A CONNECTOME INTERPRETER FOR BRAIN FOUNDATION MODEL: FROM DIRECT TO DIRECTED CONNECTIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain foundation models (BFMs) recently demonstrate strong predictive performance across various clinico-demographic traits; yet, what they internally represent remains a black box. Here, we introduce a data-driven connectome interpreter by training a canonical BFM with bias-free learnable embedding tokens on large-scale human lifespan fMRI datasets. This framework allowed us to investigate biological implication of whole-brain spatiotemporal self-attention, a core operation of BFM. Our analyses revealed that the self-attention, driven by learnable-embedding representations naturally emerging from the data, is far more than an intermediate product for downstream prediction but intrinsically embraces key information to recover the brain's "*direct*" connectivity (unmediated, non-spurious communication backbone). Moreover, integrating these attention weights with an established effective-connectivity method, a framework we formalized as AXON (Attention-informed eXtraction Of directed Networks), accurately resolves "*directionality*" (sender-receiver relationships) of these connections with high accuracy in both mouse and macaque tract-tracing connectomes as well as in the human-brain simulations where the ground-truth network is synthesized based on the empirical data. Finally, when leveraging the AXON-derived effective connectivity to predict comprehensive downstream tasks including demographics, phenotypes (58 behavioral traits) and brain states (8 task conditions), our approach broadly outperformed all state-of-the-art BFMs as well as classic yet powerful linear regression models that have recently rivaled the BFM accuracy, yielding top performance in most individual traits. These results suggest that extracting a biologically meaningful interpretation of what foundation models internally compute—rather than scaling or hand-designing their architecture—is a critical next step for developing robust BFM-based imaging markers.

1 INTRODUCTION

Brain foundation models (BFMs), a class of self-supervised transformers pretrained on large-scale fMRI data, now predict a broad range of cognitive and clinical variables (Thomas et al., 2022; Ortega Caro et al., 2024; Yang et al., 2024). Leveraging multi-cohort open-source datasets, they learn representations that transfer across datasets and phenotypes, a substantial gain for conventional neuroimaging. Yet most BFMs are evaluated almost exclusively by their utility on downstream tasks, and algorithm development has consequently been confined to only improving the accuracy of phenotypic prediction. This outcome-oriented emphasis, however, has left unexamined what may be the most important information a BFM holds, namely the "internal representation of brain dynamics" it could acquire from large volumes of fMRI data, leaving the models whose predictive performance is best established as black boxes in the field.

The key operation to unveil this black box is "*self-attention*", the core of BFMs computing representational dependencies of brain signals between cortical areas. A few reports suggest that attention weights may reflect neural information exchange (Antoniades et al., 2024; Lu et al., 2025) or network connectivity (Khajehnejad et al., 2025), an initial clue calling for more detailed examination of their principle. Yet recent approaches move in effectively the opposite direction: rather than reading out representations learned from data alone, they injected principles of brain organization into

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.

self-attention as inductive biases, such as functional connectivity (FC) (Wei et al., 2026), connectome gradients (Dong et al., 2024), anatomical coordinates (Ortega Caro et al., 2024), and geometric cortical eigenmodes (Dong et al., 2025). While beneficial for downstream performance, internal representations trained this way are inherently constrained by researchers' design intentions, masking what the data genuinely tell us about macroscale brain dynamics. Notably, not simply preventing to read the principles of brain function out of BFMs, this opacity also constrains the predictive markers the field has been developing. Indeed, when a BFM fails to predict a phenotype, we cannot tell whether the pretraining data lacks relevant variance, or the architecture simply fails to exploit what the attention has learned. Reflecting this limitation, recent studies report that BFM accuracy in phenotypic prediction is no better than, and at times even inferior to, conventional machine-learning algorithms (Marraffini et al., 2026; Lane et al., 2026). Given the immense computational resources BFMs demand, why these models fall short of expectations remains a critical, unresolved paradox.

Here, we addressed these issues by dissecting the representational contents of a BFM to build a data-driven connectome interpreter that recovers the communication architecture of the human brain. Our premise was that BFM attention can serve as an integrative framework for quantifying large-scale functional dynamics, rather than merely an intermediate product for phenotypic prediction. For such an interpretation to be meaningful, however, attention must be free of the priors the recent models tried to inject; we therefore removed these human-designed inductive biases, representing each brain ROI as a token combining its temporally segmented fMRI signal with a region-specific learnable embedding vector alone. This design followed from a simple yet untested hypothesis: if each token captures a fingerprint unique to its region's signals, relationships between tokens estimated by self-attention must encode a multi-timescale representation of dynamic connectivity. While exploring this hypothesis revealed that attention indeed excels at capturing the structural backbone of brain networks, we identified a critical theoretical boundary: self-attention alone is fundamentally incapable of resolving the directionality of neural influences—a critical limitation given that a region's functional role depends heavily on whether it acts as a sender-receiver of signals (Reid et al., 2019).

To bridge this gap, we advanced our connectome interpreter with a new framework, **AXON** (**A**ttention-informed e**X**traction **O**f directed **N**etworks), which constrains a vector autoregressive (VAR) model to the scaffold of attention-derived direct connections. AXON thus serves a dual purpose: Methodologically, it translates a mechanistic interpretability insight into a rigorous estimator of directed connectivity. Conceptually, it reveals that BFMs learn rich latent representations that conventional, prediction-oriented readouts fail to exploit—providing a mechanistic explanation for the BFM paradox. Through comprehensive evaluations across species and human lifespan cohorts, our contributions are summarized as follows:

- **Bias-free self-attention detects the structural backbone:** By removing explicit inductive biases, the learned attention recovered tract-tracing connectomes in mice and macaques, as well as the structural connectome underlying simulated human fMRI, significantly outperforming standard estimators like FC and partial correlation in detecting direct connections.
- **AXON supplies the direction that attention leaves unresolved:** By utilizing the attention-derived backbone as a physical substrate to constrain which connections exist, AXON optimally estimates each subject's directed connectivity from their own timeseries, recovering both the topology and the direction of causal neural influence.
- **Empirical validation on both generative and predictive criteria:** AXON establishes a representation-aware neuroimaging marker that excels on multiple fronts. Generatively, whole-brain simulations driven by AXON derived effective connectivity precisely reproduce the empirical resting-state FC, its temporal dynamics (FCD), and the regional time-lag structure (TD). Predictively, AXON significantly outperforms both conventional FC and state-of-the-art BFMs in predicting a wide range of individual traits (e.g., age, sex, and cognition), demonstrating that extracting mechanistically interpretable representations directly resolves the predictive shortcomings of previous models.

2 Related work

Interpreting the attention from BFMs. While rare, attempting to infer brain connectivity from BFMs is not entirely unprecedented. For instance, Neuroformer (Antoniades et al., 2024)—a generative pretrained model that processes neural data as sequences—inferred synaptic wiring among

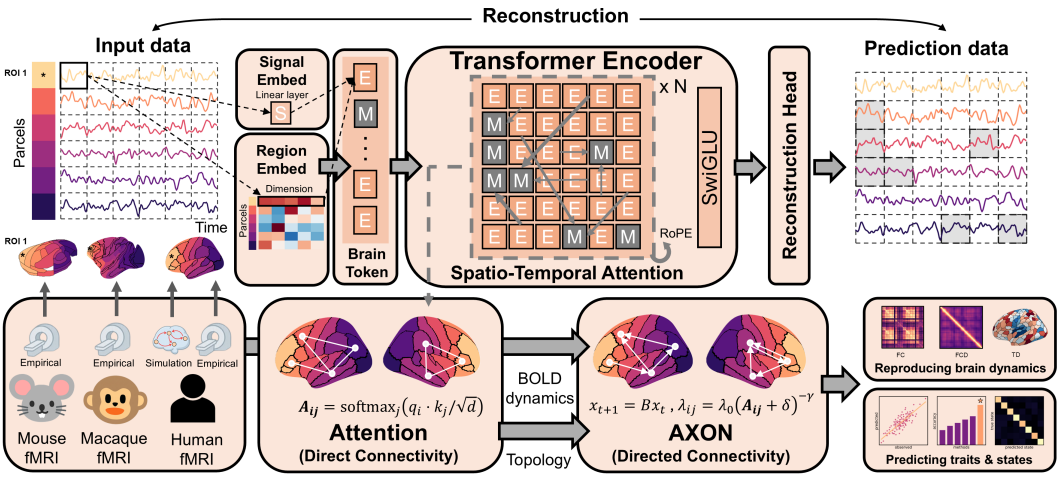

Figure 1: **Algorithmic overview.** A brain foundation model is pretrained by masked reconstruction of parcellated BOLD signals from mouse, macaque, and human fMRI, with self-attention as the only operation that crosses regions and no connectivity information at the input. It turned out that the attention could recover which regions are connected but not the direction of influence. Used as a structural constraint on a vector autoregressive model of each subject's own recording, AXON yields a directed connectome, by which we could reproduce dynamics, predict behavior and demographics, and trace directed influence between regions.

neuronal populations by analyzing internal attention weights. Similarly, NetFormer (Lu et al., 2025) tokenized individual neurons' time series to predict next-step activity and, discarding the softmax, interpreted the query–key product as a nonstationary directed connectivity matrix for biological plausibility. Despite these advances, their success relied heavily on high-resolution, microscopic neural data, while applying these methods to macroscopic fMRI dynamics is fundamentally hindered by the complexities of whole-brain bidirectional communication and intricate structure-function mappings, as well as by the noisy and indirect nature of fMRI measures. To resolve this issue, the field has already proposed a myriad of ingenious methods to infer effective connectivity (EC) from fMRI; yet they fundamentally rely on unconstrained temporal modeling of short timeseries, leaving them highly susceptible to spurious indirect connections and severely degrading the specificity of directed network mapping. In this study, we sought to address this limitation by developing AXON, a hybrid approach that leverages BFM-learned attention as a structural prior to precisely capture the macroscale EC from fMRI. By doing so, we move beyond simple attention-weight extraction for downstream tasks to reconstruct a neurobiologically grounded sender-receiver network ("*directed connectivity*").

Connectomes as predictive features. Beyond scientific insights into neural dynamics, brain connectivity could offer immense practical value for phenotypic prediction. Reflecting this, for the past two decades—well before the advent of BFMs, the field has extensively investigated connectome-based predictive models using traditional machine learning (ML) algorithms (He et al., 2020). In particular, FC which relies on simple linear correlations has served as a highly effective input for achieving the most reproducible phenotypic predictions. Paradoxically, recent studies showed that state-of-the-art BFMs frequently fail to surpass these vastly smaller, traditional algorithms, sparking skepticism about their utility. For example, Marraffini et al. (2026) showed that in predicting cognitive abilities, the 650-million-parameter BrainLM ($r = 0.003$) was drastically outperformed by a raw FC-based kernel ridge regression model ($r = 0.393$), even exhibiting an "inverse scaling" phenomenon where larger model sizes led to worse performance. Furthermore, the comprehensive neuroimaging benchmark BRAINMARKS (Lane et al., 2026) demonstrated that embeddings from state-of-the-art BFMs could not beat a basic logistic regression baseline applied to raw FC. We hypothesize that this unexpected underperformance of BFMs is not an inherent flaw of the models themselves, but rather stems from a lack of mechanistic interpretability—namely, an inadequate understanding of the semantic meaning embedded within their internal self-attention representations. To validate this, our study systematically evaluated the downstream predictive accuracy of AXON, which explicitly leverages this attention backbone to derive informed, highly predictive ECs.

