Robustness of Balanced Networks, Before and After Neuron Loss

Vassilis Alexopoulos, Stephen E. Clarke, Paul Nuyujukian

May 2024

1 Methods

Spiking neural networks (SNNs) use biologically realistic models of in-vivo learning to perform computation. SSNs accomplish this by incorporating key components of heterogeneity, sparsity, temporal adaptation, and feedback between compute units which are present in cortical neural computation.

To that end, we developed an SNN using the Neural Engineering Framework (NEF), *Nengo*, in order to simulate population-level responses to neuron loss. The NEF[1], developed at the University of Waterloo by Chris Eliasmith and Terrence Stewart, seeks to probe the bases of cognitive function by providing a platform to simulate large-scale neural circuits.

We sought to enable rapid system iteration across a vast space of model hyperparameters in our implementation. These hyperparameters reflected foundational biological mechanisms in neural computation, allowing us to explore the causal impact of these distinct mechanisms on downstream metrics, like decoder performance.

We imposed the following set of constraints on our network:

1. Every SNN instantiation featured a pool size of both excitatory (E) and inhibitory (I) neurons constrained such that the size of the excitatory population was never more than ninety percent of the size of the total population (hereafter referred to as the E/I ratio of the system). Population balance between E/I microcircuits captures a fundamental tradeoff between computation accuracy and representation capacity, wherein excitatory cells provide higher fidelity stimulus encoding, while inhibitory cells enable more complexity in the computations being performed. We examined the impact of a variable E/I ratio on network performance before and after neuron loss.

2. The excitatory and inhibitory pools abide by Dale's Principle[2], according to which neurons can be either excitatory or inhibitory but not both.

[1] https://compneuro.uwaterloo.ca/files/publications/stewart.2012d.pdf
[2] https://pubmed.ncbi.nlm.nih.gov/24945778/

3. Heterogeneous and sparse connective structure between excitatory and inhibitory units, featuring recurrent and feedforward pathways (Figure 1). This follows work by Hennequin et al., who showed that traditional homogeneous inhibition is insufficient to balance chaotic firing rate fluctuations in SNNs[3]. We ensured heterogeneity by randomly drawing connection weights from Gaussian distributions. We defined sparsity as neurons in each excitatory or inhibitory population communicating with a randomly selected subset of their neighbors across each connection type (Figure 2, 3)

We presented our network with two different tasks:

1. A continuous tracking task of a simple layered sinusoid of multiple frequencies. The network was fed this signal and we sought to have it match its output as closely as possible (Figure 4).

2. A blind source separation (BSS) task derived from Isomura and Firston's work on in vitro networks[4]. Blind source separation involves inferring the hidden causes of observed sensations by reversing the mapping from causes to effects. This task is often performed by humans attempting to distinguish an individual's voice from the noise of a crowd.

   Our modeled blind source separation task featured a set of m "electrode" sites discretely activated by n sources. Each source randomly stimulated m/n sites (electrode sites could be stimulated by more than one source). The n sources had a 50% source probability, while the m electrode sites had a 75% or 25% masking probability (their likelihood to accurately transmit a source's input) (Figure 6).

   We fed seemingly random activity from the m electrode sites into our network and sought to have it infer which of the n sources were responsible for those activity patterns. Our training and testing sets featured the same underlying source to electrode connectivity. We utilized the number of sources and the number of associated electrode sites as a lever for task complexity, with a lower m/n ratio associated with a more complex task.

After optimizing our generated networks for task performance, we applied a random (70% - 90%) ablation of the population (across excitatory and inhibitory neurons) that effectively zeroed those units' contributions to the entire network (shown in Figure 3). We then evaluated model performance before and after ablation across the stimuli and levers described above. BSS task complexity coupled with tight model constraints helps push the SNN into a regime where the performance of the approximation depends on underlying neural properties, like system robustness, which our simulations can be used to study. We believe that our prototypical model of cortical microcircuits allows for causal investigations of a population's response to neuron loss that is independent of the in-vivo effects of distributed computation from neighboring neural populations.

