Neurosim: Extracellular Electrophysiological Simulation and Ground-Truth Comparative Analysis of Spike Sorting Algorithms

Chenhao Bao   Mentor: Adam Charles

Dept. of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

Introduction

- The extracellular electrophysiological recording is one of the well-established methodologies tracking neuron activities *in vivo*, which harbor functional information pertaining to the workings of the brain, thereby providing insights into the underlying mechanisms governing behavior and mental processes. Despite the passage of several decades since its initial discovery and utilization, extracellular electrophysiological recording continues to be extensively employed and undergo continuous refinement in contemporary neuroscience research.

- However, spike sorting, the process of accurately and precisely inferring spikes from extracellular electrophysiological recordings, remains a critical challenge. As mathematically categorized as a blind source separation problem, its unsupervised nature gives rise to interpretability ambiguities and validation difficulties. The obstacles are further exacerbated by the growing trend towards utilizing increasingly dense high-density multi-electrode arrays (MEA), e.g. Neuropixels probes [4, 5], which typically comprise hundreds or thousands of channels.

- Several spike sorting algorithms have been proposed over the years to tackle the spike sorting challenge in high-density MEA (see Table 1). Notwithstanding the primary processing pipeline is akin, nuances in configuration parameters, as well as the different strategies employed for spike detection, feature extraction, and clustering, introduce distinct inherent biases unique to each spike sorting algorithm. These biases significantly impact the outcomes and as a consequence, lead to wide variations in sorting results when different algorithms are utilized.

- Despite the fact of continuous emergence and advancement of spike sorting algorithms along with their widespread adoption in recent research, surprisingly few studies have investigated and compared these algorithms while evaluating their performances, and a lack of established standards for spike sorting persists. Such circumstances can induce substantial perplexity among researchers, given the plethora of spike sorting algorithm choices available. In fact, the laborious and technically challenging nature of inter-algorithm comparisons results in the limitation for many researchers to selecting only one single algorithm and subsequently adhering to it, thereby leading to a fragmented software ecosystem that hinders reproducibility, benchmarking, and collaborations.

- To uniformly benchmark the performances of spike sorting algorithms, extracellular recording simulation presents a promising avenue. Compared to real recording datasets, employing simulated datasets has shown several advantages, including ground truth validation, various parameter control for systematic evaluations, and cost-effectiveness.

Table 1. Current overview of spike sorting algorithms

Name	Language	Base Processor	Clustering Algorithms
Kilosort2.5	MATLAB + CUDA	GPU	Scaled K-means clustering
Kilosort3	MATLAB + CUDA	GPU	Scaled K-means clustering
Kilosort4	Python + CUDA	GPU	Graph-based clustering
SpyKING CIRCUS	Python	CPU	DPCLUS
JRCLUST	MATLAB + CUDA	GPU	DPCLUS
IronClust	MATLAB + CUDA	GPU	DPCLUS
HerdingSpikes2	Python	CPU	Mean shift
MountainSort5	Python	CPU	ISO-SPLIT
Tridesclous	Python	CPU	sawchaincut, pruningshears
YASS	Python + CUDA	GPU	MFM

Highlights

- We conducted extracellular electrophysiological simulations to generate simulated recording datasets for systematic evaluation of different spike sorting algorithms, especially paying attention to different drift conditions.
- We then applied several prevalent spike sorting algorithms to the simulated datasets and compared their results with each other as well as with the ground truth settings during the simulation.
- Comparison results were summarized and discussed, highlighting the different biases of various spike sorting algorithms.
- Our study can provide useful guidance for researchers in selecting appropriate spike sorting algorithms and may shed light on the improvements in the next-generation spike sorting algorithms.

