THE VARIANCE BRAIN FOUNDATION MODELS FORGOT: A THIRD-ORDER BASIS FOR PREDICTING INDIVIDUAL DIFFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain foundation models (BFMs) are self-supervised Transformers pretrained on fMRI data. We posit that these models should capture each subject's cognitive performance from their fMRI signal. Yet across three published BFMs and every readout we test, they predict cognition worse than a linear regression from the $\sim$80K parameters of the functional connectivity matrix (FC). Scale does not help: BrainLM's 650M model predicts cognition worse than its 111M. We attribute this to a variance allocation problem: BFM pretraining captures the variance components that dominate fMRI but not the higher-order structure that predicts cognition. Our per-cumulant analysis of the reconstructed signal shows that the second-order covariance is partially preserved, while the third-order co-skewness tensor is largely destroyed. To recover what BFMs lose, we design a linear pipeline that projects the fMRI signal into the subspace that best preserves its co-skewness and computes FC there. This improves over raw FC and exceeds every pretrained BFM on both datasets and both parcellations we test (r from 0.303 to 0.368 on AOMIC AAL-424), matching or exceeding the closest published FC results on comparable targets with no pretraining and no GPU. We recover the raw-FC ceiling on BrainLM's forward pass by finetuning with a loss targeted at this same subspace, which the same loss outside the subspace does not reach. This points to the pretraining objective, not the model size, as the bottleneck.

1 INTRODUCTION

Spontaneous brain activity reflects individual differences in performance on standardised cognitive assessments (Finn et al., 2015; Chen et al., 2022; Ooi et al., 2022). The quantity studied is not performance on one task: it is a *cognitive-ability composite* built from a battery of such assessments, reproducible within a person and variable across people. Functional connectivity (FC), the pairwise correlations between regional timeseries, predicts this composite under kernel ridge regression (Ooi et al., 2022). Brain foundation models (BFMs) aim to provide richer summaries: pretrain a large Transformer to reconstruct masked fMRI, then predict the composite from the embedding. In principle this should outperform FC. We show instead that this route falls short of raw FC, that a linear decomposition of the third-order co-skewness tensor exceeds both, and that a finetuning loss aligned with this third-order structure brings the BFM's reconstruction back to the raw-FC baseline, pointing to the objective as the bottleneck.

BFM evaluations rarely compare against classical FC baselines, and Zhou et al. (2025) note that BFM training strategies borrowed from NLP and computer vision ignore the structure of functional connectivity. This raises a basic question: do brain foundation models capture the variance that predicts individual differences?

They do not. Under the replicated nested cross-validation protocol of Ooi et al. (2022), three state-of-the-art timeseries-based BFMs reach at most Pearson $r = 0.18$ on AOMIC and $r = 0.05$ on HCP, while raw FC on the same parcellations reaches $r = 0.31$–0.39. The gap is large and holds across every dataset, parcellation, and readout we test. In the BrainLM family the underperformance *worsens* with scale: the 650M model predicts the composite worse than the 111M.

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.

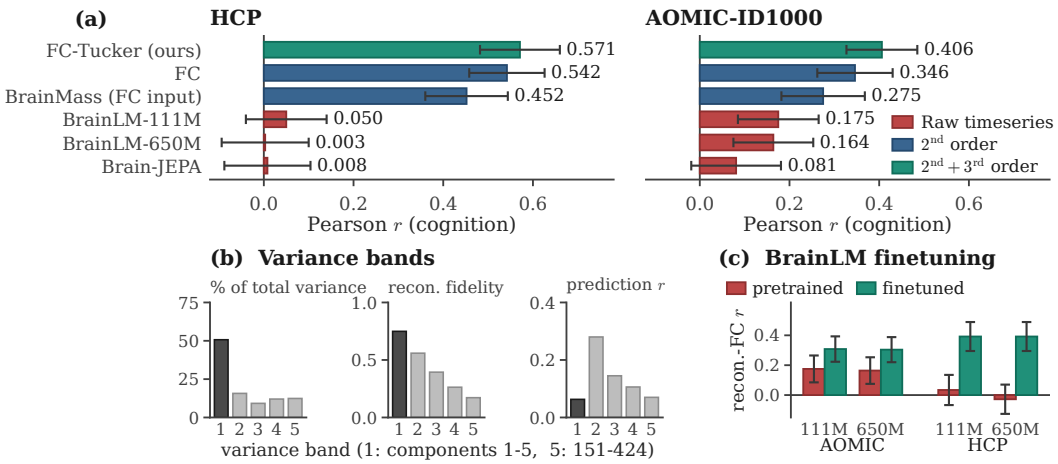


Figure 1: Bars: mean ± 1 std over 200 folds. **(a)** Pearson r between the predicted and the measured cognitive-ability composite. Pretrained models are shown at their highest-scoring readout, each on its own pretraining atlas; FC and FC-Tucker are on Schaefer-400 (Tables 1, 2). **(b)** AOMIC spatial principal components in five bands by variance rank: percentage of total signal variance, BrainLM-111M reconstruction fidelity, and r of the composite predicted from FC inside the band (Table 7). **(c)** r from the FC of the BrainLM reconstruction, pretrained and finetuned with the Tucker-subspace loss (Table 3).

We attribute this failure to a *variance allocation problem*. Reconstructing the timeseries is not the same as recovering the higher-order structure of the signal, FC included: a low reconstruction error does not imply a well-recovered FC. The fMRI signal is dominated by high-variance spatial structure that survives preprocessing: on AOMIC the top five of 424 spatial components hold half the signal variance and are what BrainLM reconstructs most faithfully, yet they predict the composite at near chance ($r = 0.06$; Section 4.2). Under this view, gradient descent prioritises that dominant variance and provides little signal for the rest. The mismatch shows at every statistical order we measure: κ_2 is partially preserved and κ_3 is largely destroyed (Section 4.3). Brain-JEPA predicts in latent space rather than signal space, but its targets come from the same variance-dominated input, so its latent inherits that same dominant variance and retains no more κ_3 than a reconstruction does (Appendix E).

Capacity is not the problem. Transformers can in principle represent arbitrary higher-order interactions among input tokens given sufficient depth and width (Yun et al., 2019); reconstruction objectives simply do not push them to. The signal that predicts the composite sits precisely where the loss is blind: synergistic information in association cortices (Luppi et al., 2022), O-information applied to resting-state fMRI (Gatica et al., 2021; 2022), partial entropy decompositions robust to pairwise controls (Varley et al., 2023), and triadic co-fluctuations that outperform pairwise FC for task decoding and brain–behaviour analyses (Santoro et al., 2023; 2024). An MSE loss does not preserve the structure that this prediction relies on.

The mismatch is also an opportunity: if what predicts the composite is tied to κ_3, an explicit κ_3 target should recover what an MSE loss discards. In our contributions we show:

- **That BFMs underperform classical FC for predicting cognitive labels.** Three timeseries-based BFMs sit below raw FC on every dataset, parcellation and readout, and stay there under a higher-capacity trained head and under end-to-end finetuning.
- **That the third cumulant improves the prediction of cognitive ability from FC.** We decompose the training subjects' co-skewness tensor, their empirical third cumulant, by Tucker/HOSVD, project each timeseries onto its R leading spatial directions, and compute FC there: κ_2 features in a κ_3-selected basis. With no pretraining and no GPU, this exceeds raw FC in all four dataset-by-parcellation cells, significantly in the two AOMIC cells, and exceeds every pretrained BFM.

- **That a cumulant-informed finetuning loss lifts the BFM to the raw-FC baseline.** A distance between FCs computed inside the κ_3-optimal Tucker subspace brings the BFM's to the raw-FC baseline.

These findings come from 200 fold-level evaluations per method across three BFM families, two datasets (AOMIC, N=876; HCP, N=955), two parcellations each (AAL-424 and Schaefer-400), and five finetuning strategies.

2 Related work

Brain foundation models. Recent rs-fMRI BFMs share a Transformer backbone but differ in their pretraining objective, which is the point of the failure we diagnose. BrainLM (Caro et al., 2023) is a ViT-MAE trained with masked fMRI reconstruction. Brain-JEPA (Dong et al., 2024) adapts the Joint-Embedding Predictive Architecture to fMRI, predicting masked-patch representations in embedding space. BrainMass (Yang et al., 2024) operates on FC matrices rather than raw timeseries and uses self-supervised mask modeling with auxiliary meta-label classification across 30 datasets. Concurrent BrainHarmonix (Dong et al., 2025) adds structural T1 morphology to a functional encoder; our function-only FC-Tucker compares favourably without pretraining or extra modality, though on a different cohort and target.[1] Standardized comparisons against classical FC baselines remain rare (Zhou et al., 2025); we provide one under the Ooi et al. (2022) protocol.

Probing studies, scaling, and failure modes. Two findings motivate our focus on the loss rather than on scale. The pretraining objective dictates what a self-supervised model captures: Man et al. (2024) show this across seven vision encoders for 3D scene understanding. Scaling does not always help: Schaeffer et al. (2023) attribute apparent LLM emergence to metric artefacts, and McKenzie et al. (2023) document inverse scaling on 11 tasks, the same pattern we report for BrainLM. Relatedly, self-supervised objectives can ignore task-relevant structure when irrelevant features dominate signal magnitude (Geirhos et al., 2020 on shortcut learning; Robinson et al., 2021 on InfoNCE feature suppression). Balestriero & LeCun (2024) show in vision that reconstruction-trained encoders allocate capacity to high-variance input components and produce features uninformative for downstream perception; Van Assel et al. (2025) prove that joint-embedding objectives inherit this failure. Both predict the BFM failure we observe in fMRI: the "shortcut" is the high-variance structure that dominates MSE reconstruction gradients.

Higher-order statistics in neuroscience. The closest prior work is Santoro et al. (2023; 2024): they introduce a higher-order co-fluctuation framework built on triadic products $z_i(t)z_j(t)z_k(t)$ (raw entries of the co-skewness tensor we decompose) and show on HCP fMRI that higher-order constructs outperform pairwise FC for task decoding, individual identification, and brain–behaviour associations. Our contribution is to work in the denoised Tucker subspace of this tensor rather than on its noisy raw entries (raw entries as features: Appendix B). O-information (Gatica et al., 2021) and partial information decomposition (Varley et al., 2023) measure the same triadic dependence with information-theoretic quantities; we do not evaluate them as features.

3 Methods

3.1 Functional connectivity and the joint cumulant expansion

We frame our analysis through the cumulant expansion because it makes precise what FC captures and what it leaves out. Let $\mathbf{Z} \in \mathbb{R}^{P\times T}$ denote z-scored ROI timeseries across P regions and T timepoints. Functional connectivity is the matrix of pairwise Pearson correlations $\mathbf{C} = \frac{1}{T-1}\mathbf{Z}\mathbf{Z}^\top$, with $\mathbf{C}_{ij} = \mathrm{Cov}(\mathbf{z}_i, \mathbf{z}_j)/(\sigma_{\mathbf{z}_i}\sigma_{\mathbf{z}_j})$: a $P \times P$ summary of the $P \times T$ timeseries that retains how much regions co-activate, not when.

The statistical structure of the joint distribution over ROI activations is fully characterized by its cumulants (Stuart & Ord, 1994). Since $\mathbf{Z}$ is z-scored, the first three joint cumulants are $\kappa_1 = 0$,

[1] HCP-A Flanker (executive function): their function-only ρ=0.30, multimodal ρ=0.42; ours r=0.43–0.57 on the HCP composite.

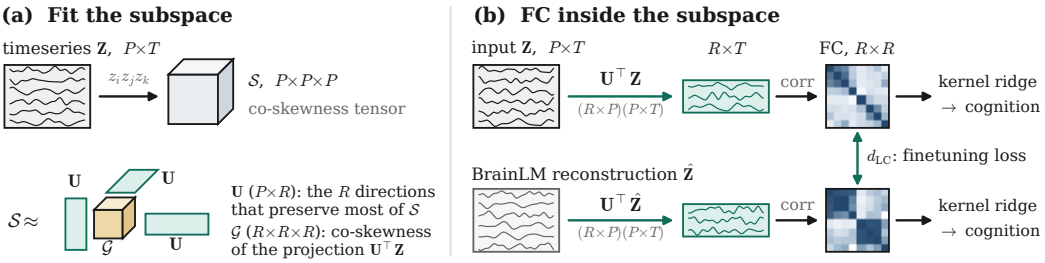


Figure 2: **(a)** HOSVD approximates the co-skewness tensor $\mathcal{S}$ of the training timeseries as $\mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{U} \times_3 \mathbf{U}$. The columns of $\mathbf{U}$ are the R spatial directions that preserve most of $\mathcal{S}$. **(b)** FC of the projected series, from the input (top, FC-Tucker) and from the BrainLM reconstruction (bottom). Both go to kernel ridge regression; their Log-Cholesky distance d_{LC} is the finetuning loss. Schematic.

$\kappa_2(z_i, z_j) = \mathbb{E}[z_i z_j]$ (pairwise covariance: whether two regions move together, one entry of FC), and $\kappa_3(z_i, z_j, z_k) = \mathbb{E}[z_i z_j z_k]$ (three-way co-skewness: whether three regions deviate together beyond what their pairwise correlations already explain). Each κ_n isolates statistical structure at order n. For Gaussian distributions $\kappa_n = 0$ for $n \geq 3$, so any non-zero κ_3 is a signature of non-Gaussian structure invisible to FC. Higher-order definitions are in Appendix S. FC is a κ_2 truncation of this hierarchy; including κ_3 is the natural next term, and the last one that needs no lower-order subtraction: from κ_4 on, each cumulant subtracts pair-products of κ_2 and so inherits any FC-estimation error (Appendix S).

3.2 The co-skewness tensor and Tucker decomposition

The raw κ_3 tensor is too noisy at typical session lengths to use directly; Tucker decomposition extracts its low-rank geometry, giving a stable subspace where covariance estimates become informative. The co-skewness tensor $\mathcal{S} \in \mathbb{R}^{P\times P\times P}$ is the empirical estimator of κ_3 over the T timepoints,

$$S_{ijk} \;=\; \frac{1}{T}\sum_{t=1}^{T} z_i(t)\, z_j(t)\, z_k(t) \;\approx\; \kappa_3(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k), \tag{1}$$

and summarises all three-way interactions. Its $O(P^3)$ entries are individually noisy when T is limited, but the low-rank geometry is more stable. Tucker decomposition (Tucker, 1966; Kolda & Bader, 2009) approximates $\mathcal{S}$ as $\mathcal{G} \times_1 \mathbf{U} \times_2 \mathbf{U} \times_3 \mathbf{U}$ with a core tensor $\mathcal{G} \in \mathbb{R}^{R\times R\times R}$ and factor matrix $\mathbf{U} \in \mathbb{R}^{P\times R}$, the same in all three modes because $\mathcal{S}$ is symmetric, obtained via higher-order SVD (HOSVD; computation in Appendix U). The columns of $\mathbf{U}$ span the subspace that approximately maximises third-order structure, analogously to how PCA maximises variance (κ_2). Projecting the timeseries as $\tilde{\mathbf{Z}} = \mathbf{U}^\top \mathbf{Z} \in \mathbb{R}^{R\times T}$ and then computing FC yields κ_2 features in a κ_3-informed basis (Figure 2). This is the representation underlying our classical result and the target of our finetuning loss.

Two properties of the tensor make its basis more useful than its entries. Its $\binom{P+2}{3} = 12{,}794{,}200$ unique entries at $P{=}424$ come from the $T \approx 200$ timepoints each model sees, so used directly as features they underperform FC by 0.07–0.08, while FC computed in the Tucker subspace exceeds FC by $+0.065$ on the same folds (Appendix B). The vectorised tensor also has magnitude $O(0.02)$, so a pure MSE on it does not train either BrainLM backbone meaningfully, whereas a Riemannian distance between covariance matrices is well conditioned (Appendix M.3).

3.3 Datasets

AOMIC-ID1000 (Snoek et al., 2021): **movie-watching** (naturalistic) fMRI, $N{=}876$ subjects, fMRIPrep-preprocessed (Esteban et al., 2019), parcellated to AAL-424 (Akiki & Abdallah, 2019; Nemati et al., 2020) ($P{=}424$) and Schaefer-400 (Schaefer et al., 2018) ($P{=}400$); a parcellation is an atlas dividing the brain into P regions whose voxel timeseries are averaged. **Human Connectome Project** (Van Essen et al., 2013), Young Adult S1200 release: $N{=}955$ subjects, four **resting-state** runs concatenated ($T{=}4{,}800$), using the minimally preprocessed ICA-FIX-denoised release, parcellated to AAL-424 and Schaefer-400. For BFM inputs, we additionally prepare the atlas each BFM was pretrained on (Schaefer-400+Tian-50 for Brain-JEPA, i.e. Schaefer-400 with the Tian

Subcortex Scale III atlas (Tian et al., 2020); Schaefer-100 FC for BrainMass). The cognitive target is dataset-specific: on HCP, the first principal component of the same 58 behavioural measures Ooi et al. (2022) use (Appendix K); on AOMIC, the first principal component of the four Intelligence Structure Test subscales (fluid, memory, crystallised, total). We call this the cognitive-ability composite, predict it throughout, and treat it as a proxy for cognitive ability rather than a measure of it. Train-only PCA fit and leakage controls in Appendix U.

3.4 Evaluation protocol

To compare BFMs and classical baselines on equal footing, we adopt the Ooi et al. (2022) protocol. Kernel ridge regression (KRR) takes as input the $P(P-1)/2$ off-diagonal FC entries with a Pearson correlation kernel (partial correlation in Appendix H). Replicated nested CV (20 repetitions $\times$ 10 folds, family-aware for HCP; Ooi uses 60 repetitions) yields 200 fold-level evaluations per method. The inner CV selects $\alpha \in \{10^{-3}, \ldots, 10^{3}\}$; all decompositions (PCA, Tucker) and standardisations are fit on training folds only. We report mean Pearson $r\pm$ std and Cohen's d for paired comparisons. Because overlapping CV training sets make folds dependent, significance uses the corrected resampled t-test of Nadeau & Bengio (2003) (written p_{NB}) with Benjamini-Hochberg FDR control across a family of tests (adjusted values written q; Appendix G).

3.5 Feature representations

Classical FC baseline. **FC-full** uses the $P(P-1)/2$ Pearson correlations of the raw timeseries. **FC-PCA(K)** and **FC-Tucker(R)** compute FC after projecting the timeseries into, respectively, the top-K PCA subspace (max κ_2) and the rank-R Tucker subspace of the training co-skewness tensor (max κ_3; Section 3.2). We sweep the Tucker rank R and PCA dimensionality K densely and additionally report the fully nested-CV optima R^*, K^* (Appendix P) to rule out post-hoc selection.

BFM readouts. For each BFM we extract four KRR readouts: the embedding (CLS token or mean-pooled patches), flattened patch tokens, the embedding-similarity matrix across timepoints, and reconstruction FC. Each BFM is evaluated on its pretraining inputs: BrainLM (AAL-424 timeseries, MAE reconstruction; best readout: reconstruction FC), Brain-JEPA (Schaefer-400+Tian-50, JEPA; best readout: embedding similarity), BrainMass (Schaefer-100 FC, mask modeling; single CLS readout). We also report **fc_input**, KRR on raw FC at the same parcellation and pipeline as the BFM input, as a no-model control (Appendix A).

Trained heads and end-to-end finetuning. On identical folds we also fit a linear ridge head and a 256-unit MLP head on the frozen embedding, and finetune each backbone end-to-end *inside* each outer fold, on that fold's training subjects only, so no test label ever reaches the finetuner (Appendix C).

3.6 Cumulant preservation and latent decoding

To localise where BFM reconstructions depart from their inputs along the cumulant hierarchy, we compute per-subject relative errors at κ_2 and κ_3 (Log-Cholesky and normalised MSE; the κ_3 metrics are evaluated inside the rank-R=80 Tucker subspace fitted on the training set; Appendix R). This needs a decoder, so for encoders without one we ask instead how much of the input's κ_2 and κ_3 is linearly decodable from the frozen embedding, both built in one shared training-fitted Tucker basis (Appendix E).

3.7 FC-preservation finetuning

To test whether a κ_3-aware loss can recover what pretraining destroys, we finetune BrainLM end-to-end with FC-preservation objectives. All runs start from the released pretrained weights, unfreeze backbone and decoder, and use the pretraining input pipeline (AAL-424 timeseries, masking ratio 0.2, patch length 20). Training uses the training split (Appendix U) with a 90/10 within-train val split for early stopping; the test split is never seen.

Log-Cholesky FC loss. FC matrices live on the SPD manifold. The Log-Cholesky metric (Lin, 2019) decomposes $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$ and computes a Euclidean distance after log-transforming the diagonal,

$$d_{\mathrm{LC}}(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2) \;=\; \sqrt{\|\mathrm{strict}(\mathbf{L}_1 - \mathbf{L}_2)\|_F^2 \;+\; \|\log \mathrm{diag}(\mathbf{L}_1) - \log \mathrm{diag}(\mathbf{L}_2)\|_2^2}, \tag{2}$$

with $\mathrm{strict}(\cdot)$ the strictly lower-triangular part. The baseline finetuning loss is $\mathcal{L}_{\mathrm{LC}} = d_{\mathrm{LC}}(\mathrm{FC}(\mathbf{Z}), \mathrm{FC}(\hat{\mathbf{Z}}))$.

Dual-moment loss. A natural extension aligns both κ_2 and κ_3:

$$\mathcal{L}(\lambda) \;=\; \lambda\, d_{\mathrm{LC}}\big(\mathrm{FC}(\mathbf{Z}), \mathrm{FC}(\hat{\mathbf{Z}})\big) \;+\; (1-\lambda)\, d_{\kappa_3}(\mathbf{Z}, \hat{\mathbf{Z}}), \tag{3}$$

with two surrogates for d_{κ_3}. The `coskew_mse` variant is the MSE between the vectorised coskewness tensors projected into the rank-R Tucker subspace of the training set (a direct κ_3 loss). The `fc_tucker` variant is the Log-Cholesky distance between FCs computed on timeseries projected into the same Tucker subspace: *κ_2 inside the κ_3-optimal basis*. The Tucker factor matrix $\mathbf{U} \in \mathbb{R}^{P \times R}$ is fit once on the training set and frozen during finetuning. We sweep λ and report the best configuration (Appendix M.3). At the reported setting, `fc_tucker` with $\lambda = 10^{-3}$, almost all the weight sits on the subspace distance, so the objective is a second-order distance taken inside the κ_3 basis rather than a constraint on κ_3 itself. These FC-preservation losses are label-free; cognition-aware variants (KRR alignment, cognition-similarity, direct regression) are reported in Appendix M.5 as architectural-plasticity probes.