[3] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009642
[4] https://www.nature.com/articles/s41598-018-35221-w

2 Results/Claims

1. While performing our blind source separation task, our model's spike-rate activity had an eigenspectrum akin to what is seen in in-vivo neural dynamics (Figures 7, 8). This eigenspectrum is considerably more impoverished in the simpler tracking task, where more than 80% of variance can be explained by one or two principal components (Figure 9).

2. Regardless of input task complexity, network ablation causes a heavy right shift in model eigenspectra (Figures 7, 9). These right shifts accompany seemingly affine transformations of network activity in state space (Figures 10, 11, 12, 13, 14). Squashing of neural trajectories following ablation decreases state separability (Figures 12, 14) negatively affecting BSS classification performance (Figures 15, 16).

3. For the continuous tracking task, we saw that the operating point under which the model performed most similarly before and after ablation was around 70%/30% E/I (Figure 18). This optimum is not evident in solely excitatory SNNs (Figure 17) and remains constant across neuron types; using spiking Izhikevich[5] neurons instead of leaky integrate and fire (LIF) neurons for compute yields the same 70%/30% ratio as most optimal (Figures 19, 20).

4. For either task, a family of similar coactivity matrices enables latent dynamics capable of leading to the same behavioral output (Figures 21, 22, 24, 25). The size and spread of this family is constrained by the E/I balance of the network (Figures 23, 27).

5. For the BSS task, an excitatory pool size in the range of 60% to 70% optimizes pre-ablation model performance (Figure 26) while simultaneously maximizing the distance between coactivity spaces for that performance (Figure 27). Put simply; within this regime the model performs *best* with the *largest family of possible coactivities*.

6. **Future work ideally shows** that within a given constraint space, behavioral deficits due to network ablation can be resolved by imprinting neighboring coactivity on the circuit of interest. Though ablation shifts systems to new coactivity regimes, the size of the possible solution set at a given performance is related to the pre-ablation coactivity. This implies that ablation is simply a mapping of matrix families from one stationary point to another. Computing such a mapping could allow for solving its inverse and restoring network activity by selectively perturbing a few network nodes.

[5] https://www.izhikevich.org/publications/spikes.pdf

3 Discussion

We developed a spiking neural network to simulate cortical dynamics under conditions of excitatory and inhibitory (E/I) balance. In the absence of noise from adjacent neural circuitry, our in vitro network provides a causal platform for understanding how neural ablation affects population level dynamics. Leveraging this ability while varying biologically meaningful model hyperparameters (connective structure, balance conditions, neuron type, etc) yielded results that mimic in vivo data and potentially explain the origins of in vivo morphology.

A trove of literature supports that roughly 30% of the brain is comprised of inhibitory cells. Hendry et al.[6] showed that 25% of the total neural population in monkeys' cerebral cortex is inhibitory. Meanwhile, Alreja et al.[7] showed that inhibitory interneurons occupy a nominal range of 10% to 25% of the total neural population across species and brain regions. Neural spike data also supports these results – applying a spike sorting algorithm to activity from *Rhesus Macaques* during a reaching task yields that **blank%** of data is inhibitory (**Stephen Clarke**).

Our findings add insight to these understandings in three mains ways:

First, spiking neural networks designed from the ground up with biologically informed (and plausible) constraints have an *emergent* optimum in this 30% inhibitory, 70% excitatory regime. This optimum exists in some way across tasks and task complexities (Figures 4, 6), which not only supports the validity of our model, but also highlights a potentially fundamental optimum in neural computation across in vitro and in vivo systems.

Second, at this 70%/30% E/I point, our model was maximally robust to network ablation. Alreja et al.[8] have already shown that in vitro networks biased to be roughly one third inhibitory yield optimal stimulus representation, sensory sparseness, and metabolic efficiency. However, we are the first to map these findings to causal neuron loss. In the continuous task, we defined robustness as the ability of a network's error to remain unchanged following an ablation event (Figures 18, 20). In the BSS task, we defined robustness as the model having maximal pre and post ablation classification performance, while occupying the largest set of possible coactivity spaces (Figures 26, 27). Functionally this means that we've added another in vitro explanation – robustness to lesioning – for why this regime exists in the brain. We believe that this property emerges from population level spike rate adaptation following ablation (Figure 5).

