BrainBERT: Self-supervised representation learning for Intracranial Electrodes

ICLR 2023 Under Review

Track: Neuroscience and Cognitive Science (e.g., neural coding, brain-computer interfaces)

Score: 8 8 6 5

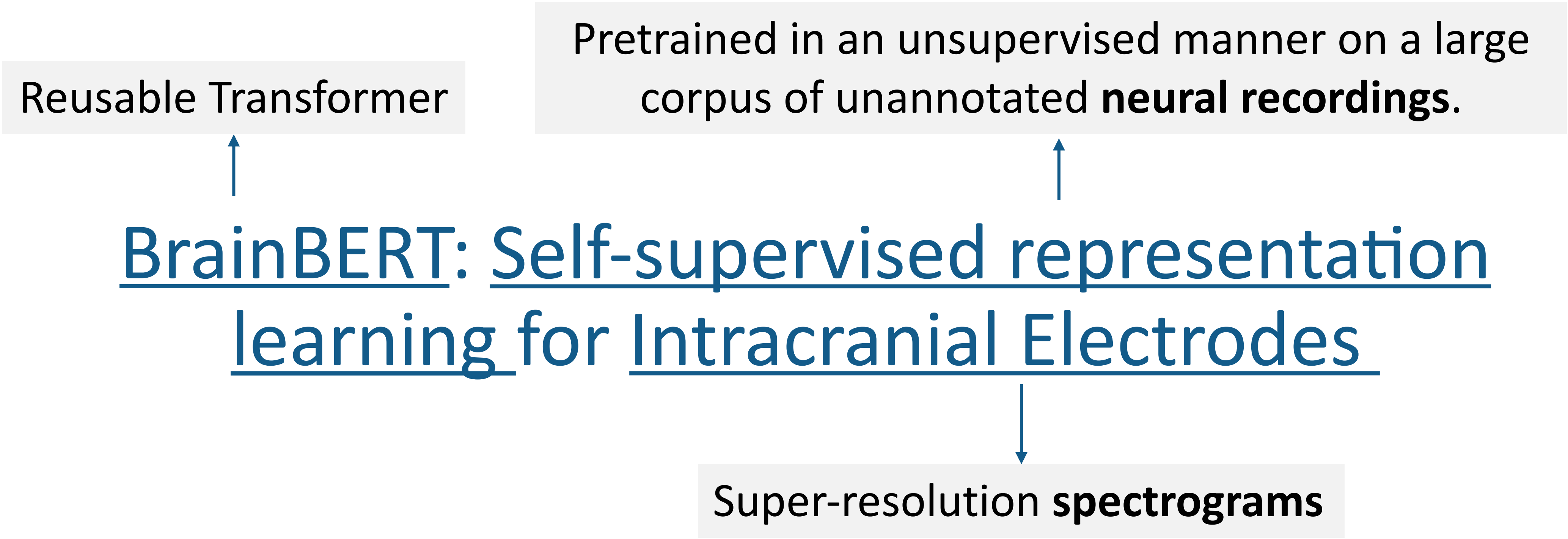
HIT Xiachong Feng

2022.11.29

ICLR Neuroscience and Cognitive Science Track

Incremental Learning of Structured Memory via Closed-Loop Transcription
A probabilistic framework for task-aligned intra- and inter-area neural manifold estimation
A Theoretical Framework for Inference and Learning in Predictive Coding Networks
Real-time variational method for learning neural trajectory and its dynamics
Words are all you need? Language as an approximation for representational similarity
Disentangling with Biological Constraints: A Theory of Functional Cell Types
Representational Dissimilarity Metric Spaces for Stochastic Neural Networks
Simplicial Hopfield networks
Backpropagation at the Infinitesimal Inference Limit of Energy-Based Models: Unifying Predictive Coding, Equilibrium Propagation, and Contrastive Hebbian Learning
Interneurons accelerate learning dynamics in recurrent neural networks for statistical adaptation
Training language models for deeper understanding improves brain alignment
GAMR: A Guided Attention Model for (visual) Reasoning
BrainBERT: Self-supervised representation learning for Intracranial Electrodes

Begin



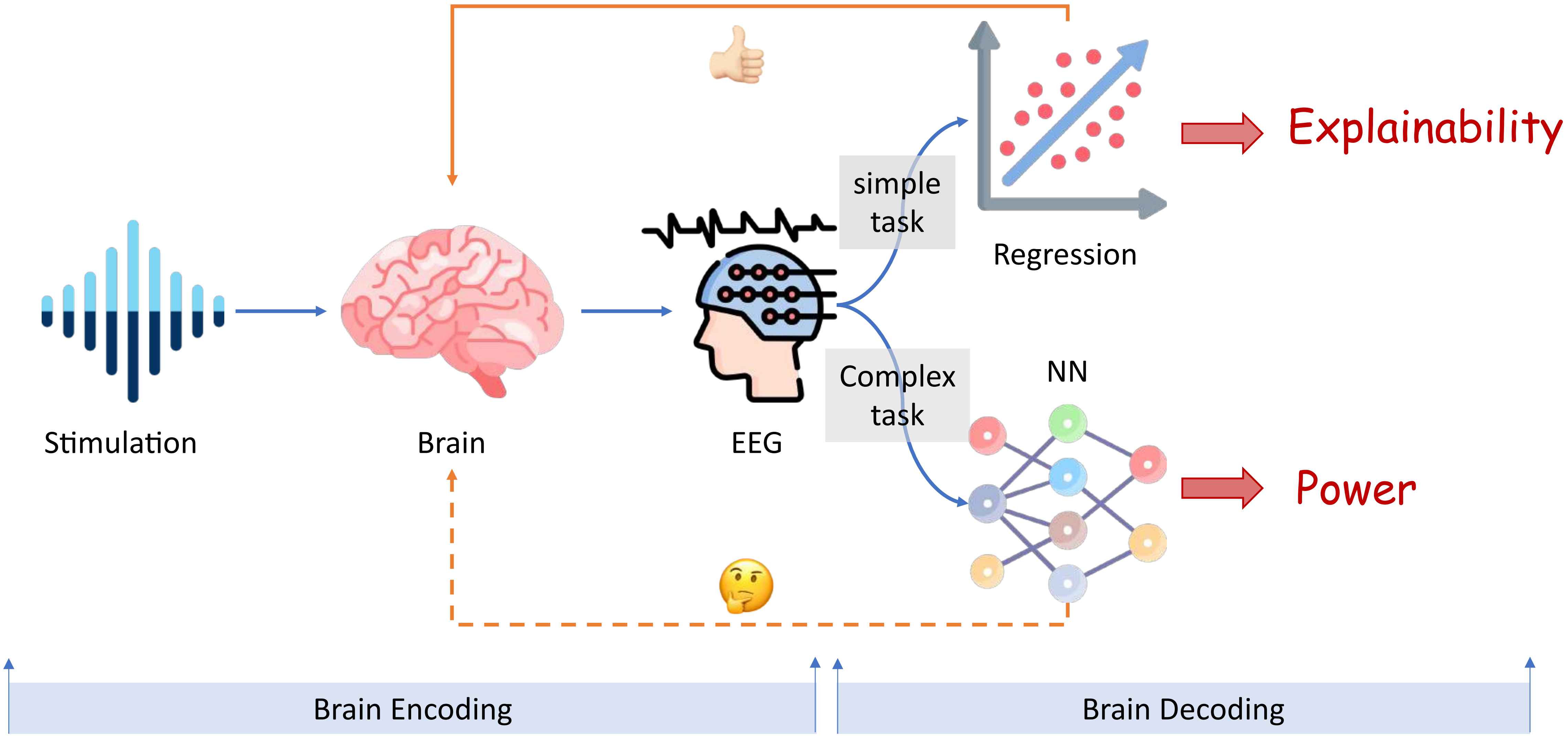
Reusable Transformer

Pretrained in an unsupervised manner on a large corpus of unannotated **neural recordings**.

BrainBERT: Self-supervised representation learning for Intracranial Electrodes

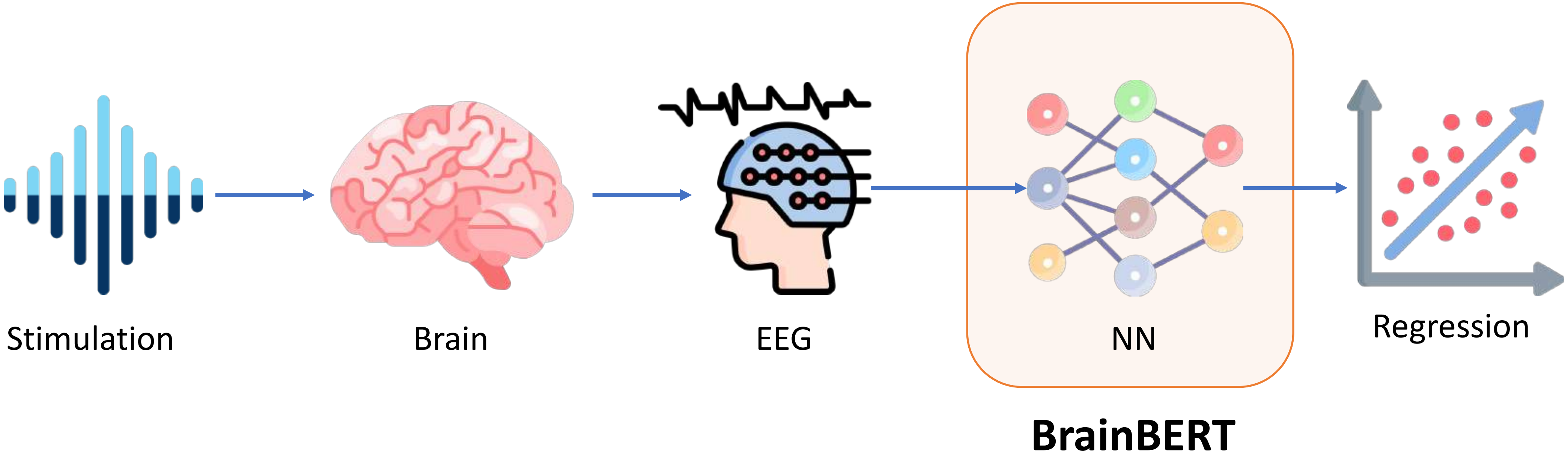
Super-resolution **spectrograms**

Motivation of BrainBERT



Motivation of BrainBERT

- How to balance both power and explainability?



Advantages

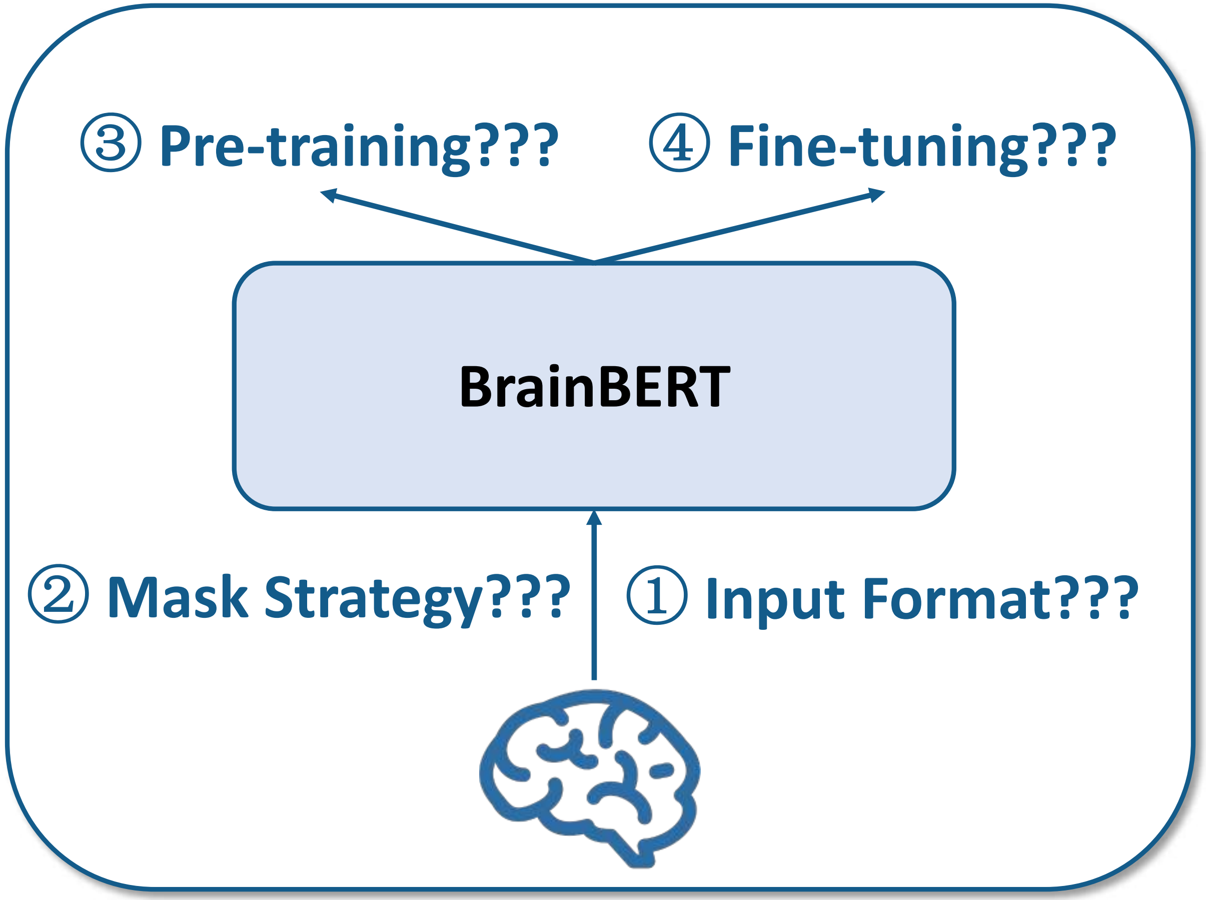
- BrainBERT is pretrained once **across a pool of subjects**, and then provides off-the-shelf capabilities for analyzing new subjects with **new electrode locations** even when data is scarce.

- Neuroscientific experiments tend to have **little data** in comparison to other machine learning settings, making additional sample efficiency critical.

- Other applications, such as **brain-computer interfaces** can also benefit from **shorter training regimes**, as well as from BrainBERT's significant performance improvements.

- In addition, the embeddings of the neural data provide a new means by which to investigate the brain.

BrainBERT



③ Pre-training???

