularised partial correlation last among the transforms it tried [6,7]. A second screen also placed a shrinkage partial correlation below plain correlation [8].

Two practical problems make such comparisons hard to read. The first is scoring. The benchmark averages Pearson correlations computed inside each test fold and regresses age and sex from the target inside each training fold [3]. Pipelines are often scored by pooling all predictions of a split instead, on the raw target. The second is development on a fixed benchmark. A pipeline tuned over many rounds on the same subjects can overfit them, information can leak between folds in subtle ways [9], and counting how many resampled splits favour a method overstates certainty, because the splits share subjects [10,11,12].

We report a pipeline that conditions each subject's connectome and then filters the spectrum of the between-subject kernel. We score it the benchmark's way against the benchmark's own models and a re-run of the benchmark's kernel. We then ask which parts of the gain survive four tests: a factorial ablation that separates conditioning from filtering, a version of the filter computed from training subjects only, a pre-registered comparison on HCP subjects outside the

benchmark, and a pre-registered test in an independent cohort never used in development. We also compare the pipeline with a released fMRI foundation model on identical splits and in three further cohorts. Finally, we list the changes tried after the headline settings were fixed and the pre-registered contrasts, met or not.

2 Results

2.1 Scoring choices move the numbers and can reverse an ordering

The same predictions gave different accuracies under different scoring choices (Table 1). Pooling all predictions of a split gave 0.7357 for our pipeline on the raw target. Averaging per-fold correlations gave 0.7326. Regressing age band and sex inside each training fold, as the benchmark does, gave 0.7215. The covariate step cost every arm about the same, 0.0111 for ours.

A second artefact came from the ridge-penalty grid. Our kernels are normalised to unit trace, and our original grid stopped at 1.0×10^{-3}. On that grid the pipeline without the kernel filter scored 0.6789, below the benchmark's kernel (0.6902 on the raw target). On a grid extended to 1.0×10^{-12} it scored 0.7055, so the order of the two reversed. On the extended grid it chose the smallest value in all 600 folds, which is the ridgeless limit: a floor of 1.0×10^{-6} gave 0.7054. The filtered pipeline chose 1.0×10^{-3} in all 600 folds, an interior value of the extended grid, so it did not change. Every comparison of benchmark arms below uses the extended grid.

Our re-run of the benchmark's kernel, the correlation kernel of the CBIG toolbox, matched the benchmark's published single-modality accuracies within 0.0086 when scored on the raw target. The published values are covariate-adjusted, however. Under our covariate step, the resting-state re-run scored 0.5170 against the published 0.5276. For the six fMRI modalities combined, the re-run scored 0.6798, against 0.6779 for the benchmark's best published combination and 0.6665 for its multi-kernel KRR.

Table 1. The same predictions under different scoring choices.

Scoring	Ours, fixed settings	Benchmark kernel, re-run here
Pooled r over all predictions of a split, raw target	0.7357	0.6936
Mean of per-fold r, raw target	0.7326	0.6902
Mean of per-fold r, age band and sex regressed inside each training fold (the benchmark's scoring)	**0.7215**	**0.6798**
Ours without the kernel filter, mean per-fold r, raw target, smallest ridge penalty 1.0×10^{-3}	0.6789	
The same, smallest ridge penalty 1.0×10^{-12}	0.7055	

Mean over 60 splits of 10 family-aware folds, 753 subjects. Raw target: the cognition score as given. The benchmark's scoring regresses sex and age from the target inside each training fold; we used a four-level age band, the benchmark used exact age. The last two rows show an arm whose best penalty lies at the boundary of the grid. The benchmark's kernel also chose the smallest penalty in most folds (517 of 600, covariate-adjusted), but its score was the same with floors of 1.0×10^{-3}, 1.0×10^{-6} and 1.0×10^{-12} (0.6902 on the raw target), so it too sits at the ridgeless limit.

2.2 Under the benchmark's scoring the pipeline exceeds the benchmark's models

Under the benchmark's scoring, the pipeline reached 0.7215 (SD across splits 0.0090) with fixed settings (Figure 1, Table 2). That is 0.0436 above the best model in the benchmark paper and 0.0417 above our re-run of its kernel. The corrected resampled t for the second difference, over all 600 per-fold differences, was 3.23 ($p = 0.0013$). The pipeline was ahead in 524 of 600 folds and in all 60 splits. A naive t over the split means would have been 52.5; the gap shows how much counting splits overstates certainty when the splits share subjects.

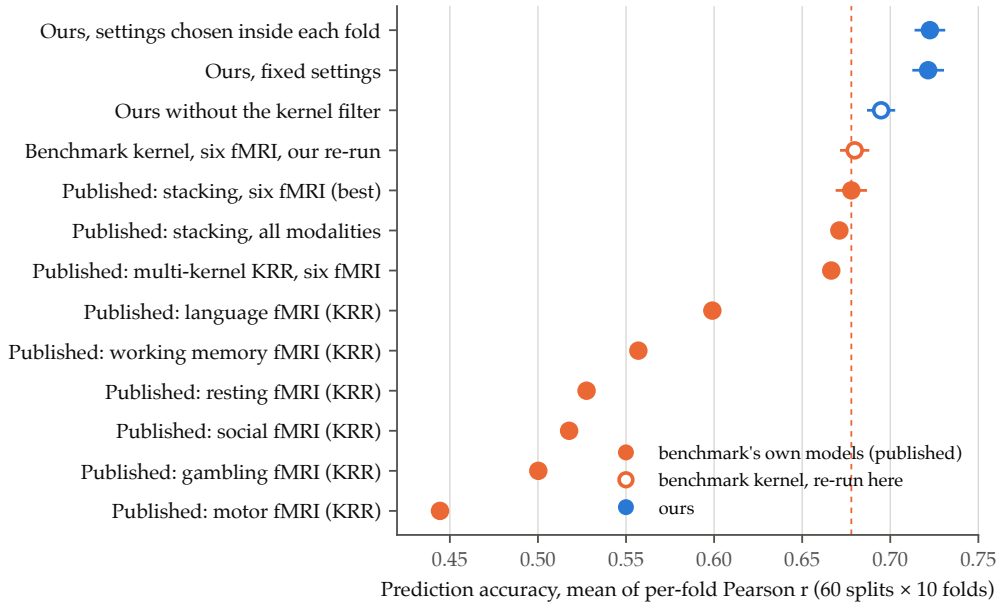


Figure 1. Accuracy on the benchmark under its own scoring. Orange: the benchmark's published models (filled) and our re-run of its kernel (open). Blue: our pipeline with fixed settings, with shrinkage, kept components and ridge penalty re-chosen inside each fold, and without the kernel filter (open). Bars, SD across the 60 splits, shown for the arms compared statistically. The published models used exact age as a covariate and our arms a four-level age band. Dashed line, the best published model.

The pipeline was developed over many rounds on these subjects; the repository holds 544 result files. The in-sample tests in this and the next section are therefore exploratory. To measure part of the selection cost, we re-chose the shrinkage, the number of kept kernel components and the ridge penalty inside every training fold (4 settings by 8 penalties, by inner cross-validation). This gave 0.7225, 0.0009 away from the fixed settings (t = 0.29), and 0.0426 above the re-run benchmark kernel (t = 3.14). The in-fold rule picked one setting in 511 of 600 folds; fixed, that setting scores 0.7252, which we report only as an observation. The re-selection covers four neighbouring settings and cannot bound the selection over the many other choices made during development; only held-out subjects can.

Table 2. Differences under the benchmark's scoring, with Nadeau–Bengio corrected resampled t.

Comparison (benchmark's scoring)	Difference in r	Corrected t	p
Ours, fixed − best published model	0.0436	—	—
Ours, fixed − benchmark kernel re-run	0.0417	3.23	0.0013
Ours, chosen in-fold − benchmark kernel re-run	0.0426	3.14	0.0018
Ours, chosen in-fold − ours, fixed	0.0009	0.29	0.769
Ours without the filter − benchmark kernel re-run	0.0150	1.33	0.184
Kernel filter (ours − ours without the filter)	0.0267	2.55	0.011
Resting state alone, ours without the filter − benchmark kernel re-run	0.0525	4.15	3.9×10^{-5}
Resting state alone, ours − benchmark kernel re-run	0.0629	4.17	3.5×10^{-5}

Differences in mean per-fold r over 600 paired fold scores; t uses the correction for test-to-training ratio 0.1111 [10]. The published models have no fold scores we can pair with ours, so the first row has no test. Exploratory: the pipeline was developed on these subjects.

2.3 The margin is transductive and needs conditioning and filtering together

A factorial ablation scored the benchmark's way separates the two parts (Table 3). Applied to a kernel on full correlations, the kernel filter lowered accuracy (−0.0105, t = −2.15). Applied to the benchmark's own kernel, normalised like ours, it also lowered accuracy, though not significantly (−0.0101, t = −1.85). Only after conditioning did

it help (0.0267, t = 2.55). The interaction, the filter's effect with conditioning minus its effect without, was 0.0372 (t = 3.47). Conditioning without the filter added 0.0160 over our full-correlation kernel (t = 1.42).

Table 3. What the kernel filter needs, under the benchmark's scoring.

Kernel	Without the filter	Filter from all subjects	Filter from training subjects only
Full correlations (no conditioning)	0.6788	0.6683	
Conditioned (shrinkage, partial correlation, eigenvalue power, full-correlation mix)	0.6948	**0.7215**	0.6886
The benchmark's kernel, double-centred and trace-normalised	0.6822	0.6722	

Mean per-fold r, age band and sex regressed in-fold, 60 splits, same ridge grid for every cell. The filter keeps the leading 150 components after a soft threshold. "From training subjects only" computes the eigenvectors and threshold on each training fold and extends them to the test subjects (Nyström extension). Post hoc analyses requested in review.