Third, by coupling this optimal E/I balance with our findings showing that a large spread of coactivity families can yield the same behavioral performance (Figures 21, 22, 24, 25 we are well on our way toward coming up with a meaningful metric for a) the extent of a network's ablation and b) our ability to restore an ablated network by selectively perturbing a few network nodes. This is the direction in which we will steer future work.

[6] https://www.jneurosci.org/content/7/5/1503
[7] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009642
[8] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009642

4 Figures

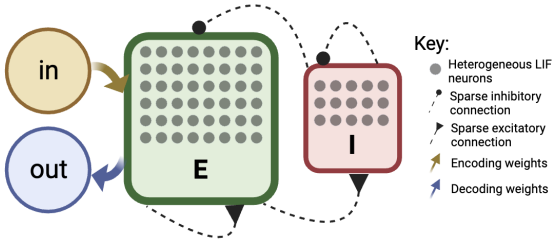

Figure 1: Model SNN structure. Compute is performed by the excitatory population which communicates with layer two inhibitory neurons

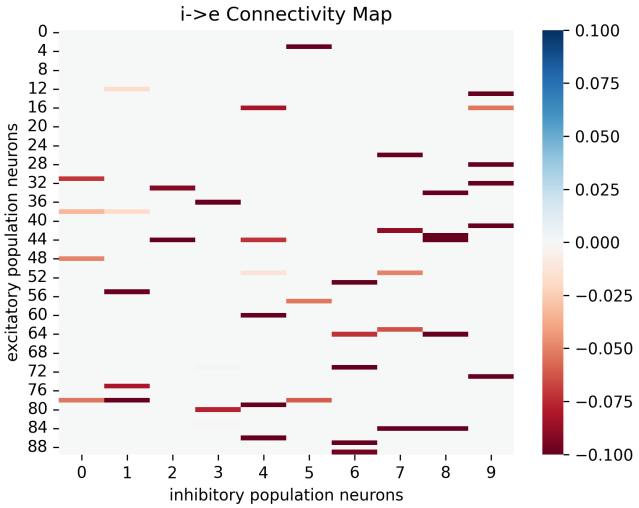

Figure 2: Example of heterogeneous structure in feedforward interaction from inhibitory to excitatory population. This network has an E/I ratio of 90/10, where each inhibitory unit communicates with a random set of its neighbors with each distinct connection having normally distributed weightings.

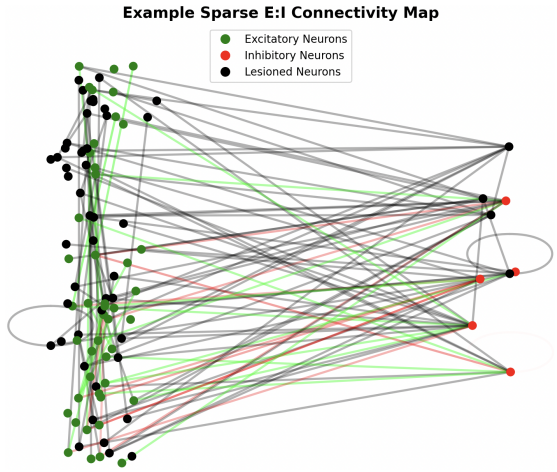

Figure 3: Example of overall structure of network in Figure 2. This includes all feedforward connections (E–I, I–E) and recurrent connections (E–E, I–I). Green units are excitatory (leftmost layer), red units are inhibitory (rightmost layer). Connection strength defined by shading. Black represents ablated neurons/connections.

