RETHINKING SELF-SUPERVISED LEARNING FOR INTRACRANIAL EEG: ELECTRODE-ORIGIN PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intracranial EEG (iEEG) foundation models aim to learn general neural representations that transfer across subjects, tasks, and institutions. While existing models pursue this goal using objectives adapted from other modalities, we design the pretraining target around iEEG's persistent implanted electrodes. PUPPET learns through **electrode-origin prediction**: it identifies which training electrode produced each channel's waveform across changing neural activity. Its shallow encoder, comprising **one main Transformer block and a lightweight local-context module**, achieves state-of-the-art overall iMindBench performance with frozen features and linear readouts, **reaching 0.661 Overall AUROC versus the prior leader's published 0.645**. Separate models exceed that leader's published mean using **ten minutes of training recordings (0.41% of its pretraining duration)** or **71k transferred encoder parameters (0.36% of its encoder size)**. Matched ablations show stronger transfer from electrode-origin prediction than from the tested reconstruction objectives under the same architecture. Scaling analyses show aggregate gains with more data and larger models, but the gains vary by task. These results suggest that objective design, not scale alone, matters for iEEG foundation models. Code: `https://anonymous.4open.science/r/puppet-0926/README.md`.

1 INTRODUCTION

Self-supervised foundation models learn representations from unlabeled data that can be adapted to many downstream tasks. This capability is particularly appealing for intracranial electroencephalography (iEEG), whose direct recordings from cortical and subcortical structures provide millisecond resolution, high signal-to-noise ratio, and fine spatial specificity for studying cognition and developing neural interfaces (Johnson & Knight, 2015; Kuruppu et al., 2026). Yet these recordings are difficult to bring together: clinical implants differ in electrode type, channel count, and electrode placement across brain regions, sometimes sampling only one hemisphere. Data are further divided among experimental tasks, clinical cohorts, and institutions (Chau et al., 2025; 2026). Learning representations that transfer across these settings could make individual recordings useful beyond the particular implantation or experiment in which they were collected.

Realizing this potential requires an effective self-supervised learning (SSL) objective: the task used to learn from recordings without downstream labels. Vision research has explored objectives ranging from spatial puzzles to contrastive learning, DINO self-distillation, and masked reconstruction, demonstrating the value of testing different learning signals (Noroozi & Favaro, 2016; Chen et al., 2020; Caron et al., 2021; He et al., 2022). Existing iEEG models similarly adapt reconstruction, corruption detection, and future prediction (Wang et al., 2023; Zhang et al., 2023; Chau et al., 2025; Carzaniga et al., 2026). These objectives reward different signal properties, not all equally useful for downstream decoding.

Reconstruction rewards signal recovery, whereas electrode-origin prediction rewards discrimination between persistent recording sites. We investigate whether this complementary target improves transfer.

We introduce **electrode-origin prediction**, an iEEG-specific objective based on implanted contacts that persist while neural activity changes. Each contact has a unique pretraining label shared by its recordings over time. PUPPET predicts that label from each channel's waveform; the label is

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.

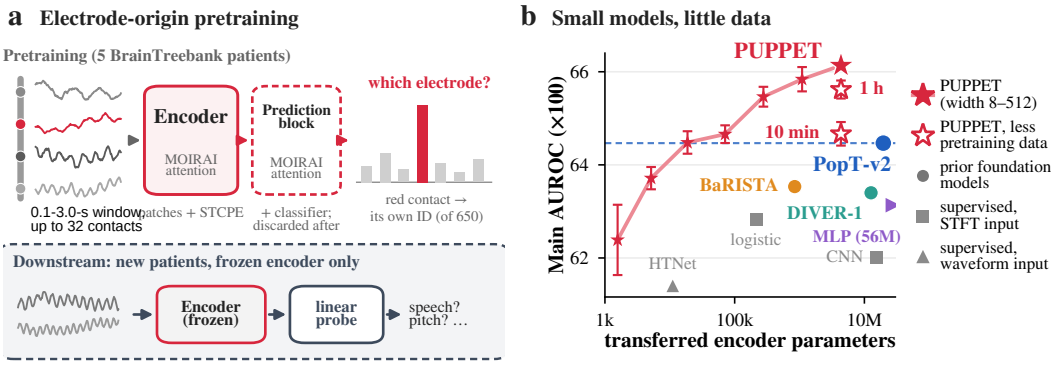


Figure 1: Electrode-origin pretraining and iMindBench transfer. (a) PUPPET's encoder (one MOIRAI-attention block with STCPE) feeds a pretraining-only prediction block and classifier identifying which of 650 training electrodes produced each waveform (illustrative traces). After pretraining, the prediction block is discarded and the frozen encoder is read out on new patients. (b) Main-subset Overall AUROC ($\times 100$) versus transferred encoder parameters. Filled/open stars vary model width/pretraining duration (three-seed mean $\pm$ SD). The 71k-parameter and ten-minute variants use 0.36% of PopT-v2's encoder parameters and 0.41% of its unique pretraining duration. Prior-model points are published scores. The multi-STFT MLP (55.7M parameters) lies beyond the plotted range and is marked at the right edge; waveform-input MLP, CNN, and logistic baselines (55–58) lie below it. All models appear in Table 1.

a target, not an input. We hypothesize that recognizing a site across changing activity encourages features useful for later decoding. After pretraining, we discard the electrode classifier and evaluate the frozen encoder on patients excluded from pretraining. With one main Transformer block and a lightweight local-context module, PUPPET reaches **0.661 Overall AUROC on iMindBench versus PopT-v2's published 0.645** (Chau et al., 2026). Separate reduced-resource variants exceed that published mean using **ten minutes of training recordings (0.41% of PopT-v2's duration)** or **71k transferred encoder parameters (0.36% of its encoder size)** (Figure 1).

Our contributions are as follows:

1. **We introduce an iEEG-specific SSL objective that enables state-of-the-art transfer even at small scale.** Electrode-origin prediction yields the highest overall iMindBench performance with a frozen one-block encoder; separate ten-minute and 71k-parameter variants also exceed the prior foundation-model leader's published mean.
2. **We assess electrode-origin supervision against reconstruction controls.** On iMindBench, electrode-origin pretraining outperforms waveform-only, Fourier-only, and DIVER-1's joint waveform–Fourier reconstruction (Han et al., 2025) **using the same encoder, pretraining data, and downstream readout**. Synthetic $1/f$ probes show more accurate oscillatory-peak readout, and cross-model peak-readout accuracy correlates with downstream performance, suggesting a possible link.
3. **We characterize how pretraining scale translates into downstream gains.** We vary model width and pretraining-data amount, comparing aggregate and task-level transfer with a joint-reconstruction control. Downstream error follows approximate power-law scaling with an irreducible term, while task-level returns vary, motivating evaluations with greater measurable headroom.

2 Related Work

Foundation models for intracranial recordings. Existing iEEG models explore several reconstruction and prediction targets. BrainBERT reconstructs masked time–frequency representations (Wang et al., 2023); Brant reconstructs masked waveform patches using temporal and spatial encoders with frequency-informed inputs (Zhang et al., 2023); and DIVER-1 combines waveform and Fourier-derived reconstruction targets (Han et al., 2025). BaRISTA reconstructs masked latent representations, varying the spatial scales used for encoding and masking (Oganesian et al., 2025). PopT

and PopT-v2 use temporal-consistency and channel-replacement detection tasks, while MVPFormer uses contrastive future-embedding prediction (Chau et al., 2025; 2026; Carzaniga et al., 2026).

Channel identity and source discrimination. Time-series research has explored using source information and channel relationships to learn better representations. TimesBERT detects replaced channels and predicts dataset identity (Zhang et al., 2025), while Channel Clustering and Channel Normalization exploit channel similarity or preserve channel distinctions for forecasting (Chen et al., 2024; Lee et al., 2025). Related work in SEEG learns functional embeddings through anatomical region-defined contrastive pairs (Javadzadeh et al., 2025). Unlike these indirect or group-level approaches, PUPPET directly predicts individual electrode identity across changing neural activity. This target fits the recording structure of iEEG: each implantation repeatedly samples a fixed set of spatially distinct sites, encouraging representations to retain site-specific signal characteristics despite temporal variation, without requiring anatomical annotations or a shared electrode layout across patients.

3 METHOD

3.1 ELECTRODE-ORIGIN PREDICTION

We form each pretraining example by selecting electrodes from one recording and taking a short time segment shared across those electrodes. A minibatch of these examples is $\mathbf{X} \in \mathbb{R}^{B\times C\times T}$, where B is the number of examples, C the number of electrode channels, and T the number of time samples. Each of the K physical electrodes pooled across all pretraining subjects has a unique class label. We denote the label of the electrode that produced channel c in example b by $y_{b,c} \in \{1, \dots, K\}$.

For channel c in example b, the model outputs $\boldsymbol{\ell}_{b,c} \in \mathbb{R}^K$, where $\ell_{b,c,k}$ is the classification score for electrode k. We minimize the average cross-entropy over all input channels in the minibatch:

$$\mathcal{L} = -\frac{1}{|\mathcal{V}|} \sum_{(b,c)\in\mathcal{V}} \log \frac{\exp(\ell_{b,c,y_{b,c}})}{\sum_{k=1}^{K} \exp(\ell_{b,c,k})}. \tag{1}$$

Here $\mathcal{V}$ is the set of input-channel indices (b, c).

3.2 MODEL ARCHITECTURE

We split each channel into non-overlapping 50-sample patches (0.1 s at 500 Hz), giving $N = T/50$ patches per channel. A linear projection maps each patch to a d-dimensional vector and produces a token grid of shape $B \times C \times N \times d$. The base model uses $d = 512$; width varies in the model-scaling experiments. Compared with DIVER-1, we use this linear patch embedding instead of a convolutional one and omit its auxiliary spectral, anatomical-position, channel-type, and register embeddings (Han et al., 2025).

Because iEEG channel numbers have no shared anatomical meaning across patients, we use DIVER-1's STCPE to add signal-dependent local temporal and cross-electrode context before the main encoder, without requiring anatomical coordinates or fixed channel positions (Han et al., 2025). The encoder and pretraining prediction block then use MOIRAI any-variate attention across electrode–time tokens: temporal RoPE represents patch order, while each attention head learns separate biases for same-electrode and cross-electrode pairs (Woo et al., 2024; Han et al., 2025). These relations are unchanged by reordering channels, so the design preserves channel-order equivariance. Appendix B details both components.

During pretraining, the encoder's output patch features pass through a Transformer prediction block. We average that block's outputs across time for each electrode to obtain $\mathbf{g}_{b,c} \in \mathbb{R}^d$, which a linear classifier maps to the electrode scores $\boldsymbol{\ell}_{b,c}$ in Equation 1. After pretraining, we discard the prediction block and classifier and freeze the encoder. For downstream decoding, we concatenate the encoder's patch features and per-electrode temporal means into a feature vector and fit a linear readout (Appendix C).[1]

[1]Encoder depth counts the main block used downstream; STCPE contains a separate auxiliary block, and the Transformer prediction block is used only during pretraining.

4 EXPERIMENTAL SETUP

4.1 PRETRAINING DATA AND PREPROCESSING

We pretrain on the BrainTreebank dataset (Wang et al., 2024), also used by the BrainBERT, PopT, PopT-v2, and BaRISTA foundation models (Wang et al., 2023; Chau et al., 2025; 2026; Oganesian et al., 2025). One of our downstream benchmarks, iMindBench, evaluates recordings from five BrainTreebank patients in its Neuroprobe tasks. To evaluate transfer to unseen patients, we exclude all recordings from these patients from pretraining, including sessions not used in the benchmark. The remaining five patients provide 650 electrode classes and 24.7 h of training recordings after the training–validation split. Appendix A lists the excluded recordings and full corpus counts.

Following DIVER-1's preprocessing pipeline (Han et al., 2025), we high-pass filter recordings at 0.5 Hz, apply notch filters at 60, 120, and 180 Hz, and resample to 500 Hz. We then apply same-shaft Laplacian re-referencing, subtracting the mean signal of the available immediately adjacent contacts. Finally, we clip amplitudes to $\pm 200\,\mu$V. Shaft-edge cases and corpus-specific preprocessing details are provided in Appendix A.

4.2 PRETRAINING PROTOCOL

We divide recordings into non-overlapping 30-s records (90% training, 10% validation). For each example, DIVER-1's Spatio-Temporal Resampling (STR) randomly selects up to 32 electrodes and a shared 0.1–3.0-s segment from one record (Han et al., 2025); padding is masked in attention, pooling, and loss. Training and validation records may share sessions, but neither contains downstream evaluation patients.

We pretrain for 64 epochs using the Muon optimizer (Jordan et al., 2024), with a batch size of 256 and weight decay of 10^{-4}, and select the epoch with the lowest validation loss. We tune the learning rate using validation loss; the values for each objective are given in Section 4.3. Appendix B provides the full configuration.

4.3 MATCHED RECONSTRUCTION CONTROLS

To isolate the effect of the pretraining objective, we compare electrode-origin prediction with DIVER-1-style joint waveform–Fourier reconstruction (Han et al., 2025) and waveform-only and Fourier-only variants. Here "Fourier" denotes the evaluated log-compressed absolute *real* rFFT component, not the full complex spectral magnitude (Appendix B).

All four objectives use the same depth-1 encoder, pretraining Transformer prediction block, corpus, sampling procedure, and training budget. The reconstruction controls replace 50% of input patches with zeros and minimize mean-squared error at those positions; electrode-origin prediction uses unmasked inputs. The joint control weights its waveform and real-component Fourier losses equally. Appendix B gives the exact transform.

The selected peak learning rate is 0.02 for electrode-origin prediction and joint reconstruction, and 0.01 for waveform-only and Fourier-only reconstruction.

4.4 DOWNSTREAM EVALUATION

We evaluate on iMindBench (Chau et al., 2026), which comprises 15 language, auditory, and visual decoding tasks on 1-s movie-watching iEEG windows from BrainTreebank-derived Neuroprobe (Zahorodnii et al., 2026; Wang et al., 2024), Bang! You're Dead (BYD) (Keles et al., 2024), and Pippi (Berezutskaya et al., 2022). These datasets include 5, 16, and 5 evaluation patients, respectively. An evaluation unit is a dataset–subject/session–task combination scored over up to two within-session folds. Main contains supported units passing the benchmark's prespecified validation-decoding screen; Challenge contains the remaining supported units, and All is their union. Published-model comparisons and ablations use Main; scaling uses All (75 Neuroprobe, 435 BYD, and 75 Pippi units). Label efficiency uses eligible All units with a scoreable 10% labeled-trial subset. For each evaluation unit, a readout is fitted on its labeled training trials while PUPPET's

encoder remains frozen; evaluation patients were unseen in encoder pretraining, but task decoding is not zero-shot. Full protocols and coverage are in Appendices D and A.

4.5 Baselines and Evaluation Protocol

We compare with iMindBench's published foundation-model and supervised baselines under their reported preprocessing and adaptation protocols (Chau et al., 2026). All are scored by AUROC, but PopT-v2 and BaRISTA fine-tune their encoders whereas DIVER-1 uses a frozen encoder with a trained head. Appendix D.2 details baseline protocols; Appendix C.1 compares frozen encoders using one logistic readout (Table 4).

Unless stated otherwise, PUPPET uses a width-512 encoder pretrained on 24.7 h of BrainTreebank, with mean $\pm$ SD over three seeds. We freeze the encoder and fit the same logistic-regression estimator used by iMindBench's multi-STFT logistic baseline (Appendix C). Test AUROC is averaged over folds and evaluation units within each dataset; Overall equally weights the three dataset means. Paired comparisons resample patients and use Holm correction (Appendix D.3).

4.6 Scaling Experiments

We independently scale model size by varying width from 8 to 1,024—widening the encoder and prediction block together—and pretraining-data duration from 1 minute to 24.7 hours. Every record-count and width run plotted in Figure 5 trains for 704 updates; the exploratory corpus-fraction rungs in Appendix H include an 808-update exception. Downstream evaluation uses the minimum-validation-loss checkpoint, which can occur earlier for small datasets. The joint waveform–Fourier reconstruction control also has model-size and data-duration ladders; it has three seeds at widths 32, 128, and 512, and one seed at its other rungs. Section 5.4 presents the completed All-subset comparisons from five minutes of PUPPET pretraining data and ten minutes of reconstruction data; shorter exploratory rungs are documented in Appendix H. We examine downstream AUROC both overall and by task.

5 Results

5.1 iMindBench transfer and efficiency

With a shallow frozen encoder and linear readouts, PUPPET reaches 0.661 Overall iMindBench AUROC versus PopT-v2's published 0.645 (Table 1). Figure 2(a) shows task-wise AUROC averaged across datasets; per-dataset scores are in Appendix Tables 6–11. In panel (b), PUPPET stands apart from the next-ranked models: its mean rank is 3.54, versus 4.88 for PopT-v2 and 4.92 for DIVER-1. Thus its 1.34-position lead over second place is much wider than the 0.05-position gap between second and third. Each baseline retains its published adaptation protocol (Chau et al., 2026). Its descriptive lead also holds on the Challenge and All subsets (Appendix Table 12).

PUPPET also performs well at smaller scale (Table 1). Using ten minutes of training recordings, it reaches 0.647 ± 0.003 Overall AUROC across three seeds, above PopT-v2's published mean. The held-out validation partition remains fixed across record-count rungs and is used for checkpoint selection (Appendix H). In a separate experiment, a width-64 encoder reaches 0.6466 ± 0.0019 with 71,222 transferred encoder parameters, 0.36% of PopT-v2's encoder size (Appendix B.1). Full scaling curves appear below.

Downstream label efficiency. With its pretrained encoder fixed, PUPPET matches the multi-STFT MLP trained with all labels using 23%, 33%, and 27% of downstream labeled training trials on Neuroprobe, BYD, and Pippi, respectively (26% Overall; Figure 3).