Methods

Simulation
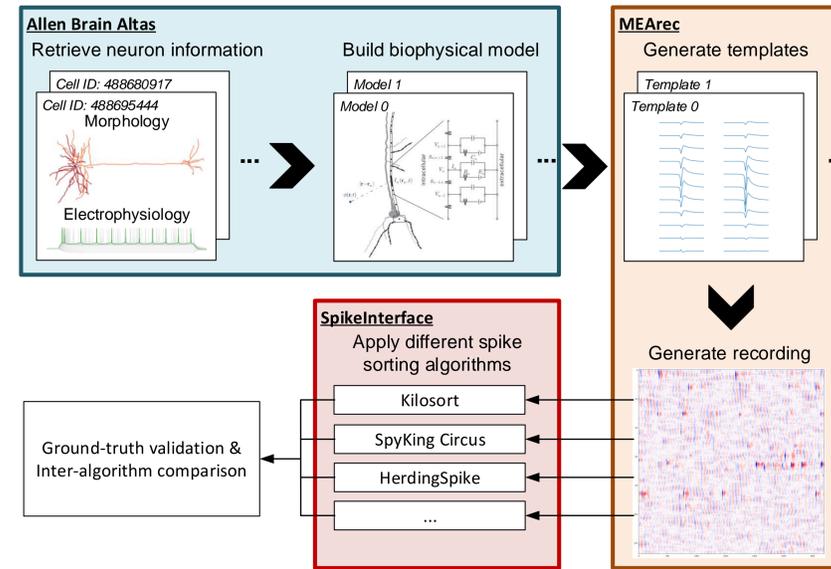
Figure 1. Block diagram of Neurosim.

Evaluation
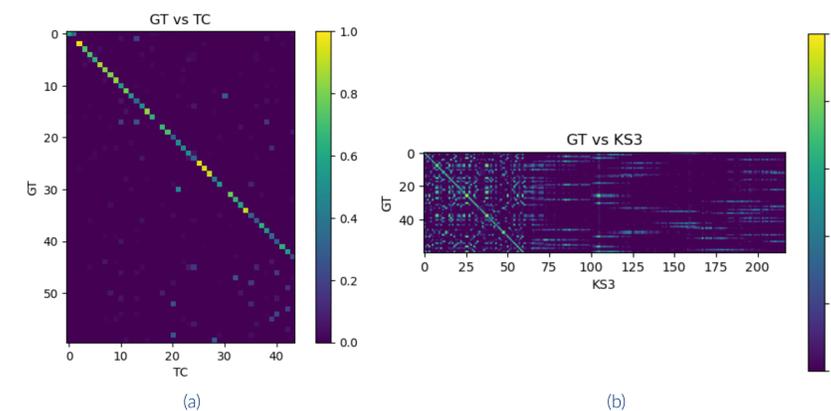
Figure 2. Examples of ground truth validation. (a) Spike time agreement matrix between ground truth units and detected units from Kilosort2.5. (b) Template waveform cosine similarity matrix between ground truth units and detected units from Kilosort3.

Computation of spike time agreement score and waveform cosine similarity score between l-th ground truth unit and k-th detected unit:

$$n_{l,k}^{\text{match}} = \# \left\{ i \mid \exists j, \left| t_i^{(l)} - t_j^{(k)} \right| < \Delta \right\}, \tag{1}$$

$$\text{spike\_time\_agreement\_score}_{l,k} = \frac{n_{l,k}^{\text{match}}}{\#\text{spike}_l^{GT} + \#\text{spike}_k^{D} - n_{l,k}^{\text{match}}}, \tag{2}$$

$$\text{waveform\_cosine\_similarity\_score}_{l,k} = \frac{flatten\left(\text{template}_l^{GT}\right) \cdot flatten\left(\text{template}_k^{D}\right)}{\left| flatten\left(\text{template}_l^{GT}\right) \right| \cdot \left| flatten\left(\text{template}_k^{D}\right) \right|}. \tag{3}$$

