iMINDBENCH: iEEG MULTI-INSTITUTION NEURAL DECODING BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Intracranial electroencephalography (iEEG) records electrical activity directly from the human brain, making it attractive for neural decoding. However, progress toward general-purpose iEEG foundation models is difficult to measure because fragmented datasets and processing pipelines obscure whether model gains persist across institutions, arise from modeling rather than preprocessing, or grow with broader supervision. Thus, we introduce iMINDBENCH, an iEEG Multi-Institution Neural Decoding Benchmark that aligns fifteen naturalistic decoding tasks across three independently collected institutional datasets to measure progress across heterogeneous recordings. We define standardized spectral and waveform preprocessing tracks and find that pretrained systems generally outperform non-pretrained baselines within their respective tracks, yet an engineered spectral baseline remains competitive with pretrained systems, especially on the two datasets with less downstream training data. The benchmark's harmonized targets and electrode metadata also enable a matched-task scaling study across subjects and institutions. We find that additional same-dataset subjects yield modest gains, whereas adding data from other institutions provides little aggregate benefit despite up to a 25-fold expansion of the supervised dataset. These findings highlight the need for general-purpose iEEG decoders that deliver consistent gains across institutions and make more effective use of cross-institution supervision.

1 INTRODUCTION

Neural decoding seeks models that generalize across subjects, institutions, and recording conditions while requiring minimal patient-specific calibration. Intracranial electroencephalography (iEEG) is an attractive setting for this goal because it records neural activity directly from electrodes inside the human brain with high temporal resolution. However, progress in iEEG decoding remains difficult to evaluate because existing studies often differ not only in dataset and task selection, but also in preprocessing pipelines and split construction. As a result, apparent benchmark improvements can reflect changes in evaluation setup rather than advances in neural modeling itself.

Existing resources establish important parts of this problem. Multi-institution iEEG benchmarks such as Omni-iEEG (Duan et al., 2026) and Mayo/FNUSA (Nejedly et al., 2020) show that clinical recordings can be harmonized across sites, but primarily study seizure, pathology, high-frequency oscillation, or related epileptology targets. Naturalistic decoding resources such as Neuroprobe (Zahorodnii et al., 2026), Brain Treebank (Wang et al., 2024a), and AJILE12 (Peterson et al., 2022) provide rich tasks and strong baselines, but their evaluations are tied primarily to a single institutional setting. This leaves open whether model gains on naturalistic iEEG tasks persist across independently collected datasets from different institutions.

We therefore introduce iMINDBench, an iEEG Multi-Institution Neural Decoding Benchmark that aligns fifteen naturalistic decoding tasks and electrode metadata across three independently collected institutional datasets while preserving their anatomical and acquisition heterogeneity. Standardized spectral and waveform preprocessing tracks and fixed target evaluations provide a common basis for measuring progress toward general-purpose iEEG decoding through cross-institution model comparisons and matched-task scaling.

The results reveal why these controls matter. First, standardized tracks clarify model comparisons: systematically varying common preprocessing parameters within both the spectral and waveform

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.

tracks shows that baseline ROC-AUC can shift by more than 0.10 in some settings. Under these conditions, pretrained systems generally outperform non-pretrained baselines within their respective tracks, yet at the complete-system level a strong engineered spectral baseline remains competitive across institutions, especially on the two datasets with less downstream training data. Second, our scaling study reveals limited benefit from broader cross-institution supervision: additional same-dataset subjects yield modest gains, while adding data from other institutions provides little aggregate benefit despite up to a 25-fold expansion of the supervised dataset. Together, these findings identify preprocessing sensitivity and limited use of cross-institution supervision as concrete challenges for general-purpose iEEG decoding.

Contributions.

- **Benchmark artifact.** We introduce a multi-institution iEEG decoding benchmark spanning three independently collected movie-watching datasets with fifteen aligned decoding tasks.
- **Evaluation contract.** iMINDBench defines fixed splits, decodability subsets, standardized spectral and waveform preprocessing tracks, and supervised scaling domains.
- **Benchmark findings.** Strong spectral baselines remain competitive with pretrained systems across institutions, providing a reference point for learned representations. In our scaling study, additional same-dataset subjects yield modest gains, whereas adding data from other institutions provides little aggregate benefit despite substantially more supervision.

2 Related Work

Datasets and benchmarks. Intracranial electroencephalography is attractive for naturalistic neural decoding because it records human neural activity at high temporal resolution during behavior, but clinical recordings are sparse, nonuniform, and institution-specific (Parvizi & Kastner, 2018). Existing iEEG benchmarks cover important parts of this landscape. The Mayo/FNUSA graphoelement dataset (Nejedly et al., 2020) and Omni-iEEG (Duan et al., 2026) show that multi-institution clinical iEEG can be harmonized, but focus on seizure, pathology, high-frequency oscillation, or related epileptology targets. Naturalistic resources such as Brain Treebank (Wang et al., 2024a), Neuroprobe (Zahorodnii et al., 2026), BYD (Keles et al., 2024), Pippi (Berezutskaya et al., 2022), and AJILE12 (Peterson et al., 2022) provide movie, language, audiovisual, motor, or activity labels, but do not jointly control task, institution, and preprocessing heterogeneity in one evaluation protocol.

Large-scale, multi-dataset benchmarks exist for scalp EEG (Jayaram & Barachant, 2018; Kuruppu et al., 2026; Yang et al., 2026), but their protocols do not transfer directly to iEEG. Scalp EEG measures volume-conducted signals and generally emphasizes lower-frequency activity, whereas iEEG records closer to neural sources and provides access to high-frequency activity (roughly 70–250 Hz) important for cognitive decoding (Parvizi & Kastner, 2018; Lachaux et al., 2012). Clinically determined electrode placement also produces sparse, participant-specific anatomical coverage rather than a shared sensor montage, while institution-specific acquisition and referencing practices compound this heterogeneity (Parvizi & Kastner, 2018; Duan et al., 2026). These properties motivate an iEEG-specific benchmark for evaluating models across heterogeneous recordings.

Neural foundation models and preprocessing. Recent electrophysiology foundation models make shared evaluation increasingly necessary. iEEG models such as BrainBERT (Wang et al., 2023), Brant (Zhang et al., 2023), PopT (Chau et al., 2025), BaRISTA (Oganesian et al., 2025), MVP-Former (Carzaniga et al., 2026), and DIVER-1 (Han et al., 2026), together with EEG foundation models such as LaBraM (Jiang et al., 2024) and EEGPT (Wang et al., 2024b), address scale, subject variability, channel mismatch, spatial layout, and low signal-to-noise ratio. However, their evaluation conditions differ across waveform inputs, spectrograms or superlets, learned tokens, frozen embeddings, spatial metadata, pretraining corpora, and split units. A recent EEG foundation-model review argues that evaluations remain heterogeneous and limited (Kuruppu et al., 2026), and a broad EEG-FM benchmark finds that fine-tuned foundation models can perform well in data-rich settings but often do not clearly outperform compact neural or classical decoders in data-scarce regimes (Yang et al., 2026). This benchmark targets the corresponding iEEG problem: model comparisons should be made within explicit task, institution, split, and preprocessing specifications, with matched baseline reruns whenever the input representation changes.

3 BENCHMARK DESIGN

The benchmark is designed to evaluate whether decoding improvements survive changes in institution, preprocessing track, and behavioral target under shared evaluation settings. To support this, iMIND-Bench standardizes preprocessing and evaluation across multiple naturalistic iEEG datasets while exposing performance differences across tasks and institutions. It also supports controlled scaling experiments that test whether additional supervised data improves decoding across sessions, subjects, and institutions. Figure 1 summarizes the benchmark workflow, including datasets, preprocessing tracks, splits, and decoding domains.

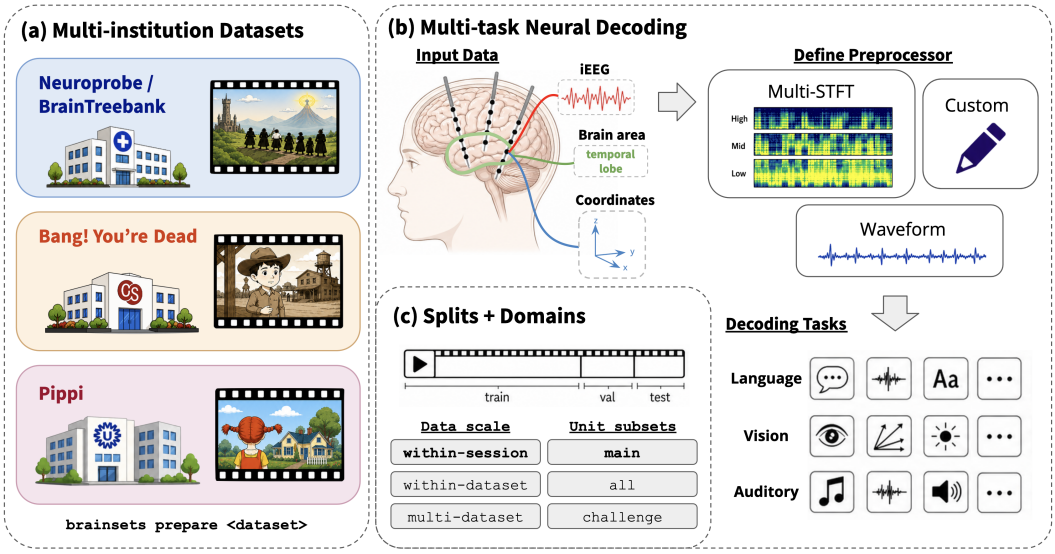


Figure 1: **Multi-institution, multi-task neural decoding benchmark.** The benchmark aligns naturalistic decoding tasks and electrode metadata across three institutions and defines preprocessing tracks, splits, and decoding domains. (a) Three naturalistic movie-watching intracranial electroencephalography (iEEG) datasets are included: Neuroprobe, Bang! You're Dead (BYD), and Pippi. (b) iEEG neural recordings are routed through explicit preprocessing tracks, which may include multi-resolution short-time Fourier transform (Multi-STFT), waveform, and custom preprocessors to be used as inputs to models that decode language, vision, and auditory task families. (c) The benchmark defines challenging temporal splits, data-scale regimes, and unit decodability subsets.

