Application of a point process GLM with precise spiking information to Neural Latent Inference*

Greggory R. Heller

Abstract— **Experimentally collected extracellular recordings are becoming increasingly high dimensional, however successful analysis methods often drastically reduce both the temporal resolution and dimensionality of the data. Spike times are reduced to rates, and the number of meaningful dimensions is often reduced to a lower dimensional latent space. This is done seemingly in direct defiance of the apparent importance of spike timing and the apparent complexity of the brain. The "Neural Latents Benchmark[1]"provides standardized datasets and prediction criteria to allow the direct, unbiased comparison of such latent variable models. This offers a novel opportunity to compare to both superior and baseline latent variable models. We use this dataset to compare the authors baseline "gaussian smoothed" rate based model to a model that more flexibly infers the appropriate relationships to precise spike times near the predicted time bin of interest. We find that allowing precise spike timing and history information to be incorporated into a poisson process GLM can in some cases lead to an increase in resolution of predictions relative to models built using smoothed data, however the overall performance is still best with a simple gaussian smoothed model. Further optimization is necessary to make any conclusions.**

I. Introduction

The importance of spike timing is one of the great unanswered questions in neuroscience. It is easy to acknowledge that spike timing is probably important when you focus on the single neuron, yet there is scant in-vivo or functional evidence that the precise timing of spikes matters. (See 2 for a broader review of this debate) Reducing the temporal resolution of a dataset to a rate code can greatly simplify the subsequent analysis of the dynamics. We will certainly not settle this debate here, however we believe that assessing the importance of spike timing in latent variable mzodels may provide a path forward, both in the development of interpretable latent variable models and in our understanding of how spike timing is incorporated into the neural code.

In one view, the gaussian filters are producing an estimate of the latent "true" or "instantaneous" firing rate that the neuron is attempting to capture, but doomed to fail because of the spiking nature of its output[3]. In this view the smoothed signal is actually the important one to the system and the spikes are only necessary out of biological reality. Proponents often note the trial to trial variability of spiking of single neurons are highly variable, and that the firing rate is more likely to be a reliable downstream readout. It is unclear how the downstream neurons would access this latent signal. Perhaps the redundance of the signal across neurons enables downstream neurons to similarly infer this latent "rate," or perhaps the temporal filtering of chemical signaling at synapses, receptor and membrane time constants and dendrites results in a downstream signal that is approximately "rate-like." Whichever method is utilized by downstream areas, this view posits that the the "rate" is the communication currency of the brain.

In another view, what the gaussian smoothing is doing, and the reason it is mildly successful, is that smoothing incorporates spike time information from both before and after the predicted time bin into the present signal. Smoothing is likely to do better than a latent method that includes only the binary spike/no-spike status of all the regressor units at a single time point simply because it incorporates more information. However, gaussian smoothing incorporates temporally distant information in a rather crude and inflexible way. While it is reasonable to assume that temporally-near activity will carry more predictive weight than temporally distant activity, forcing the filter to be gaussian is unnecessarily limiting. We know that the excitatory post-synaptic potential, which is our best estimate of the downstream impact of a spike on post synaptic partner neurons, has a consistent fast rise time and longer decay[4]. Moreover, spiking interactions between neurons can be detected in-vivo on this timescale as "sharp-peaks" in the cross-correllagram between the spike times of two neurons[5]. Thus the rate-code vs spike code debate highlights a tension between what we know the downstream effect is constrained by biology to be, and what we think the system might be trying to accomplish with that signal.

We choose to incorporate information from both before and after the predicted time bin, as with gaussian smoothing, but to allow our model to fit separate parameters to each time offset, creating custom "coupling filters" for each regressor unit - target unit pair.

Our goal was inspired by the work of Felipe Gerhard, along with Mark Kramer and Uri Eden who showed that precise spike timing information can be flexibly incorporated into a generalized linear poisson process model to infer anatomy connectivity from functional activity alone[6]. While most attempts to perform weight matrix inference are doomed to fall short and land in the realm of "functional connectivity" (which would more aptly be named functional correlations)[7,8], Uris work showed that complete knowledge of the spiking activity of a small network can be used to infer bio-realistic "coupleing filters" that mirror the known anatomical connections.

There are many differences between the Crab STG where the technique was first applied, and the neural latents benchmark. The complexity of the model organism, the number of neurons recorded, and perhaps most importantly

*Research supported by Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences.