Table 1: Transfer on iMindBench Main. Overall equally weights Neuroprobe, BYD, and Pippi (Chau et al., 2026). Prior-model rows retain published summary scores and adaptation protocols; paired tests use released per-unit scores (Appendix D.3). PUPPET uses a frozen encoder and three-seed mean. Headers show task and patient counts. Parameter counts refer to transferred encoders or representative full supervised classifiers, and durations to unique recordings (Appendix B.1). Task-wise scores and protocols are in Appendix D.

Model	Model parameters	Pretraining data (unique duration)	Neuroprobe 15 tasks 5 patients	BYD 15 tasks 16 patients	Pippi 14 tasks 5 patients	**Overall**
Supervised baselines: raw waveform						
Logistic	54.5k	—	0.617	0.519	0.525	0.554
MLP	6.99M	—	0.653	0.532	0.534	0.573
HTNet	11.1k	—	0.693	0.597	0.552	0.614
CNN	1.06M	—	0.666	0.537	0.526	0.576
Supervised baselines: multi-STFT						
Logistic	214k	—	0.697	0.624	0.564	0.628
MLP	55.7M	—	0.702	0.625	0.567	0.631
CNN	15.4M	—	0.710	0.611	0.539	0.620
Prior pretrained iEEG models						
DIVER-1	12.7M	AJILE12 + private, 5310h	0.728	0.605	0.569	0.634
BaRISTA	842k	BTB, 30.0 h	0.716	0.632	0.558	0.636
PopT-v2	19.6M	BTB, 40.5 h	0.723	0.632	0.579	0.645
PUPPET (frozen encoder + linear probe)						
71k variant	71.2k	BTB, 24.7 h	0.716	0.636	0.587	0.647
10-min variant	4.35M	BTB, 10 min	0.720	0.638	0.582	0.647
Main model	**4.35M**	**BTB, 24.7 h**	**0.742**	**0.649**	**0.593**	**0.661**

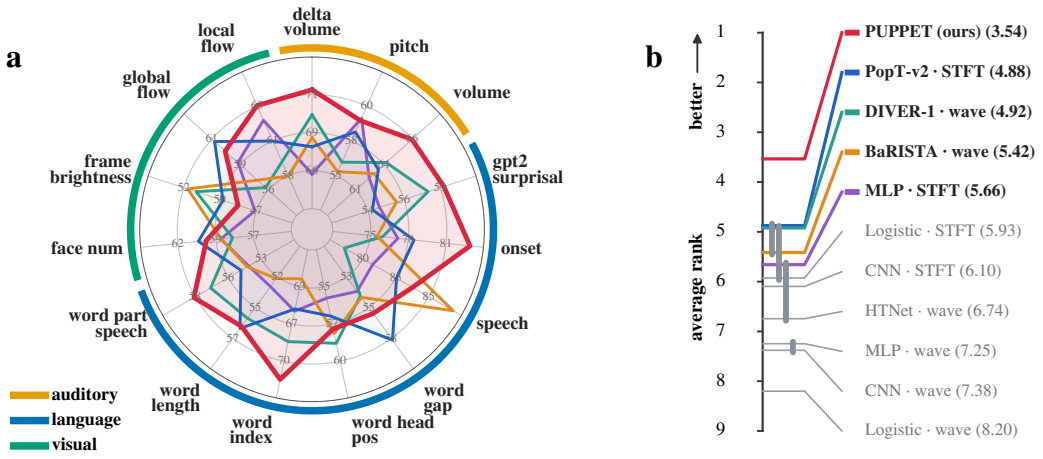


Figure 2: Task-wise transfer and average rank on iMindBench Main. (a) AUROC (×100) for 15 tasks, averaged within and then across available datasets; spoke scales differ, and arcs mark task domains. (b) Mean rank across 11 models (excluding the two small PUPPET variants), computed within evaluation units and then equally across datasets (lower is better). Grey bars mark nonsignificant Overall AUROC differences (patient-stratified bootstrap, 20,000 replicates; Holm over 55 pairs). STFT denotes multi-STFT input; wave denotes waveform input. Appendix Tables 6–11 give per-dataset task scores.

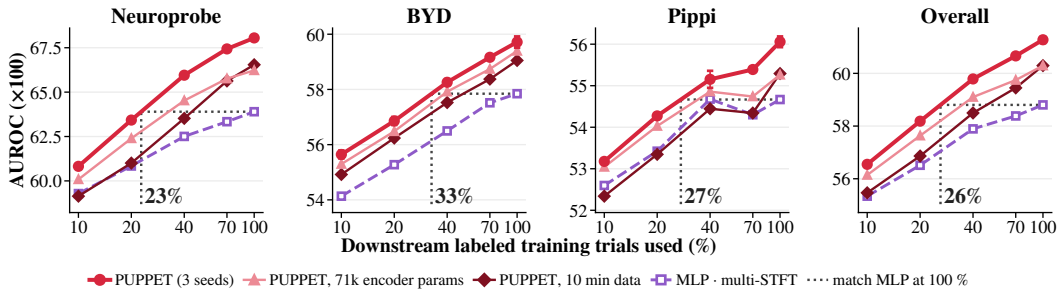


Figure 3: Downstream label efficiency. Dotted guides show how few labeled training trials PUPPET needs to match the multi-STFT MLP trained with all labels: 23% on Neuroprobe, 33% on BYD, 27% on Pippi, and 26% Overall. Percentages are relative to each evaluation unit's available training trials; both readouts use identical stratified subsets (three draws per fraction). The comparison includes 75 Neuroprobe, 435 BYD, and 50 Pippi All units scoreable for both models even at 10% of labels; lighter curves show one-seed PUPPET variants.

5.2 Pretraining-objective and architecture ablations

Under the same encoder, data, and readout protocol, matched joint waveform–Fourier reconstruction outperforms waveform-only and Fourier-only reconstruction, but electrode-origin prediction achieves the highest Overall AUROC (Table 2). Replacing MOIRAI attention with vanilla attention or removing STCPE lowers Overall scores. Together, these ablations support the benefit of electrode-origin supervision within the tested architecture and training budget; bootstrap differences and confidence intervals are in Appendix E.

Table 2: Pretraining-objective and architecture ablations on iMindBench Main. AUROC with the frozen readout of Table 1, averaged over three pretraining seeds; best/second-best values are bold/underlined. The last column gives Holm-adjusted p-values for each Overall difference from PUPPET (patient-stratified bootstrap, 20,000 replicates; 12-arm family). Differences and 95% CIs are in Appendix E.

	iMindBench Main				
Variant	Neuroprobe	BYD	Pippi	**Overall**	Holm-adjusted p
Matched pretraining objectives					
Waveform reconstruction	0.715	0.618	0.570	0.634	< 0.001
Fourier-target reconstruction	0.722	0.631	0.584	0.646	< 0.001
Joint waveform–Fourier reconstruction	0.729	0.630	0.584	0.648	< 0.001
Architecture controls					
MOIRAI attention → vanilla attention	0.733	<u>0.645</u>	0.585	0.654	< 0.001
Without STCPE	**0.743**	0.641	<u>0.588</u>	<u>0.657</u>	< 0.001
Reference model					
Electrode-origin prediction (PUPPET)	<u>0.742</u>	**0.649**	**0.593**	**0.661**	—

5.3 Information preserved by the representations

5.3.1 Spectral information in matched-objective representations

Kommineni et al. (2026) reported that reconstruction-trained EEG models recovered the $1/f$ background more accurately than higher-frequency oscillatory properties. Such higher-frequency features matter for iEEG decoding: beta- and low-gamma-band activity can contribute to speech decoding (Proix et al., 2022). We therefore probe frozen representations from electrode-origin prediction and three matched reconstruction controls using synthetic peak ranges calibrated to BrainTreebank recordings.

In this one-checkpoint probe, frozen PUPPET features predict peak center, height, and width more accurately than all three reconstruction controls. The held-out R^2 scores are 0.365, 0.329, and 0.623, compared with 0.162, 0.144, and 0.383 for joint reconstruction. Background slope and level also

remain accurately predicted ($R^2 = 0.998$ and 0.997). Appendix F.1 gives the prediction plots and methods.

5.3.2 Relating spectral recovery to downstream decoding

The matched-objective probe shows that PUPPET's representation makes oscillatory peak information more accessible. We next ask whether this information also tracks decoding on real recordings. Across twelve pretrained systems and three reconstruction controls, better prediction of synthetic peak height and width accompanies higher mean AUROC on four BYD tasks (Pearson $r = 0.57$ and 0.74, respectively; Figure 4). The relationship is positive for speech, pitch, and volume, but not word index. Although not definitive evidence of causation, this pattern is consistent with oscillatory peak information contributing to downstream decoding.

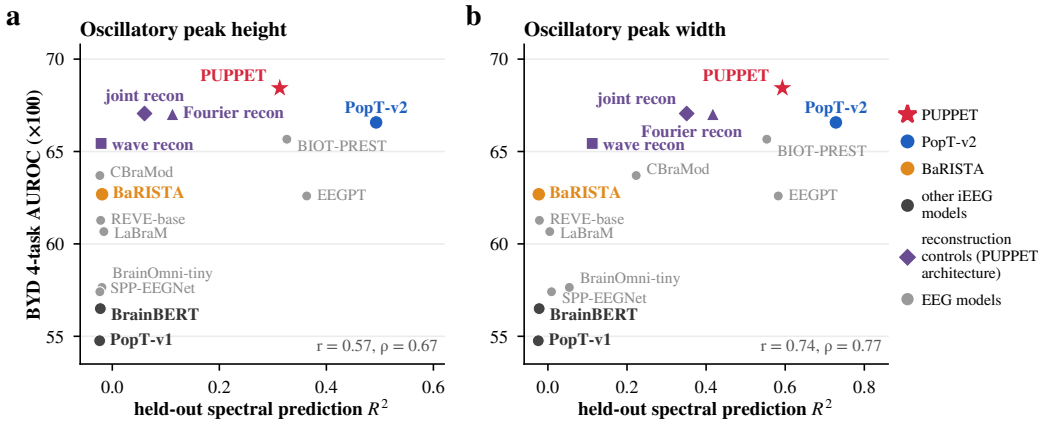


Figure 4: Spectral peak recovery and BYD decoding across fifteen frozen encoders. Panels (a) and (b) plot held-out R^2 for synthetic peak height and width against mean AUROC ($\times 100$) on four BYD tasks: speech, pitch, volume, and word index. Purple controls use PUPPET's encoder to reconstruct the masked waveform (wave recon), Fourier target (Fourier recon), or both (joint recon); other iEEG models are dark grey and EEG models light grey (SPP-EEGNet is EEG-pretrained). All systems use ridge readouts. Printed r and ρ are descriptive Pearson and Spearman correlations across systems, not evidence that peak recovery causes decoding gains. This four-task diagnostic is separate from the primary 15-task BYD benchmark; Appendix F gives input adaptations, seed coverage, and task-specific results.

5.4 Scaling pretraining and downstream performance

We ask whether electrode-origin prediction's transfer advantage changes with available pretraining data and encoder capacity. Figure 5 compares its data and width ladders with the joint reconstruction control under the same backbone and evaluation protocol. At the first shared rung shown (10 min), joint reconstruction has slightly higher Overall AUROC; by 1 h, PUPPET leads and remains ahead at the later shared rungs, while reconstruction plateaus earlier. PUPPET also scores higher at the compared model widths. Several reconstruction rungs have one seed, so these objective-wise trends remain descriptive.

To summarize each trajectory, we fit a Kaplan-style scaling law, $q(x) = E + Ax^{-\alpha}$ (Kaplan et al., 2020), separately to held-out pretraining loss and downstream error $1 - \text{AUROC}$. Here x is unique training duration or transferred encoder parameters. The fitted downstream ceiling $1 - E$ describes this tested model family, not an intrinsic limit on iEEG decoding. Pretraining-loss curves bend on log–log axes because only $q - E$ is a pure power law. The fits capture aggregate and dataset-level All-unit trends ($R^2 \geq 0.91$).

Aggregate gains mask two distinct patterns (Figure 6). In Neuroprobe and BYD, data-scaling gains largely flatten after the early rungs, whereas model-size gains continue across the tested widths. Scaling also depends on task domain and dataset: auditory tasks gain most in Neuroprobe and BYD, while Pippi auditory scores respond little to either axis. Appendix Figures 14–16 give task-wise slopes, exploratory ceilings, and trajectories.

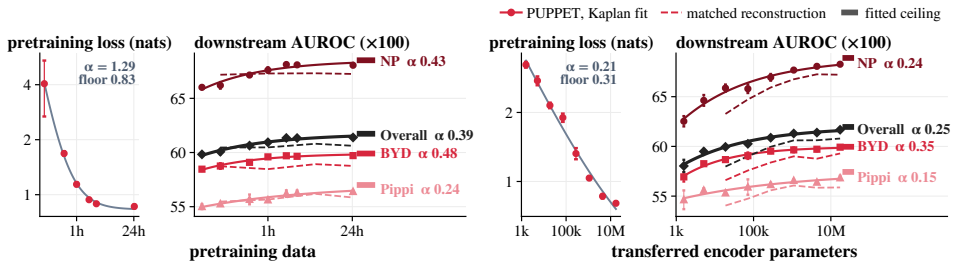


Figure 5: Pretraining loss and frozen transfer across scale. Left: data scaling at width 512; right: model-width scaling at full data. PUPPET points show three-seed means ± SD. Dashed joint waveform–Fourier reconstruction starts at ten minutes or width 32; it has three seeds at widths 32, 128, and 512, and one seed at its other rungs. Solid curves are floor-adjusted power-law fits; thick bars mark fitted AUROC ceilings. Downstream panels use iMindBench All units, with Overall equally weighting the three datasets. Figure 6 shows task-domain trajectories; Appendix H gives individual-task and reconstruction details.

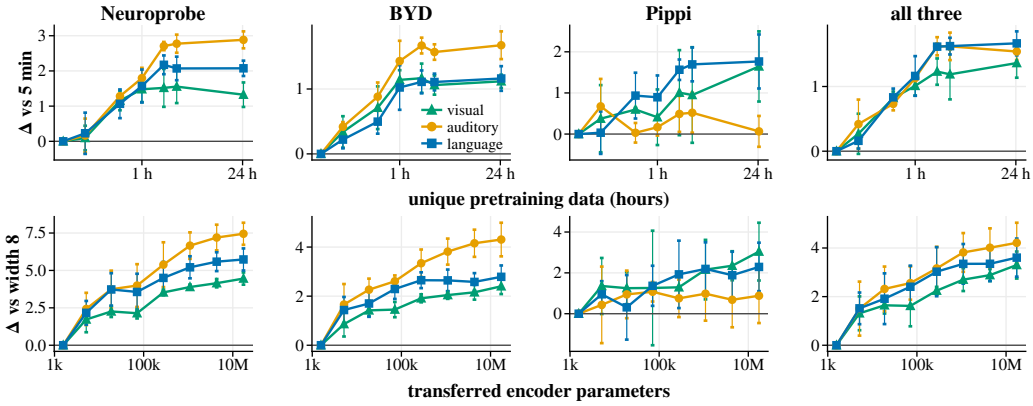


Figure 6: PUPPET scaling by task domain. Panels show Δ AUROC (×100) relative to 5 min of pretraining data (top) or width 8 (bottom). Curves give three-seed mean ± SD changes for auditory, language, and visual tasks in Neuroprobe, BYD, and Pippi using iMindBench All units. The "all three" panel pools dataset–task pairs. In Neuroprobe and BYD, data gains flatten earlier than model-size gains; the size of each gain varies by domain. Individual-task analyses and domain unit counts are in Appendix H.

6 Discussion

We introduced electrode-origin prediction as a self-supervised target built around a distinctive property of iEEG: implanted contacts persist while the activity they record changes. With one main Transformer layer and a frozen readout, PUPPET transfers more strongly than the tested matched reconstruction objectives and achieves the highest Overall iMindBench score among the compared baselines. The synthetic probes suggest one possible reason: electrode-origin training makes oscillatory peak properties more accessible to a readout, while preserving information about the aperiodic background. The scaling results also show that gains differ across task domains, so aggregate improvements should not be assumed to apply equally to every decoding task.

PUPPET reaches these results without anatomical electrode positions—even against baselines that use spatial information. That makes it useful when localization metadata are missing or implants differ, but leaves potentially valuable spatial relationships unmodeled. Combining electrode-origin prediction with position-aware modeling or channel reconstruction is a natural next test of whether those signals provide complementary gains.

GENERATIVE AI USE STATEMENT

We used generative AI tools to propose and refine hypotheses; design and critique experiments; help implement methods, preprocess data, and generate synthetic datasets; and assist in interpreting results. We also used them for literature discovery and summarization, code development and debugging, figure and table prototyping, and drafting and editing manuscript text. The authors reviewed AI-assisted research ideas against the literature, inspected and tested generated code, cross-checked reported results against available experiment records and citations against source papers, and reviewed the manuscript. AI-generated content was not treated as independent evidence. The authors take responsibility for the final text, claims, code, and artifacts in this work.

ETHICS STATEMENT

This work is a secondary analysis of publicly released intracranial recordings; we recruited no participants and collected no new recordings. We use each dataset under its release terms and report aggregate decoding results rather than attempting to identify participants. Intracranial recordings are nevertheless sensitive and potentially identifying: model and representation releases should therefore exclude raw signals, subject identifiers, electrode coordinates linked to individuals, and any artifact that enables reconstruction of protected data. Potential applications to neural decoding should be interpreted with attention to consent, privacy, population coverage, and the risk of overstating generalization beyond the clinical cohorts represented here.