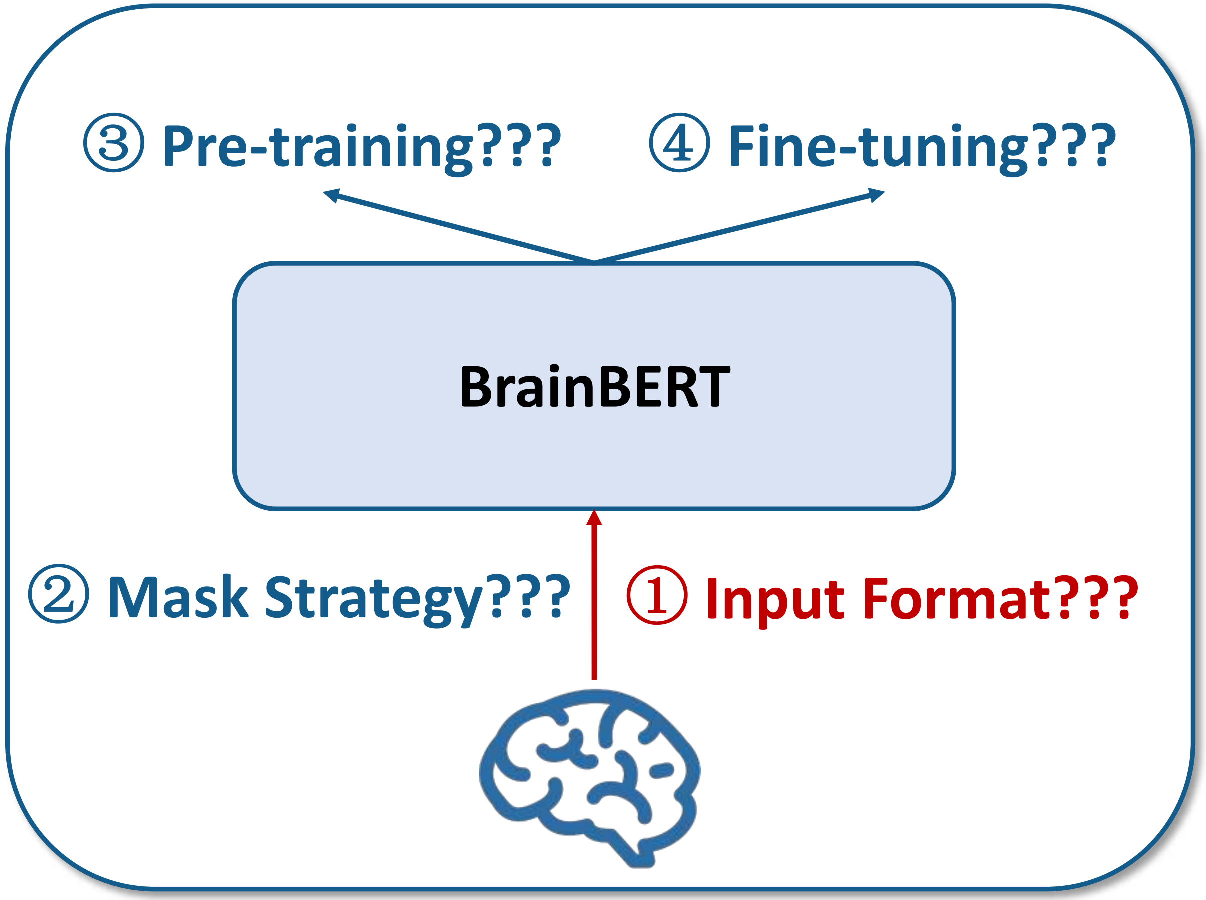
④ Fine-tuning???

BrainBERT

② Mask Strategy???

① Input Format???

BrainBERT



③ Pre-training???

④ Fine-tuning???

BrainBERT

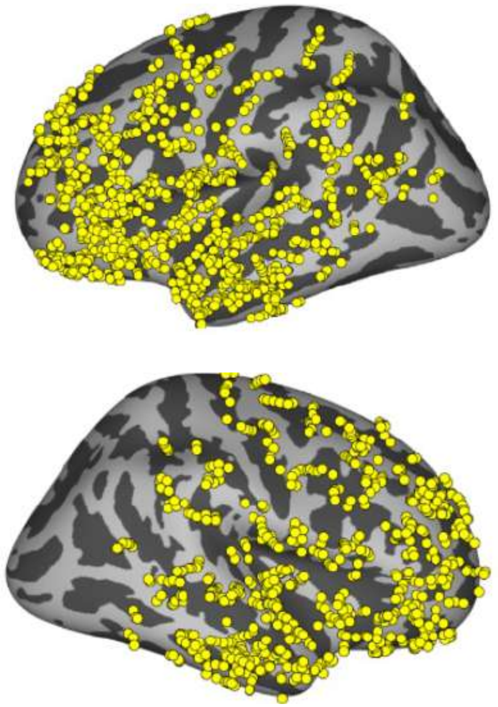
② Mask Strategy???

① Input Format???

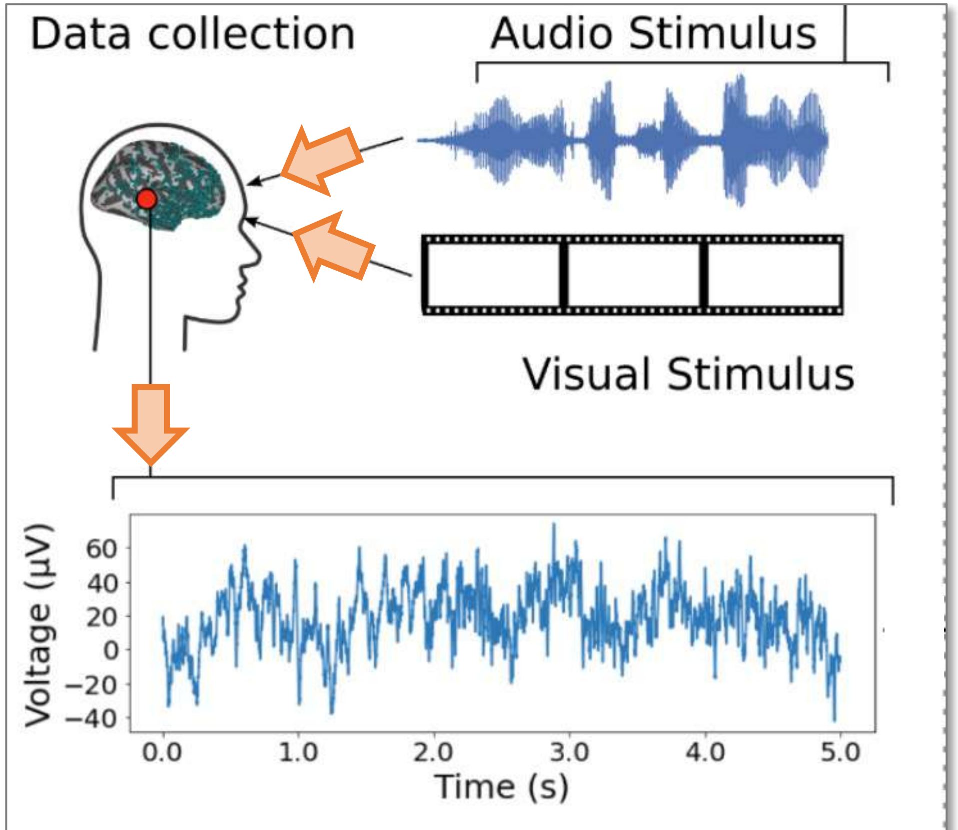
Input: SEEG

- Stereo-electroencephalographic (SEEG, 立体脑电图)

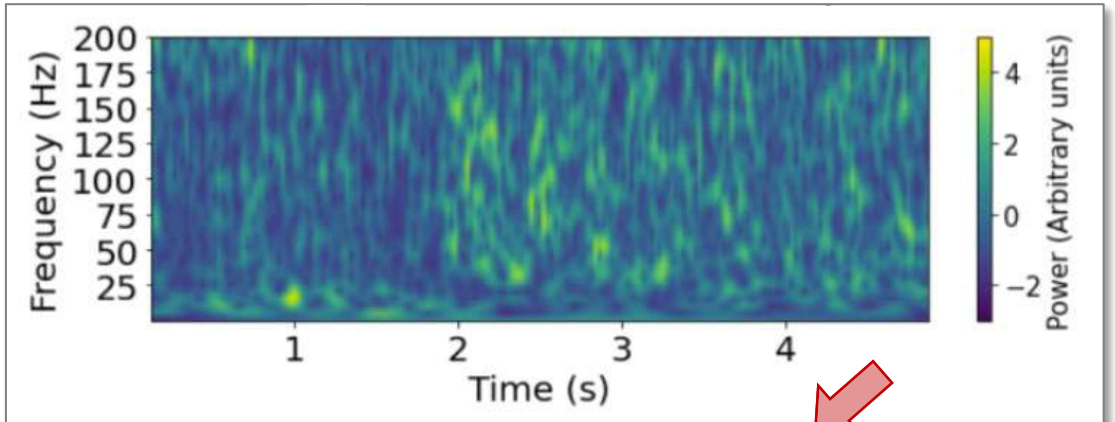
Electrode Placements



Input: SEEG

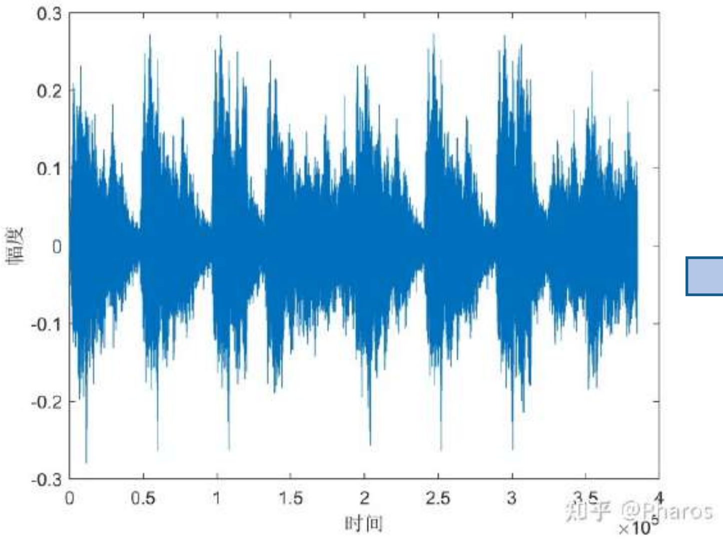


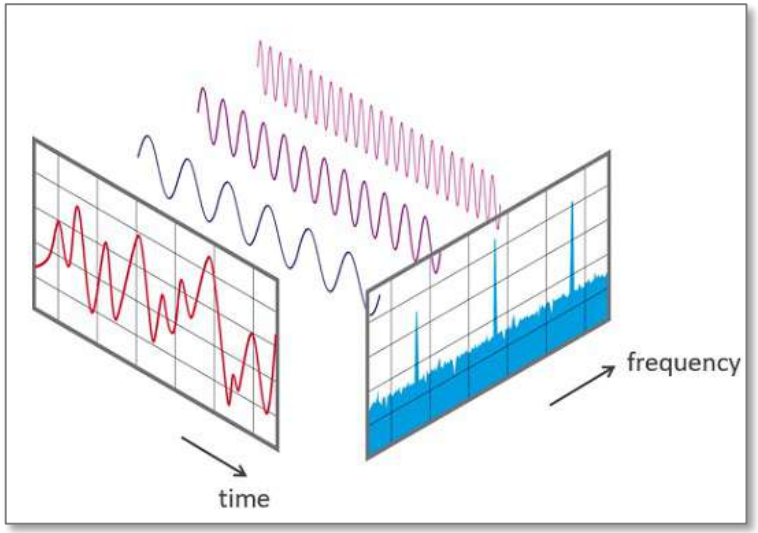
Voltage at one electrode

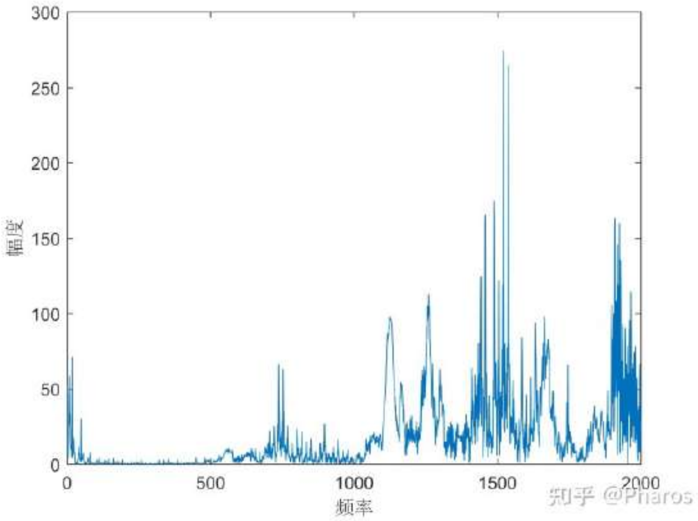


Time Frequency Representation (Superlet)

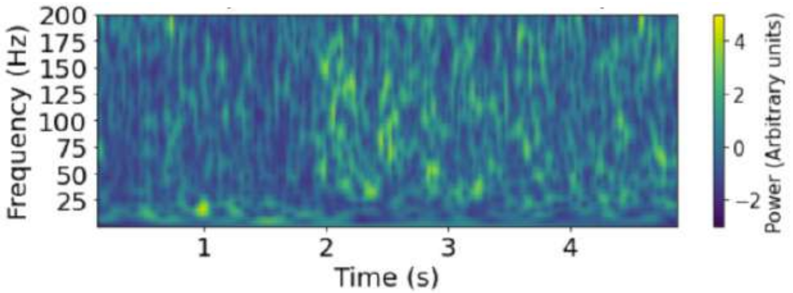
Fourier Transform (傅里叶变换)