3.1 DATASETS AND TASKS

Datasets. The current release combines Neuroprobe/Brain Treebank (Zahorodnii et al., 2026; Wang et al., 2024a), the Bang! You're Dead (BYD) movie-watching dataset (Keles et al., 2024), and Pippi (Berezutskaya et al., 2022). Compared with the original Neuroprobe benchmark, we restrict evaluation of the Neuroprobe/Brain Treebank component to five held-out subject/session units so that remaining Brain Treebank recordings can be used for model pretraining without overlapping benchmark targets. All three datasets contain neural recordings acquired during movie viewing, but differ in institution, stimulus content, recording duration, electrode coverage, and subject/session structure. Neuroprobe/Brain Treebank provides longer recordings and larger within-session training sets, whereas BYD and Pippi provide smaller within-session training sets. Appendix Figure 4 summarizes their scale, electrode coverage, and task support.

Harmonization. iMINDBench aligns decoding targets and electrode metadata to support comparisons across institutions. BYD and Pippi use the Neuroprobe/Brain Treebank stimulus-feature extraction pipeline, with dataset-specific annotation and timing alignment. Available electrode coordinates and anatomical labels are harmonized to shared conventions. Harmonizing electrode metadata does not eliminate differences in electrode contact size and geometry, anatomical coverage, or acquisition referencing across subjects and institutions. These differences remain part of the benchmark's

evaluation setting. Adding a dataset can require manual stimulus annotation and electrode-metadata alignment. Appendix A describes the alignment procedures and documents the source datasets' acquisition metadata.

Tasks. The benchmark includes fifteen decoding tasks following the original Neuroprobe task definitions (Zahorodnii et al., 2026), spanning language, auditory, and visual domains. These include speech presence, sentence onset, GPT-2 surprisal, pitch, volume, optical flow, brightness, and face-count decoding. Both binary and multiclass label modes are supported; this paper reports binary classification results. Each example pairs a 1-second neural window with a stimulus label. Continuous annotations are binarized by contrasting low- and high-value ranges within each session, while categorical annotations are mapped to binary labels. Complete task definitions and dataset-specific label-construction details are provided in Appendix B.

3.2 Splits and Domains

Splits. We follow Neuroprobe's two-fold within-session splitting procedure for each dataset–subject/session–task combination, termed an evaluation **unit**. Applying this procedure across datasets maintains compatibility with Neuroprobe, including reuse of compatible runs for Neuroprobe submission. Within each fold, training, validation, and test examples come from largely contiguous portions of the movie. Preprocessing statistics are estimated from the training split and applied unchanged to validation and test data. Each unit's final score is averaged across the two folds.

Scaling domains. Scaling domains vary the source of supervised fine-tuning data while preserving the same target units and evaluation splits. **Within-session** uses only target-session data, representing adaptation to a new subject/session using its calibration labels. **Within-dataset** adds supervised examples from other subject/session units in the same dataset. **Multi-dataset** further adds supervised examples from other institutions. These domains test whether broader supervision improves decoding beyond target-session calibration.

Unit decodability subsets. Many evaluation units remain near chance under current baselines. This may reflect weak task-relevant signal at the recorded electrode locations, poor recording quality, or limitations of current models in extracting task-relevant information. We therefore report Main, Challenge, and All unit decodability subsets to distinguish improvements on units that are decodable by baseline within-session models from progress on those that are presently near chance.

We operationalize near-chance performance using two screening decoders: an STFT decoder and HTNet at 500 Hz. The STFT decoder uses logistic regression with the single-STFT configuration that achieves the highest mean validation ROC-AUC for each unit. A supported unit u is assigned to **Main** when

$$\max\left\{\overline{\mathrm{AUC}}_{\mathrm{val}}^{\mathrm{STFT}}(u), \overline{\mathrm{AUC}}_{\mathrm{val}}^{\mathrm{HTNet\text{–}500\,Hz}}(u)\right\} > 0.60,$$

where each score is averaged across the two validation folds. **Challenge** contains the remaining supported units, classified as near chance because both screening scores are at or below 0.60. **All** contains both subsets. These labels describe decodability under the screening models; they do not establish that Challenge units lack task-relevant neural signal. The resulting Main/Challenge performance separation, task-level Main-unit model breakouts, and Main subject/session coverage are summarized in Appendix Figures 5, 6, and 7.

3.3 Preprocessing Tracks

Track definitions. Decoding performance depends on both the model and how neural recordings are preprocessed, making it necessary to account for both when interpreting improvements. Each preprocessing track specifies how neural recordings are transformed into model inputs, helping control for preprocessing variation. We provide two complementary tracks. The Multi-STFT track incorporates established signal-processing operations and spectral features (Buzsáki et al., 2012) that enable strong performance with simple baseline models. The Waveform track instead provides inputs close to the raw recorded waveforms, preserving phase and temporal structure that the Multi-STFT representation does not explicitly represent. It tests whether models can learn representations that match or exceed the benefits of engineered spectral features. Comparisons under matched

preprocessing assess model improvements using the same inputs, while comparisons across tracks assess complete decoding systems, including their preprocessing. The benchmark also supports custom preprocessing routes, allowing compatible reference baselines to be rerun on new inputs for matched comparisons.

Multi-STFT track. The Multi-STFT track is the benchmark's standardized spectral preprocessing route. The spectral track first applies 60 Hz + harmonics notch filtering for line noise removal and Laplacian rereferencing for common electrode noise removal, following prior iEEG pipelines (Wang et al., 2023; Li et al., 2018). Because neural activity spans multiple temporal and frequency scales, a single fixed STFT resolution can unevenly represent different frequency bands and tasks. The Multi-STFT representation instead uses separate STFT resolutions for low-, mid-, and high-frequency ranges before combining them into a spectrogram-like representation compatible with common neural decoding models (Appendix Figure 10). The Multi-STFT sections use duration-defined windows and an approximately 62-ms hop consistently across dataset sampling rates; Appendix Figure 10 shows the frequency ranges, and Table 7 gives the exact sample settings and retained upper bins. The Multi-STFT track selects frequency bins within 2–250 Hz, excluding the lowest-frequency components while retaining bands including high gamma (Lachaux et al., 2012). For each fold, the Multi-STFT track estimates normalization statistics exclusively from training samples, separately for each channel and frequency bin, and applies those fixed statistics to validation and test samples. This preserves cross-temporal information while standardizing the signals for machine-learning models. We compare normalization strategies and their downstream performances on Multi-STFT baselines in Figure 2a.

Waveform track. The Waveform track uses a 0.5 Hz high-pass filter, notch filtering for line noise, and Laplacian rereferencing, while permitting model-specific sampling rates and input scaling and normalization conventions. Supervised baselines use a standardized 500 Hz recipe with 60 Hz and harmonic notch filtering and robust global normalization fitted exclusively on the training split. This normalization uses a single estimated median center and a scaled median absolute deviation across training channels and timepoints, applied unchanged to validation and test samples, preserving relative amplitude differences across windows and channels. We use 500 Hz for the supervised waveform baselines because matched comparisons did not show consistent gains at higher sampling rates. Model-specific sampling, filtering, and normalization variations are documented in Table 8. Within-track comparisons involving model-specific variants evaluate complete decoding systems; they do not isolate architecture or pretraining effects. We visualize how preprocessing changes affect baseline performance on our strongest supervised waveform model, HTNet, in Figure 2a.

Custom preprocessing routes. Users can evaluate custom preprocessing routes by supplying preprocessing code and configuration files. These routes must be documented explicitly and, where compatible inputs or embeddings are available, evaluated with a simple matched baseline. Leaderboard entries report preprocessing details to make these comparisons transparent (Appendix Figure 11). Appendix D illustrates this approach for BrainBERT, comparing its pretrained representations with simple decoders using the corresponding single-STFT inputs.

3.4 Models and Reporting

Model families. We evaluate generic decoder baselines alongside pretrained and specialized iEEG models. Logistic regression, multilayer perceptron (MLP), and convolutional neural network (CNN) baselines are evaluated on both Multi-STFT and Waveform inputs. The Multi-STFT suite additionally includes pretrained PopT-v2 (Chau et al., 2025; Patel et al., 2025). The Waveform suite includes HTNet (Peterson et al., 2021), trained from scratch, and pretrained BaRISTA (Oganesian et al., 2025) and DIVER-1 (Han et al., 2026). We also examined MVPFormer and Brant, whose inclusion requires further validation of adaptations from their released temporal input representations to one-second decoding windows (Appendix I.1). Appendix I provides model and training details.

Reporting. Results are reported with their preprocessing track, model family, and dataset-level performance. For pretrained models, we also report the datasets used for pretraining to contextualize their training exposure. For each dataset and model, we report the mean ROC-AUC across the available evaluation units. The Overall score is the unweighted mean of the three dataset-level

scores. The benchmark website and leaderboard break down model comparisons by dataset, task, preprocessing track, scaling domain, and unit decodability subset, helping identify where model advantages are consistent and where they depend on the evaluation setting.

4 BENCHMARK RESULTS

4.1 DECODING PERFORMANCE

Track	Model	Pretraining	Overall	Neuroprobe	BYD	Pippi
Multi-STFT	Logistic	not pretrained	0.628	0.697	0.624	0.564
	MLP	not pretrained	<u>0.631</u>	0.702	<u>0.625</u>	<u>0.567</u>
	CNN	not pretrained	0.620	<u>0.710</u>	0.611	0.539
	PopT-v2	BrainTreeBank	**0.645**	**0.723**	**0.632**	**0.579**
Waveform	Logistic	not pretrained	0.554	0.617	0.519	0.525
	MLP	not pretrained	0.573	0.653	0.532	0.534
	CNN	not pretrained	0.576	0.666	0.537	0.526
	HTNet	not pretrained	0.614	0.693	0.597	0.552
	BaRISTA	BrainTreeBank	**0.636**	<u>0.716</u>	**0.632**	<u>0.558</u>
	DIVER-1	private+ AJILE12	<u>0.634</u>	**0.728**	<u>0.605</u>	**0.569**

Table 1: **Main benchmark results.** Within-session ROC-AUC performance on Main units, grouped by preprocessing track. Cells are shaded with a normalized color scheme within column and track. Best per column and track are bolded, and second-best values are underlined.

Table 1 compares within-session decoding performance on Main units across the three datasets. Within their respective tracks, pretrained systems outperform non-pretrained baselines: PopT-v2 achieves the highest Overall ROC-AUC of 0.645 versus 0.631 for the strongest non-pretrained Multi-STFT baseline, while BaRISTA and DIVER-1 achieve 0.636 and 0.634 versus 0.614 for HTNet in the Waveform track. Across tracks, however, the non-pretrained Multi-STFT MLP is within 0.005 ROC-AUC of both pretrained waveform systems overall and remains particularly competitive on BYD and Pippi, the two datasets with smaller downstream training sets. The evaluated pretrained waveform systems therefore do not maintain a clear advantage over this engineered spectral baseline across institutions. Because the systems use different input representations, this cross-track result compares complete decoding systems and does not isolate architecture or pretraining effects.

