Towards robust and generalizable representations of extracellular data using contrastive learning

Ankit Vishnubhotla*
Columbia University
New York
av3016@columbia.edu

Charlotte Loh*
MIT
Massachusetts
cloh@mit.edu

Liam Paninski
Columbia University
New York
liam@stat.columbia.edu

Akash Srivastava
MIT-IBM
Massachusetts
Akash.Srivastava@ibm.com

Cole Hurwitz
Columbia University
New York
ch3676@columbia.edu

Abstract

Contrastive learning is quickly becoming an essential tool in neuroscience for extracting robust and meaningful representations of neural activity. Despite numerous applications to neuronal population data, there has been little exploration of how these methods can be adapted to key primary data analysis tasks such as spike sorting or cell-type classification. In this work, we propose a novel contrastive learning framework, **CEED** (**C**ontrastive **E**mbeddings for **E**xtracellular **D**ata), for high-density extracellular recordings. We demonstrate that through careful design of the network architecture and data augmentations, it is possible to generically extract representations that far outperform current specialized approaches. We validate our method across multiple high-density extracellular recordings. All code used to run CEED can be found at https://github.com/ankitvishnu23/CEED.

1 Introduction

High-density extracellular recordings now allow for simultaneous recordings of large populations of neurons across multiple brain regions with high temporal and spatial resolution. [1–4]. These large-scale recordings are essential for gaining insights into key biological processes such as vision, decision-making, and behavior which are distributed across brain regions [5]. Along with gaining insights into brain function, these technologies also promise to improve the scalability and accuracy of brain-computer interfaces which can restore motor function to paralyzed individuals [6].

A major bottleneck for interpreting neural population activity is processing the raw extracellular signal [7]. Although extracellular recordings contain a precise record of the coordinated neural activity, it must be extracted algorithmically through a processing step called *spike sorting*. A crucial assumption in spike sorting is that each recorded neuron has a unique spatiotemporal extracellular waveform based on its morphology and position relative to the recording device [8]. Using this unique identifier, it is possible to assign a detected extracellular action potential (spike) back to its putative neuron (unit). Along with spike sorting, another important task in extracellular analysis is cell-type classification. As neural circuits are diverse and heterogeneous, it is becoming increasingly important to profile detected units using their morphoelectrical features [9]. At a coarse level, it is

*Equal contribution

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

https://doi.org/10.52202/075280-1832

possible to classify units as inhibitory or excitatory based on their extracellular profile [10]; however, it may be possible to divide units into finer subgroups [11–13].

For both spike sorting and cell-type classification, it is important to extract low-dimensional, meaningful features from extracellular waveforms. Despite the importance of feature extraction, current approaches are often ad hoc and lack robustness to common nuisance variables in extracellular recordings. By far the most common featurization method for spike sorting is principal components analysis (PCA) [3, 14–22]. Although PCA is relatively effective and scalable, it suffers from a few key drawbacks including: (1) a lack of robustness to extracellular nuisance variables such as spatiotemporally overlapping spikes (collisions), (2) an inability to model non-linear data, and (3) an objective function that aims to find features that explain variance rather than features that discriminate different waveforms. To improve the robustness of PCA, [21] introduced a supervised waveform denoiser that is able to 'clean' the waveforms before featurization. For morphoelectric cell-type classification, most feature extraction methods rely heavily on manually extracted features including action potential width, peak-to-peak amplitude, and the ratio of pre-hyperpolarization peak to the post-hyperpolarization peak [13]. Again, while scalable and effective, these features are too simple and ad hoc to fully capture morpho-eletrical differences [13]. Recently, a non-linear approach to cell-type classification, WaveMap, was introduced, utilizing UMAP [23] and Louvain community detection [24] to automatically find cell-type clusters [13].

In this work, we introduce a robust and generalizable feature learning method, CEED, for extracellular datasets. Our main hypothesis is that embeddings of extracellular waveforms that are invariant to both common and task-specific nuisance variables will be more useful for spike sorting and morphoelectric cell-type classification than current specialized feature extraction methods. For example, in extracellular recordings, there are a number of nuisance variables including collisions, correlated background noise, or variability in the time at which a spike is detected [21]. Each of these confounds pose a problem for traditional representation learning methods like PCA or manually extracted features. Along with these common nuisance variables, there are also task-specific nuisance variables such as the spatial position of detected spikes. For spike sorting, this information is crucial as spike locations are highly informative of neuron identity [25, 26]. However, for cell-type classification, neurons with different locations may still share a cell-type. In order to extract representations that are invariant to these nuisance variables, we utilize contrastive learning, which has been shown to approximately induce invariance in the representation space to a set of transformations [27]. We utilize a recent contrastive learning framework [28] for our training and implement a stochastic view generation module for extracellular waveforms. We validate our approach on multiple high-density extracellular recordings. Surprisingly, for cell-type classification, our representations appear to be more informative than state-of-the-art specialized methods even when performing zero-shot learning on an unseen animal and probe geometry. Our contributions are as follows:

1. We introduce a novel framework, CEED, for analyzing extracellular recordings based on invariance learning.
2. We implement a stochastic view generation module for both single-channel and multi-channel extracellular waveforms.
3. We demonstrate that CEED works well with multiple neural network architectures, including a novel transformer-based architecture with a spatiotemporal causal attention mask (SCAM).
4. We show that CEED outperforms specialized featurization methods for spike sorting and morphoelectric cell-type classification.

2 Background

Contrastive Learning. Contrastive representation learning [28] falls under a broad class of self-supervised learning methods [29–31] whose goal is to learn robust and generalizable representations by encouraging invariances to a prior-known set of transformations or nuisance variables. While contrastive learning has mainly been used to extract effective and transferable representations from image-based data [27, 28], more recently, it has also become a powerful tool in the sciences to learn invariances for physical systems [32, 33].