The filter is computed from the eigenvectors of the kernel over all 753 subjects. It uses no labels, but it uses the test subjects' features. The benchmark's kernel does not: each entry depends only on the two subjects it compares. When we computed the filter from the training subjects of each fold and extended it to the test subjects, the pipeline scored 0.6886. That is −0.0063 against no filter (t = −1.70) and 0.0087 against the benchmark's kernel (t = 0.72). In this implementation, the in-sample margin therefore depends on computing the filter with the test subjects present. The inductive arm still centres the kernel over all subjects, and it tests one way of extending a truncated spectrum to test subjects.

Adding the steps one at a time, in one implementation (Figure 2), our kernel on full correlations scored −0.0010 against the benchmark's kernel (t = −0.20). Shrinkage and partial correlation added 0.0194 (t = 1.69). Raising each subject's partial-correlation eigenvalues to a small power then changed accuracy by −0.0076 (t = −3.84). Mixing back 5% of the full-correlation kernel added 0.0042 (t = 2.10), and the filter 0.0267. A sequential decomposition depends on the order of the steps, and with the filter absent the power step only hurt, so we do not attribute the margin to any single per-subject step. Weighting the language task double, a choice made during development, added 0.0016 (t = 0.56); with equal weights the margin over the benchmark's kernel was 0.0401 (t = 3.04).

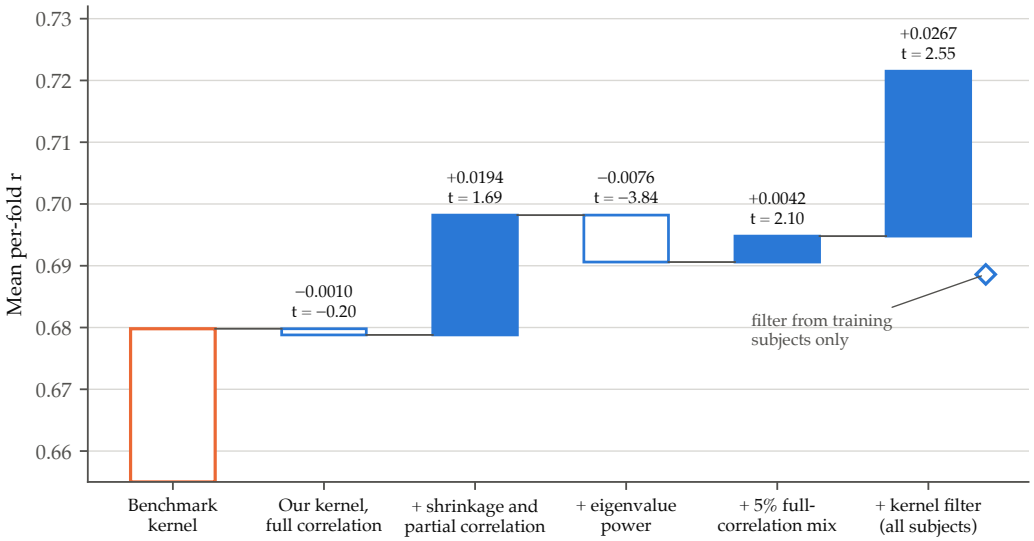


Figure 2. The steps of the pipeline under the benchmark's scoring, in one implementation. Each bar adds one step to the previous one; labels give the change and the corrected resampled t over 600 fold differences (filled, gain; open, loss). The diamond shows the final step with the filter computed from training subjects only. The vertical axis starts above zero. Post hoc.

2.4 Resting state alone

On resting-state connectivity alone, the pipeline without the filter, which uses no labels but centres the kernel over all subjects, scored 0.5695 under the benchmark's scoring. That is 0.0419 above the benchmark's published resting-state model (0.5276) and 0.0525 above our re-run of its kernel (0.5170; $t = 4.15$, $p = 3.9 \times 10^{-5}$). With the filter it scored 0.5799. On resting state, conditioning without the filter gave a significant margin, unlike in the six-modality ensemble.

2.5 On held-out subjects the pipeline beats the benchmark's kernel; the filter's share is unresolved

To test the advantage over the benchmark's kernel on subjects outside the benchmark's 753, we pre-registered one comparison on the 263 HCP subjects outside the benchmark that have the target and all six modalities. Their connectivity came from our own extraction (450 regions), the extraction also used for the training subjects. Each arm was trained once on the 753 benchmark subjects, with settings chosen by 10-fold cross-validation inside them. D1 was run as in our benchmark analysis: its kernels and filter were computed over the training and held-out subjects' features together, without labels, so the held-out subjects' unlabelled features entered it. D0 and the benchmark's kernel are inductive: a held-out subject's kernel entries depend only on that subject and the training subjects. The decision rule was fixed in advance: confirmed only if the lower bound of the 95% bootstrap interval of the difference is above zero. Two arms had to reproduce an earlier held-out run exactly before any comparison was read, and both did.

This test was not blind for its main arm. The filtered pipeline (D1) and its unfiltered variant (D0) had been scored on these subjects two days earlier, in a pre-registered comparison with a foundation model. Only the three baseline arms were new. Nor were the subjects new to the project. In the first days of development, 242 of the 263 were part of a larger cross-validated cohort of 974 subjects, the benchmark's subjects with features plus 246 outside it, whose target values were used for training and testing. Partial correlation and the mixing in of full correlations were first selected in that cohort; at least 43 result files come from it and from explorations of all HCP subjects. Later, at least 27 result files used subjects outside the benchmark as held-out test subjects for candidate changes (dropping the motor task, a second atlas, a seventh task, a joint prediction, clipping extreme edges), none of which was adopted; the motor result informed keeping the motor task. In 10 files their target values entered training, in analyses that enlarged the training cohort and in one change that used their mean target as an intercept prior (Table S1), and in 5 only their unlabelled features were used. The held-out test is therefore a test on subjects outside the benchmark's splits, not on subjects unseen during development. One earlier run, on 233 of these subjects with a second extraction, widened its ridge grid after a poor held-out score (0.6741 before, 0.7721 after) and fixed a bug in its inductive arm after seeing results.

The held-out pipeline also differs from the one scored on the benchmark. It uses our extraction rather than the benchmark's features and equal modality weights rather than language ×2. Its shrinkage was chosen from a grid that excludes the benchmark value, and both arms picked an edge of it (0.9 for D1, 0.98 for D0). Its ridge grid ran from 1.0×10^{-6}, and D0 and the benchmark's kernel both chose that floor. It uses the raw target and a pooled correlation over the held-out subjects.

D1 scored $r = 0.7863$ on the held-out subjects, against 0.7222 for the benchmark's kernel (Figure 3). The difference of 0.064 (95% interval 0.030 to 0.099) met the pre-registered rule. D0, which centres with training-subject statistics only and has no filter, scored 0.7696, 0.047 above the benchmark's kernel (0.021 to 0.076; a secondary contrast). Against a plain correlation kernel under the same protocol, the margins were larger (0.079 and 0.060). D1 minus D0, which adds the filter and the transductive kernel together, was 0.017 (−0.008 to 0.041) and was not resolved. Among the 244 held-out subjects least similar to any training subject, the primary difference was 0.062 (0.027 to 0.102). This similarity filter removes only a few subjects. On the benchmark's own features, a threshold of this kind caught 8.4% of the family pairs inferred from the benchmark's splits (Methods).

Conditioning without the filter gained more on these subjects than on the benchmark's own features, where the gain was not significant (0.0160, $t = 1.42$, Table 3; on the raw target 0.0175, $t = 1.54$). The difference appears to follow the extraction and the settings chosen with it rather than the held-out setting. On our extraction, within the benchmark's subjects and splits, D0 led a plain correlation kernel by 0.0568 in mean per-fold r (corrected $t = 5.65$; raw target, equal modality weights) and by 0.0512 in pooled r, the metric of the held-out test. That is close to its margin over the plain kernel on the held-out subjects (0.060). On our extraction D0 chose a shrinkage of 0.98 in 577 of 600 folds, against the fixed 0.92 used on the benchmark's features.

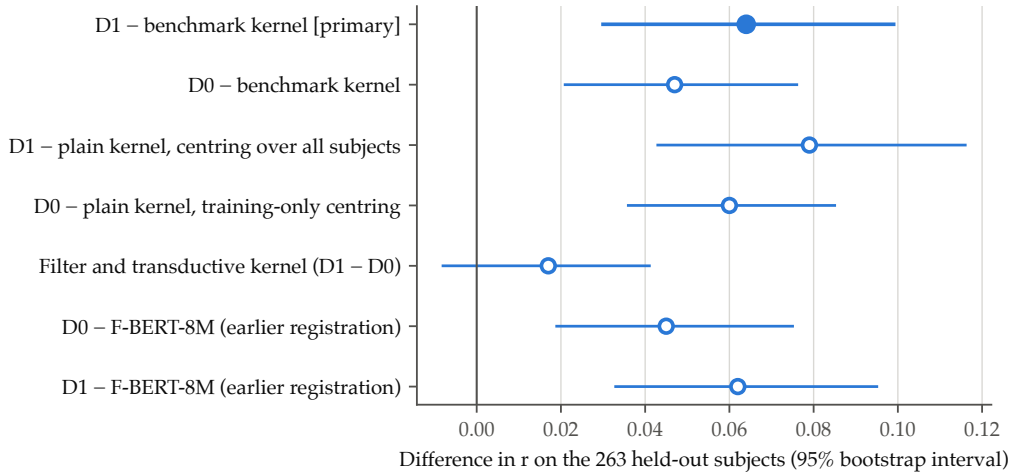


Figure 3. Paired differences on the 263 held-out subjects. Points, difference in held-out Pearson r; lines, 95% subject-bootstrap intervals (10,000 resamples). Filled point: the primary comparison of this paper's pre-registration; the two comparisons with F-BERT-8M come from an earlier pre-registration on the same subjects. D1 is the filtered pipeline on our extraction with equal modality weights, its kernels and filter computed over training and held-out subjects together; D0 is its unfiltered variant with training-only centring. The bootstrap treats subjects as independent and ignores families.