G. T. Heller is with the Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences, Cambridge, MA 02139 USA (corresponding author to provide phone: 361-596-3822; e-mail: greggh@mit.edu).

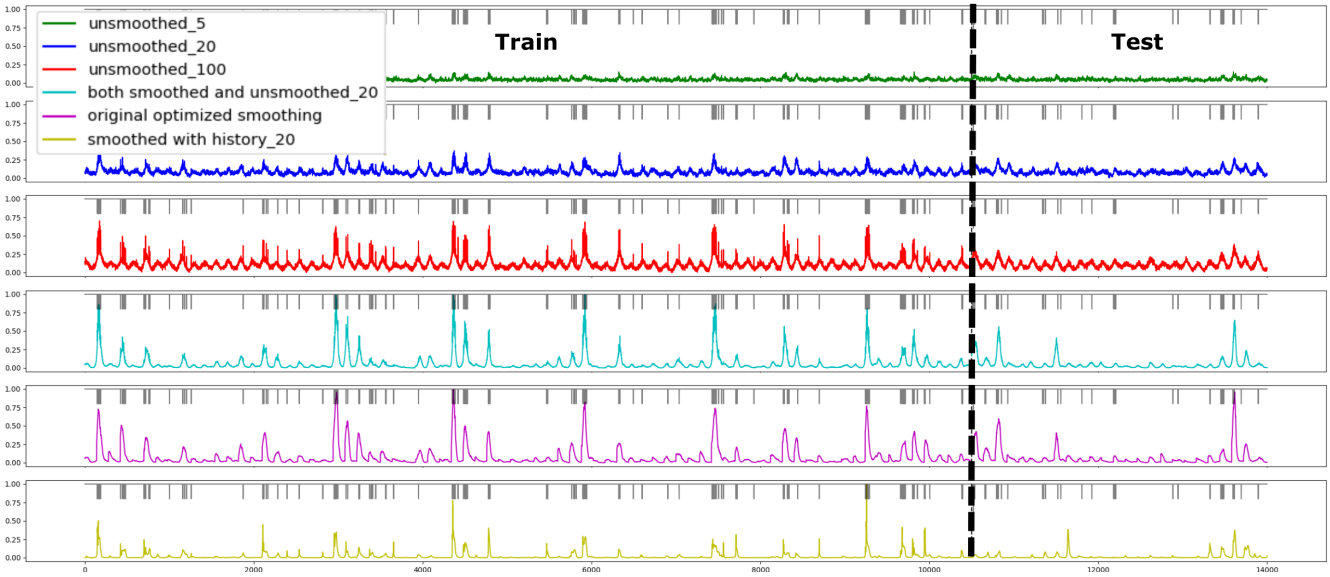

Figure 1. Example predictions. Predicted rates from each of the 6 for models for the same target neuron (spikes in gray, duplicated for each trace for reference). Rates have ben scaled and shifted to show detail.

the number of unobserved neurons in the circuit. However we believe that the successful application of the method in a simpler system indicates that precise spike times combined with appropriate coupling filters can be a powerful technique for predicting held-out target units.

It should be noted that smoothing filters, or even gaussian smoothing filters are in the solution space of our model. If we find primarily coupling filters that are performing simple smoothing operations this would suggest that gaussian smoothing does indeed preserve all necessary information and thereby support the idea of a "rate-based" neural code. However if we find coupling filters with higher complexity, or high frequency elements are indicative of specific dependencies that happen at specific timescales and this would support the idea that the precise timing of spikes has a direct impact on downstream neurons.

Another line of evidence that led us to pursue a model with access to precise spike timing information was that the benchmark was won by a Neural Data Transformer (NDT)[9]. Other baseline models that had access to unsmoothed spiking data also did well compared to the smoothing baseline. It is unclear whether the access to precise spike time and spike history information for each regressor neuron was valuable to these models, or whether they primarily benefit from more parameters and architectural complexity. Our aim was to use a model that was not a deep neural net, with many fewer parameters but nevertheless had access to precise spike time and history information for each regressor unit.

To this end, we generated a number of alternative point process models that attempted to incorporate additional information about the spike timing or history of neural activity. None of the models were able to surpass the performance of the original optimized gaussian smoothing,. This may suggest that spike timing is indeed not important for information transmission in the brain, however we would caution against this conclusion given how overly simplistic and un-optimized our attempts were. A careful investigation of the alternative models may give insight to why this was the case and illuminate a path forward toward interpretable yet accurate models that retain access spike timing information.