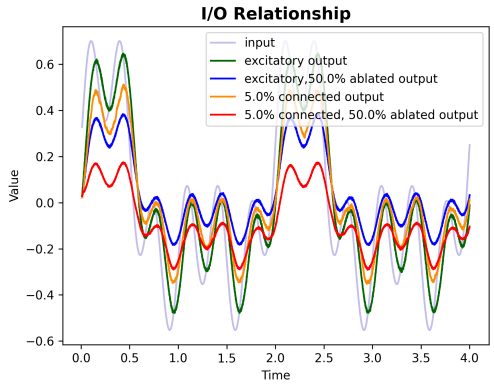

Figure 4: Task 1 – continuous input is in light blue, network tracking of input in warm colors: orange (before ablation) and red (after ablation). Green and dark blue are purely excitatory population controls.

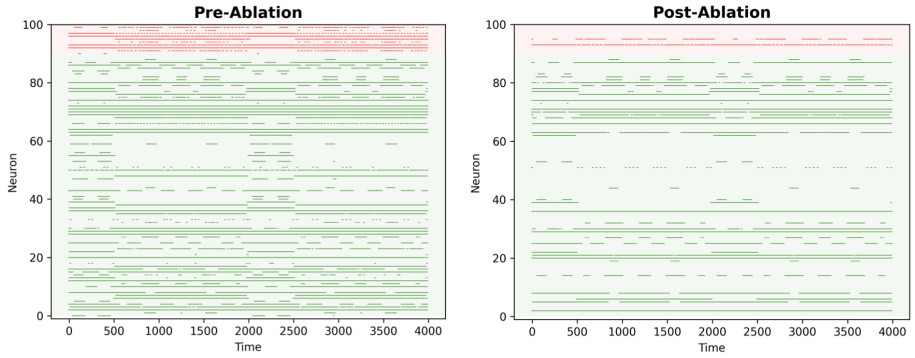

Figure 5: Underlying spiking activity while model performs the continuous tracking task seen in Figure 4. Red shading is inhibitory activity, green shading is excitatory activity. Left and right plots show spike activity before and after a *70%* ablation. It is clear that aside from the deletion of specific units, the surviving neurons modulate their post-ablation activity in entirely novel ways.

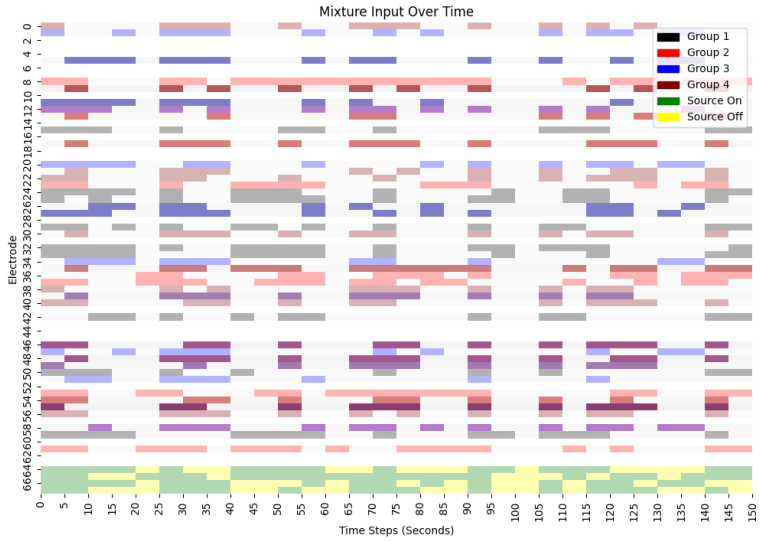

Figure 6: Task 2 – example BSS input for a n = 4, m = 64 stimulus. Each column represents a point in time. Each row represents the discrete activity of an electrode site, with the last n rows dedicated to showing source activity. Colors represent which source triggered activation of that electrode with overlapping colors signaling multiple activation events.