4 Results

4.1 Brain foundation models underperform classical FC

Every timeseries BFM performs below the parcellation-matched input FC at every readout (Table 1). On HCP, BrainLM-650M reaches $r = 0.003$ from its embedding and -0.028 from reconstruction FC, both indistinguishable from zero, while input FC on the same AAL-424 input reaches 0.393; Brain-JEPA behaves similarly ($0.004/0.008$ vs. 0.352). Within the BrainLM family, the 650M model underperforms the 111M on both datasets and at both readouts, and is uniformly worse at the two cumulant orders we measure (Table 29). Two sizes cannot establish a trend, so we report the ordering without a mechanism.

The readout is not what limits this. A linear ridge or 256-unit MLP head on the frozen embedding leaves it where it was: across the twelve timeseries-BFM cells the higher-capacity head is below the kernel readout in nine and never above it by more than $+0.017$, and the gap to the best of the same three heads on raw FC stays at 0.18–0.37 ($p_{\mathrm{NB}} < 0.001$ everywhere). We scope the claim to practical extractability: the target may be encoded, but not in a form any readout we tried can reach. Finetuning end-to-end within the fold does no better: on AOMIC the best result is 0.138, the conventional heads reach 0.034–0.059 for BrainLM-111M and 0.008–0.032 for Brain-JEPA against raw FC at 0.267–0.313, and BrainLM-650M goes negative on every head (Appendix C). The gap is larger on resting-state HCP (0.34–0.37) than on movie-watching AOMIC (0.18–0.24).

BrainMass is a partial exception: taking the FC matrix directly as input (already a κ_2 summary), it is not subject to the variance-allocation bottleneck. Under the KRR readout it sits below its input-FC baseline on both datasets (0.452 vs. 0.504 on HCP, 0.275 vs. 0.311 on AOMIC); with a ridge head on AOMIC it matches raw FC (0.311, $p_{\mathrm{NB}} = 0.97$) while trailing by 0.052 on HCP. We treat it as a control rather than a direct comparison.

4.2 The dominant variance survives preprocessing and does not predict the composite

The variance allocation hypothesis predicts that the already-preprocessed signal contains structure that dominates its variance, that the model reconstructs well, and that carries none of the target. We rank the spatial principal components by variance, group them into five log-spaced bands from B1 (components 1–5) to B5, and per band measure its variance share, its reconstruction fidelity, the composite predicted from FC inside it, and how much of it a standard confound model explains (Appendix D).

Table 1: Prediction of the cognitive-ability composite per BFM readout vs. the parcellation-matched input FC baseline (Ooi et al., 2022 protocol). Pearson r (mean $\pm$ std, $n = 200$). AOMIC is movie-watching, HCP resting-state. Each *Input FC* row uses that BFM's own input pipeline and context length, so it is not comparable with the full-scan FC of Table 2. Higher-capacity heads and end-to-end finetuning are in Appendix C. BrainMass has a single readout.

	AOMIC r		HCP r	
Model / Readout	Embedding	Best readout	Embedding	Best readout
BrainLM-111M	0.126 ± 0.101	0.175 ± 0.090	0.050 ± 0.090	0.034 ± 0.101
BrainLM-650M	0.065 ± 0.110	0.164 ± 0.089	0.003 ± 0.097	-0.028 ± 0.098
Input FC (AAL-424)	$\mathbf{0.306 \pm 0.084}$		$\mathbf{0.393 \pm 0.097}$	
Brain-JEPA	0.069 ± 0.091	0.081 ± 0.100	0.004 ± 0.096	0.008 ± 0.096
Input FC (Schaefer-400+Tian-50)	$\mathbf{0.313 \pm 0.083}$		$\mathbf{0.352 \pm 0.091}$	
BrainMass	0.275 ± 0.093		0.452 ± 0.092	
Input FC (Schaefer-100)	$\mathbf{0.311 \pm 0.095}$		$\mathbf{0.504 \pm 0.088}$	

On AOMIC the three quantities dissociate (Figure 1b): band B1 carries 50.7% of all signal variance, is the band BrainLM-111M reconstructs best (fidelity 0.749), is not attributable to preprocessing (in-sample $R^2 = 0.277$ against a null of 0.287; cross-validated $R^2 = -0.185$), and predicts the composite at $r = 0.063$, while prediction peaks at B2 ($r = 0.280$) and variance and prediction are anti-correlated across bands on this dataset (-0.363; near zero on HCP, where prediction also loads the high-index tail). The dominant variance is therefore second-order structure that survives preprocessing and is blind to the target, and a reconstruction loss commits capacity to it because that is where the signal variance is. The reconstruction result and the confound test are AOMIC-only, since on HCP BrainLM's reconstruction is near noise at both sizes and that release carries no movement regressors; B1 predicts the composite at 0.063 and 0.068 on the two datasets regardless.

4.3 Cumulant preservation localises a failure at κ_3

The prediction failure sits at a specific statistical order, not diffuse across the embedding. BrainLM reconstructions partially preserve κ_2 (29–47% relative Log-Cholesky error) but fail to preserve κ_3 (normalised MSE 2.6–$6.2\times$ input signal energy; Log-Cholesky relative error above 1.0 on HCP and at 650M on AOMIC, so predicting zero structure would be better), uniformly worse at 650M than at 111M (Table 29), which is exactly the order where composite-relevant information concentrates (Section 2).

Brain-JEPA has no decoder, so we ask the same question of the frozen latent (Appendix E). Decoding into a shared Tucker basis with dimensionality matched across the two targets, Brain-JEPA's latent carries κ_2 above its shuffled null ($+0.076$ against -0.018 on AOMIC, $p_{\mathrm{NB}} < 10^{-4}$) and carries κ_3 at the null ($+0.015$ against $+0.014$, $p_{\mathrm{NB}} = 0.93$; HCP $+0.008$ against -0.002); the same κ_2-over-κ_3 asymmetry holds for BrainLM at both sizes ($p_{\mathrm{NB}} \leq 0.006$). From the same Brain-JEPA inputs, the raw input FC decodes those κ_3 targets at $+0.223$ and $+0.162$, so third-order structure is recoverable from a classical representation of the same data and not from the learned latent, and a latent-prediction objective retains no more of it than a reconstruction objective. Fingerprinting agrees. We split each HCP session into segments and count how often a segment's nearest neighbour in cosine distance belongs to the same subject, among $\sim$950 subjects. Raw FC identifies subjects at 87% ($d = 0.91$), and still at 84% after PCA to the 768 dimensions of a BrainLM embedding, while the three timeseries BFM embeddings sit at 0.2–0.4% (Appendix T). The embeddings lose what separates people in general, of which the composite is one part.

4.4 Second- vs. third-order spatial subspace selection

If the signal lives in κ_3, the basis used to compress the timeseries should matter more than the amount of compression. PCA selects the K directions of maximum κ_2 (covariance); Tucker (Section 3.2) selects the R directions of maximum κ_3 (co-skewness). We report both in Table 2 at the sweep-optimal dimensionality, alongside FC-full. **PCA is inconsistent.** Gains over FC-full are modest where present (AOMIC Schaefer $+8\%$, HCP AAL $+6\%$, HCP Schaefer $+3\%$) and absent on AOMIC AAL

at $K^* = 154$; at that cell's Tucker optimum ($K = 117$) PCA underperforms FC-full ($d = -0.24$). Variance magnitude is not a reliable proxy for behavioural relevance. **Tucker is consistent**: it beats both FC-full and FC-PCA on all four dataset×parcellation combinations. On HCP Schaefer-400, FC-Tucker(R^*=333) reaches $r = 0.571 \pm 0.089$. The closest published rs-FC KRR numbers on the same composite are $r = 0.520 \pm 0.011$ (principal gradients) and $r = 0.513 \pm 0.013$ (Schaefer-2018) by Kong et al. (2023); Dubois et al. (2018) report $r = 0.457$ for a bifactor general-intelligence score (Glasser MMP, elastic net, N=884). FC computed inside our κ_3 Tucker subspace is above all three, on comparable though not identical targets, with no pretraining and no GPU.

Under the corrected resampled test with FDR, Tucker-over-FC-full is significant in the two AOMIC cells ($q = 0.034$ each) and in neither HCP cell (AAL $q = 0.092$; Schaefer $q = 0.153$, $d = 0.49$). These q are upper bounds, since the correction discounts the 20 repeats entirely: the uncorrected test on the same folds gives $p = 6.6 \times 10^{-11}$ on HCP Schaefer. Clearing $q = 0.05$ needs $N \approx 2{,}000$ subjects rather than our 955 (Appendix G).

Paired Tucker-vs-PCA tests at matched dimensionality (R=K) isolate the subspace-criterion effect and confirm the third-order basis as the driver, with the largest effect on AAL (d=1.57 on AOMIC, 1.21 on HCP) and a smaller one on Schaefer (Appendices I, O). FC-Tucker exceeds FC-full across a broad plateau ($R \gtrsim 80$ up to full rank) and at 25× compression (R=80) still beats it on HCP Schaefer-400 (+0.020, $d = 0.40$); a matched temporal reduction is indistinguishable from FC-full in all four cells, so the advantage is spatial (Appendices J, L); and a per-variable scan over the 57 HCP behavioural variables splits 17 for FC-Tucker on both parcellations, 20 for FC-full on both, and 20 mixed (Appendix K).

Table 2: Spatial subspace FC: PCA (max-κ_2 criterion) vs. Tucker (max-κ_3 criterion), reported at the sweep-optimal dimensionality alongside FC-full. FC-full is refit inside this sweep, so AOMIC AAL reads 0.303 here against 0.306 in Table 1. Pearson r (mean ± std, $n = 200$); Cohen's d for the paired Tucker-vs-FC-full comparison. Tucker beats FC-full in direction in all four cells, clearing FDR under the corrected test in the two AOMIC cells (Appendix G); PCA is inconsistent.

Dataset	FC-full	FC-PCA(K^*)	K^*	FC-Tucker(R^*)	R^*	d
AOMIC AAL	0.303 ± 0.086	0.307 ± 0.088	154	$\mathbf{0.368 \pm 0.086}$	117	**0.90**
AOMIC Schaefer	0.346 ± 0.084	0.375 ± 0.079	57	$\mathbf{0.406 \pm 0.079}$	122	0.82
HCP AAL	0.393 ± 0.097	0.418 ± 0.101	417	$\mathbf{0.431 \pm 0.092}$	167	0.62
HCP Schaefer	0.542 ± 0.084	0.557 ± 0.081	87	$\mathbf{0.571 \pm 0.089}$	333	0.49

4.5 Cumulant-informed finetuning recovers the raw-FC ceiling on BFM reconstruction

The cumulant preservation analysis (Table 29) points to a concrete fix: align the finetuning loss with the κ_3 structure that pretraining destroys. Table 3 reports BrainLM finetuned with (i) Log-Cholesky on ambient FC and (ii) the dual-moment loss (Log-Cholesky on FC inside the rank-$R = 80$ Tucker subspace, $\lambda = 10^{-3}$).

Dual-moment reaches the raw-FC level at both scales, on AOMIC from the AOMIC-trained checkpoint and on HCP from the HCP-trained one. BrainLM-111M/650M reach $0.308/0.304$ on AOMIC vs. raw 0.306, robust to 5× longer training, and finetuning on the HCP training split closes the HCP gap at both scales ($r = 0.392$ vs. raw 0.393) (Appendices M.4, M.6). Each checkpoint also transfers across the difference in task state, at or just below the raw baseline (AOMIC-trained 0.379–0.380 on HCP, HCP-trained 0.305–0.306 on AOMIC). What makes the loss work is the basis, not the moment: the same distance in the ambient basis reaches only 0.247–0.350, and a pure MSE on the κ_3 tensor does not train meaningfully (Appendix M.3). CLS embeddings stay near zero under every FC-preservation loss (Appendices M.1, M.2): the predictive signal is present in the reconstructed output but is not linearly decodable from the CLS token.

5 Discussion and conclusions

What predicts the cognitive-ability composite lives in a small, higher-order component of fMRI, and reconstruction-trained BFMs do not capture it: their loss is dominated by second-order structure that

Table 3: FC-preservation finetuning (Pearson r on the reconstruction FC readout, mean $\pm$ std, $n = 200$). *FT data* is the dataset a checkpoint was finetuned on; the two columns are the dataset it is evaluated on. Dual-moment is the Log-Cholesky distance on FC inside the rank-R=80 Tucker subspace of the training co-skewness tensor, at $\lambda = 10^{-3}$; the HCP rows use the checkpoint finetuned on HCP (Appendix M.6). CLS readouts in Appendices M.1, M.2.

Condition	FT data	AOMIC r	HCP r
BrainLM-111M			
Pretrained	–	0.175 ± 0.090	0.034 ± 0.101
Log-Chol FT	AOMIC	0.247 ± 0.087	0.325 ± 0.091
Dual-moment FT	AOMIC	$\mathbf{0.308 \pm 0.085}$	$\mathbf{0.379 \pm 0.096}$
Dual-moment FT	HCP	$\mathbf{0.306 \pm 0.084}$	$\mathbf{0.392 \pm 0.097}$
BrainLM-650M			
Pretrained	–	0.164 ± 0.089	-0.028 ± 0.098
Log-Chol FT	AOMIC	0.277 ± 0.089	0.350 ± 0.100
Dual-moment FT	AOMIC	$\mathbf{0.304 \pm 0.084}$	$\mathbf{0.380 \pm 0.100}$
Dual-moment FT	HCP	$\mathbf{0.305 \pm 0.084}$	$\mathbf{0.392 \pm 0.097}$
raw FC baseline		0.306 ± 0.084	0.393 ± 0.097

survives preprocessing and does not track the target, so gradient descent fits that surface and sheds κ_3, and the larger BrainLM is better at that objective and worse at the composite (650M below 111M). Finetuning with a cumulant-informed loss (Log-Cholesky distance between FCs inside the κ_3-optimal Tucker subspace) recovers the raw-FC ceiling from the BFM's forward pass, which the same distance outside that subspace does not: the objective, not model size, is the bottleneck. Three takeaways follow. First, evaluate foundation models against classical domain-knowledge baselines before scaling further; per-order cumulant preservation localises any failure at a specific order. Second, when the task-relevant subspace is known, an explicit decomposition is cleaner (cf. CEBRA (Schneider et al., 2023)). Third, when the subspace is unknown, a cumulant-informed auxiliary loss preserves the task-relevant orders, a concrete alternative to scaling wherever dominant variance is not task-relevant (physiological recordings, climate, single-cell dynamics, finance).

The neuroscience take-away is concrete: the structure that predicts individual differences is present in the third cumulant, but the raw κ_3 tensor is too noisy at typical session lengths to use directly. Linear decompositions (Tucker, HOSVD) project the timeseries into a low-rank spatial subspace where covariance estimates become stable; the resulting κ_2-in-κ_3-basis is simple, GPU-free, and beats every pretrained BFM we tested: a default worth adding to analyses predicting individual differences from FC.

6 Limitations and future work

Support and targets. The two HCP cells do not clear the corrected bar (Section 4.4); on seven OpenNeuro cohorts FC-Tucker is ahead in six, none significant, four positive under both rank rules (Appendix F). UK-Biobank-scale validation is the fix. We evaluate two composites; other target classes have different variance structure and may respond differently, and clinical cohorts (ADHD, schizophrenia, Alzheimer's) are immediate future work. **Output versus latent.** Every FC-preservation objective closes the output gap and leaves the CLS embedding near zero (Appendices M.1, M.2); aligning a pooled latent to the input's Tucker-subspace FC is the open problem, as is whether $\kappa_{\geq 4}$ helps and how Tucker directions map to physiology.

Ethics statement

This work is a secondary analysis of two openly distributed human neuroimaging cohorts. AOMIC-ID1000 (Snoek et al., 2021) is released under CC0 on OpenNeuro (`ds003097`); HCP S1200 (Van Essen et al., 2013) is governed by the WU-Minn HCP Open Access Data Use Terms, under which we registered and accepted the no-re-identification clause. Both cohorts were collected under the ethical approvals of their original studies and are distributed de-identified; we performed no new data collection and made no attempt to re-identify participants. Appendix T quantifies how

identifiable individuals are from their connectivity; we release no subject-level representations. AOMIC, HCP, and the BFM pretraining cohorts (e.g. UK Biobank) over-represent Western, healthy, well-educated participants; the methods here inherit those biases, require re-validation before any clinical deployment.

REPRODUCIBILITY STATEMENT

The evaluation protocol, fold seeds, hyperparameter grids, and leakage controls are specified in Section 3.4 and Appendix U; Appendix U also lists exact model-weight sources, dataset accessions, preprocessing steps, train/test splits, and the compute budget. Statistical procedures, including the corrected resampled t-test and the FDR correction, are given in Appendix G with per-test values for every comparison cited in the main text. Rank selection is by fully nested cross-validation (Appendix P), so no reported number depends on post-hoc hyperparameter choice. An anonymised snapshot of the code is available at `https://anonymous.4open.science/r/E4C0/`, released as a research artefact rather than a clinical or decision-making tool.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, formulate mathematical claims, propose and refine hypotheses, design and provide feedback on research methodology and experiments, implement methods, clean and reformat datasets, and interpret results. We have not used generative AI tools to generate synthetic data sets or to discover research topics or identify gaps. Proving mathematical claims and writing proofs, translation, qualitative and thematic data analysis, questions for surveys or interviews and transcription of recordings are not applicable to this work. Additionally, we used generative AI tools to create and modify scientific figures, suggest experimental parameters, create and edit software code, run and monitor experiments on our compute cluster, draft parts of this paper, summarise and analyse existing literature, brainstorm, search for information, edit this paper to improve readability, identify relevant literature, format references, suggest a structure for this paper and propose its title and keywords. We have reviewed all AI-assisted work: AI-assisted code was reviewed and tested by the authors, every number in the paper was checked by the authors against the stored experiment outputs, and AI-assisted text and cited literature were checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Teddy J Akiki and Chadi G Abdallah. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific Reports*, 9(1):19290, 2019. doi: 10.1038/s41598-019-55738-y.

Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024.

Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, 37(1):90–101, 2007.

Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5:1089–1105, 2004.

Remco R Bouckaert and Eibe Frank. Evaluating the replicability of significance tests for comparing learning algorithms. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 3–12. Springer, 2004.

Josue Ortega Caro, Antonio H de O Fonseca, Christopher Averill, Syed A Rizvi, Matteo Rosati, James L Cross, Prateek Mittal, Emanuele Zappala, Daniel Levine, Rahul M Dhodapkar, et al. Brainlm: A foundation model for brain activity recordings. *BioRxiv*, pp. 2023–09, 2023.

Jianzhong Chen, Angela Tam, Valeria Kebets, Csaba Orban, Leon Qi Rong Ooi, Christopher L Asplund, Scott Marek, Nico U F Dosenbach, Simon B Eickhoff, Danilo Bzdok, Avram J Holmes, and B T Thomas Yeo. Shared and unique brain network features predict cognitive, personality,

and mental health scores in the abcd study. *Nature Communications*, 13(1):2217, 2022. doi: 10.1038/s41467-022-29766-8.

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. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2017.03.020. URL https://www.sciencedirect.com/science/article/pii/S1053811917302288. Cleaning up the fMRI time series: Mitigating noise with advanced acquisition and correction strategies.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan T Nguyen, Joanna S Chong, Fang Ji, Nathanael R Tong, Christopher L Chen, and Juan H Zhou. Brain-jepa: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. *Advances in Neural Information Processing Systems*, 37: 86048–86073, 2024.

Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, B T Thomas Yeo, and Juan Helen Zhou. Brain harmony: A multimodal foundation model unifying morphology and function into 1d tokens. *arXiv preprint arXiv:2509.24693*, 2025. URL https://arxiv.org/abs/2509.24693.

Julien Dubois, Paola Galdi, Lynn K Paul, and Ralph Adolphs. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756), 2018.

Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, Hiroyuki Oya, Satrajit S Ghosh, Jessey Wright, Joke Durnez, Russell A Poldrack, and Krzysztof J Gorgolewski. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16(1):111–116, 2019. doi: 10.1038/s41592-018-0235-5.

Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica S 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.

Marilyn Gatica, Rodrigo Cofré, Pedro A M Mediano, Fernando E Rosas, Patricio Orio, Ibai Diez, Stephan P Swinnen, and Jesus M Cortes. High-order interdependencies in the aging brain. *Brain Connectivity*, 11(9):734–744, 2021. doi: 10.1089/brain.2020.0982. PMID: 33858199.

Marilyn Gatica, Fernando E Rosas, Pedro A M Mediano, Ibai Diez, Stephan P Swinnen, Patricio Orio, Rodrigo Cofré, and Jesus M Cortes. High-order functional redundancy in ageing explained via alterations in the connectome in a whole-brain model. *PLOS Computational Biology*, 18(9): 1–21, 2022. doi: 10.1371/journal.pcbi.1010431.

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.

Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. doi: 10.1137/07070111X. URL https://doi.org/10.1137/07070111X.

Ru Kong, Yan Rui Tan, Naren Wulan, Leon Qi Rong Ooi, Seyedeh-Rezvan Farahibozorg, Samuel Harrison, Janine D Bijsterbosch, Boris C Bernhardt, Simon Eickhoff, and BT Thomas Yeo. Comparison between gradients and parcellations for functional connectivity prediction of behavior. *NeuroImage*, 273:120044, 2023. doi: 10.1016/j.neuroimage.2023.120044.

Zhenhua Lin. Riemannian geometry of symmetric positive definite matrices via cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40(4):1353–1370, 2019. doi: 10.1137/18M1221084. URL https://doi.org/10.1137/18M1221084.

Thomas T. Liu. Noise contributions to the fmri signal: An overview. *NeuroImage*, 143:141–151, 2016. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2016.09.008. URL https://www.sciencedirect.com/science/article/pii/S1053811916304694.

Andrea I Luppi, Pedro A M Mediano, Fernando E Rosas, Negin Holland, Tim D Fryer, John T O'Brien, James B Rowe, David K Menon, Daniel Bor, and Emmanuel A Stamatakis. A synergistic core for human brain evolution and cognition. *Nature Neuroscience*, 25(6):771–782, 2022. doi: 10.1038/s41593-022-01070-0.

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *Advances in Neural Information Processing Systems*, 37:76819–76847, 2024.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*, 2023.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine Learning*, 52(3): 239–281, 2003.

Samaneh Nemati, Teddy J Akiki, Jeremy Roscoe, Yumeng Ju, Christopher L Averill, Samar Fouda, Arpan Dutta, Shane McKie, John H Krystal, JF William Deakin, et al. A unique brain connectome fingerprint predates and predicts response to antidepressants. *iScience*, 23(1):100800, 2020. doi: 10.1016/j.isci.2019.100800.

Leon Qi Rong Ooi, Jianzhong Chen, Shaoshi Zhang, Ru Kong, Angela Tam, Jingwei Li, Elvisha Dhamala, Juan Helen Zhou, Avram J Holmes, and BT Thomas Yeo. Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity mri. *NeuroImage*, 263:119636, 2022.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? *Advances in neural information processing systems*, 34:4974–4986, 2021.

