Self-supervised Feature Learning Using Human Neural Data at Scale (a.k.a. “brain foundation models”)

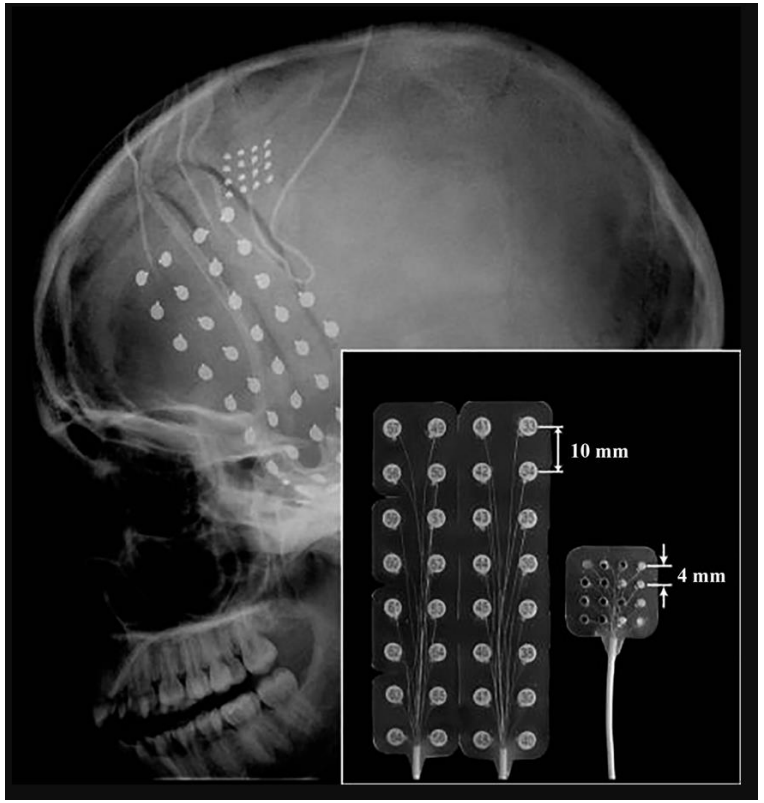
Andrii Zahorodnii
Fiete Lab, MIT Brain and Cognitive Sciences
3/2025

1. Motivation and Problem Statement

Andrii Zahorodnii
Fiete Lab, MIT Brain and Cognitive Sciences
3/2025

Background: Human intracranial EEG (ECoG/SEEG)

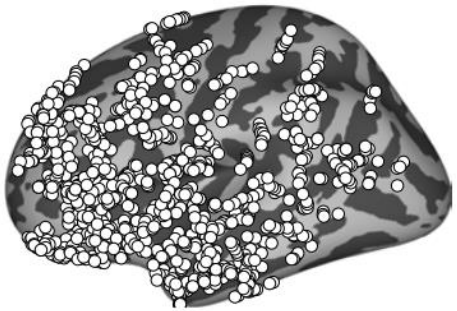
A diagnostic procedure performed on individuals with drug-resistant epilepsy to precisely localize seizure origins, when non-invasive methods are inconclusive. Assesses whether patient is a candidate for neurosurgical epilepsy treatment.



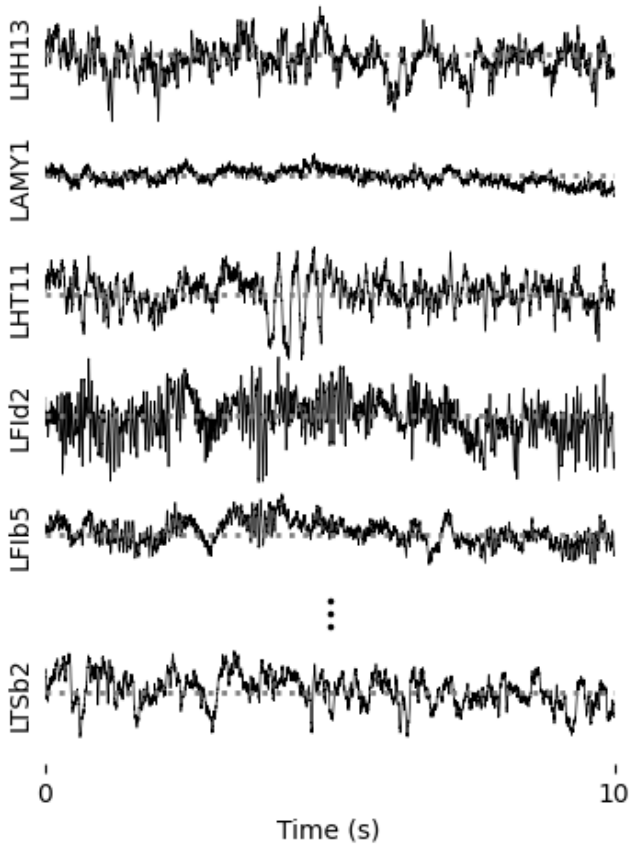
2025: Data from ~300-500 human subjects, ~10TB total, available online / stored in research labs.
~an order of magnitude more being thrown away (hundreds of patients/year in the U.S)

Challenge: decrypting the neural code

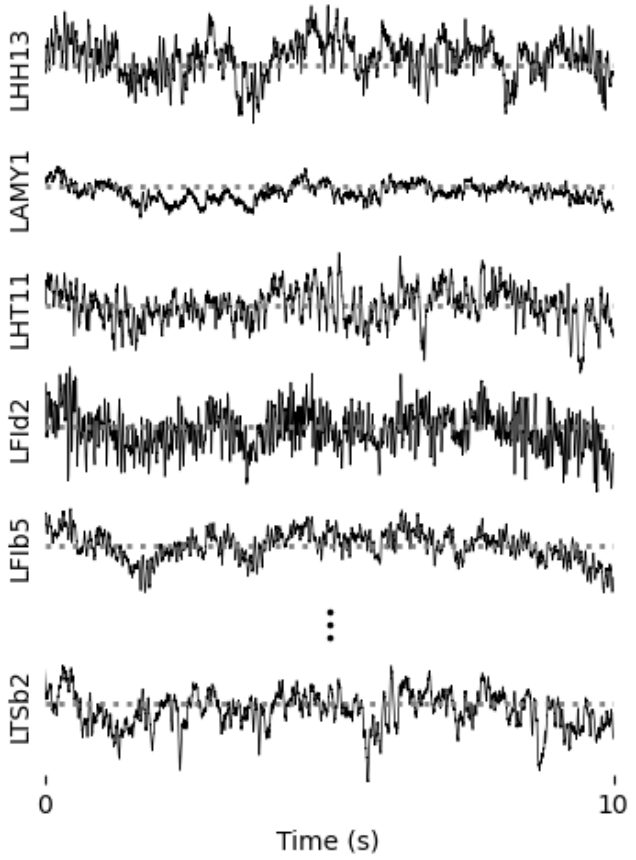
(Assuming that given enough coverage and sampling, all of the information is present in the neural recordings)



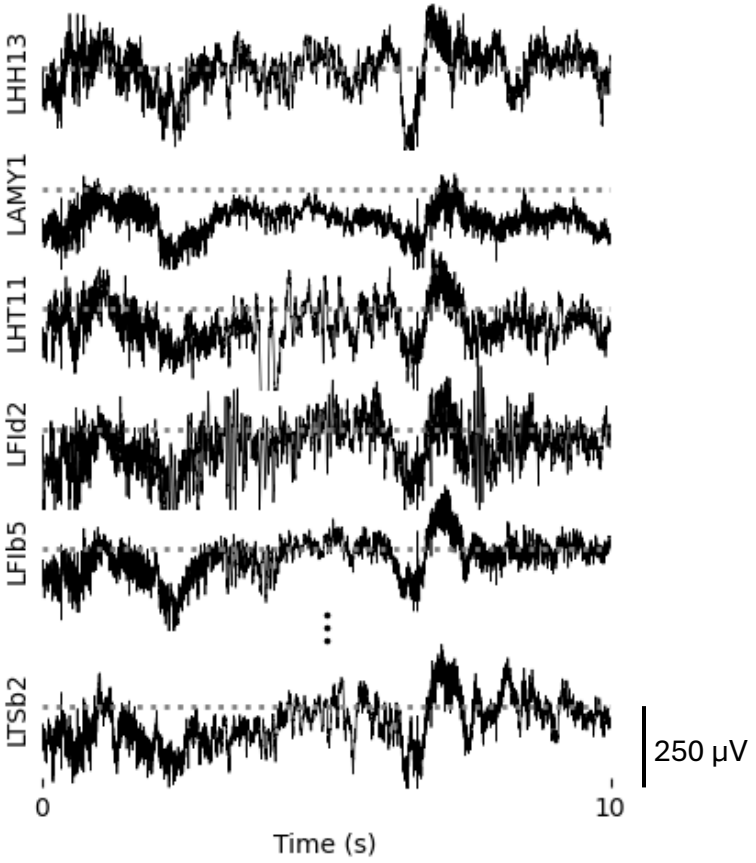
remembering a story



feeling inspired

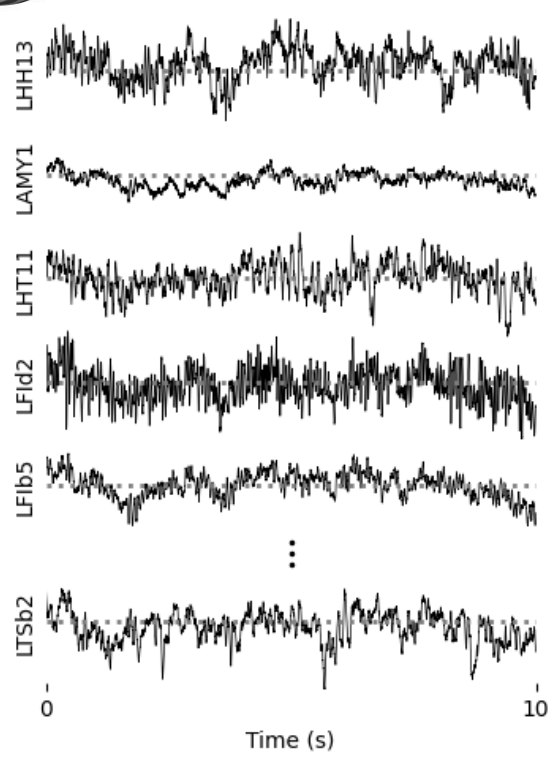


seizure



Goal: Create a useful feature extractor

Learn a transformation from raw measurement space to a latent space where features of interest are readily (linearly) available for extraction



model →

$x(t)$

$\vec{\lambda}_1 \cdot x(t) =$ “imminence of seizure”

$\vec{\lambda}_2 \cdot x(t) =$ “intensity of felt emotion”

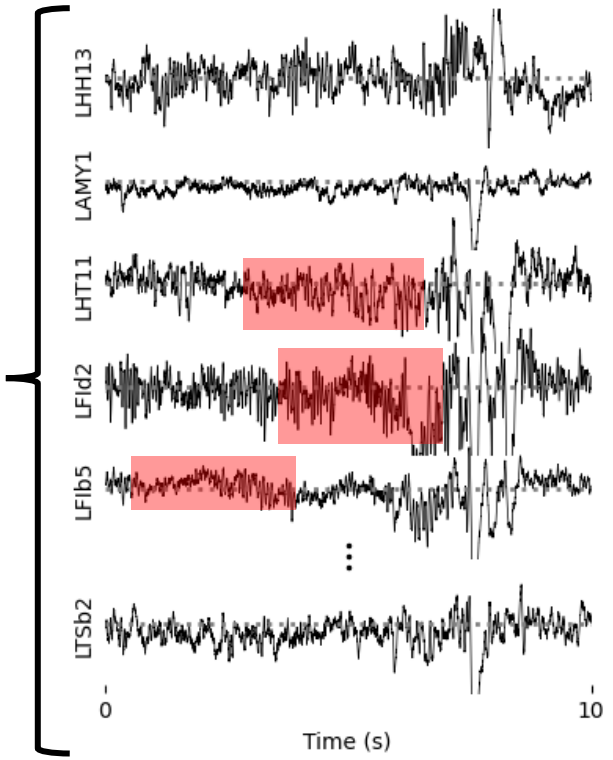
$\Lambda x(t) =$ “content of inner vocalization”

Goal: Create a useful feature extractor

if $\vec{\lambda}_1 \cdot x(t) =$ "imminence of seizure", then

$$\frac{\partial\left(\vec{\lambda}_1 \cdot x(t)\right)}{\partial(\text{model input})}$$