The benefit of more complex spectral decoders varies across datasets. On Neuroprobe, CNN outperforms both logistic regression and MLP, whereas on BYD and Pippi, logistic regression and MLP outperform CNN. This pattern suggests that engineered spectral features support effective decoding with relatively simple models when labeled data are limited, highlighting the need for learned feature extractors that generalize reliably under limited supervision.

Waveform inputs show a different pattern. HTNet consistently outperforms generic logistic, MLP, and CNN baselines across all three datasets, indicating that the choice of architecture matters for extracting useful information from waveform inputs. Pretrained BaRISTA and DIVER-1 further outperform these within-session-trained baselines, although their relative strengths vary across datasets: DIVER-1 performs better on Neuroprobe and Pippi, while BaRISTA performs better on BYD.

Together, these results show that model advantages depend on the input representation and institutional dataset, motivating evaluation across both dimensions. Additional scorecards for All and Challenge units are provided in Appendix C.

4.2 PREPROCESSING

Figure 2 evaluates how preprocessing choices affect benchmark conclusions under otherwise matched evaluation settings. Baseline performance varies across preprocessing routes, indicating that neural decoding comparisons can depend strongly on input representation and normalization choices.

In Figure 2a, we compare two normalization strategies that encode different assumptions about signal scale. *Sample Norm*, following approaches used in wav2vec 2.0 and BrainBERT (Baevski

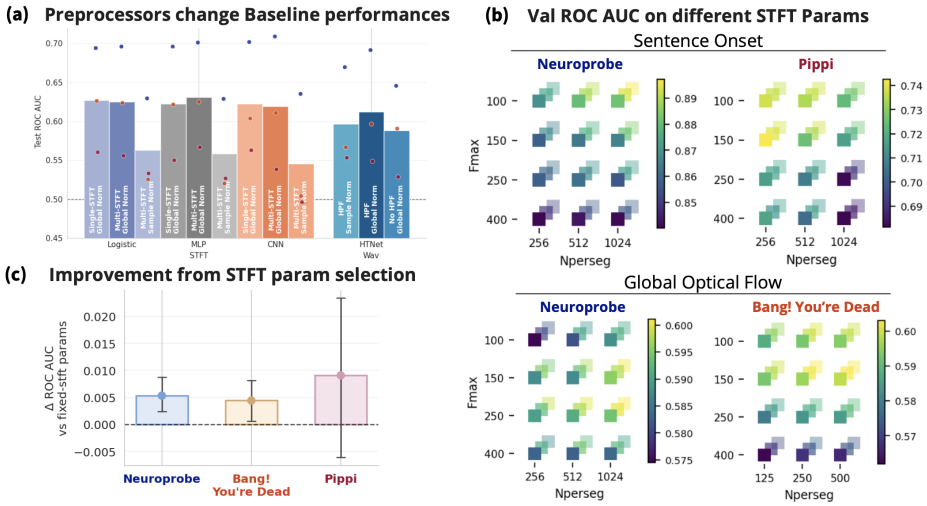


Figure 2: **Preprocessing choices have large effects on baseline performance.** (a) Baseline model performance under different STFT and Waveform preprocessing choices. Bars show the unweighted mean of dataset-level averages; dots show individual dataset averages. Standardized spectral and waveform tracks use the darkest shade in each model family. (b) Validation ROC-AUC for STFT parameter sweeps. We sweep max frequency (y-axis), number of timepoints per segment (x-axis) and percent overlap (depth) at 50%, 75%, and 87.5%. Different datasets (columns) may have different optimal single-STFT params for the same task (rows). (c) Mean paired test ROC-AUC change for logistic regression using validation-selected versus fixed single-STFT parameters (150 Hz maximum frequency, 250 ms segments, 75% overlap). Error bars indicate 95% bootstrap confidence intervals. All panels report results on the Main unit decodability subset.

et al., 2020; Wang et al., 2023), standardizes each input window independently. This may provide invariance to nuisance variation in recording gain and overall amplitude, but it can also remove between-window amplitude differences that are informative for decoding. *Global Norm* instead uses statistics estimated from the training set, preserving relative amplitude variation across examples while applying the same transformation to validation and test samples. Global Norm yields consistently stronger baseline performance, improving ROC-AUC by more than 0.10 in some evaluated settings. This result suggests that information removed by sample-wise normalization is often task-relevant in our evaluation settings. We therefore select Global Norm for the standardized baseline recipes in our Multi-STFT and Waveform preprocessing tracks; Section 3.3 details its implementation for each track.

We next test the Waveform track's sensitivity to high-pass filtering while holding the remaining HTNet preprocessing fixed. High-pass filtering suppresses slow baseline drift and other low-frequency variation that can dominate waveform amplitude, although it may also attenuate slowly varying task-related signals. Omitting the 0.5-Hz high-pass filter reduces mean ROC-AUC from 0.612 to 0.589, with lower performance in all three datasets (Figure 2a). This consistent reduction suggests that, under our evaluation protocol, the benefit of suppressing very-low-frequency variation outweighs any useful information attenuated by the filter. We therefore retain the 0.5-Hz high-pass filter in the standardized Waveform baseline recipe.

Finally, we test the spectral track's sensitivity to replacing Multi-STFT with a single STFT resolution. Aggregate performance is comparable (Figure 2a), but validation-selected Single-STFT parameters vary across datasets and tasks (Figure 2b,c). We therefore retain Multi-STFT as the flagship spectral track: its band-specific window lengths favor frequency resolution at lower frequencies and temporal resolution at higher frequencies without selecting a task- or dataset-specific resolution.

These results show that preprocessing choices can substantially affect decoding performance and support our standardized tracks through both empirical comparisons and signal-processing considerations. Comparing models under matched preprocessing helps distinguish model improvements from gains due to preprocessing changes.

4.3 Data Scaling

Transfer question. Data scaling tests whether supervised recordings from increasingly heterogeneous sources improve decoding for fixed target units. If broader supervision is useful, it should improve decoding for a target unit when target-session labels are scarce. Figure 3 separates that question into three data sources. Within-session scaling measures the value of additional examples from the same target unit. Within-dataset scaling asks whether other Main units from the same dataset help once institution, acquisition practice, stimulus family, and label construction are mostly shared. Multi-dataset scaling is the harder test: whether data from other institutions and stimulus contexts improve decoding for the same target units. We evaluate within-session scaling across spectral models and use PopT-v2, which supports joint supervised training across subjects, for within-dataset and multi-dataset scaling.

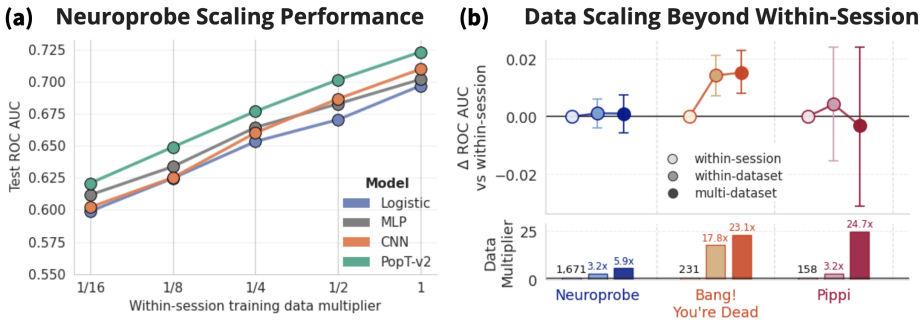


Figure 3: **Data scaling highlights a boundary in current model generalization.** (a) Neuroprobe within-session decoding performance (y-axis) as the available training data increase (x-axis) across models evaluated on the Multi-STFT track (legend). (b) For each dataset (columns and colors), the top subplot shows the paired change in test ROC-AUC relative to within-session training (y-axis) when PopT-v2 is provided supervised data from within the same dataset (medium-shaded circles) or from multiple datasets (dark-shaded circles). The bottom subplot shows the corresponding training-data multipliers relative to within-session training (colored labels); the black label above the leftmost bar gives the mean number of within-session training samples per task. Error bars indicate 95% confidence intervals for paired changes in ROC-AUC. Task-level decompositions are provided in Appendix Figures 8 and 9. All panels report results on the Main unit decodability subset.

Scaling pattern. The scaling analysis exposes an important asymmetry between target-session and broader supervised data. Across evaluated models, within-session scaling produces the largest and most consistent gains, with mean test ROC-AUC increasing by approximately 0.025 for each doubling of the training data (Figure 3a). Using approximately 3–18 times as many training samples as within-session training, within-dataset scaling improves mean ROC-AUC by 0.001 on Neuroprobe, 0.014 on Bang! You're Dead, and 0.004 on Pippi (Figure 3b). Multi-dataset training uses approximately 6–25 times as many samples as within-session training, yet underperforms within-dataset training on Neuroprobe and Pippi. On Bang! You're Dead, multi-dataset training improves mean ROC-AUC over within-dataset training by only 0.001 despite using approximately 30% more training samples.

Within-dataset training increases mean test ROC-AUC over within-session training in 30 of 44 task–dataset pairs; adding other datasets yields further increases in 19 of 44 pairs (Appendix Figure 9). The benefit of broader supervision therefore depends on the decoding task. More effective cross-subject and cross-institution learning remains an important direction for improving downstream neural decoding.

5 Release and Extensibility

We will release evaluation code, preprocessing configurations, and baseline implementations in the iMINDBench repository, with dataset preparation pipelines and split metadata through TorchBrain/Brainsets in the Neuro-Galaxy ecosystem (Azabou et al., 2023). A public website and leaderboard will support decomposed comparisons and validated submissions (Appendix J). Additional

model-native systems can use custom preprocessing routes when coverage is complete and matched baseline reruns are available. Future releases may add datasets, tasks, preprocessors, models, or hidden-test evaluation while preserving archived benchmark versions.

6 Discussion

The practical value of learned representations depends on both preprocessing and recording conditions. Strong Multi-STFT baselines remain competitive with pretrained waveform systems, particularly on datasets with smaller downstream training sets, suggesting room to improve how models learn useful representations from heterogeneous iEEG recordings. Our preprocessing analyses further show that changes in normalization and filtering can substantially reduce baseline performance. We therefore recommend using the standardized Multi-STFT and Waveform tracks to support comparisons with the evaluated models and tuned baselines. When a model requires different preprocessing, we strongly recommend rerunning compatible reference baselines through the same input pipeline to help distinguish model improvements from preprocessing effects. Matched preprocessing does not eliminate differences in electrode coverage and acquisition across recordings. Evaluating models across institutional datasets therefore helps establish whether their advantages persist across heterogeneous recording conditions.

