
Geometry-Aware Self-Supervised Electrophysiology Representation Learning

Anonymous Author(s)

Affiliation

Address

email

Abstract

Precise localization of recording electrodes is fundamental to systems neuroscience; yet, current methods depend on labor-intensive post-hoc histology, which is incompatible with real-time feedback or chronic implants. We test the hypothesis that the electrical signals being recorded themselves carry sufficient anatomical information to localize their recording site in 3D atlas coordinates, without the need for histology. To probe this, we introduce a self-supervised framework that learns geometry-aware representations of single-channel local field potentials (LFP), leveraging probe channel geometry as a self-supervisory signal. We compare speech-based SSL objectives (e.g., masked prediction) against geometry-aware objectives that constrain the embedding space to mirror physical channel layout along the probe. These representations outperform masked-prediction baselines, and the gains compound at larger data scales. We test downstream analysis with 3D coordinate regression and brain region classification, benchmarking three SSL backbones (Wav2Vec 2.0, Whisper, and Data2Vec) against supervised baselines (AnyNet, ViT, and classical spectral classifiers) across datasets, laboratories, species, and probe technologies. Our method achieves state-of-the-art results in brain region classification and 3D coordinate regression from raw single-channel LFP, outperforming all baselines, affirming our hypothesis.

1 Introduction

The ability to precisely determine where in the brain a recording electrode resides is a prerequisite for almost every analysis in systems neuroscience. Without accurate localization, inferences about cell-type specificity [2], circuit motifs [3], and disease biomarkers [4] become unreliable. In clinical settings such as deep brain stimulation (DBS), the precision of electrode placement can significantly affect therapeutic outcomes [5, 6]. Current practice relies on atlas-based surgical planning [8] combined with post-hoc histological verification [9], an approach that is labor-intensive, error-prone, destructive, and fundamentally incompatible with chronic recordings, flexible probes, or real-time clinical feedback.

Recent work has begun to show that neural signals themselves encode anatomical identity. Tolossa et al. [10] demonstrated that spike trains contain a “conserved code for anatomy,” achieving moderate single-neuron region classification via ISI-distribution analysis. The NEMO framework [11] applies multimodal contrastive learning to waveforms and autocorrelations for cell-type and brain-region classification on the International Brain Laboratory (IBL) Brain-Wide Map dataset. Most relevant to the present work, Lfp2vec [12] showed that self-supervised learning on raw local field potentials (LFP) can decode brain regions with high accuracy, transferring representations from an audio-pretrained wav2vec 2.0 model [14]. Wang et al. [13] independently demonstrated that LFP spectrograms exhibit a modular, brain-wide organization through “deep electrospectroscopy” using the same IBL dataset.

Despite these advances, critical gaps remain. Lfp2vec was evaluated on coarse classification tasks involving only five brain regions. Wang et al. achieved only 39% cross-trial accuracy on a 472-class problem using hand-crafted spectrograms rather than learned features. No existing method predicts continuous stereotaxic coordinates from single-channel raw LFP, nor do current approaches incorporate explicit spatial priors into the self-supervised objective to encourage anatomically grounded representations. Furthermore, the choice of audio backbone has not been systematically evaluated: wav2vec 2.0 was the sole architecture explored by Lfp2vec, while Whisper [15] and Data2vec [17] offer distinct architectural inductive biases and pretraining strategies that may better suit neural signals.

In this work, we test the hypothesis that local field potentials carry sufficient anatomical information to localize their recording site in 3D atlas coordinates, without histology. To probe this, we introduce a self-supervised framework that addresses the limitations above through three contributions:

1. **Geometry-aware SSL objectives.** We compare standard masked-prediction against two spatial contrastive losses that use probe channel geometry as a free supervisory signal, requiring no atlas registration.
2. **Systematic backbone comparison at scale.** We evaluate three audio foundation models as LFP encoders across data scales spanning over two orders of magnitude (5K to 1M samples) and compare against supervised deep learning baselines (AnyNet, ViT) and classical spectral features.
3. **Fine-grained brain-wide localization.** We show that single-channel raw LFP supports both 309-class brain region classification within the Allen Brain Atlas hierarchy and continuous CCF coordinate regression with millimetric precision. These capabilities generalize across laboratories, probe technologies, and species.

Our results affirm the hypothesis: LFP signals encode anatomically localizable information that is recoverable without histology, and geometry-aware representations substantially outperform both masked-prediction and supervised baselines, with gains that compound at larger data scales.

2 Datasets

We constructed our pretraining corpus from two complementary large-scale open Neuropixels datasets. The **Allen Brain Observatory Visual Coding** dataset [37] provides Neuropixels 1.0 recordings from 58 mice (~51 sessions, up to six probes per session, 384 channels per probe at 1,250 Hz). After removing out-of-brain electrodes, we retained 80,313 valid channels registered to the Allen Common Coordinate Framework (CCF) [38] with full 3D coordinates (AP, DV, ML in μm), spanning 115 brain regions across the Allen CCF hierarchy. We segment continuous recordings into 3-second chunks (3,750 samples) and construct pretraining sets at five scales: 5K, 25K, 75K, 240K, and 1M samples. The **IBL Brain Wide Map dataset** [40] adds cross-laboratory variability from multiple institutions worldwide, with standardized insertions through frontal cortex, striatum, hippocampus, and thalamus, providing both complementary brain coverage and an independent transfer evaluation set, increasing the amount to 309 regions.

For evaluation, we hold out 94,317 samples across 309 regions using a strict across-session split (12 sessions reserved for testing, 5-fold evaluation). We additionally test cross-species generalization on macaque Neuropixels-NHP recordings targeting prefrontal and motor cortex.

3 Methods

Our approach takes a single 3-second LFP chunk from one channel, encodes it into a geometry-preserving embedding, and maps that embedding to both a continuous 3D coordinate in atlas space and a discrete brain region label. The full pipeline is illustrated in Figure 1 and involves three stages: signal preprocessing, self-supervised representation learning, and downstream localization.

3.1 Signal Preprocessing

Raw LFP recordings undergo standard destriping [42]: high-pass filtering at 2 Hz, bad-channel interpolation, and common-mode noise subtraction. The cleaned signal is segmented into 3-second windows, each treated as an independent sample.

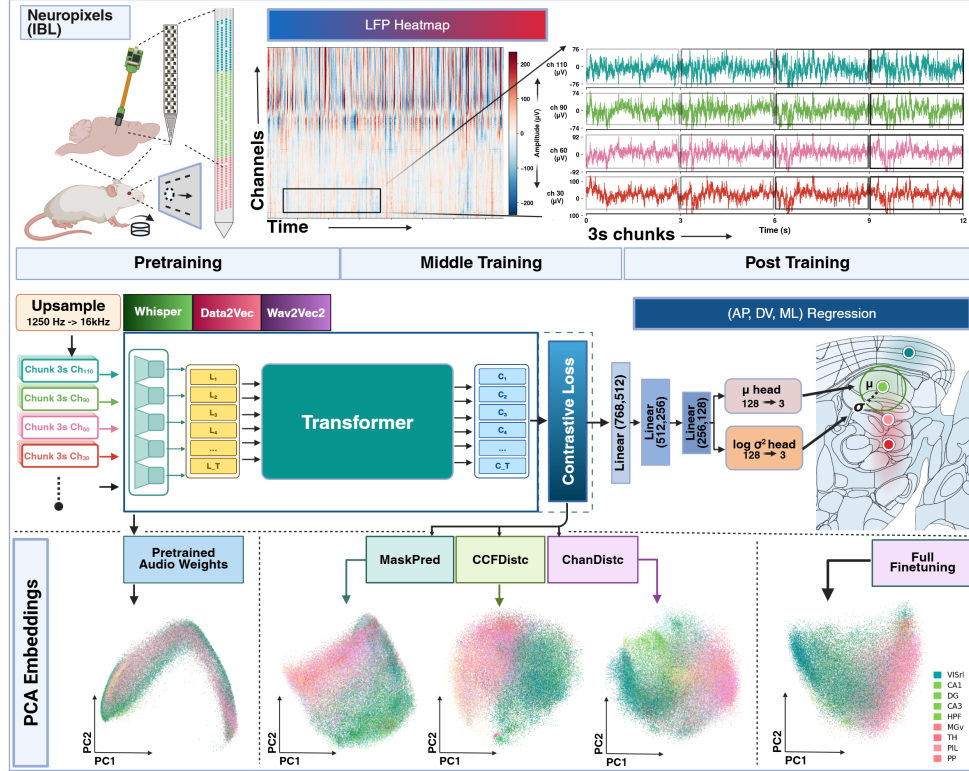


Figure 1: **Pipeline overview.** **Top (Pretraining):** Neuropixels probes record multi-channel LFP; the channel $\times$ time heatmap is segmented into 3-second single-channel chunks, each yielding LFP Time series. (right). **Middle (Training):** Chunks are upsampled from 1,250 Hz to 16 kHz and encoded by one of three audio foundation model backbones (Whisper, Data2Vec, Wav2Vec2). The Transformer produces contextualized token sequences ($C_1, \dots, C_T$) that are trained with one of three objectives (MaskPred 3.3.1, CCFDistic 3.3.2, or ChanDistic 3.3.3). The pooled 768-d embedding passes through a shared MLP trunk (768 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128) that branches into a μ head (128 $\rightarrow$ 3, predicting AP/DV/ML coordinates) and a $\log \sigma^2$ head (128 $\rightarrow$ 3, predicting per-axis uncertainty), shown on the sagittal atlas inset. **Bottom (Post Training):** PCA projections of the embedding space at five stages: pretrained audio weights, after MaskPred, after CCFDistic, after ChanDistic, and after full finetuning. Geometry-aware objectives (CCFDistic, ChanDistic) produce visibly more structured embeddings than MaskPred or audio-only weights.

86 A key design choice is resampling each 3,750-sample segment (1,250 Hz) to 48,000 samples (16 kHz)
87 before encoding. Audio foundation models were pretrained with convolutional front-ends tuned to
88 16 kHz input; feeding them lower-rate signals would leave most learned filters operating outside
89 their intended frequency range, effectively discarding pretrained knowledge. Upsampling via sinc
90 interpolation preserves the original spectral content below the Nyquist frequency (625 Hz) while
91 filling in intermediate samples, allowing the convolutional layers to engage their full filter bank at the
92 temporal resolution they were designed for.

93 3.2 Backbone Architectures

94 We systematically evaluate three audio foundation models as LFP encoders, each bringing different
95 inductive biases:

- 96 • **wav2vec 2.0** [14]: A 7-layer 1D CNN feature extractor followed by a 12-layer Transformer with
97 768-dimensional hidden states. Pretrained with a contrastive masked-prediction objective over
98 quantized speech tokens.

