Assessing cross-contamination in spike-sorted electrophysiology datasets

Jack Vincent[1, 2, 3], Michael Economo[1, 2, 3, 4, 5]

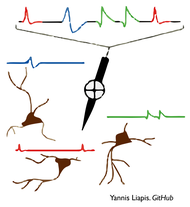
[1] Biomedical Engineering, Boston University, Boston, MA [2] Training Program in Quantitative Biology & Physiology, Boston University, Boston, MA

[3] Neurophotonics Center, Boston University, Boston, MA [4] Center for Systems Neuroscience, Boston University, Boston, MA [5] Photonics Center, Boston University, Boston, MA

Motivation

- Technical advances in extracellular electrophysiology probes have massively expanded neuron counts in recording datasets[1]
- The work of sorting these units is being increasingly offloaded to automated sorting algorithms
- Current metrics for assessing the performance of these algorithms provide only indirect and heuristic measures of sorting quality[2]
- Foremost among these metrics is the interspike interval (ISI) violation fraction, yet the exact relationship between ISI violations and underlying unit isolation has never been thoroughly characterized

0 *What is spike sorting?*



Yannis Liapis. GitHub

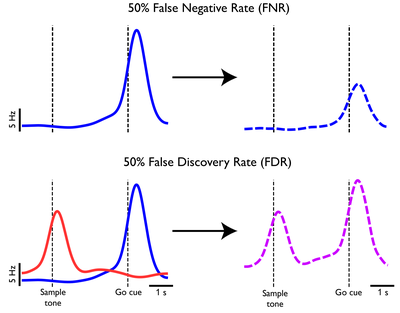
- Demultiplexing of extracellular voltage recordings into component neuronal sources
- Relies on individual differences in spike waveform templates within and across electrodes

1 *What is "good" spike sorting?*

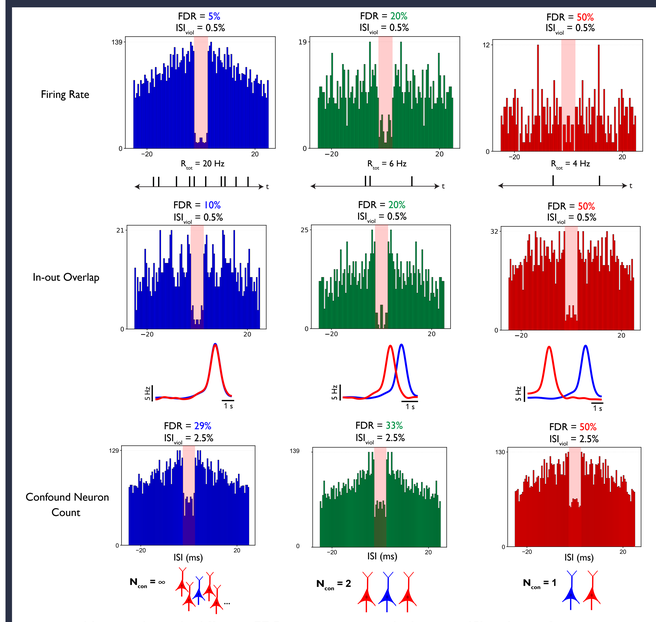
- The avoidance of false negatives and, most critically, **false positives** in sorted units

False Negative Rate (FNR) = $\frac{FN}{FN+TP}$ = fraction of true spikes missed

False Discovery Rate (FDR) = $\frac{FP}{FP+TP}$ = fraction of misassigned spikes

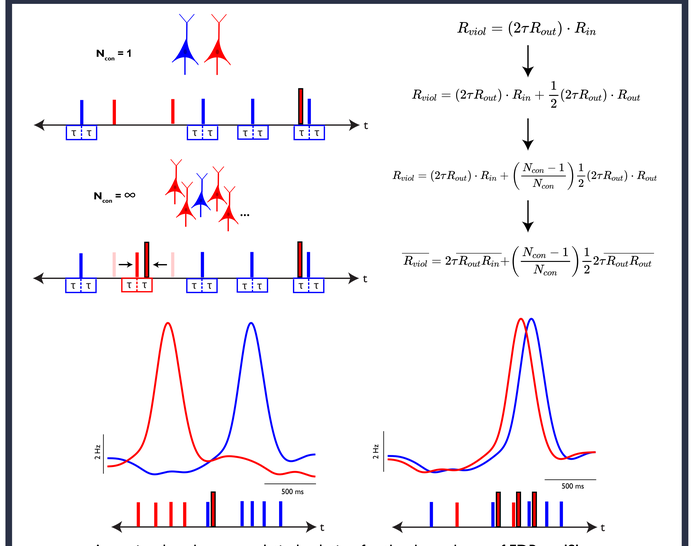


2 *The danger of using ISI violations as a spike sorting quality metric*



- Units with vastly different FDRs can present with the same ISI violation fractions, given variation in underlying conditions