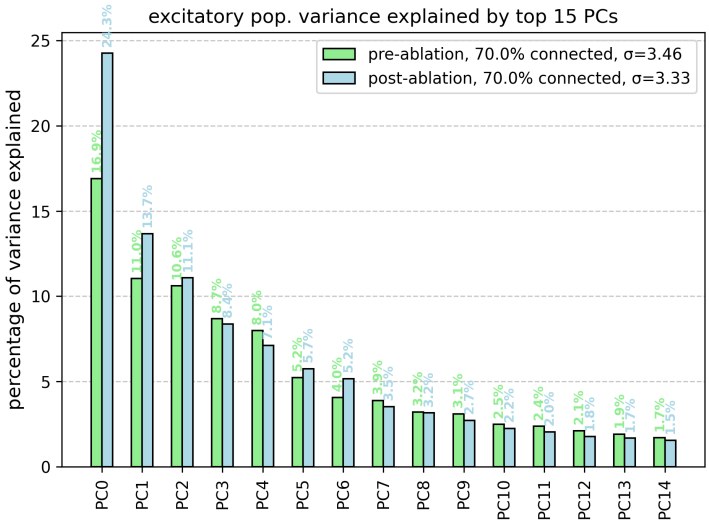

Figure 7: Eigenspectrum from a BSS with n = 4, m = 64 stimulus, as shown in Figure 6. We performed PCA on 100ms bins of network spiking activity. This shows the precentage of variance explained by the top 15 principal components before and after *60%* of the population was ablated. We see that each PC contributes substantially to model variance, even as the number of PCs increases.

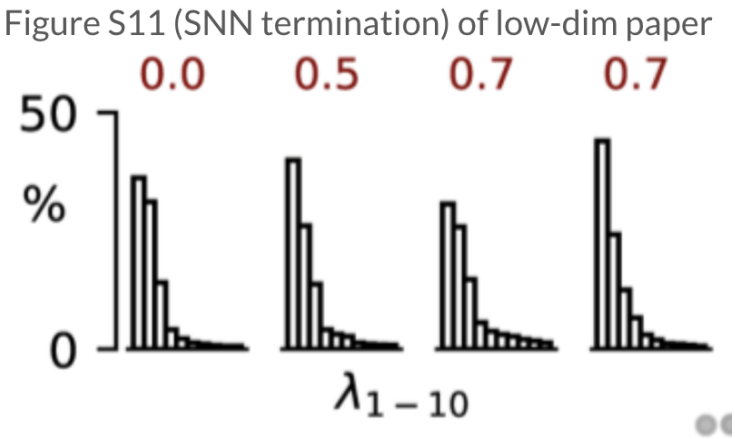

Figure 8: Similar plot as Figure 7, this time with PCA performed on in-vitro binned spiking activity from Rhesus Macaques performing a reaching task. We see that both the Figure 7 and Figure 8 distributions are heavily right-skewed with long tails. Credit: Stephen E. Clarke.

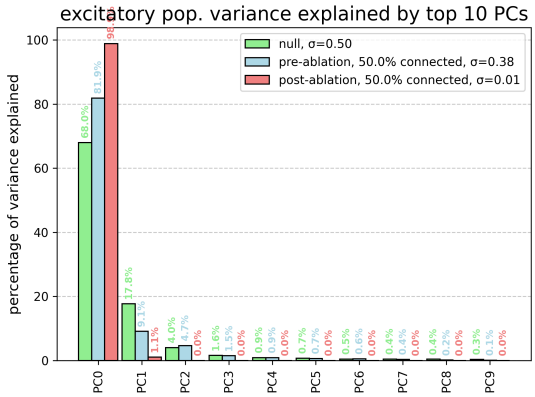

Figure 9: Similar plot as Figure 10, with input changed to the continuous tracking task and *90%* ablation. Green bars represent the eigenspectra of a a completely unconnected control population. We see that, compared to Figure 7, the spectrum is substantially more impoverished for both the pre and post ablation networks, with more than 80% of the total variance captured in just one or two PCs.