Our scaling results show that the source of additional supervision matters for target-session decoding. Within-session scaling produces the largest and most consistent gains, while broader supervision yields smaller, task-dependent improvements. Residual differences in electrode configurations and acquisition conditions may limit the benefits of pooling data across institutions, even for shared decoding targets. Our experiments do not isolate the contributions of these individual factors, but they motivate representations that are robust or adaptable to heterogeneous recordings.

7 Limitations and Future Work

Evaluation scope. The current evaluation covers three naturalistic movie-watching datasets and assumes access to target-session calibration labels. Extending the benchmark to other stimulus paradigms and evaluation without target-session labels would test broader forms of generalization.

Scaling coverage. The within-dataset and multi-dataset scaling experiments currently evaluate a single pretrained model, whose implementation supports joint supervised fine-tuning across subjects. The other pretrained systems are evaluated using linear decoding heads over output embeddings, without a corresponding mechanism for joint cross-subject fine-tuning in our implementation. Extending these evaluations to additional models would test how broadly the observed scaling pattern holds.

Model tuning. Our model comparisons also reflect the training recipes and tuning budgets used in this study; further optimization of pretraining and fine-tuning could change their relative performance. Appendix K explores this issue through a human-guided AutoResearch-style feasibility study, in which tuning a pretraining recipe improves downstream decoding. Such workflows could help establish stronger model baselines, provided that tuning budgets and validation-based selection procedures are documented consistently across models.

8 Conclusion

iMINDBench provides a shared experimental setting for measuring progress toward general-purpose iEEG decoding across naturalistic tasks and institutional datasets. Across standardized preprocessing tracks, pretrained systems generally lead within tracks, yet a strong spectral baseline remains competitive across institutions. Our scaling study finds modest gains from additional same-dataset subjects and little aggregate benefit from other institutions despite substantially more supervision. Together, these results identify consistent cross-institution gains and effective use of heterogeneous supervision as measurable targets for future general-purpose iEEG decoders.

DISCLOSURE OF AI USE

We used generative AI tools to assist with writing and polishing the manuscript, retrieve information and identify relevant literature, and draft portions of the paper. We also used a human-guided AutoResearch-style agentic workflow to tune pretraining hyperparameters, as described in Appendix K. The authors take responsibility for the final content of this work, including its claims, citations, and AI-assisted text and artifacts.

REFERENCES

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460, 2020.

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*, 9:91, 2022. doi: 10.1038/s41597-022-01173-0.

György Buzsáki, Costas A Anastassiou, and Christof Koch. The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6):407–420, 2012.

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

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

Chenda Duan, Yipeng Zhang, Sotaro Kanai, Yuanyi Ding, Atsuro Daida, Pengyue Yu, Tiancheng Zheng, Naoto Kuroda, Shaun A. Hussain, Eishi Asano, Hiroki Nariai, and Vwani Roychowdhury. Omni-iEEG: A large-scale, comprehensive iEEG dataset and benchmark for epilepsy research. In *International Conference on Learning Representations (ICLR)*, 2026.

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, 2026. arXiv preprint.

Vinay Jayaram and Alexandre Barachant. MOABB: trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018. doi: 10.1088/1741-2552/aadea0.

Weibang Jiang, Liming Zhao, and Bao-liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024.

Andrej Karpathy. autoresearch: AI agents running research on single-GPU nanochat training automatically, 2026. GitHub repository.

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*, 11:214, 2024. doi: 10.1038/s41597-024-03029-1.

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.

Jean-Philippe Lachaux, Nikolai Axmacher, Florian Mormann, Eric Halgren, and Nathan E. Crone. High-frequency neural activity and human cognition: Past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, 98(3):279–301, 2012. doi: 10.1016/j.pneurobio.2012.06.008.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. doi: 10.1088/1741-2552/aace8c.

Guangye Li, Shize Jiang, Sivylla E Paraskevopoulou, Meng Wang, Yang Xu, Zehan Wu, Liang Chen, Dingguo Zhang, and Gerwin Schalk. Optimal referencing for stereo-electroencephalographic (seeg) recordings. *NeuroImage*, 183:327–335, 2018.

Petr Nejedly, Vaclav Kremen, Vladimir Sladky, Jan Cimbalnik, Petr Klimes, Filip Plesinger, Filip Mivalt, Vojtech Travnicek, Ivo Viscor, Martin Pail, Josef Halamek, Benjamin H. Brinkmann, Milan Brazdil, Pavel Jurak, and Gregory Worrell. Multicenter intracranial EEG dataset for classification of graphoelements and artifactual signals. *Scientific Data*, 7:179, 2020. doi: 10.1038/s41597-020-0532-5.

Lucine L. Oganesian, Saba Hashemi, and Maryam M. Shanechi. BaRISTA: Brain scale informed spatiotemporal representation of human intracranial neural activity. In *The Thirty-ninth Conference on Neural Information Processing Systems*, 2025.

Josef Parvizi and Sabine Kastner. Promises and limitations of human intracranial electroencephalography. *Nature Neuroscience*, 21(4):474–483, 2018. doi: 10.1038/s41593-018-0108-2.

Eshani Patel, Yisong Yue, and Geeling Chau. Learning time-scale invariant population-level neural representations. In *NeurIPS 2025 Workshop on Foundation Models for the Brain and Body*, 2025. doi: 10.48550/arXiv.2511.13022.

Steven M. Peterson, Zoe Steine-Hanson, Nathan Davis, Rajesh P. N. Rao, and Bingni W. Brunton. Generalized neural decoders for transfer learning across participants and recording modalities. *Journal of Neural Engineering*, 18(2):026014, mar 2021. doi: 10.1088/1741-2552/abda0b.

Steven M. Peterson, Satpreet H. Singh, Benjamin Dichter, Michael Scheid, Rajesh P. N. Rao, and Bingni W. Brunton. AJILE12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific Data*, 9: 184, 2022. doi: 10.1038/s41597-022-01280-y.

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*, 2023.

Christopher Wang, Adam Uri Yaari, Aaditya K. Singh, Vighnesh Subramaniam, Dana Rosenfarb, Jan DeWitt, Pranav Misra, Joseph R. Madsen, Scellig Stone, Gabriel Kreiman, Boris Katz, Ignacio Cases, and Andrei Barbu. Brain treebank: Large-scale intracranial recordings from naturalistic language stimuli. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-3060.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Liuyin Yang, Qiang Sun, Ang Li, and Marc M. Van Hulle. Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=5Xwm8e6vbh`.

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. arXiv preprint.

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, pp. 26304–26321. Curran Associates, Inc., 2023.

APPENDIX

A DATASET HARMONIZATION AND ACQUISITION METADATA

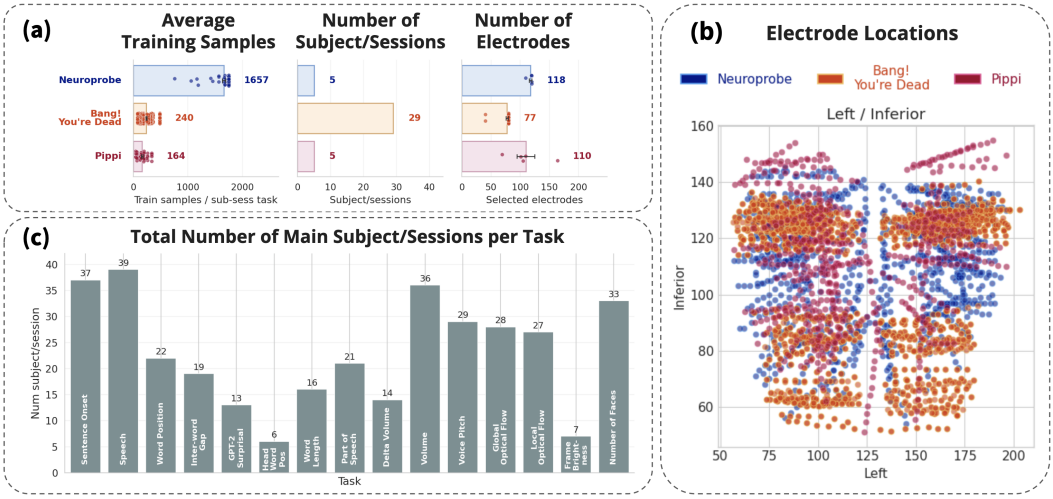


Figure 4: **Dataset coverage is heterogeneous across institutions and tasks.** (a) Average number of training samples, Main subject/session–task units, and electrodes per subject. (b) Electrode locations are shown in left/inferior coordinate space for the same datasets. (c) Task bars report the number of Main subject/session–task units for each aligned decoding task.

Controlled cross-institution evaluation requires alignment along two benchmark-specific dimensions: decoding labels derived from independently collected movie stimuli and iEEG electrode metadata reported under different localization conventions. Together, these steps support common decoding targets and spatial metadata without treating the source datasets as uniform.

Task and temporal alignment. For the BYD and Pippi components, task and label alignment starts from the Neuroprobe/Brain Treebank stimulus-feature extraction pipeline, applied to each dataset's video stimulus where source annotations support it. Task labels, scalar thresholds, and lexical/nonverbal row construction are defined in Appendix B. Sentence-level linguistic attributes are labeled using the Brain Treebank annotation protocol, and sentence and word onset times are initialized with Whisper automatic speech recognition (ASR) before manual alignment using 4× slowed video playback. We manually correct `face_num` labels because the automatic face-count detector is noisy on these stimuli.

Electrode metadata alignment. Electrode layouts and cortical coverage vary by patient, and localization metadata use different coordinate systems and anatomical labels. Electrode locations are transformed into Brain Treebank's left-posterior-inferior (LPI) coordinate space by first aligning coordinate dimensions to a sensible brain orientation and then matching the electrode coordinate ranges provided in Brain Treebank. For brain-area metadata, we harmonize labels to the Destrieux parcellation used by Brain Treebank: Pippi electrodes are overlaid onto Destrieux areas with a FreeSurfer alignment pipeline using the publicly available MRI scans, and BYD electrode labels are manually converted to the same Destrieux area set.

Remaining acquisition differences. Table 2 summarizes the remaining acquisition and stimulus differences relevant to multi-dataset model transfer, using metadata reported in the source dataset papers.

Table 2: **Institution-specific stimulus and acquisition metadata.** Dataset components differ in stimulus duration and language, electrode type and coverage, sampling, reference conventions, and coordinate reporting. BYD and Pippi entries focus on macroelectrode and clinical sEEG recordings, respectively.