2.6 Against an fMRI foundation model on identical splits

F-BERT encodes resting-state time series with a transformer pre-trained to reproduce an eigenvalue-powered connectome [6]. We ran its released weights frozen on the same subjects, time series and splits as our pipeline, from our 450-region extraction, and scored every arm on the raw target by its mean per-fold score (Table 4). In these comparisons D0 and D1 centre the correlation kernel over all subjects of the cohort, without labels, as in our benchmark analysis; only on the held-out subjects is D0 inductive. The corrected t values come from reruns that stored every fold's score and reproduced the original runs (Methods).

Across the 8 settings of Table 4, D0 was ahead of F-BERT-8M by mean in 7, significantly in 4: resting state, the held-out subjects, ABIDE diagnosis and, just above the threshold, AOMIC. It was behind only for ADHD-200 IQ, not significantly. This count is descriptive, because the HCP settings share subjects. On resting state, the unfiltered variant D0 scored 0.5913, against 0.5235 for F-BERT-8M (difference 0.0677, corrected $t = 3.88$). D0 tied F-BERT's pre-training target, the connectome with its eigenvalues raised to the power 0.35 (0.5911; $t = 0.03$). That target beat the encoder by the same margin ($t = 4.14$), and the encoder was ahead of a plain correlation kernel (0.4853; $t = -2.00$). The kernel filter did not help on resting state (-0.0084, $t = -0.87$). On a short scan, the first 200 time points of one resting run, D0 was ahead of F-BERT-8M in all 60 splits, which met the pre-registered split-count rule, but not significantly by the corrected test (0.0307, $t = 1.24$); it was level with F-BERT-1M ($t = 0.08$), and F-BERT-8M was ahead of a plain correlation kernel ($t = -2.31$).

With six modalities, neither variant was significantly ahead of F-BERT's concatenated embeddings (0.7026). D0 led by 0.0049 ($t = 0.52$) and the filtered pipeline D1 by 0.0166 ($t = 1.40$). Scored as one pooled correlation per split, as first computed, D0 was behind F-BERT (-0.0074), and the pre-registered contrast was not met (Table S2). This is a further case in which the scoring choice changed an ordering. On the 263 held-out subjects, both variants beat F-BERT (Table 4 and Figure 3).

In ABIDE-I [13], with whole sites held out, D0 beat F-BERT on autism diagnosis (AUC 0.773 against 0.734; $t = 3.20$). On full-scale IQ it was ahead in all 20 repeats but not significantly ($t = 1.89$). In AOMIC-ID1000 [14], whose subjects were part of F-BERT's pre-training data, D0 beat F-BERT on intelligence (0.4176 against 0.3752; $t = 2.04$, just above the threshold for significance). On ADHD-200 [15], with leave-one-site-out evaluation across 6 sites, neither difference was resolved: 0.022 AUC for diagnosis (-0.023 to 0.068) and -0.035 for IQ (-0.124 to 0.050). Within its largest site (256 subjects), a post hoc analysis found F-BERT ahead on diagnosis (0.622 against 0.550 AUC). A pre-registered diagnostic

did not trace this edge to head motion: a motion index alone reached an AUC of 0.417, and removing it left 0.060 of the 0.072 margin. No closed-form temporal feature reached F-BERT (best 0.562), so by the rule fixed in advance the edge is read as temporal or motion signal that the static connectome lacks. In every external cohort the filtered pipeline scored below the unfiltered one, significantly so in AOMIC (−0.0389, t = −2.44). Brain-JEPA [16], also run frozen, scored 0.067 on HCP resting state (mean of pooled per-split r). Finally, we pre-trained encoders with F-BERT's architecture on its public corpus, with our conditioned connectome as the teacher, at three sizes. The first was ahead of F-BERT-8M by mean in 4 of 9 cells, and combined systems that included our encoders met their pre-registered rule in none of their cells (Table S2).

Table 4. Our pipeline against F-BERT-8M on identical subjects, inputs and splits.

Setting	Subjects	D0	D1	Powered FC	F-BERT-8M	D0 vs F-BERT	D1 vs F-BERT
HCP resting state, benchmark splits	753	0.5913	0.5829	0.5911	0.5235	t = 3.88	t = 3.13
HCP six modalities, benchmark splits	753	0.7075	0.7193	0.6849	0.7026	t = 0.52	t = 1.40
HCP held-out subjects, six modalities	263	0.7696	0.7863		0.7243	0.019 to 0.075	0.033 to 0.095
AOMIC-ID1000, intelligence	879	0.4176	0.3788	0.4225	0.3752	t = 2.04	t = 0.16
ABIDE-I, autism diagnosis (AUC)	1,035	0.773	0.771	0.781	0.734	t = 3.20	t = 3.38
ABIDE-I, full-scale IQ	966	0.257	0.250	0.261	0.206	t = 1.89	t = 1.79
ADHD-200, diagnosis (AUC)	435	0.570	0.536	0.566	0.548	−0.023 to 0.068	−0.057 to 0.032
ADHD-200, IQ	369	0.105	0.064	0.130	0.139	−0.124 to 0.050	−0.182 to 0.034

Mean per-fold Pearson r on the raw target (AUC where marked); for the 263 held-out subjects and ADHD-200, pooled r or AUC over the subjects held out. HCP rows use the benchmark's subjects and splits with our 450-region extraction. D0, unfiltered variant; D1, filtered pipeline; both are centred over all subjects of the cohort, without labels, except D0 in the held-out row, which is inductive; powered FC, each subject's connectome with its eigenvalues raised to the power 0.35, which is F-BERT's pre-training target. The last two columns give the corrected resampled t over all paired per-fold differences. For the held-out subjects and ADHD-200, they give instead the 95% subject-bootstrap interval of the difference in pooled r or AUC; ADHD-200 is leave-one-site-out, and its intervals ignore site clustering. AOMIC-ID1000 and ABIDE-I used 10-fold cross-validation repeated 20 times, ABIDE with whole sites per fold. The six-modality rows compare against F-BERT's concatenated embeddings, its stronger six-modality option.

2.7 In an independent cohort, conditioning holds and the filter adds nothing

The held-out HCP subjects were not new to the project, so we pre-registered a test in a cohort that had never been used: the first visits of the Aging Adult Brain Connectome (AABC) Release 2, which extends the HCP in Aging [17,18]. The protocol, five arms, analysis code and decision rule were committed before any phenotype file was downloaded. The connectivity was the provider's resting-state full correlation on 379 regions, the HCP multimodal cortical parcellation plus subcortical structures [19], rather than the benchmark's features. The target was the NIH Toolbox total cognition composite [20]. Of 1,390 first visits, 1,290 had the target and covariates; they were aged 36 to 89 (mean 60.2) and came from 1,271 families at 4 sites. Every arm was scored the benchmark's way: 10-fold splits that keep families together, repeated 20 times, with age, age², sex and site regressed inside each training fold and the corrected t over all 200 per-fold differences. Unlike in the benchmark analysis, the primary arm D0-I is inductive. Its edges are centred on the training-subject mean before each subject's vector is scaled to unit norm, and its shrinkage is chosen inside each training fold.

D0-I scored 0.5861, against 0.5257 for the benchmark's kernel (Table 5). The difference of 0.0604 (corrected t = 5.47, p = 1.3×10^{-7}) met the pre-registered rule. Conditioning accounts for most of the margin: D0-I led a plain correlation kernel with the same training-only centring by 0.0527 (t = 5.53). The kernel filter added nothing, whether it was computed from training subjects only (−0.0091, t = −1.21) or over all subjects together with all-subject centring, as on the benchmark (0.0024, t = 0.32). The shrinkage chosen for D0-I was the lowest value of its grid, 0.80, in 80 of 200 folds and the next, 0.85, in 100; the benchmark's fixed 0.92 was chosen in 12.

Table 5. Pre-registered contrasts in the independent cohort (AABC Release 2).

Arm	Mean per-fold r
Benchmark kernel (CBIG correlation kernel)	0.5257
Plain correlation kernel, training-only centring	0.5334
Conditioned, training-only centring (D0-I)	**0.5861**
Conditioned, filter from training subjects only (D1-I)	0.5770
Conditioned, filter from all subjects (D1-T, transductive)	0.5884

Contrast	Difference in r	Corrected t	p	Repeats positive
D0-I − benchmark kernel (primary)	0.0604	5.47	1.3×10^{-7}	20 of 20
D0-I − plain kernel (conditioning)	0.0527	5.53	1.0×10^{-7}	20 of 20
D1-I − D0-I (filter, inductive)	−0.0091	−1.21	0.226	1 of 20
D1-T − D0-I (filter and all-subject centring)	0.0024	0.32	0.747	14 of 20
D1-T − benchmark kernel	0.0628	4.89	2.1×10^{-6}	20 of 20

Mean per-fold Pearson r over 10 family-aware folds repeated 20 times, 1,290 first visits, resting state; age, age², sex and site regressed inside each training fold. Corrected resampled t over all 200 paired per-fold differences, test-to-training ratio 0.1111 [10]. The primary contrast was confirmed only if the difference was above zero and t was at least 1.972; the others were declared secondary, with no claim of their own. I, edges centred on the training-subject mean; T, kernel centring and filter computed over all subjects, without labels.

2.8 What did not help

After the headline settings were fixed, we tried at least 14 further changes to the chain, some in several variants (Table S1). Two hand-set equal mixtures, of the chain kernel with a rank-based kernel and with a clipped-edge kernel, reached nominal significance (t = 2.17 and 2.33), without correction for the number of changes tried; no other change improved accuracy significantly. Two lowered it: averaging predictions over several ridge penalties (t = −4.22) and, nominally, replacing the conditioned kernel with one built from eigenvalue-powered connectomes, F-BERT's target (t = −1.99, the largest of four weights tried). Mixed into the chain at a small weight instead, that kernel changed nothing (−0.0003, t = −0.73). The largest positive change, clipping extreme edges, was 0.0081 (t = 1.57). A planned test of it on held-out subjects was voided after its result had been recorded and interpreted, when an execution error was found: every modality had been built from the resting-state features. Adding a seventh HCP task at a fixed share gave 0.0026 (approximate t = 1.47). On the benchmark's other two behavioural components, an earlier version of the chain, tuned for cognition, scored below a plain correlation kernel, though not significantly: 0.2821 against 0.3157 for emotion recognition (approximate t = −1.10) and 0.0743 against 0.0845 for life satisfaction (t = −0.32; pooled r, raw target, original ridge grid).