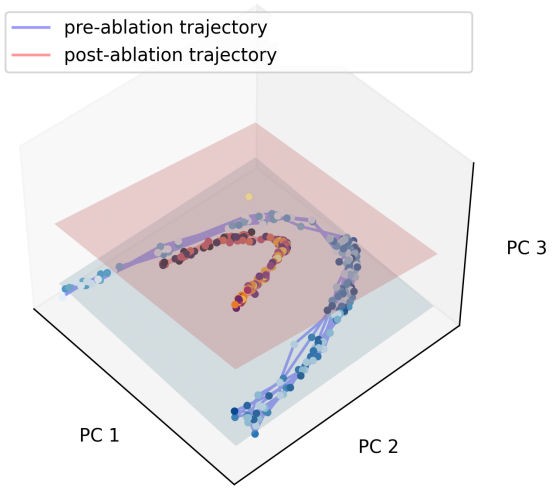

Figure 10: Continuous task network evolution in state space. Blue trajectory corresponds to pre ablation performance, red trajectory corresponds to post ablation performance. Darker hues are farther in time than lighter hues. We see that the overall structure of the trajectory is conserved following ablation, with some shift and squashing function applied.

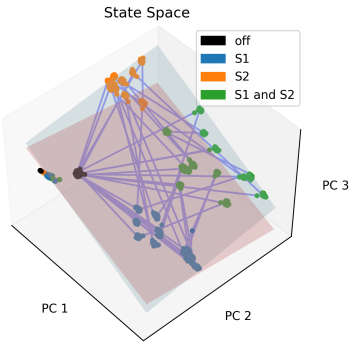

Figure 11: State space trajectory from a BSS with n = 2, m = 64 stimulus (n = 2 case easier to visualize). Black dots represent network activity with neither of the two input sources was active, colored dots represent activity with either source active individually or with both sources active together. Pre-ablation dynamics are easily visible with light blue interconnects, post-ablation dynamics are squashed and much smaller in magnitude than those before ablation. We see that each set of activity patterns is easily separable when examining the model's state space before ablation.

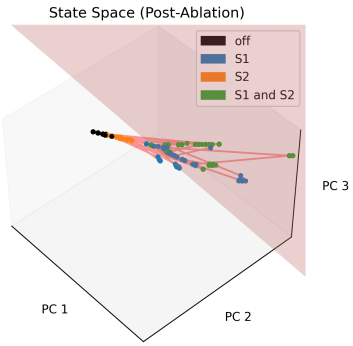

Figure 12: Zoom in onto post ablation activity present in the top left of Figure 11. Same coloring scheme. We see that ablation functionally squashes network states more tightly together, making them much less seperable than they were before ablation.

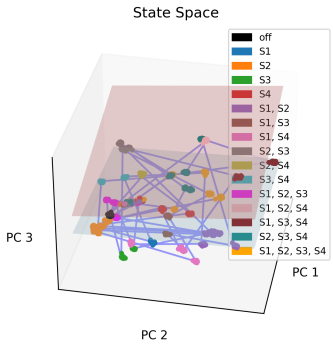

Figure 13: Same as Figure 11 except from a BSS with n = 4, m = 64. We again see that each set of activity patterns is easily separable when examining the model's state space before ablation.

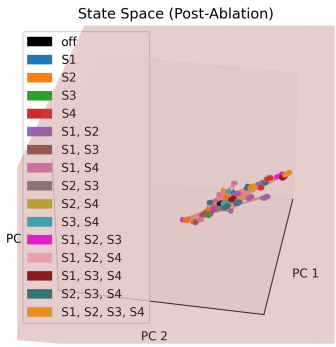

Figure 14: Zoom in onto post ablation activity present in the middle left of Figure 13. We again see that ablation squashes network states more tightly together, making them much less seperable than they were before ablation (effect pronounced for a higher-source-count input).

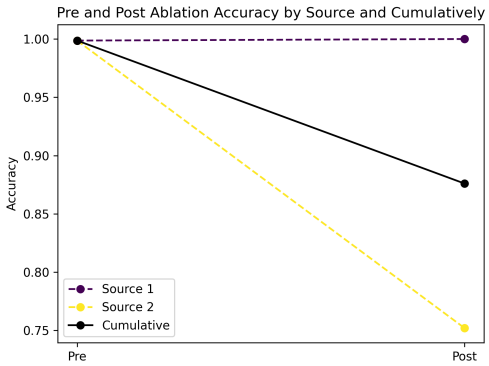

Figure 15: Classification accuracy for model corresponding to Figures 11 and 12 before and after ablation. We see that source 2 performance decrease accompanies affine transformation in state space which causes contamination between the "off" state and the "Source 2 active" state.