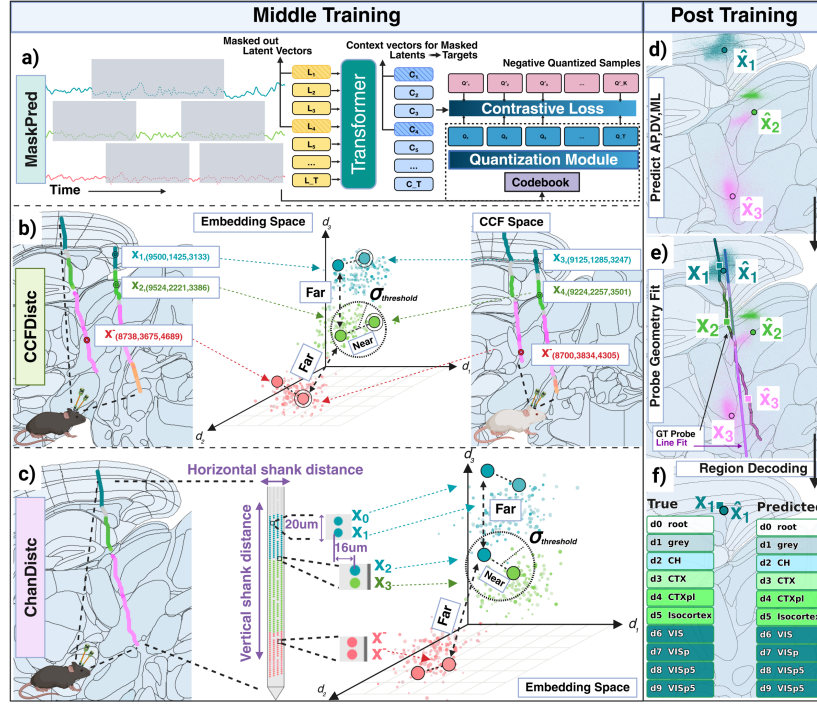


Figure 2: **Middle and Post Training Objectives.** **a) (MaskPred 3.3.1):** Random temporal spans of CNN-encoded LFP latents are masked (grey boxes); the transformer learns, via contrastive loss, to identify each masked position’s product-quantized codebook target among distractors drawn from other masked positions across channels. **b) (CCFDistc 3.3.2):** Pairs are defined using 3D atlas coordinates from two probes. Channels at nearby CCF locations (x_1, x_2) are pulled together in embedding space while anatomically distant channels (x^-) are pushed beyond margin σ . **c) (ChanDistc 3.3.3):** Positive and negative pairs are defined by physical distance along the probe shank ($20 \mu\text{m}$ vertical pitch, $16 \mu\text{m}$ horizontal pitch), requiring no atlas registration. Nearby channels (x_0, x_1) form positives; distant channels (x^-, x) form negatives. **d) (Predict (AP, DV, ML)) 3.4:** A Gaussian regression head decodes each channel’s transformer embedding into 3D Allen CCF coordinates ($\hat{x}_1, \hat{x}_2, \hat{x}_3$ in AP/DV/ML), trained against ground-truth voxels with negative log-likelihood. **e) (Probe Geometry Fit) 3.5:** Per-channel predictions across a probe are jointly aligned (Kabsch [22]) to the ground-truth shank line, exploiting the prior that all channels of a probe are collinear to refine each estimate posthoc. **f) (Region Decoding) 3.5:** Refined coordinates are voxelized and looked up in the Allen CCF annotation, then walked up the structure tree to yield a depth-resolved region label (d0 root $\rightarrow$ d9 leaf), enabling hierarchical classification.

- 99 • **Whisper** [15]: An encoder-decoder architecture pretrained on 680K hours of weakly-supervised
100 speech. We use only the encoder, which processes log-mel spectrograms through convolutional
101 downsampling and Transformer layers.
- 102 • **Data2vec** [17]: Uses a student-teacher framework where the student predicts contextualized
103 representations from the teacher rather than discrete tokens, avoiding the quantization bottleneck.

104 For all backbones, we extract the last hidden state and mean-pool across the time dimension to obtain
105 a single 768-dimensional embedding per input chunk.

106 3.3 Representation Learning Objectives

107 We compare three objectives that differ in the kind of invariance they encourage:

108 3.3.1 Masked Prediction (MaskPred)

109 This is the standard wav2vec 2.0 objective ported from speech. Given an LFP segment $x \in \mathbb{R}^T$,
 110 we compute tokens $z = E(x)$ and discretize them via Gumbel-softmax quantization $q = Q(z)$. A
 111 random subset of time steps $\mathcal{M}$ is masked, and the Transformer predicts the correct quantized target
 112 at each masked position:

$$\mathcal{L}_{\text{SSL}} = - \sum_{m \in \mathcal{M}} \log \frac{\exp(\text{sim}(c_m, q_m)/\tau)}{\sum_{k \in K(m)} \exp(\text{sim}(c_m, q_k)/\tau)} \quad (1)$$

113 where c_m is the Transformer output at masked step m , q_m is the target quantized vector, and $K(m)$
 114 contains both the target and distractor tokens sampled from other positions. A diversity loss $\mathcal{L}_{\text{div}}$
 115 encourages full codebook utilization.

116 3.3.2 CCF Distance Contrastive (CCFDiste)

117 This objective uses full anatomical supervision to shape the embedding space. For an anchor sample
 118 i with CCF coordinate $\mathbf{x}_i \in \mathbb{R}^3$, we define the positive set $\mathcal{P}_i = \{j : \|\mathbf{x}_i - \mathbf{x}_j\|_2 < \sigma\}$ as channels
 119 within a spatial radius σ , and the negative set $\mathcal{N}_i = \{j : \|\mathbf{x}_i - \mathbf{x}_j\|_2 > \sigma\}$ as channels beyond that
 120 radius (see Figure 2, middle row). The loss follows a margin-based contrastive formulation:

$$\mathcal{L}_{\text{CCF}} = \sum_i \left[\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 + \frac{1}{|\mathcal{N}_i|} \sum_{k \in \mathcal{N}_i} \max(0, m - \|\mathbf{z}_i - \mathbf{z}_k\|_2)^2 \right] \quad (2)$$

121 where $\mathbf{z}_i = f_\theta(x_i)$ is the embedding of sample i , σ is the distance threshold that separates near from
 122 far pairs, and m is a margin hyperparameter. This explicitly forces the latent space to mirror the
 123 physical geometry of the brain: channels at nearby atlas locations (e.g., $\mathbf{x}_1$ and $\mathbf{x}_2$ from the same
 124 probe trajectory) are pulled together while channels from distant regions (e.g., $\mathbf{x}^-$ from a different
 125 brain structure) are pushed apart.

126 3.3.3 Within-Probe Distance Contrastive (ChanDiste)

127 This variant replaces atlas-registered CCF distances with physical distances along the probe shank,
 128 requiring zero anatomical annotation. As shown in Figure 2 (bottom row), Neuropixels electrodes
 129 are arranged with a known 20 μm vertical pitch and 16 μm horizontal pitch. For anchor channel
 130 i at probe position p_i , we define positive and negative sets using the on-probe Euclidean distance
 131 $d_{ij}^{\text{probe}} = \|p_i - p_j\|_2$:

$$\mathcal{L}_{\text{Chan}} = \sum_i \left[\frac{1}{|\mathcal{P}_i^p|} \sum_{j \in \mathcal{P}_i^p} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 + \frac{1}{|\mathcal{N}_i^p|} \sum_{k \in \mathcal{N}_i^p} \max(0, m - \|\mathbf{z}_i - \mathbf{z}_k\|_2)^2 \right] \quad (3)$$

132 where $\mathcal{P}_i^p = \{j : d_{ij}^{\text{probe}} < \sigma_p\}$ and $\mathcal{N}_i^p = \{j : d_{ij}^{\text{probe}} > \sigma_p\}$, with σ_p set in micrometers of probe
 133 distance. Because probe geometry is known at recording time, this objective provides a spatial
 134 inductive bias at zero additional annotation cost.

135 Figure 2 illustrates how all three objectives operate on probe recordings and their qualitatively
 136 different effects on the learned embedding space.

137 3.4 Downstream Localization Head

138 We attach a heteroscedastic Gaussian head to the pooled embeddings that jointly predicts a 3D CCF
 139 coordinate $\hat{\mu} = (\hat{x}, \hat{y}, \hat{z})$ and per-axis log-variance $\log \hat{\sigma}^2$. As shown in Figure 1 (middle), the head
 140 consists of a shared MLP trunk ($768 \rightarrow 512 \rightarrow 256 \rightarrow 128$, with BatchNorm, GELU activations,
 141 and dropout) that branches into two parallel linear projections: a μ head ($128 \rightarrow 3$) producing the
 142 coordinate prediction and a $\log \sigma^2$ head ($128 \rightarrow 3$) producing the per-axis uncertainty. Training
 143 minimizes the negative log-likelihood of a heteroscedastic Gaussian:

$$\mathcal{L}_{\text{loc}} = \sum_{a \in \{x, y, z\}} \left[\frac{(\mu_a - \hat{\mu}_a)^2}{2\hat{\sigma}_a^2} + \frac{1}{2} \log \hat{\sigma}_a^2 \right] \quad (4)$$

where μ_a is the ground-truth coordinate along axis a . The predicted variance acts as a learned uncertainty estimate. To obtain a discrete brain region label, we simply look up the predicted coordinate in the Allen CCF annotation volume.

We evaluate the localization head under three regimes: (1) **fully frozen**, where the entire backbone is fixed and only the head trains; (2) **CNN-frozen**, where the convolutional feature extractor is fixed but the Transformer and head are trainable; and (3) **full finetuning**, where all parameters are updated end to end.

3.5 Inference-Time Refinement

At test time, we apply a two-step post-hoc refinement using only recording metadata. First, predictions across all time chunks for the same (probe, channel) pair are aggregated via confidence-weighted averaging (weights $\propto 1/\hat{\sigma}^2$). Second, we enforce the known $20\ \mu\text{m}$ electrode pitch by fitting a line through consensus coordinates along the probe trajectory using Kabsch Refinement. This consistently tightens localization error without retraining.

4 Experiments and Results

4.1 Scaling Behavior

We studied how each objective scales with pretraining corpus size by training at four data scales: 25K, 75K, 240K, and 1M samples (Figure 5c). At the smallest scales, all objectives perform comparably. As data grows, the geometry-aware objectives show accelerating gains in downstream localization accuracy, while pure MaskPred plateaus. Adding ChanDistc to MaskPred consistently reduces 3D RMSE across all pool sizes, and the benefit increases with scale, confirming that physical-distance supervision becomes more, not less, valuable as data grows.

Baselines. We compare against supervised deep learning and classical methods evaluated under the same splits and post-hoc refinement. **AnyNet** [13] is a convolutional encoder trained end-to-end on LFP spectrograms; **ViT** [13] applies a Vision Transformer to the same spectrograms. On the classical side, **HistGBM/band** [1] uses histogram gradient-boosted trees on band-power features (delta through high-gamma); **kNN/band** and **kNN/PCA-PSD** are non-parametric baselines on raw and PCA-reduced spectral features; and **Ridge/PSD** is an L_2 -regularized linear decoder on power spectral density.

4.2 Continuous Localization Results

Figure 3 presents 3D coordinate regression results. Our method (Wav2Vec 2.0 + ChanDistc, full finetuning) achieves $202\ \mu\text{m}$ RMSE on a best-case probe and $735\ \mu\text{m}$ on a difficult trajectory, outperforming AnyNet ($439 / 1,837\ \mu\text{m}$), ViT ($857 / 2,303\ \mu\text{m}$), and HistGBM ($1,065 / 2,232\ \mu\text{m}$). Panel (a) shows pairwise per-class RMSE at four aggregation levels (region, subject, probe, lab); points above parity favour ours. At every level, the majority of points fall above parity, confirming a consistent advantage.

Panel (b) visualizes the prediction cloud and two reference probes per model. Ours produces a tight point cloud following ground-truth CCF geometry; baselines show greater scatter and systematic bias in deeper structures.

4.3 Cross-Species and Cross-Probe Transfer

We evaluate generalization along two axes: across species and across probe technologies (Figure 4a, rightmost two panels).

Cross-probe (NeuroNexus, 5 classes). Models trained on Neuropixels 1.0 data are tested on NeuroNexus recordings covering hippocampus and cortex. Despite the substantial difference in probe geometry (NeuroNexus electrodes have different impedance, spacing, and shank configuration), our method achieves the highest balanced accuracy among all models, outperforming ViT, AnyNet, and all classical baselines. This result is consistent with the hypothesis that geometry-aware objectives learn anatomy-dependent features rather than probe-specific artifacts.