REPRODUCIBILITY STATEMENT

The Method and Experimental Setup specify the preprocessing pipeline, leakage controls, architecture, pretraining objective, optimizer, and frozen-feature evaluation protocol. The appendices provide detailed data, implementation, readout, benchmark, control, attribution, scaling, and provenance material.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020. doi: 10.18653/v1/2020.acl-main.385.

Julia Berezutskaya, Mariska J. Vansteensel, Erik J. Aarnoutse, Zachary V. Freudenburg, Giovanni Piantoni, Mariana P. Branco, and Nick F. Ramsey. Open multimodal ieeg-fmri dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, 2022. doi: 10.1038/s41597-022-01173-0. URL https://doi.org/10.1038/s41597-022-01173-0.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Francesco Carzaniga, Michael Hersche, Abu Sebastian, Kaspar Schindler, and Abbas Rahimi. A foundation model with multi-variate parallel attention to generate neuronal activity. In *International Conference on Learning Representations (ICLR)*, 2026.

Geeling Chau, Christopher Wang, Sabera Talukder, Vighnesh Subramaniam, Saraswati Soedarmadji, Yisong Yue, Boris Katz, and Andrei Barbu. Population transformer: Learning population-level representations of neural activity. In *International Conference on Learning Representations (ICLR)*, 2025.

Geeling Chau, Saba Hashemi, Yonghyeon Gwon, Eshani Patel, Jan DeWitt, Christopher Wang, Andrii Zahorodnii, Sabera J. Talukder, Danny Dongyeop Han, Chun Kee Chung, Maryam M. Shanechi, and Yisong Yue. iMINDBench: iEEG multi-institution neural decoding benchmark, 2026. URL https://arxiv.org/abs/2609.18104.

Jialin Chen, Jan Eric Lenssen, Aosong Feng, Weihua Hu, Matthias Fey, Leandros Tassiulas, Jure Leskovec, and Rex Ying. From similarity to superiority: Channel clustering for time series forecasting. *arXiv preprint arXiv:2404.01340*, 2024. URL https://arxiv.org/abs/2404.01340.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. URL https://www.jmlr.org/papers/v7/demsar06a.html.

Thomas Donoghue, Matar Haller, Erik J. Peterson, Paroma Varma, Priyadarshini Sebastian, Richard Gao, Torben Noto, Antonio H. Lara, Joni D. Wallis, Robert T. Knight, Avgusta Shestyuk, and Bradley Voytek. Parameterizing neural power spectra into periodic and aperiodic components. *Nature Neuroscience*, 23(12):1655–1665, 2020. doi: 10.1038/s41593-020-00744-x. URL https://doi.org/10.1038/s41593-020-00744-x.

Danny Dongyeop Han, Yonghyeon Gwon, Ahhyun Lucy Lee, Taeyang Lee, Seong Jin Lee, Jubin Choi, Sebin Lee, Jihyun Bang, Seungju Lee, David Keetae Park, Shinjae Yoo, Chun Kee Chung, and Jiook Cha. DIVER-1: Scaling intracranial EEG foundation models for transferable representations, 2025. URL https://arxiv.org/abs/2512.19097.

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.

Sina Javadzadeh, Rahil Soroushmojdehi, S. Alireza Seyyed Mousavi, Mehrnaz Asadi, Sumiko Abe, and Terence D. Sanger. Functional embeddings enable aggregation of multi-area SEEG recordings over subjects and sessions. *arXiv preprint arXiv:2510.27090*, 2025. URL https://arxiv.org/abs/2510.27090.

Elizabeth L. Johnson and Robert T. Knight. Intracranial recordings and human memory. *Current Opinion in Neurobiology*, 31:18–25, 2015. doi: 10.1016/j.conb.2014.07.021.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Umit Keles, Julien Dubois, Kevin J. M. Le, J. Michael Tyszka, David A. Kahn, Chrystal M. Reed, Jeffrey M. Chung, Adam N. Mamelak, Ralph Adolphs, and Ueli Rutishauser. Multimodal single-neuron, intracranial eeg, and fmri brain responses during movie watching in human patients. *Scientific Data*, 2024. doi: 10.1038/s41597-024-03029-1. URL https://doi.org/10.1038/s41597-024-03029-1.

Aditya Kommineni, Emily Zhou, Kleanthis Avramidis, Simon Bock Segaard, Jeppe Roden Münster, Andreas Peter Juhl Hansen, Takfarinas Medani, Tiantian Feng, Richard Leahy, and Shrikanth Narayanan. Aperiodic and low-frequency spectral bias in reconstruction based EEG foundation models, 2026. URL https://arxiv.org/abs/2605.26434.

Gayal Kuruppu, Neeraj Wagh, Vaclav Kremen, and Yogatheesan Varatharajah. EEG foundation models: a critical review of current progress and future directions. *Journal of Neural Engineering*, 23(2):021001, 2026. doi: 10.1088/1741-2552/ae4455.

Seunghan Lee, Taeyoung Park, and Kibok Lee. Channel normalization for time series channel identification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 33655–33687. PMLR, 2025. URL https://proceedings.mlr.press/v267/lee25ad.html.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, pp. 69–84. Springer, 2016. doi: 10.1007/978-3-319-46466-4_5.

Lucine L. Oganesian, Saba Hashemi, and Maryam M. Shanechi. BaRISTA: Brain scale informed spatiotemporal representation of human intracranial neural activity. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/pdf?id=LDjBDk3Czb.

Timothée Proix, Jaime Delgado Saa, Andy Christen, Stephanie Martin, Brian N. Pasley, Robert T. Knight, Xing Tian, David Poeppel, Werner K. Doyle, Orrin Devinsky, Luc H. Arnal, Pierre Mégevand, and Anne-Lise Giraud. Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nature Communications*, 13:48, 2022. doi: 10.1038/s41467-021-27725-3.

Christopher Wang, Vighnesh Subramaniam, Adam Uri Yaari, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. BrainBERT: Self-supervised representation learning for intracranial recordings. In *International Conference on Learning Representations (ICLR)*, 2023.

Christopher Wang, Adam Yaari, Aaditya Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph Madsen, Scellig Stone, Gabriel Kreiman, et al. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. *Advances in Neural Information Processing Systems*, 37:96505–96540, 2024.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024.

Andrii Zahorodnii, Christopher Wang, Geeling Chau, Bennett Stankovits, Charikleia Moraitaki, Eli Gross, Alexander Brady, Andrei Barbu, Boris Katz, and Ila R. Fiete. Neuroprobe: Evaluating intracranial brain responses to naturalistic stimuli, 2026. URL https://arxiv.org/abs/2509.21671.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/file/535915d26859036410b0533804cee788-Paper-Conference.pdf.

Haoran Zhang, Yong Liu, Yunzhong Qiu, Haixuan Liu, Zhongyi Pei, Jianmin Wang, and Mingsheng Long. TimesBERT: A BERT-style foundation model for time series understanding. *arXiv preprint arXiv:2502.21245*, 2025. URL https://arxiv.org/abs/2502.21245.

APPENDIX TABLE OF CONTENTS

A Data, Preprocessing, and Leakage **13**

B Architecture and Pretraining Implementation **14**

B.1 Transferred Parameter Counts 16

C Frozen Evaluation **16**

C.1 Common-Readout Comparison 17

D Extended Benchmark and Statistical Analyses **17**

D.1 iMindBench Evaluation Units and Subsets 17

D.2 Baselines Omitted from the Main Comparison 17

D.3 Main-Table Uncertainty and Paired Comparisons 18

D.4 Evaluation Coverage and Statistical Reporting 19

E Objective and Architecture Controls **26**

E.1 Additional Architecture and Implementation Controls 26

F Spectral Analyses **26**

F.1 Synthetic signals and matched-objective comparison 26

F.2 Cross-model comparison with downstream performance 29

F.3 Sensitivity and interpretation of the association 31

G Representation and Attribution Analyses **32**

H Extended Scaling **34**

A DATA, PREPROCESSING, AND LEAKAGE

BrainTreebank exclusion. The Neuroprobe evaluation manifest contains five subject/session units: `sub1_sess1`, `sub3_sess0`, `sub4_sess0`, `sub7_sess0`, and `sub10_sess0`. These correspond to BrainTreebank recordings `sub_1_trial001`, `sub_3_trial000`, `sub_4_trial000`, `sub_7_trial000`, and `sub_10_trial000`. Our pretraining subset excludes every recording from subjects 1, 3, 4, 7, and 10, retaining subjects 2, 5, 6, 8, and 9. Its 13 sessions yield 3,290 records of 30 s each (27.4 h) and 650 distinct electrode identities. The 10% record-level validation partition leaves 24.7 h of records for training. The held-out Neuroprobe subjects contribute neither recordings nor electrode classes to pretraining.

Comparison of pretraining exposure. iMindBench reports that PopT-v2 was pretrained on PopT's BrainTreebank subset of 10 subjects and 19 movie recordings (Chau et al., 2026). It does not enumerate whether those recordings include any of the five Neuroprobe evaluation sessions, so we do not claim such overlap. Our exclusion is stricter than the benchmark's stated session-level separation because it removes every recording from each evaluation subject. Our subject count is half as large as PopT-v2's reported count, but that alone does not establish a ratio of training hours or examples. For data-efficiency comparisons, *effective unique training hours* counts distinct training records before resampling during optimization. For PopT-v2, we use the measured 40.47-hour raw duration of its 19 pretraining recordings, including pre/post-movie time, rather than the full BrainTreebank corpus or optimization runtime. Ten minutes is approximately 0.41% of this duration.

Signal preparation. The pretraining pipeline applies a zero-phase FIR high-pass filter at 0.5 Hz and zero-phase FIR notches at 60, 120, and 180 Hz at the source rate (2048 Hz), resamples to 500 Hz with polyphase filtering, and applies the Laplacian below. Amplitudes are clipped to $\pm 200\,\mu$V and scaled at model input. The retained 13 sessions yield 3,290 non-overlapping 30-s records (27.42 h), of which 2,961 (24.68 h) are used for training and 329 for validation; training windows sampled from these records are not additional unique hours. The session manifest and preprocessing builder are recorded in the source audit.

Laplacian re-referencing. After stripping the non-contact markers $\star$ and # from BrainTreebank electrode labels, we identify each shaft by its alphabetic stem and each contact by its numeric suffix. A channel is re-referenced to its voltage minus the mean voltage of whichever immediately adjacent contacts, numbered $n-1$ and $n+1$ on the same shaft, are present. Interior contacts use two neighbors; shaft edges or contacts adjacent to a missing contact use the single available neighbor. Contacts with neither neighbor are dropped. The same label-neighbor rule is used for our Neuroprobe, BYD, and Pippi iMindBench waveform evaluations. Evaluation filtering uses a 15-s window centered on each target interval before cropping to the 1-s encoder input; BYD originates at 1000 Hz, while Neuroprobe and Pippi originate at 2048 Hz, and all are resampled to 500 Hz.

B Architecture and Pretraining Implementation

Table 3: Canonical architecture and pretraining configuration. The two main Transformer blocks comprise one downstream encoder block and one pretraining prediction block; STCPE contains its own auxiliary block.

Component	Setting
Patch embedding	Linear projection, 50 samples to 512 features; stride 50
Main Transformer blocks	2 during pretraining; first block used downstream
Main width / heads / FFN width	512 / 16 / 2,048; SiLU activation; dropout 0.1
STCPE	Width 64; one auxiliary block, 2 heads, FFN width 256; SiLU; dropout 0; 7-patch window
Input context	Up to 32 electrodes and 0.1–3.0 s; 500-Hz waveform input
Electrode-origin classifier	650 classes; valid-patch temporal averaging; no masking
Reconstruction controls	50% random zero masking; raw output 50, Fourier output 200
Batch / base training	Global batch 256; 64 epochs (704 updates)
Optimizer	Muon for eligible linear weights, auxiliary AdamW for other parameters
Peak learning rate	0.02 for electrode-origin/joint; 0.01 for raw-only/Fourier-only
Schedule / regularization	5% warm-up; cosine decay to 1% of peak; weight decay 10^{-4}
Numerics / selection	bfloat16; gradient-norm clip 1.0; lowest-validation-loss epoch
Pretraining seeds	41, 42, 43 for the base and three reconstruction controls

STCPE construction. DIVER-1's STCPE adds signal-dependent temporal and cross-electrode context before the main encoder without explicit 3D electrode coordinates (Han et al., 2025). For one example, let $\mathbf{U} \in \mathbb{R}^{C\times N\times d}$ be the patch-token grid. In the base model, a projection $P_\downarrow : \mathbb{R}^d \to \mathbb{R}^{d_s}$ reduces $d = 512$ to $d_s = 64$. An auxiliary MOIRAI block processes all sampled electrodes jointly within sliding windows of $w = 7$ temporal patches. If Ω_t is the set of windows containing patch t, $a_t = |\Omega_t|$, and $P_\uparrow : \mathbb{R}^{d_s} \to \mathbb{R}^d$ projects back to encoder width, the overlapping window outputs are averaged and added to the patch embedding:

$$\begin{aligned} \widetilde{\mathbf{U}} &= P_\downarrow(\mathbf{U}), \qquad r_w = (w-1)/2, \\ \mathbf{H}_\tau &= \mathrm{MOIRAI}_{\mathrm{STCPE}}\Big(\widetilde{\mathbf{U}}_{[:,\,\tau-r_w:\tau+r_w,\,:]}\Big), \\ \overline{\mathbf{H}}[:,t,:] &= \frac{1}{a_t}\sum_{\tau\in\Omega_t} \mathbf{H}_\tau[:,\,t-\tau+r_w,\,:], \\ \mathbf{U}^{(0)}[:,t,:] &= \mathbf{U}[:,t,:] + P_\uparrow(\overline{\mathbf{H}}[:,t,:]). \end{aligned} \tag{2}$$

Here τ indexes a window center. The released configuration uses stride-one full temporal padding: seven-patch windows extend over six zero-padded positions at each end, and outputs overlapping a real patch are averaged. When an example carries a validity mask, padded positions are excluded

from window attention. Permuting input electrodes permutes STCPE outputs in the same way because the auxiliary block attaches no fixed identity to a channel index.

MOIRAI attention. Following DIVER-1, the encoder and pretraining prediction block apply MOIRAI attention jointly across electrode and temporal patches (Woo et al., 2024; Han et al., 2025). For input token $\mathbf{h}_{c,n}$ at electrode c and patch n, one head uses temporal RoPE and two learned same-versus-cross-electrode biases:

$$\begin{aligned} e_{cn,c'n'} &= \frac{(\mathbf{W}^Q\mathbf{h}_{c,n})^\top \mathbf{R}_{n-n'}(\mathbf{W}^K\mathbf{h}_{c',n'})}{\sqrt{d_h}} + u^{(1)}\mathbf{1}_{\{c=c'\}} + u^{(2)}\mathbf{1}_{\{c\neq c'\}}, \\ \mathrm{Attn}(\mathbf{h})_{c,n} &= \sum_{c',n'} \mathrm{softmax}_{c',n'}(e_{cn,c'n'})\,\mathbf{W}^V\mathbf{h}_{c',n'}. \end{aligned} \tag{3}$$

Here d_h is the head width, $\mathbf{R}_{n-n'}$ represents relative temporal position through RoPE, and $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are learned projections. The two scalar biases are learned per layer and head. The same-versus-different test is unchanged by electrode reordering, preserving channel-order equivariance without fixed channel-position embeddings.

Reconstruction targets. Each raw patch contains 50 samples at 500 Hz. The raw head predicts those 50 samples. The Fourier transform zero-pads the patch to a 500-point real FFT and selects indices 0–199, corresponding to 0–199 Hz. The Fourier head's output dimension is inferred from a dummy transformed target and is therefore 200. Because the implementation casts the complex FFT to the waveform's real dtype before applying absolute value, the actual target is $\log[1 + |\,\mathrm{Re}(\mathrm{rFFT}_{500}(\mathbf{r}))|]$, not the full complex magnitude. The configured normalization flag is not applied by the transform; the FFT uses the `backward` normalization convention. The raw and Fourier mean-squared-error terms each have unit weight in the joint objective. All three controls randomly replace 50% of patches with a fixed zero vector before embedding and score only masked valid patches. This description preserves the target used to train the evaluated checkpoints rather than silently substituting a different Fourier objective.

Window sampling and padding. Records are split by a run-seeded 90/10 training–validation partition, so validation is record-disjoint but not necessarily session-disjoint. The collator first samples on a grid of 1-s blocks, capped at 32 channels and three temporal blocks. With equal probability it samples the channel count or temporal-block count first. Each count is obtained by scaling its available maximum by a $\mathrm{Beta}(3, 1)$ draw, rounding, and clipping to a valid positive count; the second count respects the remaining token budget. Channels are selected without replacement and sorted, and a shared contiguous block interval is selected with a uniformly sampled valid start. The final selected 1-s block retains a uniformly drawn multiple of 50 samples, from 50 to 500, giving total durations from 0.1 to 3.0 s. Thus durations are not drawn uniformly over that entire range. Unused samples and channels are zero-padded and marked invalid; validity masks are propagated to attention, temporal averaging, and loss. The model subsequently divides the sampled signal into 50-sample tokens.

Optimization. The active record-count and width runs use 64 epochs, with 11 updates per epoch and a global batch of 256, for 704 training updates; the transferred checkpoint is selected by minimum validation loss and can precede the last update. Muon updates eligible nondegenerate `Linear.weight` matrices, including hidden STCPE matrices. Output-head weights, embedding tables, and other ineligible weights use auxiliary AdamW. Biases, normalization parameters, and other parameters with fewer than two dimensions receive no weight decay. Muon momentum is 0.95; auxiliary AdamW uses $\beta_1 = 0.9$ and $\beta_2 = 0.95$. The launcher sets both optimizer groups to the same objective-specific peak learning rate. Both follow 5% linear warm-up and cosine decay to 1% of their peak rate. Training uses bfloat16 and gradient clipping at norm 1.0, and selects the epoch with minimum validation loss. The instantiated width-512 pretraining model has 8,545,280 Muon-updated parameters, 1,425,476 decayed AdamW parameters, and 6,816 non-decayed AdamW parameters; its transferred encoder contains 4,353,956 parameters, including 131,908 in STCPE.