Results

Spike time validation
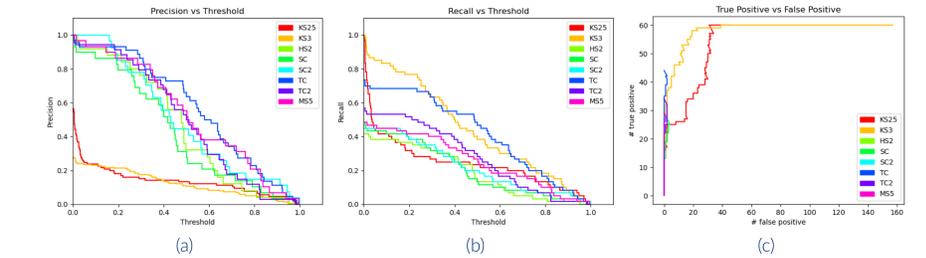
Figure 3. Evaluation of spike sorting algorithms on a simulated recording based on spike time agreement matrices. The recording data used was simulated with fine-tuned parameters at comprehensive consideration of traditional recording obstacles, including probe drifting, neuron bursting, spike diffeomorphism, recording jittering, noise, etc. The precision (a) and recall (b) were computed at different matching thresholds on the spike time agreement matrices from 1 to 0 to determine matching (true positive) between ground truth units and detected units. Precision is defined as # true positive units divided by # ground truth units, and recall is defined as # true positive units divided by # detected units. Note that a theoretically-perfect spike sorting algorithm should always keep precision = 1 and recall = 1. (c) The number of true positive units vs. the number of false positive units at different detection thresholds.

Template waveform validation
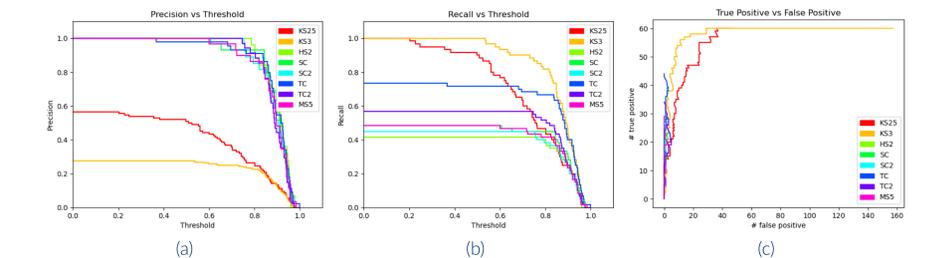
Figure 4. Evaluation of spike sorting algorithms on a simulated recording based on template waveform cosine similarity matrices.

Conclusions

- Kilosort biased towards "over-splitting" and KS3 splitted even more aggressively than KS2.5, achieving higher recall with the trade-off of lower precision. Yet, manual curation (mostly merges rather than splits) can mitigate the side effects of lower precision.
- Spike sorting algorithms except for Kilosort biased towards higher precision, paying more efforts to guaranteeing the correctness of positive-reported units, at the cost of lower recall. They have similar performance in extracting the template waveform shape.
- Tridesclous showed good performance with both high precision and recall in extracting the spike trains.

References

[1] Allen Institute (2022). Allen cell types database.

[2] Buccino, A. P. and Einevoll, G. T. (2021). Mearec: A fast and customizable testbench simulator for ground-truth extracellular spiking activity. *Neuroinformatics*, 19(1):185–204.

[3] Buccino, A. P., Hurwitz, C. L., Garcia, S., Magland, J., Siegle, J. H., Hurwitz, R., and Hennig, M. H. (2020). Spikeinterface, a unified framework for spike sorting. 9:e61834.

[4] Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç. Barbic, M., Blanche, T. J., Bonin, V., Couto, J., Dutta, B., Gratiy, S. L., Gutnisky, D. A., Häusser, M., Karsh, B., Ledochowitsch, P., Lopez, C. M., Mitelut, C., Musa, S., Okun, M., Pachitariu, M., Putzeys, J., Rich, P. D., Rossant, C., Sun, W.-l., Svoboda, K., Carandini, M., Harris, K. D., Koch, C., O'Keefe, J., and Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236.

[5] Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539):eabf4588.