3 METHOD

3.1 A PRIOR-FREE BRAIN FOUNDATION MODEL AND ITS ATTENTION

Prior-free masked pretraining. To build the BFM, we employed a canonical encoder-only transformer by masked reconstruction of parcellated BOLD signals. Let $X \in \mathbb{R}^{V \times T}$ be the timeseries of V parcel regions over T timepoints, divided into P non-overlapping temporal patches of P_s timepoints. Each token corresponds to one region observed over one time window. We called this a *brain token*, and its embedding combines the BOLD signal in that window of a given brain parcel together with learnable parameters designed to act as a spatial anchor, which we posit will eventually encode the unique identity of that region,

$$\underbrace{z_{i,p}}_{\text{brain token}} = \mathrm{LN}(W x_{i,p}) + \mathrm{LN}(E_i)\,, \tag{1}$$

where W is a shared signal projection, $E \in \mathbb{R}^{V \times D}$ a region-specific learnable spatial embedding, and LN a layer normalisation. The embedding E was initialized at random and carried no relational information between regions. Using these brain tokens, self-attention ran jointly over regions and time windows, which is what made it spatiotemporal. These tokens were then trained by standard masked reconstruction (He et al., 2022). A random subset $\mathcal{M}$ of tokens was masked at ratio m, and the signal term was replaced by a learned mask embedding,

$$z_{i,p} = (1 - M_{i,p})\,\mathrm{LN}(W x_{i,p}) + M_{i,p}\, z_{\text{mask}} + \mathrm{LN}(E_i)\,, \tag{2}$$

so that the masked brain token retained its regional placeholder but none of its signal. Finally, the model reconstructed the masked patches, with the loss computed on masked tokens only,

$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\mathcal{M}} \left[\frac{1}{|\mathcal{M}|} \sum_{(i,p)\in\mathcal{M}} \left\| g\left(h_{i,p}^{(L)}\right) - x_{i,p} \right\|^2 \right], \tag{3}$$

with $h_{i,p}^{(L)}$ the final hidden state and g a linear head back to patch space. We randomly masked the tokens over regions and time windows. The objective was what makes the attention worth reading: a masked token carries only the spatial anchor of its region, therefore the only way to reconstruct its signal is to draw on the other regions, forcing the model to intrinsically learn their inter-regional latent dependencies. The architecture and optimisation process are detailed in **Appendix A**.

Attention as connectivity. As attention is computed between brain tokens, its weight between two tokens structurally encodes a representational similarity between the two corresponding brain regions, following a canonical self-attention equation:

$$a((i,p),(j,q)) = \underset{j,q}{\mathrm{softmax}} \left(\frac{q_{i,p} \cdot k_{j,q}}{\sqrt{d}} \right), \qquad q_{i,p} = W_Q z_{i,p}, \quad k_{j,q} = W_K z_{j,q}. \tag{4}$$

Since $z_{i,p}$ carries the BOLD signal of the region i during patch p, together with its spatial identity, this weight (or similarity) mechanistically dictates how much the BOLD of region j during patch q is utilized to reconstruct region i during patch p. Because a masked brain token contributes no signal of its own, its reconstruction is made entirely of the BOLD of other regions, weighted by a; and self-attention is the only operation in the encoder that crosses regions, so any inter-regional dependency the model learns must reside in a. To distill these spatiotemporal dependencies into a unified measure of inter-regional interaction, we pooled the attention weights over layers, heads, and patch pairs into a single spatial matrix:

$$\underbrace{A_{ij}}_{\text{latent connectivity}} = \frac{1}{LHP^2} \sum_{l,h} \sum_{p,q} a^{(l,h)}((i,p),(j,q))\,, \tag{5}$$

We therefore hypothesize that this A encapsulates the latent connectivity the model has learned. In the analyses that follow, we empirically unpack this mechanism to validate whether this attention-derived matrix can indeed be interpreted and utilized as the whole brain's connectome.

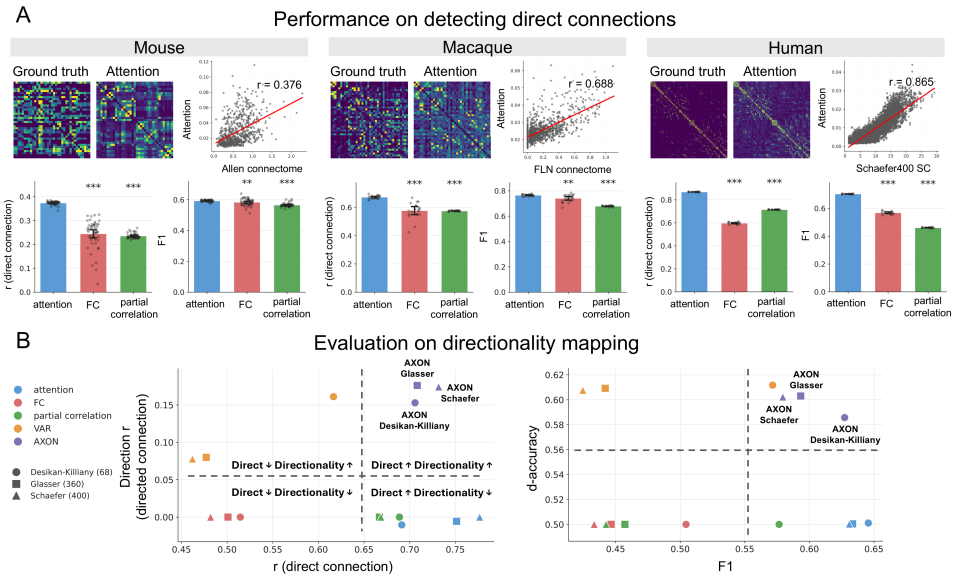

Figure 2: **Attention recovers direct connections only, while AXON reveals even their directionality. (A)** Attention, functional connectivity (FC), and partial correlation scored against the ground-truth connectome in mouse, macaque, and human simulations (Schaefer-400; other atlases in Appendix C): ground-truth and attention matrices with their correlation (top), and its r and edge recovery (F1) across held-out subjects (bottom). Asterisks indicate differences from attention (Wilcoxon signed-rank test; $^{**}p < 0.01$, $^{***}p < 0.001$). **(B)** The same estimators, together with VAR and AXON, to detect directionality on simulations driven by a rewired directed connectome at three atlases. In both panels, AXON occupied the upper-right quadrant across all three atlases, accurately recovering both the direct and directed connections.

3.2 From direct to directed connectivity

At its core, our BFM relies on a self-attention mechanism that processes dependencies simultaneously across both spatial and temporal dimensions. From a computational perspective, this formulation shares a structural resemblance to the key principles underlying effective connectivity (EC) estimation, as both inherently quantify how the temporal dynamics of one region contribute to predicting the state of another. At the same time, however, given that the masked reconstruction objective leverages bidirectional temporal context rather than strict autoregressive prediction, the attention mechanism may act more conservatively—capturing only a robust structural skeleton of direct interactions without fully resolving directed temporal causality. To systematically investigate these possibilities, we accommodated both scenarios in our framework. While we first evaluated the raw attention matrix as a potential standalone estimator of EC, we also introduced a synergistic approach that leverages attention strictly as an informative prior: by integrating this attention-derived skeleton into a VAR model, we explicitly combined the structural robustness of attention with the statistical directionality of VAR to optimally estimate EC.

A first-order VAR with ridge regularization has been a standard instrumental to estimate a directed matrix in a linear model (Chiang et al., 2017), as follows:

$$x_{t+1} = Bx_t + \epsilon_t, \qquad \hat{B} = \arg\min_B \sum_t \|x_{t+1} - Bx_t\|^2 + \lambda \sum_{ij} B_{ij}^2, \tag{6}$$

where B_{ij} quantifies how much the state of region i at time $t+1$ is explained by region j at time t, naturally allowing $B_{ij} \neq B_{ji}$. In our fMRI setting with $V = 400$ brain parcels and only a few-hundred timepoints, estimating entire B coefficients (=400×400=160k) is a highly ill-posed problem. While standard ridge regression stabilizes this via a global penalty, it shrinks all coefficients uniformly—blindly penalizing genuine brain interactions just as stringently as spurious noise. This is precisely where the attention matrix A (Equation 5) could be informative: operating under the hypothesis that A captures the robust structural skeleton of cortical interactions, we leveraged it as a data-driven prior to make the regularization *edge-specific*. To this end, we formulated an

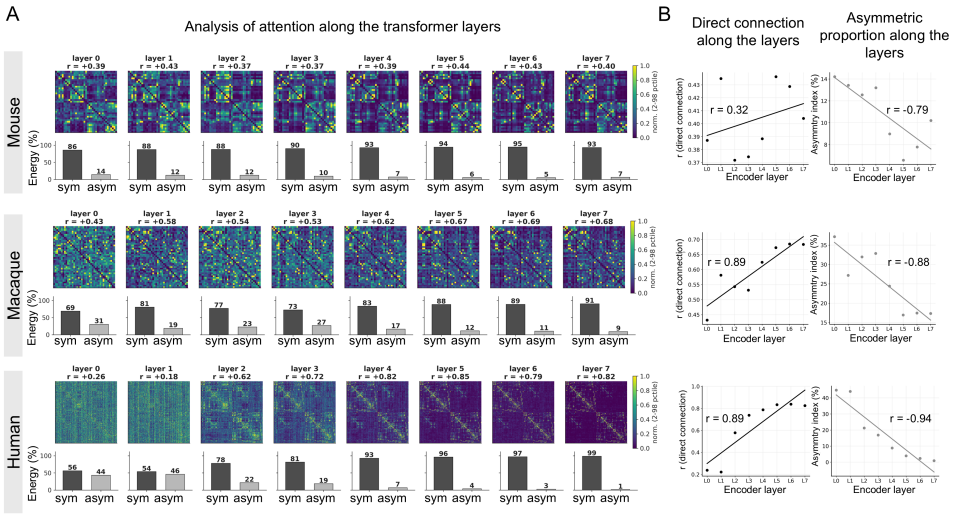

Figure 3: **Emergence of symmetric direct connections across transformer layers. (A)** Attention of each encoder layer for one held-out subject, with its r to the ground truth network above and the symmetric-asymmetric shares of its energy below (Glasser-360 for human). **(B)** Direct r and asymmetry index (Appendix B.2) across 8 layers, mean over held-out subjects.

attention-informed VAR (AXON) by scaling the penalty inversely with the attention score:

$$\lambda_{ij} = \lambda_0 \left(A_{ij} + \delta\right)^{-\gamma}, \qquad \hat{B} = \arg\min_{B} \sum_{t} \|x_{t+1} - Bx_t\|^2 + \sum_{ij} \lambda_{ij} B_{ij}^2, \tag{7}$$

such that an edge with a high score of attention (=strong direct connection) receives a weak penalty and is left free for the data to set, while one with low attention score (=weak direct connection) is shrunk toward zero. δ keeps the penalty finite, and γ sets how far the attention is trusted. At $\gamma = 0$ the penalty becomes uniform, and Equation 7 reduces to the plain VAR baseline.

3.3 Data

We validated AXON with two following criteria: 1) how well does it detect direct connections that genuinely exist in brain networks? and 2) how well does it capture the directionality of those identified connections? (Table 1).

Table 1: Validation steps. Step-1 and -2 compared the estimated connectome with a known ground truth; Step-3, lacking a ground truth for empirical data, evaluated the estimated connectome by the dynamics it generates in a whole-brain simulation.

		Data	Ground truth	Evaluated by
Step-1	Direct connection	Mouse rs-fMRI Macaque rs-fMRI Simulated BOLD	Allen Mouse Connectivity Atlas M132 tract-tracing ENIGMA SC	direct r, F1, AUC
Step-2	Directionality	Simulated BOLD	ENIGMA SC (direction rewired)	direction r, d-accuracy
Step-3	Empirical dynamics	HCP-YA, -A, -D BOLD	Not applicable	FC, FCD, TD

SC: structural connectivity; FCD: FC dynamics; TD: temporal delay.

Ground truth and fMRI data for Step-1, 2 To evaluate the model's ability to capture direct connections (Step-1), we utilized **mouse** resting-state fMRI from the Grandjean dataset (Grandjean et al., 2020) ($n = 253$; 202 pretraining, 51 held-out; Allen 37-region atlas), using the Allen Mouse Connectivity Atlas as ground-truth EC (Oh et al., 2014). We also employed the **macaque** PRIME-DE dataset (Milham et al., 2018) ($n = 19$; M132 atlas), evaluating attention weights via a leave-one-out approach due to sample size, with ground-truth directed connectivity defined by histological tract-tracing (Markov et al., 2014). For **human** connectivity, where invasive tract-tracing

is impossible, we generated synthetic BOLD signals for 1,100 subjects (1,000 pretraining, 100 held-out) using a Hopf model (Deco et al., 2017). This model operated on empirical symmetric structural connectomes from the ENIGMA toolbox (Larivière et al., 2021) across the Desikan–Killiany-68, Glasser-360, and Schaefer-400 atlases. To test directed connectivity (Step-2), we simulated BOLD signals from an asymmetric network created by randomly rewiring edge orientations while preserving node degrees (Rubinov & Sporns, 2010), thereby isolating the capacity to detect directionality without altering the structural backbone. Acquisition details and splits are provided in Appendix B.