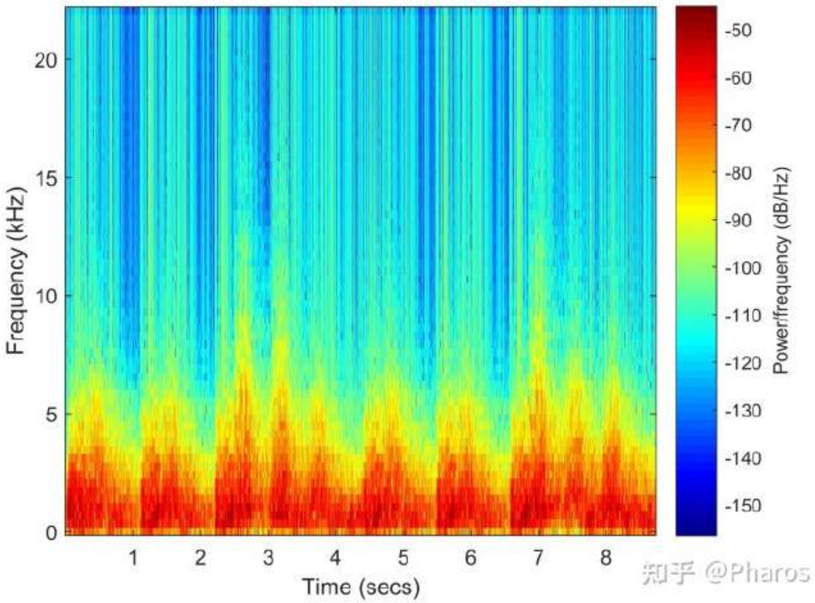


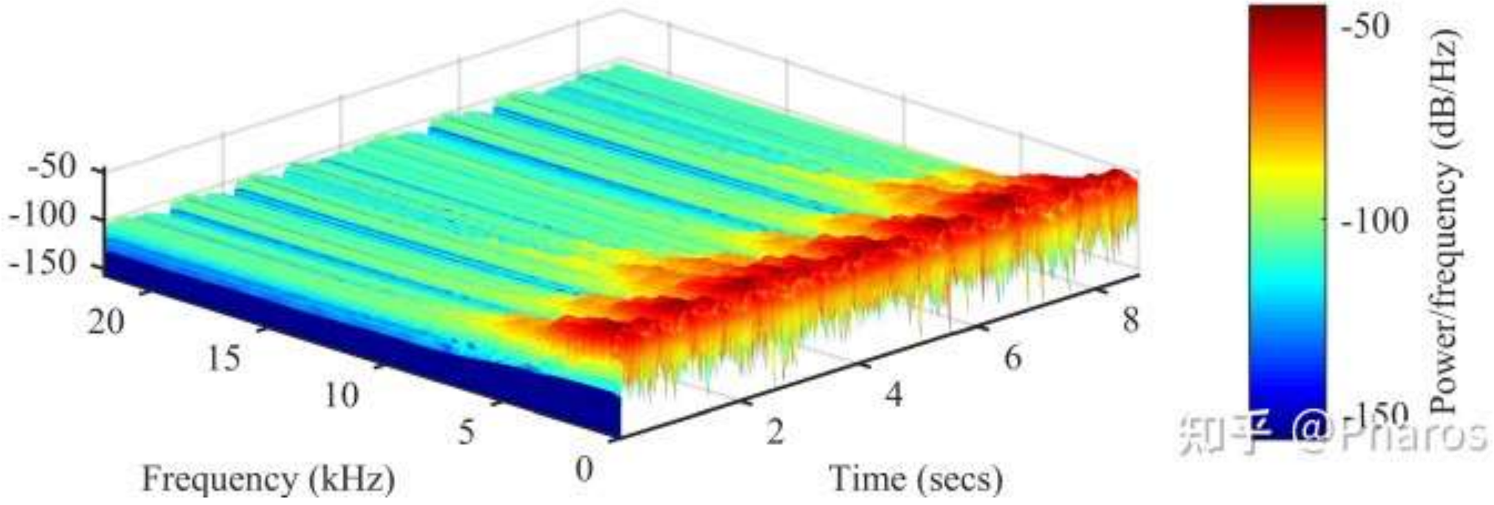




Input: Time-Frequency Representation





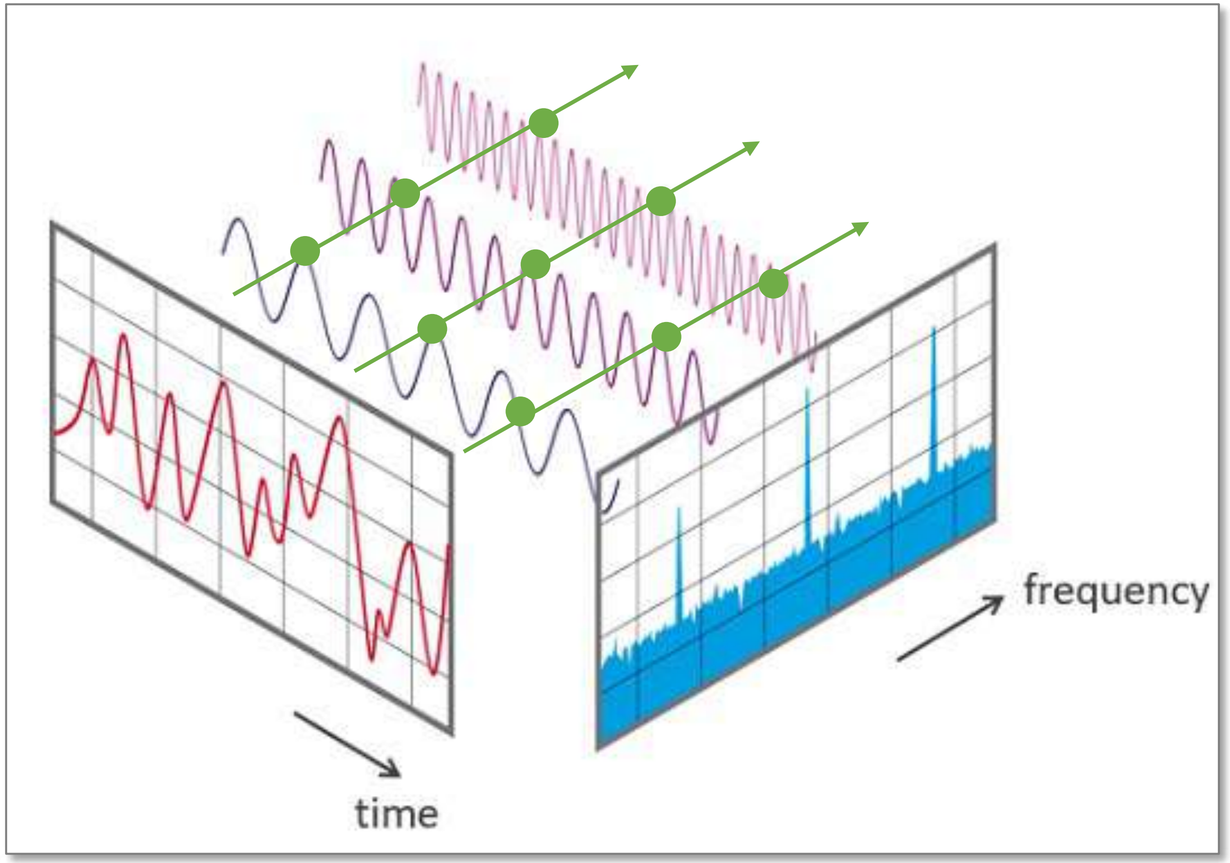


$$Y \in \mathbb{R}^{n \times m}$$

- ***n*** frequency channels
- ***m*** time frames

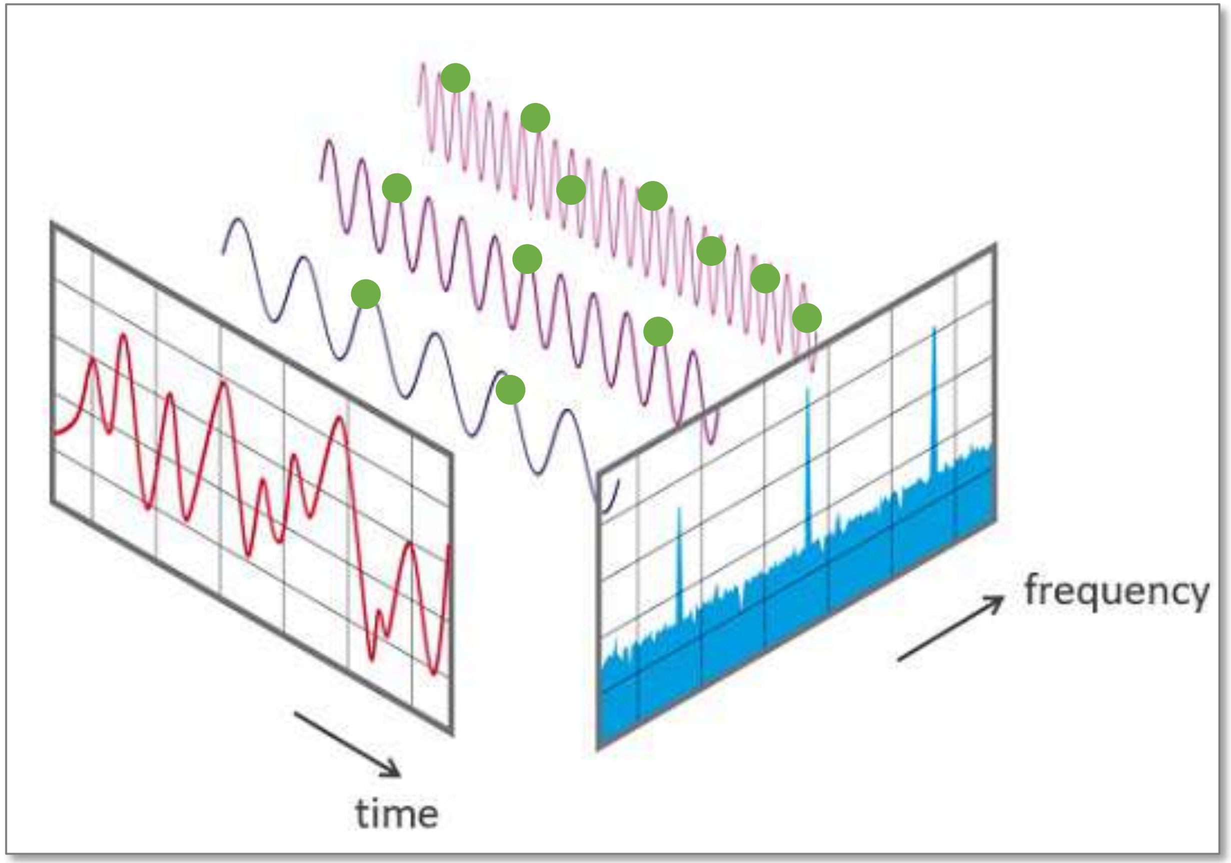
Short-Time Fourier Transform (STFT)

Temporal resolution is fixed for all frequencies



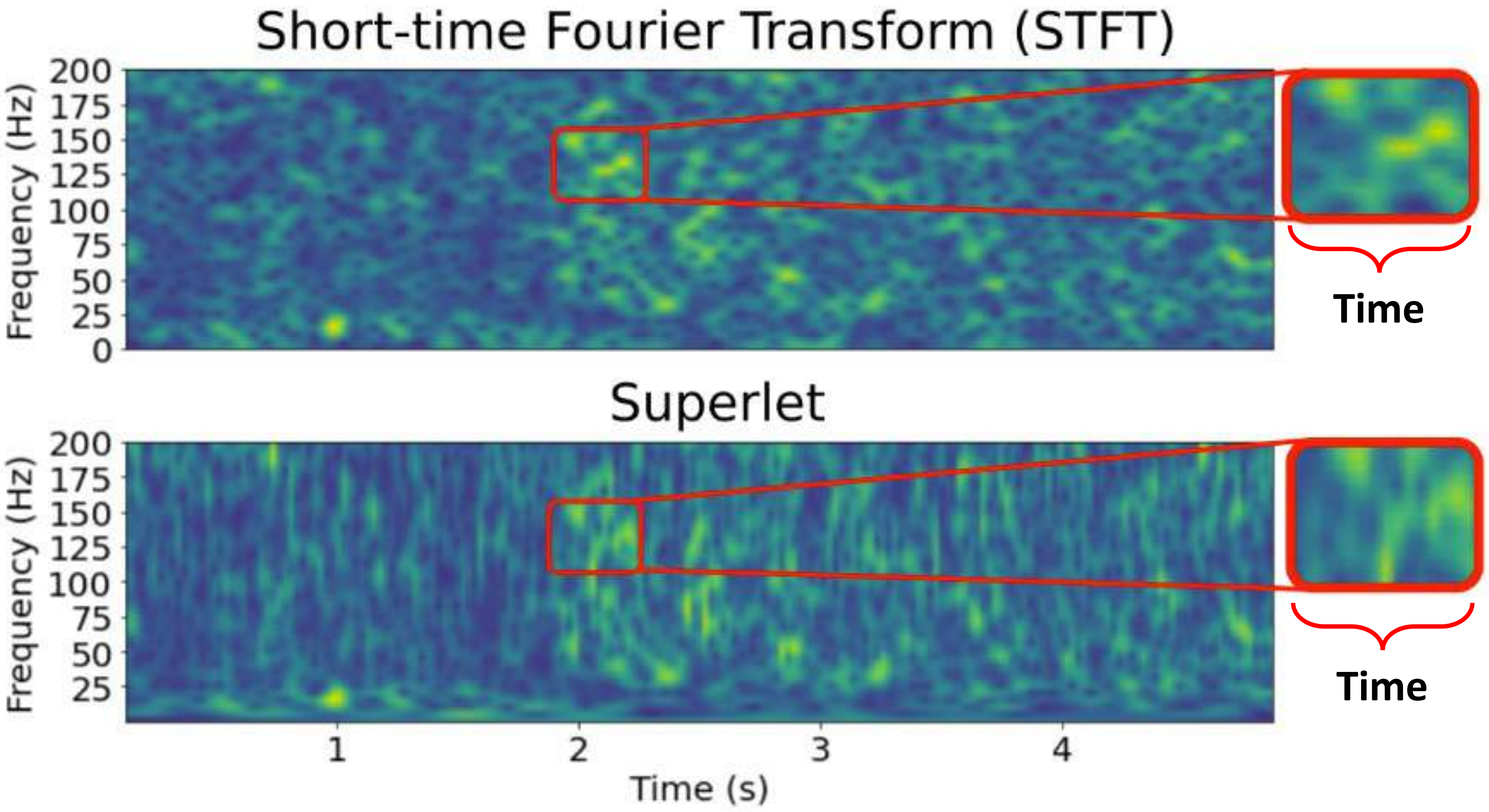
Superlet Transform

Temporal resolution increases with frequency

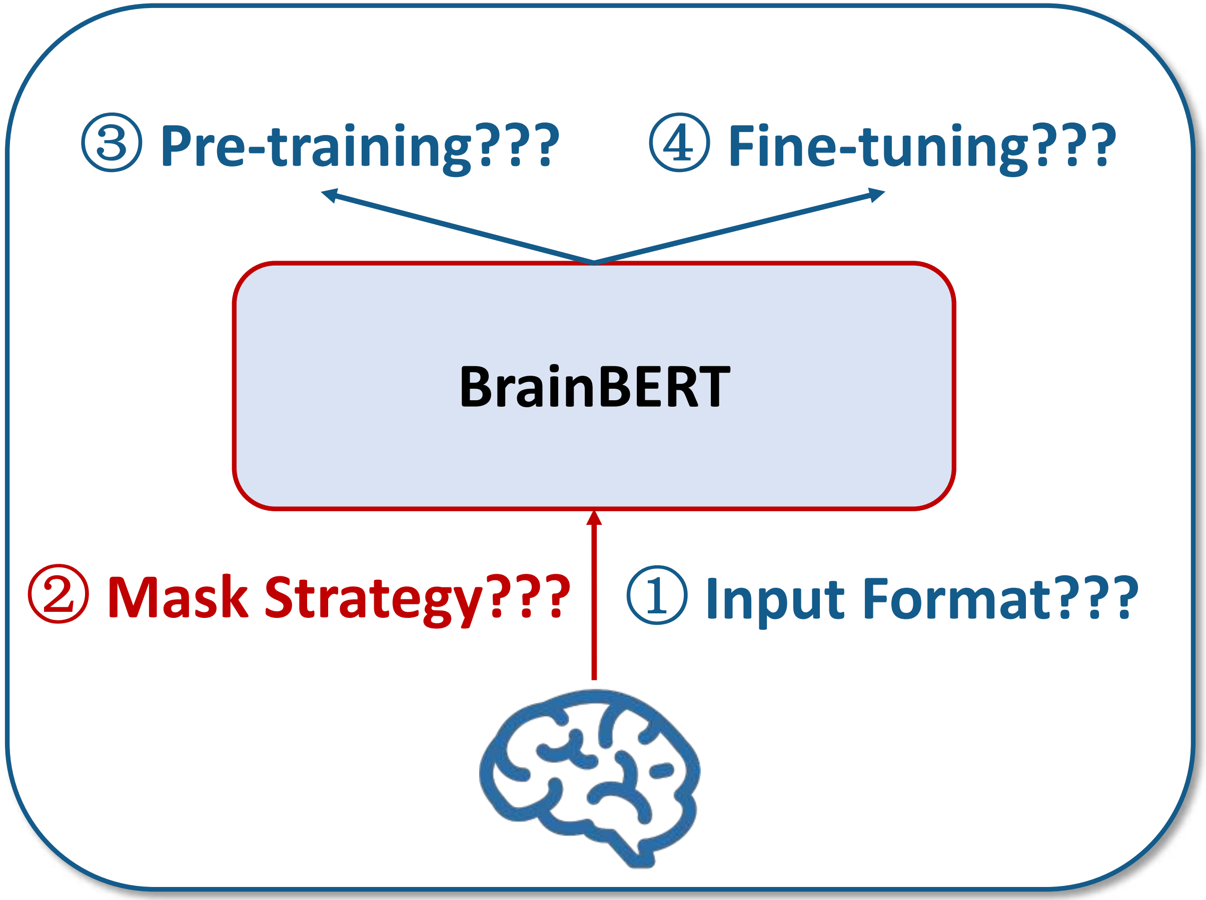


(Moca et al., 2021)

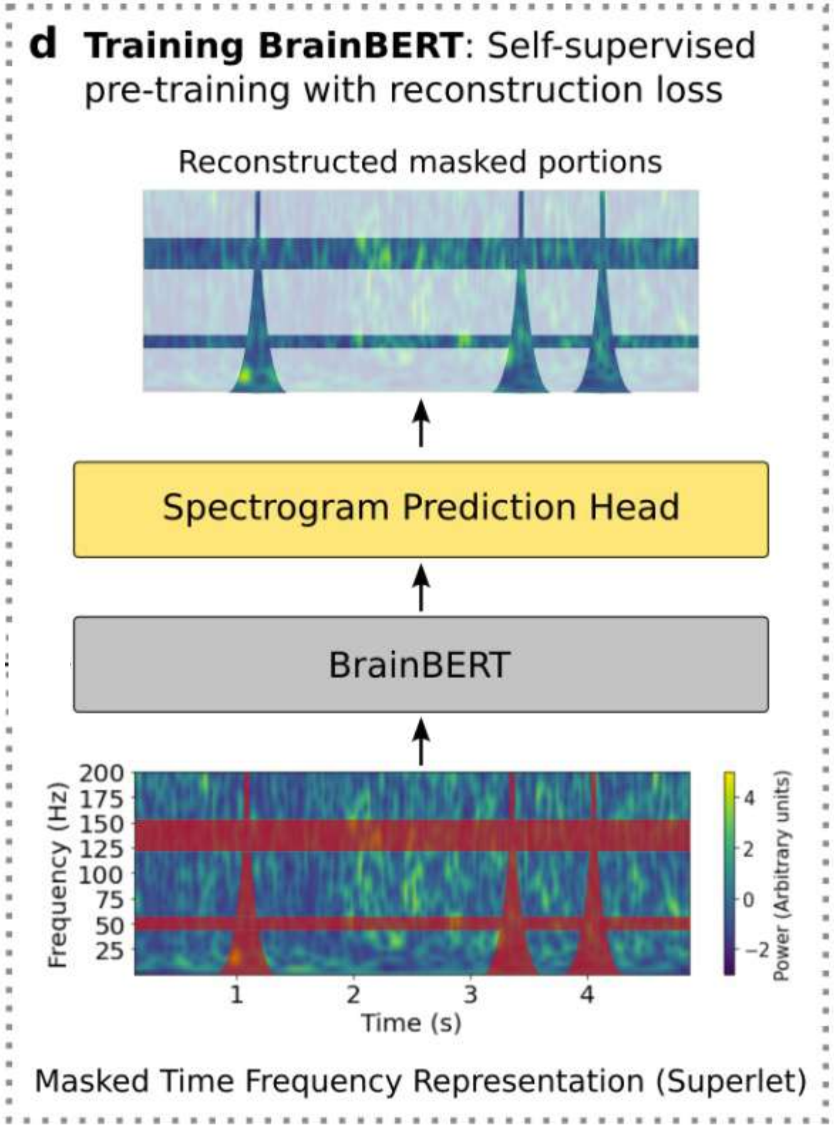
Two methods: STFT and Superlet Transform



BrainBERT


③ Pre-training???
④ Fine-tuning???
BrainBERT
② Mask Strategy???
① Input Format???