II. Results

A. Model Description

We created 6 different models in an attempt to investigate whether of spike timing and history help predict the activity of co-active neurons. We first recreated the published model with gaussian smoothing paired with a GLM. Briefly, we used the same filter parameters (std = 60 ms) to convert spike trains to rates and then used a poisson process Generalized Linear Model to fit a single coupling strength between each regressor unit and each target unit.

Next we create models that included different amount of spike timing information. To match the properties of the gaussian filter we include spike timing from both the future and the past for 3 different window sizes (+/-25ms, +/-100ms and +/-500ms coupling windows) and fit a separate coupling strength for each unit at each time lag within the coupling window, see (1).

$$\lambda(t) = \beta_0 + \sum_{n=1}^{N} \sum_{t=-T}^{T} \beta_{(n,t)} x_n(t) \quad (1)$$

for N regressor units, T - the halfwidth of the coupling window $(25, 100, 500)$, and $X_n(t)$ the activity of the n th regressor unit at time offset t.

The coupling filter is then the vector of coupling strengths between a single regressor unit and a single target unit ordered by time.

We also created a combined model with access to both smoothed (std=60ms) and unsmoothed data (+/-25ms) and a control model that included the history of the smoothed trace (+/-25ms coupling window) that would have the same number of parameters as the unsmoothed model with the same coupling window.

We then fit each model using the smallest dataset in the benchmark - mc_maze_small. We used the benchmarks "eval split" to partition the data resulting in 107 regressor units, 35 target units, 75 trials for training, and 25 trials for testing. To further reduce the size and sparcity of the dataset, we binned the data into 5ms bins (as recommended by the benchmark founders). Most bins contained one or zero spikes with a few exceptions.

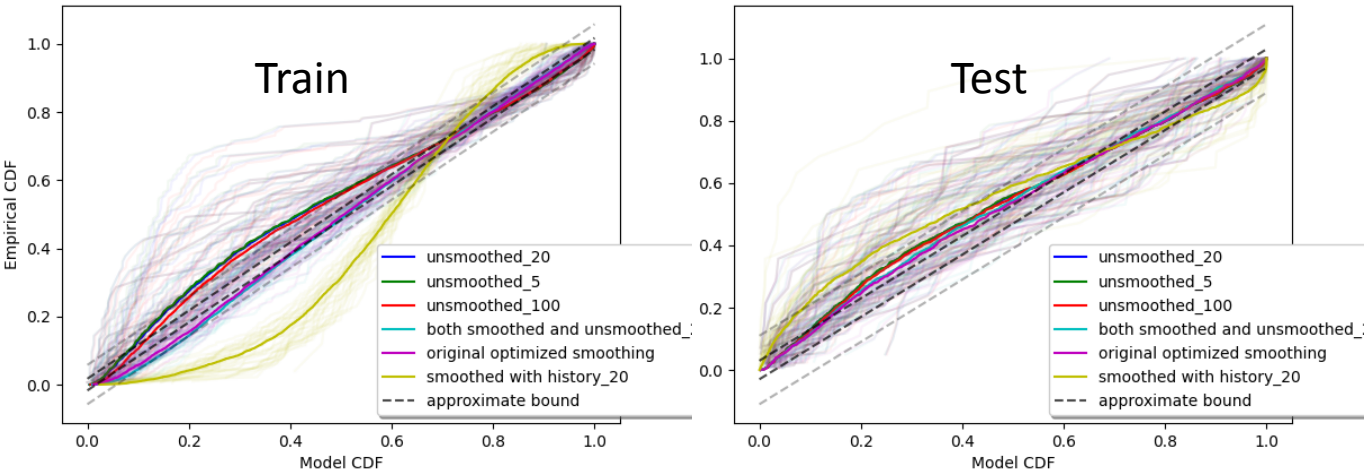

Figure 2. Model performance fails to capture neural activity. KS plots for all 6 models using the time rescaling theorum (same colors as Fig.1) for both single target units (transparent lines) and combining all target units (dark lines). Dashed lines are approximate bounds on the model fit. (again dark for combined and transparent for single units.