Human empirical fMRI data for Step-3. We utilized a large-scale HCP fMRI data from three lifespan cohorts (Van Essen et al., 2013; Bookheimer et al., 2019; Somerville et al., 2018), partitioning each into an 80:20 split for pretraining and held-out evaluation: young adult ($n = 1093$), aging ($n = 722$), and development ($n = 630$). Because empirical fMRI has no ground-truth connectome, our validation framework shifted from direct connectome recovery to generative simulation: we parameterized the whole-brain Hopf model with each held-out subject's estimated connectome and compared the simulated dynamics against their empirical recordings using FC, functional connectivity dynamics (FCD), and Temporal Delay (TD) (Appendix B.2). For downstream tasks, we also included the 7 task-fMRI conditions of HCP-YA (Barch et al., 2013). Each condition was pretrained as an independent model, ensuring the attention backbone exclusively represented condition-specific dynamics.

3.4 Evaluation for downstream tasks

We evaluated a practical utility of our framework, comparing their predictive capacity on various individual traits as well as on brain states to that of other algorithms, based on the HCP-YA data. The final feature set to predict included demographics (age and sex), 58 behavioral measures across 8 different domains (NIH Toolbox cognition, Penn cognition, sensory and physical, reward and sleep, task performance, NEO-FFI personality, emotion recognition, and affect and well-being) and the brain states (rest vs. 7 task conditions). Before prediction, all pretrained BFM encoders were frozen without further fine-tuning, as it could alter model parameters for each prediction feature, which can confound whether performance reflects the pretraining representation or target-specific optimization. Given this context, an identical linear probe was applied across all methods we compared (our and other state-of-the-art BFMs and linear models leveraging the estimated connectome, i.e., FC, VAR, and AXON), using ridge regression for continuous features and logistic regression for categorical features (Pedregosa et al., 2011). This ensured that any difference in performance would result from strictly the input representation alone. The probes were trained on the 80% of subjects used for pretraining, and the remaining 20% was split equally into validation and test sets, with hyperparameters selected on validation and scores computed on the test cases (Appendix E).

4 Results

4.1 Step-1, 2: Attention recovers direct connections but not directionality

Our evaluation started with checking the ability of our BFM in identifying the direct connections purely based on fMRI signals. When evaluating this across the mouse mesoscale connectome, the macaque M132 tract-tracing matrix, and the structural connectome underlying simulated human-brain BOLD signals, the learned attention consistently outperformed FC and partial correlation in recovering both the structural backbone and coupling strength across all species (Figure 2A). Intriguingly, layer-wise inspection of the transformer (layers 0 to 7) revealed a progressive abstraction of this topology. As network depth increased, the attention matrix yielded higher accuracy in detecting direct connections, yet its degree of asymmetry—a proxy for directedness—steadily decreased (Figure 3). This divergence suggests that deeper layers distill an increasingly abstract representation of the connectome, becoming highly tuned to the mere existence of structural pathways rather than the specific direction of information flow. Our framework achieved these findings without receiving any explicit inductive bias about brain organization, highlighting the power of fully unbiased data representation. Next, we investigated whether the attention also encodes directionality, using BOLD signals generated from a Hopf model constructed by a synthetically rewired directed connectome. Our analyses revealed that the attention retained its advantage in detecting direct connections but fell at chance for directionality. The BFM's attention therefore functions as a backbone detector rather

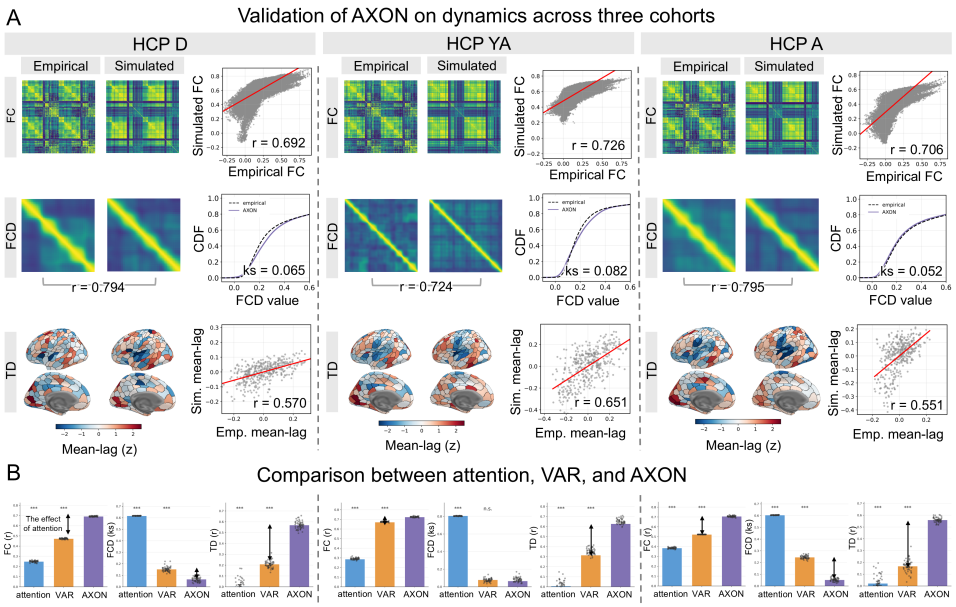

Figure 4: **AXON reproduces resting-state dynamics. (A)** For one held-out subject per cohort, ordered by age, the whole-brain simulation was driven by the AXON directed connectome and compared with that subject's recording on functional connectivity (FC), its fluctuation over time (FCD), and regional time-lag structure (TD). **(B)** The same comparison across all held-out subjects, for the attention, VAR, and AXON, in the cohort order of (A); Asterisks indicate differences from AXON (Wilcoxon signed-rank test; $^{***}p < 0.001$; n.s., not significant).

than a directed estimator. This challenge, however, was solved through our new addition, AXON, which could recover both the direct connections and their directionality with high fidelity across the Desikan–Killiany, Glasser, and Schaefer atlases (see Section 4.2; Figure 2B).

4.2 Step-3: AXON reproduces functional dynamics via directed connections

Because empirical fMRI provides no ground-truth connectome, we evaluated each directed connectome estimated by AXON through the dynamics it generated. Specifically, we drove simulated BOLD signals from a whole-brain model with each held-out subject's AXON-derived EC and compared them with that subject's empirical resting-state data. AXON achieved the highest accuracy on all evaluation metrics in every cohort, reproducing the static FC, its fluctuation over time (FCD), and the regional time-lag structure (TD) more closely than both raw attention and plain VAR (Figure 4). AXON therefore reproduced not only the static direct connections but also their temporal fluctuations and the regional lead-lag structure across the lifespan data, suggesting its high fidelity in capturing the brain's intrinsic spatiotemporal dynamics, which result is even highlighted given that neither the pure attention nor an uninformed VAR sufficed on its own.

4.3 Step-3: Downstream prediction

Having validated the connectome against ground truth and empirical dynamics, we next examined the predictive power of BFMs and AXON, together with all baseline models, on the 58 behavioral measures of HCP-YA (Section 3.4). Consistent with Marraffini et al. (2026), the connectome-based estimators (i.e., FC, VAR, AXON) outperformed all pretrained BFMs read out from their embeddings. Notably, AXON ranked first across all 8 behavioral categories and on 35 of the 58 individual measures (Table 2; Figure 5A). In our BFM, the global-average-pooled embedding yielded a mean r of 0.092, features from the attention 0.138, and AXON 0.210, indicating that the model predicts better when reading out as an interpretable connectome than as black-box features. We further tested whether a connectome estimated from the task itself improves prediction of task performance. Across 7 in-scanner performance measures from 5 tasks, AXON-derived EC from the task scan outperformed its resting-state counterpart by 0.07 to 0.17 in correlation, significantly so for 5 of the

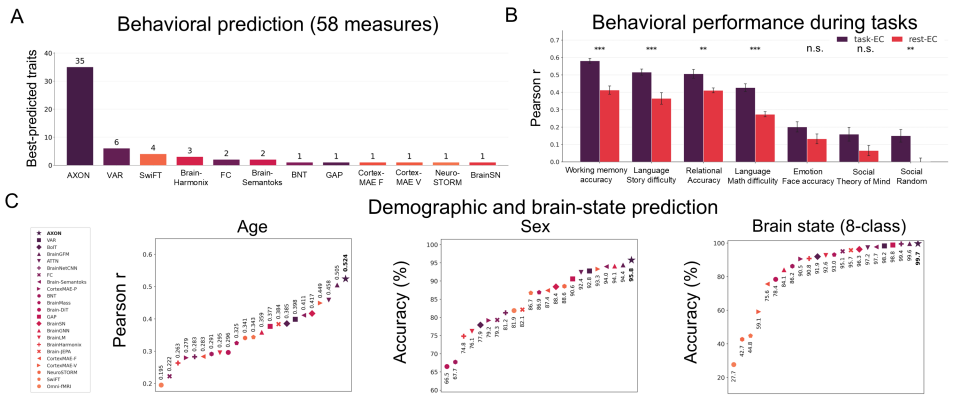

Figure 5: **Downstream prediction in HCP-YA. (A)** Number of the 58 behavioral measures on which each method achieved the highest prediction accuracy. **(B)** Prediction of in-scanner task performance from AXON-derived EC estimated during the corresponding task (task-EC) or at rest (rest-EC); error bars are s.d. over 10 folds ($^{**}p < 0.01$, $^{***}p < 0.001$; paired t-test). **(C)** Age, sex, and 8-class brain-state prediction, with all methods ranked and AXON marked with a star.

7 measures (Figure 5B). Beyond behavior, AXON also ranked first on age, sex, and brain-state classification (rest and 7 tasks; Figure 5C; full tables in Appendix E).

Table 2: Per-category behavioral prediction on HCP-YA (Pearson r averaged within each category; mean $\pm$ s.d. over ten random splits). *Conn*: connectome features; *DL*: deep neural networks (selected as standard ANN baselines, generally known for strong state decoding); *BFM*: pretrained BFMs with frozen encoders; *Ours*: features from our model, i.e., mean-pooled final hidden states (GAP), the attention matrix (ATTN), or AXON. **All**: mean over all 58 measures. **Best** and second per column. †Significantly lower than AXON (paired test, Holm–Bonferroni corrected, $p < 0.05$).