In computational neuroscience, representation learning has mostly been performed using generative models [34]. While this has led to many interesting insights into behavior [35] and decision-making [36], this paradigm is sensitive to nuisance variables and will not capture subtle changes in the

https://doi.org/10.52202/075280-1832

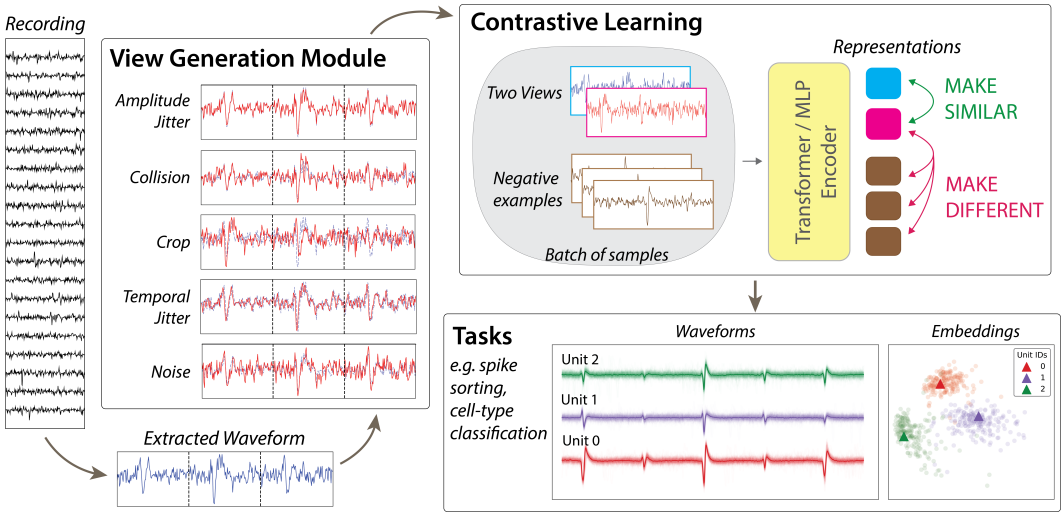

Figure 1: **CEED framework.** In CEED, we assume that waveforms are already extracted from an extracellular recording. Each waveform is then passed through our stochastic view generation module, where different views are obtained by applying transformations. These transformations induce a pre-defined set of invariances (see Section 3.2.1). Using these views, a neural-network based encoder, which can take the form of a multi-layer perceptron (MLP) or a transformer, is then trained to produce representations that respect the desired invariances. This is accomplished via contrastive learning, where representations from views of the same waveform are encouraged to be similar and views of different waveforms are encouraged to be dissimilar. When training completes, the learned representations can then be used for a series of downstream neuroscience tasks, such as spike sorting or morphoelectrical cell-type classification.

observation space that do not contribute much to the explained variance. While recent research into label-guided generative models [37–39] has somewhat alleviated these problems, it is still unclear if good generative performance is important for representation learning. Most recently, a number of contrastive learning methods have been introduced for learning robust and generalizable representations of neural population data [40–42]. To our knowledge, however, contrastive learning has not been applied to key primary data analysis tasks such as spike sorting or cell-type classification.

Spike sorting. In spike sorting, the goal is to extract a precise record of which neurons spiked at which time steps based on the raw extracellular data. For each electrode, the contributions of an unknown number of neurons are mixed together, making spike sorting a challenging blind-source separation task akin to the cocktail party problem [43]. While a number of spike sorting methods have been developed for microelectrode arrays (MEAs), almost all of these methods currently utilize PCA for feature extraction For example, the popular algorithms Klusta, HerdingSpikes2, Mountainsort4, SpykingCircus, Trideclous, and Kilosort all use a form of SVD/PCA in their processing pipelines. Despite the ubiquity of PCA as a feature extraction method for spike sorting, it lacks robustness to common nuisance variables in extracellular recordings, cannot capture non-linear features of the observed data, and prefers features that explain variance rather than those that discriminate different waveforms. More recently, it has been shown that non-linear autoencoders can have higher performance than PCA despite suffering from similar drawbacks [44].

Cell-type classification. Morphoelectric cell-type classification has many similarities to spike sorting. Similarly to spike sorting, you are also interested in grouping detected extracellular waveforms together. However, the granularity of these clusters should be lower, reflecting shared morphologies among the units in the recording. Classically, pre-defined morphoelectrical features (e.g., action potential width) were extracted from the waveforms to group together similar units [10, 45]. More recently, a non-linear method based on UMAP was introduced for cell-type classification [13]. In this work, it was shown that automatically discovering relevant features from extracellular waveforms can lead to more informative cell-type groupings than when simply clustering pre-defined features. It is important to note that, along with morphoelectric cell-type classification, it is also possible to classify cell-types by their function, morphology, physiology, molecular properties [46–48].

42273 https://doi.org/10.52202/075280-1832

3 Contrastive Embeddings for Extracellular Data (CEED)

We introduce Contrastive Embeddings for Extracellular Data (CEED), a novel framework for extracting robust and generalizable representations of extracellular data via contrastive learning. CEED consists of three main components (see Figure 1) — (1) a stochastic view generation module that augments waveforms with both common and task-specific nuisance variables. (2) A neural network-based encoder that can extract low-dimensional representations from both single-channel and multi-channel waveforms. (3) A contrastive loss function that encourages the representations of views of the same waveform to be similar while forcing the representations of views of different waveforms to be dissimilar. We postulate that the representations extracted by CEED, which are approximately invariant to nuisance variables, can outperform specialized feature extraction methods for both spike sorting and cell-type classification.

3.1 Notation

To introduce our method, we must first define some notation for extracellular recordings. Let $P_M \coloneqq \{p_m | m \in M\}$ be the position of all M channels on the multi-electrode array (MEA), where each $p_m \in \mathbb{R}^3$ is the location of channel m. Now define $P_{\tilde{M}}$ to be the set of positions for a subset of channels, $\tilde{M} \subseteq M$, on the MEA.

Let $S \coloneqq \{s_n\}_{n=1}^N$ be the set of N spiking events that are detected in an extracellular recording. Now, let the extracellular waveform of a spiking event s_n on a channel m be defined as $W_{n,m} \coloneqq \{r_{n,m}^{(0)}, r_{n,m}^{(1)}, ..., r_{n,m}^{(T)}\} \in \mathbb{R}^T$, where T contiguous samples from the recording define a waveform and $r_{n,m}^{(t)}$ is the value of the waveform at sample t. This can be thought of as a window of samples of size T on channel m which includes the spike s_n.

The amplitude of a spike s_n on a channel m can be defined as $A_{n,m} \coloneqq \max_t |W_{n,m}|$. This is also known as the absolute maximum voltage recorded on channel m. The timestep t at which the absolute maximum voltage occurs can vary for each channel. Therefore, for a spike, we center the waveform on each channel using the timestep at which the maximum $A_{n,m}$ occurred (i.e., the timestep of the maximum amplitude channel). For all spikes in the recording, waveforms are aligned in this manner so that the amplitude on the maximum amplitude channel occurs at the same timestep t.