B.1 TRANSFERRED PARAMETER COUNTS

Supervised-classifier counts in Table 1. Because these models are trained separately for each evaluation unit, their full classifier size varies with the available channels and input representation. The starred table entries describe one 109-channel binary evaluation unit. For the published waveform route, the leaderboard identifies the 500-Hz preprocessing track; a one-second input therefore has $109 \times 500 = 54{,}500$ features. The upstream logistic builder fits one binary coefficient per feature plus an intercept (54,501 parameters). The MLP configuration uses two 128-unit hidden layers, giving 6,992,898 parameters; the waveform CNN uses three temporal convolutions and a 128-unit classifier, giving 1,057,602. The representative multi-STFT classifier measurements are approximately 214,400 (logistic), 55,690,000 (MLP), and 15,390,000 (CNN); the 2048-Hz HTNet configuration has approximately 11,100. All seven values count the complete supervised classifier for one evaluation unit, not an encoder shared across units.

We count learned parameters used to compute the transferred representation, excluding discarded pretraining heads, downstream readouts, optimizer state, and fixed buffers. For width-64 PUPPET, the count is 71,222: 3,264 in the linear patch embedding, 2,162 in STCPE (including its auxiliary Transformer and projections), 65,732 in the first main Transformer block, and 64 in the final normalization used by the intermediate-depth evaluation path. The second main block and all pretraining output heads are excluded.

The evaluated PopT-v2 checkpoint contains 654,336 learned input-encoding parameters and 18,914,304 parameters in its six-layer Transformer, totaling 19,568,640. Its positional-encoding tensor of 640,000 entries is a fixed buffer, not a learned parameter; its two 513-parameter pretraining classifiers are also excluded. Thus the width-64 encoder uses $71{,}222/19{,}568{,}640 \times 100 = 0.36396\%$ of PopT-v2's transferred parameters.

Supervised classifier sizes in Table 1 use one representative binary, 109-channel, 500-sample waveform evaluation unit; they are not shared pretrained encoders. Flattening gives 54,500 input features. A binary logistic model has 54,501 fitted coefficients including its intercept. The benchmark MLP's two 128-unit hidden layers give 6,992,898 learned parameters. The benchmark one-dimensional CNN's three convolutional layers and 128-unit classifier give 1,057,602. These counts vary with channel count and are shown only to contextualize this representative evaluation unit; the benchmark scores aggregate many units. The model definitions and hidden dimensions follow the iMindBench evaluation harness (Chau et al., 2026).

C FROZEN EVALUATION

For each 1-s downstream window, the encoder emits ten 0.1-s patch vectors per electrode. We concatenate each electrode's valid-token temporal mean with its ten patch vectors and flatten the electrode, token, and feature axes. The encoder remains frozen; the only task-specific learned mapping is an affine linear readout. Temporal averaging introduces no learned pooling parameters, and no nonlinear adapter or encoder update is used.

For evaluation unit u, frozen encoder features $\boldsymbol{\phi}_i$ are standardized using only its training partition. Writing the binary labels as $y_i \in \{-1, +1\}$, an L_2-regularized logistic readout is fitted by

$$(\mathbf{w}_u^\star, b_u^\star) = \arg\min_{\mathbf{w},b} \sum_{i \in \mathcal{D}_{\text{train}}^u} \log\Big(1 + \exp[-y_i(\mathbf{w}^\top \widetilde{\boldsymbol{\phi}}_i + b)]\Big) + \frac{1}{2C}\|\mathbf{w}\|_2^2, \qquad C = 1. \tag{4}$$

The value of C is fixed across all units, with no task-specific hyperparameter optimization. The scikit-learn solver is L-BFGS with a maximum of 10,000 iterations and tolerance 10^{-3}; it uses no GPU acceleration.

The primary analysis uses the same logistic-regression estimator as iMindBench's Multi-STFT logistic baseline, holding the readout method fixed when comparing learned embeddings with engineered spectral features. This does not make all published model comparisons protocol-identical: those scores retain their reported preprocessing and adaptation procedures.

C.1 Common-Readout Comparison

To check whether PUPPET's lead reflects the choice of readout, we also fit our logistic readout to frozen PopT-v2 and BaRISTA features. PUPPET reaches 66.13 Overall AUROC, compared with 63.00 for PopT-v2 and 63.81 for BaRISTA under this readout. Their published fine-tuned scores are 64.5 and 63.6, respectively; the main comparison retains those published scores. A common readout does not make the models' feature construction or preprocessing identical.

Table 4: Published and common-readout comparisons on iMindBench Main. Overall AUROC (%); published rows use the models' reported fine-tuning protocols, while common-readout rows use frozen encoders and our standardized logistic readout. PUPPET averages three pretraining seeds.

Model	Evaluation	Overall AUROC (%)
PopT-v2	Published fine-tuning	64.5
PopT-v2	Frozen encoder + our logistic readout	63.00
BaRISTA	Published fine-tuning	63.6
BaRISTA	Frozen encoder + our logistic readout	63.81
PUPPET	Frozen encoder + our logistic readout	**66.13**

For the frozen-feature comparisons, PopT-v2 uses concatenated token features and BaRISTA uses tokens before its learned pooling. The logistic readout standardizes features and uses $C = 1$; the prior-model feature dumps are fixed rather than independently pretrained seeds.

D Extended Benchmark and Statistical Analyses

D.1 iMindBench Evaluation Units and Subsets

Coverage audit. The evaluation manifests contain 5 Neuroprobe patients / 5 recordings, 16 BYD patients / 29 recordings, and 5 Pippi patients / 5 recordings. Main / Challenge / All contain 49 / 26 / 75, 254 / 181 / 435, and 44 / 31 / 75 evaluation units, respectively.

iMindBench evaluation protocol. An evaluation unit is one dataset–subject/session–task combination, evaluated over up to two within-session folds, each with training, validation, and test splits. Three frame-brightness units in Pippi's Main subset have only one saved fold in the PUPPET and frozen PopT-v2/BaRISTA comparison rows, so their unit means use that available fold. Preprocessing statistics are fitted only on each training partition. Our waveform route uses same-shaft Laplacian re-referencing and 500-Hz sampling, with 15 s of context for filtering before extracting each 1-s window; Appendix A gives dataset-specific details. Main contains 49 Neuroprobe, 254 BYD, and 44 Pippi units. All contains 75, 435, and 75 units, respectively; Challenge contains the All units not assigned to Main. The published-comparison models have task-wise scores for both subsets in Tables 6–11. Published iMindBench comparisons retain model-native preprocessing and adaptation; they compare complete systems rather than isolating the SSL objective.

iMindBench tasks and subsets. Its language tasks decode sentence onset, speech presence, first-versus-second word position, inter-word gap, GPT-2 surprisal, dependency-head direction, word duration, and noun-versus-verb part of speech. Auditory tasks decode change in volume, absolute volume, and pitch; visual tasks decode global and local optical flow, frame brightness, and face presence. A validation-selected STFT logistic decoder and a 500-Hz HTNet-style waveform decoder (derived from EEGNet) screen each supported unit; Main contains a unit when either screen exceeds fold-averaged validation AUROC 0.60. Supported units below that threshold form Challenge, and All is their union (Chau et al., 2026). These labels describe performance under the prespecified screens, not the presence or absence of neural signal.

D.2 Baselines Omitted from the Main Comparison

The published PopT-v2, BaRISTA, and DIVER-1 scores follow model-native preprocessing and adaptation. BrainBERT was evaluated with single-STFT inputs and a frozen encoder plus linear readout in iMindBench Appendix D, not its main comparison table. MVPFormer uses five-second

segments and Brant uses six-second temporal/frequency patches; iMindBench Appendix I.1 defers their main-table inclusion pending a validated one-second adaptation (Chau et al., 2026).

BrainBERT's reported Overall AUROC is 0.619, compared with 0.630, 0.634, and 0.620 for the matched single-STFT logistic, MLP, and CNN baselines, respectively. The benchmark notes that its one-second inputs and training-split normalization differ from BrainBERT's five-second pretraining windows and original normalization. These scores characterize the released representation under that protocol; they do not establish that BrainBERT is generally ineffective. MVPFormer's released configuration additionally combines 25 five-second segments; retaining that context would extend beyond the target window and require explicit protection of chronological fold boundaries. Brant's temporal and frequency features likewise require a validated construction rather than simply stretching one-second signals to six seconds (Chau et al., 2026).

D.3 Main-Table Uncertainty and Paired Comparisons

The Neuroprobe Main subset comprises 49 evaluation units from five patients and 15 task names; BYD has 254 units from 16 patients and 15 task names; Pippi has 44 units from five patients and 14 task names. These are not counts of sessions. Repeated units from one patient are correlated. Table 5 gives the session and subset counts.

Pretraining-seed uncertainty. The following are mean $\pm$ sample SD across three pretraining seeds, in AUROC points ($100\times$ AUROC). These SDs measure pretraining-seed variability, not uncertainty across patients, and are not tests against the published baseline.

PUPPET variant	iMindBench Overall
Main	66.13 ± 0.03
Ten minutes	64.67 ± 0.25
Compact (71k)	64.66 ± 0.19

For the main width-512 model, the finalized per-unit export gives the following corpus-level scores by seed; published baselines have no comparable pretraining-seed SD in this analysis.

Subset	Seed	Neuroprobe	BYD	Pippi	Overall
Main	42	74.18	65.08	59.24	66.17
	41	74.30	64.98	59.07	66.12
	43	74.08	64.76	59.52	66.12
	Mean $\pm$ SD	74.19 ± 0.11	64.94 ± 0.16	59.28 ± 0.23	66.13 ± 0.03
All	42	68.03	59.90	56.36	61.43
	41	68.15	59.75	56.39	61.43
	43	67.98	59.50	56.45	61.31
	Mean $\pm$ SD	68.05 ± 0.08	59.72 ± 0.20	56.40 ± 0.05	61.39 ± 0.07

Patient-stratified bootstrap. Within each shared iMindBench evaluation unit, folds and PUPPET pretraining seeds are averaged before computing paired AUROC differences. We resample patients with replacement within Neuroprobe, BYD, and Pippi, recompute the unit mean for each dataset and the equal-weight Overall difference, and repeat this 20,000 times with a fixed seed. Reported intervals are the 2.5th and 97.5th percentiles; two-sided bootstrap p-values are Holm-corrected within the ten-model Table 1 family, the 55-pair Figure 2 family, or the 12-arm ablation family, as appropriate. PUPPET exceeds all ten main-table comparators under the first family: versus PopT-v2 the difference is $+1.9$ [95% CI $+1.0$, $+2.9$], versus BaRISTA $+2.7$ [$+1.0$, $+4.3$], and versus DIVER-1 $+2.9$ [$+2.3$, $+3.6$] AUROC points. A mixed-effects model, per-dataset patient-level tests, and unit-level Wilcoxon tests were examined as sensitivity checks; all agreed on effect direction, but their p-values are not used for manuscript significance claims. Neuroprobe and Pippi each have only five patients, so per-dataset confidence intervals are reported without significance stars. Table 1 quotes the benchmark's printed summary scores; the paired analyses instead use its released per-unit scores. The released Pippi means differ from the printed summaries (PopT-v2: 0.5733 versus 0.579), so the paired effect is reported separately rather than attached to a displayed Table 1 mean.

D.4 Evaluation Coverage and Statistical Reporting

Table 5: Coverage in the iMindBench manifests and completed per-unit score export. Counts denote supported evaluation units, not independent patients. The Challenge subset is the All subset minus Main.

Dataset	Patients	Recordings	Main	Challenge	All
Neuroprobe	5	5	49	26	75
BYD	16	29	254	181	435
Pippi	5	5	44	31	75

Paired analyses. Average ranks use the fixed 11-model set on 49 Neuroprobe, 254 BYD, and 44 Pippi shared Main units (Demšar, 2006). The three corpus mean ranks are equally weighted; PUPPET ranks first at 3.54. Rank is descriptive; significance is determined from patient-bootstrap AUROC differences as specified in Appendix D.3.

Task-wise breakdown. Tables 6–11 give AUROC for every task, dataset, and Main or All subset across all 11 published-comparison models. These are means over available evaluation units for each dataset–task pair; All contains Main and Challenge units. PUP, PopT, BaR, and Div denote PUPPET, PopT-v2, BaRISTA, and DIVER-1; S/W denote multi-STFT/waveform input for supervised models, and HTN denotes HTNet.

Table 6: Neuroprobe task-wise AUROC: pretrained models. Main and All results (AUROC $\times 100$); n is the number of evaluation units for that task.

Task	n	PUP	PopT	BaR	Div
Main: Neuroprobe					
delta_volume	4	82.0	80.1	79.0	81.4
face_num	1	57.0	61.5	58.1	58.1
frame_brightness	1	57.1	52.0	58.9	61.3
global_flow	2	71.7	71.1	68.2	68.2
gpt2_surprisal	3	65.3	62.5	63.0	64.9
local_flow	3	61.5	59.6	55.8	58.8
onset	5	93.3	92.7	94.9	92.3
pitch	4	62.2	59.4	55.1	58.7
speech	5	89.8	90.1	92.4	87.3
volume	4	76.3	74.1	69.1	71.9
word_gap	4	68.2	67.3	67.3	67.4
word_head_pos	2	64.1	65.3	62.0	65.0
word_index	4	85.1	80.0	83.7	84.9
word_length	4	66.0	60.8	61.4	64.2
word_part_speech	3	64.7	62.9	58.7	66.1
All: Neuroprobe					
delta_volume	5	77.8	74.9	75.4	77.6
face_num	5	52.5	50.6	51.9	51.2
frame_brightness	5	51.0	48.4	50.1	49.5
global_flow	5	66.3	63.1	61.1	63.8
gpt2_surprisal	5	61.7	60.9	60.6	62.6
local_flow	5	62.4	59.4	56.2	59.1
onset	5	93.3	92.7	94.9	92.3
pitch	5	60.0	58.7	54.0	57.2
speech	5	89.8	90.1	92.4	87.3
volume	5	72.3	71.5	66.1	67.1
word_gap	5	64.8	63.1	64.2	64.5
word_head_pos	5	61.0	60.0	58.9	60.1
word_index	5	82.4	76.4	81.1	82.3
word_length	5	63.8	59.4	60.1	62.1
word_part_speech	5	61.5	59.8	56.0	61.6

Table 7: Neuroprobe task-wise AUROC: supervised baselines. Main and All results (AUROC $\times 100$); n is the number of evaluation units for that task.

Task	n	LogS	LogW	MLPW	MLPS	CNNS	CNNW	HTN
Main: Neuroprobe								
delta_volume	4	76.2	73.0	77.5	76.0	76.9	78.0	75.3
face_num	1	59.3	50.2	49.5	60.6	63.1	52.5	56.9
frame_brightness	1	59.0	49.3	52.2	59.5	59.8	49.8	65.6
global_flow	2	69.0	53.6	56.0	69.9	73.0	59.2	63.7
gpt2_surprisal	3	62.3	58.3	62.6	62.7	61.4	62.0	58.3
local_flow	3	60.8	52.6	53.1	61.0	62.2	53.9	55.9
onset	5	88.5	75.8	84.0	89.9	88.7	85.5	89.7
pitch	4	58.0	53.1	52.9	58.2	61.6	51.5	58.3
speech	5	86.2	57.4	64.5	86.9	87.1	72.6	87.0
volume	4	68.9	54.4	57.7	69.6	74.3	58.8	73.1
word_gap	4	62.3	63.7	66.2	62.2	61.2	66.1	63.3
word_head_pos	2	63.2	57.7	62.2	63.5	63.8	63.6	62.8
word_index	4	76.4	78.3	82.8	77.0	75.6	83.2	78.1
word_length	4	58.8	59.9	61.6	59.3	58.7	61.7	56.4
word_part_speech	3	61.9	58.2	59.9	62.1	65.7	59.4	58.9
All: Neuroprobe								
delta_volume	5	71.2	70.2	74.3	71.6	72.7	75.0	70.7
face_num	5	50.6	50.2	49.1	51.0	53.6	47.9	52.2
frame_brightness	5	49.8	51.1	48.7	50.2	51.7	47.3	54.1
global_flow	5	64.3	52.8	55.1	65.7	65.7	55.8	59.3
gpt2_surprisal	5	59.1	56.9	60.2	59.6	59.1	59.7	56.1
local_flow	5	59.9	52.7	53.2	60.8	61.2	53.3	55.5
onset	5	88.5	75.8	84.0	89.9	88.7	85.5	89.7
pitch	5	57.6	52.6	52.6	57.6	60.2	51.0	56.5
speech	5	86.2	57.4	64.5	86.9	87.1	72.6	87.0
volume	5	66.8	54.1	55.8	67.4	71.7	56.7	67.9
word_gap	5	60.1	61.5	63.5	59.9	59.4	63.3	61.8
word_head_pos	5	59.5	56.4	59.1	59.3	60.4	59.1	57.6
word_index	5	72.5	76.0	80.3	73.3	72.7	79.7	73.5
word_length	5	57.0	59.0	60.2	57.7	57.2	60.3	55.1
word_part_speech	5	58.2	56.5	57.5	58.0	59.8	56.5	56.9

Table 8: BYD task-wise AUROC: pretrained models. Main and All results (AUROC $\times 100$); n is the number of evaluation units for that task.