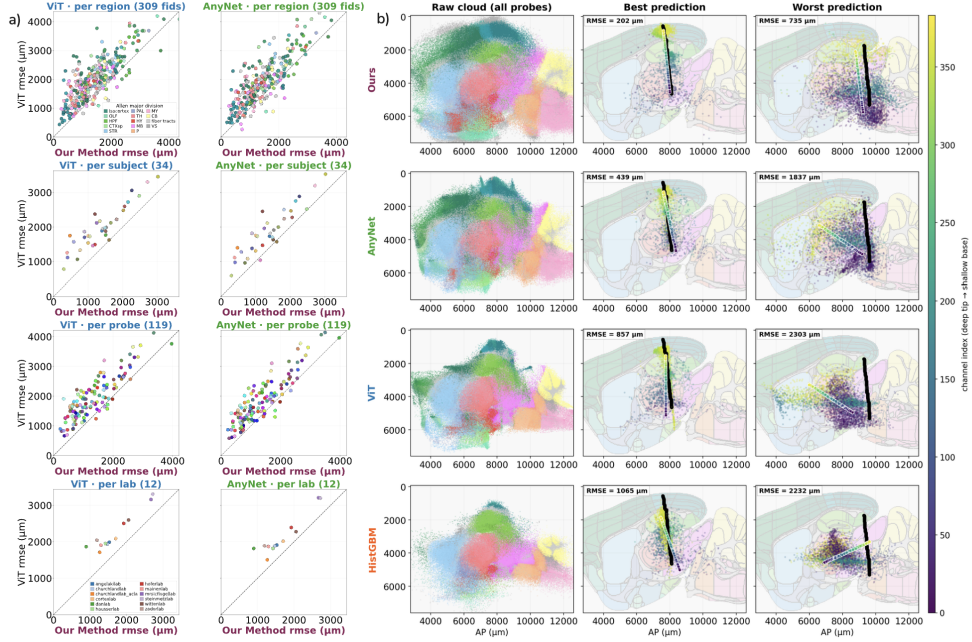


Figure 3: **3D coordinate regression on IBL: Our Method of (MaskPred + ChanDistc 75k (Full finetune)) outperforms ViT, AnyNet and classical PSD/band baselines across every aggregation level.** The metric throughout is 3D RMSE (μm) on the predicted (AP, DV, ML) atlas coordinates. (a) Pairwise per-class 3D RMSE of each baseline (y) against Wav2Vec 2.0(ours) (x), aggregated per region, subject, probe, and lab; points above parity favour ours. (b) Predicted vs. ground-truth CCF coordinates for four models (rows) over the full prediction cloud and two reference probes (best- and worst-case under the anchor) with per-probe Kabsch refinement and 3D RMSE annotated; the same two probes are shown across rows so trajectories are directly comparable. For the allen results look at the supplementary section H

191 **Cross-species (Macaque, 3 classes).** Following the evaluation protocol of Lfp2vec [12], we
 192 apply models trained on mouse data to macaque prefrontal cortex recordings (areas 45, 8, 9/46;
 193 NMT/CHARM-L5 atlas). Our method achieves the highest balanced accuracy (~ 0.59), substantially
 194 exceeding ViT, AnyNet, and all spectral baselines (≤ 0.42). The ROC curves in Figure 4b (rightmost
 195 panel) confirm that this advantage holds across all operating points. These results demonstrate that
 196 the representations learned from mouse LFP transfer meaningfully to primate cortex, suggesting that
 197 the spatial structure captured by geometry-aware SSL is at least partially conserved across species.

198 5 Discussion

199 Our results affirm the central hypothesis: single-channel LFP contains enough anatomical information
 200 to localize recording sites in 3D atlas coordinates without histology. Standard audio SSL objectives
 201 fail to capture the brain’s spatial structure; masked reconstruction learns temporal coherence but
 202 ignores anatomy (Figure 1, bottom). Geometry-aware contrastive losses fix this by incentivizing the
 203 embedding space to mirror physical distances between recording sites.

204 The distinction between our two spatial objectives is worth noting. CCFDistc requires atlas-registered
 205 CCF coordinates, which are available for major public datasets but involve a non-trivial histological
 206 registration step. ChanDistc uses only the known electrode spacing on the probe, which is available
 207 for free at recording time. The fact that ChanDistc approaches the performance of CCFDistc is
 208 encouraging: it suggests that a useful spatial prior can be constructed without any anatomical
 209 annotation at all, opening the door to localization on datasets where histology was never performed.

210 Scaling experiments (Figure 5c) show that geometry-aware objectives become increasingly important
 211 with data size. At 25K all methods perform similarly, but at 240K+ the gap between ChanDistc-

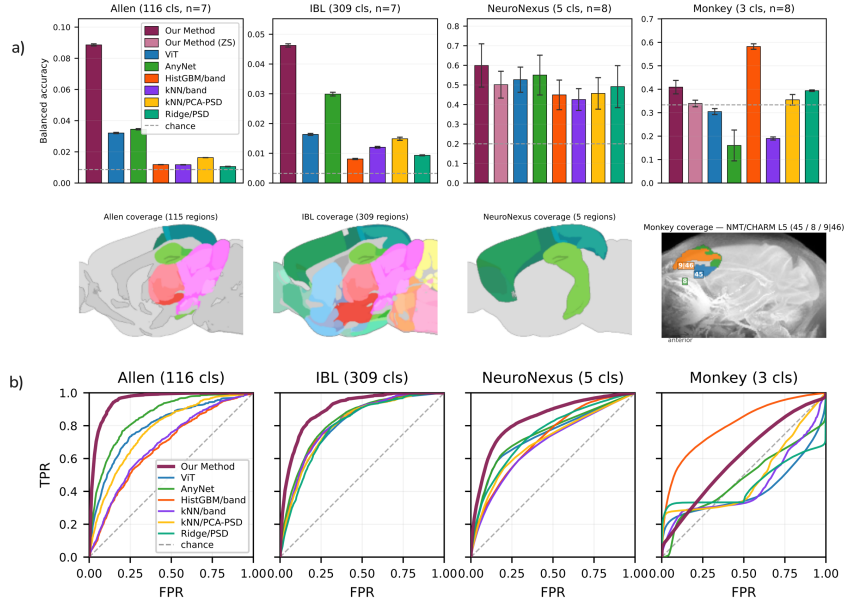


Figure 4: **Brain-region classification across datasets, probe types, and species.** 3D coordinate regression is reported in Fig. 3. **(a)** Top: balanced accuracy (chance level dashed) across four evaluation settings: Allen test set (116 classes, $n=7$ folds), IBL (309 classes, $n=7$), NeuroNexus cross-probe (5 classes, $n=8$), and macaque cross-species prefrontal cortex (3 classes, $n=8$). “Our Method (ZS)” denotes zero-shot transfer without finetuning on the target dataset. Error bars denote cross-fold standard deviation. Bottom: anatomical coverage of each dataset rendered on the Allen CCF (mouse) and NMT/CHARM-L5 (macaque) atlases; covered regions are coloured by their official atlas label, uncovered tissue is gray. **(b)** ROC curves (TPR vs. FPR, macro-averaged over classes). Our method dominates the curve at every operating point except the macaque dataset, where HistGBM outperforms our model, attributable to the single-subject training regime in which the transformer overfits to subject-specific LFP statistics in a way that subject-invariant spectral features do not. This gap is expected to close and reverse as more training subjects become available, consistent with the scaling behavior observed on the mouse dataset. All baselines are evaluated under identical splits and the same post-hoc refinement procedure.

212 augmented models and pure MaskPred is dramatic: the masked objective finds degenerate solutions,
 213 while the spatial loss prevents this by requiring the model to respect brain geometry.

214 The training-stage ablation (Figure 5b) shows that audio pretraining and geometry-aware mid-training
 215 are complementary across all three backbones. For Wav2Vec 2.0, audio-pretrained weights ($2,436 \mu\text{m}$)
 216 improve over random init ($2,267 \mu\text{m}$), MaskPred mid-training reduces error to $2,013 \mu\text{m}$, ChanDistc
 217 to $1,803 \mu\text{m}$, and full finetuning reaches $1,649 \pm 10 \mu\text{m}$. Data2Vec and Whisper-small follow the same
 218 monotonic improvement pattern, converging to $1,697$ and $2,078 \mu\text{m}$ respectively after full finetuning.

219 **Limitations.** Ground-truth labels depend on histological CCF registration quality, and errors propa-
 220 gate into training and evaluation. Rare regions (e.g., amygdala, zona incerta) remain underrepresented.
 221 Our preprocessing relies on the IBL destriping pipeline; end-to-end artifact rejection could improve
 222 robustness. Cross-species results are based on a limited macaque cohort, and clinical translation
 223 requires validation on human recordings against intraoperative biomarkers [51].

224 **Future directions.** Incorporating multi-channel context could improve predictions by exploiting
 225 local spatial correlations, and multi-modal fusion with spike features or imaging may yield richer
 226 representations. Clinically, real-time localization could benefit adaptive DBS, closed-loop BCIs, and
 227 chronic recordings where probe drift makes traditional localization unreliable.

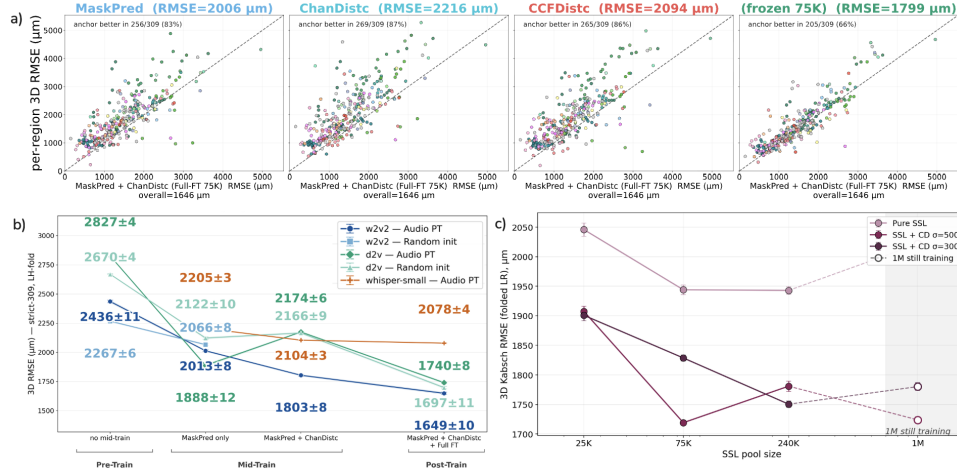


Figure 5: **SSL objective and data-scale ablations on IBL.** (a) **Objective ablation (frozen-backbone XYZ regression at 75K).** Per-region 3D RMSE (μm) of each SSL objective (y -axis) against the full-FT joint anchor (x -axis; overall RMSE = 1,646 μm). Left to right: MaskPred (2,006 μm), ChanDistc (2,216), CCFDistc (2,094), and the joint MaskPred + ChanDistc (1,799). Markers are coloured by Allen major division. Adding geometry-aware ChanDistc to MaskPred closes the bulk of the gap to the fully finetuned anchor, with the anchor outperforming in 66% of regions. (b) **Training-stage curve:** 3D RMSE (strict-309, LH-fold) across four pipeline stages for three backbones: Wav2Vec 2.0 (dark/light blue), Data2Vec (green/cyan), and Whisper-small (orange). Solid markers denote audio-pretrained initialisation; open/dashed markers denote random initialisation. Across all backbones, each successive stage, audio pretraining, MaskPred mid-training, ChanDistc mid-training, and full finetuning, yields additive gains. Wav2Vec 2.0 with audio pretraining reaches the lowest overall error (1,649 ± 10 μm). (c) **Data-scale curve:** 3D RMSE vs. SSL pool size (25K to 1M) for three conditions (Pure SSL, +ChanDistc $\sigma=500$, +ChanDistc $\sigma=300$). Geometry-aware losses keep improving with scale while pure SSL plateaus, confirming that physical-distance supervision becomes *more*, not less, valuable as data grows.