3 *Derivation of the relationship between ISI_{viol} and FDR*



$$R_{viol} = (2\tau R_{out}) \cdot R_{in}$$

$$R_{viol} = (2\tau R_{out}) \cdot R_{in} + \frac{1}{2}(2\tau R_{out}) \cdot R_{out}$$

$$R_{viol} = (2\tau R_{out}) \cdot R_{in} + \left(\frac{N_{con}-1}{N_{con}}\right)\frac{1}{2}(2\tau R_{out}) \cdot R_{out}$$

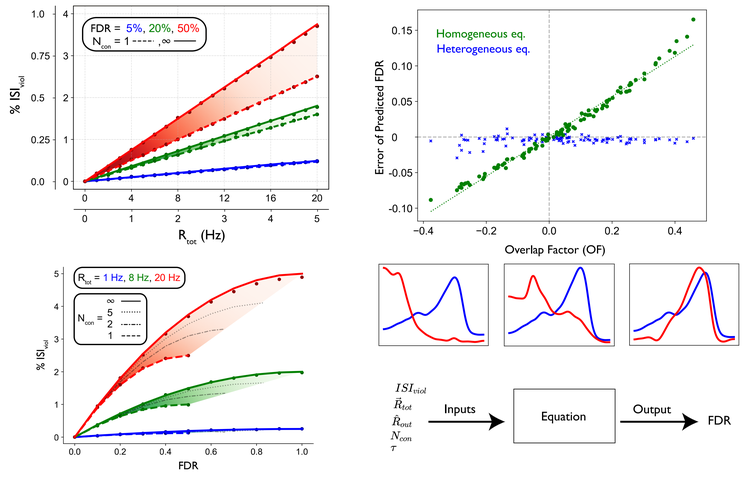
$$\overline{R_{viol}} = 2\tau \overline{R_{out}R_{in}} + \left(\frac{N_{con}-1}{N_{con}}\right)\frac{1}{2}2\tau\overline{R_{out}R_{out}}$$

- A previously unknown analytical solution for the dependence of FDR on ISI_{viol}

Conclusions

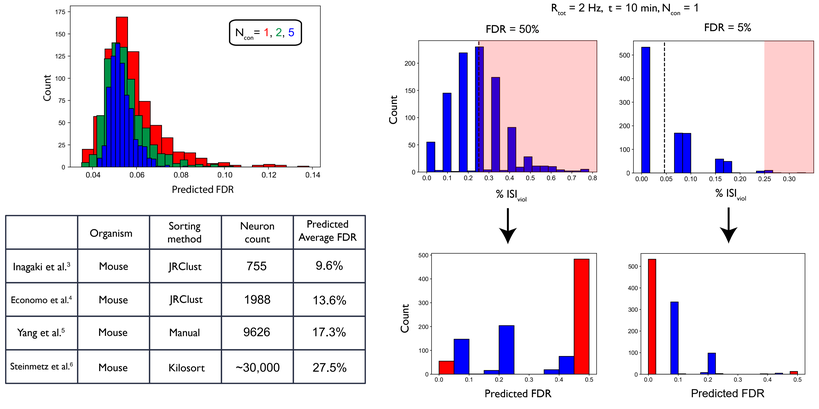
- ISI violations alone are not an accurate measure of underlying cross-contamination in spike-sorted data
- The relationship between ISI violation fraction and FDR depends, in order of importance, on the total firing rate of the sorted unit, its temporal overlap with sources of confound spikes, and the number of contaminant neurons contributing confound spikes
- An equation has been derived describing this relationship in near perfect agreement with simulated data, as well as a methodology for its application to real data
- Given the stochasticity of ISI violations, the equation is a poor predictor of unit-level FDRs, but a good predictor of population-level FDR

4 *Prediction of FDR from ISI_{viol} using ground truth simulated data*



- Across a range of different firing rates, in-out cluster overlap factors, and confounding neuron counts, predicted FDR retains high fidelity with simulated FDR
- Inputs able to be known with higher or full confidence are the most important contributors to the final prediction

5 *Prediction of FDR in real data*



	Organism	Sorting method	Neuron count	Predicted Average FDR
Inagaki et al.[3]	Mouse	JRClust	755	9.6%
Economo et al.[4]	Mouse	JRClust	1988	13.6%
Yang et al.[5]	Mouse	Manual	9626	17.3%
Steinmetz et al.[6]	Mouse	Kilosort	~30,000	27.5%

References

1. Steinmetz et al. *Science* (2021).
2. Hill et al. *Journal of Neuroscience* (2011).
3. Inagaki et al. *Nature* (2019).
4. Economo et al. *Nature* (2018).
5. Yang et al. *Nature Neuroscience* (2022).
6. Steinmetz et al. *Nature* (2019).

Economo Lab