	Methods	**Cog(NIH)** (n=7)	**Cog(Penn)** (n=5)	**Sensory** (n=9)	**Reward** (n=2)	**TaskfMRI** (n=7)	**NEO** (n=5)	**EmoRec** (n=6)	**Affect** (n=17)	**All** (n=58)
Conn	FC	0.310† ±.029	0.224† ±.036	0.178† ±.034	0.229 ±.039	0.203† ±.024	0.201 ±.048	0.045 ±.038	0.104† ±.040	0.169† ±.015
	VAR	0.345 ±.025	0.247† ±.026	0.210 ±.023	0.234† ±.057	0.225 ±.037	0.227 ±.035	0.064 ±.045	0.135† ±.018	0.197† ±.015
DL	BrainNetCNN	0.277† ±.043	0.176† ±.021	0.144† ±.016	0.173† ±.061	0.157† ±.031	0.149† ±.037	0.019† ±.045	0.042† ±.027	0.123† ±.014
	BrainGNN	0.128† ±.039	0.084† ±.022	0.090† ±.035	0.100† ±.051	0.087† ±.054	0.040† ±.029	-0.007† ±.050	0.031† ±.018	0.062† ±.016
	BolT	0.214† ±.026	0.143† ±.030	0.098† ±.042	0.137† ±.048	0.126† ±.030	0.099† ±.038	0.019† ±.049	0.026† ±.027	0.091† ±.017
	BNT	0.111† ±.062	0.085† ±.036	0.097† ±.035	0.105† ±.028	0.076† ±.046	0.070† ±.058	0.003† ±.051	0.035† ±.028	0.065† ±.012
BFM	SwiFT	0.177† ±.051	0.135† ±.037	0.171† ±.029	0.152† ±.081	0.131† ±.049	0.092† ±.059	0.053 ±.035	0.082† ±.037	0.118† ±.015
	BrainLM	0.069† ±.050	0.032† ±.057	0.053† ±.026	0.028† ±.076	0.033† ±.044	0.035† ±.028	0.027† ±.032	0.005† ±.043	0.031† ±.023
	BrainMass	0.202† ±.040	0.153† ±.048	0.131† ±.032	0.162† ±.039	0.130† ±.038	0.083† ±.039	0.026† ±.019	0.053† ±.042	0.105† ±.014
	Brain-JEPA	0.085† ±.042	0.068† ±.044	0.081† ±.033	0.046† ±.049	0.070† ±.067	0.066† ±.037	0.045 ±.038	0.026† ±.035	0.057† ±.023
	BrainSN	0.092† ±.053	0.097† ±.039	0.085† ±.025	0.091† ±.050	0.107† ±.026	0.059† ±.046	-0.013† ±.056	0.055† ±.044	0.069† ±.025
	BrainGFM	0.227† ±.041	0.138† ±.021	0.142† ±.043	0.154† ±.071	0.140† ±.048	0.118† ±.050	0.028† ±.048	0.053† ±.028	0.112† ±.010
	NeuroSTORM	0.150† ±.044	0.131† ±.035	0.165† ±.019	0.149† ±.069	0.127† ±.045	0.059† ±.056	0.028 ±.029	0.030† ±.031	0.092† ±.018
	BrainHarmonix	0.153† ±.048	0.145† ±.039	0.187 ±.039	0.157† ±.064	0.130† ±.049	0.094† ±.056	0.049 ±.047	0.064† ±.049	0.113† ±.031
	CortexMAE-P	0.175† ±.042	0.134† ±.052	0.128† ±.043	0.175† ±.042	0.153† ±.048	0.134† ±.047	0.026† ±.042	0.058† ±.028	0.108† ±.025
	CortexMAE-F	0.152† ±.040	0.119† ±.051	0.150† ±.033	0.105† ±.042	0.132† ±.040	0.081† ±.031	0.006† ±.034	0.049† ±.019	0.093† ±.007
	CortexMAE-V	0.195† ±.035	0.143† ±.060	0.172† ±.023	0.137† ±.058	0.148† ±.046	0.097† ±.035	0.044 ±.032	0.062† ±.059	0.116† ±.023
	Brain-Semantoks	0.298† ±.046	0.176† ±.032	0.177† ±.023	0.165† ±.090	0.198† ±.021	0.185† ±.043	0.038 ±.054	0.085† ±.039	0.153† ±.023
	Omni-fMRI	0.106† ±.064	0.099† ±.059	0.122† ±.031	0.066† ±.046	0.074† ±.029	0.066† ±.050	0.038 ±.047	0.039† ±.038	0.072† ±.022
	Brain-DiT	0.113† ±.051	0.058† ±.043	0.084† ±.044	0.072† ±.080	0.104† ±.051	0.028† ±.038	0.021† ±.066	0.001† ±.042	0.052† ±.030
Ours	GAP	0.143† ±.048	0.097† ±.058	0.119† ±.025	0.123† ±.071	0.113† ±.040	0.111† ±.041	0.040 ±.041	0.057† ±.032	0.092† ±.021
	ATTN	0.222† ±.039	0.186† ±.038	0.182† ±.031	0.185† ±.035	0.174† ±.035	0.156† ±.044	0.016† ±.052	0.083† ±.051	0.138† ±.017
	AXON	**0.352** ±.027	**0.272** ±.027	**0.218** ±.033	**0.262** ±.065	**0.236** ±.021	**0.235** ±.030	**0.071** ±.047	**0.153** ±.022	**0.210** ±.016

5 Conclusion

We examined what BFMs represent internally by analyzing the self-attention that couples brain regions. Without explicit inductive biases, the learned attention recovered the structural backbone of the connectome across three species more accurately than FC and partial correlation, but it could not resolve the direction of influence. By combining this attention-derived backbone with each subject's own timeseries, AXON successfully recovered directionality, reproduced empirical resting-state dynamics, and outperformed both FC-based linear models and existing BFMs in predicting demographics, behavioral traits, and brain states. The predictive potential of a BFM is thus better realized through an interpretable connectome than through black-box features. While AXON currently resolves directionality via a classical VAR model, future work will examine decoder-only architectures trained with next-step prediction, which may enable BFMs to learn directed connectivity entirely within the model itself in an end-to-end fashion.

AI USE STATEMENT

We used large language models (LLMs) throughout this work, in both the research and the preparation of the manuscript. In the research, LLMs assisted with searching and verifying the related literature, brainstorming and refining ideas, designing analyses and experiments, implementing and debugging code for model training, simulation, data analysis, and statistical testing, and generating figures. In the manuscript, LLMs were used to draft and revise parts of the main text, appendix, and figure and table captions, and to improve clarity, grammar, and consistency of terminology. The authors directed the research, made all final decisions on study design and interpretation, and reviewed, verified, and edited all AI-assisted content, including code and reported numbers.

REFERENCES

Antonis Antoniades, Yiyi Yu, Joseph Canzano, William Wang, and Spencer LaVere Smith. Neuroformer: Multimodal and multitask generative pretraining for brain data. In *International Conference on Learning Representations (ICLR)*, 2024.

Deanna M. Barch, Gregory C. Burgess, Michael P. Harms, Steven E. Petersen, Bradley L. Schlaggar, Maurizio Corbetta, Matthew F. Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, Dan Nolan, Edward Bryant, Tucker Hartley, Owen Footer, James M. Bjork, Russ Poldrack, Steve Smith, Heidi Johansen-Berg, Abraham Z. Snyder, and David C. Van Essen. Function in the human connectome: task-fMRI and individual differences in behavior. *NeuroImage*, 80:169–189, 2013. doi: 10.1016/j.neuroimage.2013.05.033.

Susan Y. Bookheimer, David H. Salat, Melissa Terpstra, Beau M. Ances, Deanna M. Barch, Randy L. Buckner, Gregory C. Burgess, Sandra W. Curtiss, Mirella Diaz-Santos, Jennifer S. Elam, et al. The lifespan human connectome project in aging: an overview. *NeuroImage*, 185: 335–348, 2019. doi: 10.1016/j.neuroimage.2018.10.009.

Sharon Chiang, Michele Guindani, Hsiang J. Yeh, Zulfi Haneef, John M. Stern, and Marina Vannucci. Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Human Brain Mapping*, 38(3):1311–1332, 2017. doi: 10.1002/hbm.23456.

Gustavo Deco, Morten L. Kringelbach, Viktor K. Jirsa, and Petra Ritter. The dynamics of resting fluctuations in the brain: metastability and its dynamical cortical core. *Scientific Reports*, 7:3095, 2017. doi: 10.1038/s41598-017-03073-5.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Joanes Grandjean, Carola Canella, Cynthia Anckaerts, Gülebru Ayrançı, Salma Bougacha, Thomas Bienert, David Buehlmann, Ludovico Coletta, et al. Common functional networks in the mouse brain revealed by multi-centre resting-state fMRI analysis. *NeuroImage*, 205:116278, 2020. doi: 10.1016/j.neuroimage.2019.116278.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Tong He, Ru Kong, Avram J. Holmes, Minh Nguyen, Mert R. Sabuncu, Simon B. Eickhoff, Danilo Bzdok, Jiashi Feng, and B. T. Thomas Yeo. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276, 2020. doi: 10.1016/j.neuroimage.2019.116276.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165, 2020.

Moein Khajehnejad, Forough Habibollahi, Devon Stoliker, and Adeel Razi. BrainSymphony: A parameter-efficient multimodal foundation model for brain dynamics with limited data. *arXiv preprint arXiv:2506.18314*, 2025.

Connor Lane, Mihir Tripathy, Leema Krishna Murali, Ratna Sagari Grandhi, Shamus Sim Zi Yang, Sam Gijsen, Debojyoti Das, Manish Ram, Utkarsh Kumar Singh, Cesar Kadir Torrico Villanueva, Yuxiang Wei, Will Beddow, Gianfranco Cortés, Suin Cho, Daniel Z. Kaplan, Benjamin Warner, Tanishq Mathew Abraham, and Paul S. Scotti. Scaling vision transformers for functional MRI with flat maps. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2510.13768.

Sara Larivière, Casey Paquola, Bo-yong Park, Jessica Royer, Yezhou Wang, Oualid Benkarim, Reinder Vos de Wael, Sofie L. Valk, et al. The ENIGMA toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nature Methods*, 18(7):698–700, 2021. doi: 10.1038/s41592-021-01186-4.

Ziyu Lu, Wuwei Zhang, Trung Le, Hao Wang, Uygar Sümbül, Eric Shea-Brown, and Lu Mi. NetFormer: An interpretable model for recovering dynamical connectivity in neuronal population dynamics. In *International Conference on Learning Representations (ICLR)*, 2025.

Nikola T. Markov, Maria M. Ercsey-Ravasz, A. R. Ribeiro Gomes, Camille Lamy, Loic Magrou, Julien Vezoli, Pierre Misery, Arnaud Falchier, Rene Quilodran, M. A. Gariel, et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex*, 24(1): 17–36, 2014. doi: 10.1093/cercor/bhs270.

Giovanni Marraffini, Gabriel Mahuas, Trinidad Borrell, Victoria Shevchenko, and Demian Wassermann. The variance brain foundation models forgot: Third-order statistics predict cognition where billion-parameter models fail. *arXiv preprint arXiv:2606.04010*, 2026.

Michael P. Milham, Lei Ai, Bonhwang Koo, Ting Xu, Céline Amiez, Fabien Balezeau, Mark G. Baxter, Erwin L. A. Blezer, et al. An open resource for non-human primate imaging. *Neuron*, 100 (1):61–74.e2, 2018. doi: 10.1016/j.neuron.2018.08.039.

Seung Wook Oh, Julie A. Harris, Lydia Ng, Brent Winslow, Nicholas Cain, Stefan Mihalas, Quanxin Wang, Chris Lau, Leonard Kuan, Alex M. Henry, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014. doi: 10.1038/nature13186.

Younghyun Oh, Yejin Ann, Jae-Joong Lee, Takuya Ito, Sean Froudist-Walsh, Casey Paquola, Michael Milham, R. Nathan Spreng, Daniel Margulies, Boris Bernhardt, Choong-Wan Woo, and Seok-Jun Hong. State-dependent signal flow hierarchy in the human cerebral cortex. *Nature Neuroscience*, 2026. doi: 10.1038/s41593-026-02389-8.

Josue Ortega Caro, Antonio Henrique de Oliveira Fonseca, Syed Asad Rizvi, Matteo Rosati, Christopher L. Averill, James L. Cross, Prateek Mittal, Emanuele Zappala, Rahul Madhav Dhodapkar, Chadi Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Jason Ramapuram, Federico Danieli, Eeshan Gunesh Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russell Webb. Theory, analysis, and best practices for sigmoid self-attention. In *International Conference on Learning Representations (ICLR)*, 2025.

Andrew T. Reid, Drew B. Headley, Ravi D. Mill, Ruben Sanchez-Romero, Lucina Q. Uddin, Daniele Marinazzo, Daniel J. Lurie, Pedro A. Valdés-Sosa, Stephen José Hanson, Bharat B. Biswal, Vince Calhoun, Russell A. Poldrack, and Michael W. Cole. Advancing functional connectivity research from association to causation. *Nature Neuroscience*, 22(11):1751–1760, 2019. doi: 10.1038/s41593-019-0510-4.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. doi: 10.1016/j.neuroimage.2009.10.003.