Property	**Neuroprobe / Brain Treebank** (Zahorodnii et al., 2026; Wang et al., 2024a)	**BYD / Cedars-Sinai** (Keles et al., 2024)	**Pippi / Utrecht** (Berezutskaya et al., 2022)
Stimulus	English feature-length Hollywood films. Brain Treebank reports 26 films and 43.5 total hours; subjects watched 2.6 movies on average.	8-minute black-and-white Hitchcock excerpt with English audio-visual content.	6.5-minute Dutch-dubbed short film from *Pippi on the Run*, with interleaved speech and music blocks.
Electrodes	sEEG depth probes with 6–16 contacts per probe. Contacts: 0.8 mm diameter, 2 mm long.	Macroelectrodes on hybrid Behnke-Fried depth electrodes.	Clinical sEEG recordings.
Sampling	2048 Hz.	1000 Hz.	2048 Hz.
Source filtering	60 Hz notch filtering and harmonics.	60 Hz notch and 0.1 Hz high-pass filtering.	Source analysis applies 50 Hz notch filtering.
Referencing	Neuroprobe evaluates Laplacian-rereferenced inputs for several benchmark models.	Source processing uses common-average rereferencing.	External mastoid reference during acquisition; common-average rereferencing in source analysis.
Localization	Electrode positions localized to an average cortical atlas.	Native-space and MNI152 coordinates provided.	Native-space electrode locations; MNI projection used for visualization.

sEEG: stereoelectroencephalography; MNI: Montreal Neurological Institute.

B Decoding Tasks

Task definitions follow the Neuroprobe task suite (Zahorodnii et al., 2026) and Population Transformer benchmark protocol (Chau et al., 2025) where source annotations support them. For scalar binary tasks, Neuroprobe uses bottom-versus-top quartiles, while BYD and Pippi use bottom-versus-top terciles to improve class support in the shorter stimuli. Lexical and surprisal tasks are restricted to word rows.

For BYD and Pippi visual and auditory tasks, feature tables include lexical word rows and sampled nonverbal rows. Nonverbal rows are sampled in 500 ms windows stepped every 500 ms from gaps longer than 2.0 s between the preceding word's offset and the next word's onset. They inherit the preceding word row's sentence identifier and provide the negative class for speech and sentence-onset tasks.

Table 3: **Aligned visual, auditory, and language decoding tasks.** All tasks use binary classification on 1-second neural windows aligned to word onset or sampled nonverbal-interval onset. Classes are rebalanced before training.

#	Feature	Description	Benchmark task
1	`onset` *(language)*	Whether a new sentence starts.	Sentence onset vs. nonverbal interval.
2	`speech` *(language)*	Whether speech is present.	Word interval vs. nonverbal interval.
3	`word_index` *(language)*	Word index within its sentence.	First word (0) vs. second word (1); later words excluded.
4	`word_gap` *(language)*	Time from the preceding word's offset to the current word's onset within the same sentence (ms).	Low vs. high value.
5	`gpt2_surprisal` *(language)*	Negative log probability of the word under GPT-2, conditioned on the preceding 20 s of language context.	Low vs. high value.
6	`word_head_pos` *(language)*	Relative position of the word's dependency-tree head.	Head to the left/root/self vs. head to the right.
7	`word_length` *(language)*	Word duration in milliseconds.	Low vs. high value.
8	`word_part_speech` *(language)*	Universal part-of-speech (UPOS) tag of the word.	Noun vs. verb; other tags excluded.
9	`delta_volume` *(auditory)*	Difference in average RMS audio power between the 500 ms windows before and after target onset.	Low vs. high value.
10	`volume` *(auditory)*	Average root mean square (RMS) audio power.	Low vs. high value.
11	`pitch` *(auditory)*	Average audio pitch.	Low vs. high value.
12	`global_flow` *(visual)*	Camera-motion proxy computed as the maximal average dense optical-flow vector magnitude.	Low vs. high value.
13	`local_flow` *(visual)*	Large-displacement proxy computed as the maximal individual optical-flow vector magnitude.	Low vs. high value.
14	`frame_brightness` *(visual)*	Mean HSV value (V) over pixels.	Low vs. high value.
15	`face_num` *(visual)*	Maximum number of faces per frame during the target interval.	No face (0) vs. at least one face ($\geq$ 1).

C Additional Unit Decodability Subset Scorecards

Table 4 reports the same model and track rows as the Main scorecard on the All subset. Table 5 reports the same rows on the Challenge subset.

Track	Model	Pretraining	Overall	Neuroprobe	BYD	Pippi
Multi-STFT						
	Logistic	not pretrained	0.587	0.641	0.576	0.545
	MLP	not pretrained	0.592	0.646	0.578	0.551
	CNN	not pretrained	0.583	0.654	0.570	0.525
	PopT-v2	BrainTreeBank	**0.599**	**0.659**	**0.585**	**0.553**
Waveform						
	Logistic	not pretrained	0.541	0.589	0.515	0.520
	MLP	not pretrained	0.554	0.612	0.523	0.527
	CNN	not pretrained	0.554	0.616	0.524	0.523
	HTNet	not pretrained	0.574	0.636	0.556	0.530
	BaRISTA	BrainTreeBank	0.593	0.655	**0.583**	0.540
	DIVER-1	private+ AJILE12	**0.593**	**0.666**	0.569	**0.545**

Table 4: **All benchmark results.** Within-session ROC-AUC performance on all supported units, grouped by the same fixed-track and model-native families as Table 1. Cells are shaded with a normalized color scheme within column and track. Best per column and track are bolded, and second-best values are underlined.

Track	Model	Pretraining	Overall	Neuroprobe	BYD	Pippi
Multi-STFT						
	Logistic	not pretrained	0.521	0.536	0.509	0.519
	MLP	not pretrained	**0.527**	0.541	0.512	**0.529**
	CNN	not pretrained	0.523	**0.550**	0.512	0.506
	PopT-v2	BrainTreeBank	0.525	0.539	**0.521**	0.515
Waveform						
	Logistic	not pretrained	0.519	0.535	0.510	0.513
	MLP	not pretrained	0.520	0.534	0.510	0.518
	CNN	not pretrained	0.515	0.522	0.505	**0.518**
	HTNet	not pretrained	0.509	0.528	0.500	0.500
	BaRISTA	BrainTreeBank	0.523	0.541	0.513	0.514
	DIVER-1	private+ AJILE12	**0.526**	**0.548**	**0.518**	0.512

Table 5: **Challenge benchmark results.** Within-session ROC-AUC performance on Challenge units, grouped by the same fixed-track and model-native families as Table 1. Cells are shaded with a normalized color scheme within column and track. Best per column and track are bolded, and second-best values are underlined.

D BrainBERT Model-Native Scorecard

Table 6 evaluates BrainBERT (Wang et al., 2023) through its model-native single-STFT route and includes corresponding non-pretrained single-STFT decoders on the Main unit decodability subset. We evaluate the released BrainBERT checkpoint as a frozen feature extractor with a trained linear head. Direct single-STFT baselines achieve higher overall ROC-AUC under the benchmark protocol. The evaluation uses one-second windows and training-split normalization, differing from the checkpoint's five-second pretraining windows and original normalization; it therefore characterizes the released representation under these conditions.

Track	Model	Pretraining	Overall	Neuroprobe	BYD	Pippi
BrainBERT STFT						
	Logistic	not pretrained	0.630	0.695	**0.627**	0.569
	MLP	not pretrained	**0.634**	0.696	0.622	**0.583**
	CNN	not pretrained	0.620	**0.702**	0.603	0.553
	BrainBERT (frozen + linear head)	BrainTreeBank	0.619	0.664	0.619	0.575

Table 6: **BrainBERT model-native preprocessing scorecard on Main units.** Within-session ROC-AUC performance on Main units for BrainBERT and corresponding non-pretrained single-STFT decoders. BrainBERT is evaluated using frozen representations pretrained on BrainTreeBank, while Logistic, MLP, and CNN are fit directly on the dataset-specific single-STFT route. Overall is the unweighted mean of the three dataset-level scores; each dataset score is a unit-weighted mean. Cells are shaded within column, with the best value bolded and the second-best value underlined.

E MAIN AND CHALLENGE UNIT COVERAGE

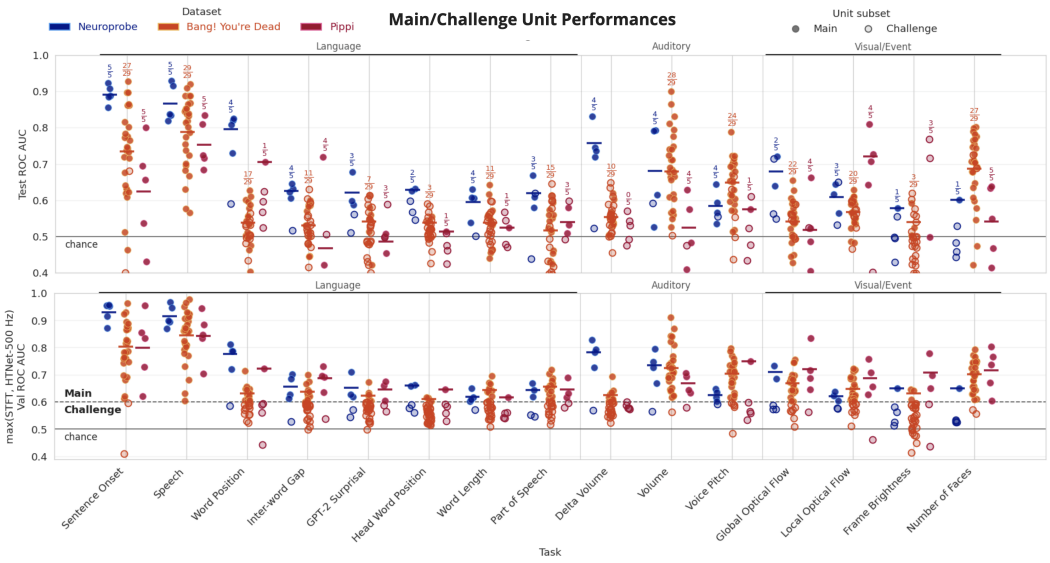


Figure 5: **Main and Challenge unit performance across tasks.** Each point shows an evaluation unit's fold-averaged ROC-AUC. The top panel reports test performance for the STFT screening decoder, while the bottom panel reports the validation screening score: the maximum mean validation ROC-AUC across the STFT screening decoder and HTNet at 500 Hz. The dashed line marks the Main-unit threshold (validation ROC-AUC > 0.60), and the solid line marks chance performance.