6 Broader Impact

Our framework can reduce reliance on histology in neuroscience, lowering the barrier to large-scale electrophysiology. Clinically, real-time electrode localization from the signal itself could improve deep brain stimulation safety by providing continuous position feedback without additional imaging. Any clinical deployment must, however, include calibrated uncertainty estimates (provided by our heteroscedastic head) and human oversight, as incorrect localization could lead to inappropriate stimulation targets. For the ML community, our results establish a new benchmark for adapting transformer-based foundation models to noisy, nonstationary biomedical time series via geometry-aware inductive biases.

7 Reproducibility

Code for all experiments, including SSL pretraining, downstream evaluation, and baseline reproduction, is available at <https://anonymous.4open.science/r/repo-9687/>.

References

- [1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [2] U. Knoblich, L. Huang, H. Zeng, and L. Li. Neuronal cell-subtype specificity of neural synchronization in mouse primary visual cortex. *Nature Communications*, 10(1):2533, 2019.

- [3] A. Nandi, T. Chartrand, W. Van Geit, A. Buchin, Z. Yao, et al. Single-neuron models linking electrophysiology, morphology, and transcriptomics across cortical cell types. *Cell Reports*, 40(6), 2022.
- [4] A. T. Lee, E. F. Chang, M. F. Paredes, and T. J. Nowakowski. Large-scale neurophysiology and single-cell profiling in human neuroscience. *Nature*, 630(8017):587–595, 2024.
- [5] I. H. Kratter, A. Jorge, M. T. Feyder, et al. Depression history modulates effects of subthalamic nucleus topography on neuropsychological outcomes of deep brain stimulation for Parkinson’s disease. *Translational Psychiatry*, 12(1):213, 2022.
- [6] V. D’Onofrio, L. Weis, L. Rigon, et al. Local field potentials survey to guide DBS programming in Parkinson’s disease: a clinical-neurophysiological longitudinal study. *npj Parkinson’s Disease*, 2025.
- [7] M. Muller, S. Scafa, I. Hanafi, et al. Online prediction of optimal deep brain stimulation contacts from local field potentials in Parkinson’s disease. *NPJ Parkinsons Dis.*, 11(1):234, 2025.
- [8] E. S. Lein, M. J. Hawrylycz, N. Ao, et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, 445(7124):168–176, 2007.
- [9] L. D. Liu, S. Chen, H. Hou, et al. Accurate localization of linear probe electrode arrays across multiple brains. *eNeuro*, 8(6), 2021.
- [10] G. B. Tolossa, A. M. Schneider, E. L. Dyer, and K. B. Hengen. A conserved code for anatomy: Neurons throughout the brain embed robust signatures of their anatomical location into spike trains. *eLife*, 13:RP101506, 2025.
- [11] H. Yu, H. Lyu, E. K. Xu, et al. In vivo cell-type and brain region classification via multimodal contrastive learning. In *ICLR*, 2025.
- [12] T. He, M. Patel, C. Li, A. Maslarova, M. Vöröslakos, N. Ramanathan, W.-L. Hung, G. Buzsáki, and E. Varol. Self supervised learning for in vivo localization of microelectrode arrays using raw local field potential. In *NeurIPS*, 2025.
- [13] X. Wang, J.-C. Béique, and R. Naud. Modular organization of electrical fluctuations in the mouse brain. *bioRxiv*, 2025.10.01.679820, 2025.
- [14] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 33:12449–12460, 2020.
- [15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *ICML*, 2023.
- [16] S. Chen, C. Wang, Z. Chen, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE JSTSP*, 16(6):1505–1518, 2022.
- [17] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022.
- [18] C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu. Brain-BERT: Self-supervised representation learning for intracranial recordings. In *ICLR*, 2023.
- [19] D. Zhang, Z. Yuan, Y. Yang, J. Chen, J. Wang, and Y. Li. Brant: Foundation model for intracranial neural signal. In *NeurIPS*, 36:26304–26321, 2023.
- [20] Z. Yuan, F. Shen, M. Li, Y. Yu, C. Tan, and Y. Yang. Brainwave: A brain signal foundation model for clinical applications. *arXiv preprint arXiv:2402.10251*, 2025.
- [21] M. Azabou, V. Arora, V. Ganesh, et al. A unified, scalable framework for neural population decoding. In *NeurIPS*, 36:44937–44956, 2023.
- [22] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976.
- [23] J. Ye, J. Collinger, L. Wehbe, and R. Gaunt. Neural data transformer 2: Multi-context pretraining for neural spiking activity. In *NeurIPS*, 36:80352–80374, 2023.
- [24] M. Azabou, K. Pan, V. Arora, I. J. Knight, E. L. Dyer, and B. Richards. Multi-session, multi-task neural decoding from distinct cell-types and brain regions. In *ICLR*, 2025.
- [25] Y. Zhang, Y. Wang, M. Azabou, et al. Neural encoding and decoding at scale. *arXiv preprint arXiv:2504.08201*, 2025.

- [26] Y. Chen, M. Orlandi, M. Rapa, et al. PhysioWave: A multi-scale wavelet-transformer for physiological signal representation. In *NeurIPS*, 2025.
- [27] T. Fiedler, L. Hermann, F. Müller, S. Cohen, et al. Teaching Wav2Vec2 the language of the brain. *arXiv:2501.09459*, 2025.
- [28] N. Vattikonda, A. R. Vaidya, R. J. Antonello, and A. G. Huth. BrainWavLM: Fine-tuning speech representations with brain responses to language. *arXiv:2502.08866*, 2025.
- [29] A. Défossez, C. Caucheteux, J. Rapin, O. Kabeli, and J.-R. King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5:1097–1107, 2023.
- [30] G. Buzsáki. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus*, 25(10):1073–1188, 2015.
- [31] A. Berényi, Z. Somogyvari, A. J. Nagy, et al. Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals. *Journal of Neurophysiology*, 111(5):1132–1149, 2014.
- [32] Y. Senzai, A. Fernandez-Ruiz, and G. Buzsáki. Layer-specific physiological features and interlaminar interactions in the primary visual cortex of the mouse. *Neuron*, 101(3):500–513, 2019.
- [33] R. T. Canolty, K. Ganguly, S. W. Kennerley, et al. Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. *PNAS*, 107(40):17356–17361, 2010.
- [34] M. A. Belluscio, K. Mizuseki, R. Schmidt, R. Kempter, and G. Buzsáki. Cross-frequency phase–phase coupling between theta and gamma oscillations in the hippocampus. *Journal of Neuroscience*, 32(2):423–435, 2012.
- [35] G. Buzsáki, C. A. Anastassiou, and C. Koch. The origin of extracellular fields and currents, EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, 13(6):407–420, 2012.
- [36] S. W. Oh, J. A. Harris, L. Ng, et al. A mesoscale connectome of the mouse brain. *Nature*, 508(7495):207–214, 2014.
- [37] J. H. Siegle, X. Jia, S. Durand, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.
- [38] Q. Wang, S.-L. Ding, Y. Li, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D reference atlas. *Cell*, 181(4):936–953, 2020.
- [39] International Brain Laboratory, K. Banga, J. Benson, et al. Reproducibility of in vivo electrophysiological measurements in mice. *eLife*, 13:RP100840, 2025.
- [40] The International Brain Laboratory et al. A brain-wide map of neural activity during complex behaviour. *Nature*, 2025.
- [41] G. N. Angotzi, M. Vöröslakos, N. Perentos, et al. Multi-shank 1024 channels active SiNAPS probe for large multi-regional topographical electrophysiological mapping of neural dynamics. *Advanced Science*, 12(16):e2416239, 2025.
- [42] K. Banga, J. Boussard, G. A. Chapuis, et al. Spike sorting pipeline for the International Brain Laboratory. *Figshare*, 2022.
- [43] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.
- [44] A. Das, W. Kong, R. Sen, and Y. Zhou. A decoder-only foundation model for time-series forecasting. In *ICML*, 2024.
- [45] H. Banville, I. Albuquerque, A. Hyvärinen, G. Moffat, D.-A. Engemann, and A. Gramfort. Self-supervised representation learning from electroencephalography signals. *arXiv preprint arXiv:1911.05419*, 2019.
- [46] A. Jackson and T. M. Hall. Decoding local field potentials for neural interfaces. *IEEE TNSRE*, 25(10):1705–1714, 2016.
- [47] A. Schilling, R. Gerum, C. Boehm, et al. Deep learning based decoding of single local field potential events. *NeuroImage*, 297:120696, 2024.
- [48] A. T. Rao, K. L. Chou, and P. G. Patil. Localization of deep brain stimulation trajectories via automatic mapping of microelectrode recordings to MRI. *J. Neural Eng.*, 20(1):016056, 2023.

- 347 [49] D. Tebaykin, S. J. Tripathy, N. Binnion, B. Li, R. C. Gerkin, and P. Pavlidis. Modeling sources
348 of interlaboratory variability in electrophysiological properties of mammalian neurons. *J.*
349 *Neurophysiol.*, 119(4):1329–1339, 2018.
- 350 [50] H. G. Rey, C. Pedreira, and R. Q. Quiroga. Past, present and future of spike sorting techniques.
351 *Brain Res. Bull.*, 119:106–117, 2015.
- 352 [51] M. De Riva et al. Deep brain stimulation-induced local evoked potentials outperform spectral
353 features in spatial and clinical STN mapping. *medRxiv*, 2025.06.14.25329308, 2025.
- 354 [52] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica. Tune: A research
355 platform for distributed model selection and training. *arXiv:1807.05118*, 2018.
- 356 [53] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparam-
357 eter optimization framework. In *KDD*, pages 2623–2631, 2019.
- 358 [54] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, J. Ben-tzur, M. Hardt, B. Recht, and A. Tal-
359 walkar. A system for massively parallel hyperparameter tuning. In *MLSys*, 2020.

360 **Supplementary Material**

361 **A Dataset Details**

362 This appendix provides the full dataset construction, splitting strategy, and evaluation protocol for
363 each of the four benchmark corpora used in our study. Figure 6 summarizes sample counts, split
364 sizes, and fold structure across all four settings.

365 **A.1 Allen Brain Observatory (Neuropixels Visual Coding)**

366 The Allen Brain Observatory Visual Coding dataset [37] contains Neuropixels 1.0 recordings from 58
367 mice across approximately 51 experimental sessions, with up to six probes per session (384 channels
368 per probe, sampled at 1,250 Hz). After removing channels flagged as outside the brain by the Allen
369 Institute’s CCF registration pipeline, we retained 80,313 valid channels spanning 116 brain regions in
370 the Allen CCF hierarchy, each annotated with full 3D atlas coordinates (AP, DV, ML in μm).

371 **Chunking and pretraining scales.** Continuous recordings are segmented into non-overlapping
372 3-second windows (3,750 samples at 1,250 Hz). We construct five pretraining scales by subsampling
373 from the full training pool: 5K (4,999 chunks), 25K (21,709), 75K (60,192), 240K (197,181), and
374 1M (870,143). Scale selection is stratified by brain region to preserve class balance at each tier.