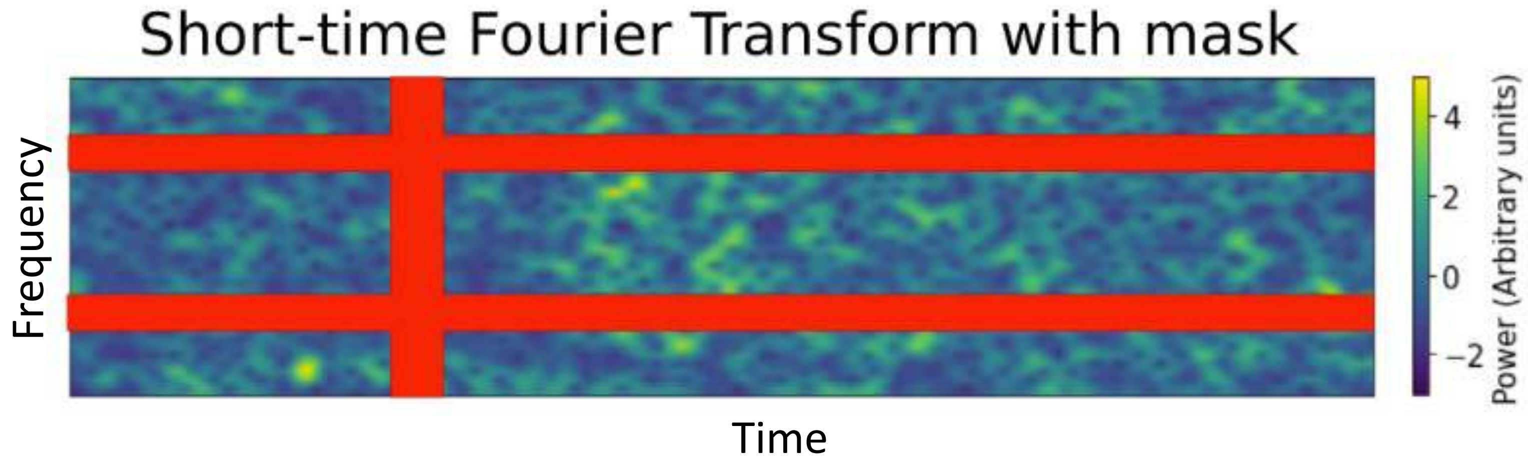
Mask



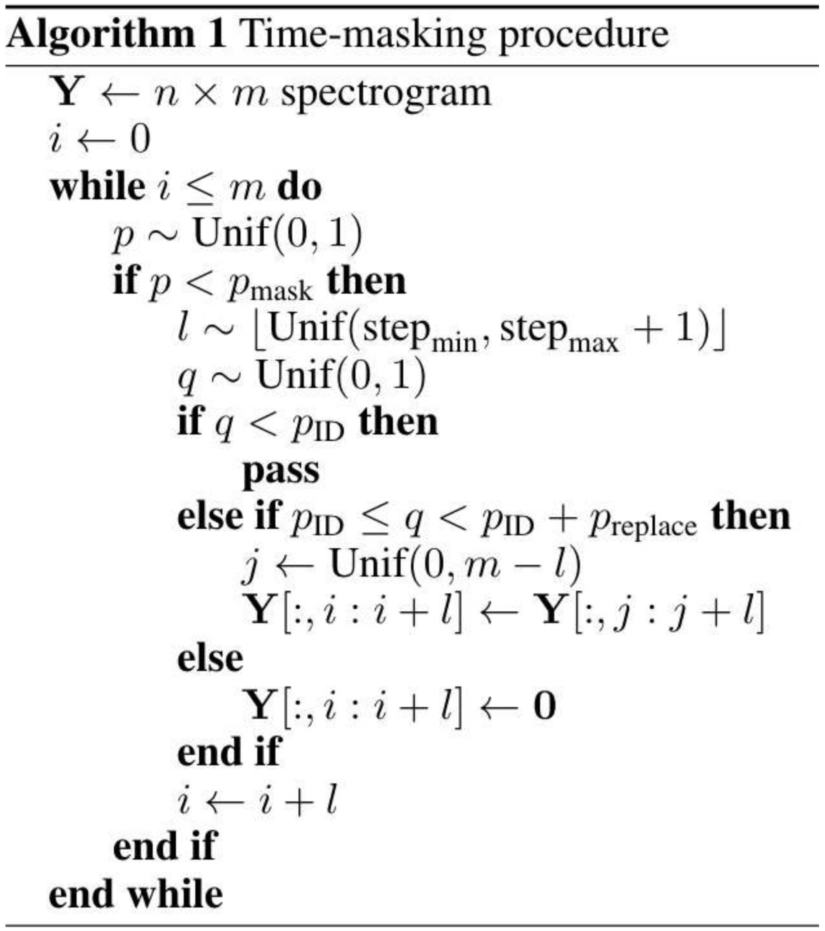
d **Training BrainBERT**: Self-supervised pre-training with reconstruction loss

Masked Time Frequency Representation (Superlet)

Static masking for STFT



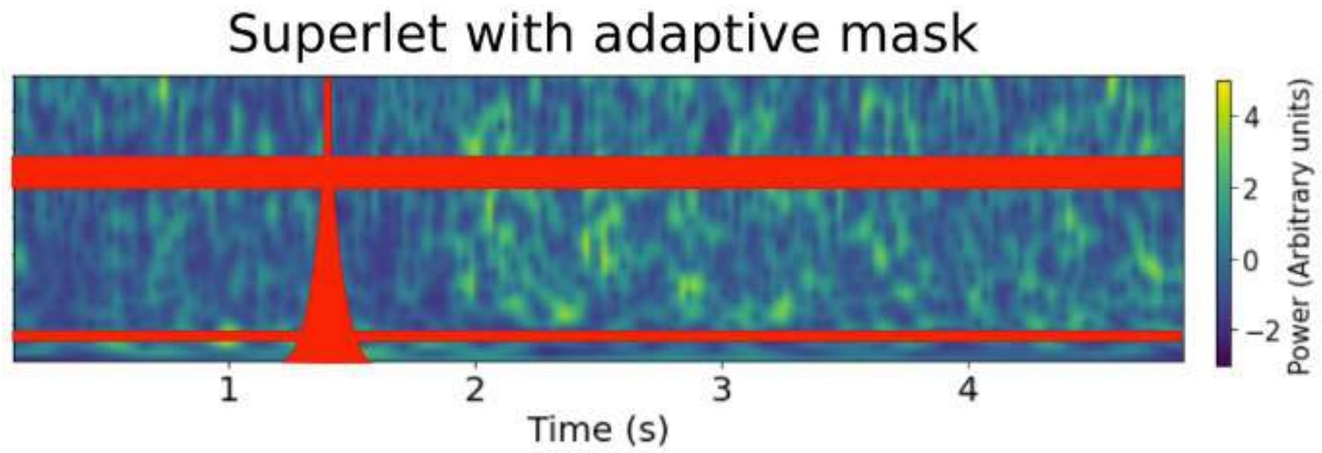
- Randomly chosen time and frequency intervals.
- The width of each time-mask is a randomly chosen integer from the range $[\mathrm{step}_{min}^{time}, \mathrm{step}_{max}^{time}]$
- The width of each frequence-mask is a randomly chosen integer from the range $[\mathrm{step}_{min}^{freq}, \mathrm{step}_{max}^{freq}]$



Algorithm 1 Time-masking procedure

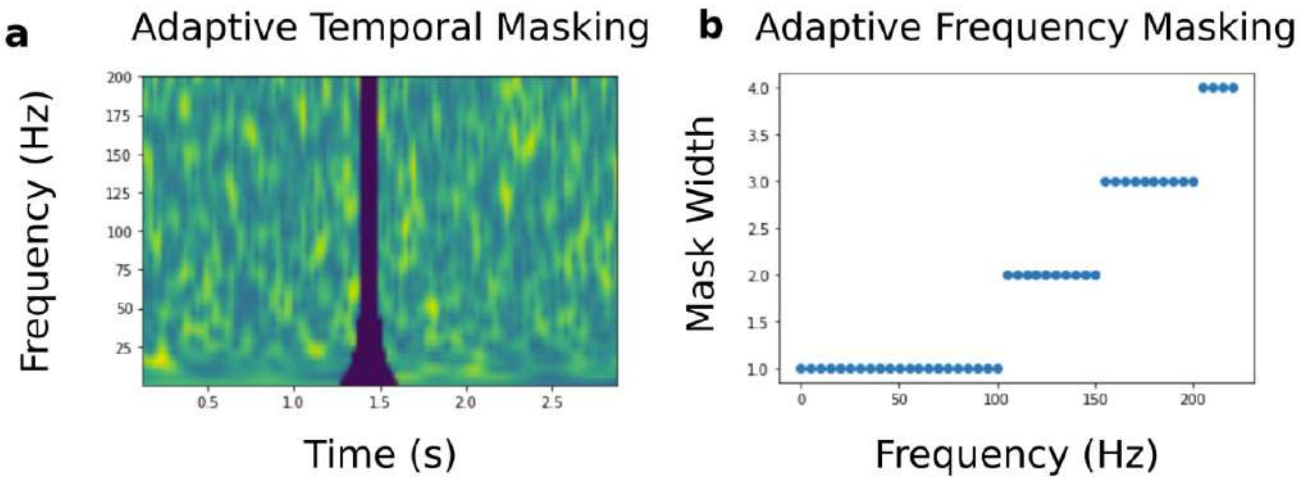
```
$\mathbf{Y} \leftarrow n \times m$ spectrogram
$i \leftarrow 0$
while $i \leq m$ do
    $p \sim \mathrm{Unif}(0, 1)$
    if $p < p_{\mathrm{mask}}$ then
        $l \sim \lfloor \mathrm{Unif}(\mathrm{step}_{\mathrm{min}}, \mathrm{step}_{\mathrm{max}} + 1) \rfloor$
        $q \sim \mathrm{Unif}(0, 1)$
        if $q < p_{\mathrm{ID}}$ then
            pass
        else if $p_{\mathrm{ID}} \leq q < p_{\mathrm{ID}} + p_{\mathrm{replace}}$ then
            $j \leftarrow \mathrm{Unif}(0, m - l)$
            $\mathbf{Y}[:, i : i + l] \leftarrow \mathbf{Y}[:, j : j + l]$
        else
            $\mathbf{Y}[:, i : i + l] \leftarrow \mathbf{0}$
        end if
        $i \leftarrow i + l$
    end if
end while
```

Adaptive Masking for Superlet Transform

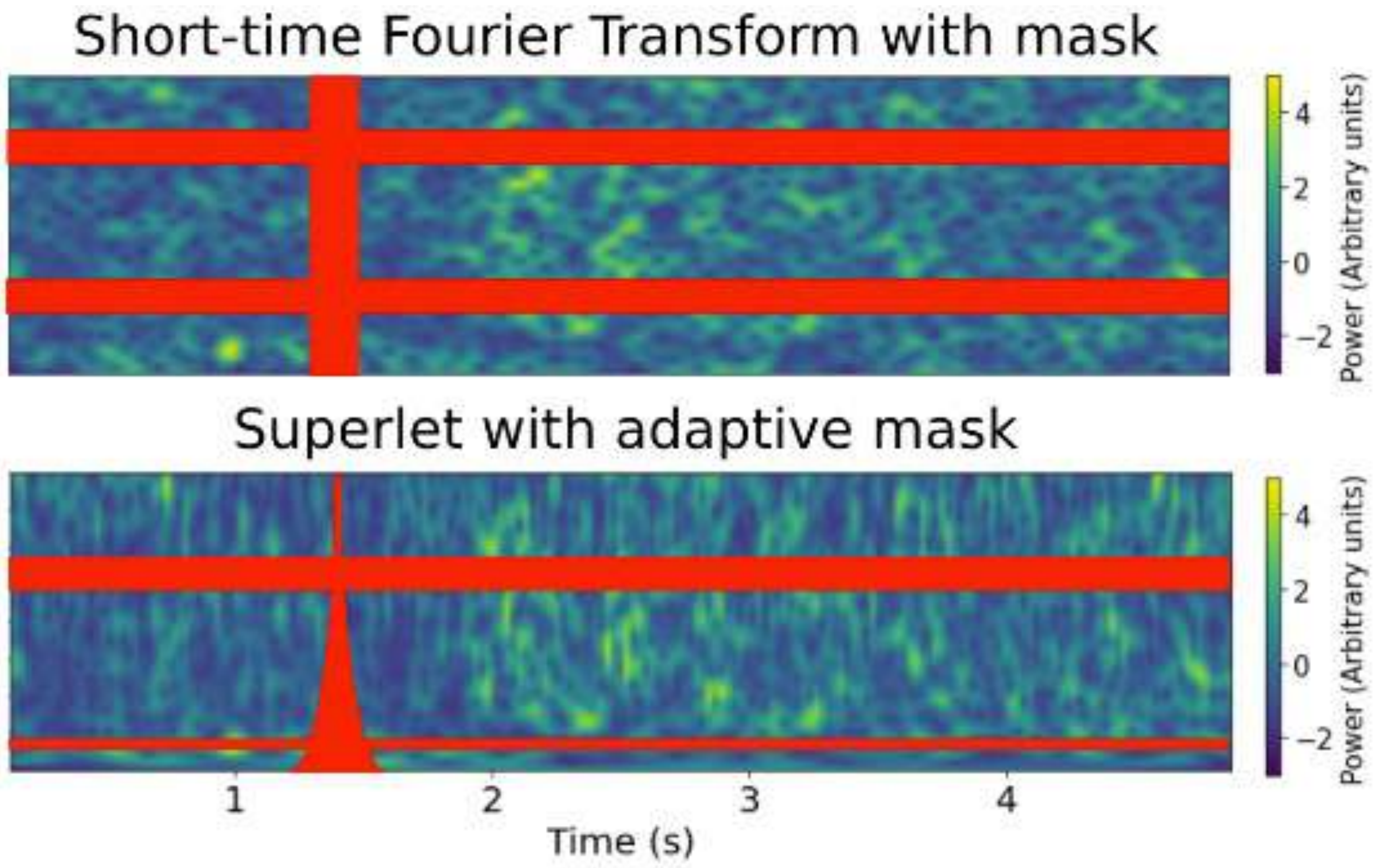


$$w_t(f) = 2\max\left(\mathbf{m}, \frac{200}{20+\boldsymbol{f}}\right)$$

$$w_f(f) = \max\left(1, \left\lfloor \frac{4.9\boldsymbol{f}}{250} \right\rfloor\right)$$