Let $\mathbf{W}_{n,\tilde{M}} \in \mathbb{R}^{T \times |\tilde{M}|}$ be the set waveforms for a spike on a subset of channels and let $\mathbf{A}_{n,\tilde{M}} \in {\mathbb{R}_{>0}}^{|\tilde{M}|}$ be the set amplitudes for a spike on a subset of channels, where $|\tilde{M}|$ is the number of channels.

3.2 Stochastic view generation module

3.2.1 Common invariances

Let us define the representation of the waveform $\mathbf{W}_{n,\tilde{M}}$ for spike s_n as $\mathbf{z}_n \in \mathbb{R}^D$. There are a number of nuisance variables that change the observed waveform of a spike without changing the underlying signal. Therefore, we can define a set of invariances we wish to impose on our representations,

1. **Amplitude voltage jitter** - We want representations that are invariant to "small" amplitude variability. To this end, if we scale the amplitude of the waveform $\mathbf{W}_{n,\tilde{M}}$ for spike s_n such that $\widetilde{\mathbf{W}}_{n,\tilde{M}} = \mathbf{W}_{n,\tilde{M}} \times (1 \pm \epsilon)$, then the representation $\tilde{\mathbf{z}}_n$ for $\widetilde{\mathbf{W}}_{n,\tilde{M}}$ should be such that $\tilde{\mathbf{z}}_n = \mathbf{z}_n$.
2. **Correlated background noise** - We want representations that are invariant to the correlated background noise found in real extracellular recordings. We assume extracellular noise can be modeled as a spatiotemporal matrix Gaussian $\mathcal{MN}(0, U, V)$, where $U \in \mathbb{R}^{T \times T}$ models the temporal correlation of a waveform and $V \in \mathbb{R}^{|M| \times |M|}$ models the spatial correlation between channels. For a spike s_n, we want a representation such that if we sample $n \sim \mathcal{MN}(0, U, V)$ and form $\widetilde{\mathbf{W}}_{n,\tilde{M}} = \mathbf{W}_{n,\tilde{M}} + n_{\tilde{M}}$, where $n_{\tilde{M}}$ is the noise on the channels subset $\tilde{M}$, then $\widetilde{\mathbf{W}}_{n,\tilde{M}}$ should still yield a representation $\tilde{\mathbf{z}}_n = \mathbf{z}_n$.
3. **Spike collisions** - We want representations that are invariant to the voltage distortions caused when spikes "collide" spatiotemporally. A collision is when two or more spikes occur at similar times in the same location on the probe leading to observed waveforms that are distorted. To this end, consider two spikes s_{n_1} and s_{n_2} with waveforms $\mathbf{W}_{n_1,\tilde{M}_1}$ and $\mathbf{W}_{n_2,\tilde{M}_2}$ such that $\tilde{M}_1$

https://doi.org/10.52202/075280-1832

and $\tilde{M}_2$ share some number of channels. Now let us shift s_{n_2} by k time steps ($-T/2 < k < 0$ or $0 < k < T/2$) such that for any channel $m \in \tilde{M}_2$, $W_{n_2,m} \coloneqq \{r_{n_2,m}^{(0+k)}, r_{n_2,m}^{(1+k)}, ..., r_{n_2,m}^{(T+k)}\}$. We refer to this temporally shifted waveform as $\mathbf{W}_{n_2,\tilde{M}_2}^{\text{shift}}$. A new waveform that is a linear sum of these two waveforms $\tilde{\mathbf{W}}_{n,\tilde{M}_1} = \mathbf{W}_{n_1,\tilde{M}_1} + \mathbf{W}_{n_2,\tilde{M}_2}^{\text{shift}}$ should yield a representation $\tilde{\mathbf{z}}_n = \mathbf{z}_{n_1}$.

4. **Channel subsets** - We want representations that are invariant to changes in the subset of channels $\tilde{M}$ used to define $\mathbf{W}_{n,\tilde{M}}$. So for a new subset of channels $\tilde{M}'$, as long as the channel with the highest amplitude is still contained in $\tilde{M}'$, then the waveforms $\mathbf{W}_{n,\tilde{M}'}$ should still yield a representation $\tilde{\mathbf{z}}_n = \mathbf{z}_n$.

3.2.2 Task-specific invariances

While all feature extraction methods suffer from common extracellular nuisance variables, there are also task-specific nuisance variables. For example, the position and orientation of neurons in the recording are essential information for spike sorting. However, for cell-type classification, this information can be a confound when trying to find shared morphoelectrical features. Therefore, we propose an additional set of invariances for cell-type classification,

1. **Cell position** - For cell-type classification, we want representations that are invariant to the channel positions P_M at which a spike is detected. So for a waveform $\mathbf{W}_{n,M}$, any uniform changes to the channel positions $P_{\tilde{M}}$ (without changing the waveforms at each channel) should still yield a representation $\tilde{\mathbf{z}}_n = \mathbf{z}_n$.
2. **Cell amplitude** - As amplitude is mainly a function of cell position and orientation, we want representations that are invariant to any uniform changes in the amplitudes. To this end, for any positive value a, a change in the amplitudes for the spike s_n, $\tilde{\mathbf{W}}_{n,\tilde{M}} = \mathbf{W}_{n,\tilde{M}} \times a$, should still yield a representation $\tilde{\mathbf{z}}_n = \mathbf{z}_n$.

To achieve these task-specific invariances for cell-type classification, we directly transform the training data. First, we extract the max channel waveform for each spike. Then, we normalize each waveform to be between -1 and 1, thus removing any positional information [13]. A drawback of this approach is that we discard multi-channel information that may be useful for classifying different cell-types [12] (see Section 7 for a more detailed discussion). The full view generation pipeline is detailed in Supplementary Materials Section A.1.

3.3 Encoder Architecture

For the encoder of CEED, we explored two neural network architectures – a transformer-based network with a novel Spatiotemporal Causal Attention Mask (SCAM) and a simple multi-layer perceptron (MLP) which is a more computationally efficient alternative (i.e., can fit on a single GPU). For a runtime comparison of these two architectures, see Supplementary Materials Section E.

3.3.1 Transformer-based encoder with spatiotemporal causal attention mask (SCAM)

Our first proposed architecture utilizes transformers [49] which have been highly successful across a series of tasks in natural language processing [50, 51]. Transformers have a natural inductive bias towards time-series and sequence-based data and are thus highly suitable for extracellular waveforms. For this work, we designed a novel spatiotemporal causal attention mask (SCAM) to obey causality across time and channels. Specifically, we allow every recorded time step in a waveform to attend to time steps on other channels as long as those data points precede it in time. Full details and visualization of the transformer-based architecture, including implementation details are available in Supplementary Materials Section A.2.