Stephen M. Smith, Karla L. Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F. Beckmann, Thomas E. Nichols, Joseph D. Ramsey, and Mark W. Woolrich. Network modelling methods for FMRI. *NeuroImage*, 54(2):875–891, 2011. doi: 10.1016/j.neuroimage.2010.08.063.

Leah H. Somerville, Susan Y. Bookheimer, Randy L. Buckner, Gregory C. Burgess, Sandra W. Curtiss, Mirella Dapretto, Jennifer S. Elam, Michael S. Gaffrey, Michael P. Harms, Cynthia Hodge, et al. The lifespan human connectome project in development: A large-scale study of brain connectivity development in 5–21 year olds. *NeuroImage*, 183:456–468, 2018. doi: 10.1016/j.neuroimage.2018.08.050.

Armin W. Thomas, Christopher Ré, and Russell A. Poldrack. Self-supervised learning of brain dynamics from broad neuroimaging data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 21255–21269, 2022.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 5998–6008, 2017.

Xinxu Wei, Kanhao Zhao, Yong Jiao, Lifang He, and Yu Zhang. A brain graph foundation model: Pre-training and prompt-tuning across broad atlases and disorders. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2506.02044.

Mitchell Wortsman, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Replacing softmax with ReLU in vision transformers. *arXiv preprint arXiv:2309.08586*, 2023.

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. doi: 10.1109/TMI.2024.3414476.

A THE PRETRAINED MODEL

A.1 ARCHITECTURE AND CONFIGURATION

Our BFM is an encoder-only masked autoencoder. Two architectural choices were made specifically to render its attention weights interpretable as a connectome. First, each token corresponds to a parcellated region within one time window, rather than to a voxel or a surface vertex. Whereas connectome is normally estimated at a ROI level due to computational cost; interpreting such attention as connectivity therefore requires an additional post-hoc pooling into a region-by-region matrix, the outcome of which depends on how the pooling is performed. Second, we retained all tokens—both visible and masked—within the encoder. Conventional masked autoencoders discard the masked tokens for computational efficiency (He et al., 2022), which confines the resulting attention to the randomly selected visible subset and thus leaves its support varying across samples. Our encoder consequently computes attention over the same complete set of tokens in every sample, which, pooled over layers, heads, and time windows (Equation 5), yields a $V \times V$ matrix that can be interpreted directly as a macroscopic connectome.

Table 3: Architecture and pretraining configuration of the 10 human BFMs at Schaefer-400.

	Hyperparameter	Value
Architecture	Layers L	8
	Attention heads H	8
	Hidden dimension D	512 (head dimension 64)
	Feed-forward width	2048, SwiGLU
	Normalisation	RMSNorm with QK-normalisation
	Reconstruction head	Linear $\mathbb{R}^D \to \mathbb{R}^{P_s}$
Tokenization	Regions V	400 (Schaefer-400, cortical)
	Window length	120 s
	Patches per region P	10 (12 s each)
	Patch size P_s	17 (YA, TR 0.72 s); 15 (A/D, TR 0.8 s)
	Patch embedding	Linear $\mathbb{R}^{P_s} \to \mathbb{R}^D$, no bias
Positional encoding	Temporal	RoPE on Q/K (half of head dim, $\theta = 10^4$)
	Region identity	learnable table $E \in \mathbb{R}^{V \times D}$
Masking	Strategy	random over tokens
	Mask ratio m	0.70
Optimisation	Optimiser	AdamW
	Base learning rate	5×10^{-3}, scaled by batch size$/256$
	Schedule	cosine, warmup ratio 0.05
	Weight decay	0.05

Table 3 summarizes the full configuration. The 10 human BFMs correspond one-to-one to the 10 conditions examined in this study—REST and the 7 tasks of HCP-YA, and REST of HCP-A and HCP-D. All 10 models share this configuration and differ only in the data filter applied and in the temporal patch size P_s, which was determined by the TR of each cohort.

A.2 RECONSTRUCTION QUALITY

Before interpreting the attention weights or applying the model to downstream prediction, we verified that the encoder reconstructs the masked signal with sufficient fidelity. Specifically, we applied the identical masking procedure to held-out recordings and scored the reconstruction against the ground-truth signal using region-wise Pearson correlation (r) and mean squared error (MSE). The three steps reported below correspond to those listed in Table 1.

Table 4: Step-1 reconstruction fidelity of the encoders whose attention is scored against a tract-tracing or structural connectome. Per-region Pearson r and MSE between the reconstructed and true masked signal, mean $\pm$ s.d. across regions and held-out recordings.

Data	Atlas	V	$r \uparrow$	MSE $\downarrow$
Mouse	Allen isocortex	37	$0.815 \pm .153$	$0.291 \pm .223$
Macaque	Markov M132	40	$0.336 \pm .191$	$0.872 \pm .172$
Simulation	DK-68	68	$0.592 \pm .156$	$0.628 \pm .174$
	Glasser-360	360	$0.452 \pm .145$	$0.780 \pm .138$
	Schaefer-400	400	$0.451 \pm .145$	$0.781 \pm .137$

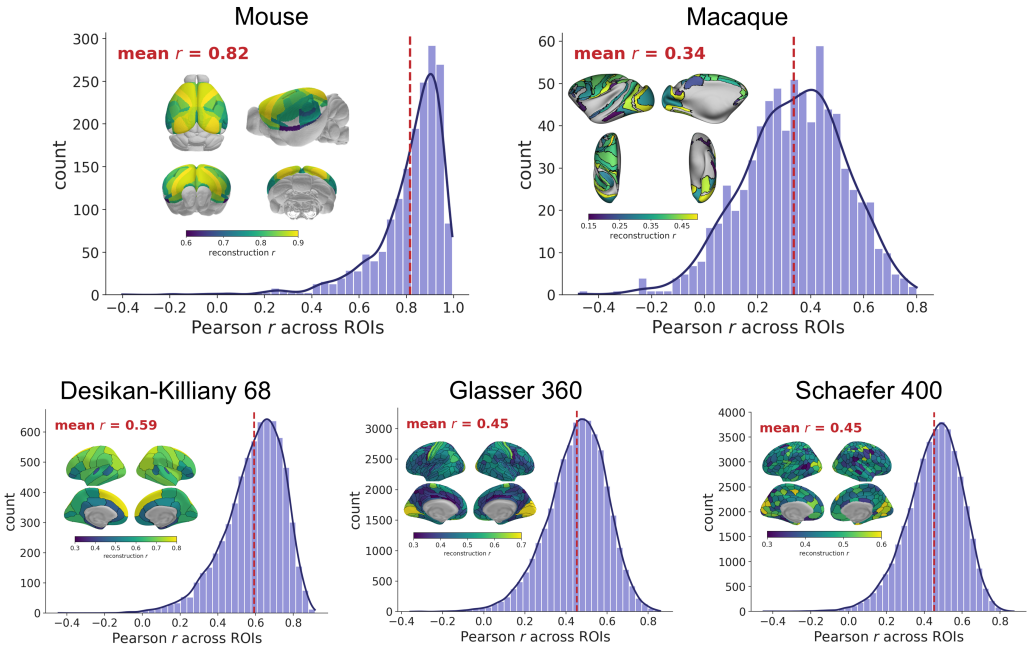

Figure 6: **Step-1 reconstruction, resolved by region.** Distribution of the Pearson r between the reconstructed and the true masked signal, taken per region and pooled over held-out recordings, for mouse, macaque, and the simulations driven by the undirected structural connectome at three atlases. The dashed line is the mean over regions, printed in each panel, and equals the value in Table 4; the inset maps the same per-region value onto the cortex.

Step-1: Structural backbone. Among the evaluated settings, the mouse data yielded the highest reconstruction accuracy ($r = 0.82$). The human simulations reached $r = 0.59$ at Desikan–Killiany 68 and $r = 0.45$ at both Glasser 360 and Schaefer 400. The macaque setting showed the lowest accuracy ($r = 0.34$; Table 4), which we attribute primarily to data scarcity: it was evaluated leave-one-out on only 19 anesthetized animals, leaving 18 pretraining recordings per model. At the regional level (Figure 6), the mouse distribution was sharply left-skewed, with most regions exceeding $r = 0.8$, whereas the macaque scores were broadly dispersed. The human simulations were tightly clustered around the mean at all three atlases.

Table 5: Step-2 reconstruction fidelity of the encoders pretrained on simulations driven by a rewired directed connectome. Format as in Table 4.

Atlas	V	$r \uparrow$	MSE $\downarrow$
DK-68	68	$0.575 \pm .161$	$0.646 \pm .175$
Glasser-360	360	$0.386 \pm .146$	$0.837 \pm .130$
Schaefer-400	400	$0.377 \pm .145$	$0.845 \pm .126$

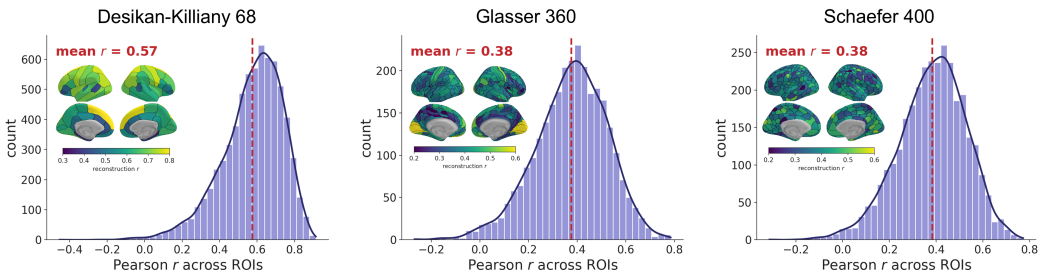

Figure 7: **Step-2 reconstruction, resolved by region.** As Figure 6, on the simulations driven by a rewired directed connectome. The dashed line equals the value in Table 5.

Step-2: directionality. The encoders pretrained on the rewired directed simulations reconstructed the masked signal slightly less accurately than their undirected counterparts at every atlas, with r lower by 0.02 to 0.07 (Table 5 vs. Table 4). At the regional level (Figure 7), the distributions were comparable across the three atlases, with the majority of regions clustered tightly around the mean. This modest decrease indicates that the rewiring did not impair the encoders' capacity to model the BOLD signal, so the direction results of Section 4.1 are not attributable to a failure of pretraining on these simulations.

Table 6: Step-3 reconstruction fidelity of the ten human encoders at Schaefer-400. Format as in Table 4.

Condition	Cohort / Task	$r \uparrow$	MSE $\downarrow$
REST	HCP-D	$0.720 \pm .191$	$0.442 \pm .234$
	HCP-YA	$0.595 \pm .205$	$0.603 \pm .229$
	HCP-A	$0.572 \pm .229$	$0.619 \pm .249$
Task (HCP-YA)	LANGUAGE	$0.686 \pm .174$	$0.497 \pm .218$
	WM	$0.675 \pm .182$	$0.504 \pm .226$
	SOCIAL	$0.705 \pm .176$	$0.469 \pm .223$
	RELATIONAL	$0.671 \pm .189$	$0.509 \pm .233$
	MOTOR	$0.706 \pm .165$	$0.470 \pm .213$
	GAMBLING	$0.655 \pm .183$	$0.535 \pm .220$
	EMOTION	$0.652 \pm .181$	$0.538 \pm .217$

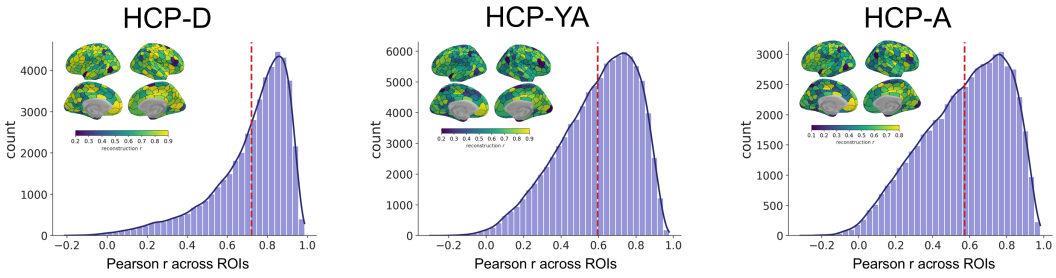

Figure 8: **Reconstruction of resting-state fMRI across the lifespan.** Distribution of the Pearson r between the reconstructed and the true masked signal, taken per region and pooled over held-out subjects, for REST of HCP-D, HCP-YA, and HCP-A (ordered by age) at Schaefer-400. The dashed line marks the mean over regions and equals the value in Table 6; the inset maps the per-region value onto the cortex. Color scales differ across panels.