B. *Qualitative model performance*

To assess the qualitative model performance we plotted the predicted rates alongside the spike rasters for example target units (Fig. 1). As expected, smoothed model predicted rates often rise well before the onset of spiking, and also require additional time to fall back to baseline after spiking has ceased. The standard deviation of the smoothing filter places an effective limit on the resolution of the rate predictions. Promisingly, the unsmoothed model with a coupling window of +/- 500 ms showed much higher resolution predictions than the smoothed models, evidenced by "spikes" in the rate trace that correspond to single spikes in the raster. This behavior disappears when applied to the test data indicating that these high resolution predictions were a result of overfitting with over 20,000 parameters (far greater than the 10500 training bins).

Another interesting qualitative feature of the unsmoothed models is that they seem to capture the trial structure of the task very well, indicated by the largely uniform oscillations. Each of the 100 oscillations corresponds to a separate trial, and this behavior does continue during the test data. This is also observed to some extent in the smoothed models, but the oscillation is not as uniform or clean and is sometimes missing altogether. Not all example target neurons showed this periodicity.

Lastly, the total rates for each model were nearly identical, and the distributions of predicted rates were skewed left (as expected for a sparse predicted signal). However for the unsmoothed models, the variance of the predicted rates was much lower, and the lower bound was well above 0, closer to the average firing rate of the target unit. We believe this reflects the relative variance of the inputs between smoothed and unsmoothed models, or perhaps the binary nature of the inputs to the unsmoothed models. It is not clear if this is desirable or not, or if a larger training dataset with more units or more time would change this.

C. *Quantitative model comparison*

To assess the models in a more quantitative way we use the time re-scaling theorem to produce KS plots[10,11] for each target unit, and for all target units combined (Fig. 2). The original gaussian smoothing model performs the best on both the training and the testing data, despite having the fewest parameters by 1-2 orders of magnitude. However, the models all appear to perform much more similarly on the testing data, despite the obvious overfitting of the unsmoothed, high parameter models. Interestingly, the smoothed models reside below the unity line for the training data and above the unity line for the testing data, while the unsmoothed reside above the unity line for both training and testing. The smoothed models are over-predicting the rates during training, but under predicting during testing. The unsmoothed models tend to under-predict everywhere.

Clearly for all models there are many single units who's activity is not well captured (translucent lines, Fig. 2). And all models deviate significantly from the unity line for the combined predictions. Unsurprisingly we these simplistic models fairly to fully explain neural activity.

Finally the control model that includes a history of the smoothed trace performs the worst, likely because it violates the assumptions that the regressors are independent. Time bins within the smoothing kernel will be highly correlated. For this reason we will disregard this model from further discussion.

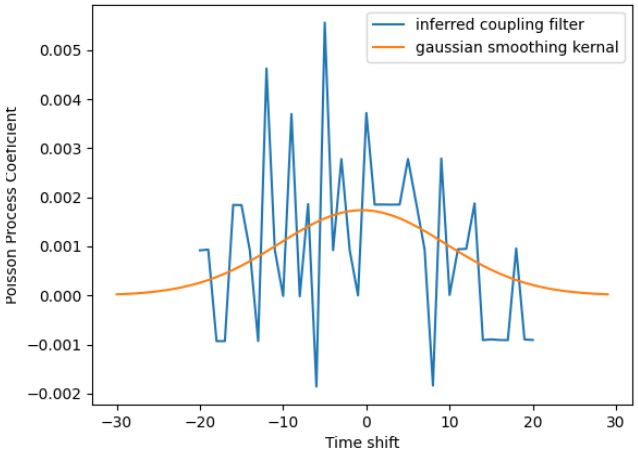

Figure 3. Inspecting the coupling filters. An example coupling filter for both the unsmoothed (+/-100 ms) and smoothed models for the same regressor unit to target unit pair.

D. *Inspecting the coupling filters*

We next assessed the coupling filters for the unsmoothed models. They demonstrated significant unstructured variability as well as high frequency transitions with no apparent patterning (Fig. 3). Some filters rarely deviated from zero. Filters corresponding to the units that were the most strongly coupled may have a slight tendency to peak around 0. Although it is not clear exactly what we should expect realistic coupling filters to look like, it seems likely that these are the result of overfitting and do not represent any meaningful structure in the biology. Clearly the unsmoothed models did not find low frequency smoothing filters (guassian or otherwise) even though they were in the solution space, perhaps indicating that smooth rates do indeed leave out relevant information

E. *The "combined" model does not effectively incorporate any spike timing information*

As a last check we compared the coupling strengths between both models that had access to smoothed data. The coupling strengths between both models were highly correlated, indicating that the inclusion of spike time data did little to account for the signal present in the smoothed data.