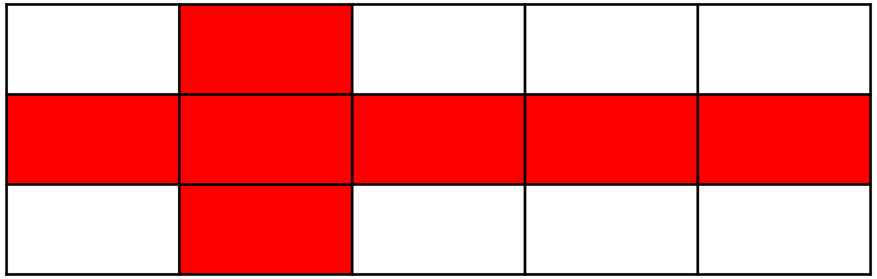
Masking Strategy for Two Methods



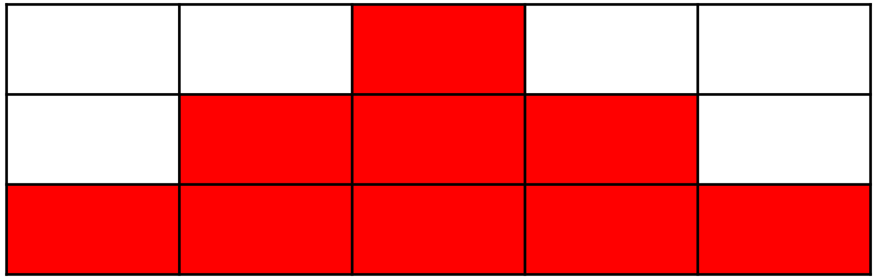
Masking Strategy for Two Methods

$$Y \in \mathbb{R}^{n \times m}$$

- ***n*** frequency channels
- ***m*** time frames

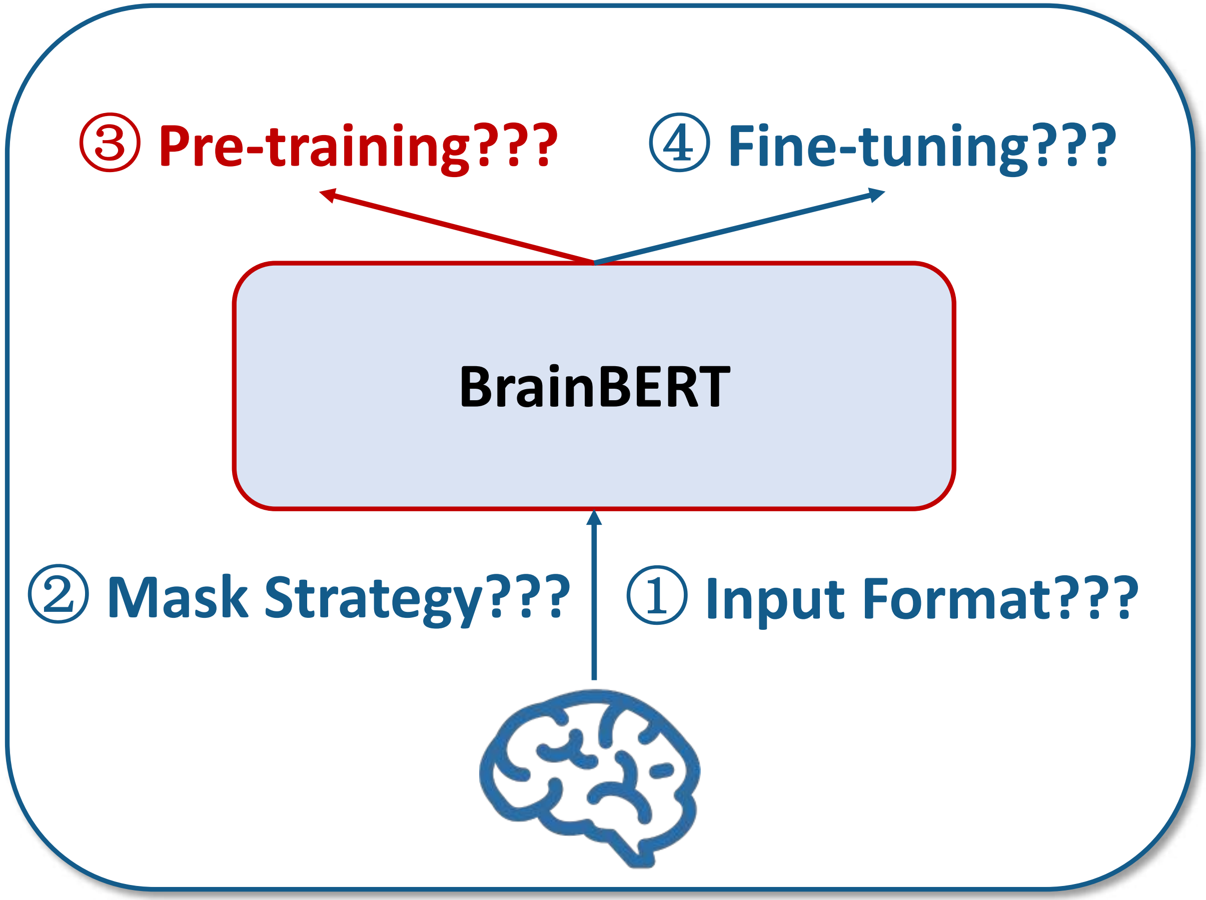


STFT



Superlet

BrainBERT



③ Pre-training???

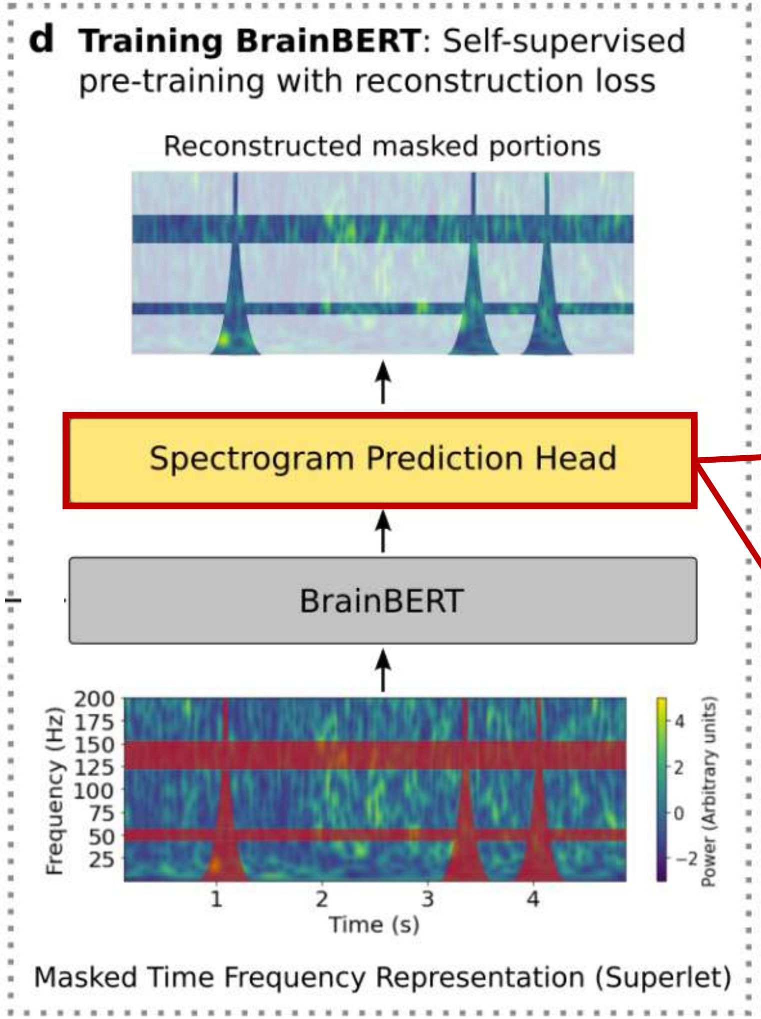
④ Fine-tuning???

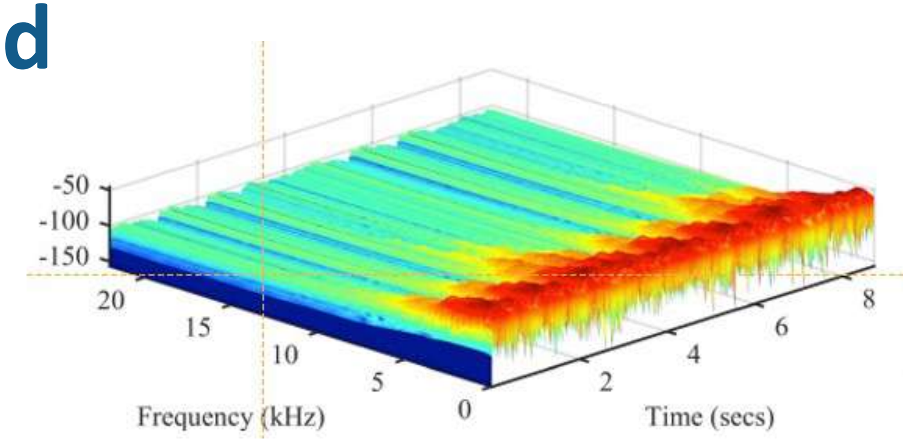
BrainBERT

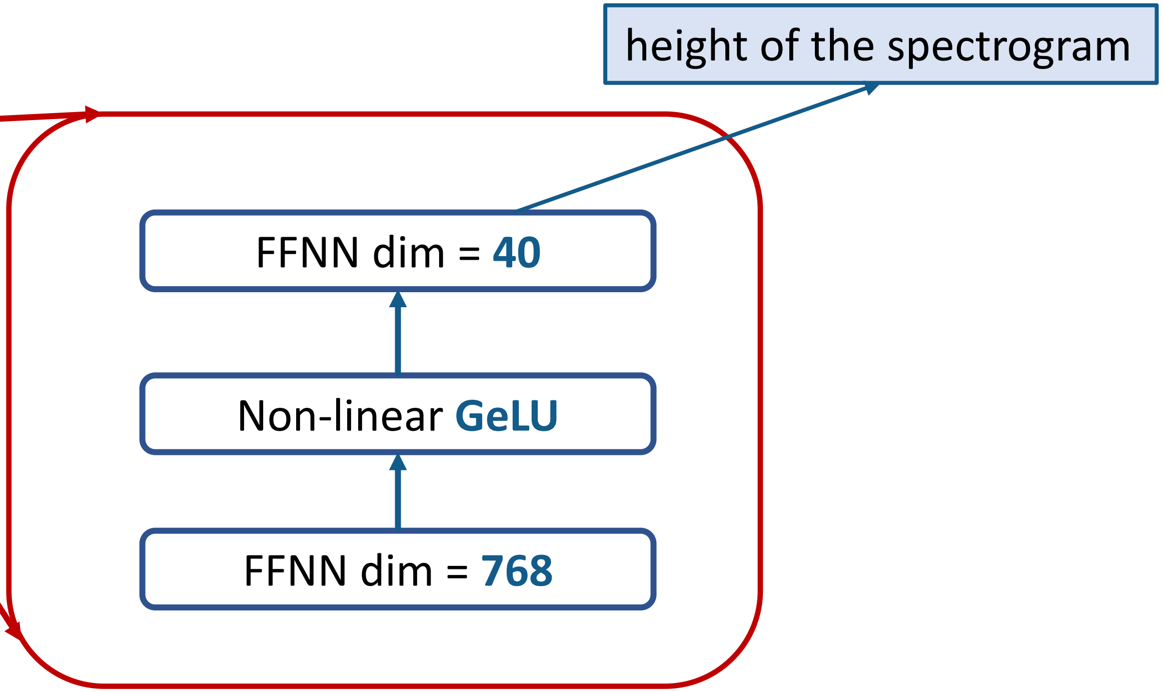
② Mask Strategy???

① Input Format???

Spectrogram Prediction Head


d Training BrainBERT: Self-supervised pre-training with reconstruction loss
Reconstructed masked portions
Spectrogram Prediction Head
BrainBERT
Frequency (Hz)
200
175
150
125
100
75
50
25
1
2
3
4
Time (s)
4
2
0
−2
Power (Arbitrary units)
Masked Time Frequency Representation (Superlet)





height of the spectrogram

FFNN dim = **40**

Non-linear **GeLU**

FFNN dim = **768**

Pretraining Loss

- **L1 reconstruction loss**

$$\mathcal{L}_L = \frac{1}{|M|}\sum_{(i,j)\in M}\left|\mathbf{Y}_{i,j}-\hat{\mathbf{Y}}_{i,j}\right|$$

M is the set of masked spectrogram positions

Since the spectrogram is z-scored along the time-axis, approximately 68% of the z-scored spectrogram is 0 or < 1.

- **Content aware loss**

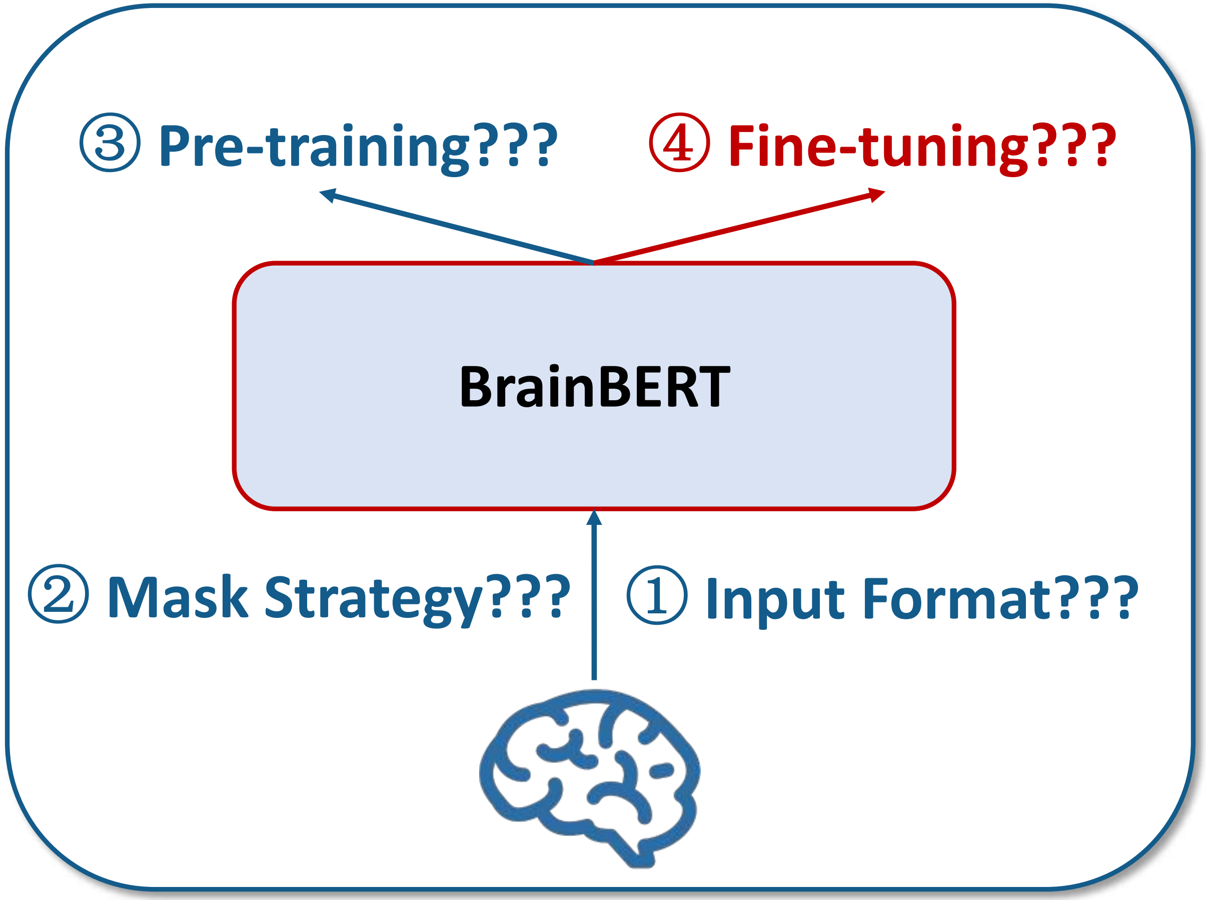
$$\mathcal{L}_C = \frac{1}{|\{(i,j) \mid \mathbf{Y}_{i,j}>\gamma\}|}\sum_{(i,j)|(i,j)\in M,\mathbf{Y}_{i,j}>\gamma}\left|\mathbf{Y}_{i,j}-\hat{\mathbf{Y}}_{i,j}\right|$$