3.3.2 Multi-layer perceptron (MLP)

While a transformer-based architecture provides a natural inductive bias towards time-series data appropriate for extracellular recordings, a downside to such an architecture is its high computational complexity and the requirement that it needs multiple GPUs to train. To demonstrate the generality of the CEED framework, we also propose a simpler MLP architecture that can be trained on a single GPU. The MLP encoder is a straightforward model that consists of three layers with sizes $[768, 512, 256]$ and ReLU activations between them.

https://doi.org/10.52202/075280-1832

3.4 Objective function

The encoder is trained as follows. Let $\tilde{\mathbf{z}}_i$ and $\mathbf{z}_i$ be representations of the two views of the input waveform i, where the two views differ by the invariances listed in Sections 3.2.1 and 3.2.2. The optimization objective for a batch of B samples follows the contrastive loss utilized by SimCLR [28] and several other representation learning methods [52, 53];

$$\mathcal{L} = \sum_{i=1}^{B} -\log \frac{\exp(\hat{\mathbf{z}}_i \cdot \hat{\tilde{\mathbf{z}}}_i)/\tau}{\sum_{k \neq i} \exp(\hat{\mathbf{z}}_i \cdot \hat{\mathbf{z}}_k/\tau)} \tag{1}$$

where $\hat{\mathbf{z}}_i$ and $\hat{\tilde{\mathbf{z}}}_i$ are the L2 normalized representations and τ is a temperature hyperparameter. Following [28], we also include a 2-layer MLP projector network after the encoder and the loss function operates on the outputs of this projector network (see Supplementary Materials Section A.3 for details and ablations of the projector network and its architecture).

4 Datasets

To train and evaluate our model, we make use of two publicly available extracellular recordings published by the International Brain Laboratory (IBL): the DY016 and DY009 recordings [54]. These multi-region, Neuropixels 1.0 recordings are taken from a mouse performing a decision-making task (see Supplementary Materials Section C for more details).

4.1 Spike sorting

To evaluate how useful the features learned by CEED are for spike sorting, we constructed three datasets using units found by Kilosort 2.5, a full spike sorting pipeline manually tuned by IBL. The first dataset was extracted from the DY016 extracellular recording. It consisted of a 10 unit train and test dataset where all 10 units were classified as "good" by IBL's quality metrics [55]. We selected these units for their high waveform diversity (qualitatively) and because they had a relatively high amplitude, i.e., peak-to-peak (ptp). For this dataset, we constructed training sets of 200 or 1200 spikes per unit with a test set of 200 spikes per unit. For each spike, we extracted waveforms from 21 channels centered on the maximum amplitude channel. Although we extract 21 channels for our data augmentations (see Supplementary Materials Section A.1), we train and evaluate our model (and baselines) on either 5 or 11 channel subsets.

The second dataset was extracted from both the DY016 and DY009 extracellular recordings. This dataset consisted of a 400 training units and 10 test units (the same 10 units evaluated in the first dataset). We extracted 200 units from each recording to create the training set with 200 or 1200 training spikes per unit. The goal of this dataset was to test how well CEED could generalize across a large set of units with varying quality. To this end, we also evaluated the performance of CEED on 100 random test sets of varying sizes (see Supplementary Materials Section D).

The third and final dataset was also extracted from both the DY016 and DY009 extracellular recordings. Unlike the first and second dataset, where the training set contained spikes from the test units (in-distribution), we purposefully excluded all units in the test set from the training set. To this end, this dataset consisted of 390 training units and 10 test units (the same 10 units evaluated in the above datasets). For all units in the training set, we utilized either 200 or 1200 training spikes per unit. The goal of this dataset was to test how well CEED could generalize to out-of-distribution (OOD) units.

4.2 Cell-type

For our cell-type classification dataset, we utilized the DY016 extracellular recording. We extracted the same 10 IBL "good" units used for the spike sorting datasets. To remove positional information from each spike, we only extract waveforms from the maximum amplitude channel and we normalize each waveform as described in Section 3.2.2.

https://doi.org/10.52202/075280-1832

5 Experiments

5.1 Spike sorting

Re-sorting units extracted using KS2.5 In this experiment, we aim to demonstrate that CEED extracts more useful features for spike sorting than both PCA and a non-linear autoencoder. We evaluate these methods on both an in-distribution (ID) and out-of-distribution (OOD) waveform discrimination task. Specifically, we train CEED, PCA, and an autoencoder on all three spike sorting datasets (see Section 5) and perform inference on the spikes from the test units. Then, we 're-sort' these test spikes back to their putative units by performing clustering on the resulting embeddings. For clustering, we use a parametric clustering algorithm, the Gaussian Mixture Model (GMM), and a non-parametric clustering algorithm, HDBSCAN [56]. To evaluate how well each clustering corresponds to the ground-truth, we use the Adjusted Rand Index (ARI) [57]. One can also compute the accuracy after optimal permutation (e.g., Hungarian algorithm), however, this is will give similar results to the ARI. To strengthen our spike sorting baselines, we also compare CEED to PCA and an autoencoder trained and tested on denoised waveforms using the YASS waveform denoiser [21, 58]. The YASS denoiser was trained to denoise single channel waveforms so we apply it independently to each channel and it improves clustering of PCA across the board. For all baselines, we sweep across (3,5,7,9) principal components and 3-11 channel subset sizes.

5.2 Cell-type classification

Cell-type classification of IBL recordings In this analysis, we aim to demonstrate that CEED can extract putative cell-type clusters from IBL extracellular recordings. For the IBL recordings, we extract average waveforms (templates) for 163 good units pooled over DY016 and DY009. We then run inference on the extracted templates using CEED and perform a GMM clustering on the resulting embeddings to find putative cell-types. As there is no ground-truth for the IBL recordings, we validate our results by visualizing the templates for each cell-type cluster.

Zero-shot cell-type classification. In this experiment, we compare the inferred representations of CEED to a state-of-the-art cell-type classification method, WaveMap, on an out-of-distribution (OOD) dataset. Not only is this dataset not seen during training, but both the animal and the probe geometry are completely different than the dataset used to train CEED. The dataset consists of 625 templates (average waveforms) extracted from units which are recorded while a monkey performs a discrimination task [13]. The probe used to record these waveforms is a Plexon U-probe. The goal of this experiment is to find putative cell-types that explain the waveform variability seen in the data. We propose to run CEED on the 625 templates and then perform a GMM clustering of the resulting embeddings. We choose the number of clusters by again utilizing the Elbow method and BIC. Given how OOD this dataset is compared to the recordings used to train CEED, success on this task would demonstrate the robustness and generalizability of CEED.