Andrea Santoro, Federico Battiston, Giovanni Petri, and Enrico Amico. Higher-order organization of multivariate time series. *Nature Physics*, 19(2):221–229, 2023. doi: 10.1038/s41567-022-01852-0.

Andrea Santoro, Federico Battiston, Maxime Lucas, Giovanni Petri, and Enrico Amico. Higher-order connectomics of human brain function reveals local topological signatures of task decoding, individual identification, and behavior. *Nature Communications*, 15(1):10244, 2024. doi: 10.1038/s41467-024-54472-y.

Theodore D Satterthwaite, Mark A Elliott, Raphael T Gerraty, Kosha Ruparel, James Loughead, Monica E Calkins, Simon B Eickhoff, Hakon Hakonarson, Ruben C Gur, Raquel E Gur, and Daniel H Wolf. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, 64:240–256, 2013.

Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT 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.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *Advances in neural information processing systems*, 36:55565–55581, 2023.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023. doi: 10.1038/s41586-023-06031-6.

Lukas Snoek, Maite M van der Miesen, Tinka Beemsterboer, Andries van der Leij, Annemarie Eigenhuis, and H Steven Scholte. The amsterdam open mri collection, a set of multimodal mri datasets for individual difference analyses. *Scientific Data*, 8(1):85, 2021. doi: 10.1038/s41597-021-00870-6.

Alan Stuart and J Keith Ord. *Kendall's Advanced Theory of Statistics, Volume 1: Distribution Theory*. Edward Arnold, London, 6 edition, 1994.

Ye Tian, Daniel S Margulies, Michael Breakspear, and Andrew Zalesky. Topographic organization of the human subcortex unveiled with functional connectivity gradients. *Nature Neuroscience*, 23 (11):1421–1432, 2020. doi: 10.1038/s41593-020-00711-6.

Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3): 279–311, 1966. doi: 10.1007/BF02289464.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*, 2025.

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. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.05.041. URL https://www.sciencedirect.com/science/article/pii/S1053811913005351. Mapping the Connectome.

Thomas F. Varley, Maria Pope, and Olaf Sporns. Partial entropy decomposition reveals higher-order information structures in human brain activity. *Proceedings of the National Academy of Sciences*, 120(30):e2300888120, 2023. doi: 10.1073/pnas.2300888120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2300888120.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. Brainmass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE transactions on medical imaging*, 43(11):4004–4016, 2024.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *IEEE Signal Processing Magazine*, 42(5):22–35, 2025.

A Comparison pipeline and per-method overview

Figure 3 diagrams the full comparison: what each method takes as input, how it forms a representation, which readout is used, and the single shared evaluator applied to all of them. Figure 4 shows the resulting per-method numbers on both datasets in one view; the same values appear in Tables 1 and 2.

B Co-skewness features in isolation

Section 4.4 of the main text reports KRR on FC computed *after* projecting the ROI timeseries into the rank-R Tucker subspace of the co-skewness tensor. A natural ablation is to skip the FC step and regress directly on the co-skewness features. Table 4 (AOMIC AAL-424) reports KRR Pearson r on two direct κ_3 feature sets and compares to FC-full and FC-Tucker at the same rank.

The gap is large. CoS-PCA and CoS-Tucker, which treat the $O(P^3)$ tensor entries as direct features, both underperform FC-full by roughly 0.07 in Pearson r. The co-skewness tensor has $\binom{P+2}{3} = 12{,}794{,}200$ unique entries for $P = 424$, estimated from $T \approx 200$ timepoints per subject: each entry has high sampling variance and individually contains little reproducible signal. What is informative is the tensor's *geometry*, captured by the Tucker factor matrix $\mathbf{U} \in \mathbb{R}^{P\times R}$. FC computed inside this subspace (FC-Tucker) outperforms FC-full by $+0.065$ in Pearson r on the same CV folds, even though the direct κ_3 features underperform it by -0.070 to -0.080. Tucker does not merely preserve third-order statistics; it denoises them via the low-rank constraint, and the denoised subspace is what carries the cognitive signal. This motivates the rest of the paper: κ_3 is the useful cue, but it enters the pipeline through the *basis* of the tensor, not through its *entries*.

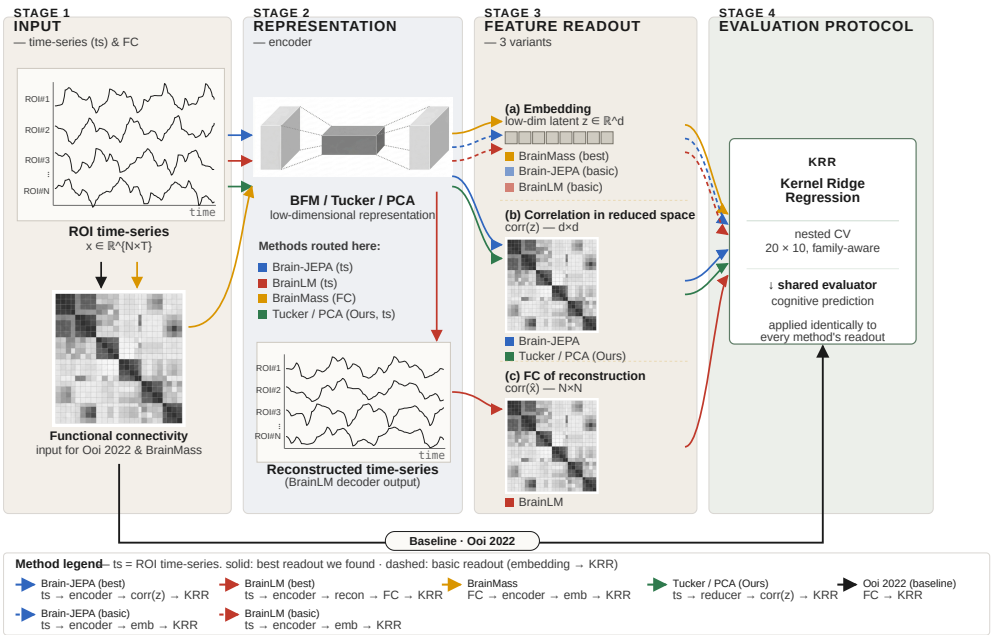


Figure 3: Feature-extraction pipelines and comparison protocol. All methods share the KRR and nested-CV evaluator (Section 3.4); they differ in (1) input (raw ROI timeseries or FC matrix), (2) representation (encoder or reducer, with or without reconstruction), and (3) readout. *(basic)* marks the CLS or embedding readout per BFM; *(best)* marks the highest-scoring readout we found (FC-recon for BrainLM, embedding similarity for Brain-JEPA). Tucker occupies the same pipeline slot as the BFM encoders.

Table 4: Direct co-skewness features vs. FC in the Tucker subspace (AOMIC AAL-424, 200 CV folds). Direct κ_3 entries underperform FC; the Tucker *basis* is informative even when the tensor entries themselves are too noisy to use directly.

Feature set	Pearson r
CoS-PCA (vectorised κ_3 entries, top-K PCA)	0.236 ± 0.089
CoS-Tucker (vectorised κ_3 entries in rank-R Tucker core)	0.222 ± 0.088
FC-full	0.303 ± 0.086
FC-Tucker ($R^* = 117$, Section 4.4)	$\mathbf{0.368 \pm 0.086}$

C HIGHER-CAPACITY PROBE HEADS AND LEAK-FREE END-TO-END FINETUNING

This appendix reports in full the experiment summarised in Section 4.1: how the BFM numbers respond to readouts of increasing capacity. Two arms answer it. Arm A keeps the backbone frozen and increases the capacity of the head. Arm B finetunes the whole backbone end-to-end. If neither lifts the embedding to raw FC, the ceiling belongs to the representation and not to the readout.

Arm A protocol. Every readout runs on identical 20×10 folds. `krr` is the paper's Pearson-kernel KRR; `ridge` is linear ridge with α selected by inner CV; `mlp256` is LayerNorm $\rightarrow$ Linear(d,256) $\rightarrow$ GELU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(256,1), trained with AdamW and early stopping on an inner 15% split. The raw-FC columns use the same folds, with the parametric heads applied to a train-fold PCA of FC.

Arm A findings. (i) A higher-capacity head does not rescue any timeseries BFM embedding. In 9 of the 12 model-by-dataset-by-head cells the parametric head is *below* the kernel readout, and it is never above it by more than $+0.017$. (ii) The embedding-to-FC gap survives correction everywhere: best

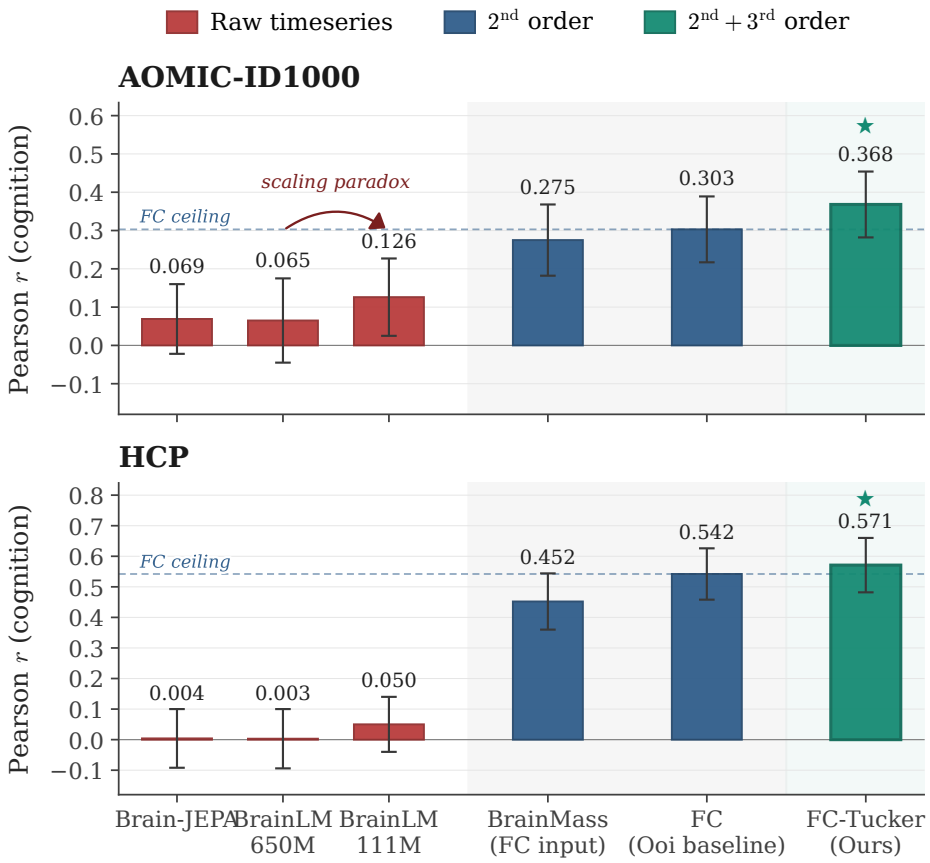


Figure 4: Prediction of the cognitive-ability composite (r, mean $\pm$ 1 std across 200 CV folds) on AOMIC and HCP. The three timeseries BFMs (left of the dashed line) sit near zero, with BrainLM-650M below BrainLM-111M; BrainMass, which takes FC as input, is higher but still below its own input-FC baseline. KRR on raw FC exceeds every BFM by a wide margin, and FC in the co-skewness Tucker subspace (right) improves on it on both datasets.

Table 5: Arm A. Frozen backbone, three readouts of increasing capacity, on identical folds (Pearson r, mean over 200 folds). For every timeseries BFM the higher-capacity heads are equal to or worse than the kernel readout.

		embedding			raw FC		
Dataset	Model	krr	ridge	mlp256	krr	ridge	mlp256
AOMIC	BrainLM-111M	0.126	0.104	0.084	**0.306**	0.252	0.200
	BrainLM-650M	0.065	0.002	0.034	**0.306**	0.252	0.200
	Brain-JEPA	0.069	0.053	0.050	**0.313**	0.300	0.237
	BrainMass	0.275	**0.311**	0.240	0.311	0.285	0.183
HCP	BrainLM-111M	0.050	0.015	0.032	**0.393**	0.385	0.361
	BrainLM-650M	0.003	0.020	0.008	**0.393**	0.385	0.361
	Brain-JEPA	0.004	0.011	−0.003	**0.352**	0.340	0.293
	BrainMass	0.452	0.440	0.423	**0.504**	0.484	0.433

embedding head against best raw-FC head is -0.180 and -0.343 for BrainLM-111M (AOMIC, HCP), -0.241 and -0.373 for BrainLM-650M, -0.243 and -0.341 for Brain-JEPA, all with $p_{\mathrm{NB}} < 0.001$. (iii) BrainMass is the exception the control framing predicts: on AOMIC ridge-on-embedding reaches 0.311 and ties raw FC ($\Delta = -0.001$, $p_{\mathrm{NB}} = 0.97$); on HCP it still trails by 0.052 ($p_{\mathrm{NB}} = 0.006$). (iv) Under every head, BrainLM-650M is at or below BrainLM-111M on the embedding.

Arm B protocol. Each backbone is finetuned *inside* each outer fold, on that fold's training subjects only, and the head then predicts the fully held-out fold. Fitting the finetuner inside the fold is what makes the comparison leak-free: a backbone finetuned once on a global training split and then scored under full-cohort CV would have seen the labels of subjects it is later tested on. Three heads: `krr` (differentiable KRR alignment on the embedding), `ridge` (linear head with MSE, the linear-probe finetune), and `mlp256` (as in Arm A, with MSE). **Arm B was run on AOMIC only**; finetuning dominates walltime (≈47 min per fold for BrainLM-650M, ≈65 min for Brain-JEPA), so BrainLM-111M ran 5 folds per head and the memory-heavy models 2–4. AOMIC is also the dataset where the frozen embeddings are strongest (0.07–0.13 against 0.00–0.05 on HCP), so this is the setting most favourable to the BFMs. The `raw FC` and `frozen` columns are Pearson-KRR on the same folds and serve only as a paired reference; the authoritative values for those methods are in Arm A.

Table 6: Arm B. End-to-end finetuning inside each fold, AOMIC (Pearson r, mean ± std over the n folds listed). Every model and head lands far below raw FC on the same folds, and BrainLM-650M finetunes to negative r.

Model	Head	n	raw FC	frozen	end-to-end finetuned
BrainLM-111M	krr	5	0.267 ± 0.030	0.140 ± 0.070	$\mathbf{0.138 \pm 0.077}$
	ridge	5	0.267 ± 0.030	0.140 ± 0.070	0.059 ± 0.072
	mlp256	5	0.267 ± 0.030	0.140 ± 0.070	0.034 ± 0.075
BrainLM-650M	krr	2	0.301 ± 0.008	0.093 ± 0.054	-0.042 ± 0.017
	ridge	3	0.285 ± 0.023	0.075 ± 0.051	$\mathbf{-0.033 \pm 0.061}$
	mlp256	3	0.285 ± 0.023	0.075 ± 0.051	-0.167 ± 0.039
Brain-JEPA	krr	4	0.313 ± 0.042	0.060 ± 0.062	$\mathbf{0.087 \pm 0.087}$
	ridge	3	0.304 ± 0.052	0.005 ± 0.038	0.008 ± 0.008
	mlp256	3	0.304 ± 0.052	0.005 ± 0.038	0.032 ± 0.054

Arm B findings. (i) End-to-end finetuning does not recover the target. The best number anywhere is 0.138; the conventional heads reach 0.034–0.059 for BrainLM and ≈0 for Brain-JEPA, against raw FC at 0.267–0.313 on the same folds. (ii) The larger model is worse under finetuning too: BrainLM-650M is negative on every head. (iii) The finetuned head can fall *below* the frozen embedding (650M mlp256 -0.167 against frozen 0.075; 111M ridge/mlp 0.059/0.034 against frozen 0.140), which is what supervised finetuning on ∼700 subjects does to a representation with little usable structure. (iv) BrainMass, run earlier on HCP with 10 folds as a harness check, shows no gain from end-to-end finetuning over frozen or raw FC.

Reading of the two arms together. Neither a higher-capacity frozen head nor full end-to-end finetuning lifts a timeseries BFM towards raw FC. The limit is on practical extractability: at the sample sizes these cohorts provide, the target cannot be recovered from these representations by the readouts in common use.

D VARIANCE BANDS: WHAT THE MODEL RECONSTRUCTS, WHAT PREPROCESSING REMOVES, AND WHAT PREDICTS THE TARGET

This appendix gives the full protocol and results behind Section 4.2. The object to isolate is variance that (a) the model captures, (b) rigorous preprocessing does not remove, and (c) does not predict the target.

Setup. BrainLM only, since a reconstruction model lets us measure "what the model captures" directly from its decoder output rather than through a proxy. Both sizes, both datasets. The inputs are the stored feature arrays: `inputs` is exactly the preprocessed timeseries the encoder saw, `reconstructions` its decoder output, both (N, P, T) with P=424.

Step 1, variance ladder. Concatenate the training subjects' z-scored timeseries over time and fit a spatial PCA with P components on the training fold only. Order components by explained variance and group them into fixed, log-spaced bands: B1 = PCs 1–5, B2 = 6–20, B3 = 21–50, B4 = 51–150, B5 = 151–P.

Step 2, three quantities per band. Project each subject's input and reconstruction onto the shared basis $\mathbf{V}$: $\mathbf{S}_{\text{in}} = \mathbf{V}^\top \mathbf{Z}_{\text{in}}$, $\mathbf{S}_{\text{rec}} = \mathbf{V}^\top \mathbf{Z}_{\text{rec}}$. Then per band b: reconstruction fidelity $F(b)$, the mean over components in b and over subjects of $\text{corr}_t(\mathbf{S}_{\text{in}}[c,:], \mathbf{S}_{\text{rec}}[c,:])$; variance share $\text{Var}(b)$, the eigenvalue mass in b; and band-wise prediction $C(b)$, the composite predicted from FC built inside band b's subspace alone under the same 20×10 protocol.

Step 3, nuisance model. High-variance fMRI structure is conventionally attributed to physiological and motion sources (Liu, 2016), which is what confound regression targets. A curated, standard confound design, not the full fMRIPrep column sink, which is rank-deficient at $T \approx 200$. On AOMIC, $m = 48$: the 36-parameter model of Satterthwaite et al. (2013) as benchmarked by Ciric et al. (2017), that is six motion parameters plus white matter, CSF and global signal, each with derivative, square and squared derivative; plus framewise displacement and standardised DVARS; plus the top five aCompCor components (Behzadi et al., 2007); plus five cosine drift terms. On HCP the movement regressors are not on disk in this release, so the design falls back to derived DVARS and global signal ($m = 2$); the AOMIC arm carries the argument and HCP is corroborating.

We report three measures per band. `cv` is ridge, 5-fold over interleaved timepoints, out-of-sample R^2: the fraction the confound model genuinely predicts. `is` is in-sample OLS R^2: what naive confound regression appears to remove, inflated by roughly q/T for *any* target. `null` is in-sample OLS R^2 after circularly shifting the same design out of alignment with the component scores, which is the chance level for `is`. If `is` $\approx$ `null`, the confounds explain nothing beyond chance.

Results. $\text{Var}(b)$, $C(b)$ and the nuisance measures are derived from the input signal, so they are identical across model sizes for a given dataset; only $F(b)$ is model-specific.

Reading. On AOMIC the top band holds 50.7% of all signal variance, is reconstructed by BrainLM-111M at the highest fidelity of any band (0.749), is not attributable to the confound model (`is` $= 0.277$ against a null of 0.287; `cv` $= -0.185$), and predicts the composite at 0.063. Prediction peaks at the lower-variance B2 (0.280), and $\text{corr}(\text{Var}, C) = -0.363$ across bands. $\text{corr}(F, C)$ is not stable in sign across cells, because C is not monotone in component rank. The robust statements are the B1 contrast and $\text{corr}(\text{Var}, C)$ on AOMIC.

Where the composite does live. On the same folds, FC restricted to the top two bands reaches 0.278 on AOMIC and 0.336 on HCP, full FC reaches 0.306 and 0.393, and FC in the rank-80 Tucker subspace reaches 0.329 and 0.400. The ordering is the same on both datasets: the high-variance, high-fidelity subspace the model reconstructs is the wrong place to look.

Table 7: Variance, reconstruction fidelity and prediction per band. AOMIC (N=876, P=424, T=200) and HCP (N=955). Var is the share of signal variance; C is the cognitive-ability composite predicted from FC inside that band alone (20 × 10 folds); F is BrainLM reconstruction fidelity.

Dataset	Band	PCs	$\mathrm{Var}(b)$	$C(b)$	$F(b)$ 111M	$F(b)$ 650M
AOMIC	B1	1–5	**0.507**	0.063	**0.749**	−0.011
	B2	6–20	0.157	**0.280**	0.559	−0.005
	B3	21–50	0.092	0.145	0.394	−0.006
	B4	51–150	0.120	0.106	0.263	−0.002
	B5	151–424	0.124	0.070	0.172	−0.001
HCP	B1	1–5	0.276	0.068	0.036	−0.013
	B2	6–20	0.097	**0.257**	0.011	−0.003
	B3	21–50	0.081	0.090	0.001	0.000
	B4	51–150	0.207	0.140	−0.001	−0.001
	B5	151–424	0.340	0.229	−0.001	−0.001

Table 8: Nuisance R^2 per band. `cv` is out-of-sample ridge R^2; `is` is in-sample OLS R^2; `null` is the circular-shift chance level for `is`. On AOMIC, with the full 48-regressor design, `is` ≈ `null` and `cv` is negative in every band: the confound model explains nothing beyond chance, including in the top-variance band.

		B1 (PCs 1–5)			B5 (PCs 151–424)		
Dataset	m	`cv`	`is`	`null`	`cv`	`is`	`null`
AOMIC	48	−0.185	0.277	0.287	−0.243	0.258	0.261
HCP	2	+0.214	0.241	0.037	−0.027	0.029	0.016

Scope. (a) The high-fidelity reconstruction result is AOMIC-only; on HCP BrainLM's reconstruction is near noise at both sizes, consistent with what the main text already reports. (b) The full 48-regressor confound test is AOMIC-only, since the HCP release we use lacks movement regressors; under the thin HCP design the top mode is partly explained (`cv` = 0.214), which is expected because PC1 of BOLD is essentially the global-signal mode, and global signal is itself contested as nuisance rather than neural. (c) $\mathrm{corr}(\mathrm{Var}, C)$ is −0.363 on AOMIC but ≈0 on HCP, where prediction also loads the high-index tail. The facts that hold on both datasets are: B1 is blind to the target (0.063, 0.068), prediction peaks at B2, and Tucker ≥ full FC > top-band FC.