$$\boxed{\mathcal{L}=\mathcal{L}_L+\alpha\mathcal{L}_C}$$

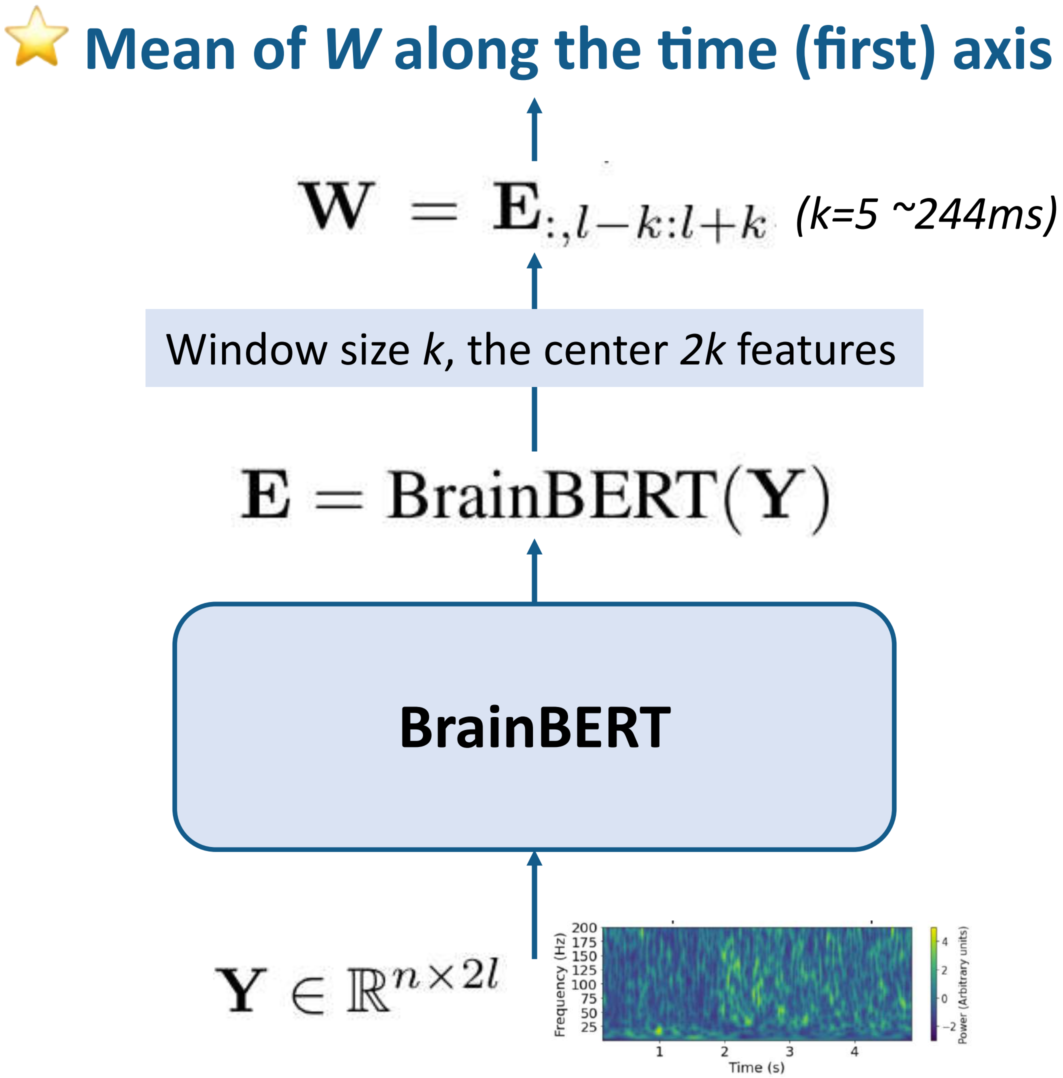
Data

- **10** subjects (5 male, 5 female; aged 4-19, µ 11.9, σ 4.6) with pharmacologically intractable epilepsy.
- **4.37 hours** of data were collected from each subject;
- Subjects watched a feature length movie in a quiet room while their neural data was recorded at a rate of **2kHz**.
- Across all subjects, data was recorded from a total of **1,688 electrodes**, with a mean of **167 electrodes** per subject
- During pretraining, data from all subjects and electrodes is **segmented into 5s** intervals, and all segments are combined into a single training pool.
- For pretraining purposes, neural recordings from **19 of the sessions** was selected, and the remaining **7 sessions** were held out to evaluate performance on decoding tasks.
- All **ten** subjects are represented in the pretraining data.

BrainBERT


③ Pre-training???
④ Fine-tuning???
BrainBERT
② Mask Strategy???
① Input Format???

Feature Extraction



Mean of *W* along the time (first) axis

$$\mathbf{W} = \mathbf{E}_{:,l-k:l+k}$$ *(k=5 ~244ms)*

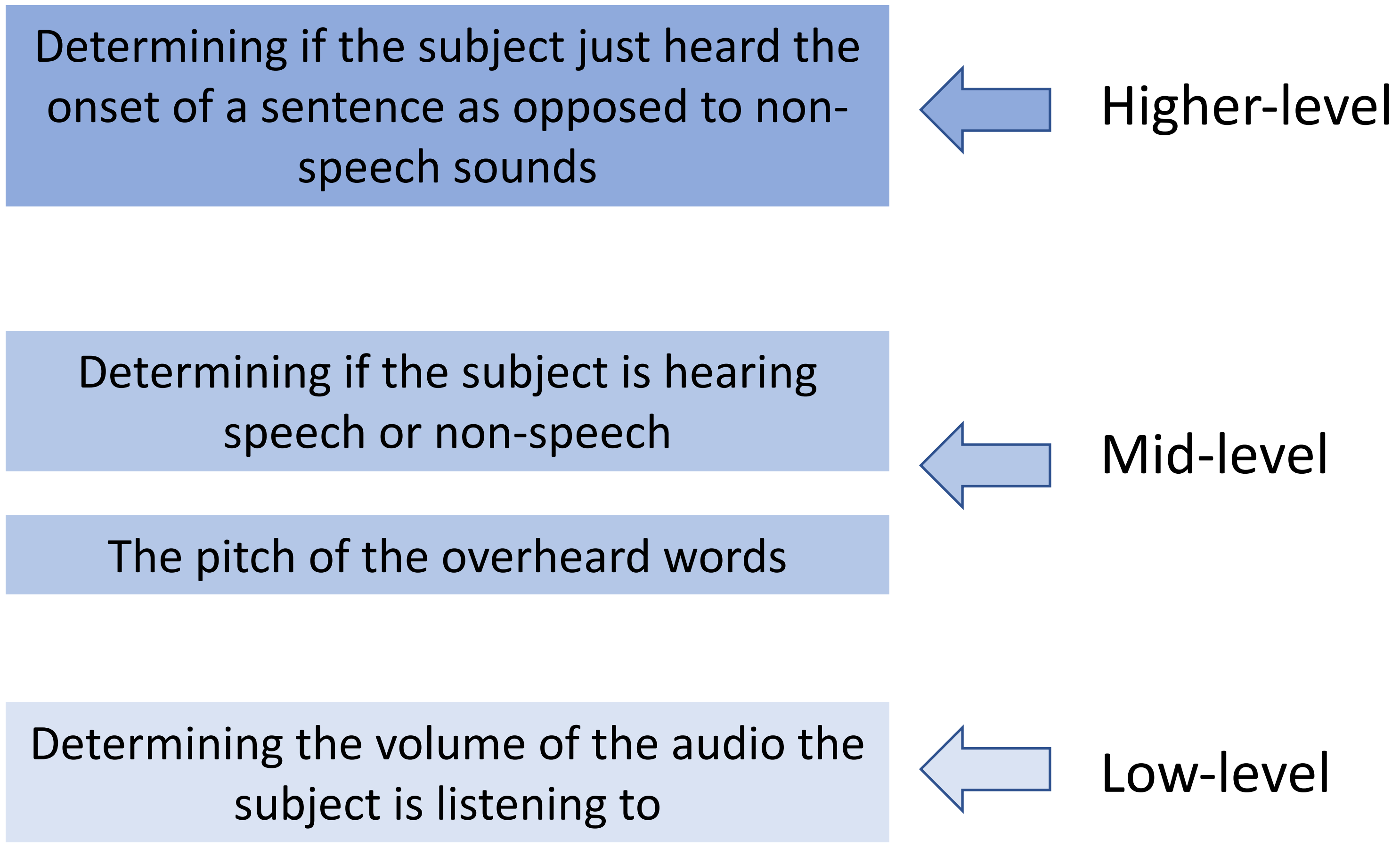
Window size *k*, the center *2k* features

$$\mathbf{E} = \mathrm{BrainBERT}(\mathbf{Y})$$

BrainBERT

$$\mathbf{Y} \in \mathbb{R}^{n \times 2l}$$

Experiments: classification tasks



Determining if the subject just heard the onset of a sentence as opposed to non-speech sounds

Higher-level

Determining if the subject is hearing speech or non-speech

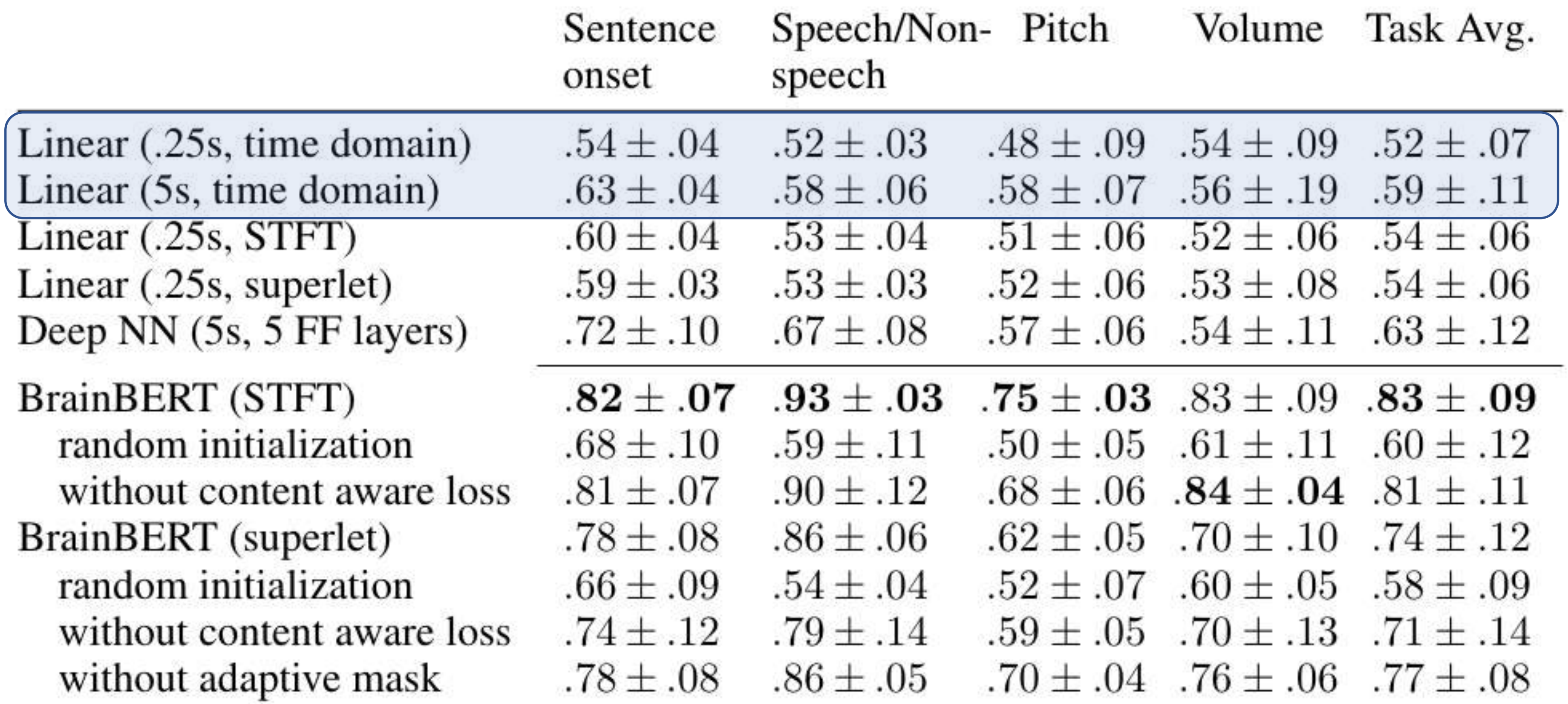
The pitch of the overheard words

Mid-level

Determining the volume of the audio the subject is listening to

Low-level

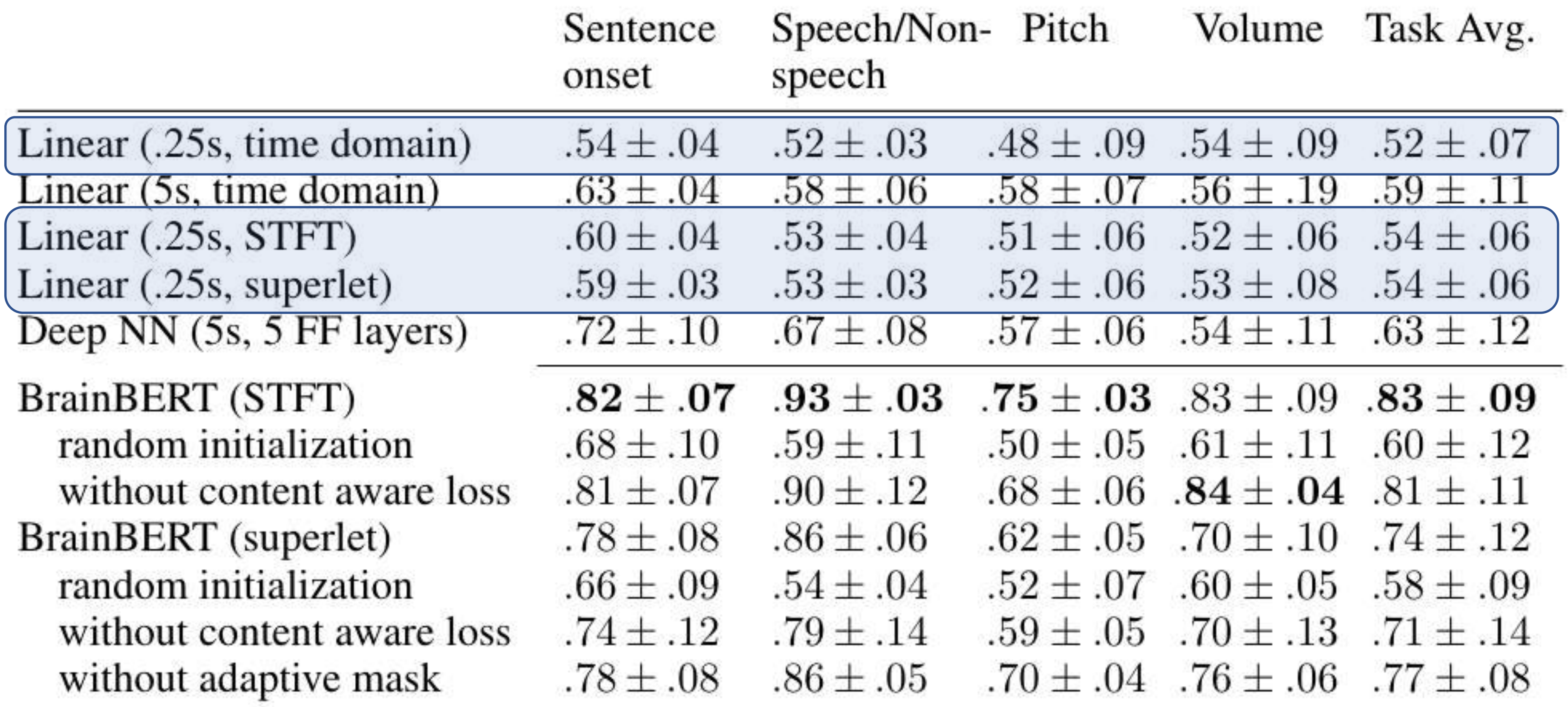
Main Results



	Sentence onset	Speech/Non-speech	Pitch	Volume	Task Avg.
Linear (.25s, time domain)	.54 ± .04	.52 ± .03	.48 ± .09	.54 ± .09	.52 ± .07
Linear (5s, time domain)	.63 ± .04	.58 ± .06	.58 ± .07	.56 ± .19	.59 ± .11
Linear (.25s, STFT)	.60 ± .04	.53 ± .04	.51 ± .06	.52 ± .06	.54 ± .06
Linear (.25s, superlet)	.59 ± .03	.53 ± .03	.52 ± .06	.53 ± .08	.54 ± .06
Deep NN (5s, 5 FF layers)	.72 ± .10	.67 ± .08	.57 ± .06	.54 ± .11	.63 ± .12
BrainBERT (STFT)	**.82 ± .07**	**.93 ± .03**	**.75 ± .03**	.83 ± .09	**.83 ± .09**
random initialization	.68 ± .10	.59 ± .11	.50 ± .05	.61 ± .11	.60 ± .12
without content aware loss	.81 ± .07	.90 ± .12	.68 ± .06	**.84 ± .04**	.81 ± .11
BrainBERT (superlet)	.78 ± .08	.86 ± .06	.62 ± .05	.70 ± .10	.74 ± .12
random initialization	.66 ± .09	.54 ± .04	.52 ± .07	.60 ± .05	.58 ± .09
without content aware loss	.74 ± .12	.79 ± .14	.59 ± .05	.70 ± .13	.71 ± .14
without adaptive mask	.78 ± .08	.86 ± .05	.70 ± .04	.76 ± .06	.77 ± .08