375 **Evaluation splits.** We use a strict cross-session split: 25 sessions for training, 14 for validation,
376 and 12 for testing, with no session overlap between partitions. Within the test set, we construct
377 5 stratified folds ($\approx 94\text{--}95\text{K}$ chunks per fold) to report mean and standard deviation of all metrics.
378 This session-level split eliminates any within-session temporal correlation from contaminating the
379 evaluation.

380 **A.2 IBL Brain Wide Map**

381 The International Brain Laboratory (IBL) Brain Wide Map dataset [40] aggregates Neuropixels
382 recordings from 12 laboratories worldwide, targeting standardized insertion trajectories through
383 frontal cortex, striatum, hippocampus, thalamus, midbrain, and other deep structures. This multi-
384 lab provenance introduces natural variability in surgical technique, probe quality, and recording
385 conditions, making it a stringent generalization benchmark.

386 **Pretraining scales.** We construct four scales from IBL data: 25K (24,750 chunks), 75K, 240K,
387 500K, and 1M, following the same 3-second windowing and region-stratified sampling as the Allen
388 corpus.

389 **Evaluation splits.** Splits are defined at the subject level to prevent cross-contamination: 76 subjects
390 for training, 29 for validation, and 34 for testing. This yields $5 \times 100\text{K}$ -chunk test folds (500K total
391 evaluation samples across 309 canonical fine-grained region IDs in the Allen CCF). The subject-level
392 split is more stringent than session-level, as it tests generalization across individual animals with
393 distinct anatomy and surgery.

394 **A.3 NeuroNexus (Cross-Probe Generalization)**

395 To evaluate whether learned representations transfer across probe technologies, we collected record-
396 ings using NeuroNexus multi-shank silicon probes in mice. These probes differ substantially from
397 Neuropixels in channel count, geometry, and impedance characteristics.

398 **Experimental protocol.** Eight recording sessions were conducted across multiple mice, targeting
399 five brain regions: Cortex, CA1, CA2, CA3, and DG (dentate gyrus). Not all regions are represented
400 in every session due to variation in probe insertion trajectory.

401 **Evaluation splits.** We use a rotating cross-session protocol: 5 sessions form a fixed training pool,
402 2 are used for validation, and 1 for testing, with 3 rotation folds. This yields $\approx 149\text{K}$ training,
403 $\approx 67\text{K}$ validation, and $\approx 33\text{--}73\text{K}$ test chunks per fold. The model is pretrained on Allen and/or IBL

Neuropixels data and evaluated on NeuroNexus recordings in a zero-shot or finetuned setting. This tests cross-probe-technology transfer: whether anatomical signatures in LFP are conserved despite differences in recording hardware.

A.4 Macaque (Cross-Species Generalization)

We evaluate cross-species transfer using Neuropixels-NHP recordings from rhesus macaques, targeting prefrontal and premotor cortex.

Dataset. 11 probes from two macaque subjects (subjects A and B) span four cortical areas: area 45, area 46, area 8, and area 9/46. Due to the limited number of subjects, we define a cross-subject split: subject A provides 7 probes for training (with 10% held out for validation), and subject B provides 4 probes for testing. This yields approximately 6.35M training chunks, 706K validation chunks, and 5.45M test chunks. The large chunk counts reflect the high channel density of NHP probes and the length of recording sessions.

Evaluation rationale. This is the most challenging transfer setting: the model must generalize across species (mouse $\rightarrow$ macaque), and cortical organization. It tests whether the LFP signatures of anatomical identity learned from mouse recordings carry over to primate cortex.

A.5 Terminology: Allen CCF Hierarchy and Depth Levels

The Allen Common Coordinate Framework organizes brain regions into a hierarchical tree with 10 depth levels (d0 = root, d9 = finest leaf). In our classification experiments, we evaluate at multiple depths to probe resolution: depth-0 assigns all channels to “brain” (trivial); depth-5 groups regions into ~ 7 coarse divisions (e.g., Isocortex, HPF, TH, HY, MB); depth-9 resolves individual cortical layers and fine nuclei (up to 309 classes on IBL). Balanced accuracy is reported at each depth to capture performance across the full anatomical hierarchy.

B Per-Region RMSE and Depth-5 Confusion Analysis

Figure 7 provides a fine-grained, region-by-region analysis of 3D localization error on the IBL evaluation set (430,759 samples across 5 folds, 309 canonical fine-grained region IDs). This analysis complements the aggregate metrics in the main paper by revealing which anatomical structures benefit most from each SSL objective and where residual error concentrates.

Per-region RMSE (Figure 7a). Each marker represents the median fold-averaged, left-right-symmetrized 3D RMSE for a single Allen fine region, ordered along the canonical Allen structure tree to expose anatomical groupings. Five model variants are compared in an ablation that adds losses sequentially on a Wav2Vec 2.0 backbone (75K pretraining chunks): ChanDistc only (gray, 2,216 μm overall), MaskPred only (light blue, 2,006 μm), CCFDistc only (orange, 2,094 μm), the joint frozen MaskPred + ChanDistc (dark blue, 1,799 μm), and the same model with end-to-end finetuning of the XYZ regression head (Full-FT, vermillion, 1,646 μm).

Several patterns emerge. Isocortical regions (leftmost block) show the lowest RMSE across all models, consistent with the distinctive laminar LFP signatures of cortical layers. Hippocampal formation (HPF) regions benefit disproportionately from geometry-aware objectives: the sharp wave-ripple and theta-gamma coupling patterns characteristic of CA1/CA3/DG [30, 34] provide strong anatomical fingerprints that are better captured when the embedding space is encouraged to mirror spatial layout. Thalamic and hypothalamic nuclei (TH, HY) show the highest residual error, likely reflecting the small spatial extent and electrophysiological similarity of neighboring nuclei in these deep structures. Notably, a handful of regions show elevated RMSE across all models (e.g., AON in olfactory areas, TTd in thalamus, NDB in basal forebrain): these tend to be small nuclei with few training samples (gray bars, right axis), where both representation quality and evaluation stability are limited.

Depth-5 confusion matrices (Figure 7b). Confusion matrices are computed at Allen depth-5 (7 coarse divisions: DORpm, DORsm, HPF, Isocortex, MY-mot, OLF, OTHER) on per-probe Kabsch-aligned Gaussian-mean predictions, restricted to classes with $\geq 2\%$ support. The progression

Datasets Overview — Pretraining + Linear-Probe / FT Eval Splits

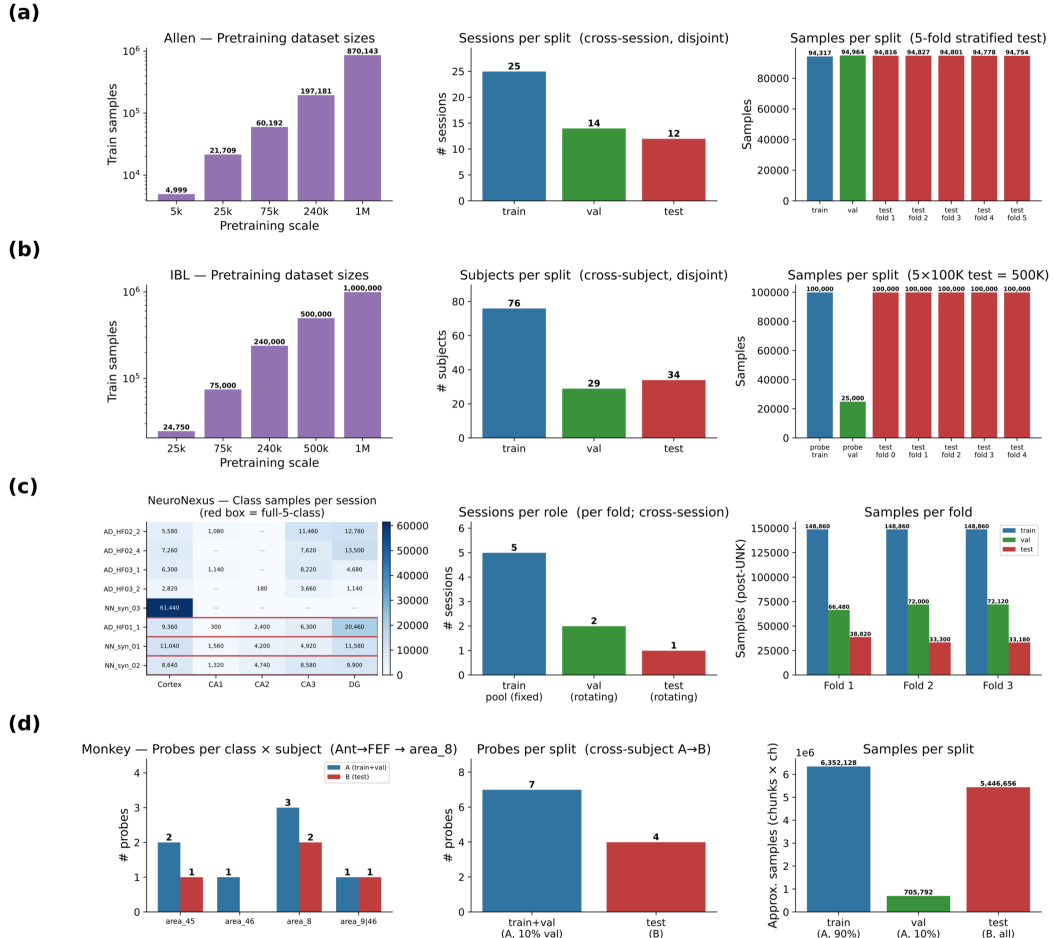


Figure 6: **Dataset overview across the four benchmark corpora.** Rows correspond to (a) Allen Brain Observatory, (b) IBL Brain Wide Map, (c) NeuroNexus cross-probe transfer, and (d) Macaque cross-species transfer. *Left column:* pretraining scales (number of 3-second LFP chunks) used for self-supervised training. *Middle column:* split sizes at the session, subject, or probe level, showing the number of units in each partition; all splits are strictly disjoint at the grouping level to prevent data leakage. *Right column:* resulting chunk counts per split and fold; Allen and IBL evaluation uses 5 stratified folds (≈95K and 100K chunks per fold, respectively), NeuroNexus uses 3 rotating folds, and macaque uses a single cross-subject split. Sample counts are reported post-filtering (out-of-brain channels removed, quality checks applied).

is striking: single-loss SSL variants (ChanDistc, MaskPred, CCFDistc) saturate near balanced accuracy 0.27–0.35, with heavy off-diagonal mass between DORpm, DORsm, and OTHER (thalamic and midbrain structures that are spatially interleaved along typical probe trajectories). Combining MaskPred with ChanDistc raises balanced accuracy to 0.324 (top-1 0.417), and end-to-end finetuning (Full-FT) reaches 0.418 balanced accuracy (top-1 0.508), a 50% relative gain over any single-loss variant. The Full-FT confusion matrix shows concentrated diagonal mass and substantially reduced off-diagonal leakage, particularly between cortical (Isocortex) and subcortical (DORpm, HPF) classes.