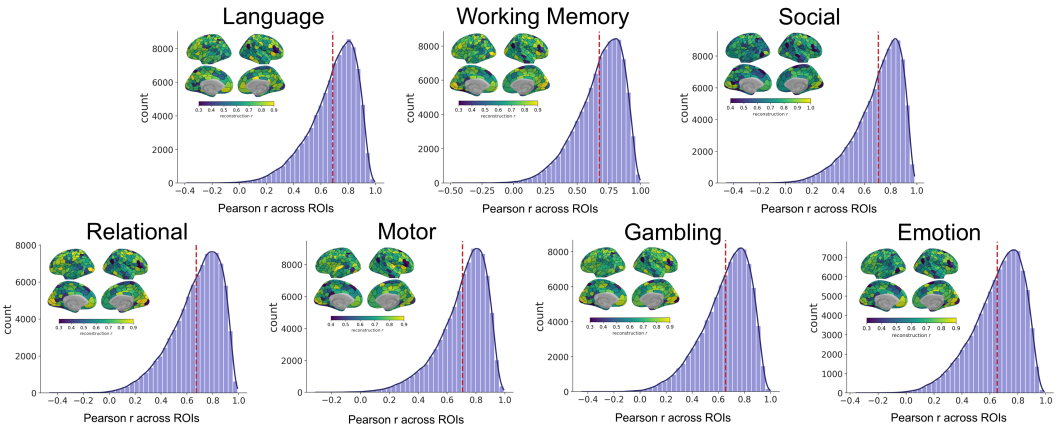

Figure 9: **Reconstruction of task fMRI in HCP-YA.** As Figure 8, for each of the 7 task conditions of HCP-YA.

Step-3: dynamics. Reconstruction accuracy for the REST condition ranged from $r = 0.57$ to 0.72 across the three lifespan cohorts, decreasing with age from HCP-D to HCP-A (Table 6; Figure 8). In HCP-YA, every task-trained encoder ($r = 0.65$–0.71) exceeded the REST encoder of the same cohort ($r = 0.60$), and at the regional level the distributions of all task conditions were shifted toward higher r than that of REST (Figure 9).

B Validation data and metrics

B.1 Data

Table 7: Validation data for the three steps.

	Step-1: backbone			Step-2: direction	Step-3: dynamics
	Mouse	Macaque	Simulation (symmetric)	Simulation (rewired)	HCP
Source	Grandjean	PRIME-DE	Hopf model	Hopf model	HCP-YA / -A / -D
Subjects	253	19	1100	1100	1093 / 722 / 630
Split	202 / 51	leave-one-out	1000 / 100	1000 / 100	80:20 per cohort
Atlas	Allen CCF (isocortex)	M132	DK-68 / Glasser-360 / Schaefer-400	DK-68 / Glasser-360 / Schaefer-400	Schaefer-400
Regions V	37	40	68 / 360 / 400	68 / 360 / 400	400
TR	1.0 s	1.6 s	0.72 s	0.72 s	0.72 / 0.8 s
Run length	$\sim$1000 TR	490 TR	1200 TR	1200 TR	1200 / 488 TR (REST)
Patch size P_s	10	8	17	17	17 / 15
Ground truth	Allen tracing	FLN tracing	SC	rewired SC	none
Evaluated by		direct r, F1, AUC		direction r, d-acc	FC, FCD, TD

Mouse. We used resting-state fMRI from the Grandjean multi-center dataset (Grandjean et al., 2020), namely 253 usable recordings parcellated into the 37 isocortical regions of the Allen Common Coordinate Framework (TR = 1.0 s). The ground truth was the ipsilateral directed weight matrix of the Allen Mouse Connectivity Atlas (Oh et al., 2014). The recordings were randomly split into 202 for pretraining and 51 held out, and all reported evaluations were performed on the held-out recordings.

Macaque. We used parcellated timeseries of 19 anesthetized macaques from PRIME-DE (Milham et al., 2018) in the 40 cortical regions of the Markov M132 atlas (TR = 1.6 s, 490 volumes per animal). The ground truth was the directed fraction-of-labeled-neurons matrix obtained by retrograde tract-tracing (Markov et al., 2014). Because this sample was too small to be partitioned, we evaluated the attention in a leave-one-out manner: each animal was scored by a model pretrained on the remaining 18.

Simulation. Since invasive tract-tracing is not feasible in humans, we simulated BOLD signals from a connectome with known ground-truth structure. Each region j was modeled as a Hopf

(Stuart–Landau) oscillator model with intrinsic frequency ω_j, coupled through the connectome C:

$$\frac{\mathrm{d}x_j}{\mathrm{d}t} = \left(a_j - x_j^2 - y_j^2\right) x_j - \omega_j y_j + G \sum_k C_{jk} \left(x_k - x_j\right) + \sigma \eta_j(t), \tag{8}$$

$$\frac{\mathrm{d}y_j}{\mathrm{d}t} = \left(a_j - x_j^2 - y_j^2\right) y_j + \omega_j x_j + G \sum_k C_{jk} \left(y_k - y_j\right) + \sigma \xi_j(t), \tag{9}$$

where η_j and ξ_j are independent Gaussian white noise and x_j is the simulated BOLD. All regions sit just below the Hopf bifurcation, so oscillations are noise-driven rather than self-sustained; parameters are in Table 8. Each subject was simulated with its own noise seed, giving 1000 pretraining and 100 held-out subjects per atlas. The two steps differ only in C:

- **Step-1: backbone.** C was the symmetric group-average structural connectome provided by the ENIGMA Toolbox (Larivière et al., 2021) at each of the three atlases. As C is symmetric, only the presence of each connection, not its orientation, could be evaluated.
- **Step-2: direction.** C was the same connectome in which the orientation of approximately 20% of edges was reassigned by degree-preserving randomization (Rubinov & Sporns, 2010), drawn independently for each subject. The set of connected region pairs was thus unchanged, while their orientations differed across subjects and were known for each.

Table 8: Hopf simulation parameters. The global coupling G was fitted per atlas, as described in the text.

	Parameter	Value
Dynamics	Bifurcation a_j	-0.01
	Noise σ	0.01
	Intrinsic frequency f_j	$\mathcal{U}(0.04, 0.07)$ Hz, per region
Coupling G	DK-68	3.0
	Glasser-360	6.0
	Schaefer-400	1.0
Integration	Scheme	Euler–Maruyama
	Step $\mathrm{d}t$	0.036 s
	Burn-in	1000 s
	Output	1200 volumes at TR 0.72 s

Empirical. Since empirical fMRI provides no ground-truth connectome, in Step-3 we evaluated each estimated connectome by the dynamics it generates: the connectome estimated from a held-out subject was used to drive the whole-brain model, and the simulated signal was compared with that subject's recording. We used resting-state fMRI of HCP-D, HCP-YA, and HCP-A at Schaefer-400, each split 80:20 into pretraining and held-out subjects (Section 3.3).

B.2 Metrics

Let M denote an estimated connectome, M^* the ground truth, and $\mathrm{asym}(M) = (M - M^\top)/2$ the antisymmetric part of M. All backbone and direction metrics were computed on the off-diagonal entries.

Step-1: backbone. These metrics assess whether an estimator identifies which regions are connected.

- **Direct r.** The Pearson correlation between M and M^*, reflecting agreement in edge *weight*.
- **F1.** The F1 score between M and M^* after thresholding both at the ground-truth edge density, reflecting agreement in the *edge set* regardless of weight.
- **AUC.** The area under the ROC curve obtained by ranking M_{ij} over all ordered pairs against the binarized ground truth; unlike F1, it requires no threshold.

Step-2: direction. These metrics assess whether an estimator identifies which of two connected regions drives the other.

- **Direction r.** The Pearson correlation between $\mathrm{asym}(M)$ and $\mathrm{asym}(M^*)$.
- **d-accuracy.** For each ground-truth connection, the sign of $M_{ij} - M_{ji}$ was compared with that of $M^*_{ij} - M^*_{ji}$, and the metric is the fraction of connections whose sign agrees; chance is 0.5 (Smith et al., 2011).
- **Asymmetry index.** $e_A/(e_S + e_A)$, where e_A and e_S are the off-diagonal sums of squares of $\mathrm{asym}(M)$ and of the mean-centered symmetric part $(M + M^\top)/2$; centering prevents a uniform offset, as in softmax attention, from inflating the symmetric share.

Step-3: dynamics. These metrics assess whether an estimated connectome reproduces the recording from which it was obtained.

- **FC r.** The Pearson correlation between the upper triangles of the simulated and empirical FC, testing whether the connectome reproduces the static inter-regional correlation structure.
- **FCD KS.** A sliding window of 60 s was applied to each recording to obtain a sequence of FC matrices, and FCD_{uv} was defined as the correlation between the FC matrices of windows u and v. The metric is the Kolmogorov–Smirnov distance $\sup_z |F^{\mathrm{sim}}(z) - F^{\mathrm{emp}}(z)|$ between the simulated and empirical distributions of FCD_{uv}, where F denotes the cumulative distribution function. It tests whether the correlation structure fluctuates over time as it does empirically; lower is better.
- **TD r.** For each pair of regions, τ_{ij} was defined as the lag at which their cross-correlation peaks, and its average over j, $\tau_i = \frac{1}{V}\sum_j \tau_{ij}$, indexes how early or late region i's signal arrives relative to the rest of the brain. The metric is the Pearson correlation between the simulated and empirical τ_i across regions.

For parameter selection, such as the γ sweep in Appendix D, we used the goodness of fit $\mathrm{GOF} =$ FC $r -$ FCD KS (Oh et al., 2026).

C Attention as a learned connectome

In AXON, the pooled attention A (Equation 5) serves as the prior that determines how strongly each connection is regularized (Equation 7). Here, we analyzed in detail how this attention represents the connectome: when the representation arises during pretraining, which component of the brain token it is built from, how it is organized across the layers of the encoder, and why it carries no directional information.

C.1 EMERGENCE OF THE CONNECTOME AND ITS BASIS IN THE REGION EMBEDDING

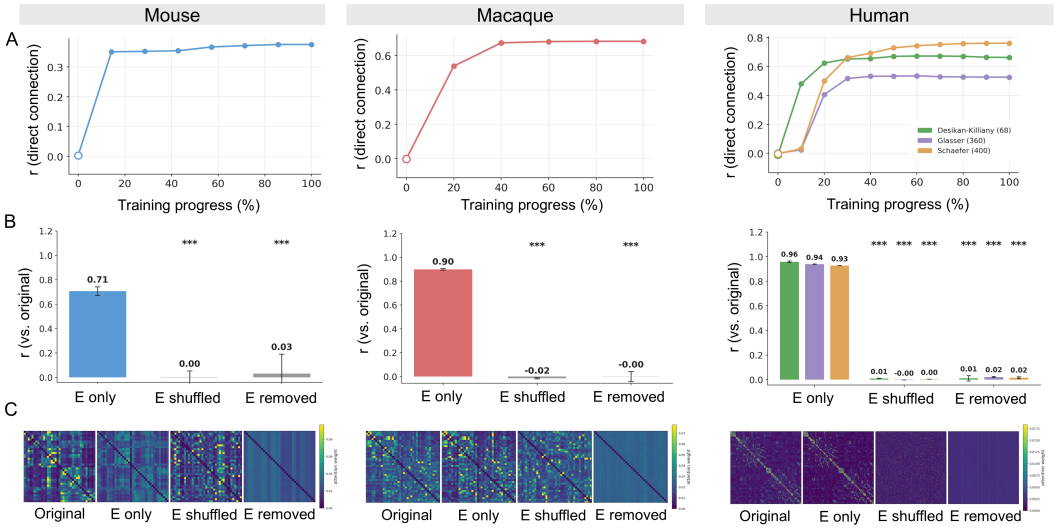

Figure 10: **The connectome emerges during pretraining and is carried by the region embedding. (A)** Direct r between the pooled attention and the ground-truth connectome over pretraining, for each species and atlas. Open circles, initialization. **(B)** Correlation between the original pooled attention and the attention obtained at inference when the model receives the region embedding E only (signal ablated), E shuffled across regions, or E removed. Bars are mean $\pm$ s.d. across subjects; *** marks a difference from the E-only condition at $p < 0.001$. **(C)** The corresponding matrices for one subject (human, Schaefer-400). Removing E leaves a near-diagonal matrix, an out-of-distribution input rather than a matched control.