3 Discussion

Of the 18 pre-registered contrasts with a single decision (Table S2), 9 were met and 9 were not; 5 of the 9 met were decided by counting the repeats in which the difference was positive, which is not a test. The unmet ones are the six-modality comparison of the per-subject pipeline with F-BERT and two rounds that tried to rescue it; three refinements tested on held-out subjects (an average with a second extraction, a three-atlas average and a seventh task); dropping the motor task, within the benchmark splits and on held-out subjects; and ADHD-200 diagnosis. Of the contrasts with a rule per cell, the in-fold choice between D0 and F-BERT's target was met in 4 of 4 cells, the combined system in 2 of 9 cells and its last attempt on ADHD-200 in 0 of 2; combined systems that included our own pre-trained encoders met their rule in none of their cells, and the encoder alone was behind F-BERT-8M by mean in 5 of 9 cells. Of 9 secondary contrasts declared with the main rule, 6 were met.

3.1 What is established

Under the benchmark's scoring, with a public age band in place of exact age, the pipeline exceeds the best model in the benchmark paper by 0.0436. It exceeds our re-run of the benchmark's kernel by 0.0417 (corrected t = 3.23). This margin is transductive. It requires the kernel filter computed over all subjects, and the filter helps only on conditioned

connectomes. Computed from training subjects alone, it gives no gain, and without it the remaining margin is not significant. On held-out subjects, with our own extraction and without covariate adjustment, the filtered pipeline, computed over training and held-out subjects together, beat the benchmark's kernel as pre-registered. The inductive unfiltered variant did so as well in a secondary comparison. The filter's share was not resolved, and the filter scored lower in all three external cohorts, significantly only in AOMIC. The inductive unfiltered variant also kept a margin on these subjects, but most of them were in the cohort in which the conditioning steps were first chosen. The independent cohort gives the stronger evidence. In 1,290 adults never used in development, scanned at 4 sites and processed with a different parcellation, the conditioned connectome fitted on training subjects only beat the benchmark's kernel as pre-registered (corrected $t = 5.47$). There the filter added nothing, whether fitted on training subjects or computed, with all-subject centring, over all subjects including the test subjects. A gain from the filter has been shown only on the benchmark's own subjects.

3.2 Where the gain comes from

The decomposition does not support a claim that partial correlation, or any single per-subject step, is decisive. Shrinking a correlation matrix towards the identity keeps its eigenvectors, so its inverse is a spectral re-weighting of the correlation matrix, apart from the final diagonal normalisation. At heavy shrinkage, the conditioned connectome therefore belongs to the same family as the eigenvalue-powered connectome that F-BERT is trained to reproduce. That would explain why our per-subject variant tied F-BERT's target on resting state. The filter itself is a transductive spectral transform of the between-subject kernel. Such transforms are established in machine learning, for example in the spectral regularisation of graph kernels and in spectral regularisation algorithms for supervised learning [21,22]. We are not aware of one in connectome-based prediction. On this benchmark it added to conditioned kernels and lowered accuracy on unconditioned ones. In the independent cohort, conditioning accounted for 0.0527 of the 0.0604 margin over the benchmark's kernel (D0-I against the plain kernel, $t = 5.53$), so the part that generalised is the per-subject chain as a whole, not the filter.

3.3 Scoring and statistics

Three lessons apply beyond this pipeline. First, scoring choices moved accuracy by amounts comparable to differences between methods. Most of them moved every arm alike, but two reversed an ordering: a boundary in the ridge grid (Table 1), and pooled against per-fold correlation in the six-modality comparison with F-BERT. A comparison across papers is only meaningful once these choices match. Second, counting the splits a method wins is not a test. Our naive t over split means was 52.5, against a corrected 3.23. Approximating the corrected t from split means also overstates it, so every corrected t in the main tables uses fold scores. Third, a correction can itself be implemented wrongly. During this work, a helper that divided the variance by the number of fold scores and also multiplied it by the correction term inflated t values by a factor of about 18 in several analyses. It was found because the corrected t exceeded the naive t, which a correction cannot do; every value here uses the correct formula.

3.4 Limitations

The benchmark margin is transductive and in-sample. Our covariate step used a public four-level age band, whereas the benchmark used exact age. The comparator is our re-implementation of the benchmark's kernel, not the authors' code. Under our covariate step, its resting-state version scored below the published value, while its six-modality version scored above the published combinations; how this affects the margin is unclear. Regressing a four-level band instead of exact age leaves within-band age variance in the target, which connectivity can predict; this may favour our arms against the published models, but not against the re-run kernel, which used the same band. The pipeline was developed over many rounds on the benchmark's subjects, and the in-fold re-selection covers only four combinations of shrinkage and kept components.

The held-out test was pre-registered and run once, but it was not blind for its main arm, and that arm was itself computed with the held-out subjects' unlabelled features present. Most of the held-out subjects (242 of 263) were in the cross-validated cohort in which the conditioning steps were first chosen, and subjects outside the benchmark appear in at least 85 earlier result files. The held-out pipeline differs from the benchmark pipeline in extraction, weights, shrinkage grid, target and metric. As its pre-registration noted, the held-out subjects' target is more spread out than

the benchmark's, and every arm scored higher on them than in-sample, so held-out and in-sample values are not comparable. Our task time series were not ICA-FIX denoised, whereas our resting-state series were. Relatives can cross the split, the bootstrap ignores family structure, and connectome similarity cannot detect relatives [23]. The benchmark's family identifiers are restricted data that we did not use. The target scores come from the benchmark's released component file; we do not know on which subjects the component model was estimated.

The independent-cohort test covers one cohort, one target and resting state only. It tests the single-modality conditioned kernel with its shrinkage chosen in each fold, not the six-modality pipeline with fixed settings that produced the benchmark number. It used the provider's parcellation and denoising rather than the benchmark's, and the NIH Toolbox composite rather than the benchmark's component score. Unlike the original protocol, it used no task modality, no global-signal sensitivity analysis and no motion-based exclusion or covariate: only the provider's resting-state connectivity was downloaded, and the phenotype file has no head-motion variable. The shrinkage of D0-I fell on the two lowest values of its grid in most folds, so a wider grid might have scored differently. The comparator is again our implementation of the benchmark's kernel. While files were being chosen, values of two cognition factor scores, which are not the target, were seen for about ten visits; the pre-registration records this.

F-BERT's forward pass was rebuilt from its released checkpoint because its code was not available. All released foundation models were evaluated frozen. ADHD-200 intervals ignore site clustering. The key numbers were reproduced by separate implementations written in separate AI-assisted sessions; they share data and design decisions and are not verification by a second research group.

3.5 Outlook

The independent cohort supplied the test that our 263 held-out subjects could not, though with the provider's processing rather than the benchmark's. It separates the two parts of the pipeline: the conditioned connectome generalised inductively in this cohort, and the filter added nothing out of sample. The same inductive comparison can now be run unchanged on other cohorts and targets, on task fMRI and in clinical populations.

4 Methods

4.1 Data and benchmark

The benchmark uses 753 HCP young-adult subjects [2] and a cognition component score as the target. It provides FC matrices for resting state and five tasks (language, social, gambling, working memory and motor) on a 400-region cortical parcellation [24], and 60 random splits of the subjects into 10 folds that keep families together [3]. We used the benchmark's released features, subject list, splits, per-split results and behavioural component scores. Every comparison with its published numbers therefore uses identical subjects and folds. The component scores of the held-out subjects come from the same released file; the component model was estimated by the benchmark's authors, and we did not refit it.

4.2 Covariates

Sex and a four-level age band for all 753 subjects came from the HCP open-access subject table. Inside each training fold, the target was regressed on sex and the age-band indicators by ordinary least squares, and the fitted coefficients were applied to the test fold. The subject-level covariate file is not redistributed; the released code builds the covariates from the open-access table that users obtain from the HCP.

4.3 Pipeline

Each subject's correlation matrix was shrunk towards the identity with a fixed weight of 0.92 and inverted, and the inverse was converted to partial correlations [7,25]. The shrinkage weight is a fixed setting, not the Ledoit–Wolf estimate. Negative eigenvalues of the partial-correlation matrix were set to zero, and the remaining eigenvalues were raised to the power 0.1. Across subjects, the vectorised upper triangles were turned into a correlation kernel: columns centred over all subjects, rows scaled to unit norm, and the kernel double-centred and scaled to unit trace.

The kernel spectral filter was applied to each modality's conditioned kernel before averaging. It subtracted 0.5 times the standard deviation of the eigenvalues, set negative values to zero, kept the leading 150 components and restored unit trace. A kernel built in the same way from the full correlations, without the filter, was mixed in with weight 5%. The six modality kernels were averaged with the language task weighted ×2, a weight chosen during development.

KRR chose its ridge penalty by 4-fold cross-validation inside each training fold, from 13 values between 1.0×10^{-12} and 10. The in-fold re-selection used values from 1.0×10^{-6} to 10. The variant without the filter is otherwise identical. On the 263 held-out subjects, D0 centres and scales with training-subject statistics only; everywhere else, and for D1 on those subjects, the kernel is centred over all subjects present. The inductive filter of Table 3 computed the eigen-decomposition of each training block, applied the same threshold and truncation, normalised by the training trace, and extended the filtered kernel to test subjects through the training eigenvectors (Nyström extension).