Task	n	PUP	PopT	BaR	Div
Main: BYD					
delta_volume	10	60.9	55.3	57.6	58.2
face_num	27	72.0	68.8	68.7	63.2
frame_brightness	3	52.0	55.7	52.0	54.4
global_flow	22	55.2	52.7	54.0	52.9
gpt2_surprisal	7	59.1	50.6	54.7	58.1
local_flow	20	56.4	56.1	55.5	54.7
onset	27	77.9	76.6	71.5	69.0
pitch	24	64.4	62.4	62.3	59.5
speech	29	81.4	81.6	88.6	75.0
volume	28	70.2	68.8	69.6	62.9
word_gap	11	53.6	52.0	50.8	52.1
word_head_pos	3	54.3	51.8	53.5	54.8
word_index	17	57.1	53.1	49.0	58.5
word_length	11	49.4	52.0	51.0	49.8
word_part_speech	15	52.4	51.2	49.2	50.1
All: BYD					
delta_volume	29	57.8	53.5	54.1	56.2
face_num	29	72.2	69.1	68.6	62.8
frame_brightness	29	47.5	50.4	51.0	49.6
global_flow	29	54.9	53.6	54.0	52.4
gpt2_surprisal	29	55.4	51.2	51.7	54.5
local_flow	29	55.5	55.9	53.9	54.3
onset	29	76.7	75.1	70.3	68.0
pitch	29	62.8	61.2	60.6	58.5
speech	29	81.4	81.6	88.6	75.0
volume	29	69.4	68.0	68.8	62.7
word_gap	29	52.6	52.0	51.1	51.4
word_head_pos	29	52.2	51.4	50.7	51.1
word_index	29	55.4	52.5	50.1	55.6
word_length	29	50.0	52.3	50.6	50.8
word_part_speech	29	51.7	50.7	50.0	50.3

Table 9: BYD task-wise AUROC: supervised baselines. Main and All results (AUROC $\times 100$); n is the number of evaluation units for that task.

Task	n	LogS	LogW	MLPW	MLPS	CNNS	CNNW	HTN
Main: BYD								
delta_volume	10	55.8	55.1	56.6	55.8	54.3	53.6	54.4
face_num	27	69.1	50.8	52.8	69.4	67.6	52.8	64.6
frame_brightness	3	52.6	45.4	44.7	49.5	53.5	51.8	57.6
global_flow	22	54.2	50.2	50.5	53.9	52.0	51.2	52.9
gpt2_surprisal	7	54.9	58.3	59.7	53.1	50.9	55.4	52.5
local_flow	20	56.1	51.3	51.6	55.5	53.7	51.7	53.1
onset	27	72.3	54.1	55.5	73.5	71.4	54.3	68.8
pitch	24	63.7	50.3	50.4	63.3	62.1	52.3	58.7
speech	29	77.7	51.7	56.0	78.5	78.1	61.0	74.3
volume	28	67.1	51.4	52.1	67.3	65.5	53.6	64.5
word_gap	11	52.3	52.4	50.6	52.2	49.9	52.6	50.7
word_head_pos	3	52.2	50.4	53.1	50.7	52.8	52.8	49.7
word_index	17	52.7	57.0	57.4	52.1	50.1	54.6	49.0
word_length	11	52.1	51.1	53.2	53.3	52.4	51.4	51.8
word_part_speech	15	50.1	48.8	51.1	51.2	50.3	50.7	50.8
All: BYD								
delta_volume	29	54.4	52.8	54.4	54.7	52.7	51.8	52.3
face_num	29	69.3	50.7	52.9	69.5	67.3	52.9	64.1
frame_brightness	29	47.7	49.5	48.4	47.4	52.7	49.1	50.9
global_flow	29	53.7	50.3	50.5	53.8	52.1	50.9	52.8
gpt2_surprisal	29	51.1	54.4	56.1	51.3	50.0	54.2	50.3
local_flow	29	55.7	51.1	51.0	55.3	53.4	51.5	52.3
onset	29	71.1	53.9	54.9	72.2	70.4	53.8	67.1
pitch	29	62.1	50.2	50.6	61.8	60.8	51.6	57.0
speech	29	77.7	51.7	56.0	78.5	78.1	61.0	74.3
volume	29	66.7	51.3	51.6	66.9	64.9	53.2	64.0
word_gap	29	51.5	51.6	50.4	51.6	49.9	51.8	48.2
word_head_pos	29	49.7	51.4	52.2	49.7	50.9	51.4	50.3
word_index	29	52.1	55.5	55.0	51.7	49.7	53.8	49.5
word_length	29	51.6	50.8	50.5	51.5	52.2	50.5	50.5
word_part_speech	29	49.9	47.9	49.6	51.1	49.6	48.4	51.1

Table 10: Pippi task-wise AUROC: pretrained models. Main and All results (AUROC $\times 100$); n is the number of evaluation units for that task.

Task	n	PUP	PopT	BaR	Div
Main: Pippi					
face_num	5	50.7	51.0	50.1	53.3
frame_brightness	3	36.3	40.8	45.0	38.5
global_flow	4	53.7	59.6	49.9	48.7
gpt2_surprisal	3	52.7	49.2	49.7	50.9
local_flow	4	72.1	66.6	63.5	64.4
onset	5	77.0	65.7	59.7	66.1
pitch	1	51.7	54.0	49.8	51.0
speech	5	80.4	76.4	80.0	73.9
volume	4	52.9	47.6	50.7	57.7
word_gap	4	45.3	54.3	45.0	43.3
word_head_pos	1	54.8	53.4	58.6	56.2
word_index	1	73.0	62.6	54.8	61.4
word_length	1	52.7	55.8	44.0	52.1
word_part_speech	3	57.6	49.8	54.7	54.7
All: Pippi					
delta_volume	5	53.0	51.7	50.0	54.5
face_num	5	50.7	51.0	50.1	53.3
frame_brightness	5	46.8	49.8	45.1	41.4
global_flow	5	52.5	56.6	51.1	48.7
gpt2_surprisal	5	51.6	48.0	50.1	48.4
local_flow	5	66.9	61.3	62.5	61.6
onset	5	77.0	65.7	59.7	66.1
pitch	5	52.3	47.9	53.3	52.0
speech	5	80.4	76.4	80.0	73.9
volume	5	52.1	48.6	51.0	57.4
word_gap	5	40.6	53.3	41.8	44.0
word_head_pos	5	50.5	53.4	53.3	52.7
word_index	5	66.3	58.8	51.4	52.6
word_length	5	50.5	50.0	48.0	53.1
word_part_speech	5	54.8	53.9	58.0	51.7

Table 11: Pippi task-wise AUROC: supervised baselines. Main and All results (AUROC $\times 100$); n is the number of evaluation units for that task.

Task	n	LogS	LogW	MLPW	MLPS	CNNS	CNNW	HTN
Main: Pippi								
face_num	5	49.6	50.6	51.0	46.9	50.5	49.7	52.4
frame_brightness	3	32.5	55.6	57.5	32.9	42.5	52.5	32.5
global_flow	4	54.3	52.5	53.1	53.4	51.6	46.8	45.8
gpt2_surprisal	3	45.7	55.1	53.4	48.0	50.6	51.9	49.1
local_flow	4	71.5	51.0	52.6	70.4	64.2	52.2	69.8
onset	5	63.7	54.2	54.6	67.6	62.5	53.8	62.4
pitch	1	58.1	50.9	49.5	57.3	49.2	50.3	46.6
speech	5	76.5	50.9	53.7	77.2	70.2	52.2	74.6
volume	4	51.0	50.1	51.7	51.1	47.0	51.1	49.6
word_gap	4	46.6	50.7	56.9	47.2	38.6	60.7	44.9
word_head_pos	1	48.5	54.5	55.2	52.7	50.9	47.8	47.6
word_index	1	63.1	57.0	51.0	67.3	48.2	56.8	59.4
word_length	1	47.0	59.7	64.3	45.9	59.4	59.2	51.4
word_part_speech	3	51.0	54.2	49.9	48.9	48.6	55.3	52.2
All: Pippi								
delta_volume	5	51.2	52.1	52.3	52.7	49.6	51.9	49.2
face_num	5	49.6	50.6	51.0	46.9	50.5	49.7	52.4
frame_brightness	5	42.1	54.4	54.1	50.3	51.2	51.8	38.8
global_flow	5	53.8	50.6	51.6	52.6	52.6	48.8	46.7
gpt2_surprisal	5	48.7	52.1	51.9	50.9	51.5	51.0	46.6
local_flow	5	64.2	51.3	53.0	63.6	57.9	52.2	64.6
onset	5	63.7	54.2	54.6	67.6	62.5	53.8	62.4
pitch	5	51.4	53.0	52.4	49.8	51.7	49.0	52.3
speech	5	76.5	50.9	53.7	77.2	70.2	52.2	74.6
volume	5	50.5	50.3	51.3	50.9	46.7	50.4	50.0
word_gap	5	45.6	52.7	56.4	51.8	44.8	64.2	49.4
word_head_pos	5	49.0	46.9	47.9	50.9	47.7	47.6	50.5
word_index	5	58.9	56.6	57.9	56.5	48.4	58.4	51.4
word_length	5	49.9	51.5	52.8	50.4	54.8	51.5	47.6
word_part_speech	5	54.2	54.1	50.6	52.7	47.3	51.4	51.9

Table 12: Aggregate iMindBench Challenge and All results. Overall AUROC is the unweighted mean of the three dataset means, computed from the released per-unit scores on the complete shared sets: Challenge has 26/181/31 and All has 75/435/75 Neuroprobe/BYD/Pippi units. PUPPET is mean ± SD over three pretraining seeds; baseline values retain their published adaptation protocols. These descriptive comparisons are not patient-bootstrap significance tests.

Model	Challenge Overall	All Overall
PUPPET	$\mathbf{0.537 \pm 0.002}$	$\mathbf{0.614 \pm 0.001}$
PopT-v2 (multi-STFT)	0.526	0.599
BaRISTA (waveform)	0.522	0.592
DIVER-1 (waveform)	0.525	0.592
Logistic (multi-STFT)	0.520	0.586
Logistic (waveform)	0.519	0.542
MLP (waveform)	0.520	0.554
MLP (multi-STFT)	0.530	0.591
CNN (multi-STFT)	0.524	0.583
CNN (waveform)	0.515	0.554
HTNet (waveform)	0.509	0.573

E Objective and Architecture Controls

E.1 Additional Architecture and Implementation Controls

The main table reports the matched objective comparison and two architecture controls. The vanilla-attention replacement removes both the binary attention bias and temporal RoPE from the main and STCPE Transformer blocks.

Overall ablation differences. Differences are PUPPET minus each ablation in Overall AUROC points; positive values favor PUPPET. Against the three matched reconstruction controls, the differences are $+2.7$ [95% CI $+2.1$, $+3.5$] for waveform reconstruction, $+1.6$ [$+1.1$, $+2.1$] for Fourier-target reconstruction, and $+1.4$ [$+0.8$, $+2.0$] for joint reconstruction. The displayed architecture controls give $+0.7$ [$+0.4$, $+1.0$] for vanilla attention and $+0.4$ [$+0.2$, $+0.7$] for STCPE removal. Holm adjustment uses a 12-arm family that additionally includes both channel-bias removals, convolutional patch embedding, both temporal-RoPE removals, STR removal, and electrode-origin prediction plus reconstruction; those additional arms are not displayed in Table 2.

F Spectral Analyses

F.1 Synthetic signals and matched-objective comparison

The matched-objective comparison tests which spectral parameters can be predicted from frozen representations (Figure 7). PUPPET predicts peak frequency, height, and width with R^2 of 0.365, 0.329, and 0.623, compared with 0.162, 0.144, and 0.383 for joint reconstruction. It also recovers background slope and level accurately ($R^2 = 0.998$ and 0.997). These values belong to the full-window probe described here, not the cross-model probe in the main figure. For each parameter, we generate 1,000 five-second signals sampled at 500 Hz. Their log-power spectra have the form

$$\log_{10} P(f) = b - \beta \log_{10} f + h \exp\left[-\frac{(f-\mu)^2}{2\sigma^2}\right] + \epsilon(f), \tag{5}$$

where b and β specify background level and slope, μ, h, and σ specify peak center, height, and width, and $\epsilon(f)$ is a zero-mean Gaussian perturbation with standard deviation s in log power. Width denotes the Gaussian standard deviation; the corresponding fitted bandwidth is 2σ. Spectra are constructed at 0.5-Hz spacing from 0.5 to 250 Hz, converted to linear power, and interpolated to the 0.2-Hz Fourier grid. An inverse Fourier transform with random phase produces each waveform. One parameter varies linearly over its specified range while the others remain at their medians. Models share the same signals and folds.

Calibration and scope. Ranges are the patient-balanced 10th–90th percentiles from 832 channel-record spectra across five BrainTreebank patients and 13 sessions. Welch spectra of 30-second records use four-second windows. Fits separate background and peaks over 2–45 Hz (Donoghue et al., 2020); accepted peaks determine peak quantiles and accepted fits determine background quantiles. Each patient contributes equal total weight. The sweeps isolate individual parameters; they do not sample their empirical joint distribution.

Table 13: Synthetic parameter ranges for the matched-objective comparison. Height, level, and residual spread are measured in $\log_{10}$ power units.

Parameter	10th percentile	Median	90th percentile
Peak center μ (Hz)	6.596	17.042	31.278
Peak height h	0.196	0.323	0.588
Peak width σ (Hz)	0.326	0.768	3.113
Background slope β	1.211	1.716	2.270
Background level b	1.016	2.182	3.187
Residual spread s	0.098	0.112	0.127

Of low-band fits, 81.9% pass quality checks, while 86.7% reach the six-peak fitting limit. The resulting ranges are a controlled low-frequency test, not a validated account of all physiological oscillations or high-gamma activity. Residual spread combines fitting error, unresolved structure, and noise; it is not a measurement-noise estimate. None of the four models predicts this sensitivity target reliably.

Encoders and readouts. The four models share a width-512 architecture, pretraining corpus, sampling procedure, and training budget. Reconstruction uses 50% masking during pretraining, whereas electrode-origin prediction uses unmasked inputs. Selected peak learning rates are 0.02 for electrode-origin prediction and joint reconstruction, and 0.01 for waveform-only and Fourier-only reconstruction. Thus, the comparison concerns these training recipes rather than the loss function independently of masking and optimization choices.

At evaluation, features from the first main Transformer block are averaged over the full five-second signal. No masking or electrode identity labels are supplied. A five-fold shuffled split assigns 800 signals to training and 200 to testing in each fold. Feature standardization and whitened principal components retaining 99% of variance are fitted on the outer training set. Five-fold cross-validation within those transformed training features selects ridge regularization from $\{0.01, 0.1, 1, 10, 100\}$. The outer test set is excluded from preprocessing and regularization selection. Reported R^2 is the mean of the five test-fold scores, each calculated relative to its test-target mean. Negative scores are retained. These results use one signal seed (20260923) and one checkpoint per objective (seed 41); the five folds do not quantify variability across pretraining runs.

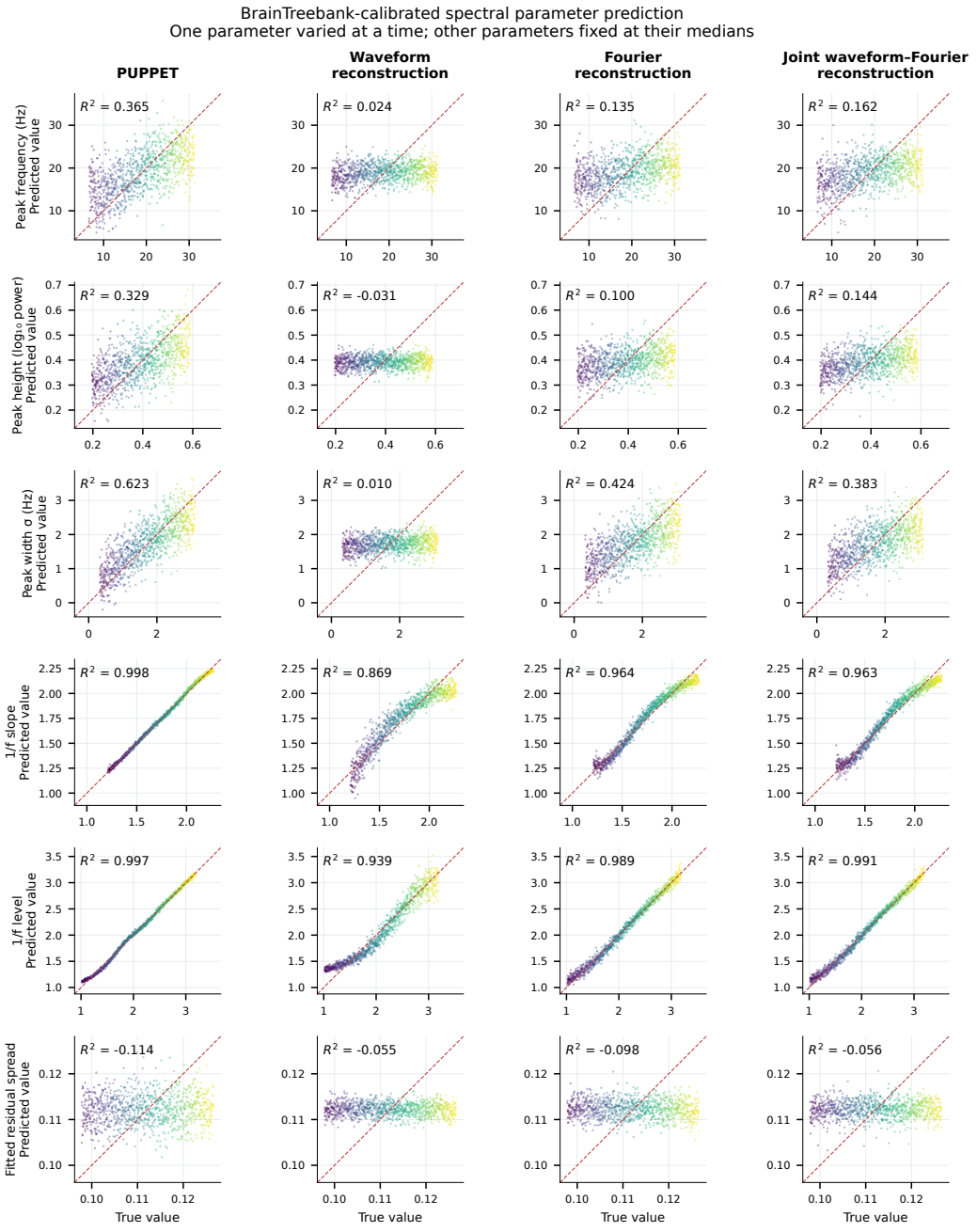


Figure 7: Spectral parameters recovered from frozen representations. True versus held-out predicted values for the BrainTreebank-calibrated sweeps. Columns compare electrode-origin prediction (PUPPET), waveform-only, Fourier-only, and joint waveform–Fourier reconstruction, using the same encoder architecture and pretraining data. Rows vary peak frequency, height, width, background slope, background level, and fitted residual spread separately; other parameters remain at their medians. Peak centers span 6.596–31.278 Hz and widths span 0.326–3.113 Hz, with medians of 17.042 Hz and 0.768 Hz. Dashed diagonals indicate perfect prediction; numbers report mean held-out-fold R^2 across five folds for 1,000 five-second signals, using one signal seed and one checkpoint per trained model. PUPPET scores higher than all three reconstruction controls on the five spectral parameters. No model reliably predicts fitted residual spread, which is not a pure measurement-noise estimate. This full-window readout differs from the cross-model probe in the main figure; absolute scores should be compared within each analysis. This subsection gives preprocessing, calibration, and control details.