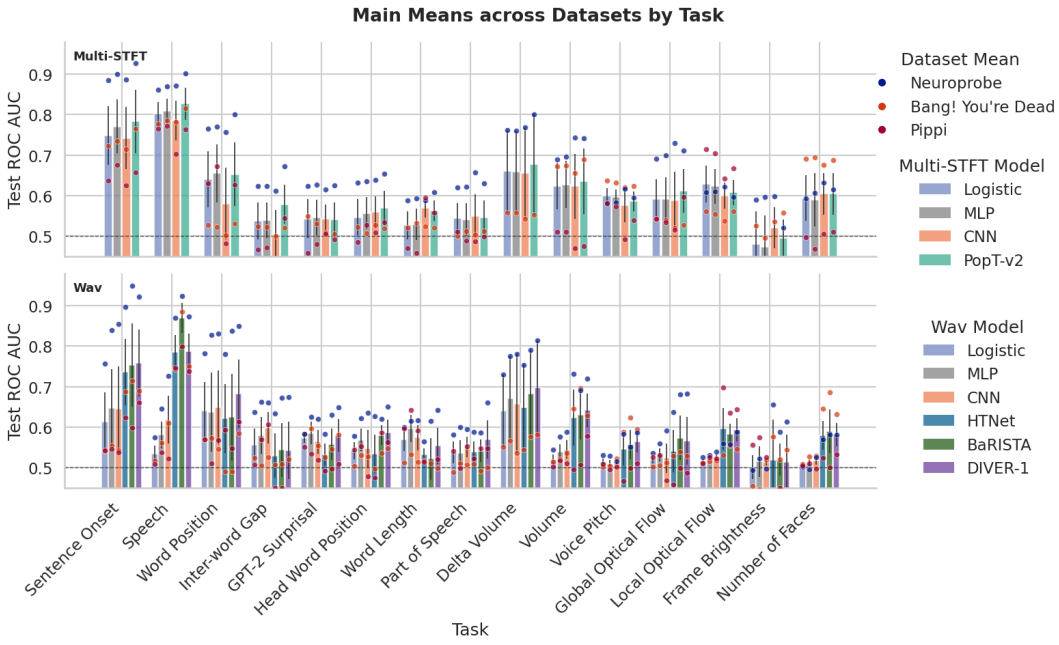


Figure 6: **Main-unit task breakout by model family.** Bars show mean test ROC-AUC on Main units for each task and model family, with points showing dataset means. The dashed horizontal line marks chance performance. Error bars indicate SEM across the 3 dataset means.

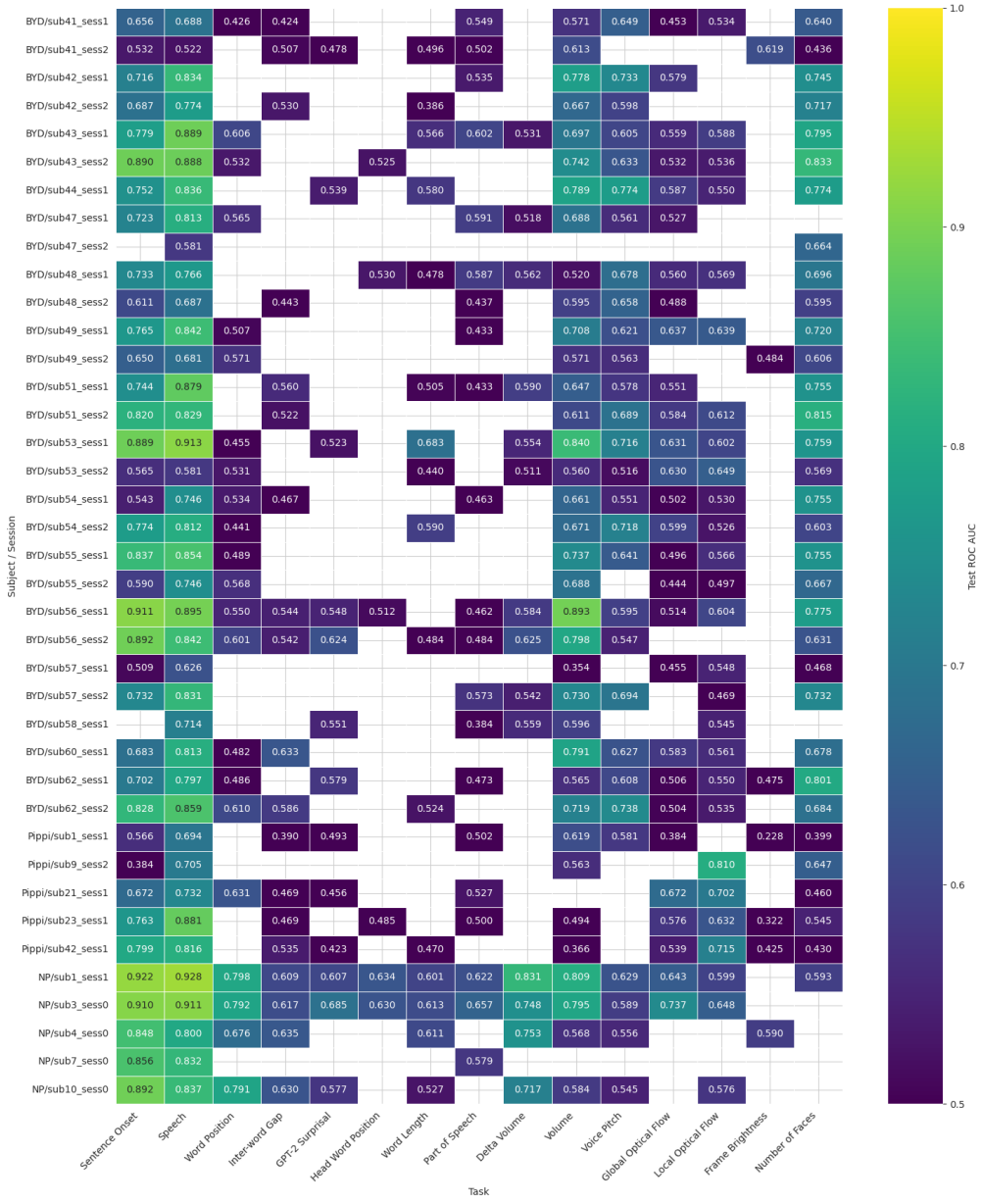


Figure 7: **Main subject/session coverage across benchmark tasks.** Cells show fold-averaged test ROC-AUC from the STFT screening decoder for Main subject/session–task units. Blank cells correspond to Challenge or unsupported units, and higher values indicate stronger screening-decoder test performance.

F TASK-LEVEL WITHIN-SESSION SCALING

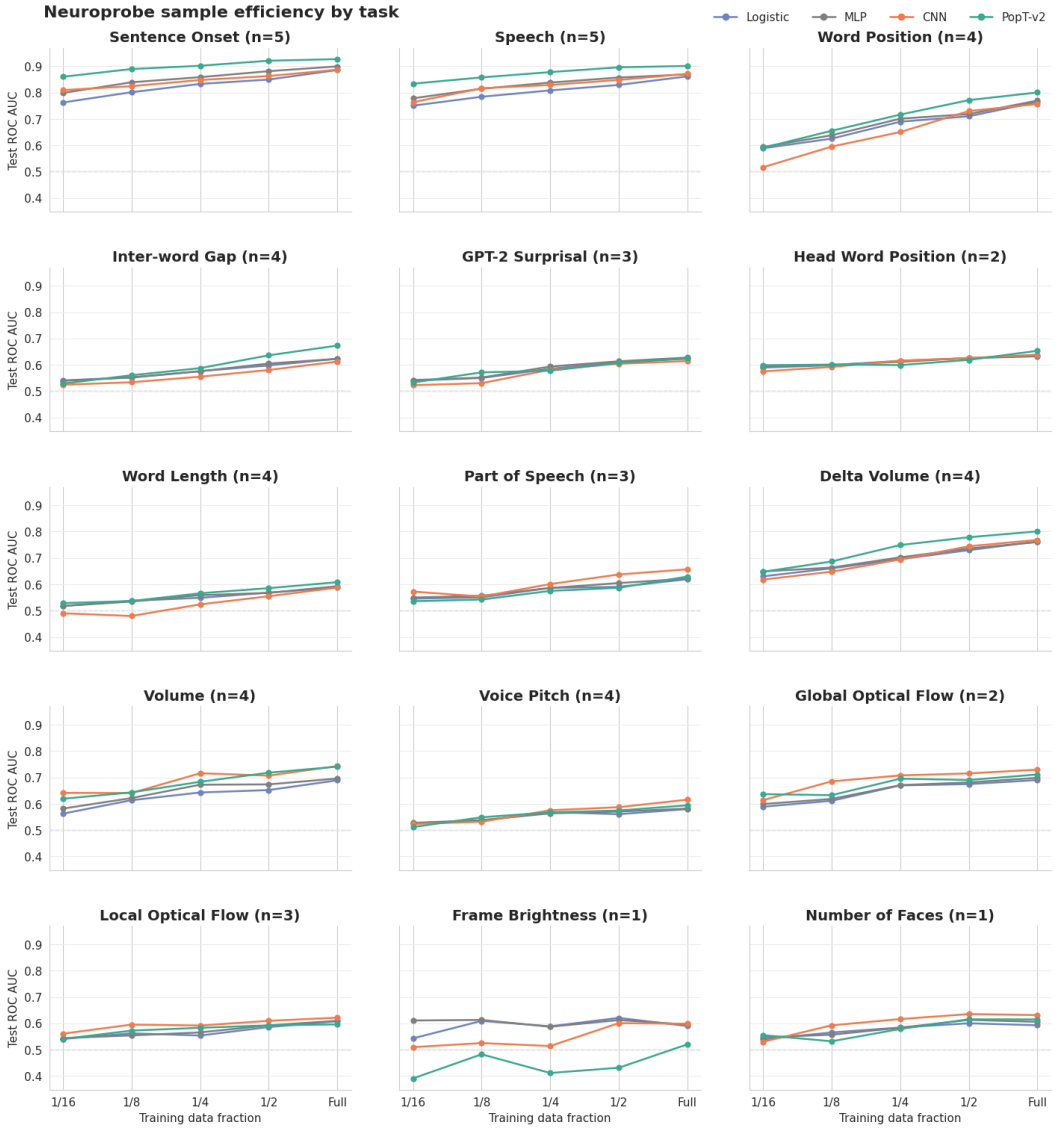


Figure 8: **Neuroprobe within-session sample efficiency by task.** Each panel decomposes Figure 3a by decoding task and model family, showing unit-weighted mean test ROC-AUC as the available target-session training data increase from one-sixteenth to the full training set. Only complete-case task/subject-session/fold units containing every model family and data fraction are retained. Panel titles report the number of unique supported subjects (n), and the dashed horizontal line marks chance performance.