6 Results

6.1 Spike Sorting

Re-sorting units extracted using KS2.5 For our re-sorting task, we find that CEED outperforms both PCA and the non-linear autoencoder using raw waveforms or denoised waveforms across all three datasets introduced in Section 4. The results for this analysis can be found in Table 1. As can be seen, in both the ID and OOD regimes, CEED has much higher performance than both baseline models. For all methods we utilize the same number of latent dimensions for the analysis (5D). We also show in Supplementary Materials Section D that the strong performance of CEED generalizes to many different sets of test units.

Along with this results table, we also quantitatively and qualitatively compare the performance of CEED to PCA on the 10 neuron train and test dataset in Figure 2. Visually, it can be seen that CEED's features are far more informative about each unit's identity than the representations found by PCA. The performance of CEED is also much higher when being clustered by either HDBSCAN or a GMM even when when more principal components are afforded for the analysis.

https://doi.org/10.52202/075280-1832

Method	10 neuron train set 10 neuron ID (ARI)	400 neuron train set 10 neuron ID (ARI)	390 neuron train set 10 neuron OOD (ARI)
CEED (1200 spikes, 11 channels)	**.89 ± .04**	**.79 ± .09**	**.78 ± .05**
CEED (1200 spikes, 5 channels)	.83 ± .03	.76 ± .07	.77 ± .08
Denoised PCA (1200 spikes, 11 channels)	.39 ± .05	.45 ± .04	.46 ± .04
Denoised PCA (1200 spikes, 5 channels)	.46 ± .07	.48 ± .04	.49 ± .04
Autoencoder (1200 spikes, 11 channels)	.47 ± .06	.35 ± .03	.28 ± .02
Autoencoder (1200 spikes, 5 channels)	.43 ± .06	.37 ± .03	.33 ± .01

Table 1: **Benchmarking CEED, PCA, and an autoencoder on in-distribution (ID) and out-of-distribution (OOD) data.** For evaluation, we fit 50 GMMs to the embeddings and compute the mean and std. of the adjusted rand index (ARI). First column: we train and test each method with spikes from 10 neurons. Second column: we train each method with spikes from 400 neurons and then test on the original 10 neurons which are **included** in the training set (ID). Third column: we train each method on spikes from 390 neurons and test on the original 10 neurons which are **not included** in the training set (OOD). This experiment demonstrates that CEED peforms well on OOD data and can outperform a non-linear autoencoder. All CEED results are generated using the MLP encoder.

6.2 Cell-type Classification

Cell-type classification of IBL recordings The results for cell-type classification of the pooled IBL recordings are shown in Supplementary Materials Section B. We choose the number of cell-type clusters by sweeping over 1-10 clusters and choosing the minimum BIC. We find that 4 clusters explain most of the waveform variability. Interestingly, we find good separation in CEED's embedding space between narrow-spiking and broad-spiking units which indicates that we may be able to discriminate between inhibitory and excitatory subtypes.

Zero-shot cell-type classification. The results for cell-type classification of OOD single unit data are shown in Figure 3. Despite training on IBL extracellular datasets from a mouse brain recorded with Neuropixels 1.0, CEED is able to generalize to unseen data from a completely different animal (monkey) and probe (Plexon U-probe). In Figure 3A, we visualize the inferred representations from CEED using a 2D UMAP and by coloring the points with the output of a GMM that is trained on the 5D contrastive representations. We choose the number of cell-type clusters by sweeping over 1-10 clusters and then using the Elbow Method on the BIC curve. With this method, we discover 6 putative cell-types (similarly to WaveMap). Upon visual inspection, our cell-types are more well-isolated from each other than those of WaveMap (Figure 3C). To quantitatively assess which cell-type classification method better reflects the 'real' differences in extracellular waveforms, we utilize a supervised classifier which is trained to predict cell-type labels using input waveforms (introduced in [13]). In Figure 3B and Figure 3D, we show the results of this data on 5 cross-validation folds. Despite never training on this dataset, CEED has a much higher accuracy (93.4%) compared to that of WaveMap (88.8%). A small caveat is that we ran WaveMap using publically available code but were unable to precisely reproduce the original result in the paper which finds 8 cell-type clusters. Despite this difference, the accuracy value reported in the WaveMap paper for the 8 cell-type clusters (91%) was still lower than that of CEED.

7 Discussion

In this paper, we introduced a novel representation learning method, CEED, for extracellular recordings. Our main hypothesis was that by finding representations of extracellular waveforms that are robust to both common and task-specfic nuisance variables, we can outperform specialized feature extraction approaches on two key tasks: spike sorting and morphoelectrical cell-type classification. We validate CEED on multiple high-density extracellular datasets. For spike sorting, we show that CEED extracts features that far outperforms those of PCA and a non-linear autoencoder on a waveform discrimination task. For cell-type classification, we show that CEED is able to extract discriminative features of extracellular waveforms that allow for finding morphological subgroups in an unsupervised manner. Surprisingly, we find that CEED even outperforms a recent non-linear cell-type classification method WaveMap, on an animal and probe geometry unseen during training.