F.2 CROSS-MODEL COMPARISON WITH DOWNSTREAM PERFORMANCE

The cross-model comparison evaluates PUPPET, BrainBERT, PopT-v1, PopT-v2, CBraMod, LaBraM, EEGPT, BIOT-PREST, REVE-base, BrainOmni-tiny, SPP-EEGNet, and BaRISTA, as well as waveform, Fourier, and joint waveform–Fourier reconstruction. The three controls use the same evaluation as the twelve pretrained systems; their values are not imported from the full-window comparison in the preceding subsection. It asks whether accessible spectral information accompanies useful downstream features under a common diagnostic. This includes scalp-EEG models adapted to iEEG and is separate from the paper's primary benchmark ranking.

Input and representation conventions. All systems receive the same one-second signal segments. Inputs are resampled to each encoder's sampling rate. Except for PopT-v1, PopT-v2, and BaRISTA's population attention, channels are encoded separately; task readouts concatenate their features. Table 14 specifies preprocessing choices that can affect the recoverability of spectral parameters. In particular, per-window normalization can remove amplitude information. Cross-model differences therefore concern the complete preprocessing and encoder pipeline, and cannot be attributed solely to learned spectral preferences. Fixed channel or position inputs are interface conventions, not anatomical labels for the iEEG electrodes.

Table 14: Encoder interfaces in the fifteen-system comparison. Rates are in Hz. Amplitude factors refer to input waveforms in microvolts.

Model	Input and feature convention
PUPPET and three reconstruction controls	500 Hz; divide by 200; first-block temporal mean.
BrainBERT	2048 Hz; per-frequency, per-window STFT standardization; mean of ten central encoded frames.
PopT-v1	BrainBERT frontend; population attention with fixed zero spatial inputs; per-electrode output tokens.
PopT-v2	2048 Hz; multi-window STFT magnitude, standardized per channel and frequency using training data only; zero spatial inputs; population attention and per-electrode output tokens.
CBraMod	200 Hz; divide by 100; single-channel, one-second patch; mean pre-head features.
LaBraM	200 Hz; divide by 100; first pretrained channel token; mean patch features.
EEGPT	256 Hz; divide by 100; fixed CZ token; temporal mean of encoder summary features.
BIOT-PREST	200 Hz; divide by 100; first pretrained channel token; encoder summary.
REVE-base	200 Hz; divide by 100; fixed zero position; mean features.
BrainOmni-tiny	256 Hz; divide by 100; fixed zero position and EEG sensor type; tokenizer pads one second to two; mean latent features.
SPP-EEGNet	500 Hz; per-window zero mean and unit variance; spatial-pyramid features.
BaRISTA	2048 Hz; per-window/channel standardization; parcel-based encoder with unknown-region embedding; population attention; four time tokens per channel.

BaRISTA uses the public `parcels_chans.ckpt` weights from the authors' repository, with all parameters loaded strictly and frozen. Its one-second input produces four 512-sample patches per channel; all four output tokens are retained. Synthetic and BYD inputs use the model's unknown-region embedding because anatomical labels are not supplied. This is an explicit transfer adaptation, not the paper's fine-tuned BaRISTA benchmark result.

PopT-v2 uses the supplied pretrained six-layer, 512-dimensional encoder (19.57 million parameters). Its three STFT windows span 0.5, 0.25, and 0.125 seconds and retain 2–40, 20–150, and 80–250 Hz, respectively. Normalization statistics are fitted separately for each training split, pooling training windows and time frames within each channel and frequency bin. Synthetic normalization uses only the 600 training signals' one-second segments; downstream normalization uses only the training windows of that task and fold. Validation and test inputs do not contribute to either fit. Spatial inputs are fixed at zero, as for PopT-v1. The common bipolar signals and filters are retained for this comparison. These are new frozen-encoder measurements on the same splits as the other systems, not published fine-tuned benchmark scores.

Spectral prediction. Each five-second synthetic waveform is encoded as five separate one-second segments and their features are averaged. Height and width use the calibrated ranges in Table 13. Background sweeps have no peak: slope varies from 1 to 2 with level fixed at 2, and level varies from 0.1 to 3 with slope fixed at 1.5. Their log-power perturbation has standard deviation 0.005, rather than the calibrated residual spread used in the peak sweeps. Consequently, scores across the four targets need not reflect equal difficulty. The same fixed waveform gain of 200 is applied to all synthetic inputs before the model-specific amplitude conventions in Table 14.

For each target and input seed, 1,000 signals are partitioned into five shuffled folds. Each rotation uses 600 training, 200 validation, and 200 test signals. Training-set feature and target standardization precedes ridge fitting; validation mean-squared error selects regularization from $10^{-6}, 10^{-5}, \ldots, 10^{2}$. No test targets select the readout. Scores average test-fold R^2 over three input seeds (20260923–20260925), and additionally three checkpoints for each of PUPPET and the waveform, Fourier, and joint reconstruction controls (41–43); other models use one checkpoint each. Unlike the matched-objective comparison, this diagnostic averages separately encoded segments and uses a different readout procedure and background sweeps. Absolute scores should be compared within each analysis.

Downstream evaluation and averaging. The four BYD tasks use one-second bipolar signals at 500 Hz, a 1-Hz high-pass filter, and 60/120/180-Hz notch filters. All encoders remain frozen. Ridge readouts use training-only feature standardization and validation AUROC to select regularization from $10^{-6}, 10^{-5}, \ldots, 10^{2}$. Two temporal splits per recording retain a two-second gap between training, validation, and test windows. Folds, recordings, and available checkpoint seeds are averaged within participant before averaging participants within task. Speech, pitch, volume, and word index include 16, 15, 16, and 10 participants, respectively, across 29 recordings in total. The plotted ordinate equally averages these four task means; it was chosen as a descriptive summary after task results were available. It is not the primary 15-task BYD score, which uses different preprocessing, task splits, and logistic readouts.

Table 15: Values in the fifteen-system scatter. The four spectral columns are held-out R^2; AUROC is the equal-weight four-task BYD mean.

Model	Slope	Level	Height	Width	AUROC
PUPPET	0.9992	0.9985	0.3129	0.5929	0.6843
BrainBERT	0.5195	-0.0225	-0.0224	-0.0220	0.5650
PopT-v1	0.5568	-0.0221	-0.0236	-0.0243	0.5476
CBraMod	0.9807	0.9972	-0.0232	0.2232	0.6370
LaBraM	0.9801	0.9890	-0.0158	0.0048	0.6067
EEGPT	0.9939	0.9993	0.3633	0.5824	0.6259
BIOT-PREST	0.9881	0.9823	0.3261	0.5534	0.6567
REVE-base	0.8938	0.6346	-0.0218	-0.0214	0.6127
BrainOmni-tiny	0.9381	0.9907	-0.0194	0.0542	0.5765
SPP-EEGNet	0.9120	-0.0214	-0.0232	0.0092	0.5741
Waveform reconstruction	0.9647	0.9637	-0.0210	0.1120	0.6544
Fourier reconstruction	0.9825	0.9849	0.1124	0.4171	0.6703
Joint reconstruction	0.9844	0.9789	0.0602	0.3509	0.6706
BaRISTA	0.9637	-0.0216	-0.0191	-0.0230	0.6269
PopT-v2	0.9997	0.9999	0.4930	0.7281	0.6657

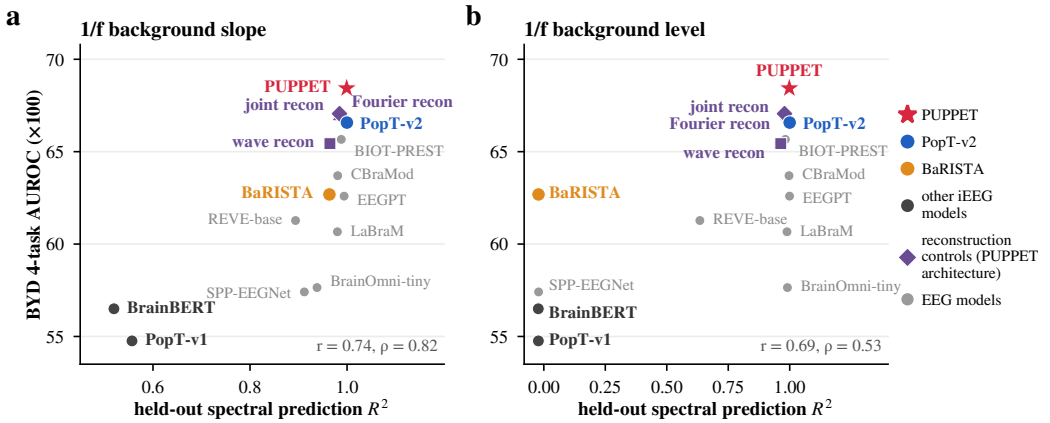


Figure 8: Background prediction and downstream performance. The fifteen systems and four-task BYD scores are the same as in the main peak-height and peak-width figure. Panels show held-out R^2 for (a) $1/f$ background slope and (b) $1/f$ background level against mean test AUROC ($\times 100$). Reconstruction controls (purple) use PUPPET's encoder trained to reconstruct the masked waveform (wave recon), its Fourier spectrum (Fourier recon), or both (joint recon). Other iEEG foundation models are dark grey and EEG foundation models light grey (SPP-EEGNet is an EEG-pretrained baseline). r and ρ are Pearson and Spearman correlations across the fifteen systems (descriptive). Scores average five folds and three synthetic input seeds, with three checkpoints for each matched objective. Each task is averaged across participants before averaging the four tasks. Near-perfect background prediction does not ensure the highest task performance. Points and correlations are descriptive; related models are not independent replications.

F.3 Sensitivity and interpretation of the association

Each system contributes one point. Across the fifteen systems, Pearson correlations for slope, level, height, and width are 0.735, 0.692, 0.568, 0.744; Spearman correlations are 0.821, 0.529, 0.668, 0.768. Table 16 separates the four tasks and checks the result of omitting each system. Because BrainBERT and PopT-v1 share a frontend, we also omit that pair together. These are descriptive checks of the same results, not independent experiments or confidence intervals.

Table 16: Pearson correlations by task and after model omissions. The last two columns use the four-task mean. "One omitted" spans fifteen single-system omissions; "Pair omitted" excludes BrainBERT and PopT-v1.

Target	Speech	Pitch	Volume	Word index	One omitted	Pair omitted
Slope	0.72	0.69	0.70	0.36	0.63–0.79	0.72
Level	0.69	0.67	0.72	0.02	0.61–0.80	0.48
Peak height	0.60	0.60	0.60	-0.27	0.51–0.62	0.53
Peak width	0.77	0.78	0.78	-0.25	0.70–0.79	0.71

Omitting PUPPET and all three reconstruction controls together gives Pearson correlations of 0.75, 0.63, 0.69, 0.75 for slope, level, height, and width. These checks account for obvious groups of related systems without making the remaining architectures independent replications.

The aggregate correlations remain positive in every single-system and shared-frontend omission check. However, peak associations are positive for speech, pitch, and volume and negative for word index. The aggregate trend therefore does not imply a positive association for every task. Nor does it establish a relationship across all foundation models: architecture, pretraining data, feature dimension, and input normalization differ within this fixed set. Fold and seed averages are descriptive point estimates; related models are not independent replications.

Background-level prediction is near its $R^2 = 1$ ceiling for both PUPPET and EEGPT despite different downstream AUROC. This shows that near-perfect prediction of one spectral parameter does not determine task performance. It does not estimate a saturation law or show that spectral recovery causes the downstream advantage. The synthetic probes measure accessibility to a specified readout, not exclusive encoding of physiological oscillations.

G Representation and Attribution Analyses

What pretraining changes. To separate learned features from architectural features, we compare the canonical PUPPET encoder with a genuinely untrained random-weight checkpoint of the same width-512 architecture. Both are frozen and evaluated with the paper's logistic readout on iMind-Bench Main units: 49 Neuroprobe, 254 BYD, and 44 Pippi. Figure 9 reports descriptive task-level AUROC differences. Speech and onset gain across all three datasets, while several fine-grained word-level tasks change little; these unit averages are not patient-bootstrap significance tests.

	Neuroprobe	BYD	Pippi
volume	**63.7 → 76.6 +12.9**	**58.1 → 70.6 +12.5**	50.8 → 53.7 +2.9
pitch	**56.5 → 62.5 +6.0**	**55.0 → 64.3 +9.3**	57.0 → 52.6 -4.4
delta volume	80.1 → 82.0 +1.8	60.7 → 60.7 -0.0	*not in Main*
speech	**76.1 → 90.0 +13.9**	**65.0 → 81.6 +16.6**	**61.4 → 80.9 +19.5**
onset	**88.0 → 93.2 +5.2**	**64.3 → 78.5 +14.2**	**62.3 → 76.8 +14.5**
word index	82.6 → 85.1 +2.5	59.3 → 56.4 -3.0	**54.0 → 71.4 +17.4**
word part speech	62.7 → 64.7 +2.0	51.5 → 52.8 +1.4	**50.7 → 57.6 +6.9**
gpt2 surprisal	61.1 → 65.5 +4.4	60.6 → 59.5 -1.2	51.3 → 52.4 +1.2
word head pos	61.1 → 64.0 +2.9	52.9 → 53.6 +0.7	54.1 → 54.1 +0.0
word gap	66.6 → 67.6 +1.0	51.9 → 53.7 +1.8	45.6 → 43.9 -1.7
word length	64.1 → 65.6 +1.5	52.8 → 49.6 -3.2	**57.1 → 52.0 -5.1**
global flow	**60.1 → 72.1 +12.0**	51.2 → 55.1 +3.9	**46.0 → 54.0 +8.0**
local flow	**55.9 → 61.6 +5.7**	53.5 → 56.5 +3.0	**58.8 → 72.4 +13.5**
face num	**51.4 → 57.1 +5.6**	**56.3 → 72.3 +16.0**	50.8 → 49.8 -1.0
frame brightness	52.3 → 57.0 +4.7	49.2 → 52.1 +2.9	**48.4 → 37.7 -10.7**
Main units	**68.6 → 74.2 +5.6 (98% units up)**	**57.3 → 65.1 +7.8 (79% units up)**	**53.4 → 59.2 +5.8 (68% units up)**

cell: untrained → PUPPET AUROC (%), gain from pretraining · paper logistic, Main units · task colour = domain

Figure 9: Pretraining gain by task. Each cell gives the AUROC (×100) of a genuinely untrained encoder and PUPPET, both frozen with the same logistic readout, followed by their difference. Both models are scored on iMindBench Main units (Neuroprobe 49, BYD 254, Pippi 44); Pippi has no Main unit for delta volume. Task names are colored by auditory, language, or visual domain. The differences are descriptive unit means, not significance tests.

Frozen embedding views. Figure 10 compares six Neuroprobe evaluation units across multi-STFT input, DIVER-1, and PUPPET under a common frozen-feature readout. These selected single-unit views show how linear accessibility varies across tasks and representations; they are descriptive and should not be read as the aggregate benchmark comparison.

Readout-free spatial allocation. PopT previously used attention rollout for regional interpretation (Chau et al., 2025; Abnar & Zuidema, 2020). Here we show only the task-free rollout from the canonical seed-42 PUPPET checkpoint on BYD movie windows (Figure 11). Each electrode's absolute rollout attribution is converted to a centered rank within its window, so lower or higher values mean relative model allocation, not neural suppression or activation. In this exploratory map, left anterior cingulate cortex and left amygdala have lower centered ranks, while left inferior frontal gyrus, pars orbitalis, has a higher rank. The map does not establish anatomical causality or explain the objective's transfer advantage.