G TASK-LEVEL SCALING TRAJECTORIES

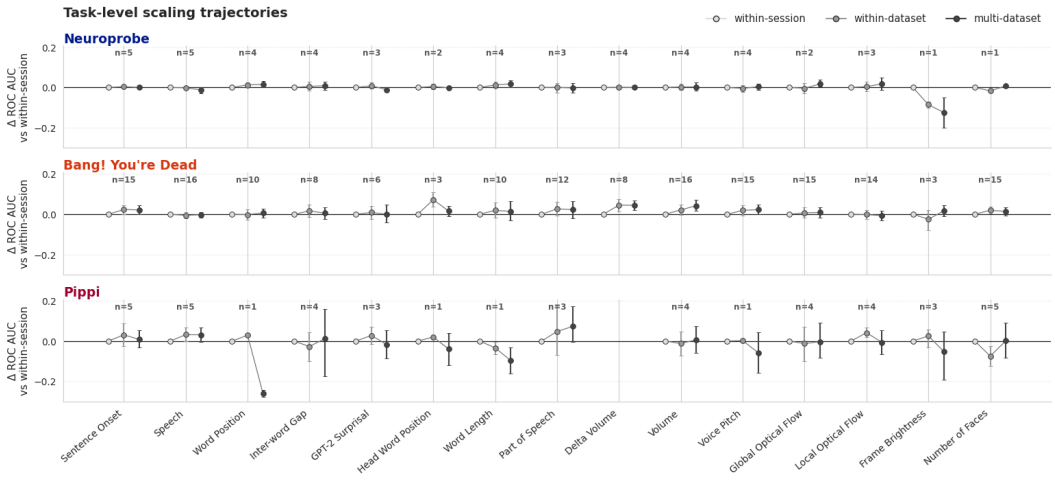


Figure 9: **Task-level scaling trajectories across datasets.** For PopT-v2, points show the change in ROC-AUC for each decoding task relative to within-session evaluation when scaling to within-dataset and multi-dataset training; within-session values are therefore fixed at zero. Rows correspond to Neuroprobe, Bang! You're Dead (BYD), and Pippi. Error bars denote 95% confidence intervals, and labels above each task report the number of unique supported subjects (n). Positive values indicate improvement over within-session training, while negative values indicate reduced performance.

H MULTI-STFT INPUT EXAMPLE

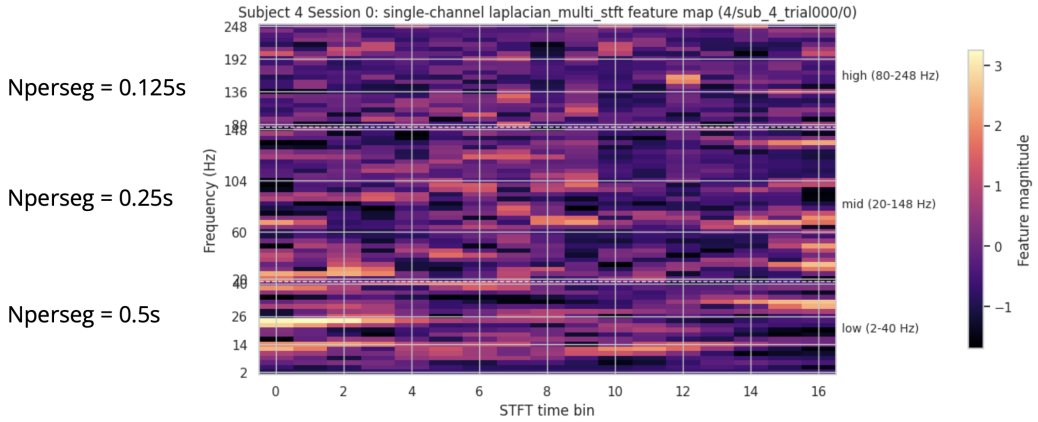


Figure 10: **Example Multi-STFT input representation.** The low-, mid-, and high-frequency sections use 0.5-, 0.25-, and 0.125-second windows with an approximately 62-ms hop and target 2–40, 20–150, and 80–250 Hz, respectively. The sections are concatenated along frequency with aligned time bins. Colors show feature magnitudes standardized per channel and frequency bin using training-split statistics.

Setting	Low	Mid	High
Window samples at 2048 Hz	1024	512	256
Window samples at 1000 Hz (BYD)	500	250	125
Retained upper frequency at 2048 Hz (Hz)	40	148	248

Table 7: **Exact Multi-STFT settings.** Window sample counts specify `nperseg`; hops are 128 samples at 2048 Hz and 62 samples at 1000 Hz. Retained upper frequencies reflect discrete-bin spacing.

I MODEL DETAILS

Generic decoders. Logistic regression uses a scikit-learn linear classifier on flattened aligned-channel inputs with maximum 10,000 iterations, tolerance 10^{-3}, and random seed 42. The MLP flattens each input sample and applies two 128-unit hidden layers with ReLU activations and 0.2 dropout. The CNN uses three convolutional blocks with 32, 64, and 128 channels, kernel size 3, max pooling, and 0.5 dropout, followed by a 128-unit fully connected layer. Convolutions are 1D for waveform inputs and 2D for STFT-style inputs.

We keep these model hyperparameters fixed across datasets, tasks, and the Multi-STFT and Waveform tracks. The MLP and CNN use Adam, learning rate 10^{-4}, batch size 200, and validation-based early stopping for up to 100 epochs. For waveform inputs, filtering uses a 15-second context around each one-second target window to reduce boundary effects; the signal is then cropped to the target window before rereferencing, resampling, and normalization.

PopT-v2. PopT-v2 uses a PopT-style Transformer to aggregate sparse and variable electrode sets across subjects (Chau et al., 2025). It has 6 layers, hidden dimension 512, 8 attention heads, 2048-dimensional feed-forward layers, multi-subject positional encoding, and a classification-token pretraining head.

We pretrain PopT-v2 from scratch on the Brain Treebank subset used by PopT: 10 subjects across 19 movie recordings (Chau et al., 2025). Following TSAP's finding that PopT pretraining benefits from matching pretraining and downstream window lengths (Patel et al., 2025), we use one-second Laplacian Multi-STFT windows with the duration-defined track settings and per-channel keyed standardization. Pretraining uses a next-segment prediction and token-replacement objective, with 1-second windows sampled every 0.2 seconds, variable channel subsets of 10–100 channels, and 10% token replacement. Training runs for 1,000,000 steps with LAMB, learning rate 10^{-4}, batch size 256, gradient clipping at 1.0, a 2.5% warmup ramp-up schedule with decay factor 0.99, and bfloat16 mixed precision. Pretraining takes 48 hours on a single A100 GPU.

For benchmark fine-tuning, we update the pretrained Transformer and a new classification head with AdamW, weight decay 10^{-2}, learning rates 5×10^{-5} and 5×10^{-4} respectively, batch size 128, and 1,000 gradient-update steps.

HTNet. We evaluate an HTNet-style implementation (Peterson et al., 2021), based on the EEGNet-derived architecture (Lawhern et al., 2018), on one-second waveform inputs resampled to 500 Hz. We retain $F_1 = 8$, depth multiplier $D = 2$, $F_2 = 16$, and dropout 0.5, and adapt the temporal and separable kernel lengths to 32 and 8 samples for one-second inputs. Training uses Adam, learning rate 10^{-3}, batch size 128, and validation-based early stopping with patience 10 for up to 100 epochs. HTNet uses the same filtering and cropping procedure as the generic waveform decoders.

BaRISTA. BaRISTA (Oganesian et al., 2025) is a self-supervised intracranial waveform model whose central design feature is spatial encoding and masking at configurable spatial scales beyond individual channels: temporal patch representations can be encoded and masked at the level of individual channels, parcels, or lobes, making the granularity of spatial representation an explicit modeling choice. It is pretrained on 30 hours of Brain Treebank data, segmented into 3-second chunks sampled at 2048 Hz with 250 ms temporal patches. Pretraining excludes all recording sessions used for testing in the benchmark. Pretraining waveforms are Laplacian rereferenced, notch filtered, and high-pass filtered at 0.5 Hz. Destrieux atlas labels define the spatial categories, following the original paper's best configuration: parcel-level encoding and channel-level masking.

For benchmark evaluation, we freeze the tokenizer and retain the 250 ms temporal patch size. Fine-tuning uses learning rate 10^{-3}, weight decay 10^{-2}, batch size 128, and early stopping on validation ROC-AUC with patience 15 for up to 100 epochs. We extend the Destrieux spatial encoding to Destrieux-ASEG to support more subcortical regions.

Evaluation filtering is performed separately within each recording and train, validation, or test split. Selected windows are ordered by start time, concatenated for 0.5 Hz high-pass and notch filtering, then divided back into the original windows; the procedure does not filter the continuous recording or mix samples across splits. BYD is filtered at 1000 Hz and upsampled to 2048 Hz; Neuroprobe and

Pippi remain at 2048 Hz. Filtering is followed by Laplacian rereferencing, robust global scaling, and per-window, per-channel temporal standardization before tokenization.

DIVER-1. We evaluate the 0.1-second-patch, Tiny-width variant of DIVER-1 (Han et al., 2026). It was pretrained to reconstruct masked raw inputs and STFT/FFT features at the patch level for 32 epochs on private iEEG data and AJILE12 (Peterson et al., 2022), totaling 5,310 hours. Pretraining inputs use a 0.5 Hz high-pass filter and notch filtering.

Benchmark evaluation uses the original paper's frozen setting: learning rate 2×10^{-3}, weight decay 10^{-2}, batch size 32, and early stopping on validation ROC-AUC for up to 40 epochs. DIVER-1 uses Montreal Neurological Institute (MNI) coordinates when available; Pippi is evaluated without coordinate inputs because the benchmark release does not include Pippi MNI coordinates.

The DIVER-native evaluation route filters a 15-second context with a 0.5 Hz high-pass filter followed by notch filters at 60, 120, and 180 Hz, crops to the one-second target window, applies Laplacian rereferencing, and resamples to 500 Hz. No additional standardization is applied.

Waveform preprocessing. Table 8 summarizes the sampling, filtering, and normalization conventions for each system. Baseline comparisons did not show consistent improvements from higher sampling rates. BaRISTA and DIVER-1 retain model-specific preprocessing and pretraining exposure, so comparisons with the fixed-track baselines evaluate complete systems.

System	Input rate	Filtering procedure	Input normalization
Logistic, MLP, CNN, HTNet	500 Hz	Filter with 15-second context, then crop to the one-second target window.	Training-fitted robust global normalization; one median and median-absolute-deviation scale across channels and timepoints.
BaRISTA	2048 Hz	Concatenate selected windows in temporal order within each recording and split for filtering; upsample BYD from 1000 Hz.	Robust global scaling followed by per-window, per-channel temporal standardization.
DIVER-1	500 Hz	Filter with 15-second context, then crop to the one-second target window.	No additional input standardization.