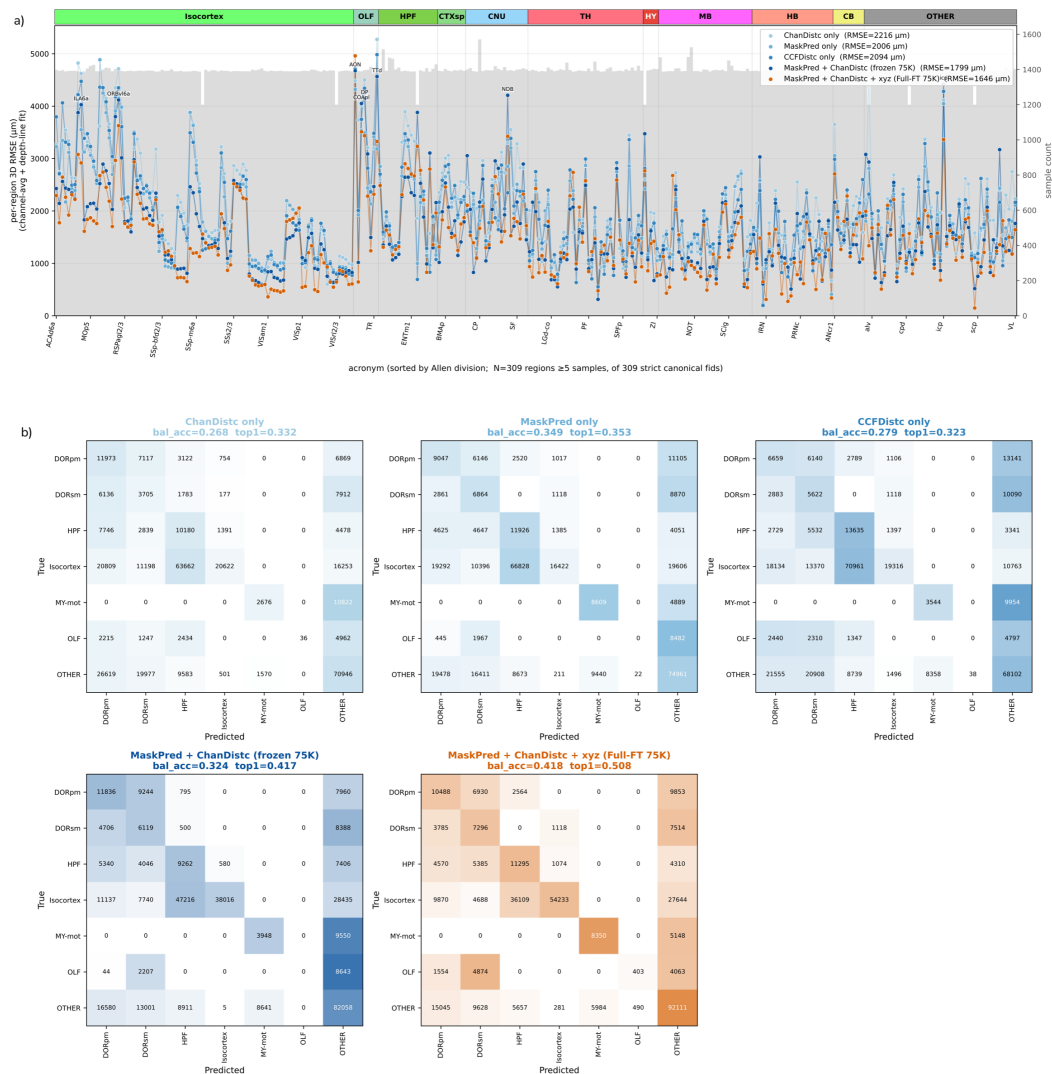


Figure 7: **Spatial losses progressively improve IBL localization, with end-to-end finetuning yielding the largest gain.** (a) Per-region 3D RMSE on IBL (100K per fold $\times$ 5 folds; 430,759 samples; 309 canonical fine-grained region IDs). Markers show the median fold-LR-symmetrized RMSE per Allen fine region for five SSL ablation stages on a Wav2Vec 2.0 backbone (75K chunks). Region acronyms are colored by Allen high-level division (top strip) and ordered along the canonical Allen tree. Background bars (right axis) indicate per-region sample count. (b) Depth-5 macro confusion matrices for the same five models. Single-loss SSL variants saturate near balanced accuracy 0.28; combining MaskPred with ChanDistc raises it to 0.324; Full-FT lifts it to 0.418 (top-1 0.508), a 50% relative gain, while concentrating mass on the diagonal across all seven coarse anatomy divisions.

458 C Extended Per-Probe Localization: Data2Vec Ablation

459 Figure 8 extends the per-probe case study from the main paper (Figure 3, which uses Wav2Vec 2.0)
 460 to the Data2Vec backbone, and compares against the supervised ViT and AnyNet baselines from
 461 Wang et al. [13]. Two matched probes are tracked across all model variants: a *best-case probe* (the
 462 insertion whose predicted trajectory most closely matches ground truth after Kabsch refinement) and
 463 a *worst-case probe* (the insertion with the highest residual 3D RMSE). Using the same two probes
 464 across all rows enables direct visual comparison of how each model handles both favorable and
 465 challenging trajectories.

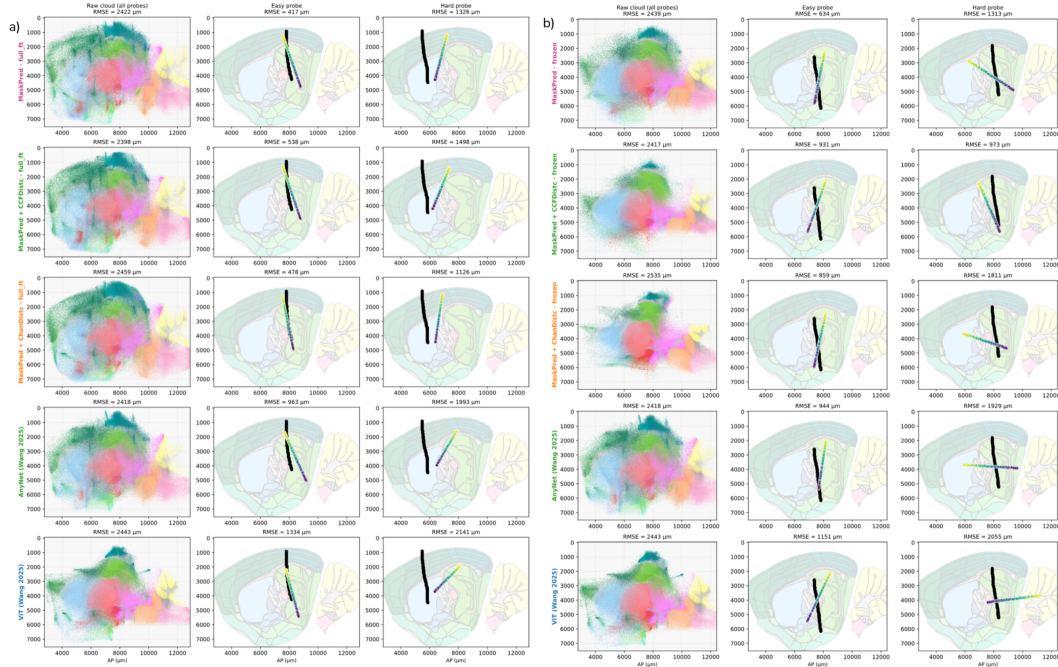


Figure 8: Per-probe localization on IBL: Data2Vec variants vs. Wang et al. (2025) ViT and AnyNet baselines. Predicted vs. ground-truth CCF coordinates for five models (rows: MaskPred, MaskPred + CCFDistc, MaskPred + ChanDistc, AnyNet, ViT) under (a) end-to-end finetuning and (b) frozen encoder with head-only regression. Each row shows the full prediction cloud (left, all probes) and two reference probes: the best-case probe (middle, lowest RMSE after Kabsch refinement) and the worst-case probe (right, highest RMSE). The same two probes are reused across all rows for direct comparison. Per-probe 3D RMSE (μm) is annotated on each panel. Geometry-aware SSL objectives (CCFDistc, ChanDistc) consistently reduce error on both probes relative to MaskPred alone, and all SSL variants outperform the supervised AnyNet and ViT baselines.

End-to-end finetuning (Figure 8a). Under full finetuning, MaskPred (top row) achieves $538 \mu\text{m}$ on the best-case probe and $1,498 \mu\text{m}$ on the worst-case probe. Adding CCFDistc tightens the best-case probe to $478 \mu\text{m}$ (11% improvement) and the worst-case to $1,126 \mu\text{m}$ (25% improvement). Adding ChanDistc instead yields $963 \mu\text{m}$ / $1,993 \mu\text{m}$, showing that the within-probe geometric prior helps more on typical trajectories than on difficult ones. The AnyNet baseline achieves $1,334$ / $2,141 \mu\text{m}$ and ViT achieves similar performance, both substantially worse than the SSL variants on both probes.

Frozen encoder (Figure 8b). When the encoder is frozen and only the regression head is trained, the gap between SSL objectives widens. MaskPred frozen yields 634 / $1,313 \mu\text{m}$; CCFDistc frozen yields 931 / $973 \mu\text{m}$ (the worst-case probe improves dramatically, suggesting that atlas-aware pretraining provides a stronger geometric scaffold when the encoder cannot adapt); ChanDistc frozen yields 944 / $1,929 \mu\text{m}$. Notably, AnyNet and ViT show minimal degradation between finetuned and frozen settings (AnyNet: $1,151$ / $2,055 \mu\text{m}$ frozen vs. $1,334$ / $2,141 \mu\text{m}$ finetuned), consistent with these supervised baselines having learned less transferable representations that plateau regardless of head capacity.

Probe difficulty terminology. Throughout this work, we refer to the insertion with the lowest post-refinement RMSE as the *best-case probe* and the insertion with the highest RMSE as the *worst-case probe*. Probe difficulty is determined by factors outside the model’s control: insertion angle relative to region boundaries, number of distinct brain structures traversed, proximity to tissue damage or vascular artifacts, and the electrophysiological distinctiveness of the traversed regions. The best-case/worst-case framing provides a natural dynamic range that reveals how robust each method is under both favorable and adversarial conditions.

D Model Architecture and Training Pipeline

D.1 Backbone Encoders

All three backbone encoders share the same signal processing front-end and differ only in the Transformer core. Table 1 summarizes the key architectural parameters.

Table 1: **Backbone architectures.** All backbones use pretrained audio foundation model weights and share the same convolutional feature extractor design (7 layers, GELU activations, group normalization). Parameter counts include the feature extractor, Transformer encoder, and quantizer/projector modules.

Backbone	Pretrained ID	Layers	Heads	d_{model}	d_{ff}	Params
Wav2Vec 2.0	facebook/wav2vec2-base	12	12	768	3072	94.4 M
Data2Vec	facebook/data2vec-audio-base	12	12	768	3072	94.4 M
Whisper	openai/whisper-base (encoder)	6	8	512	2048	72.6 M

Convolutional feature extractor. The raw waveform passes through a 7-layer 1D convolutional stack that progressively downsamples the input. All layers have 512 output channels. The kernel sizes are (10, 3, 3, 3, 3, 2, 2) and the strides are (5, 2, 2, 2, 2, 2, 2), yielding a total downsampling factor of $\prod_i s_i = 320$. Group normalization and GELU activations are applied after each convolution. A convolutional relative positional encoding layer (kernel 128, 16 groups) is added before the Transformer blocks. Biases are disabled in the convolutional layers.

Signal preprocessing. Each 3-second LFP chunk (3,750 samples at 1,250 Hz) is resampled on-GPU to 16,000 Hz using `torchaudio.functional.resample`, producing a 48,000-sample waveform compatible with the pretrained audio models. After the 7-layer CNN with total stride 320, this yields a 150-token sequence. Per-sample z-score normalization was evaluated but disabled in the final pipeline, as pretrained audio backbones expect unnormalized waveforms.