→ model (inverted)



high derivative score

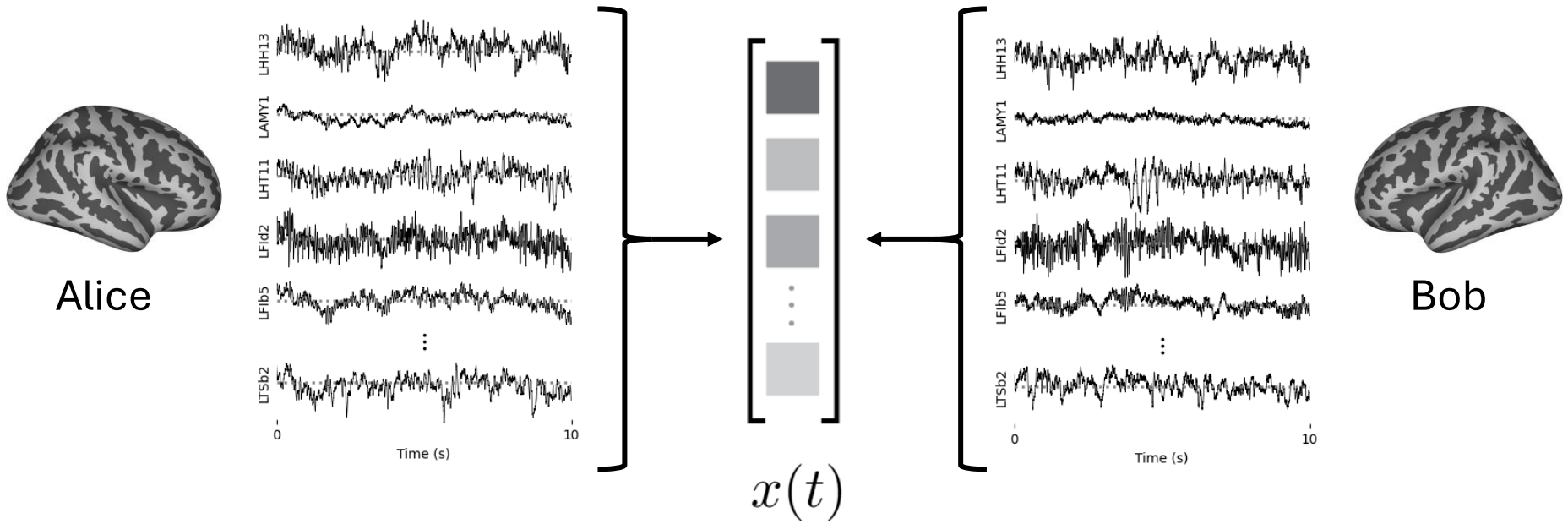
→ targeted intervention
(i.e. electrode stimulation to prevent seizure)

Goal*: Create an *invariant* feature extractor

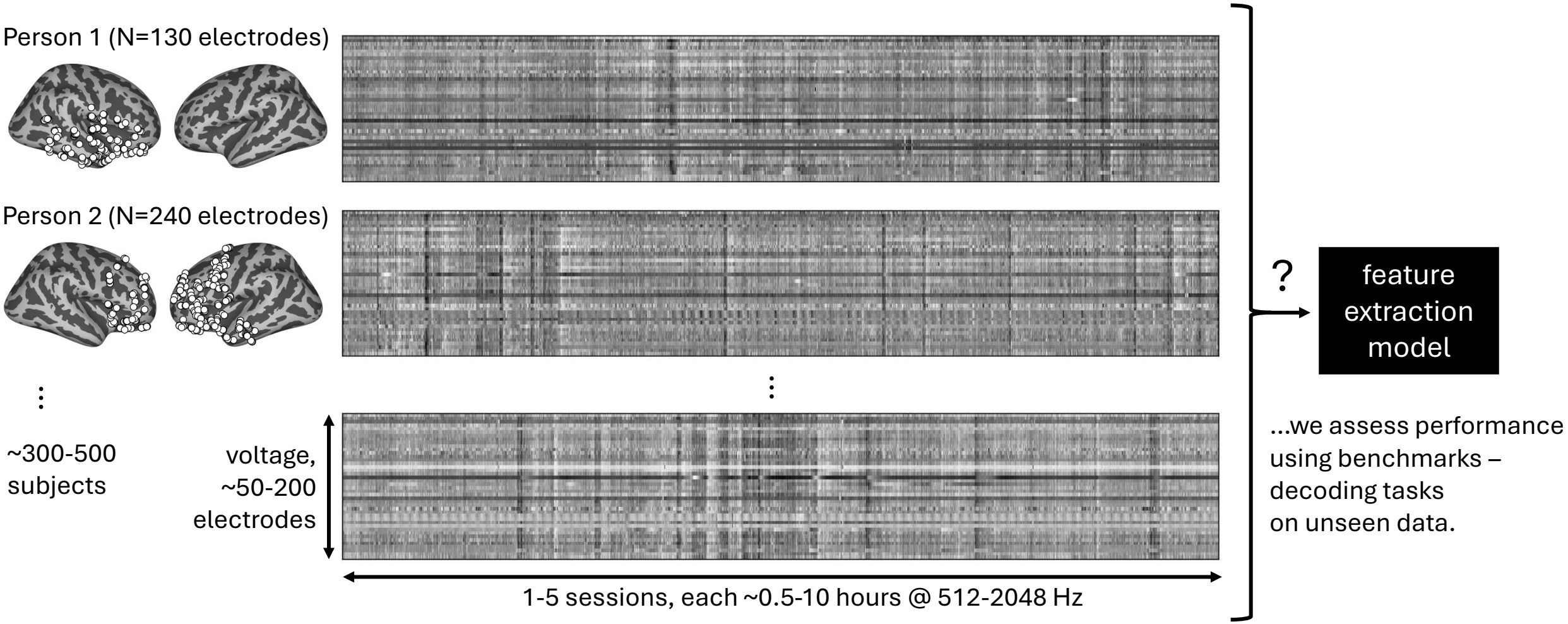
Learn a transformation from raw measurements into a latent space that can be readily aligned across individuals, across electrode placements in the brain, and/or across time within one individual

if $\Lambda x(t) =$ interesting subspace, then hopefully

$$\Lambda^{(Alice)} \approx \Lambda^{(Bob)}, \text{ or } \Lambda^{(Alice)} \approx R\Lambda^{(Bob)}$$



Problem statement:
Representation learning from human iEEG @ scale

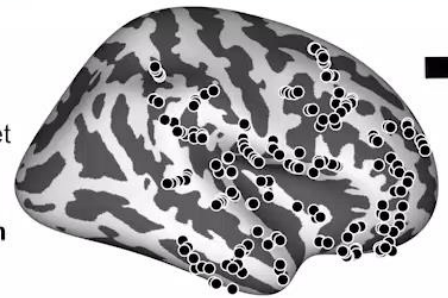


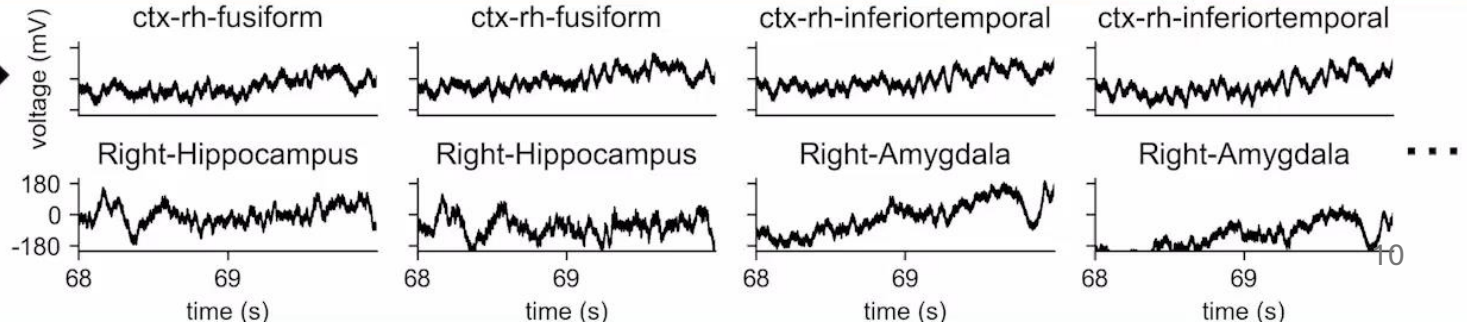
2. Benchmarking: Measuring Progress

Andrii Zahorodnii
Fiete Lab, MIT Brain and Cognitive Sciences
3/2025

Background: BrainTreebank (iEEG+Movies dataset)

BrainTreebank Dataset
Wang et al. (2024)
10 subjects
100+ electrodes each
~4.3hr / subject

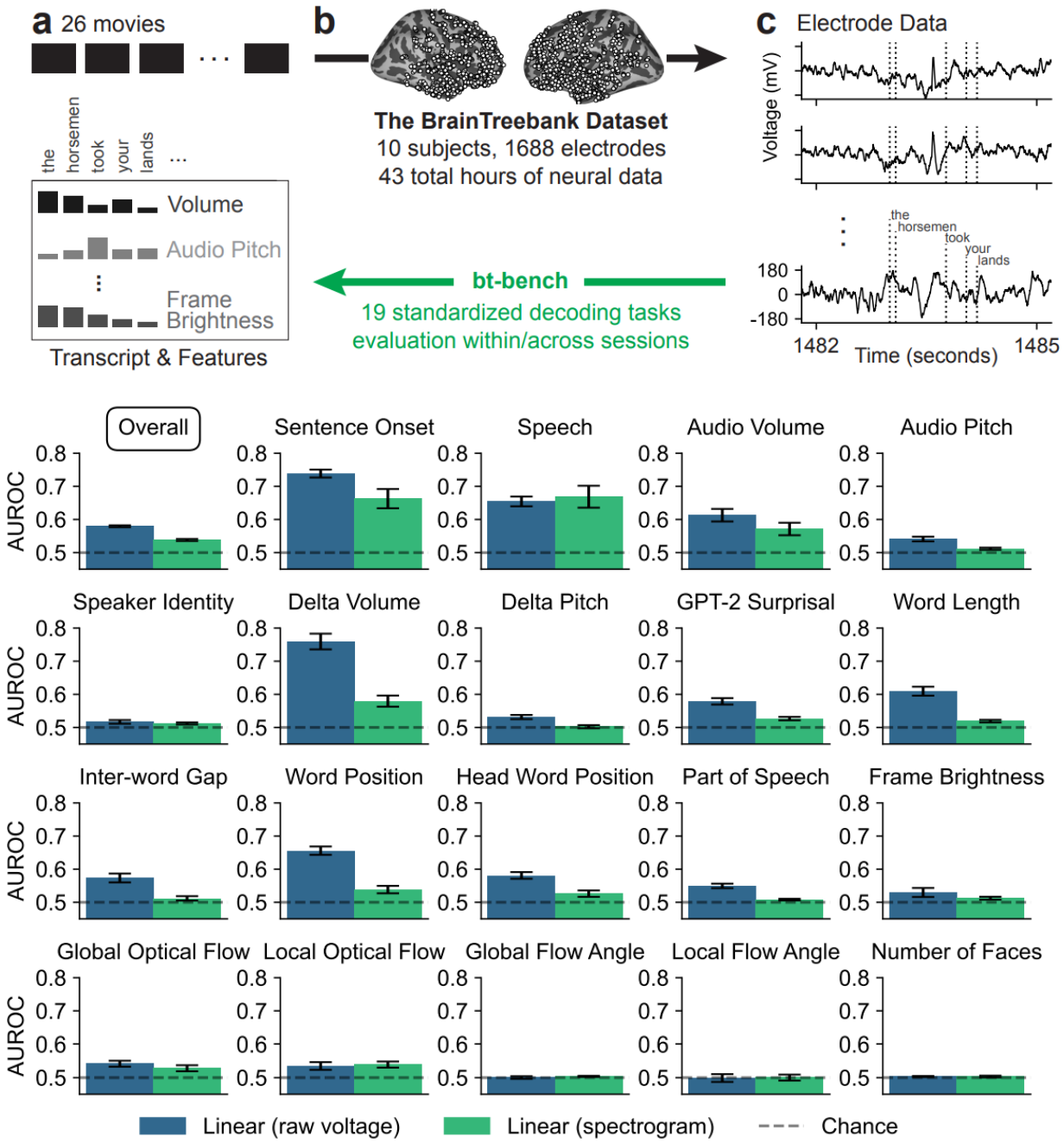




Benchmarking: BTBench

Evaluating foundation models of intracranial brain responses to naturalistic stimuli (based on the BrainTreebank dataset)

BTBench contains **19 standardized decoding tasks** (in the visual, auditory, language and multimodal categories), as well as defined train/test splits **that evaluate performance within or across recording sessions, and** within or across **human subjects.**