Nonlinear confound models. A linear model can miss a nonlinear dependence on the same regressors, so we repeat the band-prediction test with kernel ridge under three model classes: a linear kernel, a degree-2 polynomial kernel (pairwise cross-terms) and an RBF kernel (arbitrary smooth nonlinearity). The setup is otherwise the one above, on 200 randomly drawn subjects per dataset, with the ridge penalty selected per component on an inner split of the training timepoints; with a single shared penalty the several hundred noise components dominate the criterion and force maximum shrinkage, which would report "no confound effect" by construction. Each model class is read against its own circular-shift null, since held-out R^2 for an unpredictable target sits slightly below zero. Per-component penalty selection makes these numbers a different estimator from the shared-ridge `cv` column of Table 8; they are comparable within this table, not across.

On AOMIC every kernel sits at or below its own null in every band, so the dominant band is not recoverable from the 48 regressors by a nonlinear model any more than by a linear one. On HCP the top band is partly predictable from the two available regressors (+0.216 linear against a null of +0.007), but the nonlinear kernels do not improve on the linear one, so what is recoverable there is a linear dependence rather than a missed nonlinear one. Neither dataset gives a model class under which the dominant band becomes attributable to the confound design.

A note on what AOMIC preprocessing already removed. The AOMIC pipeline already regresses six rigid-body motion parameters and applies detrending and a 0.01–0.1 Hz bandpass (Appendix U). Part of the null is therefore close to tautological for the motion columns specifically. The informative

Table 9: Out-of-sample R^2 of the confound model predicting each band's component scores, by model class, against the circular-shift null (in parentheses). AOMIC uses the 48-regressor design, HCP the 2 regressors its release provides. N=200 subjects per dataset.

Dataset	Kernel	B1 (PCs 1–5)	B2 (PCs 6–20)	B5 (PCs 151–424)
AOMIC	linear	-0.018 (-0.019)	-0.037 (-0.039)	-0.045 (-0.051)
	poly-2	-0.022 (-0.024)	-0.034 (-0.039)	-0.043 (-0.049)
	RBF	-0.006 (-0.003)	-0.022 (-0.020)	-0.031 (-0.030)
HCP	linear	$+0.216$ ($+0.007$)	$+0.011$ (-0.007)	-0.023 (-0.033)
	poly-2	$+0.149$ (-0.053)	-0.062 (-0.065)	-0.094 (-0.096)
	RBF	$+0.210$ ($+0.004$)	$+0.004$ (-0.014)	-0.039 (-0.048)

part is that white matter, CSF, global signal and aCompCor, which are *not* already removed, also explain nothing beyond chance in any band.

E Decoding cumulants from the frozen latent

The cumulant-preservation analysis of Appendix R compares an input to a reconstruction, so it cannot be run on Brain-JEPA, which predicts representations in latent space and has no decoder. This appendix asks the same question in a form that needs no decoder: how much of the input's κ_2, and how much of its κ_3, is linearly decodable from the frozen embedding? If κ_2 is decodable and κ_3 is not, the latent has discarded third-order structure regardless of how the model was trained.

Protocol. For each model and dataset, 5 repetitions $\times$ 10 folds, everything fitted on the training fold only. (1) Fit one shared Tucker basis $\mathbf{U} \in \mathbb{R}^{P\times R}$, R=40, on the training subjects' co-skewness tensor; both cumulants are then measured inside this same subspace, so the two targets cannot differ because of the basis. (2) Build both targets from the exact timeseries the encoder saw: κ_2 is the lower triangle of $\mathrm{corrcoef}(\mathbf{U}^\top\mathbf{Z})$ (780 features), κ_3 the unique co-skewness triples of the standardised projection (11,480 features). (3) Reduce each target to its top K=30 principal components, fitted on the training fold, so dimensionality cannot confound the comparison. (4) Ridge-regress each of the 30 component scores from the predictor, with α by efficient leave-one-out on the training fold; the metric is out-of-fold Pearson r per component, aggregated as a variance-weighted mean. (5) Four predictors on identical folds: `embedding` (the frozen readout), `fc_input` (the raw $P \times P$ input FC, a rich but purely second-order representation of the same data), `fc_tucker` (FC computed inside the same Tucker subspace, the paper's own representation, as the reference for how much of κ_3 a good classical representation carries), and `shuffled` (the embedding matrix with rows permuted, the empirical null). (6) A split-half noise ceiling: recompute both feature sets on each half of a subject's scan in the same subspace, project onto the same component basis, correlate half against half across held-out subjects, Spearman-Brown corrected.

Table 10: Decoding r into the shared Tucker basis (mean $\pm$ std over 50 folds). Every embedding carries κ_2 above its shuffled null. Brain-JEPA carries κ_3 at the null; BrainLM carries some, most on AOMIC at 111M ($+0.095$), still well below what FC in the same subspace carries.

Model	Dataset	emb $\to \kappa_2$	emb $\to \kappa_3$	null $\to \kappa_3$	FC $\to \kappa_3$	FC-Tucker $\to \kappa_3$
Brain-JEPA	AOMIC	$+0.076 \pm 0.023$	$+0.015 \pm 0.025$	$+0.014 \pm 0.027$	$+0.223 \pm 0.041$	$\mathbf{+0.238 \pm 0.050}$
Brain-JEPA	HCP	$+0.033 \pm 0.018$	$+0.008 \pm 0.020$	-0.002 ± 0.017	$+0.162 \pm 0.021$	$\mathbf{+0.200 \pm 0.025}$
BrainLM-111M	AOMIC	$+0.342 \pm 0.017$	$+0.095 \pm 0.054$	$+0.007 \pm 0.029$	$+0.198 \pm 0.039$	$\mathbf{+0.230 \pm 0.052}$
BrainLM-111M	HCP	$+0.144 \pm 0.022$	$+0.024 \pm 0.021$	$+0.006 \pm 0.019$	$+0.175 \pm 0.028$	$\mathbf{+0.219 \pm 0.032}$
BrainLM-650M	AOMIC	$+0.216 \pm 0.019$	$+0.027 \pm 0.029$	-0.007 ± 0.017	$+0.198 \pm 0.039$	$\mathbf{+0.230 \pm 0.052}$
BrainLM-650M	HCP	$+0.171 \pm 0.020$	$+0.038 \pm 0.023$	$+0.007 \pm 0.018$	$+0.175 \pm 0.028$	$\mathbf{+0.219 \pm 0.032}$

Findings. Brain-JEPA's latent decodes κ_2 well above its null ($+0.076$ against -0.018 on AOMIC, $p_{\mathrm{NB}} < 10^{-4}$; $+0.033$ against -0.019 on HCP) and decodes κ_3 *at* the null ($+0.015$ against $+0.014$, $p_{\mathrm{NB}} = 0.93$; $+0.008$ against -0.002, $p_{\mathrm{NB}} = 0.24$). The within-embedding contrast κ_2 against κ_3 is

Table 11: Split-half reliability of the component scores (Spearman-Brown corrected). κ_3 is genuinely noisier than κ_2 at these scan lengths, which is why the `fc_input` and `fc_tucker` references matter: they show what a representation *can* carry under the same noise.

Input	reliability κ_2	reliability κ_3
AOMIC, Schaefer-400+Tian-50 $\times$ 160 TP (Brain-JEPA)	0.701 ± 0.021	0.271 ± 0.070
HCP, Schaefer-400+Tian-50 $\times$ 160 TP (Brain-JEPA)	0.513 ± 0.022	0.122 ± 0.027
AOMIC, AAL-424 $\times$ 200 TP (BrainLM)	0.699 ± 0.019	0.255 ± 0.077
HCP, AAL-424 $\times$ 200 TP (BrainLM)	0.657 ± 0.015	0.170 ± 0.028

significant under the corrected test in every model and dataset ($p_{\mathrm{NB}} \leq 0.006$). BrainLM's embedding carries more κ_3 than Brain-JEPA's but remains far below what FC in the same subspace carries. So a latent-prediction objective is no more able to retain third-order structure than a reconstruction objective.

The `fc_tucker` reference. FC-Tucker decodes κ_3 slightly better than raw FC on the full scan ($+0.015$ to $+0.043$, corrected-significant in three of four cells). `fc_tucker` and the κ_3 target are computed from the same projected timeseries in the same subspace and therefore share estimation noise. We ran the disjoint-half control that removes the shared noise, taking κ_2 from scan half 1 and the κ_3 target from scan half 2: the advantage over raw FC vanishes ($+0.012$, -0.006, $+0.003$, $+0.004$ across the four cells, none significant). `fc_tucker` is reported as the reference column for what a strong classical representation carries.

F Replication on independent OpenNeuro cohorts

Section 4.4 rests the FC-Tucker claim on effect size and on agreement in direction across four cells, because the HCP cells are underpowered under our corrected test. We test whether the direction reproduces on cohorts the method was not developed on.

Protocol. Seven OpenNeuro cohorts from our parcellated corpus with both Schaefer-400 parcellated fMRI and a genuine cognitive or behavioural measure. Target: PC1 of the cohort's cognitive variables, z-scored, median-imputed, sign-aligned so higher is better; where only one variable exists it is used directly. Comparison: raw FC against FC in the rank-R co-skewness subspace, Tucker basis fitted on the training fold only, both on identical 20×10 folds. Volumes with FD > 0.3 mm censored; scans with fewer than 120 usable volumes dropped. Two rank protocols: fixed R=80, and nested-R with $R \in \{20, 40, 80, 120, 160\}$ selected by inner 5-fold CV on the training fold, which is the paper's own protocol.

Table 12: FC-Tucker minus raw FC on seven independent cohorts, under both rank protocols. Positive is a Tucker win. No cohort clears the corrected resampled test under either protocol ($p_{\mathrm{NB}} \geq 0.10$ throughout), and the sign flips between protocols in three cohorts.

Cohort	Type	N	T	raw FC	Δ at R=80	Δ nested-R
UCLA-CNP (ds000030)	rest	257	151	$+0.277$	$+0.010$	-0.060
NeurocogAging (ds003592)	rest	287	200	$+0.408$	$+0.042$	$+0.038$
AOMIC-PIOP2 (ds002790)	rest	221	236	$+0.060$	$+0.037$	-0.026
AOMIC-PIOP1 (ds002785)	rest	202	477	$+0.243$	-0.119	$+0.022$
ds000115 (N-back)	task	96	136	$+0.358$	$+0.024$	$+0.019$
SUDMEX-CONN (ds003346)	rest	114	265	-0.090	$+0.067$	$+0.021$
ds004469	rest	50	190	$+0.031$	$+0.256$	$+0.063$

Reading. FC-Tucker is ahead in six of seven cohorts at fixed rank and five of seven under nested rank selection, across rest and task, N from 50 to 287, and four different cognitive instruments. A one-sided sign test on the fixed-rank pass gives $p = 0.063$ and on the nested pass $p = 0.227$; an inverse-variance meta-analysis over the fixed-rank cohorts gives $\Delta = +0.024 \pm 0.025$ ($p = 0.34$), with $I^2 = 4\%$, so the cohorts are consistent with one small common effect but that effect is not

distinguishable from zero. No individual cohort clears the corrected resampled test, and only four of seven are positive under *both* rank protocols. At fixed rank AOMIC-PIOP1 goes the other way by -0.119, and it is by far the longest-scan cohort ($T \approx 477$), where a 400×400 FC is already well estimated so a rank-80 truncation costs more than it denoises. Under nested-R the inner CV selects R=20 there in 153 of 200 folds and the sign flips to $+0.022$.

Three of these targets are weak proxies: NeurocogAging uses years of education in an explicitly young-versus-old sample, so its $r \approx 0.41$ probably carries age; SUDMEX-CONN uses an education and SES composite and both methods are near zero; ds004469 has N=50 and fold standard deviations near 0.5. We therefore present this as a scope result: the advantage is reproducible in direction on small, short-scan, single-instrument cohorts, is small there. Those cohorts differ from ours in sample size, scan length and target reliability.

G STATISTICAL TESTING NOTES

We adopt the statistical protocol of Ooi et al. (2022): the corrected resampled t-test of Nadeau & Bengio (2003) with Benjamini–Hochberg FDR ($q < 0.05$) per dataset. For r repeats $\times$ k folds with per-fold paired differences d_i, mean $\bar{d}$, sample variance s_d^2, and per-fold test/train sizes $n_{\text{test}}, n_{\text{train}}$, the corrected statistic is

$$t_{\text{NB}} = \frac{\bar{d}}{\sqrt{\left(\frac{1}{rk} + \frac{n_{\text{test}}}{n_{\text{train}}}\right) s_d^2}}, \qquad \text{df} = rk - 1. \tag{4}$$

For our setup (r=20, k=10, family-aware), the correction factor is $\frac{1}{200} + \frac{1}{9} \approx 0.116$, dominated by the $\frac{1}{k-1}$ term, so the effective standard error is roughly $4.8\times$ the naive standard error.

Per-test results. Table 13 reports both the naive paired t-test p (which assumes fold independence and is therefore not strictly valid here) and the NB+BH-FDR-corrected p for every comparison cited in Sections 4.4 and Appendix I. Naive p-values are reported only for transparency: they range from $\sim 10^{-11}$ to $\sim 10^{-56}$ and would all be considered significant at any conventional threshold, but the dependence between overlapping CV training sets makes those numbers unreliable as inferential evidence.

Table 13: All paired comparisons cited in the main text under the naive paired t-test (assumes independence, invalid here) and the Nadeau–Bengio corrected t-test with BH-FDR within dataset (Ooi protocol). Effect sizes (Cohen's d) are unaffected by the dependence structure and are the primary statistic.

Comparison	Cell	Δr	d	naive p	NB+FDR q
Tucker(R^*) vs FC-full (Table 2)					
FC-Tucker>FC-full	AOMIC AAL	$+0.065$	0.90	1.7×10^{-27}	$\mathbf{0.034}^{\star}$
FC-Tucker>FC-full	AOMIC Schaefer	$+0.060$	0.82	4.9×10^{-24}	$\mathbf{0.034}^{\star}$
FC-Tucker>FC-full	HCP AAL	$+0.038$	0.62	6.6×10^{-16}	0.092
FC-Tucker>FC-full	HCP Schaefer	$+0.028$	0.49	6.6×10^{-11}	0.153
Tucker(R=K) vs PCA(K=R) at matched dim (Table 15)					
FC-Tucker>FC-PCA	AOMIC AAL	$+0.085$	1.57	9.7×10^{-56}	$< \mathbf{10^{-4}}^{\star}$
FC-Tucker>FC-PCA	AOMIC Schaefer	$+0.038$	0.74	1.3×10^{-20}	$\mathbf{0.032}^{\star}$
FC-Tucker>FC-PCA	HCP AAL	$+0.041$	1.21	5.9×10^{-41}	$< \mathbf{10^{-3}}^{\star}$
FC-Tucker>FC-PCA	HCP Schaefer	$+0.004$	0.17	0.018	0.621
Inline					
FC-PCA(K=117) vs FC-full	AOMIC AAL	-0.019	-0.24	9.5×10^{-4}	0.487
FC-Tucker(R=80) vs FC-full	HCP Schaefer	$+0.020$	0.40	6.0×10^{-8}	0.244

Caveat: NB is conservative. The NB correction is conservative: the dominant $\frac{1}{k-1}$ term in the correction factor effectively treats the J=r repeats as adding no independent information beyond the k base folds, so the effective sample size collapses toward $k-1$. Bengio & Grandvalet (2004) prove

there is no universal unbiased estimator of the variance of k-fold CV, so any test for repeated CV is necessarily a bias/variance trade-off; Bouckaert & Frank (2004) empirically compare popular tests and recommend the NB-corrected r-times k-fold CV test (the statistic we use) for best replicability. The q-values above should therefore be read as upper bounds.

What would lower the q-values. For our setup, $t_{\text{NB}} \propto d\sqrt{N(k{-}1)/k}$ in the large-r limit. More repetitions (r) reduces only the $1/(rk)$ term, which is already negligible against $1/(k{-}1)$; more folds (k) reduces the correction factor but proportionally inflates per-fold variance ($s_d^2 \propto k/N$), so the two effects roughly cancel. Only more subjects (N) actually scales the test statistic: t_{NB} grows as $\sqrt{N}$. To bring HCP Schaefer (Tucker>FC-full, $d{=}0.49$) below $q{=}0.05$ would require $N \approx 2{,}000$ subjects (versus our 955); reaching $q{=}0.01$ would require $N \approx 5{,}000$. UK-Biobank-scale validation (planned future work) closes this gap trivially.

Summary. The basis-criterion claim (FC-Tucker > FC-PCA at matched dimensionality) is significant under NB+FDR in 3 of 4 cells (the cells where the basis matters most; the failed HCP-Schaefer-87 cell has $d{=}0.17$, where the Schaefer atlas already supplies a clean spatial basis). The Tucker-over-FC-full claim is significant under NB+FDR in the two larger-effect AOMIC cells and consistent in direction in all 4 cells; the HCP cells (smaller d, larger N needed) survive on effect-size and replicated-direction evidence rather than on a single ultra-low p-value.

H PARTIAL-CORRELATION ROBUSTNESS

The κ_2 measure in the main text is Pearson FC. A natural robustness check is to swap it for the other standard κ_2 summary, partial correlation (Ledoit–Wolf-shrunk precision matrix rescaled to $-P_{ij}/\sqrt{P_{ii}P_{jj}}$), which conditions each pairwise association on all other ROIs. We rerun FC-full and FC-Tucker($R{=}80$) with partial correlation on all four dataset×parcellation cells, using the same 200-fold CV.

Table 14: Pearson vs. partial correlation for κ_2 (mean ± std, $n{=}200$). Partial FC-full never beats Pearson FC-full; on HCP it is substantially worse (shrinkage cannot fully regularise $P \gg T$ regimes at short scans). The Tucker subspace narrows the gap but does not surpass Pearson. The main-text headline ($r = 0.571$) stands regardless of the κ_2 choice.

Dataset	FC-full (Pearson)	FC-full (Partial)	FC-Tucker $R{=}80$ (Partial)	FC-Tucker R^* (Pearson)
AOMIC AAL	$\mathbf{0.303 \pm 0.086}$	0.296 ± 0.092	0.237 ± 0.087	0.368 ± 0.086
AOMIC Schaefer	$\mathbf{0.346 \pm 0.084}$	0.325 ± 0.086	0.318 ± 0.086	0.406 ± 0.079
HCP AAL	$\mathbf{0.393 \pm 0.097}$	0.235 ± 0.099	0.341 ± 0.108	0.431 ± 0.092
HCP Schaefer	$\mathbf{0.542 \pm 0.084}$	0.349 ± 0.093	0.530 ± 0.087	0.571 ± 0.089

Dual-moment-finetuned BrainLM reconstructions track the same ordering: FC-recon at $\lambda{=}0$ reaches $r = 0.301 \pm 0.093$ (111M) and $r = 0.295 \pm 0.093$ (650M) on AOMIC under partial correlation, and $r = 0.309 \pm 0.081$ / 0.312 ± 0.084 under FC-Tucker($R{=}80$) partial, versus $r = 0.307 \pm 0.085$ / 0.304 ± 0.084 Pearson (main Table 3). The BFM-vs-raw story and the dual-moment gain both survive the κ_2 swap. Partial correlation is routed through `utils.evaluation.fc_vec_partial` and evaluated under the same KRR protocol as every other number in the paper.

I TUCKER VS. PCA AT MATCHED DIMENSIONALITY

To isolate the second- vs. third-order subspace-criterion effect, we compare FC-Tucker and FC-PCA at identical $R = K$: same feature count, same KRR pipeline, only the basis differs. Table 15 reports the paired comparison at the Tucker-optimal R^* of each combination. The effect is strongest on AOMIC AAL ($d = 1.57$) and HCP AAL ($d = 1.21$), and weakens on Schaefer, where the atlas already provides a clean spatial basis. At matched $R{=}K$, Tucker and PCA share the same feature count and pipeline, so their per-fold r values are highly correlated (0.79–0.95) and the SD of the difference is small; against FC-full the pipelines differ more (correlation 0.60–0.81), so even a larger Δr can yield a smaller t and d (Table 2 vs. Table 15).

Table 15: Tucker vs. PCA at matched dimensionality ($R = K$): paired fold-level comparison (n=200). Δr is the mean within-fold difference in Pearson r; significance details in Appendix G. Since the *only* difference between the two pipelines is whether the subspace is selected by κ_2 (PCA) or κ_3 (Tucker), the effect isolates third-order structure.

Dataset	Matched dim.	Δr	Cohen's d
AOMIC AAL	R=K=117	+0.085	**1.57**
AOMIC Schaefer	R=K=122	+0.038	0.74
HCP AAL	R=K=167	+0.041	**1.21**
HCP Schaefer	R=K=87	+0.004	0.17

J TEMPORAL VS. SPATIAL REDUCTION

Section 4.4 concludes that Tucker's advantage is spatial, not temporal. This appendix backs that claim with a matched ablation. We reduce the *temporal* dimension via PCA or Tucker (from T to K temporal components), keeping the full $P \times P$ FC matrix, and compare against the spatial reduction used in the main text (from P to R spatial components). Temporal PCA fits PCA(K) on the concatenated training timeseries (each subject-ROI as a row); Temporal Tucker takes the top-K eigenvectors of the group spatially-weighted temporal gram $\mathbf{M} = (1/N)\sum_n \mathbf{Z}_n^\top(\mathbf{C}_n \odot \mathbf{C}_n)\mathbf{Z}_n/P^2$ with $\mathbf{C}_n = \mathbf{Z}_n\mathbf{Z}_n^\top$ the spatial gram, so the basis is shaped by fourth-order spatial structure rather than raw temporal variance (and is therefore distinct from a single-subject 2D-timeseries SVD).

Table 16: Temporal PCA/Tucker of the ROI timeseries (full $P \times P$ FC retained), swept over K (Pearson r, mean $\pm$ std, n=200; "best K" is the sweep-optimal value among the temporal Tucker components). Across all four cells the temporal-reduced result is at most marginally below FC-full (within $\sim$0.01 in Pearson r on every cell). Temporal denoising alone does not improve prediction of the composite.

Dataset / parcellation	FC-full r	Temporal-reduced r (best K)	Best K
AOMIC AAL-424	0.303 ± 0.086	0.303 ± 0.087	72
AOMIC Schaefer-400	0.346 ± 0.084	0.346 ± 0.085	102
HCP AAL-424	0.393 ± 0.097	0.393 ± 0.097	172
HCP Schaefer-400	0.542 ± 0.084	0.534 ± 0.086	400

The cognitive signal is structured spatially (which regions co-activate), not temporally (when they co-activate). This asymmetry is also consistent with the BFM diagnosis in Section 4.1: reconstruction-based BFMs primarily model temporal dynamics of the raw signal, the dimension that carries no additional cognitive information beyond what FC already captures.

K PER-VARIABLE TUCKER BEHAVIOUR ON THE HCP BEHAVIOURAL BATTERY

The composite target in the main text aggregates the 58 HCP behavioural measures used by Ooi et al. (2022). To probe whether the Tucker advantage is a property of the composite or extends to its components, we ran FC-full and FC-Tucker (at $R \in \{87, 167, 337\}$) on each of the 57 usable variables (ER40HAP returned NaN across methods), at 10×10 family-aware CV on both AAL-424 and Schaefer-400. This is a preliminary protocol relative to the 20×10 of the main text, but the cross-parcellation pattern is consistent.