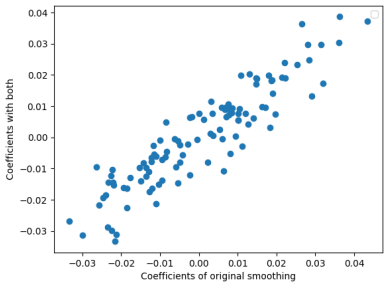

Figure 4. Correlation between the model coefficients for smoothed unit activity for the model containing smoothed activity only (X axis) and for the model containing both smoothed activity and spike times (Y axis).

III. Discussion and Future Directions

In many ways it is not surprising that the unsmoothed models were not able to outperform the published gaussian smoothing model. The Neural Latents Benchmark founders optimized both the kernel standard deviation and the poisson GLM regularization penalty before fitting their optimized model[1]. While the unsmoothed models incorporated many more parameters, we did not attempt to optimize the coupling window, or the regularization penalty.

The high variability, lack of structure and high frequency transitions of the coupling filters was something that was observed previously in attempts to recapitulate known anatomy from activity alone[6]. These filters were deemed "non-physiological" and the authors claim that fitting the coupling filters with splines instead of completely independent coupling strengths was necessary to recapitulate the anatomical results (however it is not entirely clear that it was tested without the splines). Splines require far fewer parameters than the approach we used, while still allowing much more flexibility than gaussian smoothing. This may be a promising addition moving forward, however we note that the time constants of AMPA signaling are on the order of a few milliseconds[4], and any constraints on the filters should be able to accommodate these short timescale interaction without unnecessarily smoothing them. Indeed if the goal is to increase the precision of our spike predictions, then it will likely be necessary to have high frequency transitions in the filters.

The evidence of overfitting indicates that this approach may have been ill suited to apply to such a small dataset. Detecting sharp peak interactions between neurons is relatively rare, requiring both many pairs of neurons, and high spike counts in order to separate the low probability, short time-lag coincident events from the noise floor[5]. Detecting connections that are less direct than this would benefit even more from more data. We are optimistic that applying this technique to one of the larger monkey data sets could be fruitful, or better yet to the Allen Institute mouse dataset that includes in some cases more than 1000 simultaneously spiking units recorded for nearly 3 hours[5]. The Allen Institute Neuropixels dataset was designed to maximize the probability of finding connected pairs by targeting portions of visual areas that are known to communicate mono-synaptically, which will also increase the chances of detecting both short timescale and longer timescale correlations between individual unit spiking.

Lastly, while it is unclear exactly why deep neural nets of various forms perform better in general on the neural latents benchmarks, it is still not clear how much of this is due to the incredibly high number of parameters or to the inherent features of the models. We propose that one important feature might be the ability account for interactions between individual neurons. For example when a downstream effect is only observed when to neurons are active coincidently, or conversely when an inhibitory effect is only evident as the lack of an expected excitatory effect from a specific neuron. We know that there are multiple mechanisms by which this form of interaction could be present in biological neural networks, either through the effect of an intervening, unobserved neuron, or through local interactions within dendritic compartments[12]. The ability for deep nets to account for interactions and flexibly infer hidden states and modulatory effects is something clearly lacking from simpler single layer methods like the point process GLM used here. While deepnets are highly effective at Neural Latent inference, their complexity yields little interpretability and its hard to imagine how they will be useful to scientists unless your primary goal is quite literally predicting the activity of other units. We propose instead that these deepnets should be used as an upper bound to identify when simpler models have come close to maximal performance.

There have been some efforts to incorporate interactions in simpler models, with cross dependance in the history terms [13] or what others call "higher order correlations"[14]. Both methods would be interesting to apply to larger datasets that have become available more recently. Even these methods likely have an explosion of parameters relative to the truly simple models which leads to difficulty interpreting the coupling strengths. Progressively adding complexity akin to the cascade correlation learning architecture[15], or progressively pruning unneeded nodes or hidden units[16] could help build models with the least complex structure necessary for sufficient prediction.