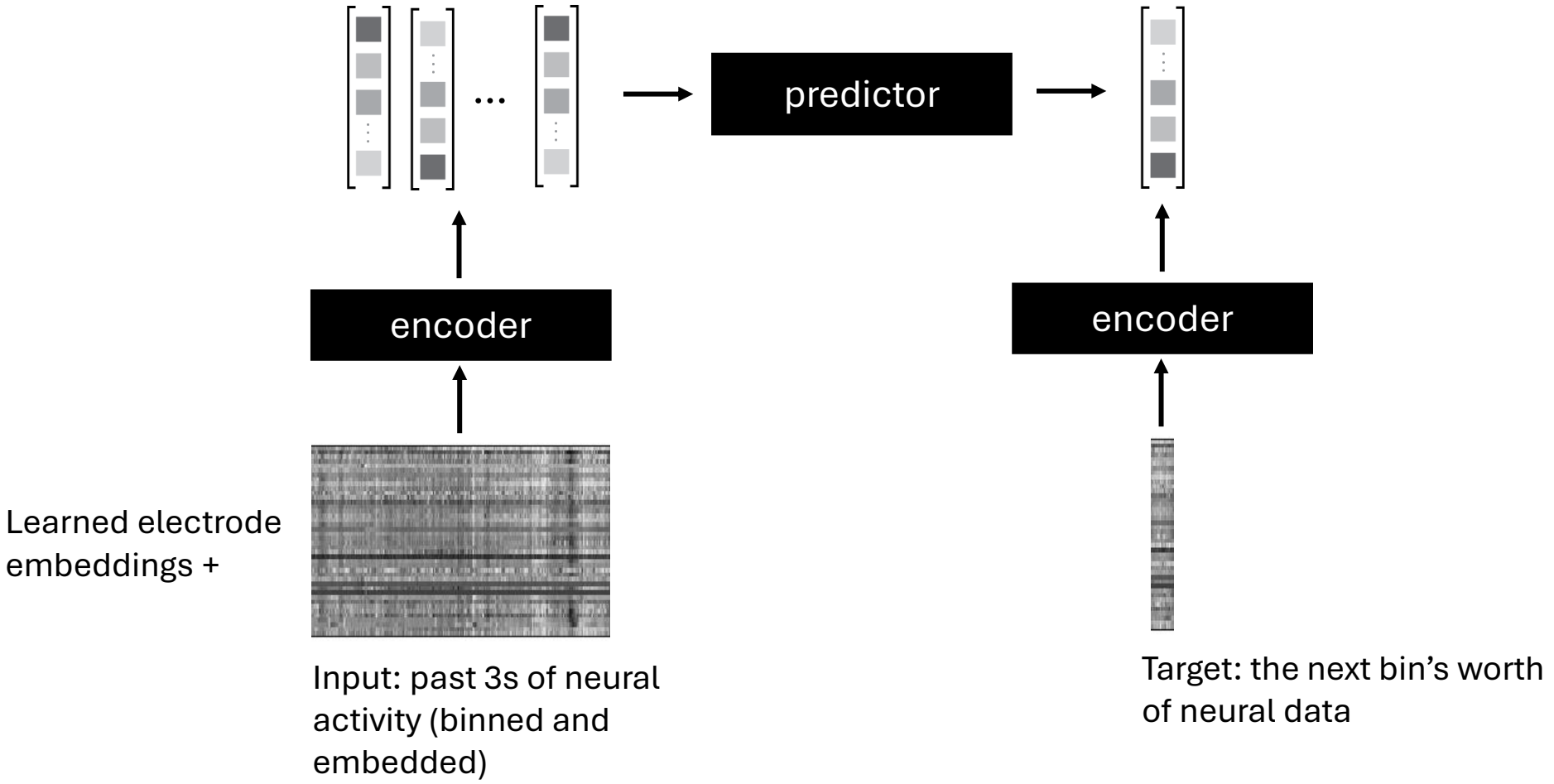


Zahorodnii, A., Stankovits, B., Wang, C., Moraitaki, C., Fiete, I. R., Katz, B., & Barbu, A. (in preparation). BrainTreebank-bench: Evaluating foundation models of intracranial brain responses to naturalistic stimuli.
https://github.com/azaho/btbench/

3. Current SOTA Solution (ours)

Andrii Zahorodnii
Fiete Lab, MIT Brain and Cognitive Sciences
3/2025

Idea: Predict in the latent space of the model

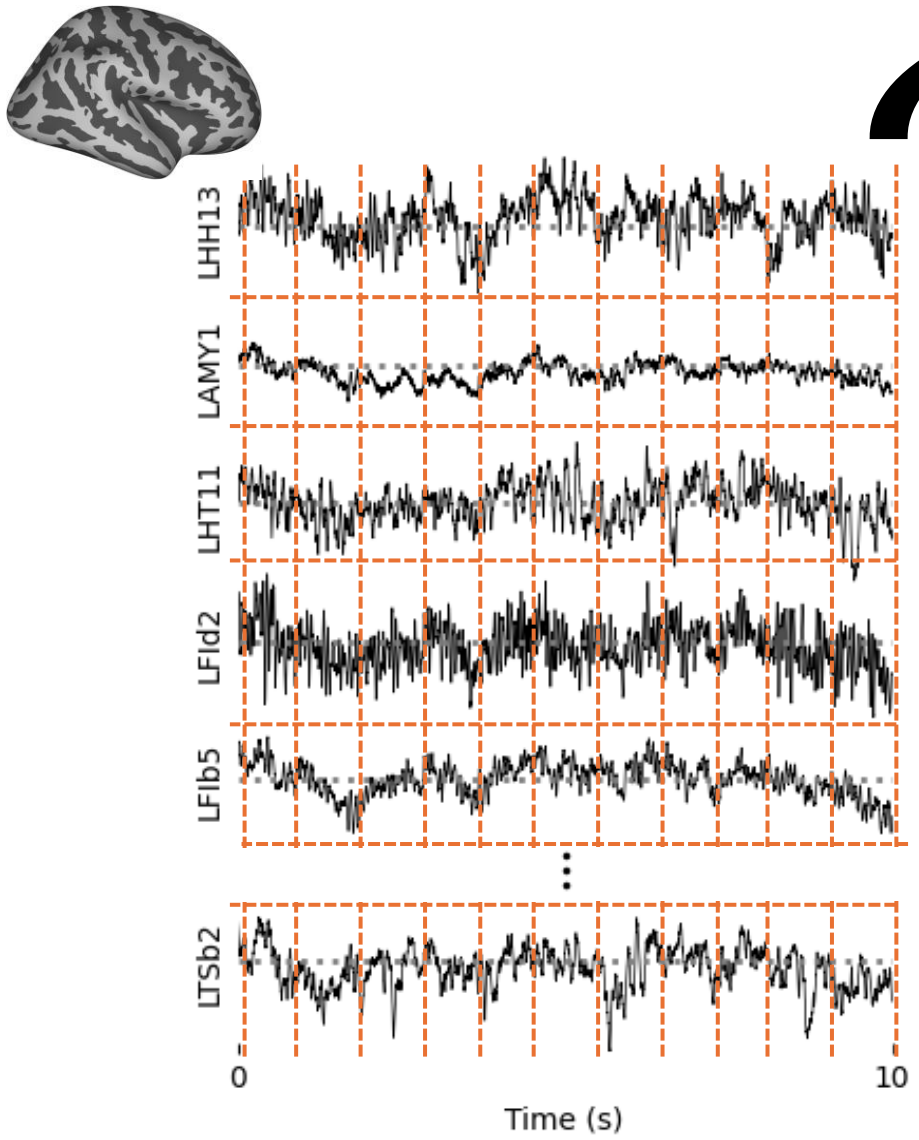


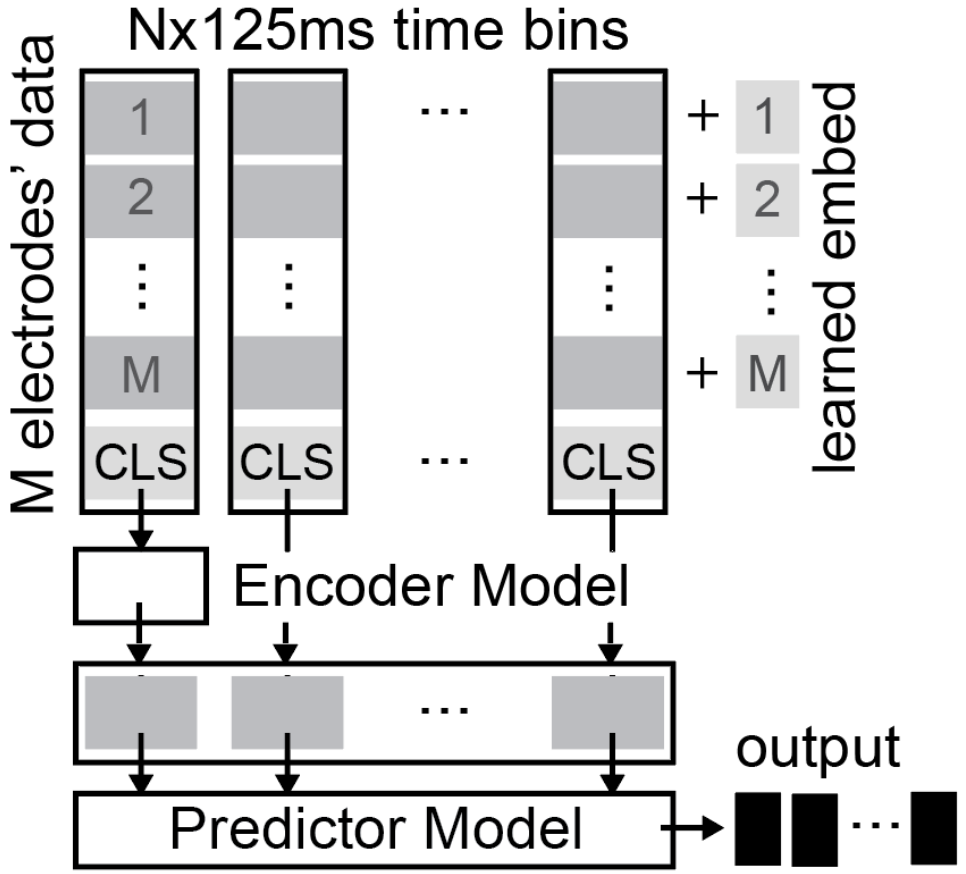
Benefit: there is no incentive to encode noise (or anything unhelpful for prediction) into the latent space.

BUT: Collapse! (a degenerate solution where encoder = constant)

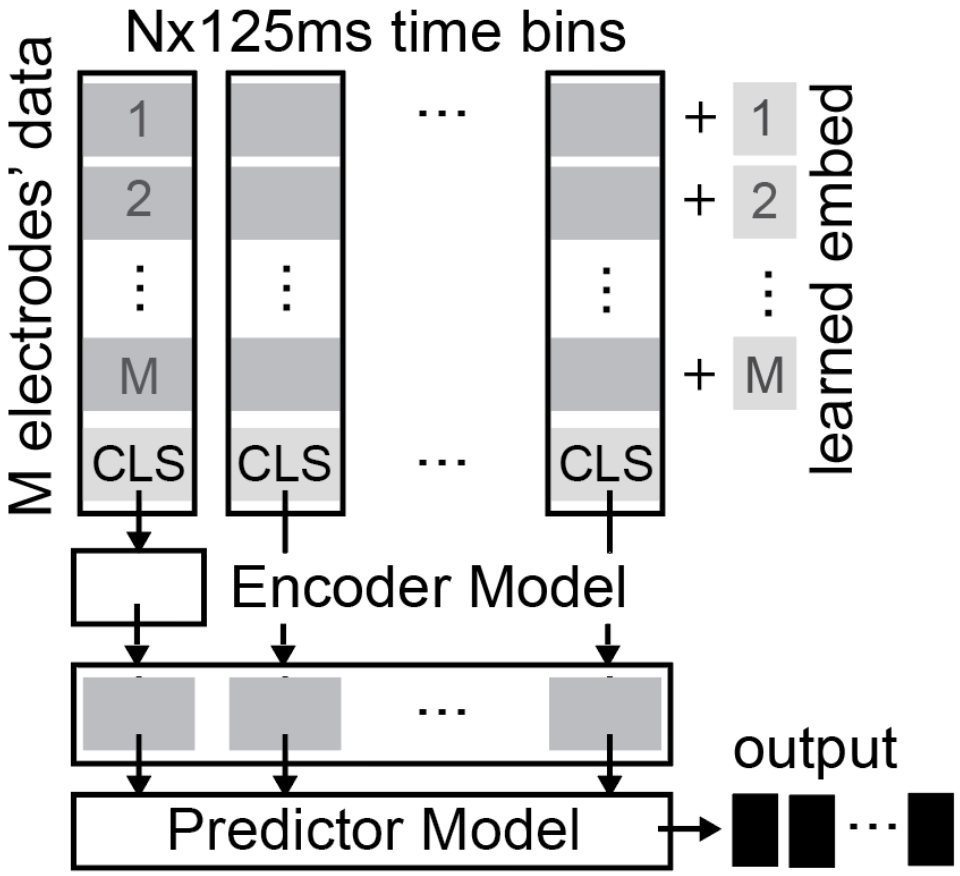
Specific Encoder / Predictor Architecture

input token = spectrogram of a patch





Specific Encoder / Predictor Architecture



Why this scheme?