Of 57 variables, 17 favour FC-Tucker on both parcellations, 20 favour FC-full on both, and 20 are mixed (FC-Tucker wins on one parcellation and loses or ties on the other). Table 17 reports the 17 consistent Tucker winners. **MMSE_Score** is the most consistent cognitive case ($\Delta r = +0.051$ on AAL, $+0.081$ on Schaefer; the Tucker advantage grows under the better parcellation). The 20 mixed cases tend to flip in the direction predicted by parcellation quality: a cleaner atlas lifts FC-full into the regime where Tucker's low-rank trade-off costs more than it gains, which is also why the Tucker Δr on Schaefer is typically smaller than on AAL.

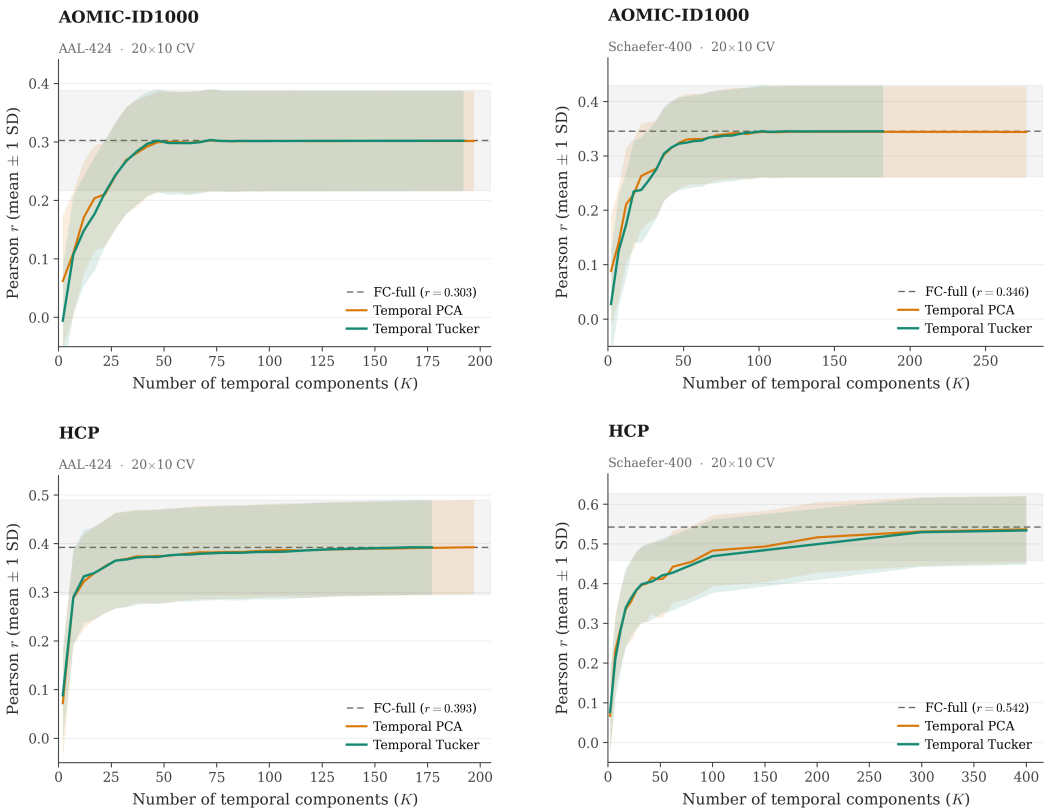


Figure 5: Temporal reduction sweeps across the four dataset×parcellation cells. Temporal PCA (orange) and temporal Tucker (green) are swept over K with the full $P \times P$ FC retained. The FC-full baseline (dashed) is indistinguishable from either temporal reduction at every K.

Table 17: HCP behavioural variables on which FC-Tucker beats FC-full on *both* AAL-424 and Schaefer-400, sorted by $\min(\Delta r_{\mathrm{AAL}}, \Delta r_{\mathrm{Sch}})$. Pearson r at the per-variable optimal $R \in \{87, 167, 337\}$, 10×10 family-aware CV (preliminary; main-text protocol is 20×10). Per-cell standard deviations across the 100 fold-level evaluations are 0.08–0.10 throughout (omitted from cells for compactness; available in the per-variable result CSVs).

	AAL-424			Schaefer-400		
Variable	FC-full	FC-Tuck	Δr	FC-full	FC-Tuck	Δr
Social_Task_Perc_Random	-0.027	$+0.047$	$+0.074$	-0.018	$+0.081$	$+0.100$
Emotion_Task_Face_Acc	-0.027	$+0.075$	$+0.101$	-0.024	$+0.046$	$+0.070$
PosAffect	-0.033	$+0.025$	$+0.058$	$+0.026$	$+0.086$	$+0.060$
MMSE_Score	$+0.033$	$+0.084$	$+0.051$	$+0.009$	$+0.090$	$\mathbf{+0.081}$
PercHostil	-0.025	$+0.056$	$+0.081$	-0.017	$+0.028$	$+0.045$
GaitSpeed_Comp	-0.048	$+0.065$	$+0.113$	$+0.043$	$+0.075$	$+0.032$
IWRD	$+0.013$	$+0.037$	$+0.024$	$+0.020$	$+0.048$	$+0.028$
SCPT_SPEC	-0.020	$+0.002$	$+0.022$	-0.005	$+0.050$	$+0.055$
LifeSatisf	$+0.046$	$+0.087$	$+0.040$	$+0.132$	$+0.155$	$+0.022$
Mars_Final	-0.026	$+0.017$	$+0.043$	-0.001	$+0.018$	$+0.019$
ER40ANG	$+0.014$	$+0.033$	$+0.018$	$+0.033$	$+0.061$	$+0.028$
MeanPurp	-0.044	-0.023	$+0.021$	$+0.052$	$+0.069$	$+0.017$
Dexterity	$+0.047$	$+0.063$	$+0.015$	$+0.102$	$+0.161$	$+0.059$
InstruSupp	-0.022	$+0.010$	$+0.032$	$+0.034$	$+0.049$	$+0.015$
ProcSpeed	$+0.021$	$+0.063$	$+0.043$	$+0.120$	$+0.133$	$+0.013$
NEOFAC_N	-0.034	$+0.000$	$+0.034$	$+0.079$	$+0.087$	$+0.008$
ER40FEAR	$+0.046$	$+0.067$	$+0.021$	$+0.055$	$+0.057$	$+0.002$

L MOTION CONFOUND CHECK ON THE TUCKER SUBSPACE

The co-skewness tensor captures non-Gaussian structure, and so does motion (transient framewise displacement, cardiac and respiratory rhythms): a natural worry is that the Tucker subspace amplifies motion-related variance rather than variance that predicts the composite. We test this directly. For each subject we project the timeseries into the rank-R=80 Tucker basis fit on training subjects (Section 3.2) and compute the per-component temporal variance. Across subjects, we then correlate each component's per-subject variance with two motion summaries: mean framewise displacement (FD; AOMIC, from fMRIPrep confounds) and mean DVARS (per-timepoint root-mean-square BOLD-derivative magnitude; from fMRIPrep on AOMIC, derived from the cleaned timeseries on HCP, since HCP `Movement_Regressors` are not retained after ICA-FIX in our local copy). PCA components fit on the same training set serve as a no-third-order control.

Median and max $|r|$ across the 80 components per cell are reported in Table 18. On the cleanest motion measure available (mean FD on AOMIC), median $|r|$ is 0.118 for Tucker and 0.085 for PCA, with maximum $|r|$ of 0.388 for Tucker and 0.436 for PCA: the most motion-loaded component is in PCA, not Tucker. Across every cell, the Tucker median tracks the PCA median to within ~ 0.03, and the Tucker maximum tracks the PCA maximum to within ~ 0.05. Whatever motion correlation the components carry is inherited from variance estimation, not introduced by the third-order subspace selection. The κ_3-informed basis is no more motion-loaded than a generic variance basis. The variance the Tucker subspace amplifies is not motion variance.

The HCP AAL-424 row reports the largest absolute correlations in the table (0.30 median, 0.73 max), but Tucker and PCA are essentially indistinguishable on it (0.301 vs 0.271 median; 0.729 vs 0.729 max). Derived DVARS on a cleaned timeseries also captures legitimate high-frequency signal, not only residual motion, which inflates correlations in both bases. The relative comparison between Tucker and PCA, the only one that bears on the κ_3-vs-κ_2 claim, is what matters here, and it shows no preferential motion-loading from the third-order subspace selection.

M FINETUNING DETAILS

This appendix expands on Section 4.5. Subsection M.1 reports the full strategy sweep on AOMIC (six losses $\times$ two model sizes). Subsection M.2 reports the HCP-side Log-Cholesky results and the

Table 18: Cross-subject correlations between per-component temporal variance and motion summaries. Median and maximum $|r|$ across the $R{=}K{=}80$ components per cell. AOMIC ($N{=}876$): mean FD and mean DVARS from fMRIPrep confounds. HCP ($N{=}955$): mean DVARS computed from the ICA-FIX-cleaned timeseries (no `Movement_Regressors` locally). PCA fit on the same training set is a no-third-order control. The AOMIC Schaefer-400 cell was not computed.

		Tucker		PCA (control)	
Cell	Motion summary	median $\|r\|$	max $\|r\|$	median $\|r\|$	max $\|r\|$
AOMIC AAL-424	mean FD	0.118	0.388	0.085	0.436
AOMIC AAL-424	mean DVARS (fMRIPrep)	0.124	0.373	0.086	0.349
HCP AAL-424	mean DVARS (derived)	0.301	0.729	0.271	0.729
HCP Schaefer-400	mean DVARS (derived)	0.133	0.240	0.121	0.235

cross-dataset transfer of AOMIC-trained dual-moment checkpoints. Subsection M.3 sweeps the dual-moment λ and compares the two κ_3 surrogates. Subsection M.4 adds a sub-λ grid and a $5\times$ longer schedule to rule out the AOMIC ceiling being a sweep or schedule artefact. Subsection M.5 discusses the KRR-alignment metric-training tautology for Brain-JEPA and BrainMass. Subsection M.6 reports Track C, direct HCP-scale training of the dual-moment loss, which closes the BFM-to-raw gap on HCP.

M.1 BrainLM: complete strategy sweep (AOMIC, AAL-424)

All six finetuning strategies tested on BrainLM, plus the pretrained baseline, under the CV protocol of Section 3.4. Embedding r is KRR on the CLS token; FC-recon r is KRR on FC computed from the reconstructed timeseries. Dual-moment rows report the overall optimum (`fc_tucker` at $\lambda{=}10^{-3}$) from the sub-λ sweep in Appendix M.4.

Table 19: BrainLM finetuning strategies on AOMIC (AAL-424). Pearson r (mean $\pm$ std, $n{=}200$). Only the dual-moment (Log-Cholesky in the κ_3-optimal Tucker subspace, $\lambda{=}10^{-3}$ from the sub-λ sweep, Appendix M.4) closes FC-recon to the raw FC baseline ($0.308 \approx 0.306$ for 111M, 0.304 for 650M). No strategy meaningfully improves the embedding.

Strategy	111M Emb. r	111M FC-recon r	650M Emb. r	650M FC-recon r
Pretrained (no FT)	0.126 ± 0.101	0.175 ± 0.090	0.065 ± 0.110	0.164 ± 0.089
Log-Cholesky FC	0.093 ± 0.114	0.247 ± 0.087	0.012 ± 0.093	0.277 ± 0.089
KRR alignment	0.116 ± 0.118	0.099 ± 0.108	—	—
Cognition similarity	0.100 ± 0.119	0.118 ± 0.105	—	—
Direct regression	0.127 ± 0.113	0.137 ± 0.102	—	—
Dual-moment (`fc_tucker`, $\lambda{=}10^{-3}$)	$\mathbf{0.082 \pm 0.118}$	$\mathbf{0.308 \pm 0.085}$	$\mathbf{0.034 \pm 0.104}$	$\mathbf{0.304 \pm 0.084}$
Raw FC baseline	—	0.306 ± 0.084	—	0.306 ± 0.084

Two patterns dominate Table 19. First, the dual-moment loss is the only objective that brings FC-recon to the raw FC baseline at both model scales. The FC reconstructed from the BFM forward pass carries essentially the same predictive signal as the raw FC the model was trained to reconstruct. Second, no strategy meaningfully changes the embedding. Direct regression to the composite (full supervision) matches the pretrained embedding. This is consistent with the main-text diagnosis: BrainLM's CLS latent geometry is committed to noise structure during pretraining and does not reorganise under losses that reach it only through the decoder.

M.2 BrainLM on HCP: Log-Cholesky and cross-dataset dual-moment

The HCP FC-reconstruction rows are reported in the main Table 3; the Log-Cholesky rows are trained directly on HCP and the dual-moment rows apply the *AOMIC-trained* `fc_tucker` $\lambda{=}0$ checkpoint to HCP timeseries with no HCP-specific finetuning. Table 20 reports the matching CLS embedding readouts.

Table 20: BrainLM CLS embedding r on HCP under FC-preservation finetuning. Embeddings stay near zero under every loss.

Model	Condition	FT data	Emb. r
BrainLM-111M	Pretrained		0.050 ± 0.090
BrainLM-111M	Log-Chol FT	AOMIC	0.025 ± 0.090
BrainLM-111M	Dual-moment FT	AOMIC	0.056 ± 0.100
BrainLM-111M	Dual-moment FT	HCP	0.015 ± 0.089
BrainLM-650M	Pretrained		0.003 ± 0.097
BrainLM-650M	Log-Chol FT	AOMIC	-0.018 ± 0.098
BrainLM-650M	Dual-moment FT	AOMIC	-0.033 ± 0.105

We draw three observations from the main-text FC-reconstruction rows. First, reconstruction gains on HCP are large in absolute terms ($+0.29$ and $+0.38$ for Log-Cholesky at 111M and 650M; $+0.34$ and $+0.41$ for dual-moment), because HCP pretrained reconstructions are near-noise. Second, the AOMIC-trained dual-moment checkpoint *transfers to HCP*: it improves FC-recon by $+0.05$ (111M) and $+0.03$ (650M) over the HCP-trained Log-Cholesky, and narrows the gap to raw FC from -0.043 (Log-Cholesky 650M) to -0.013 (dual-moment 650M). Third, the dual-moment FC-recon on HCP reaches 96% of the raw FC baseline under cross-dataset transfer, with $T \approx 200$–290 TP training data (AOMIC) and T=4,800 evaluation data (HCP). The κ_3-subspace objective is not AOMIC-specific; the remaining gap to raw FC is closed by HCP-scale training in Appendix M.6.

M.3 Dual-moment λ sweep (AOMIC)

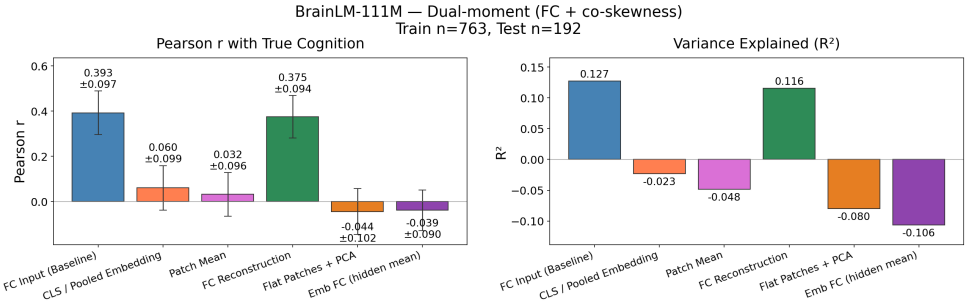
The natural follow-up to Log-Cholesky is the joint $\kappa_2 + \kappa_3$ regulariser $\mathcal{L}(\lambda)$ defined in Section 3.7, with two surrogates for d_{κ_3} (no closed-form Riemannian metric exists for third-order symmetric tensors):

- `coskew_mse`: MSE between vectorised co-skewness tensors S_{ijk} in the rank-R Tucker subspace (R=80, 88,560 unique entries). A direct κ_3 loss.
- `fc_tucker`: Log-Cholesky distance between FCs computed on timeseries projected into the rank-R Tucker subspace. A κ_2 loss in the κ_3-optimal basis.

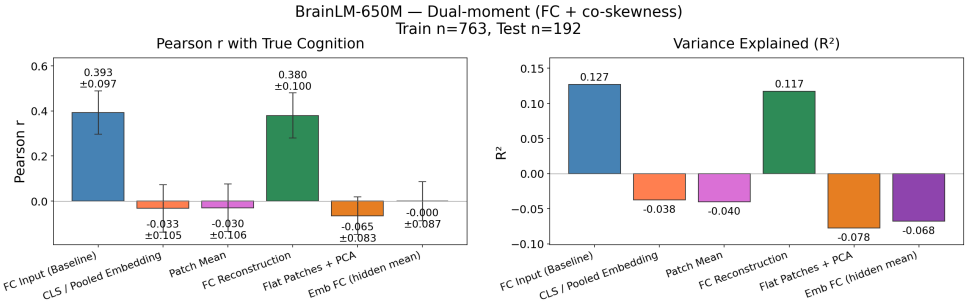
The Tucker factor matrix $\mathbf{U} \in \mathbb{R}^{424\times 80}$ is fit once on the training set (93.7% of third-order variance) and frozen. Pure co-skewness MSE (λ=0, `coskew_mse`) fails to train meaningfully (val loss reduction $< 1\%$): the vectorised κ_3 tensor has $O(0.02)$ magnitude, yielding gradients too small for either BrainLM backbone (87M and 658M parameters for the 111M and 650M checkpoints). The `fc_tucker` variant combines the rich gradient of a Riemannian SPD metric with a κ_3-informed basis, and converges cleanly (-80% val loss on both model sizes).

Table 21: Dual-moment λ sweep (AOMIC AAL-424). Pearson r (mean $\pm$ std, n=200). Within this coarse grid, `fc_tucker` at λ=0 is the winner for both model sizes; a finer sub-λ sweep (Appendix M.4) picks λ=10^{-3} as the overall optimum, which is the configuration adopted in the main text. Adding the ambient-FC term back ($\lambda > 0$) monotonically degrades reconstruction on 111M. Embeddings remain flat across all 20 configurations.

Mode	λ	111M Emb. r	111M FC-recon r	650M Emb. r	650M FC-recon r
`coskew_mse`	0.00	0.091 ± 0.118	0.032 ± 0.097	0.038 ± 0.113	0.030 ± 0.117
`coskew_mse`	0.25	0.070 ± 0.117	0.274 ± 0.088	0.034 ± 0.097	0.274 ± 0.091
`coskew_mse`	0.50	0.072 ± 0.117	0.265 ± 0.088	0.032 ± 0.099	0.270 ± 0.089
`coskew_mse`	0.75	0.079 ± 0.117	0.263 ± 0.090	0.041 ± 0.097	0.272 ± 0.090
`coskew_mse`	1.00	0.079 ± 0.118	0.267 ± 0.090	0.056 ± 0.101	0.271 ± 0.089
`fc_tucker`	**0.00**	0.082 ± 0.118	$\mathbf{0.307 \pm 0.085}$	0.034 ± 0.104	$\mathbf{0.304 \pm 0.084}$
`fc_tucker`	0.25	0.067 ± 0.117	0.303 ± 0.085	-0.029 ± 0.092	0.300 ± 0.085
`fc_tucker`	0.50	0.074 ± 0.121	0.295 ± 0.086	0.044 ± 0.094	0.287 ± 0.086
`fc_tucker`	0.75	0.079 ± 0.121	0.280 ± 0.091	0.037 ± 0.100	0.285 ± 0.087
`fc_tucker`	1.00	0.079 ± 0.118	0.267 ± 0.091	0.056 ± 0.101	0.271 ± 0.089



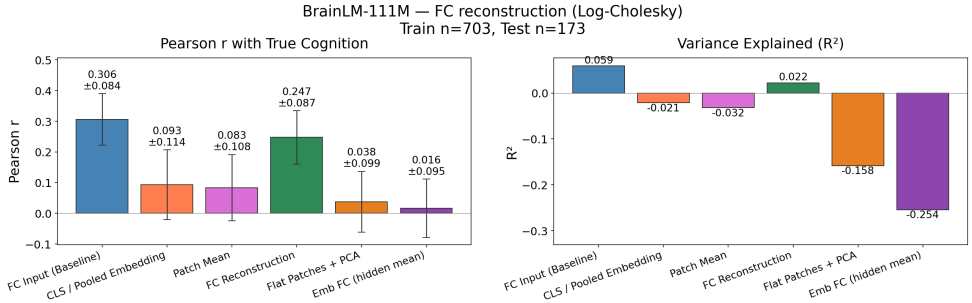
Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3926	0.0968	0.1270	0.3712	—
CLS / Pooled Embedding	768	0.0604	0.0988	-0.0232	0.0608	-0.3323
Patch Mean	768	0.0319	0.0964	-0.0482	0.0344	-0.3607
FC Reconstruction	89,676	0.3746	0.0938	0.1157	0.3589	-0.0181
Flat Patches + PCA	500	-0.0440	0.1023	-0.0799	-0.0483	-0.4366
Emb FC (hidden mean)	265,356	-0.0387	0.0901	-0.1063	-0.0295	-0.4313



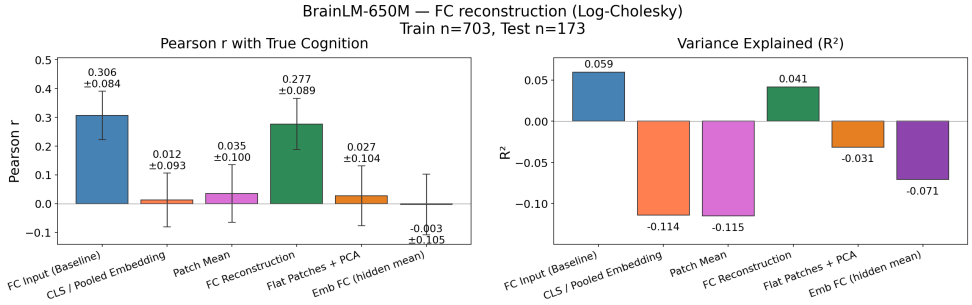
Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3926	0.0968	0.1270	0.3712	—
CLS / Pooled Embedding	1,280	-0.0328	0.1053	-0.0375	-0.0180	-0.4255
Patch Mean	1,280	-0.0296	0.1056	-0.0402	-0.0121	-0.4223
FC Reconstruction	89,676	0.3801	0.0996	0.1170	0.3644	-0.0126
Flat Patches + PCA	500	-0.0651	0.0835	-0.0778	-0.0585	-0.4578
Emb FC (hidden mean)	461,280	-0.0002	0.0868	-0.0678	0.0028	-0.3929

Figure 6: BrainLM dual-moment FT (AOMIC-trained at λ=0; coarse sweep, Appendix M.3) evaluated on HCP, 111M (top) and 650M (bottom). FC-recon on HCP reaches $r = 0.375$ and $r = 0.380$ here; the main-text reported configuration (λ=10^{-3}, Table 3) gives r=0.379/0.380.