https://doi.org/10.52202/075280-1832

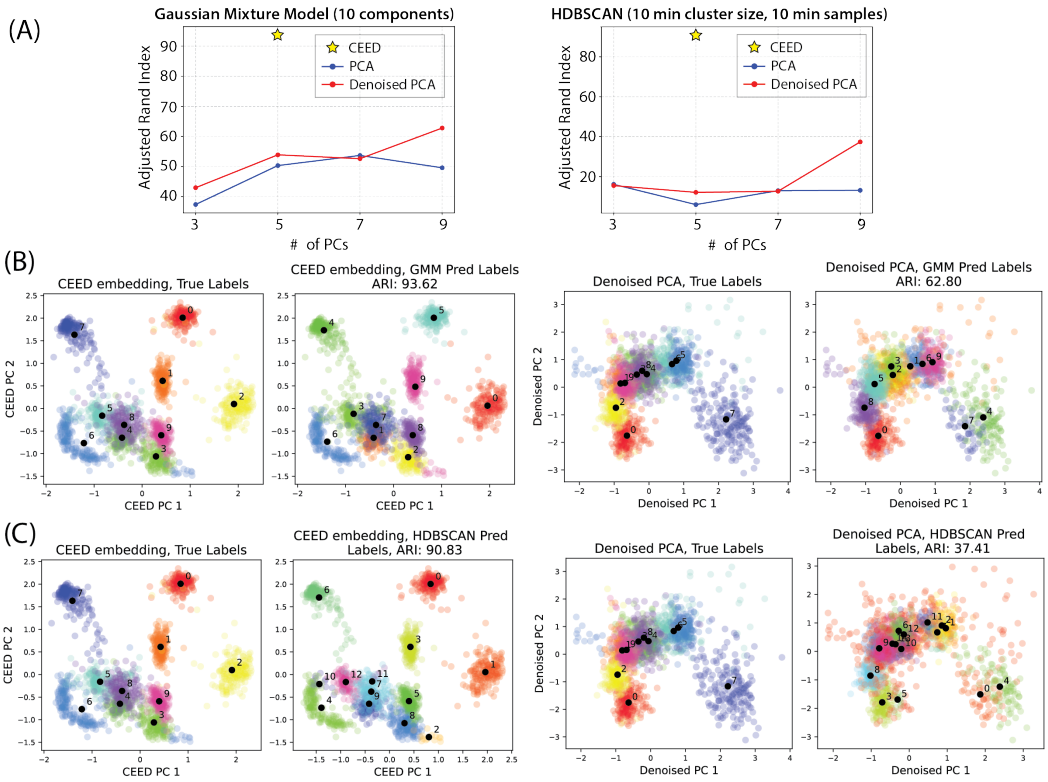

Figure 2: **CEED outperforms PCA on spike sorting featurization**. Here, we show results for CEED (using the SCAM architecture) and PCA when trained and evaluated on a 10 neuron dataset. (A) Clustering results for the two featurization methods using a parametric (GMM) and non-parametric (HDBSCAN) clustering algorithm. As can be seen, the featurization learned by CEED is much more discriminative than that of PCA for both clustering methods even when using a supervised denoiser to reduce noise in the data. (B) Visualized results of a 10 component GMM clustering on the learned embeddings from CEED (left) and denoised PCA (right). (C) Visualized results of the HDBSCAN clustering applied to the learned embeddings from CEED (left) and denoised PCA (right).

Despite the significant performance improvements of CEED, there are few limitations that must be addressed before it can become a plug-and-play method for extracellular analysis. Firstly, the training sets we utilize in this work are still quite small and lack neuron diversity. For CEED to be generalizable to multiple recordings and animals, more diverse datasets must be constructed. Secondly, the best results of CEED are obtained when using the transformer-based encoder which requires multiple GPUs and is currently quite slow to train. Recent progress in acceleration software [59] offer promising solutions to speed up computation and incorporating these methods into CEED could be a future direction. Thirdly, all spike sorting results in the paper are from re-sorting already sorted datasets; CEED must be incorporated into a full spike sorting pipeline in order to be used by many different research groups. Finally, our cell-type results do not include any functional classification which could help validate the clusters found by CEED.

8 Broader Impact

Although CEED has the potential to improve key tasks in extracellular analysis, a drawback of our approach compared to simple approaches like PCA is the additional computational resources it requires to train and run. Contrastive learning requires large batch sizes to achieve state-of-the-art performance which means, if training the SCAM transformer model, we often have to run experiments on large-scale, multi-gpu clusters. Moreover, transformer-based architectures have hundreds of thousands of parameters, which further increases CEED's compute requirements. As highly parameterized deep neural networks produce large amounts of carbon emissions [60], CEED could have a possible negative impact on the environment. We also propose an MLP-based