- Clear candidate for the "Brain State" vector – the CLS token representation
  - And it's evolving in time, like a dynamical system

- Full architecture memory footprint is:

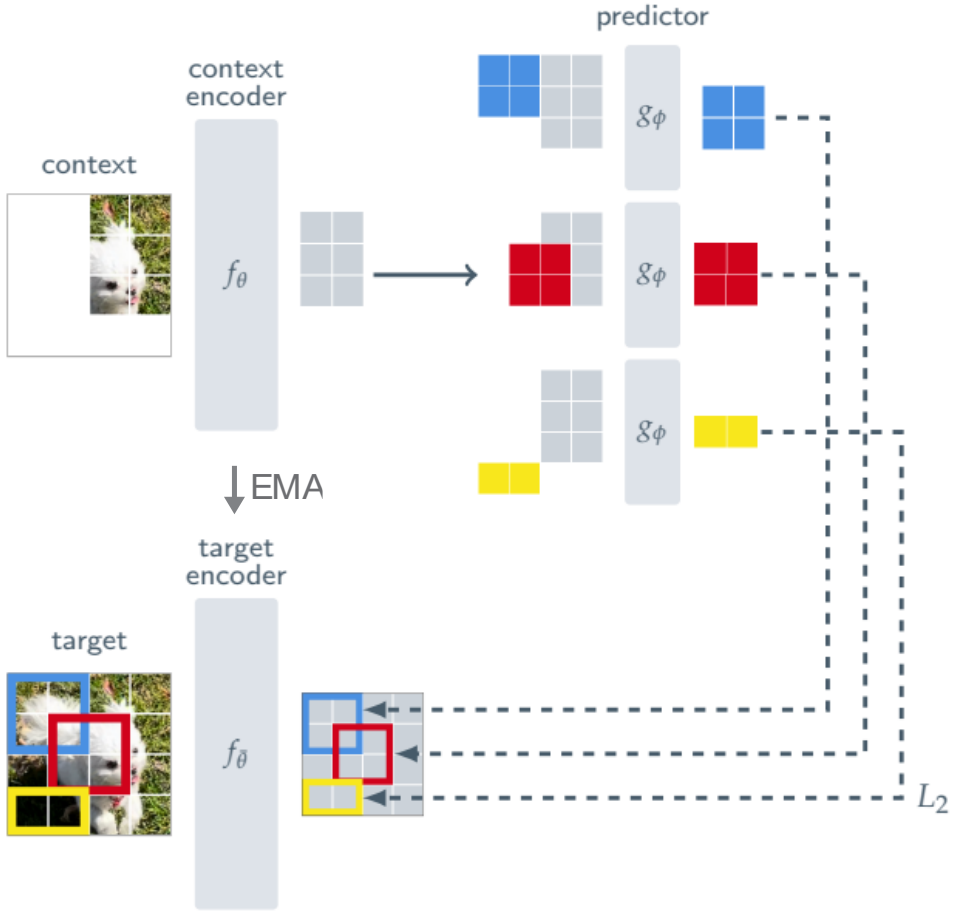
$$O((N_{context\ timebins} \times M_{electrodes})^2 \times Batch\ Size)$$

Whereas this has:

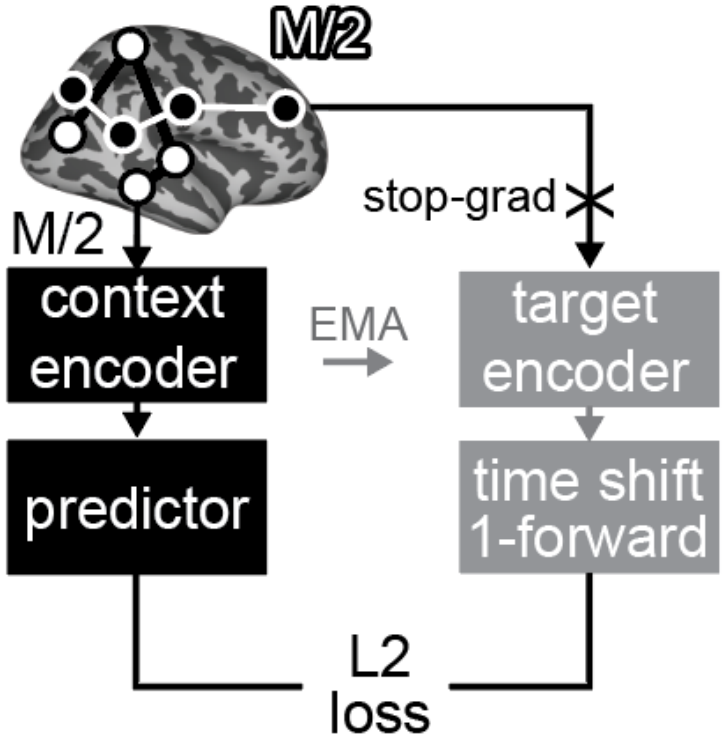
$$O(N_{context\ timebins} \times M^2_{electrode} \times Batch\ Size)$$

JEPA

I-JEPA (Assran et al. 2023)

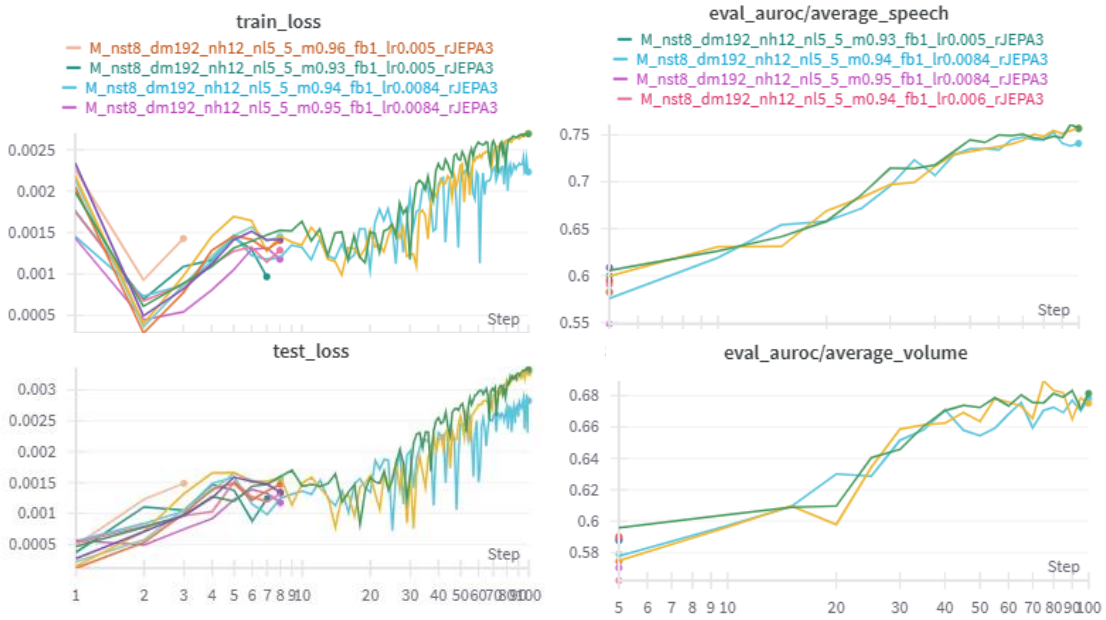


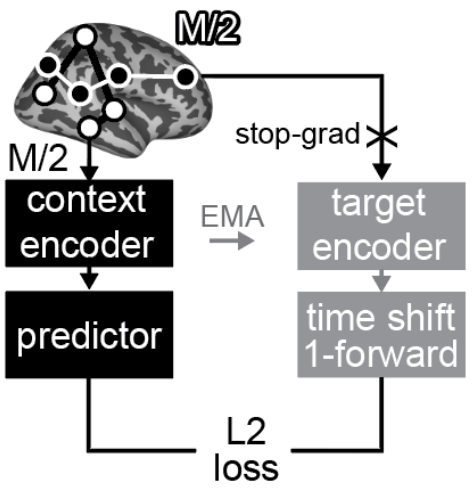
Our pretraining objective



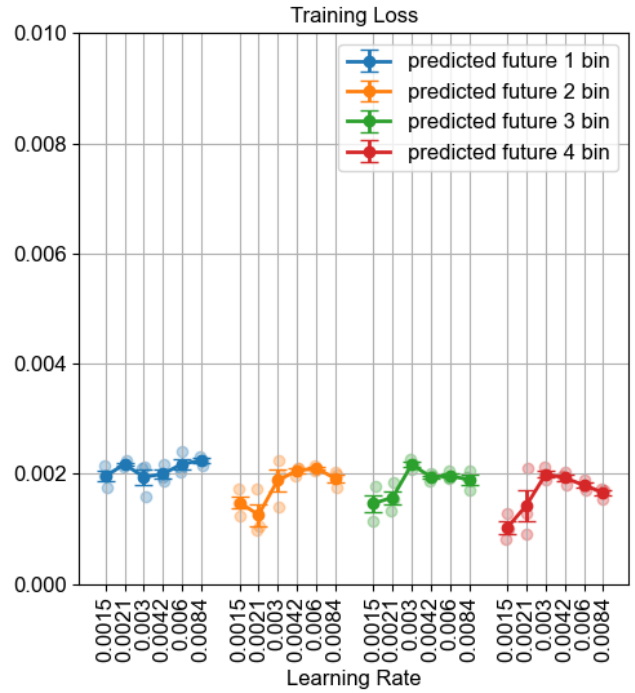
Another reference: Simple Siamese Representation Learning (Chen & He, 2020)

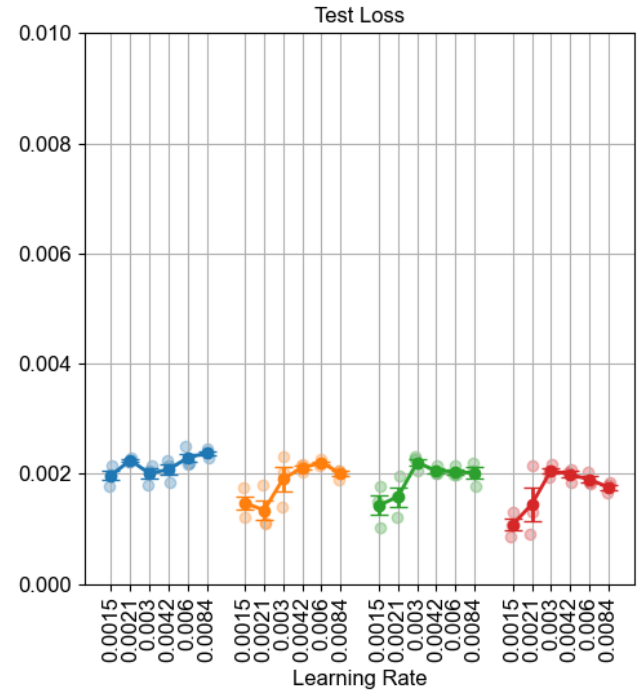
JEPA (momentum = 0.98)

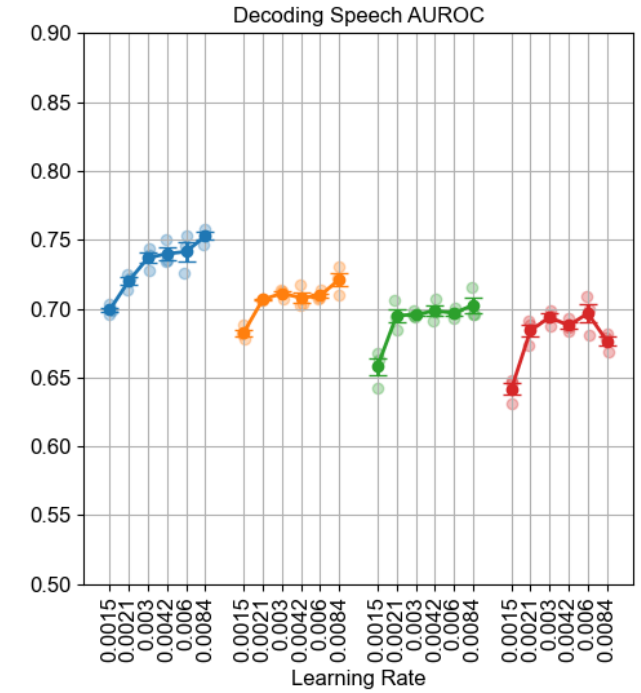


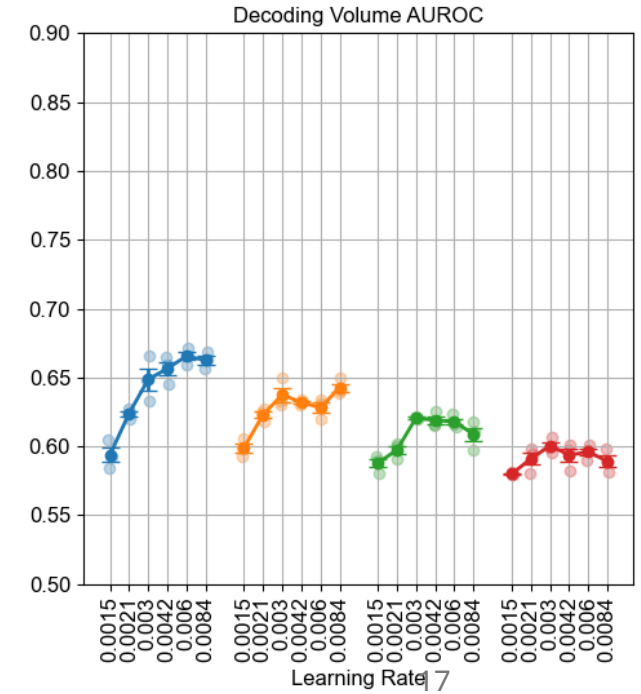


Performance after 100 epochs of training on btbank-lite:

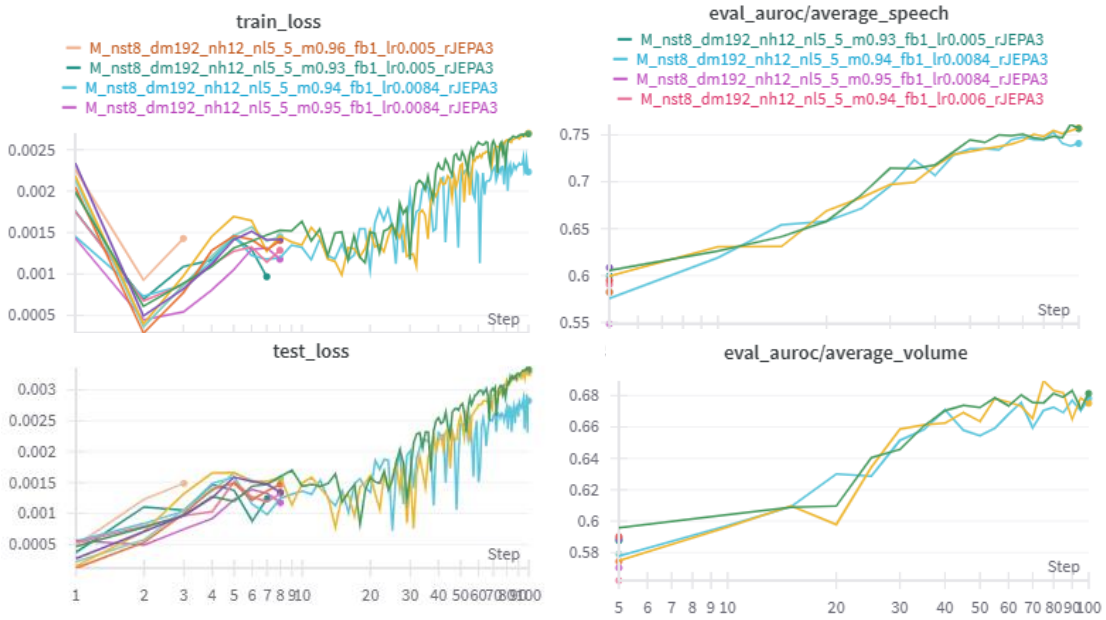


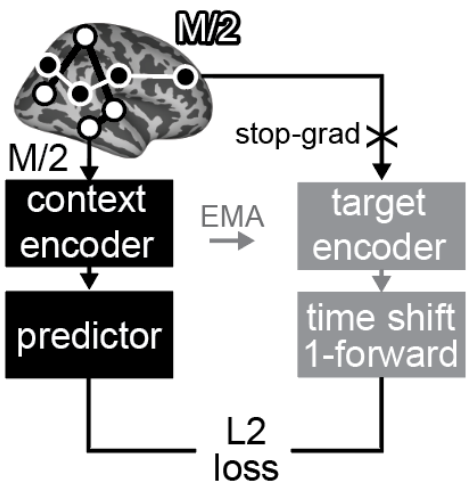




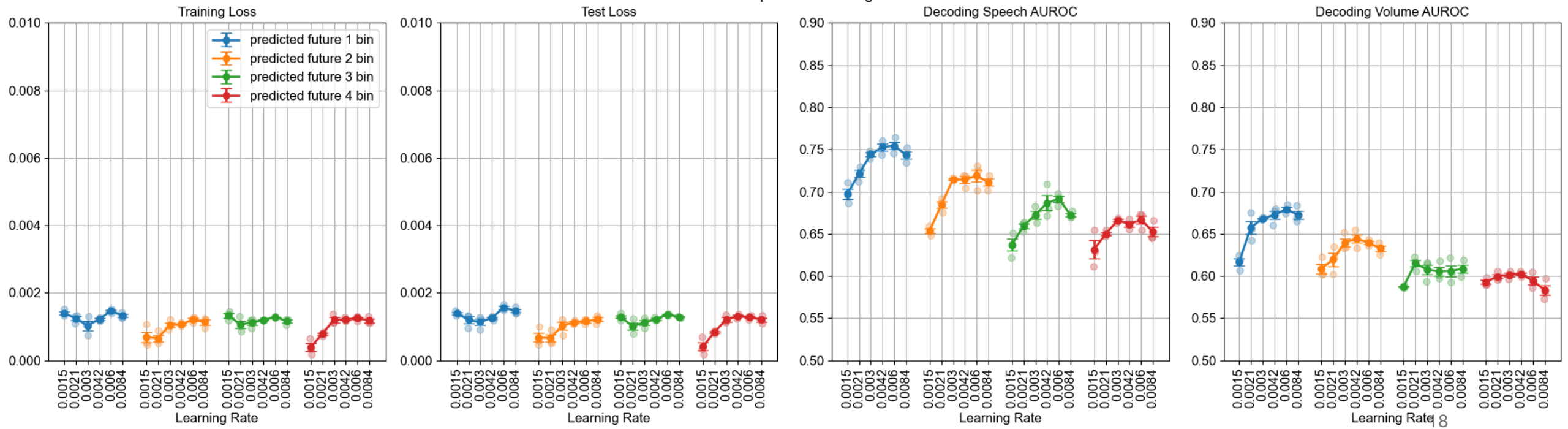


JEPA (momentum = 0.94)



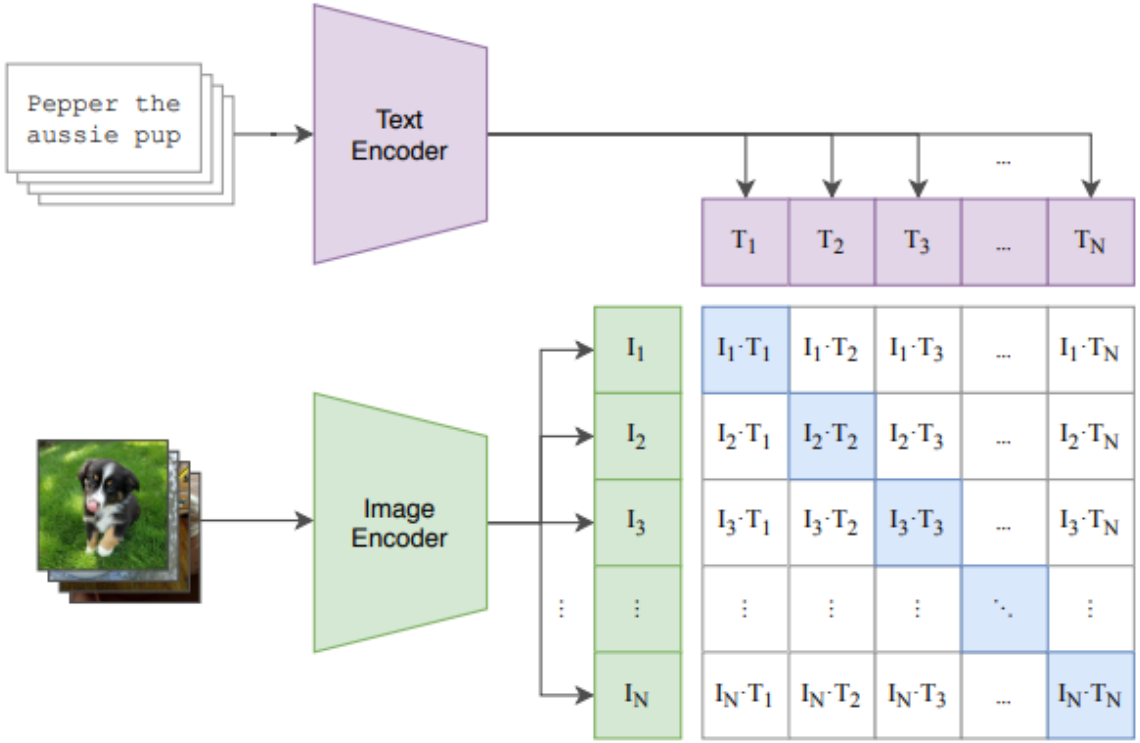


Performance after 100 epochs of training on btbank-lite:

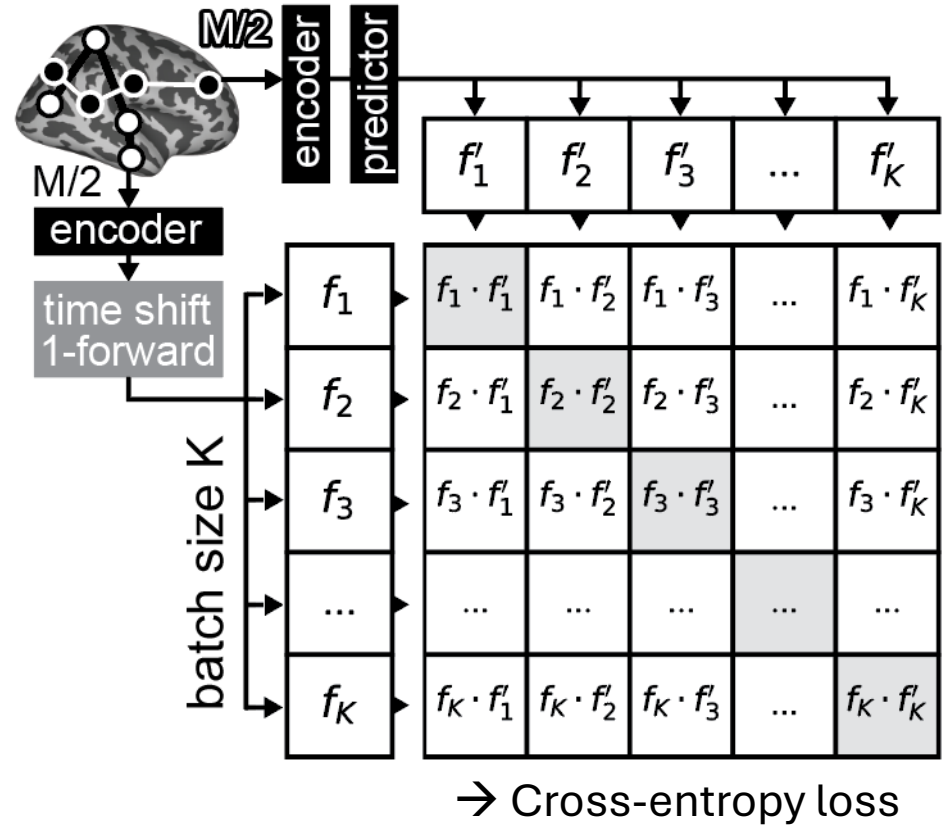


Contrastive Prediction Loss

CLIP (Radford & Kim et al. 2021)



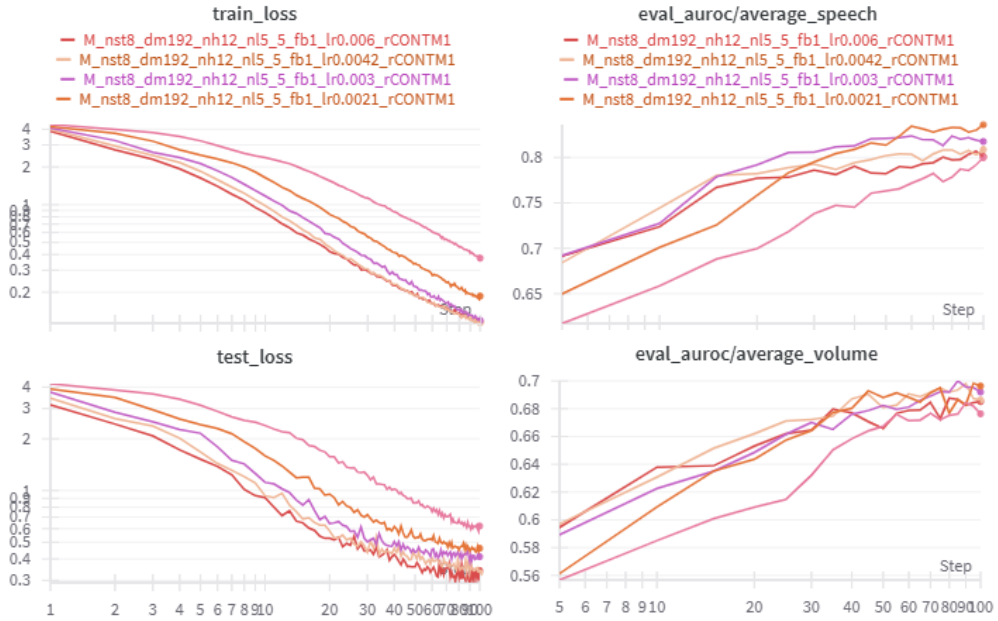
Our pretraining objective

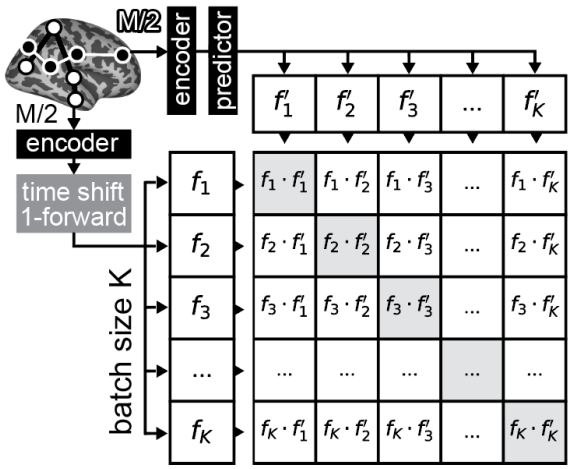


→ Cross-entropy loss

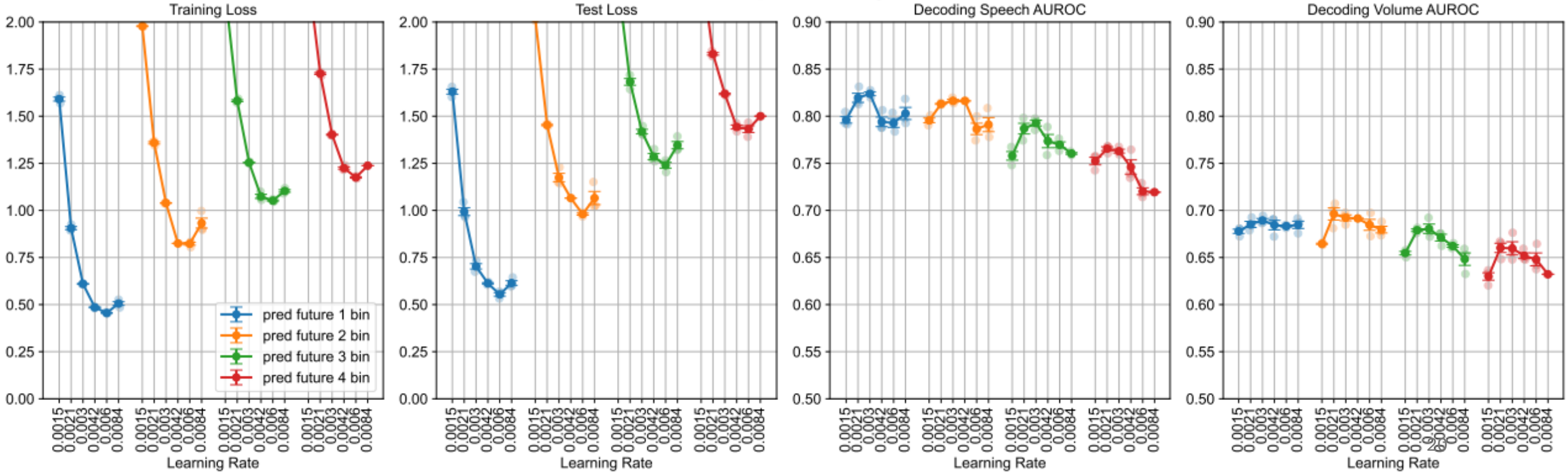
Another reference: Contrastive Predictive Coding (van den Oord, 2019)