D.2 Downstream Regression Head

The pooled 768-dimensional embedding (mean over the token sequence) passes through a shared MLP trunk:

$$\mathbf{h}_0 = \text{Pool}(\mathbf{z}) \in \mathbb{R}^{768} \rightarrow \mathbf{h}_1 \in \mathbb{R}^{512} \rightarrow \mathbf{h}_2 \in \mathbb{R}^{256} \rightarrow \mathbf{h}_3 \in \mathbb{R}^{128},$$

with ReLU activations between each linear layer. The trunk branches into two heads:

- **Mean head** (μ): $\mathbf{h}_3 \rightarrow \mathbb{R}^3$, predicting the AP, DV, ML coordinates in μm .
- **Log-variance head** ($\log \sigma^2$): $\mathbf{h}_3 \rightarrow \mathbb{R}^3$, predicting per-axis heteroscedastic uncertainty. The regression loss is the heteroscedastic Gaussian negative log-likelihood: $\mathcal{L}_{\text{reg}} = \sum_{k \in \{AP, DV, ML\}} \frac{(y_k - \mu_k)^2}{2\sigma_k^2} + \frac{1}{2} \log \sigma_k^2$, which jointly learns both the prediction and its calibrated uncertainty.

For classification experiments (e.g., depth-5 balanced accuracy), the regression head is replaced with a linear projector ($768 \rightarrow 256$) followed by a classifier ($256 \rightarrow C$, where C is the number of classes at the target depth level).

D.3 Kabsch Refinement

Per-probe predictions are post-processed via Kabsch alignment [22] to correct for systematic offsets and rotations. Given predicted coordinates $\hat{\mathbf{X}} \in \mathbb{R}^{N \times 3}$ and ground-truth coordinates $\mathbf{X} \in \mathbb{R}^{N \times 3}$ for all N channels on a single probe:

1. Center both point clouds: $\bar{\mathbf{X}} = \mathbf{X} - \mu_{\mathbf{X}}$, $\bar{\hat{\mathbf{X}}} = \hat{\mathbf{X}} - \mu_{\hat{\mathbf{X}}}$.
2. Compute the cross-covariance matrix: $\mathbf{H} = \bar{\hat{\mathbf{X}}}^\top \bar{\mathbf{X}}$.

- 520 3. Perform SVD: $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$.
- 521 4. Compute the optimal rotation: $\mathbf{R} = \mathbf{V} \text{diag}(1, 1, \det(\mathbf{V}\mathbf{U}^\top)) \mathbf{U}^\top$.
- 522 5. Apply: $\hat{\mathbf{X}}_{\text{refined}} = (\hat{\mathbf{X}} - \boldsymbol{\mu}_{\hat{\mathbf{X}}})\mathbf{R}^\top + \boldsymbol{\mu}_X$.

523 This corrects for global translation and rotation errors that are consistent across all channels on a
 524 single probe insertion, isolating the model’s ability to resolve *relative* channel positions along the
 525 probe trajectory.

526 E Hyperparameter Tuning Pipeline

527 Hyperparameters are tuned independently for each backbone using Ray Tune [52] with the Op-
 528 tuna [53] Tree-structured Parzen Estimator (TPE) as the search algorithm and ASHA (Asynchronous
 529 Successive Halving) [54] as the trial scheduler. ASHA terminates poorly performing trials early (grace
 530 period: 1 epoch), enabling efficient exploration of large search spaces. The optimization metric is
 531 `loss_per_mask`, defined as validation loss normalized by `mask_time_prob` $\times$ `mask_time_length`,
 532 which prevents the optimizer from gaming the masking rate.

Table 2: IBL SSL hyperparameters: 75K cross-negative SSL + ChanDistc joint, $\sigma=500\ \mu\text{m}$. Wav2Vec2-base initialized from facebook/wav2vec2-base and fully unfrozen (12/12 layers); jointly optimized with masked InfoNCE prediction ($\lambda_{\text{ssl}}=1.0$) and the channel-distance contrastive loss ($\lambda_{\text{dist}}=1.0$, $\sigma=500\ \mu\text{m}$, $\tau=0.2$) over Neuropixels 1.0 4-column shank coordinates. Quantizer uses 2 groups $\times$ 320 codevectors with diversity weight 0.1 (HuggingFace wav2vec2-base default). Training: 200 epochs target, batch 64 (4 probes $\times$ 16 samples), AdamW (lr= 1.02×10^{-4} , wd= 0.08), 400-step warmup, plateau LR schedule with floor $\eta_{\text{min}}=10^{-5}$, gradient clipping 1.0, AMP off.

		<i>Optimizer</i>	
<i>Backbone</i>		lr_ssl	1.02×10^{-4}
		weight_decay	0.080
		hidden_dropout	0.097
		attention_dropout	0.042
		final_dropout	0.0
		warmup_steps	400
		warmup_start_factor	0.01
		grad_clip_norm	1.0
		lr_schedule	plateau
		plateau_factor	0.5
		plateau_patience	5
		eta_min	1.0×10^{-5}
<i>Quantizer</i>		<i>ChanDistc</i>	
codevectors / group	320	dist_temperature	0.2
codevector groups	2	dist_sigma (μm)	500.0
codebook size	640	dist_target_fraction	1.0
diversity_loss_weight	0.1	samples_per_probe	16
gumbel_temp_init	2.0	num_probes	4
gumbel_temp_min	0.5	coord_dim	2
gumbel_temp_decay	0.999995	queue_size	0
<i>Loss weights</i>		<i>Training</i>	
ssl_lambda (MaskPred)	1.0	batch_size_ssl	64
dist_lambda (ChanDistc)	1.0	ssl_epochs	200
<i>Masking</i>		amp	false
mask_time_prob	0.079	atlas_depth	9
mask_time_length	5	num_workers	8
num_negatives	100	prefetch_factor	16
cross_sample_negatives	true	embed_interval	5
min_masks	2	probe_interval	5
		probe_epochs	20
		early_stop_patience	30

Table 3: SSL hyperparameters: Allen 75K (MaskPred + ChanDistc joint). Wav2Vec2-base initialized from facebook/wav2vec2-base and fully unfrozen (12/12 layers); jointly optimized with masked InfoNCE prediction ($\lambda_{\text{ssl}} = 1.0$) and the channel-distance contrastive loss ($\lambda_{\text{dist}} = 7.74$, $\sigma = 71.9 \mu\text{m}$, $\tau = 0.037$) over 2D shank coordinates. Quantizer uses 2 groups $\times$ 64 codevectors with diversity weight 1.0 and context normalization. Hyperparameters selected by HEBO (80 trials; metric = spatial correlation between embedding and channel geometry). Training: 200 epochs, batch 64, AdamW (lr= 3.15×10^{-5} , wd= 0.024), 100-step warmup, gradient clipping 1.0, AMP off.

		<i>Optimizer</i>	
<i>Backbone</i>		lr_ssl	3.15×10^{-5}
backbone	wav2vec2	weight_decay	0.0242
pretrained_model	wav2vec2-base	hidden_dropout	0.0582
rand_init	false	attention_dropout	0.0979
hidden_size	768	warmup_steps	100
num_hidden_layers	12	warmup_start_factor	0.0154
num_attention_heads	12	grad_clip_norm	1.0
intermediate_size	3072	<i>ChanDistc</i>	
unfreeze_top_k_layers	12 (full)	dist_temperature	0.0374
<i>Quantizer</i>		dist_sigma (μm)	71.92
codevectors / group	64	probe_batch	true
codevector groups	2	coord_dim	2
codebook size	128	queue_size	0
diversity_loss_weight	1.0	<i>Training</i>	
normalize_context	true	batch_size_ssl	64
<i>Loss weights</i>		ssl_epochs	200
ssl_lambda (MaskPred)	1.0	amp	false
dist_lambda (ChanDistc)	7.736	atlas_depth	9
<i>Masking</i>		num_workers	4
mask_time_prob	0.1434	prefetch_factor	10
mask_time_length	9	embed_interval	2
num_negatives	100	metrics_interval	1
		max_embed_samples	10000
		probe_epochs	5
		probe_interval	10
		max_probe_samples	5000

533 F Compute Resources

534 **Compute resources.** All experiments were run on a HPC cluster. Across the primary account,
535 between January and May 2026, the project consumed approximately 12,150 GPU-hours: $\sim 10,000$ h
536 on NVIDIA L40S (48 GB), $\sim 1,500$ h on A100 (80 GB), ~ 520 h on H100 (80 GB), and ~ 320 h on
537 H200 (141 GB). Of this total, $\sim 3,250$ GPU-hours were spent on completed final runs, ~ 990 h on
538 HEBO hyperparameter search, and the remainder on preliminary, failed, or auto-cancelled jobs (the
539 cluster auto-cancels jobs that average $< 50\%$ GPU utilization for 2 h). Including co-author accounts
540 running portions of the IBL and rival-baseline pipelines, the full project consumed roughly 15,000
541 GPU-hours of compute.

Table 4: **IBL probe-channel XYZ regression at the 75k SSL scale.** All numbers are computed on the same `eval_ibl_v2_compat` test set (500,000 samples, 5 folds $\times$ 100k each, 119 probes). Mean Euclidean error, median, R^2 , Spearman ρ and Gaussian coverage are reported as mean $\pm$ std across the 5 folds. *Kabsch fold-LR* applies per-probe rigid alignment to the predictions and folds the medio-lateral axis at the Allen midline; for the neural backbones (data2vec, wav2vec2, whisper-small); baselines without per-sample CCF region IDs use the all-region Kabsch fold-LR. Lower is better for all error columns. PT = HF audio pretrained init; RI = random init. All supervised baselines (ViT, AnyNet, HistGBM, kNN, Ridge, PCA+Ridge, RF, SVM) are trained from scratch on the IBL XYZ regression target