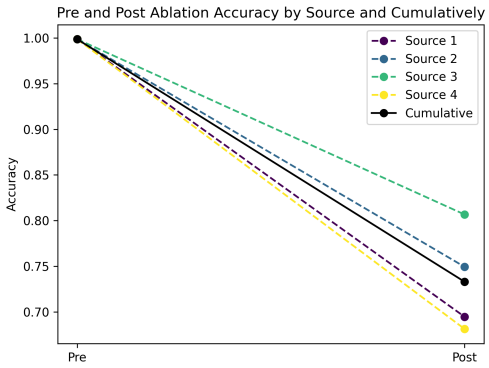

Figure 16: Classification accuracy for model corresponding to Figures 13 and 14 before and after ablation. We see that greater contamination in state space yields greatly reduces performance across sources.

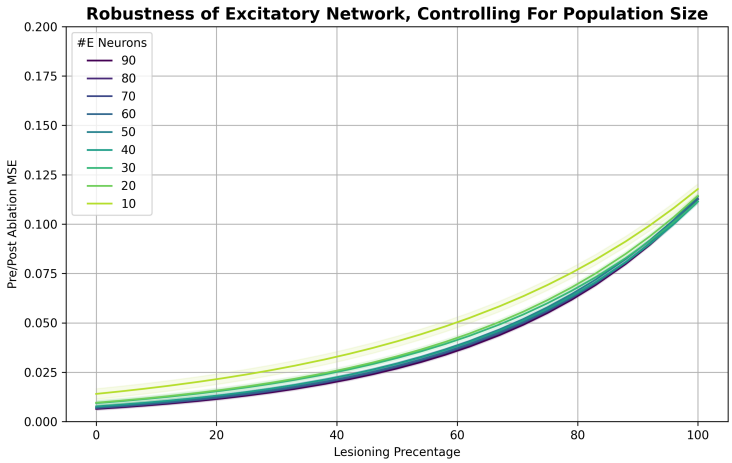

Figure 17: Solely excitatory leaky integrate and fire (LIF) SNN (as seen in Figure 1 without the inhibitory units) performing the continuous tracking task. We evaluated the mean square error between the model's outputs before and after ablation (y-axis) as a function of the ablation precentage (x-axis). Line color represents excitatory population size, with shaded regions representing 10 trial standard deviation.

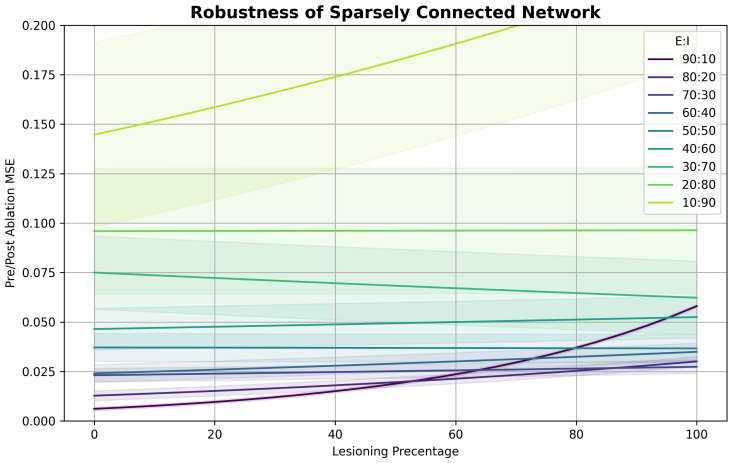

Figure 18: Same as above, except LIF SNN is coupled E/I network as shown in Figure 1. Line color represents E/I ratio for an overall population size of 100 neurons. Darker colors correspond to more excitation than inhibition. We see that the network instantiation corresponding to the dark blue line (with an E/I of 70/30) has a slope closest to zero, meaning that its output performance is the most resilient to varying levels of ablation.