Contrastive Prediction Loss





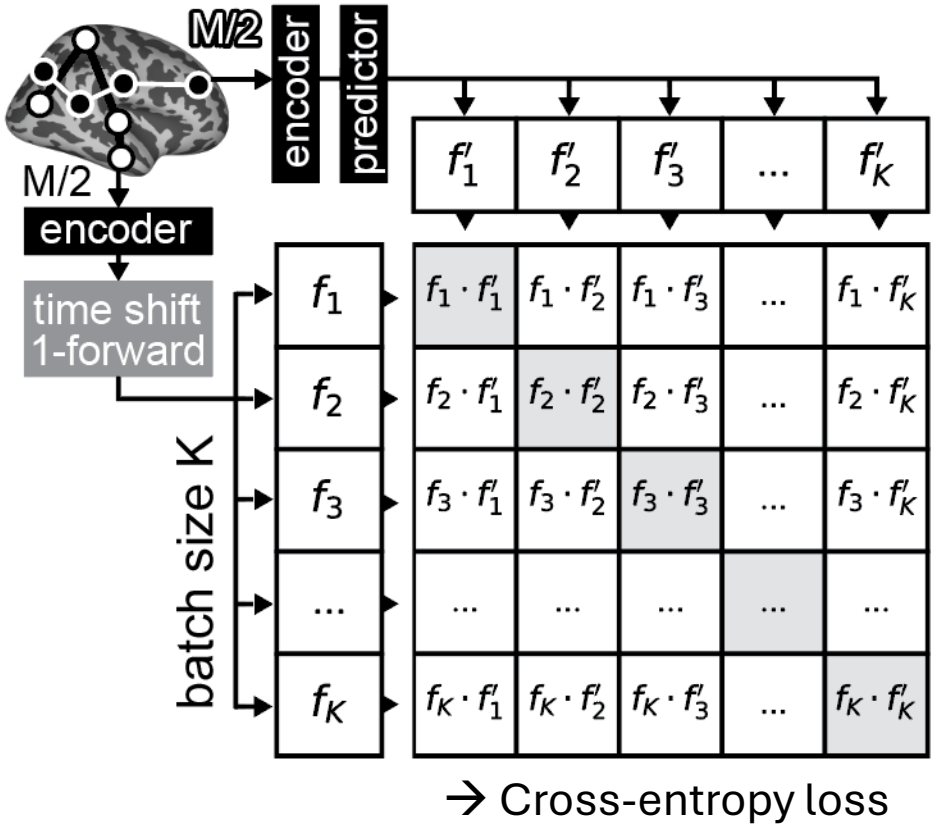
Performance after 100 epochs of training on btbank-lite:



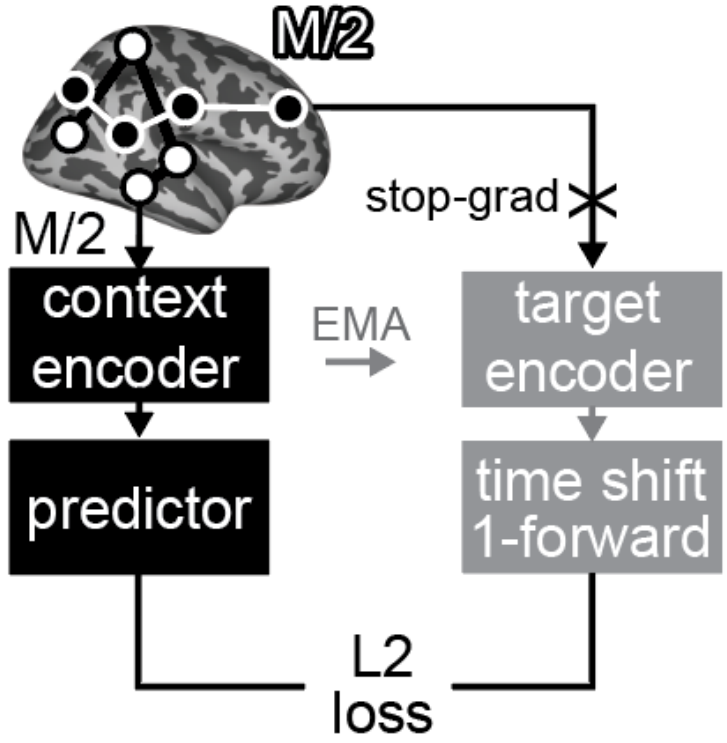
4. Preliminary Analysis Results

Andrii Zahorodnii
Fiete Lab, MIT Brain and Cognitive Sciences
3/2025

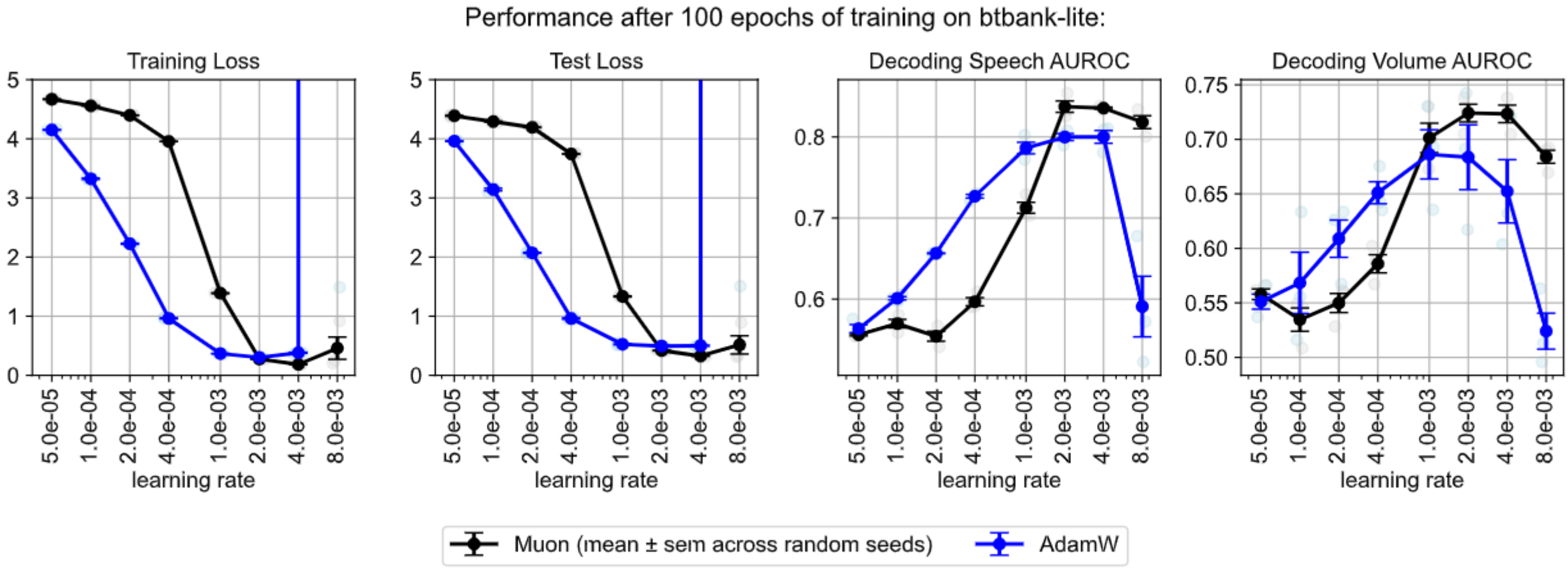
Preliminary result:
Contrastive Prediction Loss > JEPA. But why?



vs

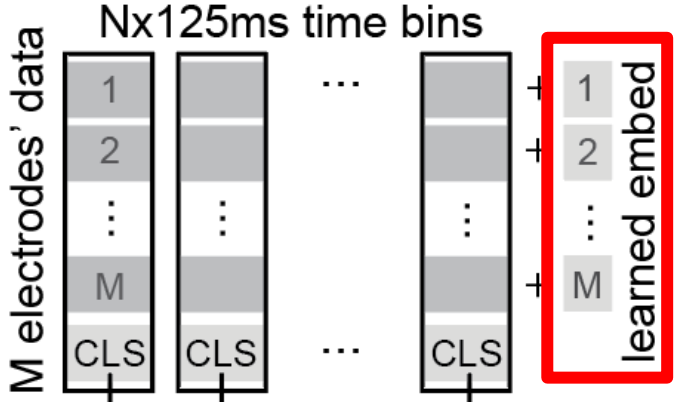


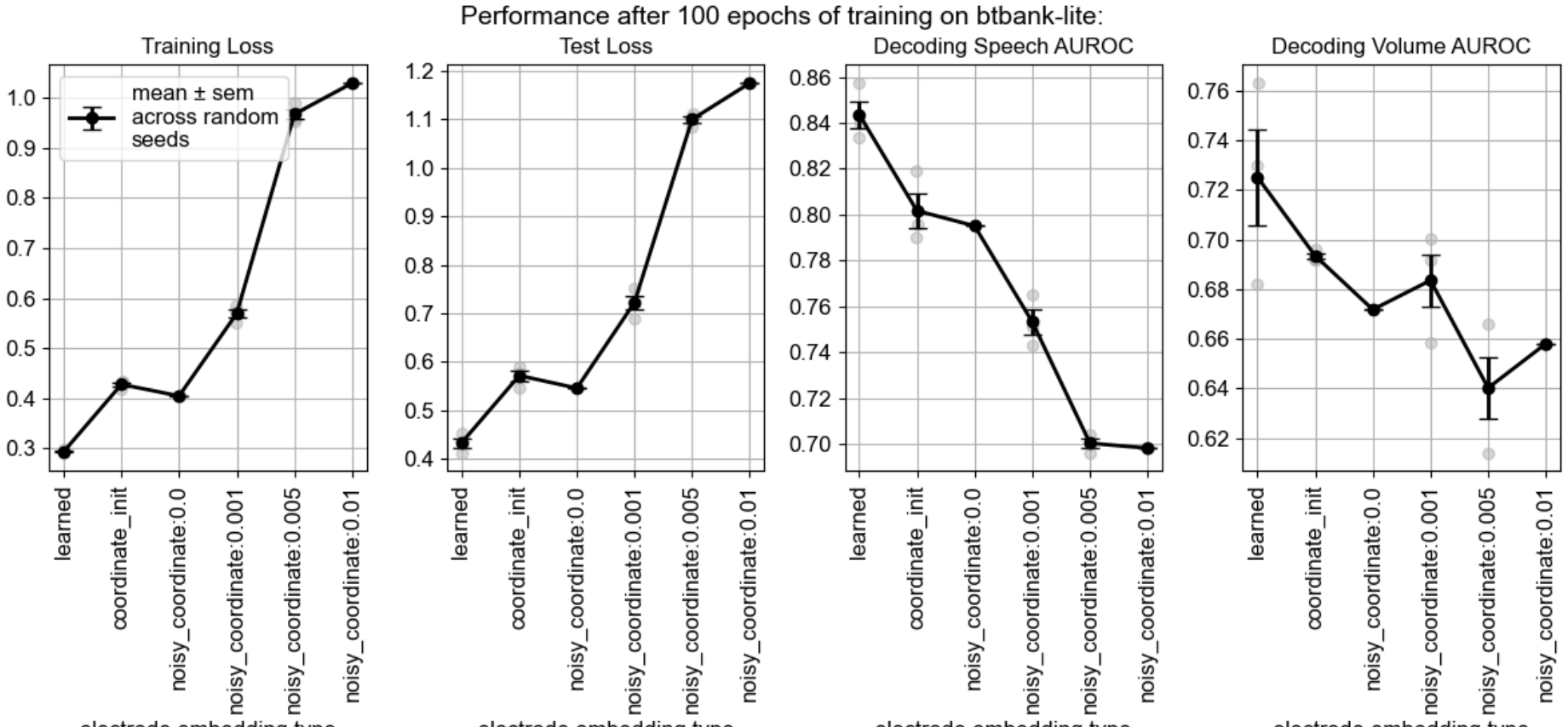
Preliminary result:
Muon is better than AdamW for learning this data!



(this is for the contrastive loss)

Preliminary result:
Electrode coordinates NOT needed(!)
Learned Embeddings are better.