4.4 The benchmark's kernel, re-run

For each modality, the CBIG correlation kernel (the Pearson correlation between two subjects' edge vectors) was computed from the benchmark's features. The six kernels were averaged with equal weights and fitted with the same KRR and penalty selection as our pipeline. In Table 3, this kernel was also double-centred and trace-normalised so that the filter could be applied to it.

4.5 Metric and statistics

For the benchmark, accuracy is the Pearson correlation between predicted and observed scores inside each test fold, averaged over the 10 folds of a split and then over the 60 splits. Paired differences were tested with the Nadeau–Bengio corrected resampled t over all per-fold differences [10,11]. The test-to-training ratio was 0.1111 for the benchmark, and the degrees of freedom were the number of fold scores minus one. The in-fold re-selection used a grid of shrinkage values 0.90 and 0.92, kept components 150 and 200, and 8 ridge penalties. Its criterion was inner 4-fold cross-validation on the residualised target of the training fold. The levers of Table S1 were scored on the raw target, except where noted. The seventh-task row uses the pooled correlation, and its t is approximated from split means.

4.6 Held-out test

The pre-registration fixed five arms, the protocol, a sanity gate and the decision rule, and was committed to the project repository before the run. The test subjects were the 263 HCP subjects outside the benchmark that had the target and all six modalities. Training used the 753 benchmark subjects with the same extraction. As fixed in the pre-registration, D1's kernels and filter were computed over training and test subjects together, without labels; D0 and the benchmark's kernel were inductive. The target was not residualised, so that the gate could reproduce the earlier held-out run. Differences in held-out Pearson r were bootstrapped over subjects (10,000 resamples). The similarity-filtered subset kept the test subjects whose highest connectome similarity to any training subject was below the 90th percentile of the highest similarities among training subjects. To gauge such filters, we inferred family pairs inside the benchmark as pairs never separated across its 60 family-aware splits and measured how many a threshold of this kind flags on the benchmark's features.

4.7 Independent-cohort test

AABC Release 2 was obtained from BALSA under the AABC Consortium Data Use Terms. The pre-registration was fixed for this cohort and committed before any phenotype file was downloaded; it is in the released repository with its outcome note. We used the provider's resting-state full-correlation matrices of the first visit (V1), computed after its removal of structured artefacts, on 379 regions (71,631 edges per subject). All 1,390 V1 matrices were complete, and 100 visits without the target, age, sex or site were excluded. The target was the uncorrected standard score of the NIH Toolbox Cognition Total Composite at V1 [20]. Inside each training fold it was regressed on age, age², sex and site indicators, and the coefficients were applied to the test fold. The phenotype file has no head-motion variable, so none was used. Folds kept pedigrees together and were balanced within site.

Five arms shared the KRR, the ridge grid (14 values from 1.0×10^{-12} to 10 on trace-normalised kernels) and inner 4-fold penalty selection on the residualised training target. The benchmark's kernel was the CBIG correlation kernel on the edges. PLAIN-I is a correlation kernel on the raw edges and D0-I the same kernel on conditioned connectomes, both with edges centred on the training-fold mean and each subject's vector scaled to unit norm. The conditioning was the benchmark pipeline's (eigenvalue power 0.1, full-correlation mix 5%), with the shrinkage chosen inside each training fold from 7 values between 0.80 and 0.99. D1-I adds the kernel filter fitted on the training block and extended to test subjects by the Nyström extension; D1-T computes the filter over all subjects, as in the benchmark analysis. The rule was fixed in advance: the primary contrast, D0-I minus the benchmark's kernel, is confirmed only if the mean per-fold difference is above zero and the corrected t is at least 1.972 (two-sided 5% over 200 fold scores). Before the run, the full analysis was executed once with a synthetic target, covariates and families (mean per-fold r 0.03 to 0.07 across arms). With 8 further synthetic targets, one repeat each, the per-target means ranged from −0.075 to 0.083 and averaged −0.0006 to 0.0095 per arm. The phenotype file's header was then reduced to variable names so that the frozen code could read it, with every row unchanged, and this step was recorded before any value was read. The analysis ran once.

4.8 Foundation-model comparison

All arms read the same resting-state and task time series from our extraction: the 400-region cortical parcellation plus a subcortical atlas [24,26], band-passed and z-scored as F-BERT's model card prescribes. F-BERT [6] is the name its authors give the encoder; it was run frozen from its released weights. We report the better of two linear probes, a choice that favours F-BERT. Among forward-pass variants consistent with the released checkpoint, we used the one whose loss on the pre-training objective came closest to the value recorded with the checkpoint (0.229 against 0.2115), a label-free choice. On HCP resting state, the rebuilt encoder scored 0.5247 as the mean of pooled per-split correlations (0.5235 as the mean per-fold r), against 0.502 reported in the preprint for a larger subject list. In the six-modality comparison, our arms average one kernel per modality, while F-BERT's embeddings were concatenated, its stronger option. D0 chose its shrinkage inside each training fold. AOMIC-ID1000 [14] and ABIDE-I [13] used 10-fold cross-validation repeated 20 times, ABIDE with whole sites per fold. ADHD-200 [15] used leave-one-site-out evaluation and a subject bootstrap. The original runs stored one pooled score per split. To test with every fold-level difference, we re-ran each script unchanged except that it also stored every arm's per-fold scores; each rerun reproduced the stored pooled scores before any of its values were used. Accuracy in these comparisons is the mean per-fold score on the raw target, and the corrected t uses all paired per-fold differences, with the test-to-training ratio taken from the recorded fold sizes for AOMIC and ABIDE-I.

4.9 Implementation and verification

All analyses used NumPy and SciPy. Every number in this paper is generated by a script from the stored result files (Table S3 lists them, their fingerprints and the scripts that produced them), and every verbal comparison is checked against those numbers when the paper is built. The headline value 0.7215 was reproduced to four decimals by two separately written implementations of the residualised scoring in the repository (0.7215 in the second).

Declarations

Funding. This research received no specific funding.

AI use. The AI3 Discovery scientific-discovery harness and large language models (LLMs) were used in this research. The author is responsible for all content, and no AI tool is listed as an author.

Competing interests. The author is affiliated with AI3 Discovery, whose scientific-discovery harness is described under AI use. No patent application has been filed in relation to this work.

Data and code availability. HCP data are available from ConnectomeDB under the HCP data use terms; the benchmark's features, splits and results are available from its authors' public repository [3]. AOMIC, ABIDE-I and ADHD-200 are public. AABC Release 2 is available from BALSA under the AABC Consortium Data Use Terms; it is not redistributed, and the released AABC result files contain only per-fold scores and aggregate sample statistics.

Code, the result files with per-fold and per-split scores, the numbers index behind every number in this paper and the pre-registrations with their addenda and outcome notes are available at https://github.com/ai3inc/connectome-cognition-benchmark and archived at Zenodo [27]; code is under the MIT licence and result files under CC BY 4.0. Subject-level fields were removed from the released result files, and rebuilding the numbers index from the released files reproduces every value in this paper. The HCP open-access table used for the covariates is not redistributed.

Ethics. This is a secondary analysis of previously collected, de-identified data; no new data were collected and no participant was contacted. Each original study obtained informed consent from its participants and approval from its institutional review boards or ethics committees, as described in the dataset papers [2,13,14,15,17]. No additional ethics review was sought for this analysis. HCP, AOMIC, ABIDE-I and ADHD-200 data were used under their data use terms. AABC Release 2 was accessed through BALSA after acceptance of the AABC Consortium Data Use Terms and was used for non-commercial research only.

Acknowledgements. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Data, methods used, and/or research reported in this publication were provided in whole or in part by the Aging Adult Vulnerability and Resiliency in the Aging Adult Brain Connectome (AABC) project (U19AG073585) and the Human Connectome Project in Aging (HCP-A, U01AG052564) funded by the National Institute of Aging of the National Institutes of Health. HCP-A was further supported by funds provided by the McDonnell Center for Neuroscience at Washington University in St. Louis.

References

1. He T, Kong R, Holmes AJ, Nguyen M, Sabuncu MR, Eickhoff SB, et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* 206, 116276 (2020). doi:10.1016/j.neuroimage.2019.116276
2. Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: An overview. *NeuroImage* 80, 62-79 (2013). doi:10.1016/j.neuroimage.2013.05.041
3. Ooi LQR, Chen J, Zhang S, Kong R, Tam A, Li J, et al. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *NeuroImage* 263, 119636 (2022). doi:10.1016/j.neuroimage.2022.119636
4. Dadi K, Rahim M, Abraham A, Chyzhyk D, Milham M, Thirion B, et al. Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage* 192, 115-134 (2019). doi:10.1016/j.neuroimage.2019.02.062
5. Pervaiz U, Vidaurre D, Woolrich MW, Smith SM. Optimising network modelling methods for fMRI. *NeuroImage* 211, 116604 (2020). doi:10.1016/j.neuroimage.2020.116604
6. Marraffini G, Shevchenko V, Barbano CA, Wassermann D. *Flattening the Connectome Spectrum: A Spectral Filter for FC Induces a Pretraining Target for fMRI Encoders*. arXiv (2026). doi:10.48550/arxiv.2609.37642
7. Smith SM, Miller KL, Salimi-Khorshidi G, Webster M, Beckmann CF, Nichols TE, et al. Network modelling methods for FMRI. *NeuroImage* 54, 875-891 (2011). doi:10.1016/j.neuroimage.2010.08.063
8. Marraffini G, Mahuas G, Borrell T, Shevchenko V, Wassermann D. *The Variance Brain Foundation Models Forgot: Third-Order Statistics Predict Cognition Where Billion-Parameter Models Fail*. arXiv (2026). doi:10.48550/arxiv.2606.04010
9. Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications* 15, 1829 (2024). doi:10.1038/s41467-024-46150-w
10. Nadeau C, Bengio Y. Inference for the Generalization Error. *Machine Learning* 52, 239-281 (2003). doi:10.1023/a:1024068626366
11. Bouckaert RR, Frank E. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. *Lecture Notes in Computer Science* 3-12 (2004). doi:10.1007/978-3-540-24775-3_3
12. Varoquaux G. Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* 180, 68-77 (2018). doi:10.1016/j.neuroimage.2017.06.061
13. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry* 19, 659-667 (2014). doi:10.1038/mp.2013.78