Table 8: **Waveform-track preprocessing recipes.** All systems use 0.5 Hz high-pass filtering, line-noise notch filtering, and Laplacian rereferencing. Model-specific variants retain their sampling and normalization conventions.

I.1 Additional model coverage

MVPFormer and Brant use temporal input representations that require additional adaptation for iMINDBench's one-second naturalistic decoding windows.

MVPFormer. The released MVPFormer configuration we examined (Carzaniga et al., 2026) encodes each channel in five-second segments and combines 25 segments of temporal context. This temporal granularity differs from iMINDBench's one-second prediction windows. Retaining the released context would provide information beyond the benchmark window and require explicit handling of chronological fold boundaries; shortening or rescaling the input changes the signal presented to the pretrained encoder. Including MVPFormer therefore requires a validated mapping between its temporal representation and the benchmark's prediction targets.

Brant. Brant's published preprocessing (Zhang et al., 2023) uses six-second patches with both temporal and frequency-domain representations. Adapting one-second recordings to this input format changes the temporal support of those representations. In particular, stretching a short window to the expected patch length changes its effective timescale and requires care when computing frequency features. A benchmark-compatible adaptation must establish how these representations are constructed while preserving the intended signal interpretation.

We defer inclusion of these models in the main comparison pending validation of these adaptations. These cases motivate documenting model-specific input requirements alongside standardized benchmark protocols.

I.2 Benchmark Runtime Estimates

Cumulative per-job timing records are normalized to four concurrent workers. Preprocessing and CPU fitting assume a node with two 32-core AMD EPYC 7513 processors; neural fitting assumes four NVIDIA A100 SXM4 GPUs (80 GB each), with one worker per GPU. Estimates cover all three datasets and both folds and are not measured end-to-end wall-clock times.

Preprocessing route	Worker-hours	Four-worker wall time
Multi-STFT	∼5.5 h	∼1.4 h
Waveform	∼80 h	∼20 h

Table 9: **Estimated preprocessing time.** Waveform preprocessing is estimated from the HTNet runs that built the cache reused by waveform Logistic, MLP, and CNN.

Track	Model	Fitting worker-hours	Four-worker wall time
Multi-STFT	Logistic	0.8 h	∼0.2 h
Multi-STFT	MLP	3.9 h	∼1.0 h
Multi-STFT	CNN	5.6 h	∼1.4 h
Multi-STFT	PopT-v2	72.6 h	∼18.2 h
Waveform	Logistic	0.4 h	∼0.1 h
Waveform	MLP	79.1 h	∼19.8 h
Waveform	CNN	64.0 h	∼16.0 h
Waveform	HTNet	14.1 h	∼3.5 h
Model-native	BaRISTA	27.4 h	∼6.8 h
Model-native	DIVER-1	68.0 h	∼17.0 h

Table 10: **Estimated model-fitting time.** Fitting excludes subject loading and preprocessing.

J LEADERBOARD

The interactive leaderboard is illustrated in Figure 11; its public link is omitted for anonymous review. The public website and leaderboard will support comparisons across datasets, tasks, preprocessing tracks, scaling domains, and unit decodability subsets. To submit a model, users will run the evaluation pipeline and submit generated result artifacts and model metadata through a pull request to the leaderboard repository, where automated checks validate the submission. Detailed instructions will be provided in the submission guide.

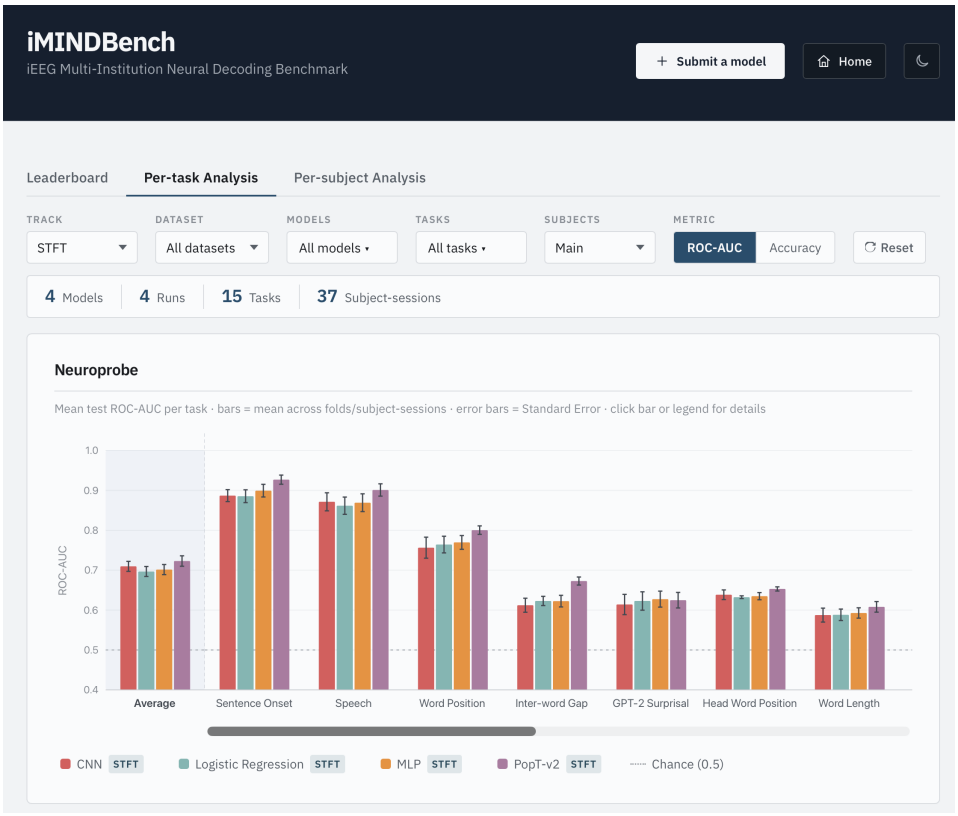


Figure 11: **Decomposed results viewer.** The leaderboard supports comparisons across datasets, tasks, preprocessing tracks, scaling domains, and unit decodability subsets.

K AutoResearch Feasibility Study

As a scoped proof of use, we applied an AutoResearch-style agentic experimentation loop (Karpathy, 2026) to tune the PopT-v2 pretraining recipe starting from a fixed 780k-step single-STFT checkpoint. The best selected variant improved Neuroprobe within-session decoding over the original 1M-step single-STFT baseline by ΔROC-AUC $= 0.0066$ (95% CI: 0.0021–0.0111, $p = 0.0049$; Figure 12). This demonstrates how automated tuning systems may support scalable, controlled model comparisons.

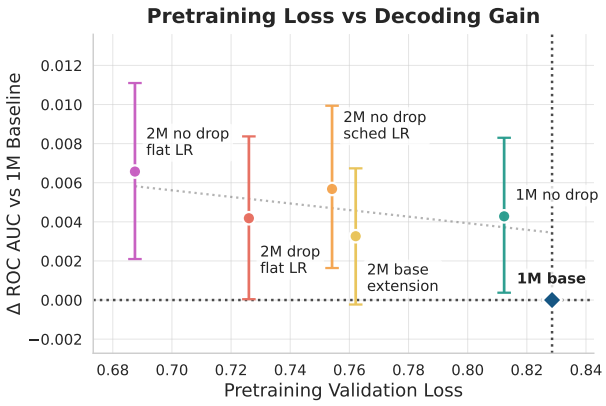


Figure 12: **AutoResearch feasibility result.** AutoResearch-selected PopT-v2 variants with lower pretraining validation loss (colors) generally improve Neuroprobe within-session decoding. Points show variants evaluated across 150 paired units; error bars show 95% bootstrap confidence intervals.

AutoResearch was run in a human-in-the-loop mode from the same 780k-step single-STFT PopT-v2 checkpoint, with human decisions limited to approving phase transitions and extending promising candidates. The search varied optimization and regularization hyperparameters (learning rate, scheduler type, batch size, gradient clipping, weight decay, and dropout) while model architecture, dataset, preprocessing track (`laplacian_stft_time_pooled`), split policy (per-subject, seed 42, 1% validation / 10% test), and validation objective (`nsp_replace_only_pretrain` loss) were held fixed. Across 26 launched jobs (25 valid), the sweep consumed approximately 228.3 GPU-hours over approximately 184.5 active wall-clock hours. Initial screening produced small or inconsistent effects; dropout removal, identified within the initial 780k–820k window (approximately 16.2 GPU-hours, compared to approximately 48 GPU-hours for a full 1M steps pretraining run), was the first robust improvement and was subsequently extended to 2M steps with variants of both scheduled and flat learning rates. As shown in Figure 13, no-dropout variants separate clearly from dropout variants beyond approximately 1.2M steps, with the best model (2M no-dropout flat-LR) achieving validation loss 0.687 versus 0.829 for the original 1M baseline. Because this model uses a single-STFT rather than the benchmark's Multi-STFT track, it is not directly comparable to Table 4; however, its within-session All-subset performance of 0.664 sits above the Multi-STFT PopT-v2 baseline (0.599 overall, 0.659 on Neuroprobe). The total compute cost was approximately $342 (228.3 GPU-hours at $1.50/A100-hour).

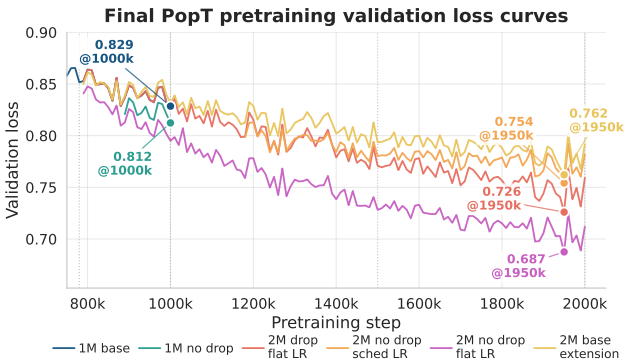


Figure 13: **AutoResearch pretraining validation curves.** All runs continue from a shared 780k-step checkpoint. The 1M baseline (dropout=0.1, flat LR) and 1M no-dropout model terminate early; the four 2M continuations separate beyond approximately 1.2M steps, with no-dropout variants consistently outperforming their dropout counterparts. The best model, 2M no-dropout flat-LR, achieves the lowest final validation loss (0.687), confirming dropout removal as the primary driver of improvement and the flat learning rate as an additional benefit at longer horizons.