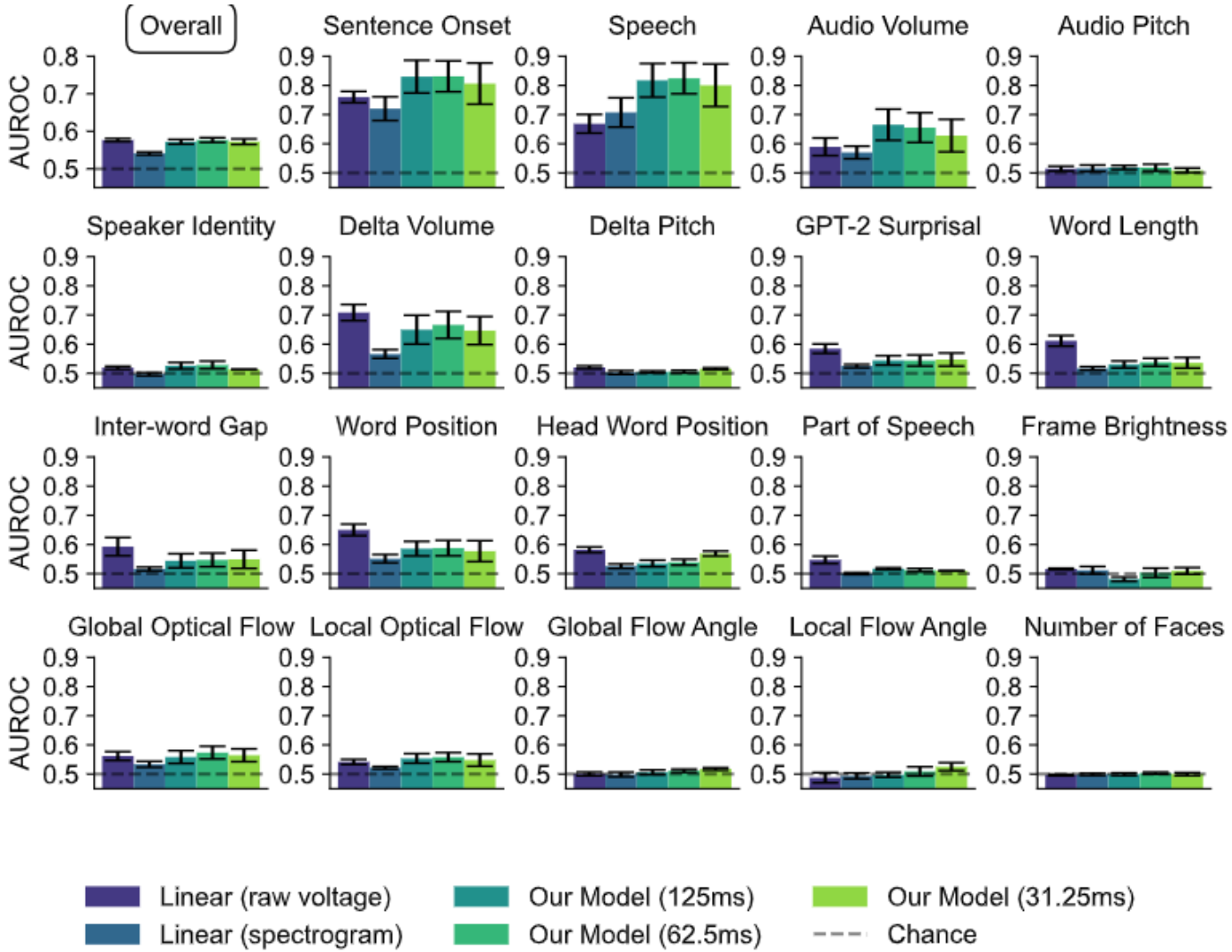


Q for future:

What about no electrode embeddings whatsoever?

Can the model learn electrode identity fully in-context?

Suspicion: Taking spectrogram of the signal might hurt performance – from evaluation on BTBench.



("Our Model" = contrastively trained model on top of spectrogram of LFP)

The model is always above the spectrogram regression baseline.

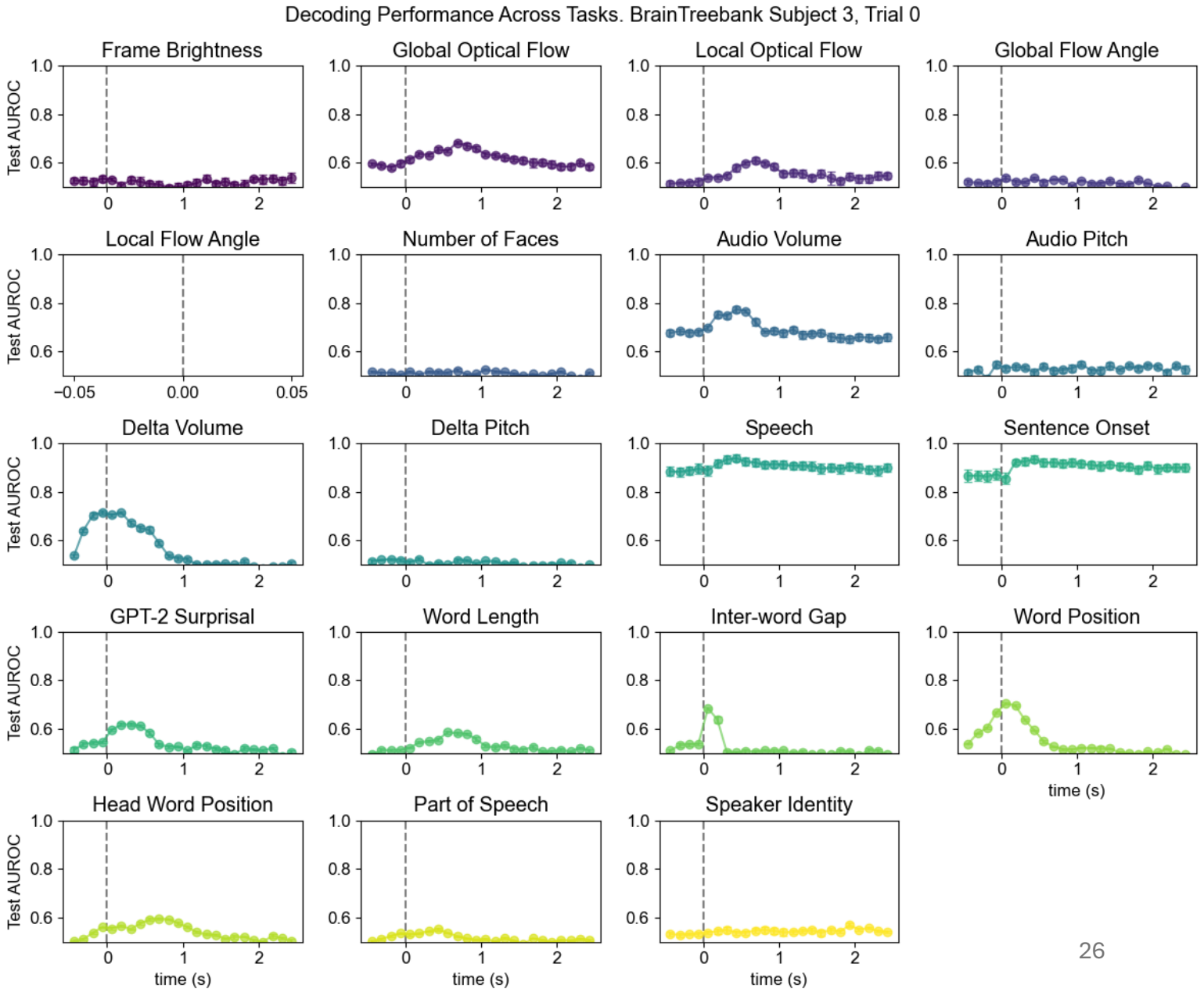
But the raw voltage regression baseline performs better than spectrogram. Often, better than the trained model too!

And...
EVERYBODY is taking the spectrogram!

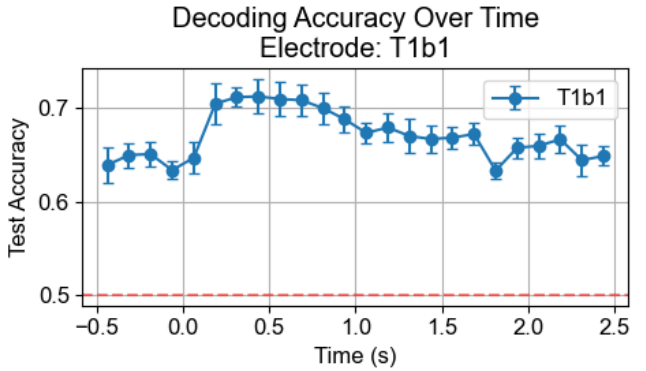
Tracking Information Processing

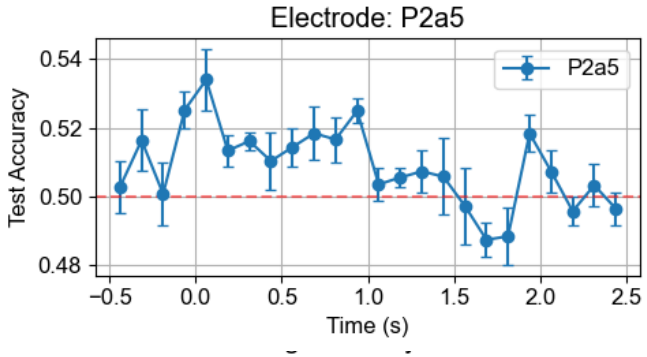
Decoding using linear regression from the latent space of the model

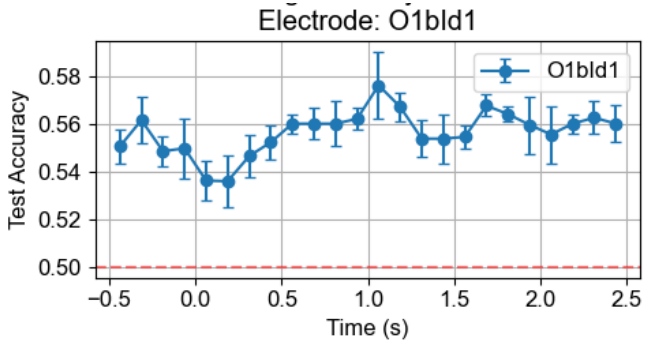
across different time bins of data (time-locked to word onset)



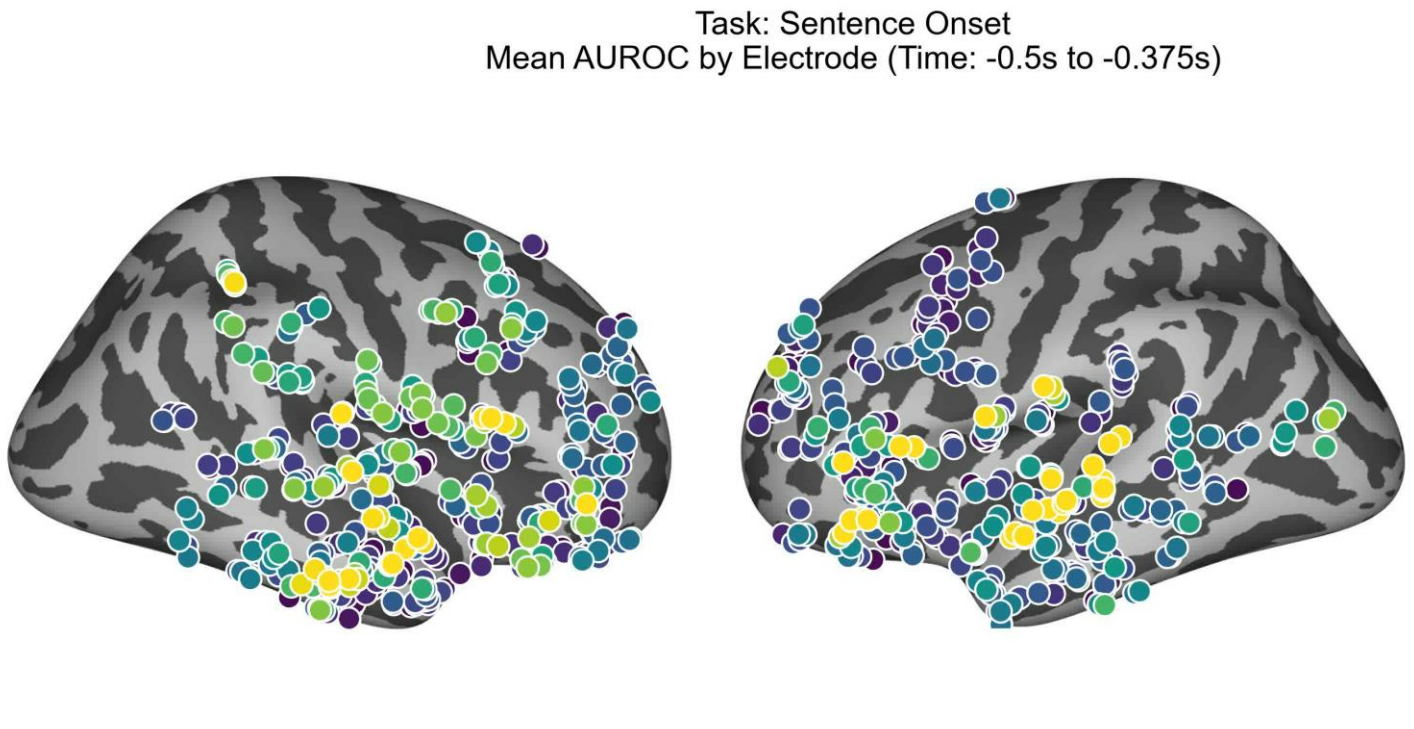
Tracking Information Processing over Electrodes





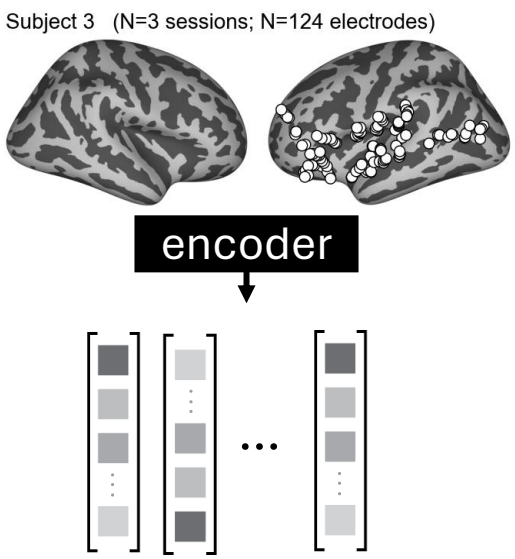


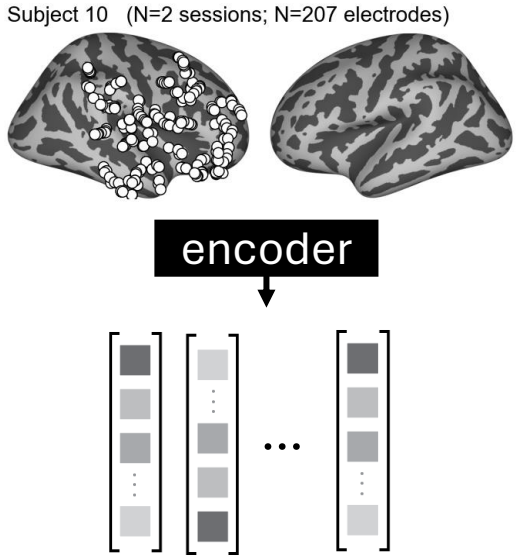
. . .