To this end we propose that next steps for this work should be to pursue an optimized model applied to an optimal dataset to

1. account for both short timescale and longer timescale interactions in the coupling filters, while avoiding overfitting

2. Account for possible intricate interaction between the regressors

3. Identify the simplest necessary structure to required approach the performance of optimized deepnets

Acknowledgment

This work relied heavily on the documentation and code provided by Uri Eden and Mark Kramer in their book "Case Studies in Neural Data Analysis," and by the Neural Latents Benchmark team. They are both cited in the references, but we would like to thank them for their well documented and easy to follow python resources.

References

1. Pei, F., Ye, J., Zoltowski, D., Wu, A., Chowdhury, R. H., Sohn, H., O'doherty, J. E., Shenoy, K. v, Kaufman, M. T., Churchland, M., Jazayeri, M., Miller, L. E., Pillow, J., Park, I. M., Dyer, E. L., & Pandarinath, C. (n.d.). *Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity*. Retrieved May 10, 2022, from http://neurallatents.github.io
2. Brette, R. (2015). Philosophy of the spike: Rate-based vs. Spike-based theories of the brain. *Frontiers in Systems Neuroscience*, *9*(November), 151. https://doi.org/10.3389/FNSYS.2015.00151/BIBTEX
3. Kass, R. E., Ventura, V., & Cai, C. (2003). *Statistical smoothing of neuronal data*.
4. Kleppe, I. C., & Robinson, H. P. C. (1999). Determining the Activation Time Course of Synaptic AMPA Receptors from Openings of Colocalized NMDA Receptors. *Biophysical Journal*, *77*(3), 1418–1427. https://doi.org/10.1016/S0006-3495(99)76990-0
5. Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., Groblewski, P. A., Ahmed, R., Arkhipov, A., Bernard, A., Billeh, Y. N., Brown, D., Buice, M. A., Cain, N., Caldejon, S., … Koch, C. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature 2021 592:7852*, *592*(7852), 86–92. https://doi.org/10.1038/s41586-020-03171-x
6. Gerhard, F., Kispersky, T., Gutierrez, G. J., Marder, E., Kramer, M., & Eden, U. (2013). Successful Reconstruction of a Physiological Circuit with Known Connectivity from Spiking Activity Alone. *PLOS Computational Biology*, *9*(7), e1003138. https://doi.org/10.1371/JOURNAL.PCBI.1003138
7. Brinkman, B. A. W., Rieke, F., Shea-Brown, E., & Buice, M. A. (2018). Predicting how and when hidden neurons skew measured synaptic interactions. *PLOS Computational Biology*, *14*(10), e1006490. https://doi.org/10.1371/JOURNAL.PCBI.1006490
8. Marc, D., Mehler, A., & Kording, K. P. (n.d.). The lure of misleading causal statements in functional connectivity research.
9. Ye, J., & Pandarinath, C. (2021). Representation learning for neural population activity with Neural Data Transformers.
10. Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., & Frank, L. M. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, *14*(2), 325–346. https://doi.org/10.1162/08997660252741149
11. *Point Process Generalized Linear Models — Case Studies in Neural Data Analysis*. (n.d.). Retrieved May 10, 2022, from https://mark-kramer.github.io/Case-Studies-Python/09.html
12. Xu, N. L., Ye, C. Q., Poo, M. M., & Zhang, X. H. (2006). Coincidence Detection of Synaptic Inputs Is Facilitated at the Distal Dendrites after Long-Term Potentiation Induction. *Journal of Neuroscience*, *26*(11), 3002–3009. https://doi.org/10.1523/JNEUROSCI.5220-05.2006
13. Okatan, M., Wilson, M. A., & Brown, E. N. (2005). Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural Computation*, *17*(9), 1927–1961. https://doi.org/10.1162/0899766054322973
14. Staude, B., Grün, S., & Rotter, S. (2010). Higher-order correlations in non-stationary parallel spike trains: Statistical modeling and inference. *Frontiers in Computational Neuroscience*, *4*, 16. https://doi.org/10.3389/FNCOM.2010.00016/BIBTEX
15. Fahlman, S. E., & Lebiere, C. (n.d.). The Cascade-Correlation Learning Architecture.
16. Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., & Guttag, J. (2020). *WHAT IS THE STATE OF NEURAL NETWORK PRUNING?* https://github.com/jjgo/shrinkbench.