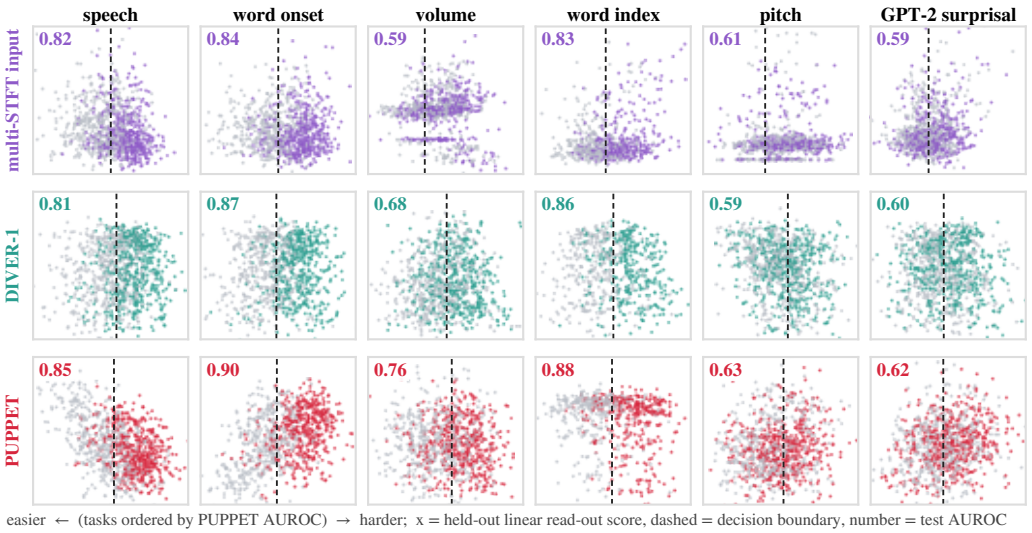


Figure 10: Held-out readout views across tasks and representations. Six selected Neuroprobe tasks are shown for multi-STFT input, frozen DIVER-1 features, and frozen PUPPET features, one evaluation unit per task. The horizontal axis is the training-fitted logistic readout score with a dashed decision boundary; the vertical axis is the leading orthogonal direction. Numbers in the panels are test AUROC. These examples illustrate varying class separation, not a patient-level significance comparison.

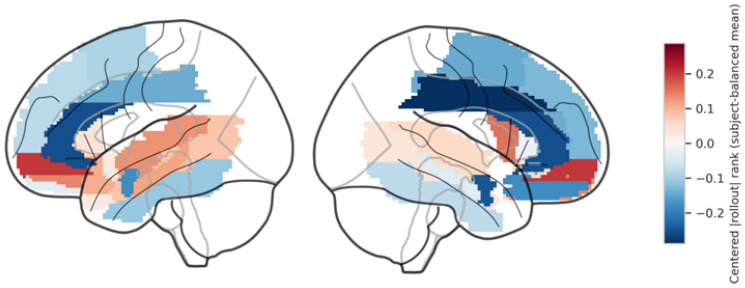


Figure 11: Readout-free backbone attention rollout on BYD. The displayed panel contains only the canonical seed-42 encoder's task-free spatial rollout, without integrated-gradient task maps. Color shows centered within-window electrode-attribution rank, not neural activity or a causal regional effect. This one-checkpoint analysis is exploratory.

H EXTENDED SCALING

Label-efficiency coverage and sampling. Figure 3 uses 75 Neuroprobe, 435 BYD, and 50 Pippi All evaluation units scoreable for both PUPPET and the multi-STFT MLP at every label fraction. At each fraction, their readouts use exactly the same stratified training-trial subsets, with three draws per fraction. The appendix Figure 19 shows the larger PUPPET-only grid across pretraining scales; its 55 Pippi units need only be scoreable for PUPPET. This downstream label-scaling analysis is distinct from the pretraining-data sweep in Figure 5.

Data subsets and training budget. The record-count and corpus-fraction sweeps use different sampling procedures. For record-count rungs, we first form the seeded training–validation split, then select a seeded subset of training records without changing validation. Subsets at different rungs are not guaranteed to be nested. Each 30-s record contributes 0.5 min of unique data. Every active record-count run trains for 704 updates, sampling examples with replacement; repeated sampling does not increase unique-data duration. The transferred `best.pt` is selected on the fixed validation set (329 records, about 2.74 h). For PUPPET rungs of at most 30 min, this checkpoint occurs after 44–308 updates, whereas the corresponding reconstruction checkpoints are selected at the final update. Thus the runs share a training budget, not necessarily the number of updates in the transferred checkpoint.

Corpus-fraction rungs instead use the ordinary shuffled loader and adjust the number of epochs as the training set shrinks. The dataset implementation shuffles record keys with seed 42 before retaining the requested fraction; the subsequent training–validation split uses the run seed. Thus these rungs do not inherit the record-count sweep's unchanged validation partition. The launch schedule specifies 704, 352, 235, 141, 101, and 79 epochs for fractions 0.125, 0.25, 0.28, 0.50, 0.72, and 0.86, respectively; the full-data run uses 64 epochs. Training logs show 704, 704, 705, 705, 808, and 711 updates for those original fraction rungs, respectively, across all three seeds. The separately corrected fraction-0.72 checkpoint uses 704 updates but exists only for seed 42; the three-seed row in Table 20 retains the original 808-update runs. Accordingly, the exploratory fraction sweep is not strictly step-matched, unlike the 704-update record-count sweep.

Width and data results. Tables 17 and 20 transcribe the current electrode-origin scaling results; Tables 18 and 19 add the matched joint-reconstruction ladders. The transferred encoder depth is fixed at one; the additional prediction block is used only during pretraining. Widths 8 and 16 use one attention head, while widths 32–1024 use head dimension 32. These tables do not constitute a completed four-objective width-and-data scaling grid.

Matched joint reconstruction improves as model width increases, but its data-scaling curve plateaus earlier than PUPPET's within the tested range. PUPPET trails joint reconstruction at the shared ten-minute rung and leads at one hour; there is no shared rung between them. Reconstruction widths 32, 128, and 512 have three seeds, while other rungs currently have one, so the objective-wise trends are descriptive. Appendix Figures 12 and 13 show the absolute scores and reconstruction-specific curves on the full All subset. Reconstruction loss is not numerically comparable with electrode-identification cross-entropy.

Domain-wise scaling. Main-text Figure 6 shows changes in domain-level AUROC relative to five minutes of pretraining or the smallest model. Neuroprobe and BYD auditory tasks gain most along both axes; Pippi's auditory tasks remain near chance across these rungs, leaving a nearly flat auditory trajectory. The numbers of evaluation units (visual/auditory/language) are 20/15/40 for Neuroprobe, 116/87/232 for BYD, and 20/15/40 for Pippi. The pooled "all three" panel combines dataset–task pairs.

Task-wise scaling. Figure 14 summarizes descriptive per-task slopes. Figure 15 shows exploratory fitted ceilings, and Figures 16–18 show three-seed AUROC trajectories. Task curves are heterogeneous, with the weakest median data-scaling slope on Pippi. A fitted ceiling describes extrapolation within this model family; a flat or noisy task curve does not establish an intrinsic limit on iEEG decoding.

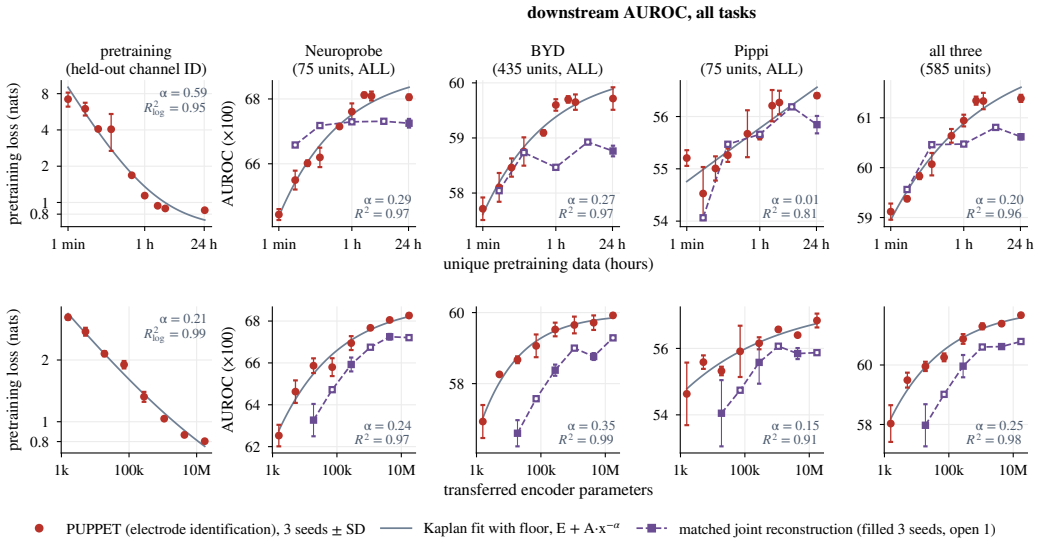


Figure 12: Dataset-level scaling for PUPPET and matched joint reconstruction. Rows vary unique pretraining duration and transferred encoder parameters; columns show held-out electrode-identification loss, downstream Neuroprobe, BYD, and Pippi AUROC, and their equally weighted mean. PUPPET points average three seeds with SD bars; dashed reconstruction trajectories show three-seed rungs as filled markers and one-seed rungs as open markers. Solid curves fit an irreducible-floor power law to PUPPET's loss or downstream error. All downstream scores use the All subset (75 Neuroprobe, 435 BYD, 75 Pippi evaluation units).

Table 17: Model-size scaling on iMindBench Main, in AUROC points. Model size is varied through width; rows use seeds 41, 42, and 43, and ± denotes their sample standard deviation. Sources: T64/T67.

Width	Heads	Neuroprobe	BYD	Pippi	iMB Overall
8	1	67.73	60.95	58.48	62.39 ± 0.75
16	1	70.19	62.69	58.27	63.71 ± 0.24
32	1	71.77	63.58	58.10	64.48 ± 0.24
64	2	71.60	63.65	58.72	64.66 ± 0.19
128	4	72.86	64.61	58.92	65.46 ± 0.21
256	8	73.73	64.75	59.05	65.84 ± 0.26
512	16	74.19	64.94	59.28	66.13 ± 0.03
1024	32	74.42	65.22	59.70	66.44 ± 0.05

Table 18: Matched model-size scaling on iMindBench Main Overall (AUROC points). Joint reconstruction uses the same encoder widths and evaluation protocol as electrode-origin prediction. A dagger marks a one-seed reconstruction estimate; other entries use three seeds. Values are descriptive means, not a model-size significance test.

Width	Electrode origin	Joint reconstruction	Difference
32	64.48	61.39	+3.09
64	64.66	$62.56^{\dagger}$	+2.10
128	65.46	63.90	+1.57
256	65.84	$64.62^{\dagger}$	+1.22
512	66.13	64.79	+1.35
1024	66.44	$65.13^{\dagger}$	+1.32

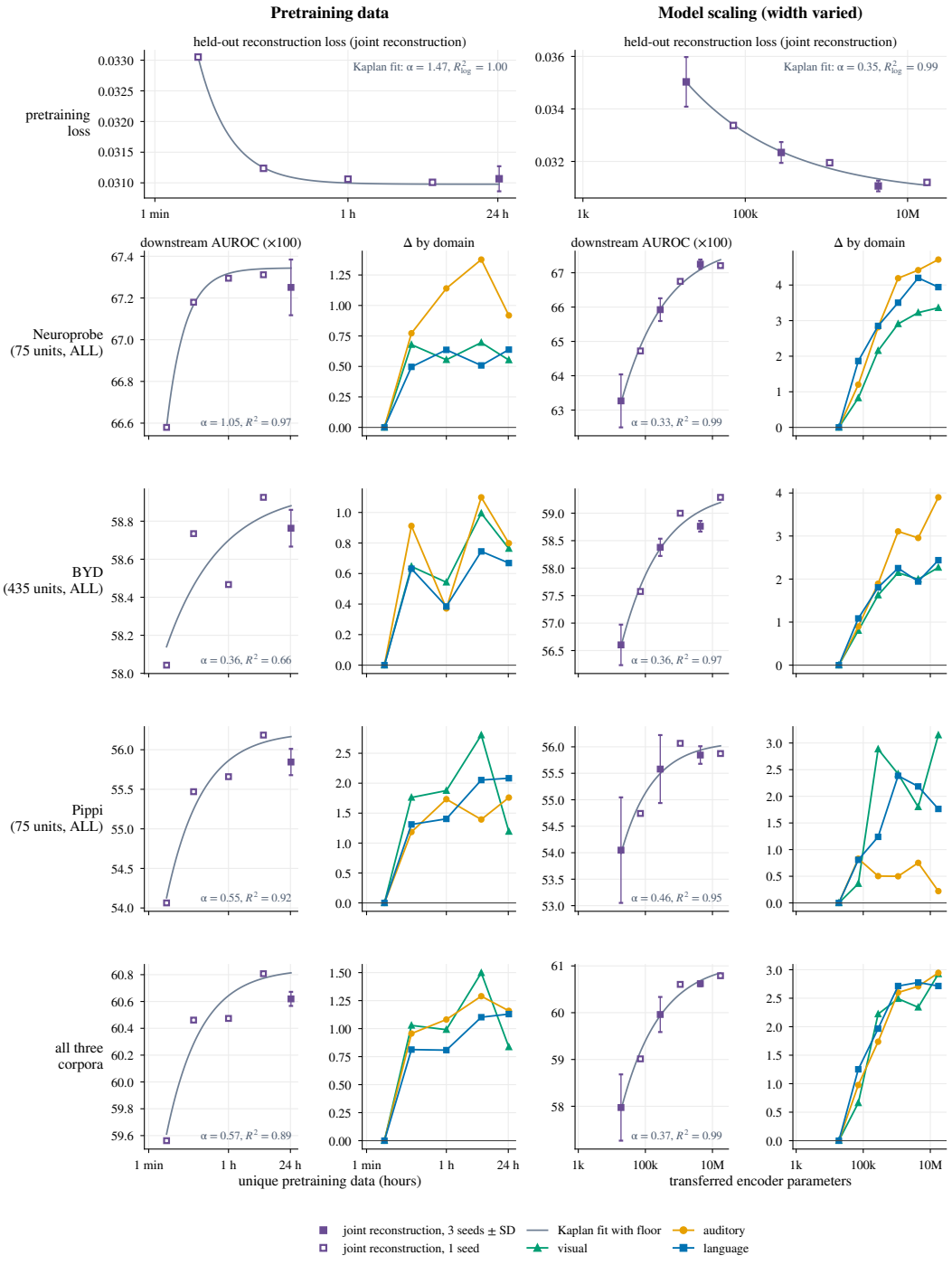


Figure 13: Matched joint waveform–Fourier reconstruction scaling. Columns vary unique pretraining duration and model size. The top row shows held-out reconstruction loss; subsequent rows show dataset-level AUROC with irreducible-floor fits and changes by auditory, language, and visual domain. Open markers denote one-seed reconstruction rungs; filled markers with bars denote three-seed rungs. All downstream scores use the All subset (75 Neuroprobe, 435 BYD, 75 Pippi evaluation units). Reconstruction loss is not numerically comparable with PUPPET's electrode-identification loss.

Table 19: Matched pretraining-data ladder on iMindBench Main Overall (AUROC points). Electrode-origin results average three seeds; joint reconstruction has one seed except at full data, which has three. Negative differences at short durations show that the ten-minute efficiency result is not unique to electrode-origin prediction.

Unique pretraining data	Electrode origin	Joint reconstruction	Difference
2.5 min	63.67	63.83†	−0.16
10 min	64.67	64.81†	−0.14
1 h	65.62	64.55†	+1.07
3.1 h	66.11	65.04†	+1.07
6 h	66.24	65.02†	+1.23
12.4 h	66.20	64.85†	+1.35
24.7 h	66.13	64.79	+1.35

Table 20: Pretraining-data scaling on iMindBench Main, in AUROC points. Means and sample standard deviations use three seeds (41/42/43). Fractions are retained as source labels rather than converted into unverified exact hours; their update counts, including the 808-update fraction-0.72 exception, are specified above. Sources: T65/T67.

Unique-data rung	Neuroprobe	BYD	Pippi	iMB Overall
1 min (2 records)	70.03	61.64	57.79	63.15 ± 0.25
2.5 min (5 records)	71.28	62.71	57.01	63.67 ± 0.16
5 min (10 records)	71.79	63.26	57.80	64.29 ± 0.15
10 min (20 records)	72.00	63.77	58.23	64.67 ± 0.25
30 min (60 records)	73.03	64.37	58.74	65.38 ± 0.11
1 h (120 records)	73.62	64.84	58.40	65.62 ± 0.19
2 h (240 records)	74.15	64.96	59.27	66.12 ± 0.03
3 h (360 records)	74.18	64.91	59.21	66.10 ± 0.21
Fraction 0.25	74.21	65.01	59.38	66.20 ± 0.14
Fraction 0.28	74.19	64.96	59.44	66.19 ± 0.24
Fraction 0.50	74.23	64.86	59.50	66.20 ± 0.10
Fraction 0.72	74.26	64.92	59.50	66.23 ± 0.10
Fraction 0.86	74.19	65.02	59.34	66.18 ± 0.14
Full corpus	74.19	64.94	59.28	66.13 ± 0.03

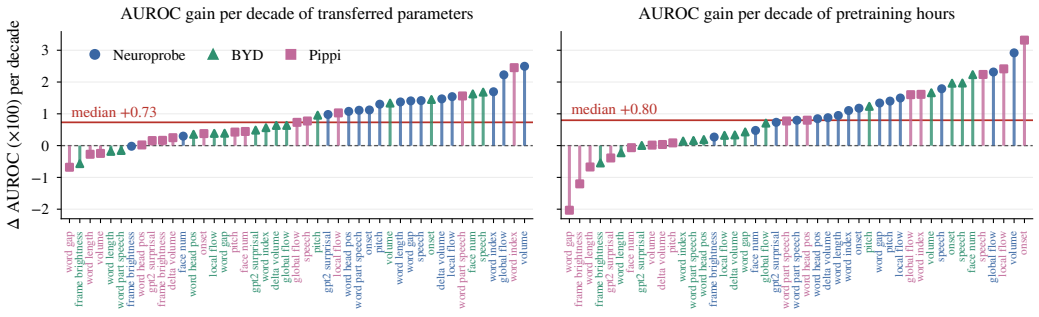


Figure 14: Individual-task scaling slopes. Each mark represents one dataset–task series on iMindBench All. Slopes summarize AUROC change per decade of transferred encoder parameters or unique pretraining duration, using three pretraining seeds. The task-level responses are heterogeneous; these slopes are descriptive rather than evidence that every task follows a common scaling law.