On 111M, adding the full ambient FC term back degrades reconstruction monotonically ($0.307 \rightarrow 0.303 \rightarrow 0.295 \rightarrow 0.280 \rightarrow 0.267$), suggesting that ambient Log-Cholesky injects noise that the κ_3-subspace filter removes. The λ=1 endpoint is the same theoretical objective as the standalone Log-Cholesky FC baseline (r=0.247, Table 19); the small offset reflects independent runs with different seeds and early-stopping epochs. Embeddings remain flat across all 20 cells, confirming that the κ_3-informed reconstruction target, like every other FC-preservation target, does not restructure the latent space.



Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3061	0.0840	0.0594	0.3115	—
CLS / Pooled Embedding	768	0.0927	0.1142	-0.0213	0.0896	-0.2133
Patch Mean	768	0.0831	0.1081	-0.0317	0.0672	-0.2230
FC Reconstruction	89,676	0.2473	0.0872	0.0222	0.2380	-0.0588
Flat Patches + PCA	500	0.0376	0.0992	-0.1582	0.0475	-0.2685
Emb FC (hidden mean)	265,356	0.0164	0.0950	-0.2541	0.0287	-0.2897

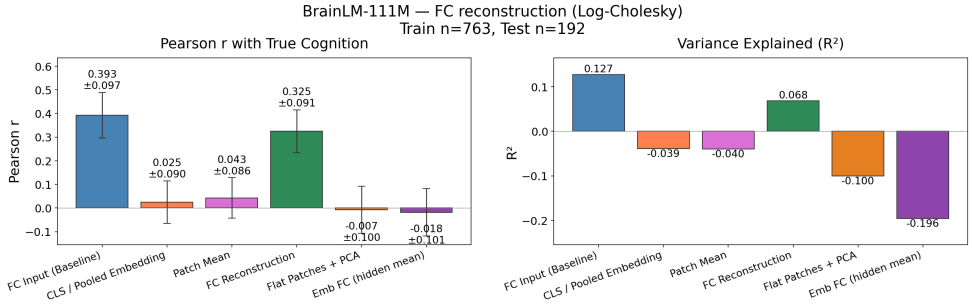


Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3061	0.0840	0.0594	0.3115	—
CLS / Pooled Embedding	1,280	0.0123	0.0934	-0.1139	0.0167	-0.2938
Patch Mean	1,280	0.0353	0.0996	-0.1149	0.0348	-0.2708
FC Reconstruction	89,676	0.2766	0.0894	0.0414	0.2726	-0.0295
Flat Patches + PCA	500	0.0272	0.1037	-0.0314	0.0336	-0.2789
Emb FC (hidden mean)	461,280	-0.0030	0.1053	-0.0707	0.0154	-0.3091

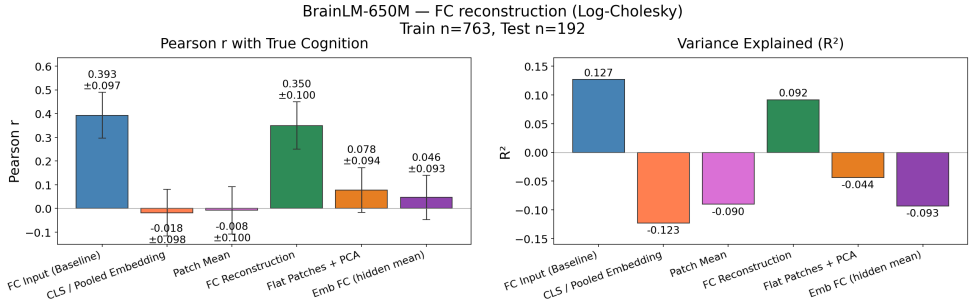
Figure 7: Readout comparison after finetuning on AOMIC: BrainLM-111M (top) and BrainLM-650M (bottom). FC-reconstruction (the decoder output reprocessed into FC) rises close to the raw FC baseline under the dual-moment loss, while the CLS embedding and patch readouts stay near zero.

M.4 Sub-λ sweep and long convergence confirm the AOMIC ceiling

The $\lambda \in \{0, 0.25, 0.5, 0.75, 1.0\}$ sweep (Appendix M.3) peaks at `fc_tucker` λ=0 with FC-recon $r = 0.307$, tying the raw FC baseline ($r = 0.306$). Two potential artefacts could inflate or deflate



Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3926	0.0968	0.1270	0.3712	—
CLS / Pooled Embedding	768	0.0248	0.0899	-0.0388	0.0340	-0.3679
Patch Mean	768	0.0427	0.0860	-0.0399	0.0467	-0.3500
FC Reconstruction	89,676	0.3245	0.0910	0.0684	0.3057	-0.0681
Flat Patches + PCA	500	-0.0074	0.1003	-0.1003	0.0012	-0.4000
Emb FC (hidden mean)	265,356	-0.0183	0.1008	-0.1956	-0.0117	-0.4109



Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3926	0.0968	0.1270	0.3712	—
CLS / Pooled Embedding	1,280	-0.0179	0.0980	-0.1229	-0.0114	-0.4105
Patch Mean	1,280	-0.0081	0.0999	-0.0903	-0.0137	-0.4007
FC Reconstruction	89,676	0.3500	0.0996	0.0917	0.3351	-0.0426
Flat Patches + PCA	500	0.0777	0.0940	-0.0436	0.0721	-0.3149
Emb FC (hidden mean)	461,280	0.0463	0.0929	-0.0933	0.0412	-0.3463

Figure 8: Readout comparison after finetuning on HCP: BrainLM-111M (top) and BrainLM-650M (bottom). Same conventions as Figure 7.

that number: (i) early termination (losses were still decreasing at the 50-epoch budget); (ii) the λ grid being too coarse, hiding a near-optimum at a small non-zero value. We address both, and find a modest sub-λ optimum at $\lambda = 10^{-3}$ that we adopt as the reported configuration in the main text.

Track B, long convergence. The `fc_tucker` λ=0 BrainLM-111M run was re-executed with a $5\times$ longer schedule (250 epochs, patience 40). The final checkpoint gives FC-recon $r = 0.307 \pm 0.088$, indistinguishable from the 50-epoch result.

Track A, sub-λ sweep. Four additional `fc_tucker` runs at $\lambda \in \{10^{-4}, 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}\}$ and two additional `coskew_mse` runs at $\lambda \in \{10^{-4}, 10^{-3}\}$.

Table 22: Tracks A and B on BrainLM-111M. The $r \approx 0.307$ AOMIC plateau is not a schedule artefact ($5\times$ training reproduces it). The λ landscape is flat near zero, and any non-trivial $\lambda \gtrsim 5 \cdot 10^{-3}$ hurts. `fc_tucker` beats `coskew_mse` at matched λ in every comparison.

Track	Mode	λ	Sched.	AOMIC FC-recon r	HCP FC-recon r (transfer)
C.3 (prior)	`fc_tucker`	0.00	50 ep	0.307 ± 0.085	0.375 ± 0.094
B	**`fc_tucker`**	**0.00**	**250** ep	$\mathbf{0.307 \pm 0.088}$	$\mathbf{0.378 \pm 0.096}$
A	`fc_tucker`	10^{-4}	50 ep	0.308 ± 0.088	0.379 ± 0.096
A	`fc_tucker`	10^{-3}	50 ep	0.308 ± 0.085	0.379 ± 0.096
A	`fc_tucker`	$5 \cdot 10^{-3}$	50 ep	0.300 ± 0.085	0.369 ± 0.092
A	`fc_tucker`	10^{-2}	50 ep	0.289 ± 0.084	0.356 ± 0.090
A	`coskew_mse`	10^{-4}	50 ep	0.298 ± 0.087	0.367 ± 0.094
A	`coskew_mse`	10^{-3}	50 ep	0.294 ± 0.086	0.363 ± 0.092
—	*Raw FC baseline*	—	—	0.306 ± 0.084	0.393 ± 0.097

Three conclusions. (1) The AOMIC plateau is not a schedule artefact: $5\times$ more training reproduces $r = 0.307$ to three decimals. (2) The λ landscape is flat near zero; sub-λ perturbations $\lambda \in \{10^{-4}, 10^{-3}\}$ both give $r = 0.308$ AOMIC and $r = 0.379$ HCP-transfer, a modest improvement over $\lambda = 0$ ($+0.001$ and $+0.004$ respectively), while any ambient-FC term beyond $\lambda \gtrsim 5 \cdot 10^{-3}$ hurts. We adopt $\lambda = 10^{-3}$ as the reported main-text configuration. (3) `fc_tucker` beats `coskew_mse` at matched λ in every comparison (λ=10^{-4}: $+0.010$; λ=10^{-3}: $+0.014$), confirming that the Riemannian SPD gradient on the κ_3-subspace FC is a strictly better training signal than MSE on raw κ_3 tensor entries. HCP cross-dataset transfer follows the AOMIC ordering exactly but sits at $r \approx 0.36$–0.38, below the HCP raw ceiling. Breaking that ceiling requires training directly at HCP scale (Appendix M.6).

M.5 Brain-JEPA and BrainMass: the KRR-alignment caveat

To check whether the BFM embedding problem is BrainLM-specific, we finetuned two architectures with distinct training paradigms.

The KRR-alignment finetuning is fitted on the canonical training split and scored by the standard 200-fold CV over the full cohort, so the test folds contain subjects whose labels the finetuner saw. The objective and the evaluation metric are the same object applied to the same target on overlapping subjects, so these values measure optimisation rather than generalisation. Every main-text finetuning objective is label-free (Section 3.7); the leak-free per-fold comparison is Arm B of Appendix C.

Table 23: Brain-JEPA and BrainMass under KRR-alignment finetuning. The numbers look large, but the objective and the evaluation metric are the same object applied to the same target, so these are training-set optimisation results and do not imply the embedding encodes transferable cognitive structure.

Model	Strategy	Emb. r (CV)
Brain-JEPA (base)	Pretrained	0.069 ± 0.091
Brain-JEPA (base)	KRR alignment FT	$\mathbf{0.301 \pm 0.094}$
Brain-JEPA (base)	Cognition similarity FT	0.101 ± 0.094
BrainMass	Pretrained	0.275 ± 0.093
BrainMass	KRR alignment FT	$\mathbf{0.324 \pm 0.095}$

Brain-JEPA's KRR-alignment FT lifts the pooled-patch representation from 0.069 to 0.301 and BrainMass from 0.275 to 0.324. These numbers look impressive, but we do not count them as evidence that the resulting embeddings are good representations of the composite. The KRR-alignment objective is a differentiable meta-KRR that, within each batch, fits a kernel ridge regressor on the current embeddings and backpropagates through the prediction correlation of held-out rows. It directly maximises, on training folds, the exact evaluation statistic we report (KRR Pearson r). The training loss and the evaluation metric are the same object, applied to the same target, on the same subjects. A high KRR-alignment r therefore states only that the objective was successfully optimised on the training target. It does not imply the embedding has learned anything transferable about the composite. The same embedding would almost certainly fail on any held-out behavioural variable not used in training (age, sex, personality factors, task performance) precisely because the finetune is shaped around a single scalar. This is the well-known metric-training tautology.

We include these numbers because they are informative about architecture: Brain-JEPA and BrainMass can be pushed to $r \approx 0.30$ under KRR-alignment while BrainLM cannot (its CLS stays at $r = 0.116$ under the same loss). JEPA's predictive objective and BrainMass's Transformer-on-FC encoder leave more latent-space plasticity than BrainLM's reconstruction-only CLS token. The absolute r values should not be read as generalisable representation gains. Even at face value, the KRR-aligned Brain-JEPA embedding ($r = 0.301$) sits at the level of its own input FC ($r = 0.313$) and well below Tucker at either parcellation ($r = 0.368$–0.406). The label-free FC-preservation finetunings (Log-Cholesky, dual-moment), which do not train on the evaluation target and therefore do test whether the representation changed, leave embeddings unchanged in every model we tested.

M.6 Track C: HCP-scale training closes the BFM-to-raw gap

The AOMIC-trained dual-moment checkpoint transfers to HCP at $r \approx 0.38$, below the raw HCP FC baseline ($r = 0.393$). To test whether this gap reflects a data-scale limit rather than an objective limitation, we trained BrainLM-111M and BrainLM-650M with the same `fc_tucker` λ=10^{-3}, R=80 configuration directly on HCP data, segmenting each of the 763 training subjects into up to 24 non-overlapping 200-TP windows (a few subjects have shorter concatenated runs and contribute fewer segments) for 18,298 training segments in total ($\sim 20\times$ the AOMIC corpus).

Training outcome. 50 epochs completed for the 111M run; train loss decreased $59\times$ ($3.524 \rightarrow 0.060$, -98.3%) and val loss $42\times$ ($2.336 \rightarrow 0.056$, -97.6%). Both losses were still monotonically decreasing at epoch 50 (no plateau), indicating the model has not converged. The 650M run uses the same configuration.

Table 24: BrainLM dual-moment FT trained directly on HCP (18,298 segments) at both scales. FC-reconstruction matches the raw FC baseline on HCP in-domain and on AOMIC cross-dataset, at both 111M and 650M. The 650M-below-111M ordering of the pretrained checkpoints is absent: 111M and 650M reach the same in-domain HCP r to three decimals and the same AOMIC transfer r to two decimals.

Model	Eval dataset	*fc_input* (raw) r	FC-recon r	Δ
BrainLM-111M	HCP in-domain (AAL-424)	0.393 ± 0.097	$\mathbf{0.392 \pm 0.097}$	-0.001
	AOMIC cross-dataset (AAL-424)	0.306 ± 0.084	$\mathbf{0.306 \pm 0.084}$	0.000
BrainLM-650M	HCP in-domain (AAL-424)	0.393 ± 0.097	$\mathbf{0.392 \pm 0.097}$	-0.001
	AOMIC cross-dataset (AAL-424)	0.306 ± 0.084	$\mathbf{0.305 \pm 0.084}$	-0.001

The BFM-to-raw gap is closed at both scales. Two factors drive the improvement over the AOMIC-trained checkpoints: (i) in-domain HCP training eliminates the cross-dataset transfer gap; (ii) the $20\times$ larger segment pool provides richer gradient signal for the dual-moment loss. The 650M run reaches the same in-domain HCP r as the 111M run and shows the same near-zero cross-dataset gap, indicating that the pretrained-model ordering (650M $<$ 111M) does not survive cumulant-aligned finetuning. Since the model has not converged, a longer schedule may push reconstruction FC above raw FC, particularly if Tucker decomposition is applied to the reconstruction (raw HCP AAL-424 FC-Tucker at R=167 reaches $r = 0.431$, Table 2).

N GRANULARITY ANALYSIS: STATISTICAL REDUCTION VS. ATLAS RESOLUTION

An alternative to the Tucker story is that a finer parcellation alone would suffice: the Tucker advantage over FC-full might simply reflect better spatial resolution. This appendix rules that out by applying each Schaefer resolution (100, 200, 300, 400) *independently* to the raw HCP CIFTI timeseries, rather than aggregating Schaefer-400 into coarser groupings, and comparing to Tucker and PCA computed on Schaefer-400 timeseries at matched feature counts.

Table 25: FC-full at each native Schaefer resolution on HCP. Each doubling of ROIs yields diminishing returns; the total gap across a $4\times$ increase in ROIs is $+0.038$ in Pearson r.

Resolution	Features	Pearson $r\pm$ std
Schaefer-100	4,950	0.504 ± 0.088
Schaefer-200	19,900	0.519 ± 0.083
Schaefer-300	44,850	0.527 ± 0.085
Schaefer-400	79,800	0.542 ± 0.084

Table 26: At exactly the same feature count, Tucker on Schaefer-400 beats FC-full at the coarser native parcellation at every resolution (HCP, 200 CV folds).

N	Schaefer-N FC r	Tucker($R \approx N$) on Schaefer-400 r	Δr
100	0.504 ± 0.088	$\mathbf{0.561 \pm 0.083}$ (R=100)	$+0.057$ ($+11\%$)
200	0.519 ± 0.083	$\mathbf{0.560 \pm 0.085}$ (R=203)	$+0.041$ ($+8\%$)
300	0.527 ± 0.085	$\mathbf{0.563 \pm 0.089}$ (R=298)	$+0.036$ ($+7\%$)

Tucker at only R=80 (3,160 features) already surpasses Schaefer-400 FC at 79,800 features ($d = 0.40$; reported in Section 4.4). A $25\times$ compression in a third-order-informed basis beats a raw $25\times$ expansion in atlas resolution. The Tucker advantage is about statistical structure, not spatial resolution.

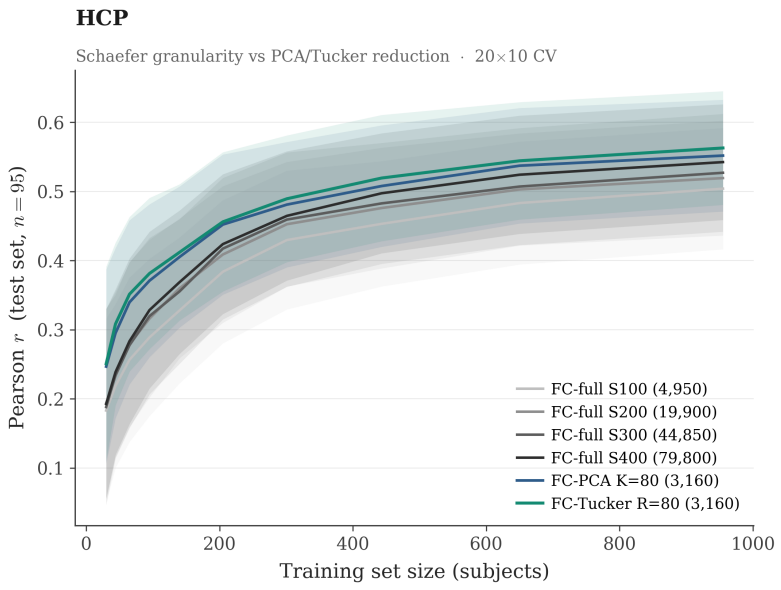


Figure 9: HCP Schaefer granularity comparison. FC-full at each native Schaefer resolution (100–400) compared to Tucker and PCA computed on Schaefer-400 timeseries at matched feature counts. Tucker is above FC-full at the native resolution and above PCA at each of the four feature budgets.

O LEARNING CURVES

This appendix reports learning curves across all four dataset×parcellation combinations, computed at the CV-optimal K/R values: AOMIC AAL-424 (K=154, R=117), AOMIC Schaefer-400 (K=57, R=122), HCP AAL-424 (K=417, R=167), HCP Schaefer-400 (K=87, R=333). For each training size N, the training fold is subsampled to N subjects, the decomposition basis is refit on the subsample, and the test fold is evaluated, giving 200 CV evaluations per point.

Three qualitative findings. First, Tucker reaches competitive performance at smaller training sizes than PCA or FC-full on AOMIC, and asymptotes at a higher plateau, consistent with better sample efficiency from the co-skewness basis. Second, on HCP the three methods converge at large N because long scans ($T = 4{,}800$) stabilise even the less-robust PCA eigenvectors. Third, the variance bands computed from the 200 fold-level r values confirm that the Tucker advantage on AOMIC is robust across folds and not an artefact of any single favourable split.

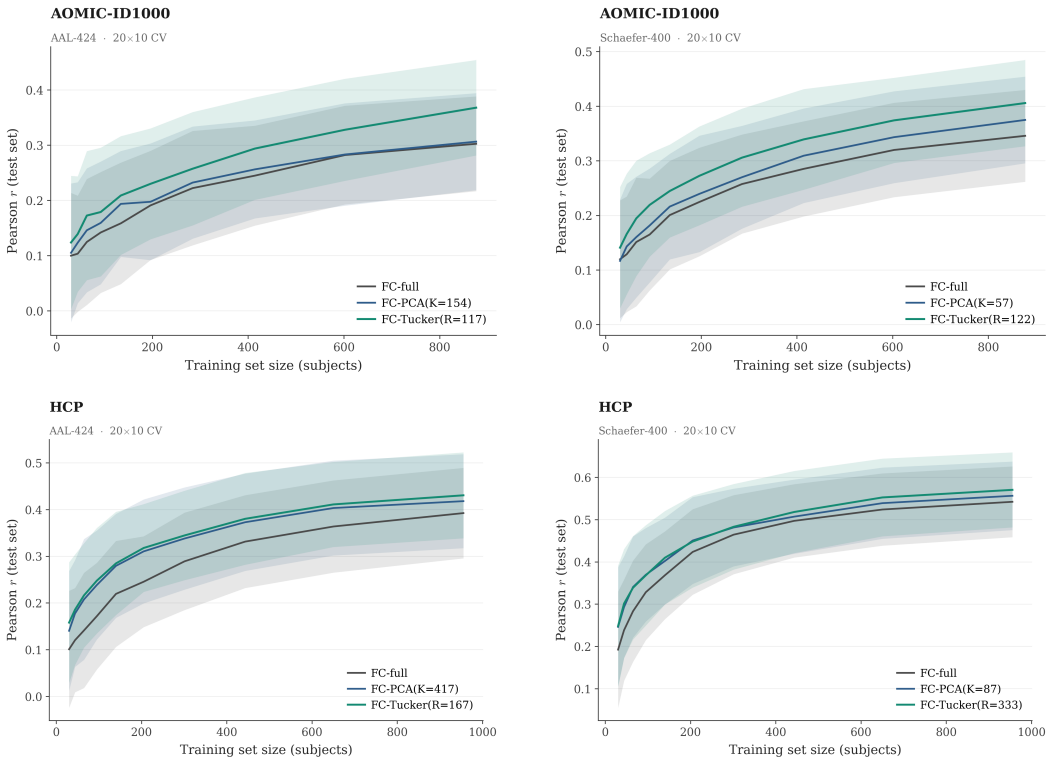


Figure 10: Learning curves across all four dataset×parcellation combinations. Top: AOMIC AAL-424 (left), AOMIC Schaefer-400 (right). Bottom: HCP AAL-424 (left), HCP Schaefer-400 (right). Shaded bands are ±1 std across the 200 CV folds.

P NESTED CROSS-VALIDATION FOR LEAK-FREE R^* SELECTION

Table 2 reports R^* and K^* values chosen from a dense sweep over $R, K \in [2, P]$ at 200 CV folds per point. A natural concern is that R^* was chosen after observing all sweep results, and a less favourable R would have yielded weaker Tucker performance. This appendix addresses the concern with fully nested cross-validation, where R selection is embedded within the evaluation loop and no information from the test fold enters R^* selection.

Protocol. For each of the 200 outer folds (20 reps × 10 folds), an inner 5-fold CV loop evaluates every R in the grid $\{40, 60, 80, 100, 120, 140, 167, 200, 250, 333\}$ on the training set only. The R^* that maximises mean inner-CV Pearson r is selected, and the model is evaluated at that R^* on the

held-out outer test fold. The same procedure is applied to PCA (K^* selection). FC-full is included as a fixed reference (no hyperparameter).