🔥**BrainBERT**

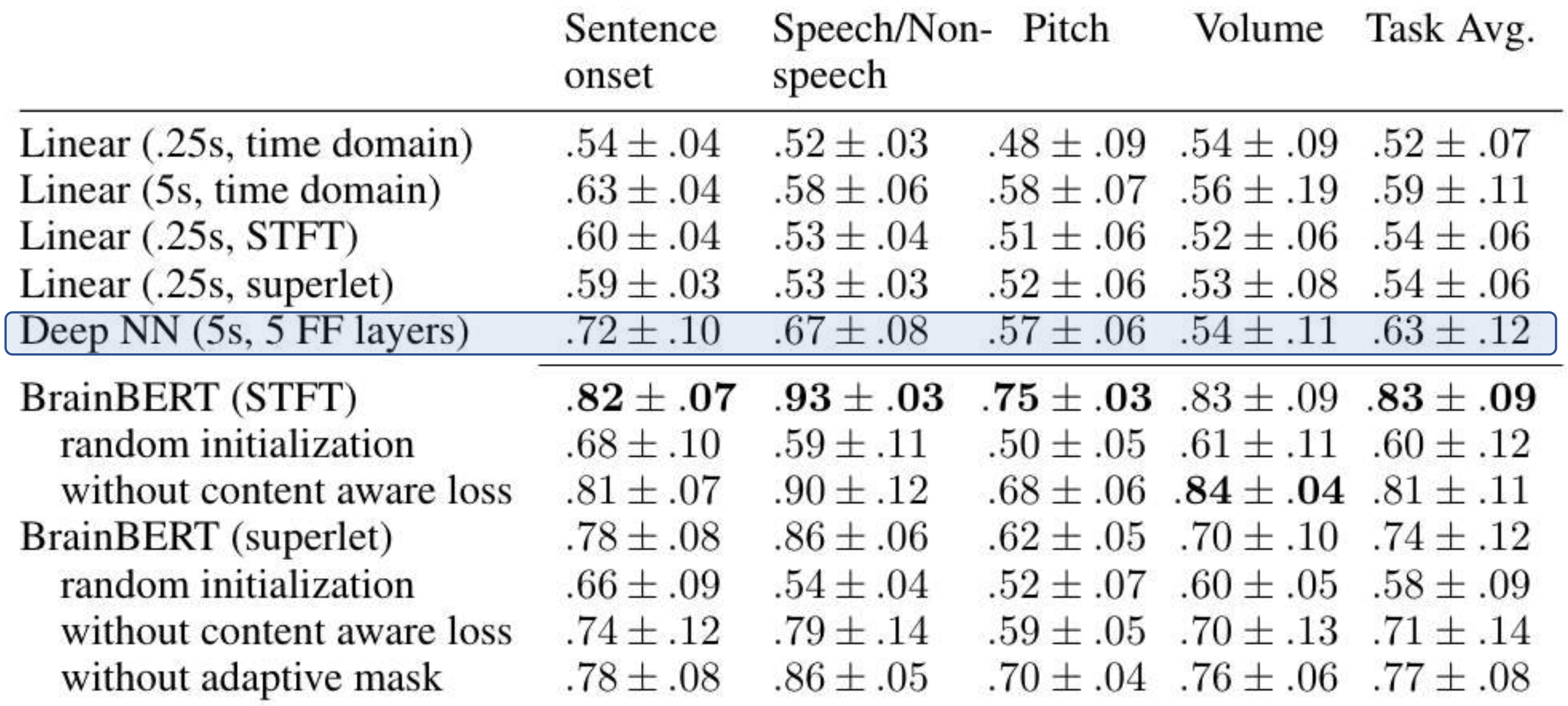
Main Results



	Sentence onset	Speech/Non-speech	Pitch	Volume	Task Avg.
Linear (.25s, time domain)	.54 ± .04	.52 ± .03	.48 ± .09	.54 ± .09	.52 ± .07
Linear (5s, time domain)	.63 ± .04	.58 ± .06	.58 ± .07	.56 ± .19	.59 ± .11
Linear (.25s, STFT)	.60 ± .04	.53 ± .04	.51 ± .06	.52 ± .06	.54 ± .06
Linear (.25s, superlet)	.59 ± .03	.53 ± .03	.52 ± .06	.53 ± .08	.54 ± .06
Deep NN (5s, 5 FF layers)	.72 ± .10	.67 ± .08	.57 ± .06	.54 ± .11	.63 ± .12
BrainBERT (STFT)	**.82 ± .07**	**.93 ± .03**	**.75 ± .03**	.83 ± .09	**.83 ± .09**
random initialization	.68 ± .10	.59 ± .11	.50 ± .05	.61 ± .11	.60 ± .12
without content aware loss	.81 ± .07	.90 ± .12	.68 ± .06	**.84 ± .04**	.81 ± .11
BrainBERT (superlet)	.78 ± .08	.86 ± .06	.62 ± .05	.70 ± .10	.74 ± .12
random initialization	.66 ± .09	.54 ± .04	.52 ± .07	.60 ± .05	.58 ± .09
without content aware loss	.74 ± .12	.79 ± .14	.59 ± .05	.70 ± .13	.71 ± .14
without adaptive mask	.78 ± .08	.86 ± .05	.70 ± .04	.76 ± .06	.77 ± .08

🔥**BrainBERT**

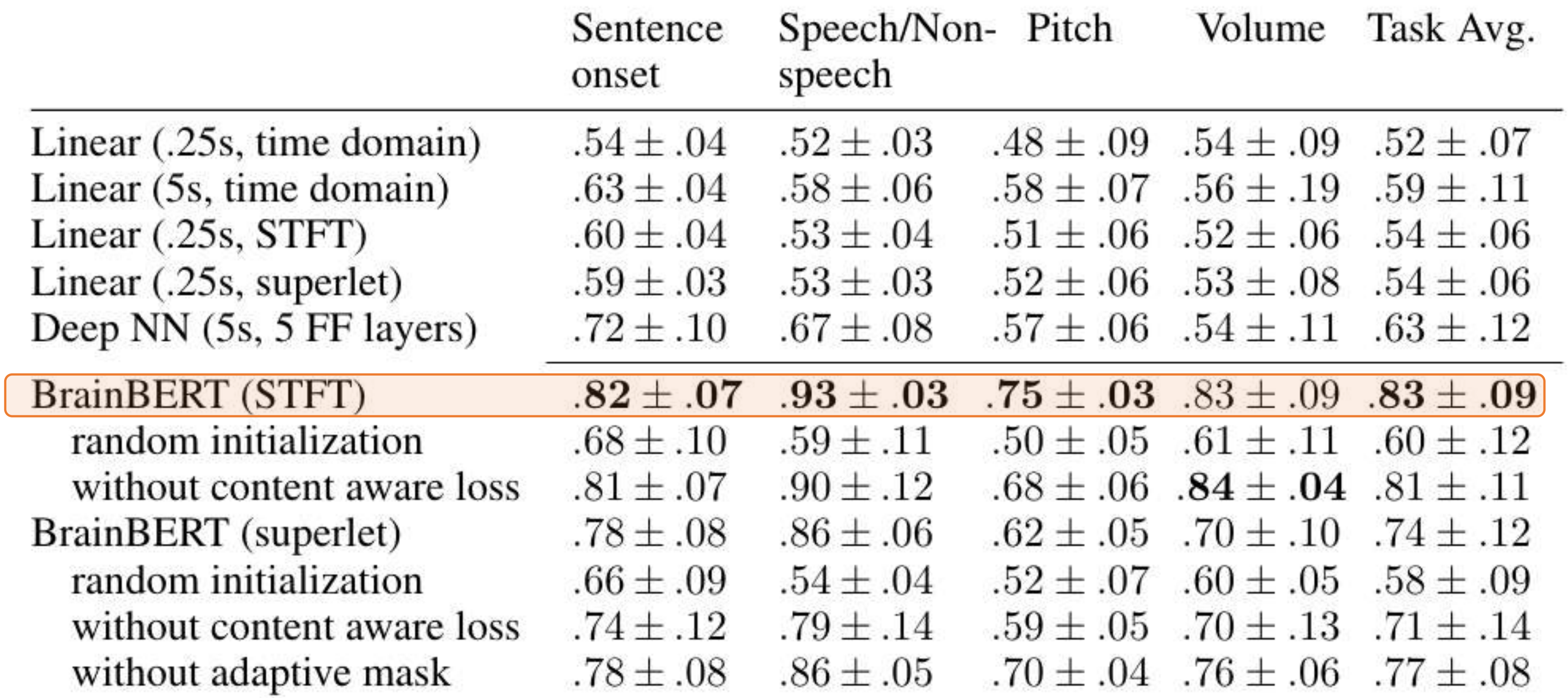
Main Results



	Sentence onset	Speech/Non-speech	Pitch	Volume	Task Avg.
Linear (.25s, time domain)	.54 ± .04	.52 ± .03	.48 ± .09	.54 ± .09	.52 ± .07
Linear (5s, time domain)	.63 ± .04	.58 ± .06	.58 ± .07	.56 ± .19	.59 ± .11
Linear (.25s, STFT)	.60 ± .04	.53 ± .04	.51 ± .06	.52 ± .06	.54 ± .06
Linear (.25s, superlet)	.59 ± .03	.53 ± .03	.52 ± .06	.53 ± .08	.54 ± .06
Deep NN (5s, 5 FF layers)	.72 ± .10	.67 ± .08	.57 ± .06	.54 ± .11	.63 ± .12
BrainBERT (STFT)	**.82 ± .07**	**.93 ± .03**	**.75 ± .03**	.83 ± .09	**.83 ± .09**
random initialization	.68 ± .10	.59 ± .11	.50 ± .05	.61 ± .11	.60 ± .12
without content aware loss	.81 ± .07	.90 ± .12	.68 ± .06	**.84 ± .04**	.81 ± .11
BrainBERT (superlet)	.78 ± .08	.86 ± .06	.62 ± .05	.70 ± .10	.74 ± .12
random initialization	.66 ± .09	.54 ± .04	.52 ± .07	.60 ± .05	.58 ± .09
without content aware loss	.74 ± .12	.79 ± .14	.59 ± .05	.70 ± .13	.71 ± .14
without adaptive mask	.78 ± .08	.86 ± .05	.70 ± .04	.76 ± .06	.77 ± .08

🔥**BrainBERT**

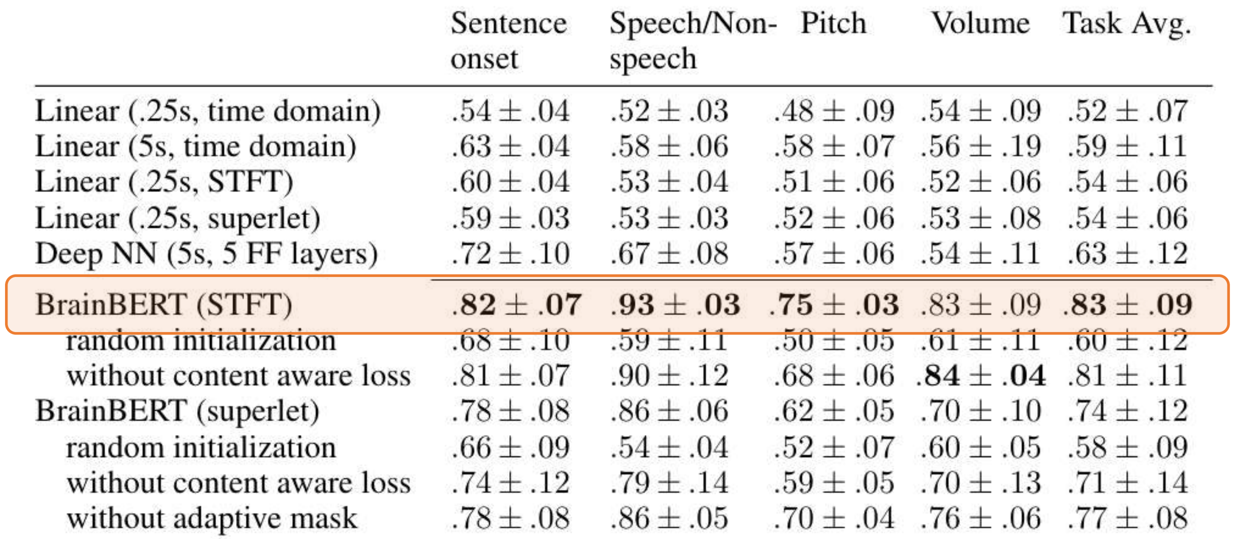
Main Results



	Sentence onset	Speech/Non-speech	Pitch	Volume	Task Avg.
Linear (.25s, time domain)	$.54 \pm .04$	$.52 \pm .03$	$.48 \pm .09$	$.54 \pm .09$	$.52 \pm .07$
Linear (5s, time domain)	$.63 \pm .04$	$.58 \pm .06$	$.58 \pm .07$	$.56 \pm .19$	$.59 \pm .11$
Linear (.25s, STFT)	$.60 \pm .04$	$.53 \pm .04$	$.51 \pm .06$	$.52 \pm .06$	$.54 \pm .06$
Linear (.25s, superlet)	$.59 \pm .03$	$.53 \pm .03$	$.52 \pm .06$	$.53 \pm .08$	$.54 \pm .06$
Deep NN (5s, 5 FF layers)	$.72 \pm .10$	$.67 \pm .08$	$.57 \pm .06$	$.54 \pm .11$	$.63 \pm .12$
BrainBERT (STFT)	$\mathbf{.82 \pm .07}$	$\mathbf{.93 \pm .03}$	$\mathbf{.75 \pm .03}$	$.83 \pm .09$	$\mathbf{.83 \pm .09}$
random initialization	$.68 \pm .10$	$.59 \pm .11$	$.50 \pm .05$	$.61 \pm .11$	$.60 \pm .12$
without content aware loss	$.81 \pm .07$	$.90 \pm .12$	$.68 \pm .06$	$\mathbf{.84 \pm .04}$	$.81 \pm .11$
BrainBERT (superlet)	$.78 \pm .08$	$.86 \pm .06$	$.62 \pm .05$	$.70 \pm .10$	$.74 \pm .12$
random initialization	$.66 \pm .09$	$.54 \pm .04$	$.52 \pm .07$	$.60 \pm .05$	$.58 \pm .09$
without content aware loss	$.74 \pm .12$	$.79 \pm .14$	$.59 \pm .05$	$.70 \pm .13$	$.71 \pm .14$
without adaptive mask	$.78 \pm .08$	$.86 \pm .05$	$.70 \pm .04$	$.76 \pm .06$	$.77 \pm .08$