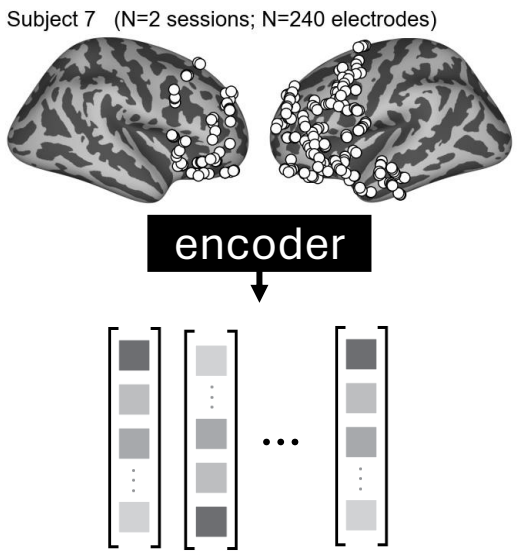


Testing alignment of the latent space across people

Three subjects watching the same movie (“Cars 2”). Features extracted at the same points in the movie

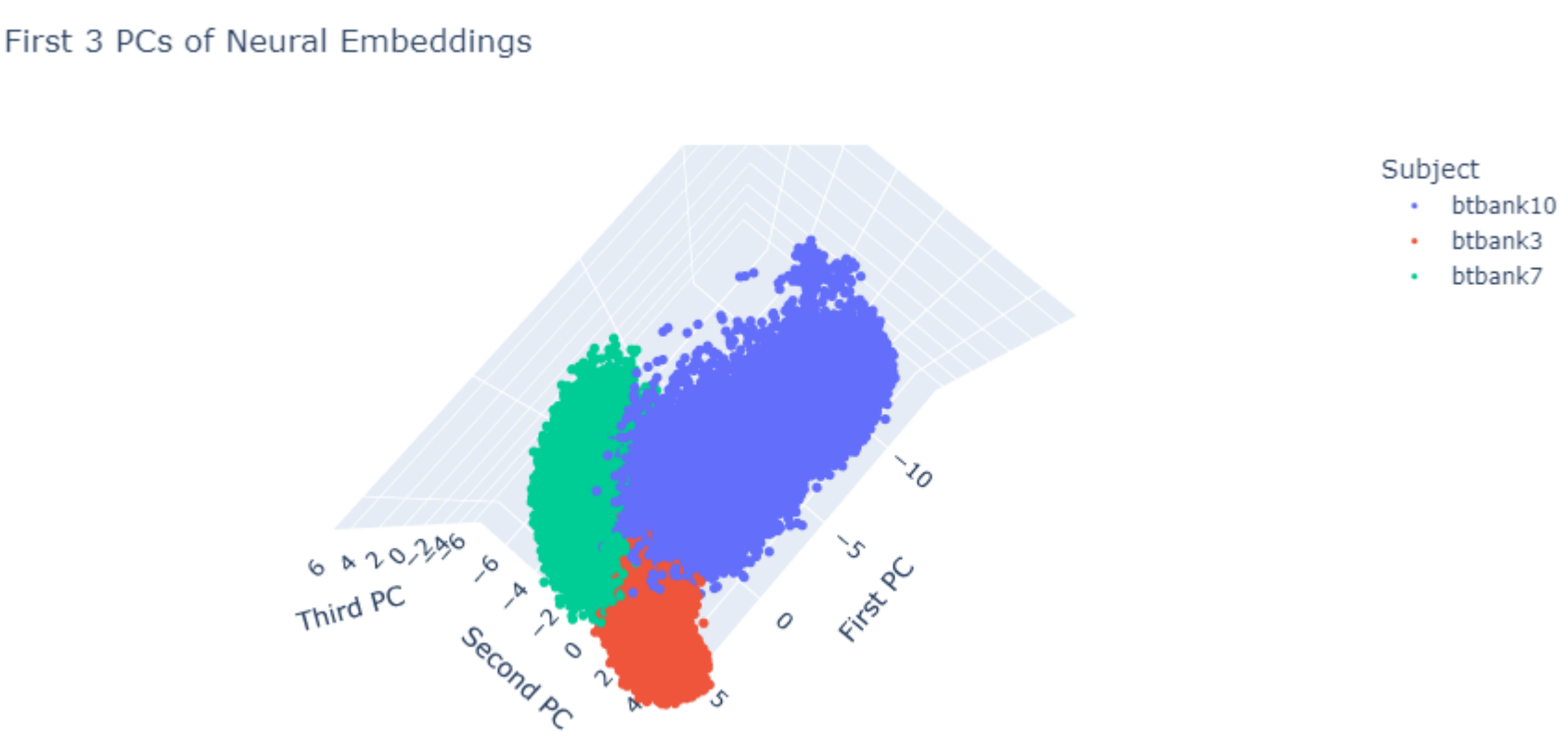






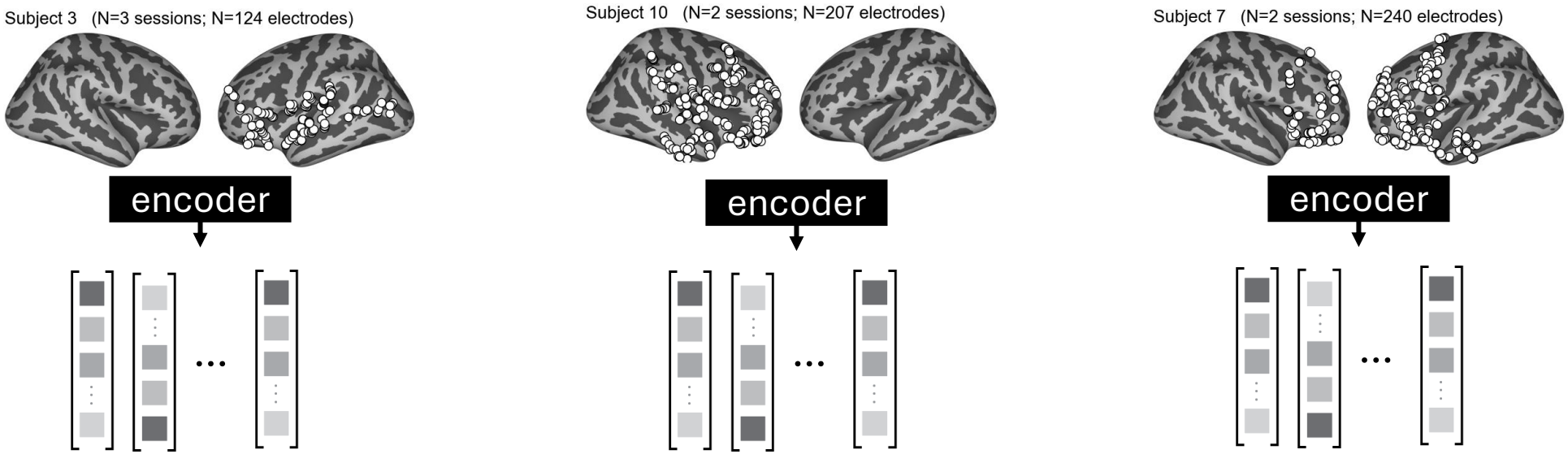
Preliminary result:
Different people watching the same movie are encoded in different subspaces of the latent space

Three subjects watching the same movie (“Cars 2”). Features extracted at the same points in the movie



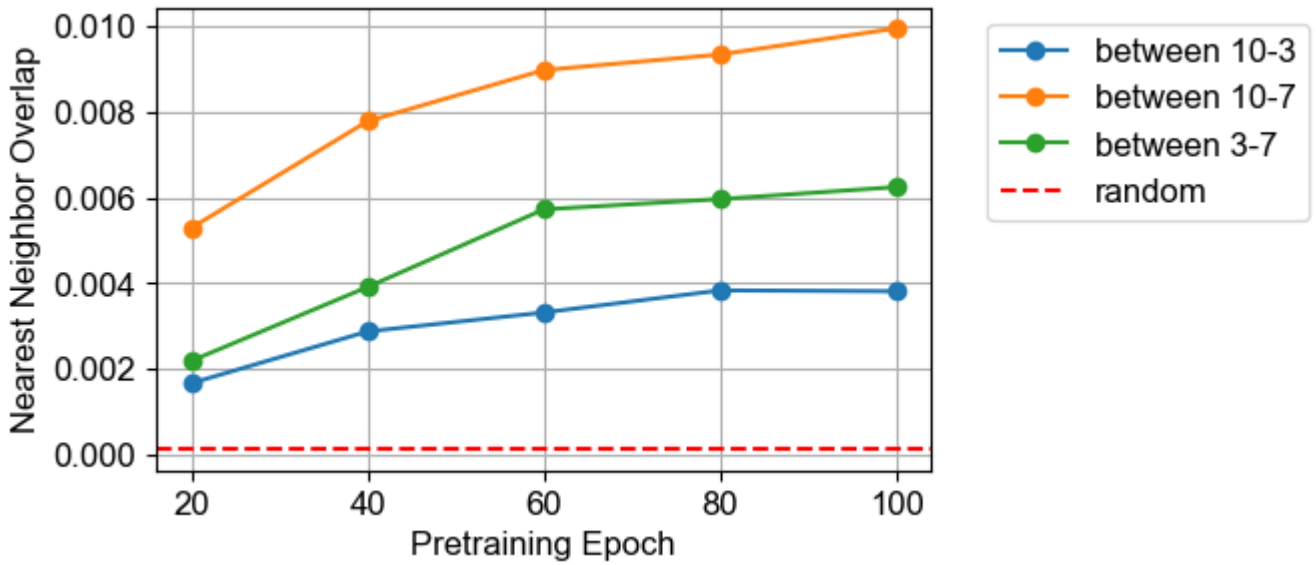
Platonic Representation Hypothesis for Brains?

Three subjects watching the same movie (“Cars 2”). Features extracted at the same points in the movie

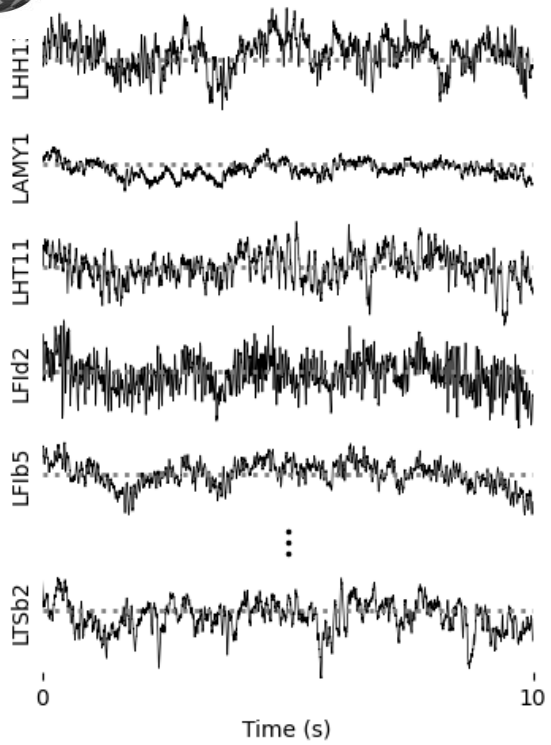


Mutual k-NN (Huh & Cheung et al, 2024)

Platonic Representation Hypothesis for Brains?



A multimodal feature extractor?



model →

$x(t)$

$\vec{\lambda}_1 \cdot x(t) =$ “imminence of seizure”

$\vec{\lambda}_2 \cdot x(t) =$ “intensity of felt emotion”

⋮

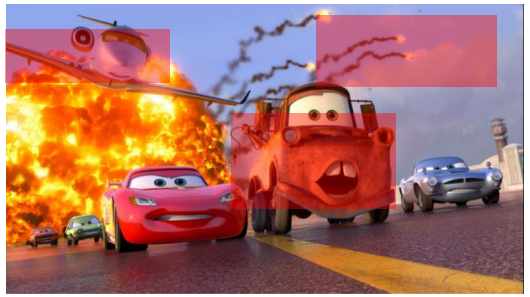
$\Lambda x(t) =$ “content of inner vocalization”

A multimodal feature extractor?

if $\vec{\lambda}_1 \cdot x(t) =$ “imminence of seizure”, then

$$\frac{\partial\left(\vec{\lambda}_1 \cdot x(t)\right)}{\partial(\text{sensory input})}$$

→ model (inverted)



high derivative score

→ targeted non-invasive intervention
(modifying the stimulus)