Table 27: Nested-CV results. Pearson r (mean $\pm$ std, $n = 200$ fold-level evaluations). R^* and K^* are selected per fold by inner 5-fold CV on the training set only. Tucker with nested R^* selection outperforms FC-full on all four combinations, confirming that the main-text Tucker advantage is not an artefact of post-hoc R^* selection.

Dataset / parcellation	FC-full	FC-Tucker (nested R^*)	FC-PCA (nested K^*)
AOMIC AAL-424	0.303 ± 0.086	$\mathbf{0.346 \pm 0.093}$	0.268 ± 0.087
AOMIC Schaefer-400	0.346 ± 0.084	$\mathbf{0.392 \pm 0.078}$	0.343 ± 0.083
HCP AAL-424	0.393 ± 0.097	$\mathbf{0.416 \pm 0.095}$	0.406 ± 0.100
HCP Schaefer-400	0.542 ± 0.084	$\mathbf{0.558 \pm 0.084}$	0.540 ± 0.081

The nested-CV Tucker results are slightly lower than the sweep-optimal results in Table 2 (e.g., AOMIC AAL-424: 0.346 nested vs. 0.368 sweep), as expected. The inner CV occasionally selects a suboptimal R for a given fold, and the 10-value grid is coarser than the dense sweep. The key result is that the Tucker advantage over FC-full is preserved in every cell, and the advantage over nested PCA is consistent and substantial.

Sensitivity to R^*. Tucker's nested-CV penalty relative to the sweep optimum is modest: -6% on AOMIC AAL-424, -4% on AOMIC Schaefer-400, -4% on HCP AAL-424, -2% on HCP Schaefer-400. This matches the broad performance plateau visible in the dense sweep curves (Figure 11). Tucker is not sensitive to the exact choice of R within a wide range.

R^* selection distributions. The distribution of R^* values selected across the 200 outer folds shows the structure of the problem.

Table 28: Most-selected Tucker R^* values across the 200 outer folds (inner-CV selection). On AOMIC, selections concentrate on moderate R (80–140), consistent with the broad plateau in the dense sweeps. On HCP, selections favour higher R, especially R=333 on HCP Schaefer-400 ($150/200$ folds), consistent with long scans (T=4,800) supporting reliable estimation of higher-dimensional structure.

Dataset / parcellation	Most-selected Tucker R^* (fold counts)
AOMIC AAL-424	R=120 (93), R=60 (37), R=100 (30), R=140 (31)
AOMIC Schaefer-400	R=120 (55), R=100 (47), R=140 (41), R=80 (28)
HCP AAL-424	R=167 (83), R=140 (38), R=333 (29), R=200 (28)
HCP Schaefer-400	R=333 (150), R=200 (19), R=167 (12), R=250 (10)

The spread across multiple R^* values confirms that the Tucker advantage is robust. Even folds that select R=60 or R=80 contribute positively to the aggregate result.

Q Full BFM evaluation plots

This appendix collects per-model, per-dataset readout plots for all four BFMs in their pretrained (frozen) state, plus the Ooi-style FC baseline and the reconstruction-quality comparison. Each bar shows mean Pearson r across the 200 CV folds for one readout strategy (CLS or pooled embedding, patch mean, FC from reconstruction, flattened patches, embedding similarity). The aggregated version of this material is Figure 4 and Table 1; the plots below show the full per-readout breakdown.

R Cumulant preservation metric definitions

This appendix expands the per-order relative errors used in Section 3.6 and reported in Table 29. Let $\mathbf{Z} \in \mathbb{R}^{P \times T}$ be the input ROI timeseries and $\hat{\mathbf{Z}} \in \mathbb{R}^{P \times T}$ the BFM reconstruction of the same subject. Let $\mathbf{C}_{\text{in}}, \mathbf{C}_{\text{re}} \in \mathbb{R}^{P \times P}$ denote the corresponding FC matrices (Pearson). Let $\mathbf{U} \in \mathbb{R}^{P \times R}$

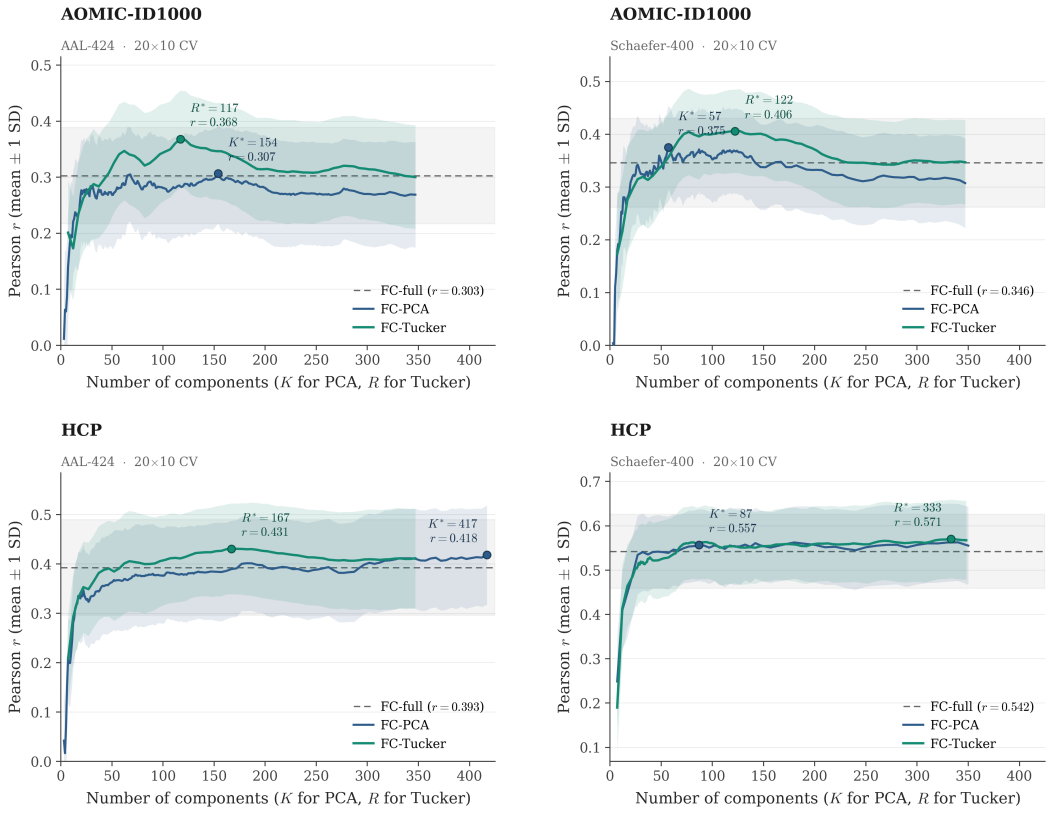
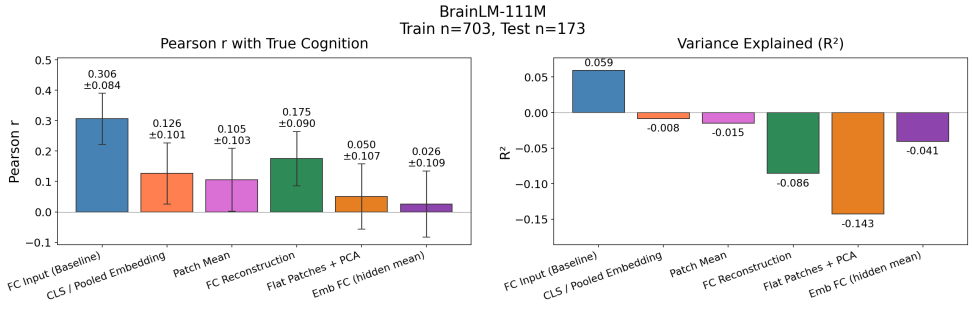
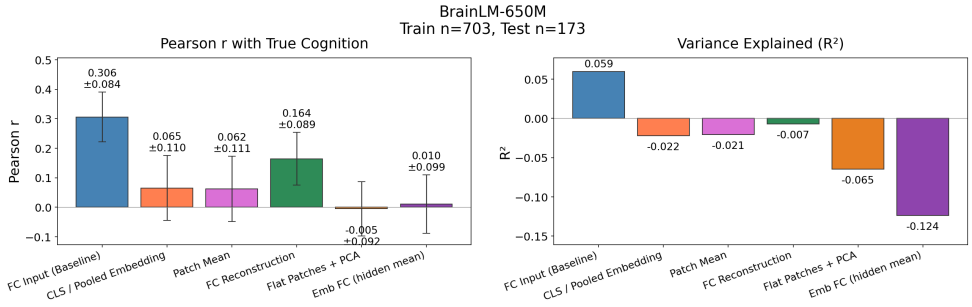


Figure 11: Dense-sweep curves for all four dataset×parcellation cells: AOMIC AAL-424 (top left), AOMIC Schaefer-400 (top right), HCP AAL-424 (bottom left), HCP Schaefer-400 (bottom right). FC-full shown as a horizontal dashed line. FC-Tucker exceeds FC-full across a broad plateau ($R \gtrsim 80$ up to full rank), not just at the sweep-optimal R^*; FC-PCA is at or below FC-full over the corresponding range. On AOMIC AAL-424 Tucker peaks at $R = 117$ with $r = 0.368$ ($+21\%$ over FC-full, $d = 0.90$).

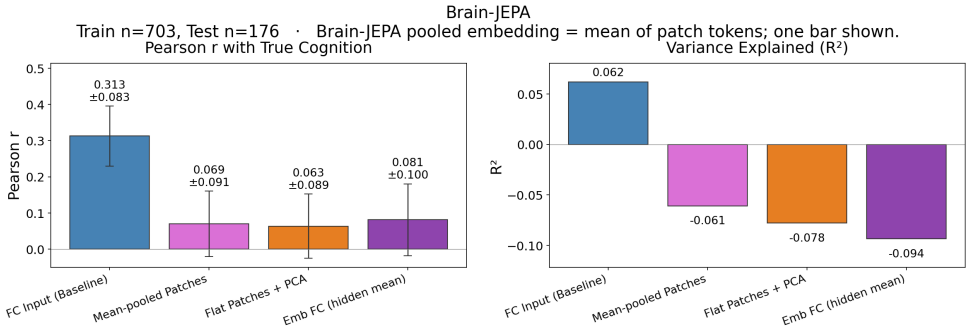


Method	Feature Dim	Pearson r	± std	R^2	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3061	0.0840	0.0594	0.3115	—
CLS / Pooled Embedding	768	0.1263	0.1009	-0.0085	0.1237	-0.1797
Patch Mean	768	0.1054	0.1034	-0.0148	0.1130	-0.2006
FC Reconstruction	89,676	0.1750	0.0897	-0.0856	0.1730	-0.1311
Flat Patches + PCA	500	0.0505	0.1070	-0.1429	0.0377	-0.2556
Emb FC (hidden mean)	265,356	0.0260	0.1087	-0.0405	0.0227	-0.2801

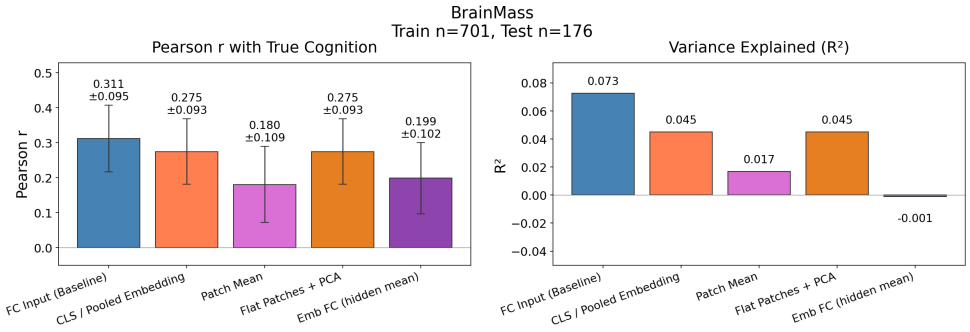


Method	Feature Dim	Pearson r	± std	R^2	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3061	0.0840	0.0594	0.3115	—
CLS / Pooled Embedding	1,280	0.0649	0.1103	-0.0223	0.0653	-0.2411
Patch Mean	1,280	0.0619	0.1109	-0.0209	0.0572	-0.2442
FC Reconstruction	89,676	0.1644	0.0891	-0.0074	0.1660	-0.1417
Flat Patches + PCA	500	-0.0054	0.0921	-0.0649	-0.0037	-0.3115
Emb FC (hidden mean)	461,280	0.0103	0.0990	-0.1237	0.0038	-0.2958

Figure 12: Pretrained BrainLM readouts on AOMIC (AAL-424): 111M (top) and 650M (bottom). All bars are KRR Pearson r over 200 CV folds.

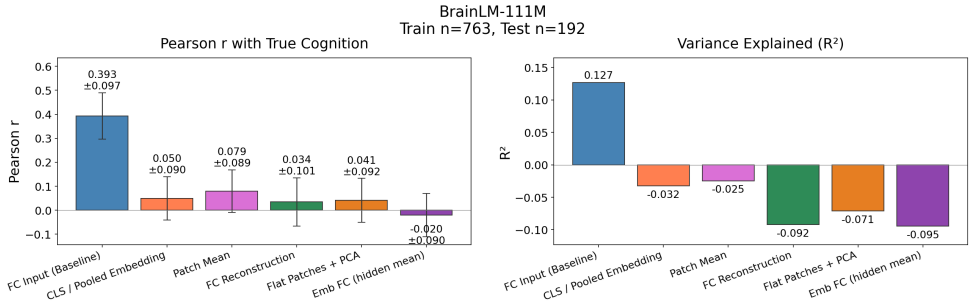


Method	Feature Dim	Pearson r	± std	R^2	Spearman	Δr vs Baseline
FC Input (Baseline)	101,025	0.3128	0.0830	0.0621	0.3111	—
Mean-pooled Patches	768	0.0695	0.0906	-0.0610	0.0760	-0.2433
Flat Patches + PCA	3,456,000	0.0635	0.0891	-0.0779	0.0657	-0.2493
Emb FC (hidden mean)	101,025	0.0810	0.0996	-0.0936	0.0882	-0.2318

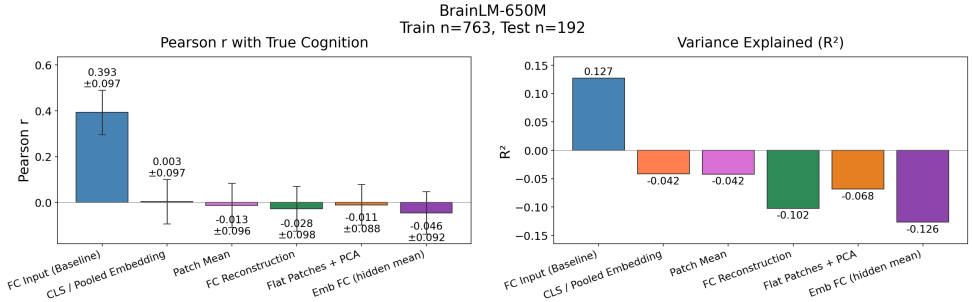


Method	Feature Dim	Pearson r	± std	R^2	Spearman	Δr vs Baseline
FC Input (Baseline)	4,950	0.3115	0.0953	0.0726	0.2990	—
CLS / Pooled Embedding	800	0.2747	0.0931	0.0451	0.2601	-0.0368
Patch Mean	8	0.1805	0.1092	0.0169	0.1450	-0.1310
Flat Patches + PCA	500	0.2747	0.0931	0.0451	0.2601	-0.0368
Emb FC (hidden mean)	4,950	0.1985	0.1017	-0.0011	0.1894	-0.1130

Figure 13: Pretrained Brain-JEPA (top, Schaefer-400+Tian-50) and BrainMass (bottom, Schaefer-100 FC) readouts on AOMIC. Same conventions as Figure 12.

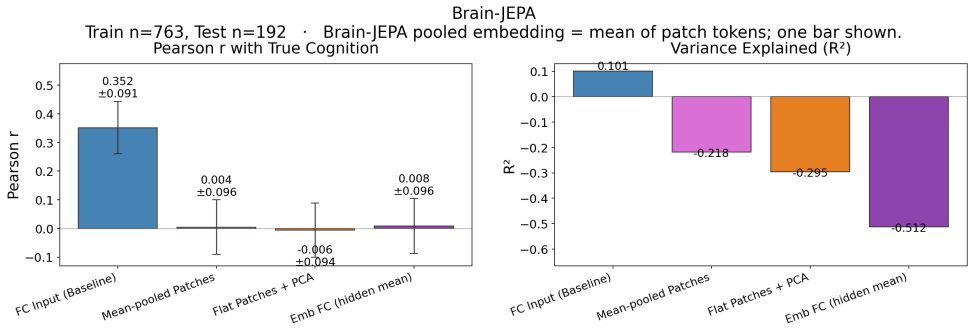


Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3926	0.0968	0.1270	0.3712	—
CLS / Pooled Embedding	768	0.0496	0.0903	-0.0325	0.0572	-0.3431
Patch Mean	768	0.0793	0.0890	-0.0248	0.0817	-0.3134
FC Reconstruction	89,676	0.0344	0.1007	-0.0924	0.0310	-0.3582
Flat Patches + PCA	500	0.0408	0.0917	-0.0714	0.0353	-0.3519
Emb FC (hidden mean)	265,356	-0.0203	0.0902	-0.0946	-0.0111	-0.4130

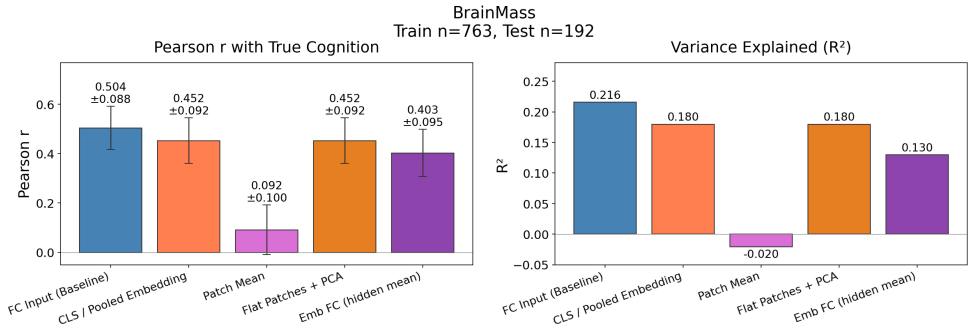


Method	Feature Dim	Pearson r	± std	R²	Spearman	Δr vs Baseline
FC Input (Baseline)	89,676	0.3926	0.0968	0.1270	0.3712	—
CLS / Pooled Embedding	1,280	0.0028	0.0969	-0.0417	0.0083	-0.3899
Patch Mean	1,280	-0.0132	0.0959	-0.0419	-0.0037	-0.4058
FC Reconstruction	89,676	-0.0276	0.0978	-0.1021	-0.0199	-0.4202
Flat Patches + PCA	500	-0.0108	0.0882	-0.0682	-0.0141	-0.4034
Emb FC (hidden mean)	461,280	-0.0461	0.0922	-0.1265	-0.0555	-0.4388

Figure 14: Pretrained BrainLM readouts on HCP (AAL-424): 111M (top) and 650M (bottom).



Method	Feature Dim	Pearson r	± std	R^2	Spearman	Δr vs Baseline
FC Input (Baseline)	101,025	0.3518	0.0912	0.1006	0.3376	—
Mean-pooled Patches	768	0.0043	0.0956	-0.2183	0.0002	-0.3475
Flat Patches + PCA	3,456,000	-0.0058	0.0944	-0.2954	-0.0134	-0.3576
Emb FC (hidden mean)	101,025	0.0085	0.0957	-0.5121	0.0036	-0.3434



Method	Feature Dim	Pearson r	± std	R^2	Spearman	Δr vs Baseline
FC Input (Baseline)	4,950	0.5040	0.0875	0.2160	0.4699	—
CLS / Pooled Embedding	800	0.4520	0.0923	0.1799	0.4293	-0.0520
Patch Mean	8	0.0916	0.1005	-0.0204	0.0876	-0.4124
Flat Patches + PCA	500	0.4520	0.0923	0.1799	0.4293	-0.0520
Emb FC (hidden mean)	4,950	0.4026	0.0954	0.1298	0.3754	-0.1014

Figure 15: Pretrained Brain-JEPA (top) and BrainMass (bottom) readouts on HCP. Same atlas-per-model convention as Figure 12.

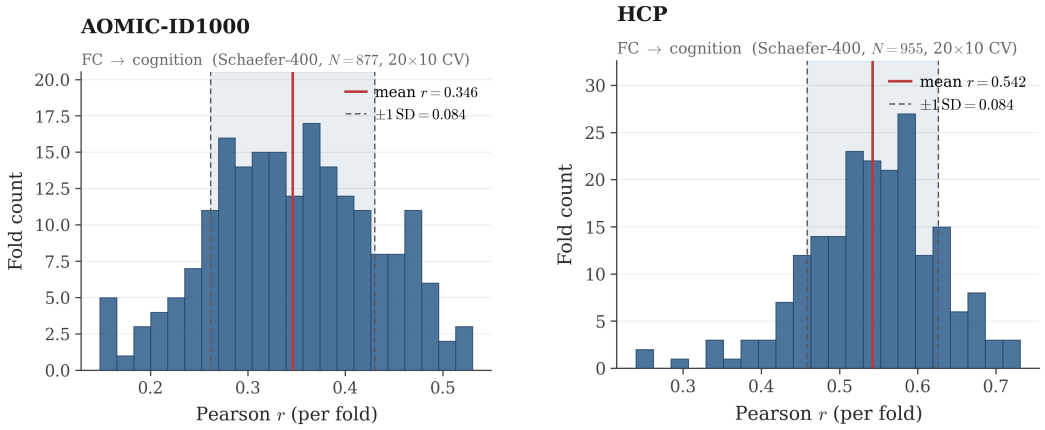


Figure 16: Ooi-style raw FC baseline plots on AOMIC (left) and HCP (right). Used throughout the paper as the parcellation-matched input FC reference.

be the rank-R Tucker factor matrix of the training co-skewness tensor (R=80 throughout the main text), fit on the training subjects only and held fixed. All four metrics are computed per subject and aggregated as mean $\pm$ std.