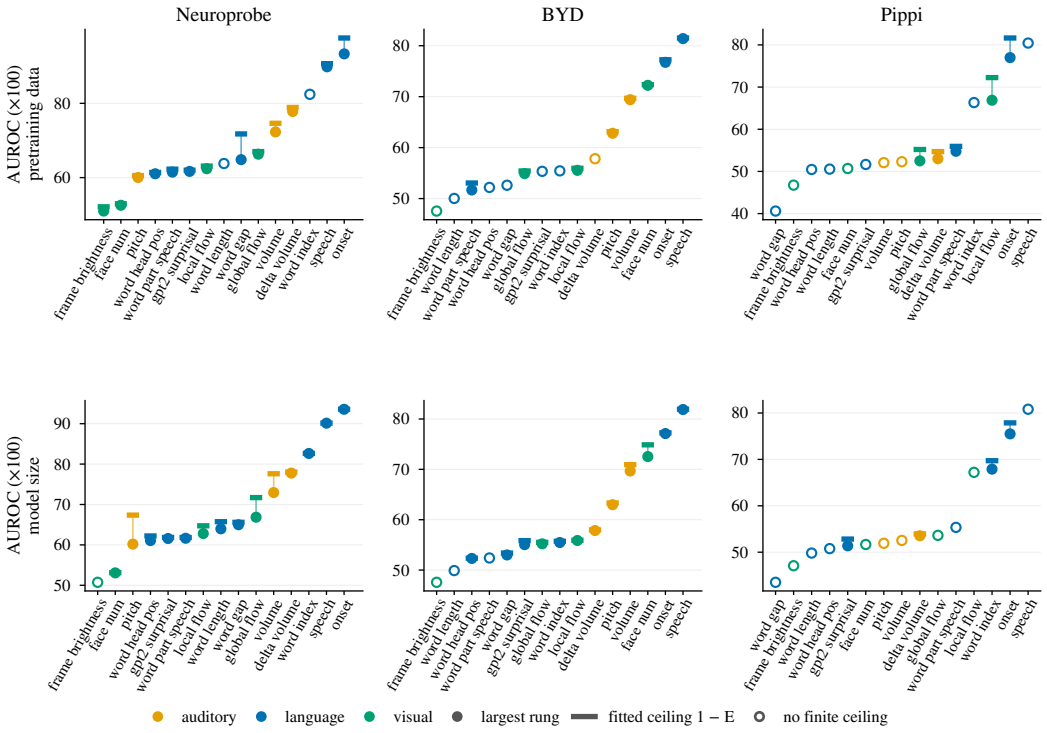


Figure 15: Exploratory per-task Kaplan fits. Rows vary pretraining duration and model size; columns show Neuroprobe, BYD, and Pippi. Each colored dot is the three-seed score at the largest rung, and a short bar marks the fitted ceiling $1 - E$. Open rings denote tasks without a finite fitted ceiling under this model. Ceilings differ substantially across tasks; for most tasks with a measurable trend, the largest model lies close to its fitted ceiling. These extrapolations are specific to the tested model family and are not limits on neural decodability. All fits use the All subset (75, 435, and 75 evaluation units, respectively).

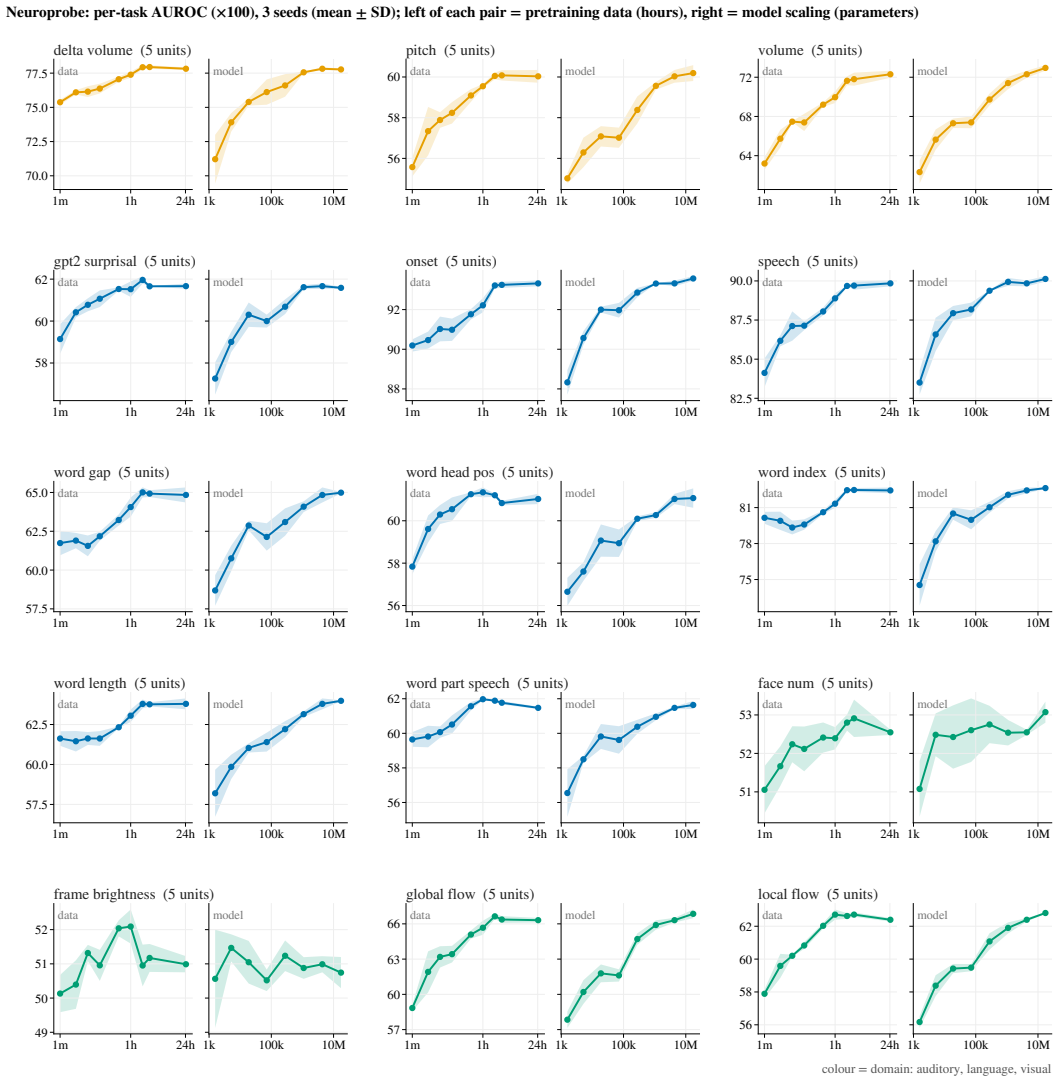


Figure 16: Neuroprobe task-wise trajectories. Each task has a pretraining-data panel and a model-scaling panel; lines and bands show three-seed mean ± SD AUROC. These are descriptive trajectories, not tests of task-specific scaling. This analysis uses all released evaluation units (Neuroprobe 75, BYD 435, Pippi 75).

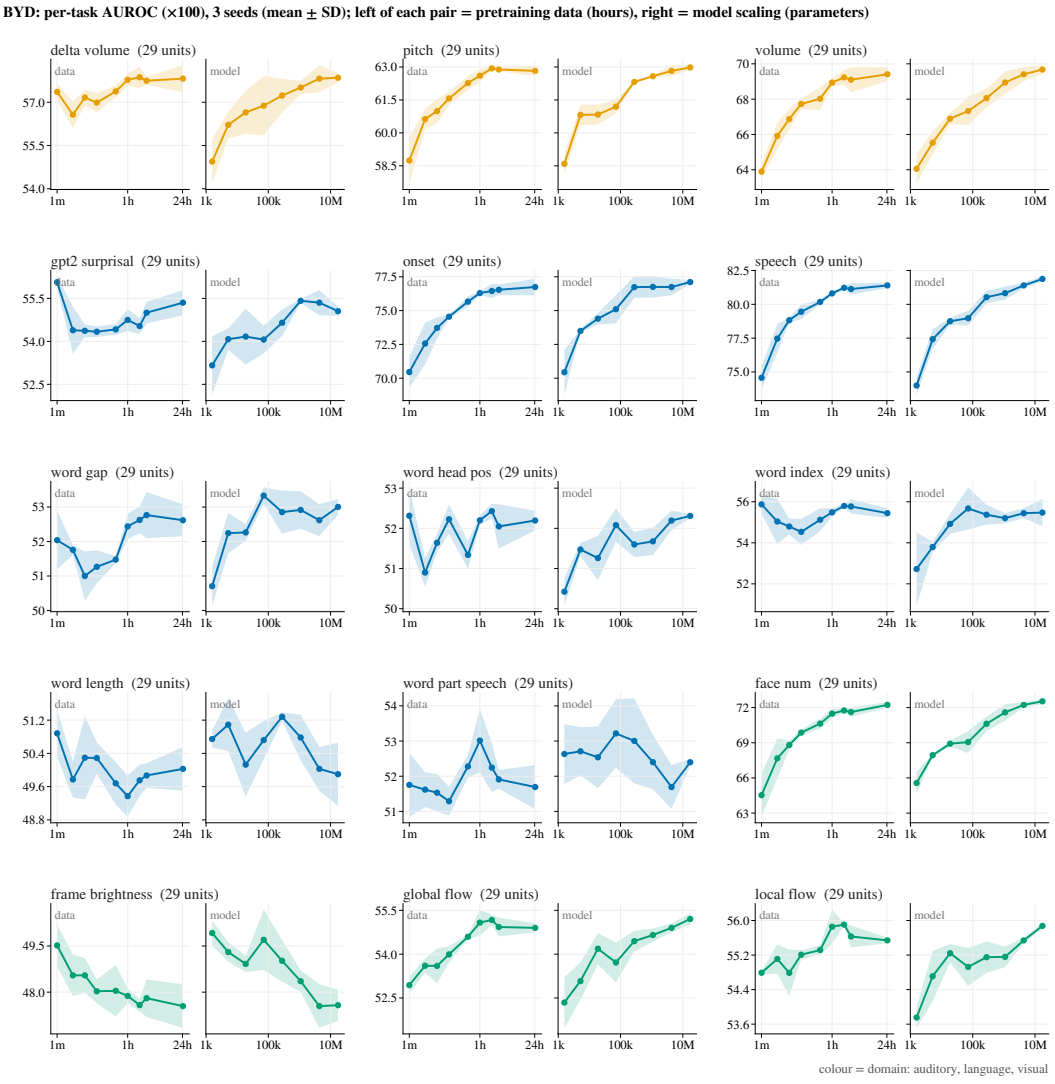


Figure 17: BYD task-wise trajectories. Same three-seed model- and data-scaling protocol as Figure 16; task panels use their own y-axis scales. This analysis uses all released evaluation units (Neuroprobe 75, BYD 435, Pippi 75).

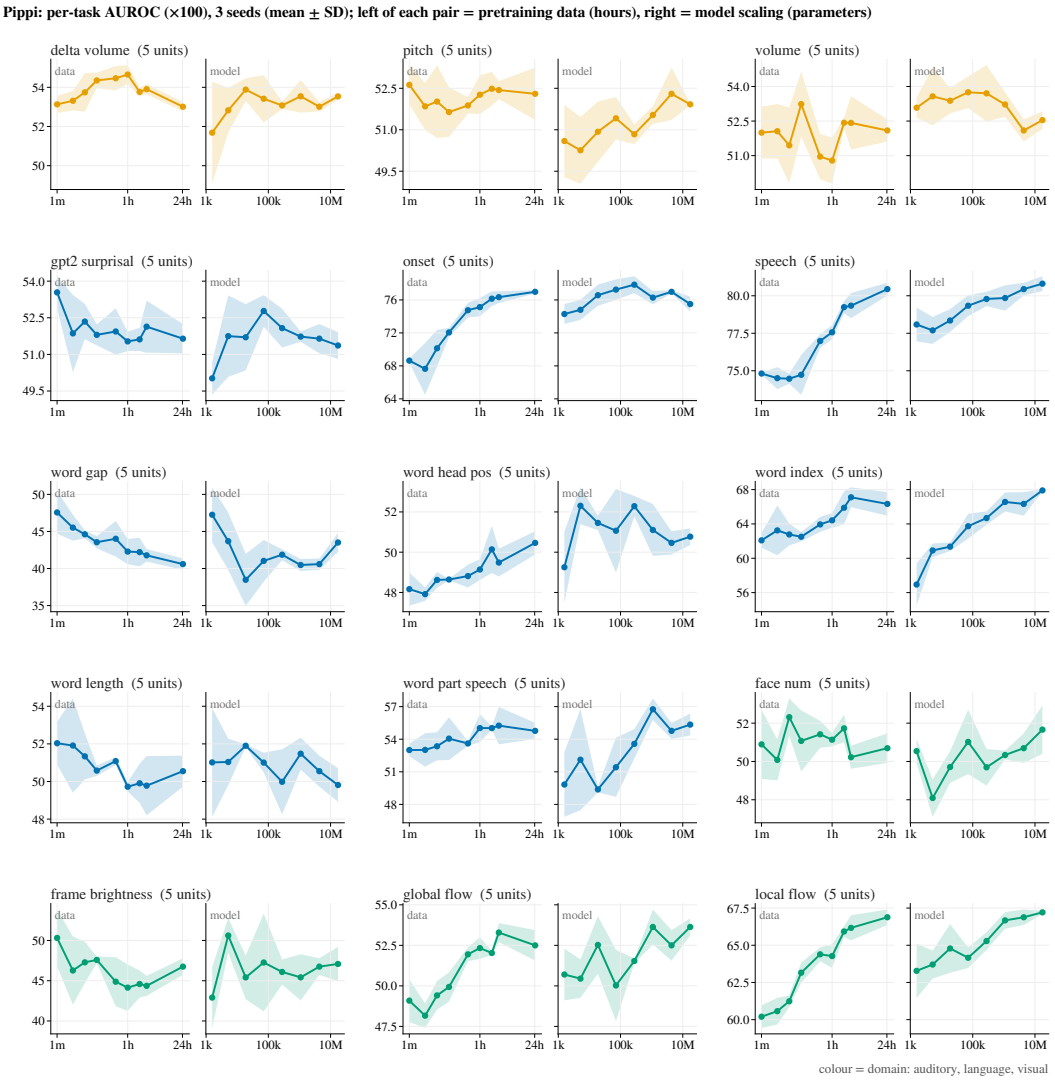


Figure 18: Pippi task-wise trajectories. Same three-seed model- and data-scaling protocol as Figure 16. The weak and sometimes negative data slopes here prevent a general per-task scaling claim. This analysis uses all released evaluation units (Neuroprobe 75, BYD 435, Pippi 75).

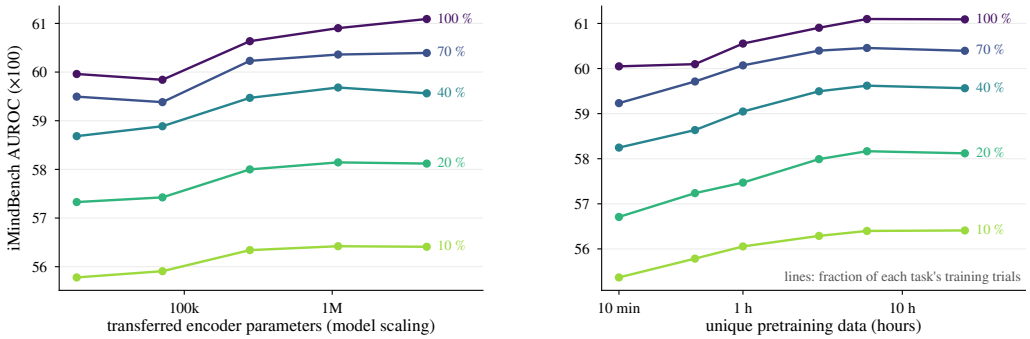


Figure 19: Full labeled-data fraction by pretraining scale. Downstream AUROC as the frozen readout uses 10, 20, 40, 70, or 100% of training trials across the model-size and unique-pretraining-data ladders. This expanded one-seed view complements the main-text Figure 3; three stratified trial draws are used at each fraction. The analysis uses 75 Neuroprobe and 435 BYD units; 55 Pippi units are scoreable for PUPPET at every fraction, versus 50 scoreable for both PUPPET and the MLP in Figure 3. The curves are descriptive, not significance tests.

Power-law analysis. Following Kaplan et al. (2020), we fit pretraining loss and downstream error $1-\mathrm{AUROC}$ with $E+Ax^{-\alpha}$ across the tested data and model-size rungs. The floor E bends the loss curves on log–log axes; only $L-E$ is straight. On the All subset and the final displayed PUPPET ranges (at least 5 min or width 8), all aggregate and dataset-level PUPPET fits have $R^2 \geq 0.91$ (Figure 5). The fitted $1-E$ values are model-family extrapolations, not task-independent limits. Main-text Figure 6 shows domain changes; Appendix Figures 14, 15, and 16 show individual-task slopes, fitted ceilings, and trajectories.

Scope of supplementary scaling. Subject-count and epoch ladders are not additional scaling axes in this paper. The label-efficiency analysis varies downstream training labels, not pretraining data.