14. Snoek L, van der Miesen MM, Beemsterboer T, van der Leij A, Eigenhuis A, Steven Scholte H. The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data* 8, 85 (2021). doi:10.1038/s41597-021-00870-6

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

16. Dong Z, Li R, Wu Y, Nguyen TT, Chong JSX, Ji F, et al. *Brain-JEPA: Brain Dynamics Foundation Model with Gradient Positioning and Spatiotemporal Masking*. arXiv (2024). doi:10.48550/arxiv.2409.19407

17. Bookheimer SY, Salat DH, Terpstra M, Ances BM, Barch DM, Buckner RL, et al. The Lifespan Human Connectome Project in Aging: An overview. *NeuroImage* 185, 335-348 (2019). doi:10.1016/j.neuroimage.2018.10.009

18. Harms MP, Somerville LH, Ances BM, Andersson J, Barch DM, Bastiani M, et al. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *NeuroImage* 183, 972-984 (2018). doi:10.1016/j.neuroimage.2018.09.060

19. Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, et al. A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171-178 (2016). doi:10.1038/nature18933

20. Weintraub S, Dikmen SS, Heaton RK, Tulsky DS, Zelazo PD, Bauer PJ, et al. Cognition assessment using the NIH Toolbox. *Neurology* 80, S54-S64 (2013). doi:10.1212/wnl.0b013e3182872ded

21. Smola AJ, Kondor R. Kernels and Regularization on Graphs. *Lecture Notes in Computer Science* 144-158 (2003). doi:10.1007/978-3-540-45167-9_12

22. Lo Gerfo L, Rosasco L, Odone F, De Vito E, Verri A. Spectral Algorithms for Supervised Learning. *Neural Computation* 20, 1873-1897 (2008). doi:10.1162/neco.2008.05-07-517

23. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience* 18, 1664-1671 (2015). doi:10.1038/nn.4135

24. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex* 28, 3095-3114 (2018). doi:10.1093/cercor/bhx179

25. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88, 365-411 (2004). doi:10.1016/s0047-259x(03)00096-4

26. Tian Y, Margulies DS, Breakspear M, Zalesky A. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature Neuroscience* 23, 1421-1432 (2020). doi:10.1038/s41593-020-00711-6

27. Pyo C. *connectome-cognition-benchmark: code, per-fold results and pre-registrations*. Zenodo (2026). doi:10.5281/zenodo.24355565

Supplementary material

Table S1. Changes to the chain tested after the headline settings were fixed.

Change to the chain, tested after the headline was fixed	Difference in r	Corrected t
RELATIONAL task added at a fixed one-ninth share; fixed setting; four fold-internal rules failed	0.0026	1.47
Edge clipping at 1.5 robust SD (covariate-matched); held-out test void (execution error)	0.0081	1.57
Spearman (rank) kernel	0.0054	1.06
Winsorised target (2.5 SD); scored against the transformed target	0.0010	0.60
Kernel-target alignment weights; 40 seeds, no paired test	−0.0010	—
Label-Gram kernel augmentation (weight 0.10)	−0.0025	−0.82
Mean target of the labelled subjects outside the 753 as intercept prior	−0.0026	−1.04
Auxiliary composite target	−0.0040	−0.79
Rank-normalised target; scored against the transformed target	−0.0056	−1.51
Chain kernel replaced by the filtered kernel of eigenvalue-powered connectomes (power 0.35, weight 1); largest of four weights tried	−0.0079	−1.99
The same kernel mixed into the chain at weight 0.10; weights 0.25 and 0.50 also not significant	−0.0003	−0.73
Equal mix of the chain kernel and the Spearman kernel (hand-set); fixed weight chosen by hand	0.0057	2.17
Equal mix of the raw-edge and clipped-edge kernels (hand-set); fixed weight chosen by hand; raw target, no covariate matching	0.0060	2.33
Averaging predictions over six ridge penalties	−0.0511	−4.22

Mean per-fold r against the fixed chain, 60 splits, raw target unless noted; corrected resampled t over per-fold differences. Exceptions: edge clipping used the benchmark's covariate step; the seventh-task row is a pooled correlation with t approximated from split means. Post hoc.

Table S2. Pre-registered contrasts and their outcomes.

Pre-registered contrast	Rule as written	Observed	Met
Single-decision contrasts			
HCP resting state, D0 vs F-BERT-8M (main contrast)	positive in at least 57 of 60 repeats	60 of 60	yes
HCP six modalities, D0 vs F-BERT-8M (A)	positive in at least 57 of 60 repeats	9 of 60	no
Six modalities, in-fold choice among eighteen kernels vs F-BERT-8M (B)	positive in at least 57 of 60 repeats	10 of 60	no
Six modalities, removing a shared cross-modality component, vs F-BERT-8M (C)	positive in at least 57 of 60 repeats	7 of 60	no
Held-out subjects, D1 vs F-BERT-8M (D)	bootstrap P(difference > 0) at least 0.95, and above both F-BERT probes	P = 1.000	yes
Held-out subjects, half D1 and half F-BERT vs F-BERT-8M (E)	bootstrap P(difference > 0) at least 0.95	P = 1.000	yes
Held-out intersection, average with a second extraction vs D1 (F; rule amended after both components' held-out r was known)	bootstrap P(difference > 0) at least 0.95	P = 0.853	no
Held-out subjects, three-atlas average vs D1 (G)	bootstrap P(difference > 0) at least 0.95	P = 0.304	no
AOMIC-ID1000, D0 vs F-BERT-8M (H)	positive in at least 19 of 20 repeats	20 of 20	yes

Pre-registered contrast	Rule as written	Observed	Met
HCP resting state, D0 vs Brain-JEPA (I)	positive in at least 57 of 60 repeats	60 of 60	yes
ABIDE-I autism, D0 vs F-BERT-8M (J)	positive in at least 19 of 20 repeats	20 of 20	yes
ADHD-200 diagnosis, D0 vs F-BERT-8M (L)	bootstrap P(difference > 0) at least 0.95	P = 0.834	no
Dropping the motor task, D1 (R)	positive in at least 57 of 60 repeats	26 of 60	no
Dropping the motor task, held-out subjects, D0 (S)	bootstrap P(difference > 0) at least 0.95	P = 0.519	no
Adding the RELATIONAL task, D1, benchmark splits (T)	positive in at least 57 of 60 repeats	57 of 60	yes
Adding the RELATIONAL task, D1, held-out subjects (T)	bootstrap P(difference > 0) at least 0.95	P = 0.716	no
Held-out subjects, D1 vs the benchmark kernel (this paper)	95% interval above zero	0.030 to 0.099	yes
Independent cohort (AABC), inductive conditioned kernel vs the benchmark kernel (this paper)	difference above zero and corrected t at least 1.972	t = 5.47	yes
Contrasts with a rule per cell			
D0 and its pre-training target chosen in-fold vs F-BERT-8M, per cohort (K)	the addendum's rule in every cell	4 of 4 cells	yes
Combined in-fold system vs F-BERT-8M and D0, per cell (N)	the addendum's rule in every cell	2 of 9 cells	no
The two ADHD-200 cells, combined system vs both F-BERT models and D0, per cell (O)	the addendum's rule in every cell	0 of 2 cells	no
Combined system with our own pre-trained encoder in the pool, per cell (P)	the addendum's rule in every cell	0 of 9 cells	no
Combined system with ensembles of encoders in the pool, per cell (Q1)	the addendum's rule in every cell	0 of 7 cells	no
Combined system with a larger own encoder in the pool, per cell (Q2)	the addendum's rule in every cell	0 of 9 cells	no
Combined system with the largest own encoder in the pool, per cell (Q3)	the addendum's rule in every cell	0 of 9 cells	no
Our own pre-trained encoder alone vs F-BERT-8M, per cell (P)	the addendum's rule in every cell	behind F-BERT-8M by mean in 5 of 9 cells (per-repeat counts not stored)	no
Secondary contrasts declared with the main rule			
HCP resting state, D0 vs F-BERT-1M	positive in at least 57 of 60 repeats	60 of 60	yes
HCP resting state, D1 vs F-BERT-8M	positive in at least 57 of 60 repeats	60 of 60	yes
HCP resting state, plain FC vs F-BERT-8M	positive in at least 57 of 60 repeats	0 of 60	no
HCP resting state, powered FC vs F-BERT-8M	positive in at least 57 of 60 repeats	60 of 60	yes
Short scan, D0 vs F-BERT-8M	positive in at least 57 of 60 repeats	60 of 60	yes
Short scan, D0 vs F-BERT-1M	positive in at least 57 of 60 repeats	31 of 60	no

Pre-registered contrast	Rule as written	Observed	Met
Short scan, plain FC vs F-BERT-8M	positive in at least 57 of 60 repeats	0 of 60	no
Short scan, powered FC vs F-BERT-8M	positive in at least 57 of 60 repeats	60 of 60	yes
Short scan, D0 vs plain FC	positive in at least 57 of 60 repeats	60 of 60	yes
Registered, not decided			
What F-BERT's ADHD-200 within-site edge reflects (M, diagnostic with a reading rule)	reading rule fixed in advance	motion index (DVARS) alone AUC 0.417; margin 0.072 before and 0.060 after removing it; best closed-form temporal feature 0.562 against F-BERT 0.622	reading: temporal or motion signal the static connectome lacks
Clipping extreme edges, held-out subjects (separate pre-registration)	decision rule fixed in advance	voided: every modality had been built from resting-state features	—
Chain vs plain FC, held-out subjects (separate pre-registration)	decision rule fixed in advance	superseded before it ran	—

Rules as written in the pre-registration before each run, except where a row notes an amendment; the letters refer to its addenda, which are in the released repository. Rows decided by repeats count the splits or repeats in which the difference is positive. This paper reports corrected tests from fold scores where they exist (Table 2, Table 4), and some rows met by counting repeats are not significant by that test. The short scan is the first 200 time points of one resting run. P and Q pre-trained encoders with F-BERT's architecture on its public corpus, with our conditioned connectome as the teacher, and added them to the combined system of N. Brain-JEPA [16] predicted little as a frozen feature extractor, and its row carries little weight.