κ_2 metrics. We report both a Riemannian (Log-Cholesky) distance and a Frobenius (NMSE) error on the FC matrices,

$$\kappa_2 \text{ LC rel.} = \frac{d_{\text{LC}}(\mathbf{C}_{\text{in}}, \mathbf{C}_{\text{re}})}{d_{\text{LC}}(\mathbf{C}_{\text{in}}, \mathbf{I})}, \tag{5}$$

$$\kappa_2 \text{ MSE NMSE} = \frac{\left\| \operatorname{triu}(\mathbf{C}_{\text{in}}) - \operatorname{triu}(\mathbf{C}_{\text{re}}) \right\|_2^2}{\left\| \operatorname{triu}(\mathbf{C}_{\text{in}}) \right\|_2^2}, \tag{6}$$

where d_{LC} is the Log-Cholesky metric of Section 3.7 and triu extracts the strictly upper-triangular entries ($P(P-1)/2$ unique correlations). The LC denominator is the Log-Cholesky norm of the input FC, i.e. the distance from $\mathbf{I}$; the NMSE denominator is the signal energy of the input FC, so dividing by the same-shape quantity makes the two denominators commensurate across subjects.

κ_3 metrics. Raw $O(P^3)$ co-skewness tensors are prohibitive to compare entry-by-entry for $P = 424$; we therefore evaluate κ_3 preservation inside the fixed Tucker subspace $\mathbf{U}$. Define the Tucker-projected and z-scored timeseries

$$\tilde{\mathbf{Z}} = \operatorname{zscore}(\mathbf{U}^\top \mathbf{Z}) \in \mathbb{R}^{R \times T}, \qquad \hat{\tilde{\mathbf{Z}}} = \operatorname{zscore}(\mathbf{U}^\top \hat{\mathbf{Z}}), \tag{7}$$

and the corresponding co-skewness vectors $\mathbf{s}_{\text{in}}, \mathbf{s}_{\text{re}} \in \mathbb{R}^{N_3}$ with

$$s_{(i,j,k)} = \frac{1}{T} \sum_{t=1}^{T} \tilde{z}_i(t)\, \tilde{z}_j(t)\, \tilde{z}_k(t), \qquad 1 \le i \le j \le k \le R, \tag{8}$$

where the index set ranges over unique unordered triples of size $N_3 = \binom{R+2}{3} = 88{,}560$ for $R = 80$. We use equal weighting on the unique-triple vector; the alternative weighting by tensor multiplicities $\{1, 3, 6\}$ (auto, two-identical, all-distinct triples) gives the Frobenius-norm NMSE on the full R^3 tensor, which up-weights all-distinct triples and would make all reported κ_3 NMSEs larger, but does not change the qualitative cross-cell ordering. The two κ_3 metrics are

$$\kappa_3 \text{ MSE NMSE} = \frac{\|\mathbf{s}_{\text{in}} - \mathbf{s}_{\text{re}}\|_2^2}{\|\mathbf{s}_{\text{in}}\|_2^2}, \tag{9}$$

$$\kappa_3 \text{ LC rel.} = \frac{d_{\text{LC}}\big(\text{FC}(\tilde{\mathbf{Z}}), \text{FC}(\hat{\tilde{\mathbf{Z}}})\big)}{d_{\text{LC}}\big(\text{FC}(\tilde{\mathbf{Z}}), \mathbf{I}\big)}, \tag{10}$$

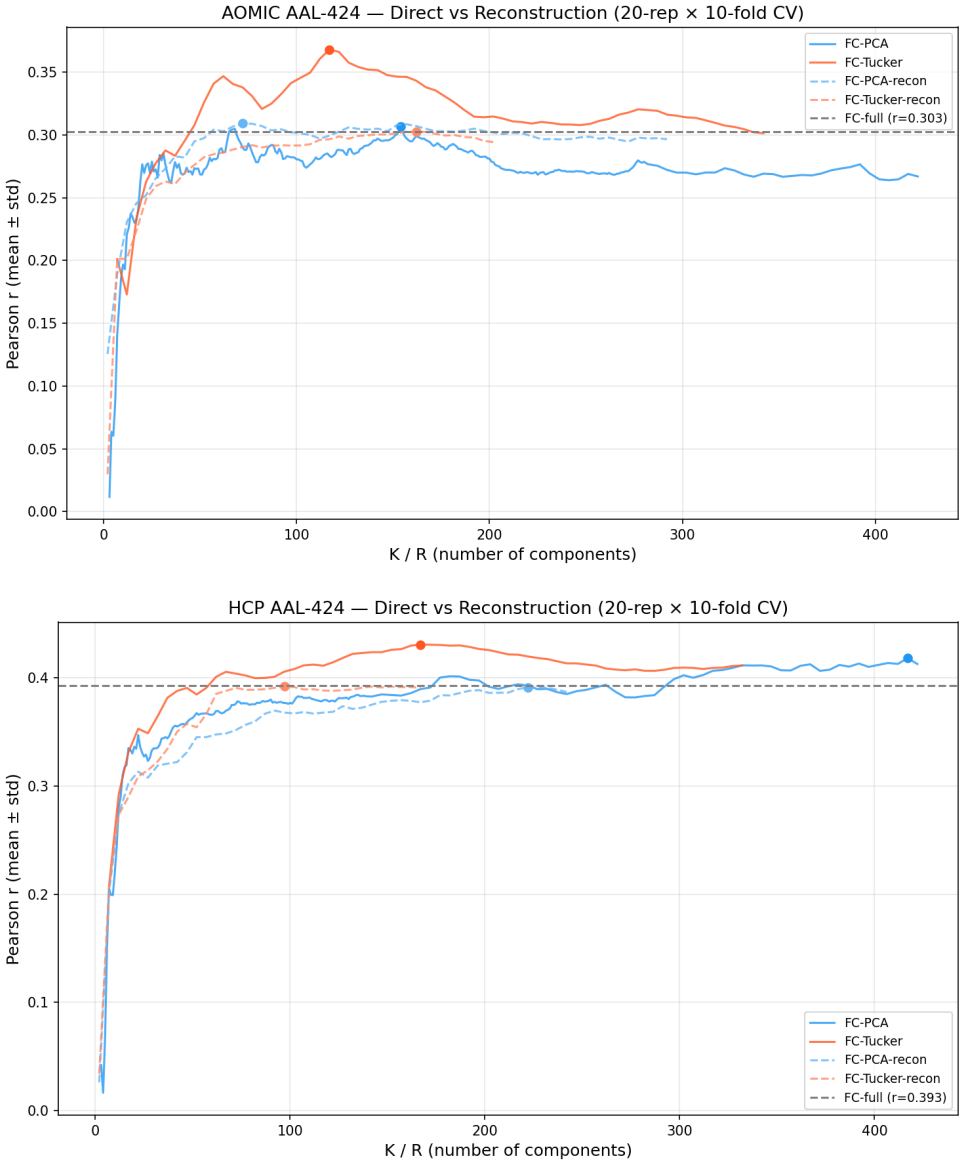


Figure 17: Reconstruction-quality comparison. FC computed from BFM-reconstructed timeseries versus raw FC, showing the degradation introduced by the reconstruction bottleneck. Top: AOMIC. Bottom: HCP. This is the visual counterpart of the per-order cumulant preservation analysis in Table 29.

where $\mathrm{FC}(\cdot)$ is the Pearson correlation matrix and d_{LC} the Log-Cholesky distance. The first is a direct NMSE on the κ_3 entries in the Tucker subspace; the second is a κ_2 distance inside the κ_3-optimal basis (the same target used by the dual-moment `fc_tucker` finetuning loss, Section 3.7), which measures whether the reconstruction preserves the covariance structure that discriminates κ_3 directions.

Interpretation. All four metrics are dimensionless relative errors. A value of 0 is perfect preservation, 1 is the energy of the input object ($\|\hat{X} - X\|^2 = \|X\|^2$, equivalent to predicting zero structure up to a sign flip), and values > 1 mean the reconstruction error exceeds the input signal energy, so predicting zero structure would be a strictly better reconstruction. Table 29 reports these four quantities per (dataset, model-size) cell.

Table 29: Cumulant preservation of pretrained BrainLM between input $\mathbf{Z}$ and reconstruction $\hat{\mathbf{Z}}$ (mean $\pm$ std across subjects; Tucker subspace $R = 80$). κ_3 is destroyed at both scales and on both datasets; scaling amplifies the damage at every order.

Cell	κ_2 LC rel.	κ_2 MSE NMSE	κ_3 MSE NMSE	κ_3 LC rel.
AOMIC 111M	0.294 ± 0.028	0.640 ± 0.328	2.637 ± 0.659	0.838 ± 0.050
AOMIC 650M	0.425 ± 0.007	1.479 ± 0.295	3.864 ± 1.044	1.058 ± 0.055
HCP 111M	0.458 ± 0.003	2.700 ± 1.056	5.663 ± 0.762	1.789 ± 0.120
HCP 650M	0.466 ± 0.004	2.897 ± 1.216	6.212 ± 0.848	1.912 ± 0.133

S CUMULANT FORMALISM

For zero-mean data $\kappa_3 = \mathbb{E}[z_i z_j z_k]$ requires no lower-order subtraction, which is why it can be estimated directly; from κ_4 on, each cumulant subtracts pair-products of κ_2, inheriting any FC-estimation error and inflating sample requirements. How the Tucker directions map onto physiology (nonlinear hemodynamic coupling, excitation and inhibition dynamics) is out of scope here and a natural follow-up for systems neuroscience.

This appendix collects the general cumulant definitions abbreviated in Section 3.1. The joint cumulant-generating function is the Taylor expansion of the log-characteristic function,

$$K(\mathbf{t}) \;=\; \log \mathbb{E}\Big[e^{i\mathbf{t}^\top \mathbf{X}}\Big] \;=\; \sum_{n \geq 1} \frac{1}{n!}\, \kappa_n \cdot (i\mathbf{t})^{\otimes n}, \tag{11}$$

where κ_n is the order-n joint cumulant tensor and $\kappa_n \cdot (i\mathbf{t})^{\otimes n}$ denotes its full contraction with $(i\mathbf{t})^{\otimes n}$. For arbitrary (not necessarily zero-mean) random variables, joint cumulants are defined via the partition formula

$$\kappa(X_1, \ldots, X_n) \;=\; \sum_{\pi} (|\pi| - 1)!\, (-1)^{|\pi|-1} \prod_{B \in \pi} \mathbb{E}\left[\prod_{i \in B} X_i\right], \tag{12}$$

summing over all partitions π of the index set $\{1, \ldots, n\}$. At order 4, cumulants no longer reduce to raw moments:

$$\kappa_4(z_i, z_j, z_k, z_l) \;=\; \mathbb{E}[z_i z_j z_k z_l] - \kappa_2(z_i, z_j)\kappa_2(z_k, z_l) - \kappa_2(z_i, z_k)\kappa_2(z_j, z_l) - \kappa_2(z_i, z_l)\kappa_2(z_j, z_k), \tag{13}$$

isolating genuinely new structure beyond pairwise combinations. The three-line definitions of $\kappa_1, \kappa_2, \kappa_3$ used in the main text are specialisations of this formula to zero-mean variables at orders ≤ 3.

T SUBJECT FINGERPRINTING

This appendix reports a subject-fingerprinting analysis on HCP that complements the main cognitive-prediction story. BFM embeddings fail to discriminate individual subjects from each other broadly, not only at axes that predict the composite, consistent with the variance allocation diagnosis of Section 4.1: the lost structure is general individual variability, of which the composite is one slice.

Protocol. To test whether BFM embeddings preserve subject identity, we split each HCP subject's 4800-TP session into non-overlapping temporal segments (12 for FC, 24–30 for BFMs matching each model's training context length), extract the representation per segment, and compute pairwise cosine distances between all segment-level representations. The raw-FC baseline uses HCP Schaefer-400; each BFM uses its pretraining atlas. We report (i) *identification accuracy*, the fraction of segments whose nearest neighbour belongs to the same subject, and (ii) *separability*, Cohen's d between within-subject and between-subject distance distributions. As a dimensionality control, we project FC to $\{50, 100, 768\}$ dimensions via training-set PCA and repeat the analysis.

Results. Table 30 shows that raw FC identifies individuals at 87% among $\sim$950 HCP subjects ($d = 0.91$), while all three BFMs sit at chance (≈ 0.2–0.4%, $d \approx 0$). Neither scale nor architecture helps: BrainLM-650M ($6\times$ the parameters, 1280-dim embedding) is no more fingerprintable than BrainLM-111M.

Table 30: Subject fingerprinting on HCP ($N = 955$). ID accuracy = fraction of temporal segments whose nearest neighbour in cosine distance belongs to the same subject; Cohen's d = separability between within- and between-subject distance distributions. All three BFMs collapse to $d \approx 0$.

Representation	Dims	ID accuracy	Cohen's d
Raw FC (full)	79,800	**0.872**	**0.910**
FC-PCA (dim. ctrl.)	768	0.843	1.199
FC-PCA (dim. ctrl.)	100	0.443	0.919
BrainLM-111M (CLS)	768	0.002	−0.005
BrainLM-650M (CLS)	1,280	0.003	0.001
Brain-JEPA (pooled)	768	0.004	−0.008

Matched-dimensionality controls rule out feature count: FC-PCA at 768 dims (matching BrainLM CLS) retains 84% and at 100 dims still reaches 44% (Table 30), two orders of magnitude above any BFM. BrainLM CLS tokens in particular collapse to near-constant vectors (cosine distance ≈ 0.002 regardless of subject or segment), a known MAE failure mode in which subject-specific content is pushed into patch tokens while the CLS converges to a generic summary.

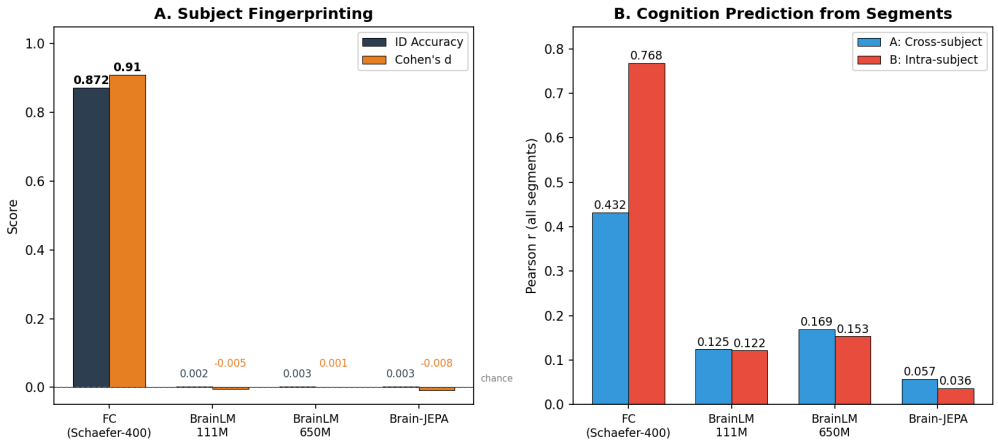


Figure 18: Subject fingerprinting on HCP. Raw FC separates within- from between-subject distances cleanly ($d = 0.91$, 87% ID); all three BFMs collapse the two distributions entirely ($d \approx 0$). FC-PCA at 768 dimensions (same as BrainLM CLS) retains 84% accuracy, ruling out a dimensionality confound.

U Reproducibility Details

Model weights. BrainLM-111M and BrainLM-650M: HuggingFace repository `vandijklab/brainlm`, subfolders `vitmae_111M` and `vitmae_650M`, loaded through

`transformers.ViTMAEForPreTraining.from_pretrained`. Brain-JEPA: checkpoint `jepa-ep300.pth.tar` and the population FC-gradient file `gradient_mapping_450.csv` from the official Brain-JEPA repository (`Eric-LRL/Brain-JEPA`). BrainMass: BNTF online-network state dict (`test.pth`) from the official BrainMass repository (`podismine/BrainMass`). All three BFMs are used in their released pretrained states; no weights were modified prior to the evaluations in Section 4.1.

Datasets. AOMIC-ID1000 (Snoek et al., 2021) is pulled from OpenNeuro accession `ds003097` via datalad; we use the movie-watching run and the fMRIPrep (Esteban et al., 2019) outputs in MNI152NLin2009cAsym space (`desc-preproc_bold.nii.gz`). The AOMIC target is the first principal component of the four Intelligence Structure Test subscales (fluid, memory, crystallised, total); the HCP target is the first principal component of the 58 behavioural measures of Ooi et al. (2022). HCP is the Young Adult S1200 release, using the minimally preprocessed, ICA-FIX-denoised resting-state NIfTI volumes `MNINonLinear/Results/rfMRI_REST{1,2}_{LR,RL}/rfMRI_REST{...}_hp2000_clean.nii.gz` (volumetric, MNI 2 mm).

Preprocessing. AOMIC: NiftiLabelsMasker parcellation to AAL-424 or Schaefer-400 (Schaefer et al., 2018) (7-network, Yeo lab distribution), detrending, bandpass 0.01–0.1 Hz, $T_R = 2.2$ s ($T \approx 290$ timepoints), six rigid-body motion regressors from fMRIPrep (Esteban et al., 2019) confounds. HCP: Schaefer-400 CIFTI parcellation on the fsLR 32k surface (Yeo lab) or AAL-424 on the volumetric NIfTI; bandpass 0.01–0.08 Hz applied per run at $T_R = 0.72$ s; the four REST runs are concatenated to T=4,800 timepoints. Brain-JEPA uses its native Schaefer-400 + Tian Subcortex Scale III (Tian et al., 2020) (50 subcortical ROIs, 450 total) at 160 TP; BrainMass takes the Schaefer-100 FC matrix directly as input.

Evaluation. KRR with a Pearson-correlation kernel, 20 repetitions $\times$ 10 folds (family-aware for HCP), shared fold seeds across methods so all pairwise tests are paired on matched folds, inner CV over $\alpha \in \{10^{-3}, \ldots, 10^{3}\}$. All decompositions (PCA, Tucker) and feature standardisations are fit on training folds only.

Computing the Tucker basis. The $P \times P \times P$ tensor is never formed. For a subject with z-scored timeseries $\mathbf{Z}_n \in \mathbb{R}^{P \times T}$, the mode-1 unfolding $\mathbf{S}_{n(1)}$ of its co-skewness tensor satisfies $\mathbf{S}_{n(1)}\mathbf{S}_{n(1)}^{\top} = \mathbf{Z}_n(\mathbf{C}_n \circ \mathbf{C}_n)\mathbf{Z}_n^{\top}/T^2$ with $\mathbf{C}_n = \mathbf{Z}_n^{\top}\mathbf{Z}_n$ the $T \times T$ Gram matrix and $\circ$ the entrywise product, at a cost of $O(PT^2)$ and a $P \times P$ matrix per subject. $\mathbf{M}$ is the mean of these matrices over the training subjects of the fold and $\mathbf{U}$ holds its top-R eigenvectors, which is the mode-1 HOSVD factor of the training subjects' tensors stacked along a fourth mode. The per-subject matrices are cached once, so each fold sums the matrices of its training subjects. The rank R is the only regularisation.

Train/test splits and leakage controls. The composite itself is a PCA fit: on each dataset the target PCA is fitted on the training split only and applied unchanged to the test split, so no test subject contributes to the definition of the target. AOMIC: 703 train / 173 test. HCP: 763 train / 192 test, family-aware. The train/test partition is used only for BFM finetuning: the reconstruction loss sees training-split inputs, composite labels are never read, and test-split subjects are never seen. All label-free finetuning objectives in the main text therefore leak no composite labels into the downstream KRR. Tucker and PCA bases are fit on training folds only. When needed we trim the timepoints to match each BFM's input length.

Compute. Tucker decomposition and FC-full run on a single 64-core node in $\sim$1–12 h per dataset$\times$parcellation (decomposition plus 200-fold KRR); none of the classical experiments require a GPU. BFM feature extraction runs on a single A100 40 GB in 0.7–1.8 h per model. BFM finetuning runs on a single A100 for 50–250 epochs, 12–72 h depending on model size and schedule. The full Tucker pipeline across all four dataset$\times$parcellation combinations costs less compute than a single pretraining epoch of BrainLM. The project total estimated from cluster accounting (Table 31) is $\approx$ 55,000 effective CPU-hours and $\approx$ 270 A100 40 GB GPU-hours over $\approx$ 9,400 jobs, including preliminary and failed experiments not reported in the paper.

Licenses, citations, and asset use. We rely on the following third-party assets and respect their licenses. *Datasets:* AOMIC-ID1000 (Snoek et al., 2021) is released under CC0 (public domain)

Table 31: Total project compute estimated from SLURM accounting, broken down by experiment family. Includes preliminary and failed runs; the experiments reported in the paper are a subset of these totals. CPU-hours are corrected for typical effective core utilisation ($\sim$25% of allocated cores on the classical KRR/Tucker workloads); GPU-hours are wall-clock A100 time on dedicated single-GPU jobs.

Experiment family	CPU-h	A100-h	# jobs
Tucker / PCA / KRR sweeps	$\approx$ 32,700	0	718
Data preprocessing (fMRIPrep on AOMIC)	$\approx$ 10,800	0	7,104
Per-variable + variability scans	$\approx$ 8,900	0	826
Dataset preparation (parcellation, segments)	$\approx$ 400	0	162
BFM finetuning	$\approx$ 230	183	119
BFM feature extraction	$\approx$ 40	80	201
BFM eval, reconstruction, misc	$\approx$ 1,200	0	$\approx$ 240
Total	$\approx$ 55,000	$\approx$ 270	$\approx$ 9,400

on OpenNeuro (`ds003097`); HCP S1200 (Van Essen et al., 2013) is governed by the WU-Minn HCP Open Access Data Use Terms, under which we registered and accept the no-re-identification clause; the required HCP funding acknowledgement is omitted for anonymity. *Atlases:* the AAL-424 parcellation was introduced by Akiki & Abdallah (2019) (Scientific Reports, CC BY 4.0) and reused by Nemati et al. (2020), and is distributed via the `emergelab/CFP-NFP` GitHub repository as well as bundled with BrainLM; Schaefer-400 (Schaefer et al., 2018) is released by the Yeo lab (CBIG repository) under an MIT-style license; the Tian Subcortex atlas (Tian et al., 2020) is distributed under the Melbourne Subcortex Atlas License, which permits free use subject to citing the original paper. *Software:* fMRIPrep (Esteban et al., 2019) (Apache-2.0); nilearn (BSD), nibabel (MIT), scikit-learn (BSD), PyTorch (BSD), and HuggingFace transformers (Apache-2.0). *BFM checkpoints:* BrainLM-111M / 650M (Caro et al., 2023) are released on HuggingFace under CC BY-NC-ND 4.0; we use them strictly for non-commercial academic research. We finetune BrainLM and report results, but **do not redistribute the resulting checkpoints** in order to comply with the NoDerivatives clause; we release only the finetuning code. Brain-JEPA (Dong et al., 2024) (NeurIPS 2024 Spotlight; codebase at `github.com/Eric-LRL/Brain-JEPA`) and BrainMass (Yang et al., 2024) (IEEE TMI; codebase at `github.com/podismine/BrainMass`) are distributed without an explicit LICENSE file in their repositories; we use the released weights for benchmark inference only, with no modification or redistribution, and cite the original publications.

Code release. An anonymised snapshot of the project code is available at `https://anonymous.4open.science/r/E4C0/`; the repository will be released under an open-source license upon publication.