Emergence during pretraining. At initialization, the direct r between the attention and the ground truth was near zero in all settings; it rose steadily over pretraining and converged before training ended (Figure 10A). The connectome in the attention is therefore not an inductive bias of the architecture but a representation learned from the data.

Basis in the region embedding. Each brain token combines the BOLD signal of a region within a time window with the region's learnable embedding E (Equation 2). Ablating each at inference showed that when the model received E only, with the signal ablated, the attention remained largely intact ($r = 0.71$–0.96 with the original across species and atlases), whereas shuffling E across regions rendered it uncorrelated with the original (Figure 10B,C). This suggests a division of roles within the brain token: the region embedding determines which regions the model draws on when reconstructing a given one, whereas the signal supplies the content to be drawn from those regions. Because the region embedding is shared across subjects, the attention remains stable across them, allowing it to serve as a prior over inter-regional connections in Equation 7.

C.2 LAYER-WISE ASSEMBLY AND SYMMETRY OF THE CONNECTOME

Table 9: Direct r to the ground-truth connectome by encoder layer, mean over subjects. **Bold**, best per row; *, targets where the uniform pooling of Equation 5 scores above every individual layer. ρ is the Spearman correlation between layer index and direct r across the eight layers, significant at $p < 0.05$ everywhere except mouse.

Target	L0	L1	L2	L3	L4	L5	L6	L7	Eq. 5	ρ
Mouse	0.387	0.435	0.372	0.374	0.388	**0.436**	0.429	0.404	0.432	+0.38
Macaque	0.432	0.581	0.543	0.532	0.624	0.673	0.685	0.684	**0.739***	+0.88
DK-68	0.633	0.642	0.691	0.742	0.770	0.793	0.760	0.757	**0.808***	+0.79
Glasser-360	0.238	0.221	0.578	0.737	0.787	0.833	0.838	0.824	**0.850***	+0.90
Schaefer-400	0.028	0.716	0.836	0.842	0.845	0.850	0.846	0.848	**0.871***	+0.93

Since the pooled attention of Equation 5 averages over all encoder layers, we examined how the connectome is distributed across them by evaluating the attention of each layer separately (Table 9; see Figure 3 for an individual example). Direct r was lowest in the earliest layers and increased with depth in four of the five targets (Spearman $\rho = 0.79$–0.93, $p < 0.05$), with mouse the exception. At Schaefer-400, for instance, the direct r rose from 0.028 at layer 0 to 0.716 at layer 1 and remained above 0.83 from layer 2 onward, indicating that the connectome is assembled progressively as the input passes through the encoder. One could instead select the single best layer, but this would require access to the ground truth, which is unavailable for empirical data. We therefore retained the uniform average, which in fact outperformed every individual layer in four of the five targets and was within 0.004 of the best layer in mouse.

Table 10: Symmetric energy share of the mean-centred attention operator by encoder layer, mean over held-out subjects, computed on the off-diagonal as $\|\mathrm{sym}\|_F^2/(\|\mathrm{sym}\|_F^2 + \|\mathrm{asym}\|_F^2)$.

Target	L0	L1	L2	L3	L4	L5	L6	L7	Eq. 5
Mouse	0.858	0.866	0.874	0.868	0.910	0.934	0.922	0.898	0.975
Macaque	0.628	0.728	0.680	0.671	0.756	0.830	0.825	0.827	0.932
DK-68	0.776	0.772	0.879	0.901	0.917	0.955	0.970	0.970	0.980
Glasser-360	0.551	0.559	0.787	0.832	0.912	0.961	0.977	0.991	0.991
Schaefer-400	0.515	0.861	0.952	0.923	0.947	0.982	0.974	0.986	0.992

The layers that best recovered the connectome were also the most symmetric. The fraction of the attention's energy in its symmetric part increased with depth in all targets (Table 10; Figure 3), from about 0.50 at layer 0 in the finer atlases to above 0.97 at the final layer, and was higher still in the uniform average. Antisymmetric components, prominent in the early layers, thus diminished with depth, leaving the pooled attention without directional information (Section 4.1).

C.3 Neither the mask nor the activation carries direction

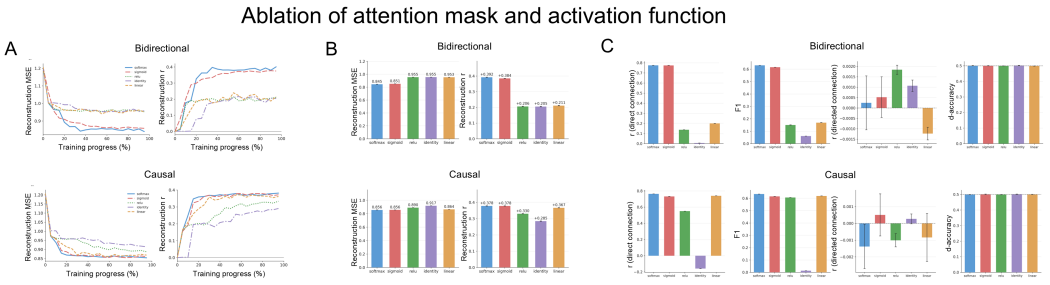

Figure 11: **Neither the attention mask nor the activation carries directionality.** 10 encoders pretrained on the Step-2 simulations at Schaefer-400, crossing two attention masks (top, bidirectional; bottom, causal over the patch index) with five activations applied to the attention scores. **(A)** Reconstruction MSE and r over pretraining. **(B)** The same metrics at convergence. **(C)** Recovery of direct connections (direct r, F1) and their directionality (direction r, d-accuracy) against the ground-truth connectome. Error bars, s.d. across held-out subjects; values in Table 11.

Table 11: Attention mask and activation ablation on the Step-2 simulations at Schaefer-400. Ten encoders differ only in the mask and the activation applied to the attention scores; all else as in Table 3. Chance is 0.5 for d-accuracy and 0 for direction r.

		Reconstruction		Direct connections		Directionality	
Mask	Activation	$r \uparrow$	MSE$\downarrow$	direct $r \uparrow$	F1$\uparrow$	direction r	d-acc.
Bidirectional	softmax	0.392	0.845	0.775	0.630	$+0.0003$	0.501
	sigmoid	0.384	0.851	0.775	0.616	$+0.0005$	0.500
	ReLU	0.206	0.955	0.137	0.150	$+0.0018$	0.499
	identity	0.205	0.955	0.006	0.061	$+0.0011$	0.502
	linear	0.211	0.953	0.200	0.169	-0.0012	0.499
Causal	softmax	0.378	0.856	0.763	0.634	-0.0014	0.499
	sigmoid	0.378	0.856	0.732	0.615	$+0.0005$	0.501
	ReLU	0.330	0.890	0.551	0.606	-0.0010	0.499
	identity	0.285	0.917	-0.159	0.013	$+0.0003$	0.502
	linear	0.367	0.864	0.739	0.619	-0.0008	0.500

Two further design choices could account for the direction-blindness of the attention (Section 4.1). First, the encoder is bidirectional, so temporal order is never enforced by the objective; second, the softmax activation constrains each row of the attention to be non-negative and to sum to one. To test whether either choice suppresses direction, we pretrained 10 encoders on the Step-2 simulations, crossing two attention masks (bidirectional and causal over the patch index) with five activation functions applied to the attention scores: softmax (Vaswani et al., 2017); sigmoid (Ramapuram et al., 2025) and ReLU (Wortsman et al., 2023), which act element-wise without normalizing each row; linear attention (Katharopoulos et al., 2020), which replaces the exponential similarity with a dot product of positive feature maps; and identity, which uses the raw scaled scores as in NetFormer (Lu et al., 2025).

None of the 10 configurations recovered edge direction (Table 11): d-accuracy remained at chance (0.499–0.502) and direction r near zero (≤ 0.002). In contrast, direct r varied widely across configurations, from 0.775 to -0.159. The mask and activation thus determined how well the connections were recovered, but not their orientation. This failure was not due to the causal mask degrading the model: the causal softmax encoder reconstructed the BOLD signal nearly as well as its bidirectional counterpart ($r = 0.378$ vs. 0.392) and recovered the connectome, indicating that what it lacked was an objective that orients the edges rather than the capacity to model the signal.

A causal mask alone is therefore insufficient to make the attention directional. Recovering direction would likely require a next-timestep prediction objective, in which each timepoint is predicted from the preceding one as in Equation 6, and hence an autoregressive, decoder-only architecture; we leave this to future work.

D AXON

Parameters. With λ_0 and δ fixed (Table 12), γ was selected for each cohort by maximizing the GOF (Appendix B.2) on the pretraining subjects, yielding 2.0 for HCP-D, 1.0 for HCP-YA, and 1.5 for HCP-A; the downstream evaluations used the HCP-YA value. Since $\gamma = 0$ reduces AXON to VAR at the same λ_0, any difference between the two arises solely from the attention prior.

Table 12: AXON parameter settings for the generative validation.

	Value
δ	10^{-3}
λ_0	10^4
γ	$\{0.5, 1, 1.5, 2\}$; selected D 2.0, YA 1.0, A 1.5

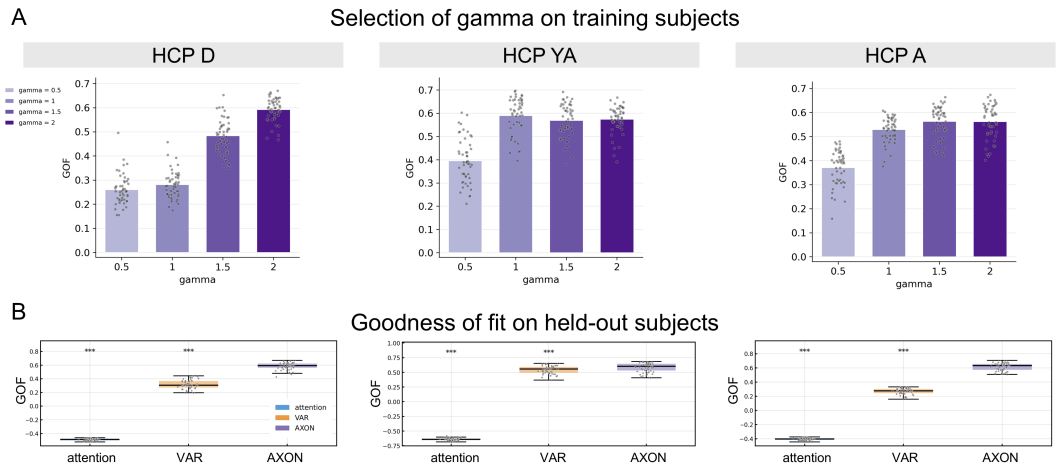

Figure 12: **Goodness of fit across cohorts. (A)** Selection of the prior exponent γ: GOF of AXON as a function of γ on the pretraining subjects. **(B)** GOF of the attention, VAR, and AXON on the held-out subjects at the selected γ. Points are subjects; stars mark a difference from AXON (Wilcoxon signed-rank test, $^{***}p < 0.001$).

Goodness of fit. At the selected γ, AXON achieved the highest GOF on the held-out subjects in every cohort, significantly exceeding both VAR and the attention (Figure 12B).

E Downstream prediction

This appendix reports the downstream results in full: demographics, brain-state decoding, each of the 58 behavioral measures, and the task specificity of task-derived EC.

E.1 Demographics

Table 13: Demographic prediction on HCP-YA. Mean ± s.d. over ten folds; **best** and second per column. †Significantly lower than AXON (paired test, Holm–Bonferroni corrected, $p < 0.05$).