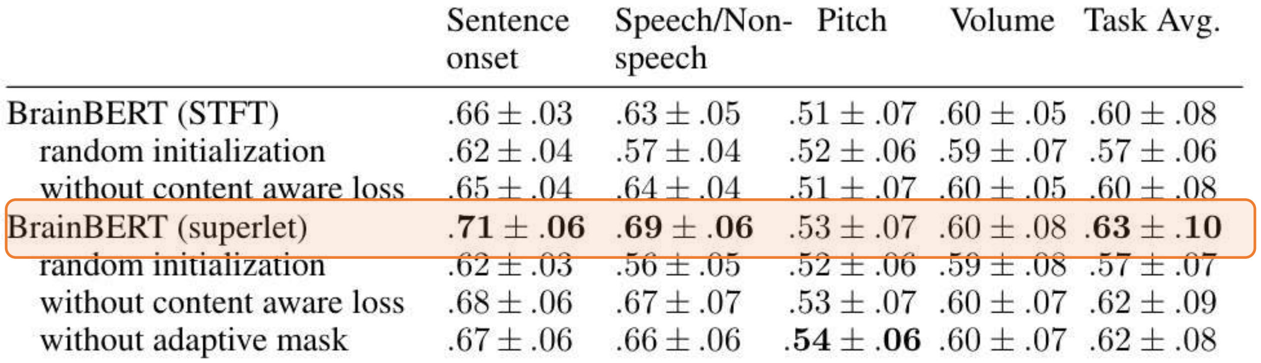
🔥**BrainBERT**

Main Results



	Sentence onset	Speech/Non-speech	Pitch	Volume	Task Avg.
Linear (.25s, time domain)	.54 ± .04	.52 ± .03	.48 ± .09	.54 ± .09	.52 ± .07
Linear (5s, time domain)	.63 ± .04	.58 ± .06	.58 ± .07	.56 ± .19	.59 ± .11
Linear (.25s, STFT)	.60 ± .04	.53 ± .04	.51 ± .06	.52 ± .06	.54 ± .06
Linear (.25s, superlet)	.59 ± .03	.53 ± .03	.52 ± .06	.53 ± .08	.54 ± .06
Deep NN (5s, 5 FF layers)	.72 ± .10	.67 ± .08	.57 ± .06	.54 ± .11	.63 ± .12
BrainBERT (STFT)	**.82 ± .07**	**.93 ± .03**	**.75 ± .03**	.83 ± .09	**.83 ± .09**
random initialization	.68 ± .10	.59 ± .11	.50 ± .05	.61 ± .11	.60 ± .12
without content aware loss	.81 ± .07	.90 ± .12	.68 ± .06	**.84 ± .04**	.81 ± .11
BrainBERT (superlet)	.78 ± .08	.86 ± .06	.62 ± .05	.70 ± .10	.74 ± .12
random initialization	.66 ± .09	.54 ± .04	.52 ± .07	.60 ± .05	.58 ± .09
without content aware loss	.74 ± .12	.79 ± .14	.59 ± .05	.70 ± .13	.71 ± .14
without adaptive mask	.78 ± .08	.86 ± .05	.70 ± .04	.76 ± .06	.77 ± .08

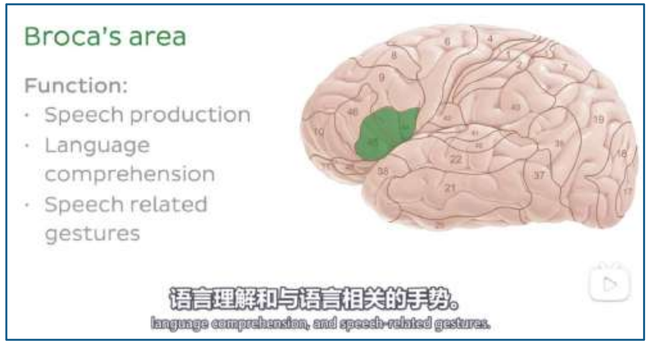
🔥BrainBERT



	Sentence onset	Speech/Non-speech	Pitch	Volume	Task Avg.
BrainBERT (STFT)	.66 ± .03	.63 ± .05	.51 ± .07	.60 ± .05	.60 ± .08
random initialization	.62 ± .04	.57 ± .04	.52 ± .06	.59 ± .07	.57 ± .06
without content aware loss	.65 ± .04	.64 ± .04	.51 ± .07	.60 ± .05	.60 ± .08
BrainBERT (superlet)	**.71 ± .06**	**.69 ± .06**	.53 ± .07	.60 ± .08	**.63 ± .10**
random initialization	.62 ± .03	.56 ± .05	.52 ± .06	.59 ± .08	.57 ± .07
without content aware loss	.68 ± .06	.67 ± .07	.53 ± .07	.60 ± .07	.62 ± .09
without adaptive mask	.67 ± .06	.66 ± .06	**.54 ± .06**	.60 ± .07	.62 ± .08

❄️BrainBERT

Map on Brain



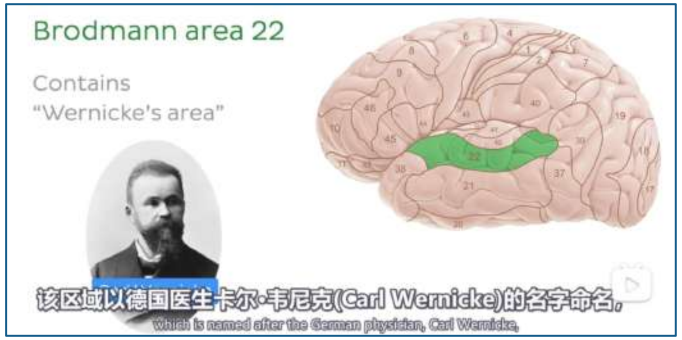
Broca's area

Function:

- Speech production
- Language comprehension
- Speech related gestures

语言理解和与语言相关的手势。

language comprehension, and speech-related gestures.



Brodmann area 22

Contains "Wernicke's area"

该区域以德国医生卡尔·韦尼克(Carl Wernicke)的名字命名,

which is named after the German physician, Carl Wernicke,

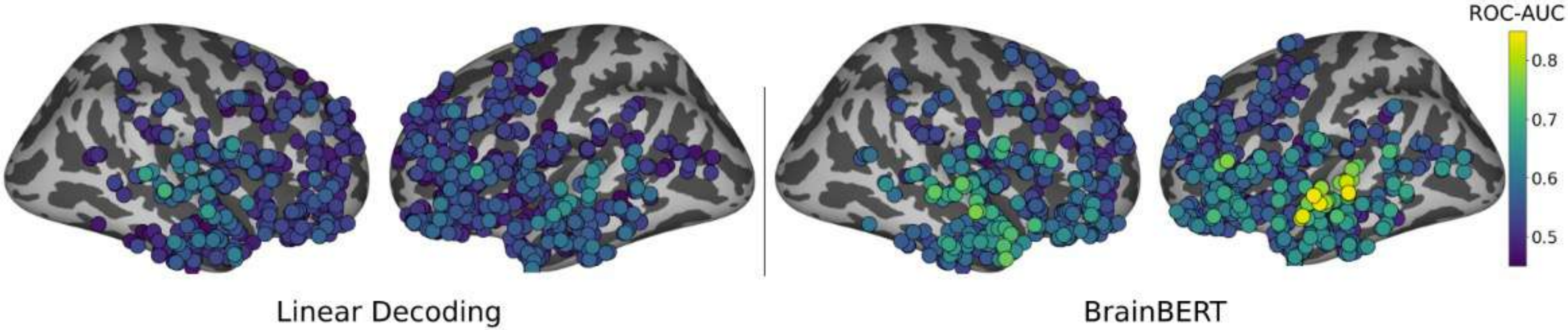


Figure 3: Using a linear decoder for classifying sentence onsets either (left) directly with the neural recordings or (right) with BrainBERT (superlet input) embeddings. Chance has AUC of 0.5. Only the 947 held-out electrodes are shown. Using BrainBERT highlights far more relevant electrodes, provides much better decoding accuracy, and more convincingly identifies language-related regions in the superior temporal and frontal regions.

Generalizing to New Subjects

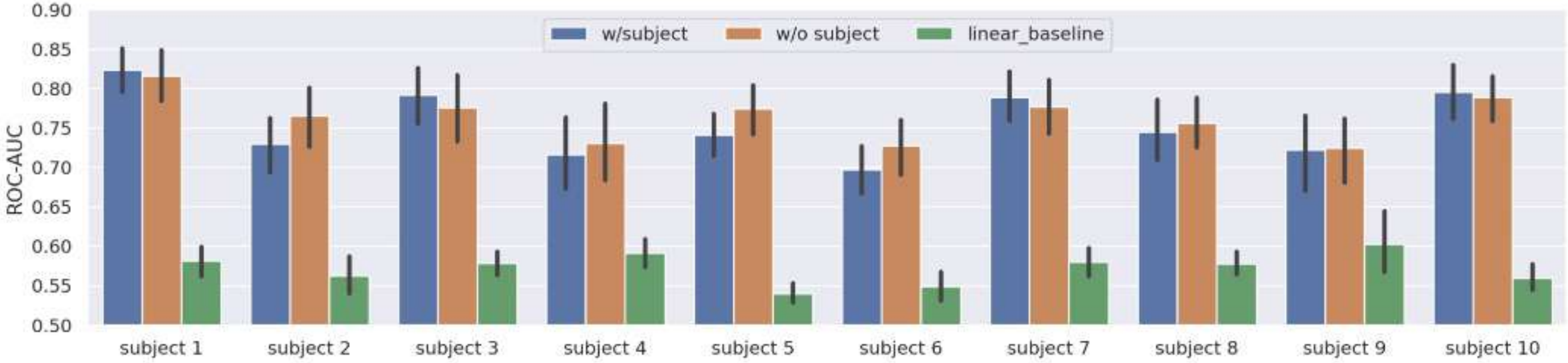


Figure 4: BrainBERT can be used off-the-shelf for new experiments with new subjects that have new electrode locations. The performance of BrainBERT does not depend on the subject data being seen during pretraining. We show AUC averaged across the four decoding tasks, in each case finetuning BrainBERT's weights and training a linear decoder. 10 held-out electrodes were chosen from the held out subject's data. As before, these electrodes have the highest linear decoding accuracy on the original data without BrainBERT. The first two columns in each group show BrainBERT decoding results when a given subject is included in the pretraining set, and when that subject is held out. The performance difference between the two is negligible, and both significantly outperform the linear decoding baseline, showing that BrainBERT is robust and can be used off the shelf.

Improved Data Efficiency

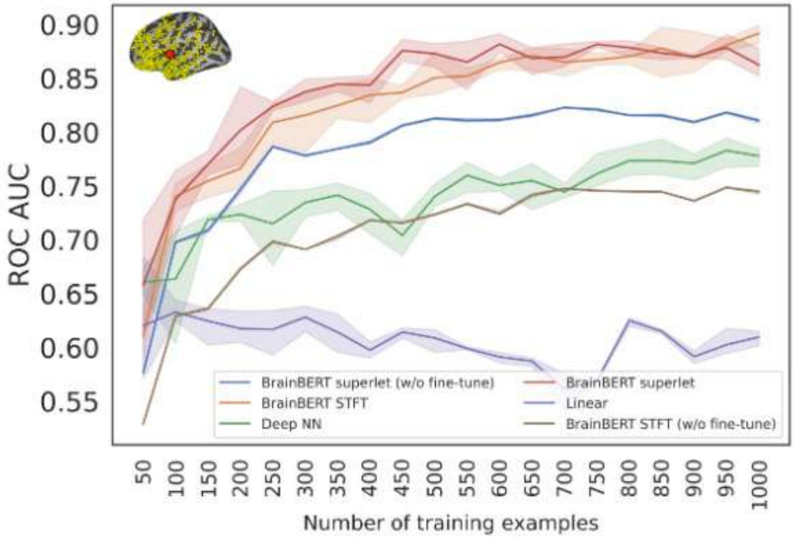


Figure 5: BrainBERT not only improves decoding accuracy, but it does so with far less data than other approaches. Performance on sentence onset classification is shown for an electrode in the superior temporal gyrus (red). Error bars show 95% confidence interval over 3 random seeds. Linear decoders saturate quickly, deep neural networks (5 FF layers, details in text) perform much better but they lose explainability. BrainBERT without fine tuning matches the performance of deep networks, without needing to learn new non-linearities. With fine-tuning BrainBERT significant outperforms, and it does so with 1/5th as many examples (deep NN peak at 1000 examples is exceeded with only 150 examples). This is a critical enabling step for other analyses where subjects may participate in only a few dozen trials as well as for BCI.

Intrinsic Dimension

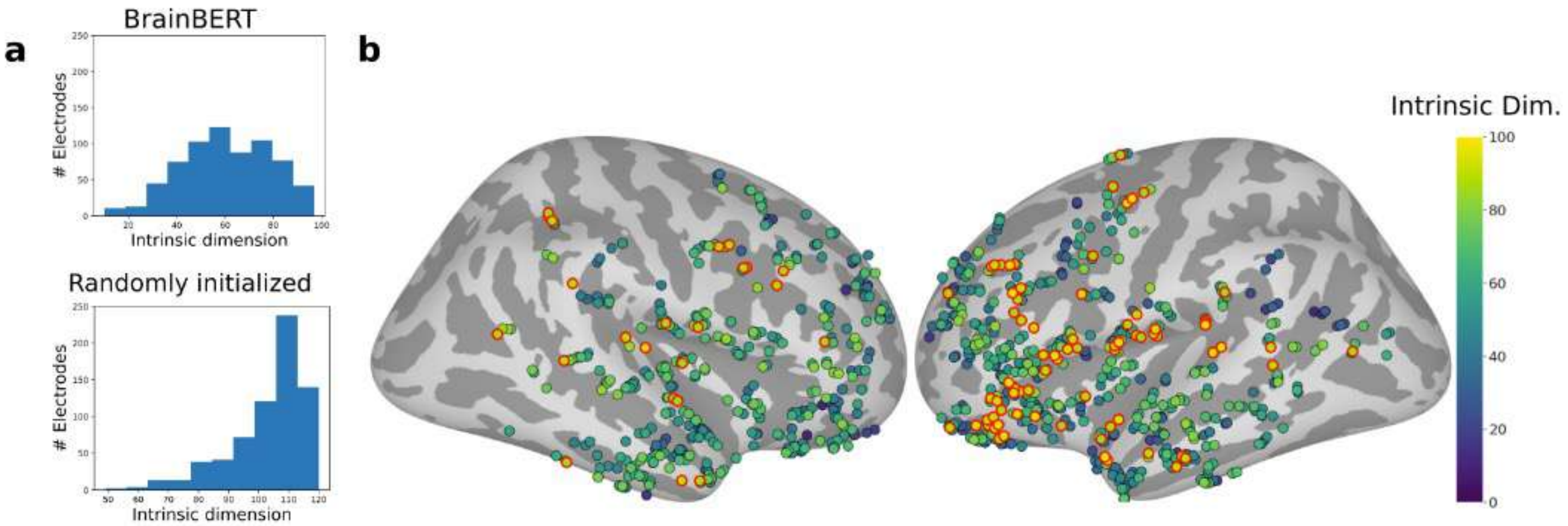


Figure 6: Given neural recordings without any annotations, we compute the intrinsic dimensonality (ID) of the BrainBERT embeddings at each electrode. (a) These embeddings lie in a lower dimensional space than those produced by a randomly initialized model. (b) The electrodes with the highest ID (top 10-th percentile; circled in red) can be found mainly in the frontal and temporal lobes, and demonstrate that electrodes that participate in similar computations on similar data will have similar ID, providing a new data-driven metric by which to identify functional regions and the relationship between them.

Thanks~