42279 https://doi.org/10.52202/075280-1832

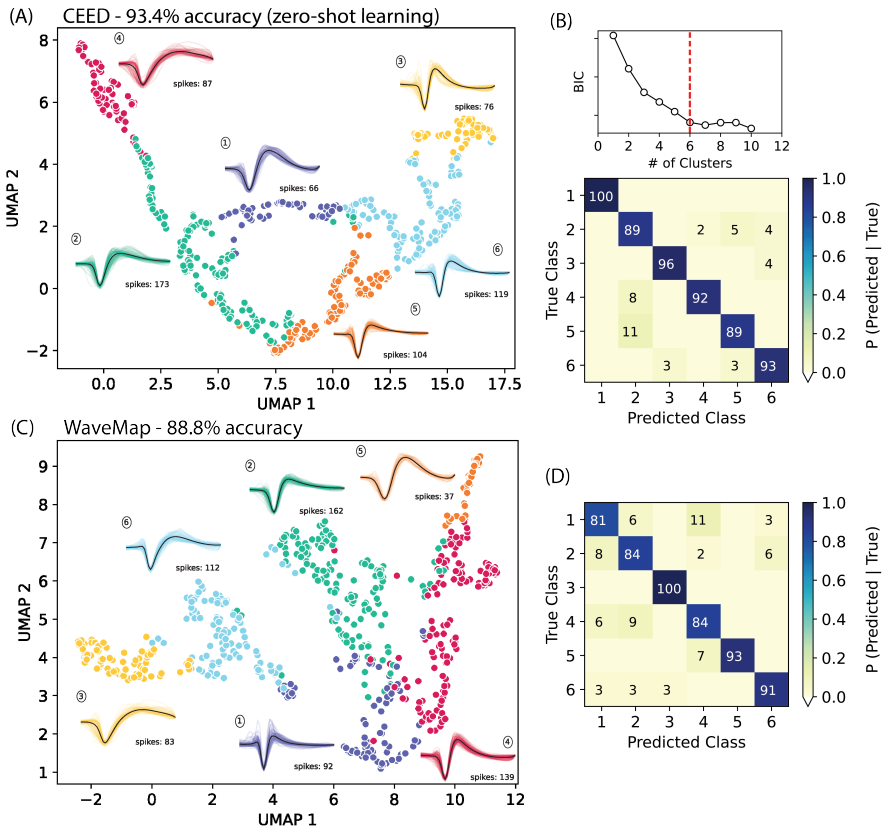

Figure 3: **CEED outperforms WaveMap on morphoelectrical cell-type classification with zero-shot learning.** (A) A 2D UMAP embedding of CEED's inferred representations for a OOD dataset which consists of 625 single units extracted from a monkey recording. We cluster the representations using a GMM and plot each cluster with a distinct color. For each cluster, we also plot the waveforms associated with the discovered cell-type. It can be seen that narrow-spiking waveforms (3,5,6) are well-separated from broad-spiking waveforms (1,2,4) which may indicate good separation of inhibitory and excitatory subtypes, respectively. Surprisingly, CEED is trained on a different animal and probe geometry, but can still generalize to this dataset, outperforming WaveMap on a classification baseline introduced in [13]. (B) On the top, we demonstrate how we chose the number of clusters for the GMM, i.e., with the Elbow Method and BIC. On the bottom, we show the confusion matrix of a gradient boosted decision tree classifier trained to map raw waveforms to the cell-type clusters extracted by CEED. The accuracy of each method is defined as the average of the diagonal. (C) A 2D UMAP embedding the single unit dataset using WaveMap. The clusters are colored according to WaveMap's outputted labels. Narrow-spiking clusters (1,2,4) are more mixed with broad-spiking clusters (3,5,6). (D) The confusion matrix of a gradient boosted decision tree classifier trained to map raw waveforms to the cell-type clusters extracted by WaveMap.

architecture that performs comparably to a transformer while only using a single GPU, which can lower CEED's impact on the environment.

References

[1] James J Jun, Nicholas A Steinmetz, Joshua H Siegle, Daniel J Denman, Marius Bauza, Brian Barbarits, Albert K Lee, Costas A Anastassiou, Alexandru Andrei, Çağatay Aydın, et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017.

[2] Ashley L Juavinett, George Bekheet, and Anne K Churchland. Chronically implanted neuropixels probes enable high-yield recordings in freely moving mice. *Elife*, 8:e47188, 2019.

[3] Nicholas A Steinmetz, Cagatay Aydin, Anna Lebedeva, Michael Okun, Marius Pachitariu, Marius Bauza, Maxime Beau, Jai Bhagat, Claudia Böhm, Martijn Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588, 2021.

[4] International Brain Laboratory, Kush Banga, Julius Benson, Niccolò Bonacchi, Sebastian A Bruijns, Rob Campbell, Gaëlle A Chapuis, Anne K Churchland, M Felicia Davatolhagh, Hyun Dong Lee, et al. Reproducibility of in-vivo electrophysiological measurements in mice. *bioRxiv*, pages 2022–05, 2022.

[5] Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

[6] Elon Musk et al. An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194, 2019.

[7] Matthias H Hennig, Cole Hurwitz, and Martino Sorbaro. Scaling spike detection and sorting for next-generation electrophysiology. *In Vitro Neuronal Networks: From Culturing Methods to Neuro-Technological Applications*, pages 171–184, 2019.

[8] Carl Gold, Darrell A Henze, Christof Koch, and Gyorgy Buzsaki. On the origin of the extracellular action potential waveform: a modeling study. *Journal of neurophysiology*, 95(5):3113–3128, 2006.

[9] Ganesh Vigneswaran, Alexander Kraskov, and Roger N Lemon. Large identified pyramidal cells in macaque motor and premotor cortex exhibit "thin spikes": implications for cell type classification. *Journal of Neuroscience*, 31(40):14235–14242, 2011.

[10] Peter Barthó, Hajime Hirase, Lenaïc Monconduit, Michael Zugaro, Kenneth D Harris, and Gyorgy Buzsaki. Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *Journal of neurophysiology*, 92(1):600–608, 2004.

[11] Alessio P Buccino, Michael Kordovan, Torbjørn V Ness, Benjamin Merkt, Philipp D Häfliger, Marianne Fyhn, Gert Cauwenberghs, Stefan Rotter, and Gaute T Einevoll. Combining biophysical modeling and deep learning for multielectrode array neuron localization and classification. *Journal of neurophysiology*, 120(3):1212–1232, 2018.

[12] Xiaoxuan Jia, Joshua H Siegle, Corbett Bennett, Samuel D Gale, Daniel J Denman, Christof Koch, and Shawn R Olsen. High-density extracellular probes reveal dendritic backpropagation and facilitate neuron classification. *Journal of neurophysiology*, 121(5):1831–1847, 2019.

[13] Eric Kenji Lee, Hymavathy Balasubramanian, Alexandra Tsolias, Stephanie Udochukwu Anakwe, Maria Medalla, Krishna V Shenoy, and Chandramouli Chandrasekaran. Non-linear dimensionality reduction on extracellular waveforms reveals cell type diversity in premotor cortex. *Elife*, 10:e67490, 2021.

[14] Samuel Garcia and Christopher Pouzat. *Tridesclous*, 2015. `https://github.com/tridesclous/tridesclous`.

[15] Cyrille Rossant, Shabnam N Kadir, Dan FM Goodman, John Schulman, Maximilian LD Hunter, Aman B Saleem, Andres Grosmark, Mariano Belluscio, George H Denfield, Alexander S Ecker, et al. Spike sorting for large, dense electrode arrays. *Nature neuroscience*, 19(4):634–641, 2016.

[16] Marius Pachitariu, Nicholas A Steinmetz, Shabnam N Kadir, Matteo Carandini, and Kenneth D Harris. Fast and accurate spike sorting of high-channel count probes with kilosort. *Advances in neural information processing systems*, 29, 2016.

[17] Jason E Chung, Jeremy F Magland, Alex H Barnett, Vanessa M Tolosa, Angela C Tooker, Kye Y Lee, Kedar G Shah, Sarah H Felix, Loren M Frank, and Leslie F Greengard. A fully automated approach to spike sorting. *Neuron*, 95(6):1381–1394, 2017.

https://doi.org/10.52202/075280-1832

[18] Gerrit Hilgen, Martino Sorbaro, Sahar Pirmoradian, Jens-Oliver Muthmann, Ibolya Edit Kepiro, Simona Ullo, Cesar Juarez Ramirez, Albert Puente Encinas, Alessandro Maccione, Luca Berdondini, et al. Unsupervised spike sorting for large-scale, high-density multielectrode arrays. *Cell reports*, 18(10):2521–2532, 2017.

[19] Pierre Yger, Giulia LB Spampinato, Elric Esposito, Baptiste Lefebvre, Stéphane Deny, Christophe Gardella, Marcel Stimberg, Florian Jetter, Guenther Zeck, Serge Picaud, et al. A spike sorting toolbox for up to thousands of electrodes validated with ground truth recordings in vitro and in vivo. *Elife*, 7:e34518, 2018.

[20] Roland Diggelmann, Michele Fiscella, Andreas Hierlemann, and Felix Franke. Automatic spike sorting for high-density microelectrode arrays. *Journal of neurophysiology*, 120(6):3155–3171, 2018.