		Age	Sex	
	Methods	r ↑	AUROC↑	ACC↑
Conn	FC	0.458† ±.054	0.972† ±.022	0.924† ±.039
	VAR	0.505 ±.068	0.985† ±.013	0.944† ±.029
DL	BrainNetCNN	0.411† ±.067	0.986† ±.014	0.940† ±.027
	BrainGNN	0.279† ±.067	0.875† ±.037	0.792† ±.040
	BolT	0.283† ±.066	0.899† ±.025	0.812† ±.036
	BNT	0.222† ±.131	0.873† ±.152	0.793† ±.200
BFM	SwiFT	0.341† ±.075	0.944† ±.016	0.886† ±.031
	BrainLM	0.291† ±.081	0.744† ±.034	0.677† ±.047
	BrainMass	0.325† ±.057	0.937† ±.032	0.869† ±.031
	Brain-JEPA	0.296† ±.052	0.737† ±.048	0.665† ±.041
	BrainSN	0.263† ±.077	0.821† ±.047	0.748† ±.053
	BrainGFM	0.377† ±.094	0.959† ±.023	0.906† ±.039
	NeuroSTORM	0.343† ±.130	0.941† ±.013	0.867† ±.021
	BrainHarmonix	0.359† ±.083	0.981† ±.010	0.941† ±.028
	CortexMAE-P	0.384† ±.066	0.897† ±.036	0.821† ±.029
	CortexMAE-F	0.283† ±.094	0.939† ±.021	0.874† ±.029
	CortexMAE-V	0.449† ±.072	0.984† ±.010	0.933† ±.019
	Brain-Semantoks	0.417† ±.068	0.951† ±.024	0.884† ±.029
	Omni-fMRI	0.195† ±.091	0.899† ±.030	0.819† ±.040
	Brain-DiT	0.295† ±.073	0.835† ±.032	0.761† ±.030
Ours	GAP	0.385† ±.044	0.867† ±.023	0.779† ±.026
	ATTN	0.398† ±.048	0.982† ±.007	0.928† ±.012
	AXON	**0.524** ±.056	**0.991** ±.008	**0.958** ±.019

AXON achieved the highest prediction accuracy for both age ($r = 0.524$) and sex (AUROC 0.991, accuracy 95.8%; Table 13). For age, AXON significantly exceeded all methods except VAR ($r = 0.505$), including FC (0.458) and the best pretrained BFM, CortexMAE-V (0.449). For sex, AXON significantly exceeded every method, including VAR. Directed connectivity estimated from the timeseries thus outperformed correlation-based FC, and the attention prior added a further gain for sex.

E.2 BRAIN-STATE DECODING

Table 14: Brain-state decoding on HCP-YA (eight classes: rest and seven tasks; chance 0.125). Mean $\pm$ s.d. over ten subject-wise splits; **best** and second per column. †Significantly lower than AXON (Wilcoxon signed-rank test, Holm-corrected within each metric, $p < 0.05$).

	Methods	Bal. acc↑	Acc↑	F1↑	AUC↑
Conn	FC	$0.969^{\dagger} \pm .006$	$0.972^{\dagger} \pm .005$	$0.969^{\dagger} \pm .006$	$0.9988^{\dagger} \pm .0004$
	VAR	$0.996 \pm .002$	$0.996 \pm .002$	$0.996 \pm .002$	$\mathbf{1.0000} \pm .0000$
DL	BrainNetCNN	$0.977^{\dagger} \pm .007$	$0.977^{\dagger} \pm .007$	$0.977^{\dagger} \pm .007$	$0.9991^{\dagger} \pm .0004$
	BrainGNN	$0.905^{\dagger} \pm .021$	$0.905^{\dagger} \pm .020$	$0.905^{\dagger} \pm .020$	$0.9938^{\dagger} \pm .0019$
	BolT	$0.995 \pm .003$	$0.995 \pm .003$	$0.995 \pm .003$	$\mathbf{1.0000} \pm .0001$
	BNT	$0.951^{\dagger} \pm .028$	$0.951^{\dagger} \pm .028$	$0.951^{\dagger} \pm .029$	$0.9978^{\dagger} \pm .0019$
BFM	SwiFT	$0.196^{\dagger} \pm .016$	$0.196^{\dagger} \pm .016$	$0.184^{\dagger} \pm .017$	$0.5962^{\dagger} \pm .0134$
	BrainLM	$0.862^{\dagger} \pm .010$	$0.862^{\dagger} \pm .011$	$0.862^{\dagger} \pm .011$	$0.9816^{\dagger} \pm .0034$
	BrainMass	$0.931^{\dagger} \pm .007$	$0.931^{\dagger} \pm .007$	$0.931^{\dagger} \pm .007$	$0.9947^{\dagger} \pm .0006$
	Brain-JEPA	$0.784^{\dagger} \pm .012$	$0.784^{\dagger} \pm .012$	$0.783^{\dagger} \pm .012$	$0.9635^{\dagger} \pm .0048$
	BrainSN	$0.908^{\dagger} \pm .008$	$0.908^{\dagger} \pm .008$	$0.908^{\dagger} \pm .008$	$0.9913^{\dagger} \pm .0013$
	BrainGFM	$0.988^{\dagger} \pm .004$	$0.988^{\dagger} \pm .004$	$0.988^{\dagger} \pm .004$	$0.9994^{\dagger} \pm .0004$
	NeuroSTORM	$0.204^{\dagger} \pm .008$	$0.204^{\dagger} \pm .008$	$0.188^{\dagger} \pm .008$	$0.6094^{\dagger} \pm .0077$
	BrainHarmonix	$0.842^{\dagger} \pm .014$	$0.841^{\dagger} \pm .014$	$0.842^{\dagger} \pm .014$	$0.9808^{\dagger} \pm .0025$
	CortexMAE-P	$0.957^{\dagger} \pm .006$	$0.958^{\dagger} \pm .006$	$0.958^{\dagger} \pm .006$	$0.9979^{\dagger} \pm .0005$
	CortexMAE-F	$0.755^{\dagger} \pm .017$	$0.756^{\dagger} \pm .017$	$0.755^{\dagger} \pm .017$	$0.9582^{\dagger} \pm .0051$
	CortexMAE-V	$0.590^{\dagger} \pm .011$	$0.591^{\dagger} \pm .011$	$0.587^{\dagger} \pm .012$	$0.9001^{\dagger} \pm .0040$
	Brain-Semantoks	$0.963^{\dagger} \pm .007$	$0.963^{\dagger} \pm .007$	$0.963^{\dagger} \pm .007$	$0.9985^{\dagger} \pm .0006$
	Omni-fMRI	$0.167^{\dagger} \pm .009$	$0.168^{\dagger} \pm .009$	$0.153^{\dagger} \pm .009$	$0.5608^{\dagger} \pm .0101$
	Brain-DiT	$0.925^{\dagger} \pm .010$	$0.925^{\dagger} \pm .010$	$0.925^{\dagger} \pm .010$	$0.9947^{\dagger} \pm .0015$
Ours	GAP	$0.919^{\dagger} \pm .011$	$0.919^{\dagger} \pm .010$	$0.919^{\dagger} \pm .011$	$0.9928^{\dagger} \pm .0023$
	ATTN	$0.983^{\dagger} \pm .003$	$0.983^{\dagger} \pm .003$	$0.983^{\dagger} \pm .003$	$0.9996^{\dagger} \pm .0002$
	AXON	$\mathbf{0.996} \pm .002$	$\mathbf{0.997} \pm .002$	$\mathbf{0.996} \pm .002$	$\mathbf{1.0000} \pm .0000$

AXON achieved the highest balanced accuracy in classifying the 8 brain states (0.996; chance 0.125; Table 14), significantly exceeding all other methods except VAR (0.996) and BolT (0.995), from which it did not differ.

E.3 PER-MEASURE PREDICTION



Figure 13: **Per-measure prediction on HCP-YA.** Each panel is one of the 58 behavioral measures, with every method ranked by its grand-average Pearson r and AXON marked with a star. Error bars are s.d. over 10 folds.

For the 58 individual behavioral traits, AXON achieved the highest prediction accuracy on 35, followed by VAR on 6; no other method ranked first on more than 4 (Figure 13; Figure 5A). Averaged over all measures, AXON ($r = 0.210$) exceeded VAR (0.197), FC (0.169), and the best pretrained BFM, Brain-Semantoks (0.153; Table 2).

E.4 TASK SPECIFICITY OF TASK-DERIVED EC

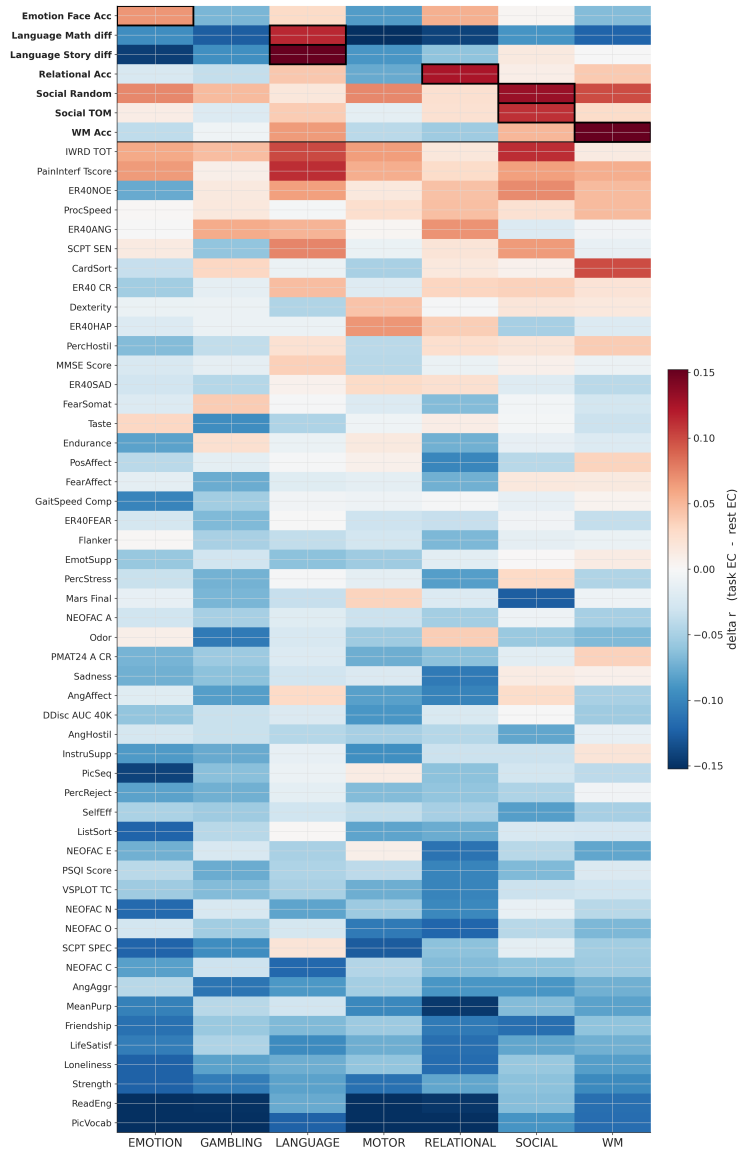

Figure 14: **Task specificity of task-derived EC.** Difference in prediction accuracy (Δr: Pearson r with task-EC minus that with rest-EC) for each of the 58 behavioral measures (rows), using AXON-derived EC estimated from each of the 7 HCP-YA tasks (columns). The top 7 rows are in-scanner performance measures; black boxes pair each measure with the task in which it was acquired. Red, task-EC better; blue, rest-EC better.

We further asked whether the advantage of task-derived EC (Figure 5B) is specific to the task in which a measure is acquired. For the in-scanner performance measures, the gains concentrated on the matching task (Figure 14, boxes), whereas EC from other tasks yielded smaller or negative gains. In contrast, most out-of-scanner traits, notably crystallized cognition (e.g., reading and vocabulary), personality, and well-being, were better predicted from resting-state EC. Task-derived EC thus carries information specific to the ongoing task, whereas resting-state EC better reflects stable individual traits.