Table S3. Result files behind the numbers in this paper.

File	SHA-256 (16)	Produced by	Earlier use
CANON_NUMBERS.json	abf6a0c0ac2cecdd	values copied from the F-BERT preprint's table, or a file count	
code/canonical.py	34791855b1fd0852	the code itself	
code/fbert/eval_holdout.py	9e87767341b8e6f9	the code itself	
code/fbert/eval_holdout_chain.py	23b7f0a1d80b3df6	the code itself	
code/fbert/eval_samefolds.py	9829c7b3ec02e83f	the code itself	
code/fbert/hcp_prep.py	c6bd29de02b7faee	the code itself	
results/ADDITIVE_CHAIN_RESIDUALISED_1118.json	7bfc4ba57bc7c857	not in the repository	
results/AGE_DECOMPOSITION_1115.json	fc3d68d472e753c9	not in the repository	early CV cohort
results/ALPHA_ON_TOP_1119.json	38da6a4167aa338d	code/alpha_on_top.py	
results/ALPHA_SWEEP_FINAL_1115.json	78402ea3e31e11ea	not in the repository	early CV cohort
results/AUX_TARGET_1118B.json	db8bae3b14d559aa	code/aux_target.py	
results/BCHO.json	ec04c7da72db0385	not in the repository	held-out
results/CANONICAL.json	cd446c53d20ace13	not in the repository	early CV cohort
results/CANON_HOLDOUT_233_1117.json	7a82be9853ba718e	not in the repository	held-out
results/CHAIN_VS_OOI_KERNEL_263_1119.json	8bdfb003dca47fe4	code/fbert/eval_holdout_chain.py	
results/CKA961_b.json	64314608ff6ab52b	not in the repository	features only
results/CKA_2X2_FILTER_G_1117.json	c764d4890cfe2957	not in the repository	features only
results/CKA_2x2.json	79c24c490d8bcf54	not in the repository	features only

File	SHA-256 (16)	Produced by	Earlier use
results/CKA_SHRINKAGE_1115.json	38cf455dabdd11ca	not in the repository	early CV cohort
results/CKA_SIX_KERNELS_961_1117.json	79f3b0b61a1ce2bc	not in the repository	features only
results/CKA_SWEEP_G.json	028c69933eaa155a	not in the repository	features only
results/CLIP_COV_GATE_1118B.json	5c23a030212379fd	code/clip_cov_gate.py	
results/CLIP_GATE_1118B.json	a2807bc4fda85987	code/clip_gate.py	
results/CLIP_HOLDOUT_VALID_1119.json	0e32989a27f355a7	code/clip_holdout_valid.py	held-out
results/CPM_BASELINE_1115.json	eb6dac6bc9702b45	not in the repository	early CV cohort
results/CPM_WITH_ALPHA_1115.json	68acd55eba6b1d57	not in the repository	early CV cohort
results/CROSSTARGET_WEIGHTING_1115.json	47c630ca9beb8fce	not in the repository	early CV cohort
results/DIMENSION_WEIGHTED_ENSEMBLE_1115.json	ad45bd8def679033	not in the repository	early CV cohort
results/EXTRACTION_PAIR_FOLD_INTERNAL_HOLDOUT_1117.json	581fd7980058d217	not in the repository	held-out
results/EXTRA_SUBJECTS_1118B.json	f244b8039a7af518	code/extra_subjects.py	labels in training
results/EXT_AABC_1119.json	bd1a7324d65f2e98	code/external/ext_eval.py	
results/EXT_AABC_DRYRUN_SYNTHETIC_1119.json	46f933d1809bd087	code/external/ext_eval.py (--synthetic)	
results/EXT_AABC_NULLCHECK_1119.json	d00c0b4b57b6fa78	code/external/ext_eval.py (--synthetic, DRY_NULL_SEEDS)	
results/EXT_AABC_SAMPLE_1119.json	eef08306611d2756	code/external/aabc_sample.py	
results/FAMILY_DETECTOR_SENSITIVITY_1117.json	ad719d1335468f0c	not in the repository	
results/FBERT_ABIDE_1117.json	76bea37c12650896	code/fbert/abide_eval.py	
results/FBERT_ABIDE_FOLDS_1119.json	3c58ae76536afd1f	code/fbert/fold_scores.py (rerun of the original script)	
results/FBERT_ADHD_1117.json	44c5e5e8239b9d8b	code/fbert/adhd_eval.py	
results/FBERT_ADHD_MOTION_1117.json	f731066d429fa043	not in the repository	
results/FBERT_ADHD_NYU_WITHIN_1117.json	9dfd51b875e4aff0	code/fbert/adhd_within.py	
results/FBERT_AOMIC_1117.json	1c73a60ef7dc8152	code/fbert/aomic_eval.py	
results/FBERT_AOMIC_FOLDS_1119.json	41cf52b8941ff1e2	code/fbert/fold_scores.py (rerun of the original script)	
results/FBERT_BJEPA_1117.json	47806912279a71a6	code/fbert/eval_bjepa.py	
results/FBERT_HOLDOUT_1117.json	b0831b710fac7da2	code/fbert/eval_holdout.py	held-out
results/FBERT_HOLDOUT_FAMILY_1117.json	53be6c84a43dfa50	not in the repository	held-out
results/FBERT_HOLDOUT_HYB_1117.json	9416ed4fe3aa6221	code/fbert/eval_holdout_hyb.py	held-out
results/FBERT_JOINT_HOLDOUT_F_1117.json	45a53ee3682b7826	not in the repository	held-out
results/FBERT_MULTIATLAS_1117.json	f7da9a92ca1a2a0e	code/fbert/eval_multiatlas.py	held-out
results/FBERT_P_HCP_rest_k0.json	b3665632445bd6ba	not in the repository	
results/FBERT_P_HCP_six_k0.json	4ee77b4090794714	not in the repository	
results/FBERT_P_abide_k0.json	7f26bbc687258704	not in the repository	
results/FBERT_P_adhd_k0.json	2403d79247f9d21d	not in the repository	

File	SHA-256 (16)	Produced by	Earlier use
results/FBERT_P_adhd_nyu_k0.json	d0cf22613c314ae1	not in the repository	
results/FBERT_P_aomic_k0.json	eb86f3b8f4cbc99b	not in the repository	
results/FBERT_Q1_abide.json	36f0f76b99e7192c	not in the repository	
results/FBERT_Q1_adhd.json	384caab9e0316ea6	not in the repository	
results/FBERT_Q1_adhd_nyu.json	0ec8afb2c947a9d4	not in the repository	
results/FBERT_Q1_aomic.json	384e1d37ef6c62d7	not in the repository	
results/FBERT_Q2_HCP_rest_k0.json	24975cbca2b9ac53	not in the repository	
results/FBERT_Q2_HCP_six_k0.json	fd2726a8c2dc8f76	not in the repository	
results/FBERT_Q2_abide_k0.json	c57dee9fefb309b9	not in the repository	
results/FBERT_Q2_adhd_k0.json	b2cdb25eadea0172	not in the repository	
results/FBERT_Q2_adhd_nyu_k0.json	5b0ab27423d2b21f	not in the repository	
results/FBERT_Q2_aomic_k0.json	0ca0bb876f68ca19	not in the repository	
results/FBERT_Q3_HCP_rest_k0.json	7d8321ec1bfd88ce	not in the repository	
results/FBERT_Q3_HCP_six_k0.json	2467b309dc884d2c	not in the repository	
results/FBERT_Q3_abide_k0.json	c276a002c0b1d5ae	not in the repository	
results/FBERT_Q3_adhd_k0.json	8f1a4cf4928eacdf	not in the repository	
results/FBERT_Q3_adhd_nyu_k0.json	f1441ce060372622	not in the repository	
results/FBERT_Q3_aomic_k0.json	e28354911177590e	not in the repository	
results/FBERT_R_MOTOR_1117.json	539877148f429ebd	code/fbert/motor_drop.py	
results/FBERT_SAMEFOLDS_1117.json	7ad51b53d0721cd6	code/fbert/eval_samefolds.py	
results/FBERT_SAMEFOLDS_FOLDS_1119.json	7d912be7016617ab	code/fbert/fold_scores.py (rerun of the original script)	
results/FBERT_SIXMOD_1117.json	e2477290be4de65c	code/fbert/eval_sixmod.py	
results/FBERT_SIXMOD_DECOR_1117.json	56eca90df1e6dbf1	code/fbert/eval_sixmod_decor.py	
results/FBERT_SIXMOD_FOLDS_1119.json	9e372e7d3a624c18	code/fbert/fold_scores.py (rerun of the original script)	
results/FBERT_SIXMOD_PLUS_1117.json	562410166cfe8474	code/fbert/eval_sixmod_plus.py	
results/FBERT_SOTA_ABIDE_1117.json	972a432e18ae97e3	code/fbert/sota_single.py	
results/FBERT_SOTA_ADHD_1117.json	cca3d4037298c76f	code/fbert/sota_single.py	
results/FBERT_SOTA_ADHD_NYU_1117.json	61342efd1e9b1afb	code/fbert/sota_single.py	
results/FBERT_SOTA_AOMIC_1117.json	b99319f8fdae780f	code/fbert/sota_single.py	
results/FBERT_SOTA_HCP_rest_1117.json	f72f5cb27b4b845d	code/fbert/sota_hcp.py	
results/FBERT_SOTA_HCP_six_1117.json	2c607041af10e745	code/fbert/sota_hcp.py	
results/FBERT_SOTA_O_adhd_1117.json	9d400976aae5ec3e	code/fbert/sota_single.py	
results/FBERT_SOTA_O_adhd_nyu_1117.json	4e08eae1b02b4c31	code/fbert/sota_single.py	
results/FBERT_S_MOTOR_HOLDOUT_1117.json	9b2e71ad60f463cd	code/fbert/motor_holdout.py	held-out
results/FBERT_T_HOLDOUT_1117.json	6e93a2eaa963310c	code/fbert/rel_eval.py (holdout)	held-out
results/FBERT_T_INFOLD_1117.json	b0018a2eaa048cd2	code/fbert/rel_eval.py (infold)	
results/FBERT_UNION_SUMMARY_1117.json	21d278bd565d3ee4	summary of the in-fold union arms of the cohort runs	
results/FBERT_VARIANTS_1117.json	99cb808417dca587	not in the repository	
results/FINAL_METHOD_1115.json	268b2776e5a659b8	not in the repository	early CV cohort