Backbone	Pretraining	Eval	Init	Mean (μ m)	Med. (μ m)	R^2	ρ	Kabsch fold-LR (μ m)
data2vec	(none, HF audio init)	frozen	PT	2873 $\pm$ 6	2820	-0.01	0.19	2827 $\pm$ 4
	MaskPred	frozen	PT	2295 $\pm$ 6	2063	0.25	0.53	1888 $\pm$ 12
		fullft	PT	2221 $\pm$ 6	1927	0.25	0.56	1725 $\pm$ 13
	MaskPred + ChanDistc ($\lambda=0.1$)	frozen	PT	2448 $\pm$ 4	2279	0.20	0.46	2174 $\pm$ 6
		fullft	PT	2267 $\pm$ 6	1979	0.22	0.54	1740 $\pm$ 8
	(none, no SSL)	frozen	RI	2809 $\pm$ 5	2609	-0.04	0.20	2670 $\pm$ 4
	MaskPred	frozen	RI	2441 $\pm$ 5	2261	0.21	0.47	2122 $\pm$ 10
	MaskPred + ChanDistc ($\lambda=1.0$)	frozen	RI	2435 $\pm$ 4	2246	0.20	0.46	2166 $\pm$ 9
	MaskPred + ChanDistc ($\lambda=1.0$)	fullft	RI	2192 $\pm$ 5	1889	0.27	0.57	1697 $\pm$ 11
wav2vec2	(none, HF audio init)	frozen	PT	2602 $\pm$ 7	2458	0.11	0.39	2436 $\pm$ 11
	MaskPred	frozen	PT	2264 $\pm$ 4	2084	0.29	0.54	2013 $\pm$ 8
	ChanDistc only	frozen	PT	2455 $\pm$ 4	2298	0.19	0.45	2214 $\pm$ 7
	CCFDistc only	frozen	PT	2302 $\pm$ 4	2115	0.26	0.51	2093 $\pm$ 7
	MaskPred + ChanDistc $\sigma=300$	frozen	PT	2233 $\pm$ 4	1982	0.29	0.54	1947 $\pm$ 7
	MaskPred + ChanDistc $\sigma=500$	frozen	PT	2117 $\pm$ 5	1889	0.35	0.58	1803 $\pm$ 8
	MaskPred + ChanDistc $\sigma=500$	fullft	PT	2084 $\pm$ 5	1816	0.34	0.60	1649 $\pm$ 10
	(none, no SSL)	frozen	RI	2311 $\pm$ 3	2119	0.27	0.52	2066 $\pm$ 8
	supervised XYZ only	fullft	RI	2335 $\pm$ 7	1950	0.17	0.53	1777 $\pm$ 7
whisper-small	MaskPred	frozen	PT	2420 $\pm$ 3	2272	0.22	0.48	2205 $\pm$ 3
	MaskPred + ChanDistc $\sigma=500$	frozen	PT	2320 $\pm$ 4	2108	0.26	0.51	2104 $\pm$ 3
	MaskPred	fullft	PT	2366 $\pm$ 5	2046	0.16	0.47	1975 $\pm$ 8
	MaskPred + ChanDistc $\sigma=500$	fullft	PT	2404 $\pm$ 4	2173	0.18	0.48	2078 $\pm$ 4
ViT	supervised XYZ	fullft	from-scratch	1867 $\pm$ 6	1671	0.50	0.70	1900 $\pm$ 5
AnyNet		fullft	from-scratch	1764 $\pm$ 4	1550	0.53	0.72	1829 $\pm$ 4
HistGBM		band feats	from-scratch	2161 $\pm$ 7	2055	0.37	0.60	2190 $\pm$ 6
kNN-5		band feats	from-scratch	2178 $\pm$ 8	2075	0.36	0.60	2210 $\pm$ 7
		PCA(50) PSD	from-scratch	2106 $\pm$ 7	1913	0.38	0.61	2170 $\pm$ 7
Ridge		band feats	from-scratch	2305 $\pm$ 3	2169	0.31	0.56	2278 $\pm$ 1
		PSD feats	from-scratch	2217 $\pm$ 2	2068	0.35	0.60	2208 $\pm$ 1
PCA+Ridge		PSD feats	from-scratch	2269 $\pm$ 2	2149	0.33	0.57	2267 $\pm$ 1
RF		band feats	from-scratch	2181 $\pm$ 7	2083	0.36	0.59	2216 $\pm$ 6
SVM		band feats	from-scratch	2341 $\pm$ 3	2152	0.27	0.53	2341 $\pm$ 1

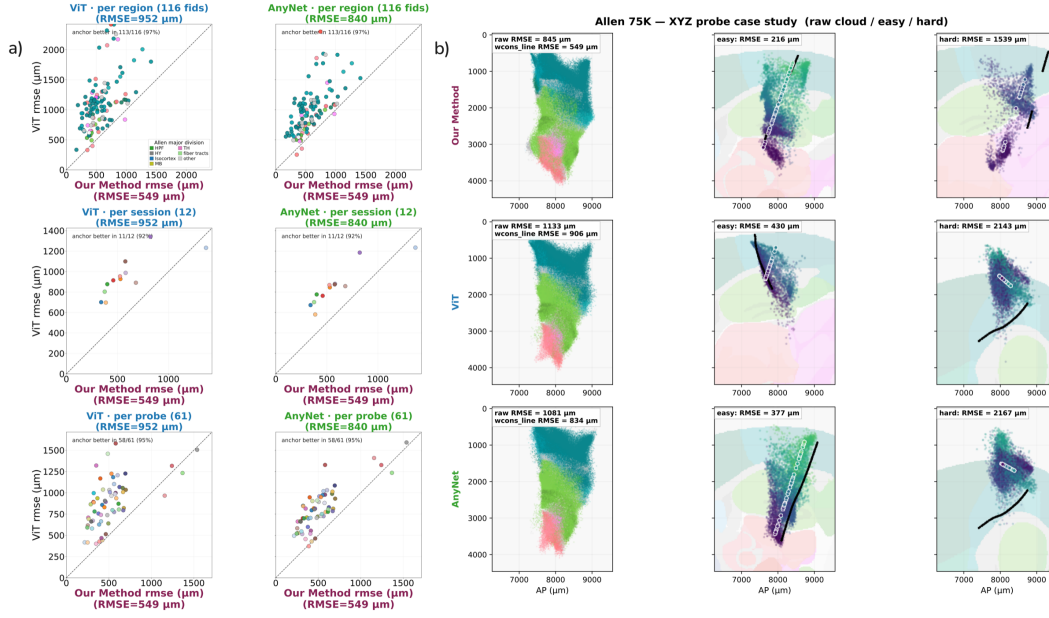


Figure 9: Allen Brain Observatory localization comparison (75K scale). (a) Per-group 3D RMSE: our method (MaskPred + ChanDistc, full-FT, 75K) vs. ViT and AnyNet rivals, all trained on Allen and evaluated on the held-out test split. Each point is one group’s RMSE; rows aggregate per region (top, colored by Allen Brain Atlas), per session (middle), and per probe (bottom). The dashed identity line marks parity; points above the diagonal indicate our method outperforms the rival on that group. Annotations report the fraction of groups on which our method is better. Headline: ours 549 μm vs. AnyNet 840 μm vs. ViT 952 μm . The same posthoc (weighted-consensus + per-probe line projection) is applied uniformly to all three models. (b) Per-model probe case study: predictions on the Allen test set (raw cloud / easy / hard probe; refinement on a sagittal slab). Our method achieves 549 μm 3D RMSE, vs. 834 μm AnyNet and 906 μm ViT.

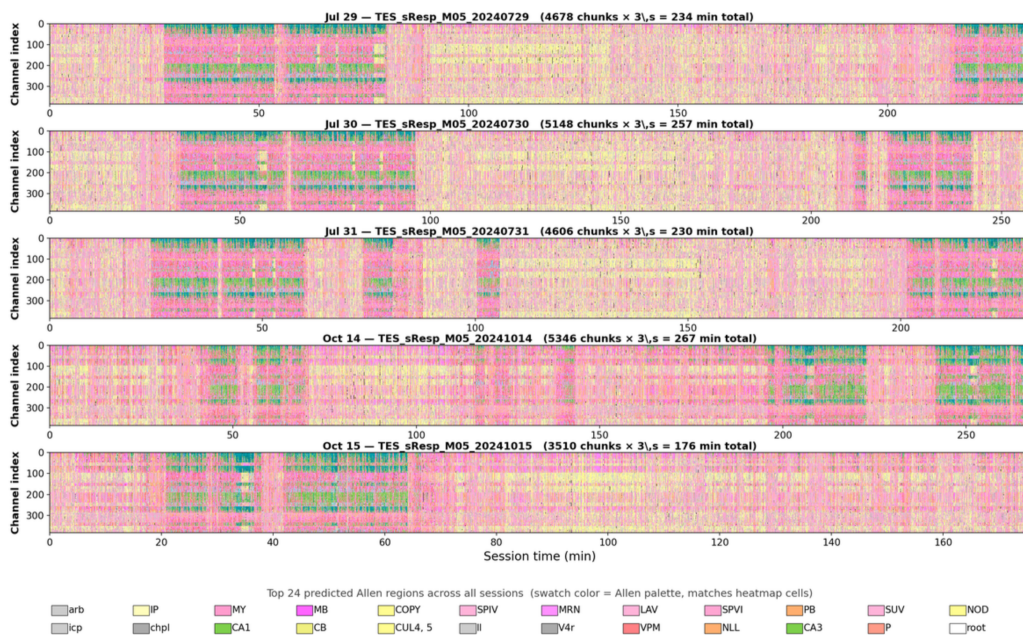


Figure 10: **Zero-shot chronic stability of Wav2Vec2 on Neuropixels 2.0 recordings.** Each row shows one full session ($\sim 3\text{--}4.5$ h) from mouse M05, recorded by an external collaborator on a probe and rig that the model has never seen during training. The three July sessions (Jul 29–31, 2024) and two October sessions (Oct 14–15, 2024) span ~ 2.5 months on the same chronically implanted probe. Each panel is a $384 \times N_{\text{chunks}}$ heatmap of per-channel, per-3 s-chunk predictions: every cell is colored by the Allen CCF region returned by voxel-lookup of the predicted xyz coordinate (color matches the Allen palette legend below). Channel index runs surface (top, ch 0) $\rightarrow$ tip (bottom, ch 383). Predictions are produced **zero-shot**: the encoder and Gaussian xyz head trained on IBL data are applied directly to the chronic NP 2.0 stream after destriping and resampling ($2500 \rightarrow 1250$ Hz). No histology-aligned ground truth is available for these recordings, so we report only the model’s own predictions; horizontal banding within a session reflects stable per-channel region assignments across hours, while differences between rows reflect day- and month-scale drift of those assignments. The day-over-day consistency within each week (intra-July, intra-October) and the visible Jul $\rightarrow$ Oct shift demonstrate that the model produces a reproducible, drift-trackable region readout on out-of-distribution hardware.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction clearly highlight and convey the contributions and scope of this paper: LFP data contains enough anatomical information to localize and decode to 3D anatomical CCF coordinates, which translates across probe technologies, labs, species through geometry aware self supervised learning objectives, all of which are substantiated by the results of our extensive experiments with scaling, backbones, training methodologies and evaluations.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We include and discuss limitations and future scope in Section 5, emphasizing data imbalance for rare classes, quality of histological registration to the CCF atlas and further scope of extension to larger cross species datasets with extended training and larger models.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present new theoretical results; it rather focuses on leveraging probe geometry for self supervised learning objectives, thus providing novel representation learning objectives.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all necessary steps and methodologies for training publicly available model architectures, pre-processing techniques, and post training procedures, including our novel representation learning contributions through detailed figures and explanations as detailed in Section 3 and Section 4 as well as the Supplementary material (A-I), ensuring full reproducibility of our results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We commit to releasing all code and pretrained weights of our models, and instructions for accessing public datasets and model backbones, along with our representation learning objectives and post hoc refinement scripts, with well documented code in the code repository in Section 7

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 2 describes data splits (12 held-out sessions, 5-fold evaluation, 94,317 test samples). Section 3 specifies backbone architectures, MLP dimensions, loss functions, and three evaluation regimes (frozen, CNN-frozen, full finetuning). Margin and distance threshold hyperparameters (σ , m) are defined in the loss formulations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our figures, plots and results include errors bars with standard deviation for all quantitative and qualitative comparisons of scaling, backbones, representation learning objectives and evaluation methods, computed with consistent train/test/validation splits with multiple folds, along with explanation and interpretation of the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have added information about our compute, GPU types, GPU hours in the supplementary in Section F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: All animal experiments were conducted under institutional IACUC or equivalent approval at collaborating labs, and we adhere to community guidelines on data handling and privacy for neural recordings.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Section 6 discusses positive impacts (reducing histology burden, enabling real-time DBS feedback) and negative impacts (risk of incorrect localization in clinical settings), along with mitigation via calibrated uncertainty estimates.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The model localizes electrodes in mouse/primate brains and poses minimal misuse risk. No high-risk generative capabilities are involved.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use the Allen Brain Observatory and IBL Brain Wide Map datasets, along with the publicly available pretrained audio based transformer backbones including Wav2Vec2, Data2Vec, Whisper, and baselines ViT, AnyNet; all cited with their original publications.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We commit to release all pretrained model checkpoints, novel representation learning methodologies and post hoc refinement, along with their code and documentation in our code repo in section 7.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No human or crowdsourced experiments are involved; our study majorly involves animal electrophysiology recordings from mice and monkeys.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human trials and subjects were involved; all animal electrophysiology recordings and experiments were conducted under appropriate institutional animal care and approvals.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of our study and methodology; we only use self-supervised audio based transformer backbones for our experimentation and results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.