[21] JinHyung Lee, Catalin Mitelut, Hooshmand Shokri, Ian Kinsella, Nishchal Dethe, Shenghao Wu, Kevin Li, Eduardo Blancas Reyes, Denis Turcu, Eleanor Batty, et al. Yass: Yet another spike sorter applied to large-scale multi-electrode array recordings in primate retina. *BioRxiv*, pages 2020–03, 2020.

[22] Keming Chen, Yangtao Jiang, Zhanxiong Wu, Nenggan Zheng, Haochuan Wang, and Hui Hong. Htsort: Enabling fast and accurate spike sorting on multi-electrode arrays. *Frontiers in Computational Neuroscience*, 15:657151, 2021.

[23] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[24] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[25] Cole Hurwitz, Kai Xu, Akash Srivastava, Alessio Buccino, and Matthias Hennig. Scalable spike source localization in extracellular recordings using amortized variational inference. *Advances in Neural Information Processing Systems*, 32, 2019.

[26] Julien Boussard, Erdem Varol, Hyun Dong Lee, Nishchal Dethe, and Liam Paninski. Three-dimensional spike localization and improved motion correction for neuropixels recordings. *Advances in Neural Information Processing Systems*, 34:22095–22105, 2021.

[27] Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant contrastive learning. *arXiv preprint arXiv:2111.00899*, 2021.

[28] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.

[30] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[31] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[32] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, mar 2022.

[33] Charlotte Loh, Thomas Christensen, Rumen Dangovski, Samuel Kim, and Marin Soljačić. Surrogate- and invariance-boosted contrastive learning for data-scarce applications in science. *Nature Communications*, 13(1), jul 2022.

[34] Cole Hurwitz, Nina Kudryashova, Arno Onken, and Matthias H Hennig. Building population models for large-scale neural recordings: Opportunities and pitfalls. *Current opinion in neurobiology*, 70:64–73, 2021.

[35] Juan A Gallego, Matthew G Perich, Raeed H Chowdhury, Sara A Solla, and Lee E Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, 23(2):260–270, 2020.

[36] Zoe C Ashwood, Nicholas A Roy, Iris R Stone, International Brain Laboratory, Anne E Urai, Anne K Churchland, Alexandre Pouget, and Jonathan W Pillow. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience*, 25(2):201–212, 2022.

[37] Ding Zhou and Xue-Xin Wei. Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-vae. *Advances in Neural Information Processing Systems*, 33:7234–7247, 2020.

[38] Omid G Sani, Hamidreza Abbaspourazad, Yan T Wong, Bijan Pesaran, and Maryam M Shanechi. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience*, 24(1):140–149, 2021.

[39] Cole Hurwitz, Akash Srivastava, Kai Xu, Justin Jude, Matthew Perich, Lee Miller, and Matthias Hennig. Targeted neural dynamical modeling. *Advances in Neural Information Processing Systems*, 34:29379–29392, 2021.

[40] Ran Liu, Mehdi Azabou, Max Dabagia, Chi-Heng Lin, Mohammad Gheshlaghi Azar, Keith Hengen, Michal Valko, and Eva Dyer. Drop, swap, and generate: A self-supervised approach for generating neural activity. *Advances in neural information processing systems*, 34:10587–10599, 2021.

[41] Mehdi Azabou, Mohammad Gheshlaghi Azar, Ran Liu, Chi-Heng Lin, Erik C Johnson, Kiran Bhaskaran-Nair, Max Dabagia, Bernardo Avila-Pires, Lindsey Kitchell, Keith B Hengen, et al. Mine your own view: Self-supervised learning through across-sample prediction. *arXiv preprint arXiv:2102.10106*, 2021.

[42] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, pages 1–9, 2023.

[43] Glen D Brown, Satoshi Yamada, and Terrence J Sejnowski. Independent component analysis at the neural cocktail party. *Trends in neurosciences*, 24(1):54–63, 2001.

[44] Eugen-Richard Ardelean, Andreea Coporîie, Ana-Maria Ichim, Mihaela Dînșoreanu, and Raul Cristian Mureșan. A study of autoencoders as a feature extraction technique for spike sorting. *Plos one*, 18(3):e0282810, 2023.

[45] David A McCormick, Barry W Connors, James W Lighthall, and David A Prince. Comparative electrophysiology of pyramidal and sparsely spiny stellate neurons of the neocortex. *Journal of neurophysiology*, 54(4):782–806, 1985.

[46] Thomas Euler, Silke Haverkamp, Timm Schubert, and Tom Baden. Retinal bipolar cells: elementary building blocks of vision. *Nature Reviews Neuroscience*, 15(8):507–519, 2014.

[47] Hongkui Zeng and Joshua R Sanes. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9):530–546, 2017.

[48] Nathan W Gouwens, Staci A Sorensen, Fahimeh Baftizadeh, Agata Budzillo, Brian R Lee, Tim Jarsky, Lauren Alfiler, Katherine Baker, Eliza Barkan, Kyla Berry, et al. Integrated morphoelectric and transcriptomic classification of cortical gabaergic cells. *Cell*, 183(4):935–953, 2020.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

https://doi.org/10.52202/075280-1832

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[51] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[52] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1857–1865, Red Hook, NY, USA, 2016. Curran Associates Inc.

[53] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[54] The International Brain Laboratory, Valeria Aguillon-Rodriguez, Dora Angelaki, Hannah Bayer, Niccolo Bonacchi, Matteo Carandini, Fanny Cazettes, Gaelle Chapuis, Anne K Churchland, Yang Dan, et al. Standardized and reproducible measurement of decision-making in mice. *Elife*, 10, 2021.

[55] Mayo Faulkner Chapuis, Kenneth D Harris, Julia M Huntenburg, Cole Hurwitz, Hyun Dong Lee, Liam Paninski, Cyrille Rossant, Noam Roth, Nicholas A Steinmetz, Charlie Windolf, et al. Spike sorting pipeline for the international brain laboratory. *channels*, 10:6, 2022.

[56] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

[57] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[58] Julien Boussard, Charlie Windolf, Cole Hurwitz, Hyun Dong Lee, Han Yu, Olivier Winter, and Liam Paninski. Dartsort: A modular drift tracking spike sorter for high-density multi-electrode probes. *bioRxiv*, pages 2023–08, 2023.

[59] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training, 2021.

[60] Payal Dhar. The carbon impact of artificial intelligence. *Nat. Mach. Intell.*, 2(8):423–425, 2020.