File	SHA-256 (16)	Produced by	Earlier use
results/FOLDNB_ABIDE_DX_1119.json	1443b30fced24221	code/fbert/fold_nb.py, from the per-fold scores of the rerun	
results/FOLDNB_ABIDE_FIQ_1119.json	bfc52a3e8621f21f	code/fbert/fold_nb.py, from the per-fold scores of the rerun	
results/FOLDNB_AOMIC_1119.json	aada6679dd03907d	code/fbert/fold_nb.py, from the per-fold scores of the rerun	
results/FOLDNB_SAMEFOLDS_1119.json	f4de2226510f5e9d	code/fbert/fold_nb.py, from the per-fold scores of the rerun	
results/FOLDNB_SIXMOD_1119.json	b49719856ce32502	code/fbert/fold_nb.py, from the per-fold scores of the rerun	
results/HO263.json	d785d5caee8df714	not in the repository	held-out
results/HO263B.json	4ae4303550ae5c15	not in the repository	held-out
results/KERNEL_COUNT_1115.json	695fa9d19f3298d8	not in the repository	early CV cohort
results/KERNEL_TRANSFORMS_1115.json	8736e5d8effba317	not in the repository	early CV cohort
results/KTA_STRUCT_PRIOR_1118B.json	64c894944dccb03b	code/kta_struct_prior.py	
results/LAMRULE.json	b17a3c51857d0c36	not in the repository	held-out
results/LITERATURE_PARCELLATION_VALIDATION_1115.json	25577e99a8e456d0	not in the repository	early CV cohort
results/LOSO_AGE_ROBUSTNESS_1115.json	757e8af8461d72a8	not in the repository	early CV cohort
results/METRIC_CHECK_1118.json	421be5cf474c047c	code/fbert/metric_check.py (original ridge grid)	
results/METRIC_CHECK_COV753_1118.json	bc42a1e9e567c17a	code/fbert/metric_check.py (VARIANTS=1)	
results/METRIC_CHECK_COV753_REST_1118.json	76e63a8323c1fb31	code/fbert/metric_check.py (REST_ONLY=1)	
results/METRIC_CHECK_LAMGRID_1118.json	8f8d16660dd34d55	not in the repository	
results/MIXED_PARTIAL_BEST_1115.json	c7b400e49b710e4b	not in the repository	early CV cohort
results/MKL_ENSEMBLE_1115.json	56758fcb412bc0c8	not in the repository	early CV cohort
results/MOTORHO.json	617e850843913984	not in the repository	held-out
results/MOTORREP.json	16c2d68b9c5087e9	not in the repository	held-out
results/MOTOR_DROP_HOLDOUT_233_1117.json	23e7e50b14a28e0b	not in the repository	held-out
results/MULTITARGET_1115.json	fbf01cea8fe5156a	not in the repository	early CV cohort
results/MULTITARGET_FIXED_1115.json	be2fe13ec21286c5	not in the repository	early CV cohort
results/MULTITASK_REJECTED_1115.json	ad04bb1d51c5befa	not in the repository	early CV cohort
results/NESTED_EXPONENT_1115.json	470de5b15a1ae59f	not in the repository	early CV cohort
results/NESTED_MATCHED_1118.json	02315af3ec7f7096	code/fbert/nested_matched.py	
results/PAIRED_TEST_1115.json	615dc92f57ac62fc	not in the repository	early CV cohort
results/PARTIAL_ONLY_DISCOVERY_1115.json	6e5d72129b5a36a5	not in the repository	early CV cohort

File	SHA-256 (16)	Produced by	Earlier use
results/POOL_OVERLAP_1119.json	20ddde88a0ca85ad	code/fbert/pool_overlap.py	
results/POPULATION_LEAK_CONTROL_1117.json	a72e260a28e48f2b	not in the repository	labels in training
results/POPULATION_SCALING_1117.json	b7eeba8bf6649849	not in the repository	labels in training
results/POP_LEAKCONTROL_a.json	8800e64eee14052f	not in the repository	labels in training
results/RANKLAM_1118B.json	dcdfd2707501884a	code/rank_lam.py	
results/RANK_HOLDOUT_1118B.json	9fe7d3abe0eacf75	not in the repository	held-out
results/RELATIONAL_BAGGED_SUBJECT_BOOTSTRAP_1117.json	c37d2e8bf8ed5cf3	not in the repository	held-out
results/RELATIONAL_BEST_SUBSET_1117.json	083f845613605e1a	not in the repository	held-out
results/RELATIONAL_REPEATED_HOLDOUT_1117.json	59e607b8dc5b75a7	not in the repository	held-out
results/RELBAG.json	c37d2e8bf8ed5cf3	not in the repository	held-out (identical copy)
results/RELHO.json	85cd87619afcb4df	not in the repository	held-out
results/RELHO2.json	37bc0a2b4c26bedb	not in the repository	held-out
results/RELREP.json	c076cef22c1a37a4	not in the repository	held-out
results/REVIEW_ARMS_1119.json	185be36b0f89fd15	code/fbert/review_arms.py	
results/REVIEW_CORRECTED_T_1118.json	6cb91ec28e0242a7	code/fbert/corrected_t_review.py	
results/ROBUSTNESS_1115.json	67ee6062f339d032	not in the repository	early CV cohort
results/ROBUSTNESS_BATTERY_1115.json	42a7658d6c550d58	not in the repository	early CV cohort
results/RS_AT_THEIR_COHORT_SIZE_1117.json	5dccc2f42ac95109	not in the repository	labels in training
results/SEVENTH_TASK_RELATIONAL_1117.json	5b2af9399b2e5a1b	not in the repository	held-out
results/SIX961_LEAK_CONTROL_1117.json	13e0202c0ddbfb66	not in the repository	labels in training
results/SIX961_NESTED_1117.json	ece47c6ba1c8999c	not in the repository	labels in training
results/SIX961_WIDE_GRID_1117.json	2e07a9a81b3097f3	not in the repository	labels in training
results/SIX974_a.json	5e6561b1365e11db	not in the repository	labels in training
results/SIX_MODALITY_961_1117.json	081fec56d03ee3a1	not in the repository	labels in training
results/SOTA_TARGET_CORRECTED_1115.json	c5598dadcaf7ba61	not in the repository	early CV cohort
results/SURFACE_PARCEL_TRANSFER_1115.json	943d254533d16e14	not in the repository	early CV cohort
results/SURFACE_TRANSFER_DECISIVE_1115.json	922aeede4519198b	not in the repository	early CV cohort
results/TARGET_TRANSFORMS_1115.json	bbfe0bb7ce2ae553	not in the repository	early CV cohort
results/TARGET_TRANSFORM_1118B.json	ab43e334a86c360d	code/target_transform.py	
results/THREEHO.json	e94fd5f6bceb9d0a	not in the repository	held-out
results/THREE_EXTRACTION_HOLDOUT_233_1117.json	dc0eb788aed83803	not in the repository	held-out

File	SHA-256 (16)	Produced by	Earlier use
results/THREE_TARGETS_1116.json	87dec61b6fa7f1e7	not in the repository	
results/TOPK_VS_WEIGHTING_1115.json	8bebc1f065b00799	not in the repository	early CV cohort
results/TWENTYFIVE_SEED_VERDICT_1115.json	f79ebc442000019f	not in the repository	early CV cohort
results/final_200.json	a5624f10cc980a9a	not in the repository	early CV cohort
results/full_feat.json	b6847aa19cf15dab	not in the repository	early CV cohort
results/fusion_hcp.json	b424706bf272d2be	not in the repository	early CV cohort
results/hcp_cognition_full1003.json	63cc0874ed022578	not in the repository	early CV cohort
results/hcp_netmats_report.json	fa1b8f967816c372	not in the repository	early CV cohort
results/hcp_perm_1003.json	a568ffff429c9ca1	not in the repository	early CV cohort
results/int_full.json	156574fc3a491a1d	not in the repository	early CV cohort
results/intervals_hcp.json	3d7ba384bc46efa6	not in the repository	early CV cohort
results/loso_hcp.json	0d2acafadd57396c	not in the repository	early CV cohort
results/nonlinear_hcp.json	b51609338a2a0dfb	not in the repository	early CV cohort
results/residual_hcp.json	db107d6358df096e	not in the repository	early CV cohort
results/sota_crossed.json	7d0e3bda4034ab62	not in the repository	early CV cohort
results/struct_from_fc.json	f582596d91009fea	not in the repository	early CV cohort
results/sweep_hcp.json	9e8edd49f61fe901	not in the repository	early CV cohort

Every file the build script opens, including files read only for checks and the earlier analyses that used subjects outside the benchmark (cited for their count and kind); the fingerprint identifies the exact version used. "Not in the repository" means that the script ran on a compute node and was not kept; the file records its own settings. Earlier use: how a file used subjects outside the benchmark (early CV cohort, a cross-validated development cohort that included them with labels; held-out, as test subjects; labels in training; features only, unlabelled). The private repository holds 544 result files in all; the released versions of these files have subject-level fields removed, so their fingerprints differ.