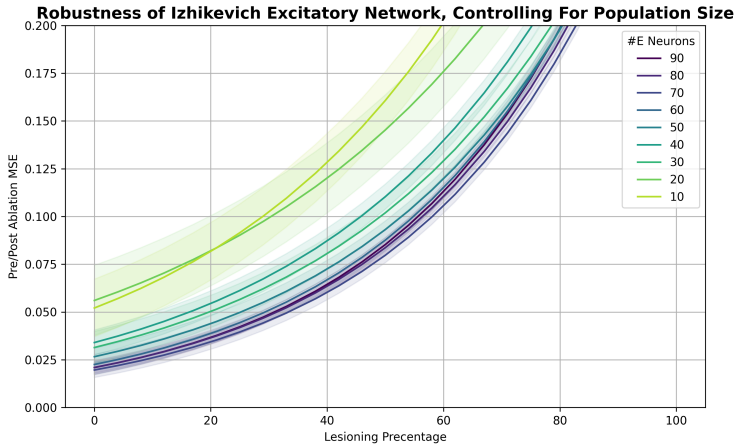

Figure 19: Same control as Figure 17 above, except Izhikevich spike-rate adaptive neurons were used instead of LIF neurons.

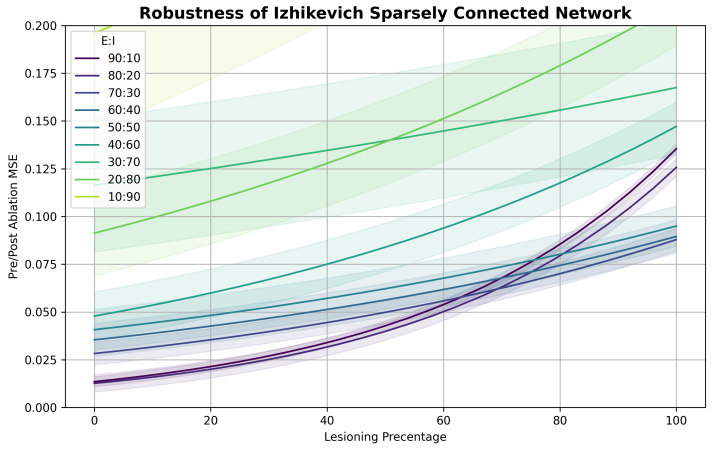

Figure 20: Same as Figure 18 above, except Izhikevich spike-rate adaptive neurons were used instead of LIF neurons. We see that the dark blue line (with an E/I of 70/30) has a slope closest to zero – a result that's preserved from the LIF network instantiation.

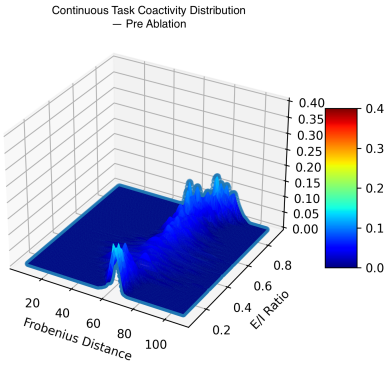

Figure 21: Frobenius distance of entire network coactivity for differing instantiations of a continuous tracking task. We performed a 81 point sweep across the E/I ratio of the network and plotted a 210 point distribution of Frobenius distance between pairs of network solutions at each E/I. Warmer colors represent higher Frobenius distance densities.

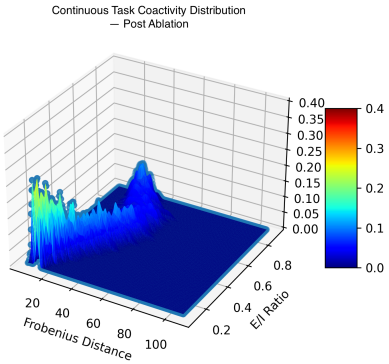

Figure 22: Frobenius distance distribution from continuous task coactivity following *90%* ablation.

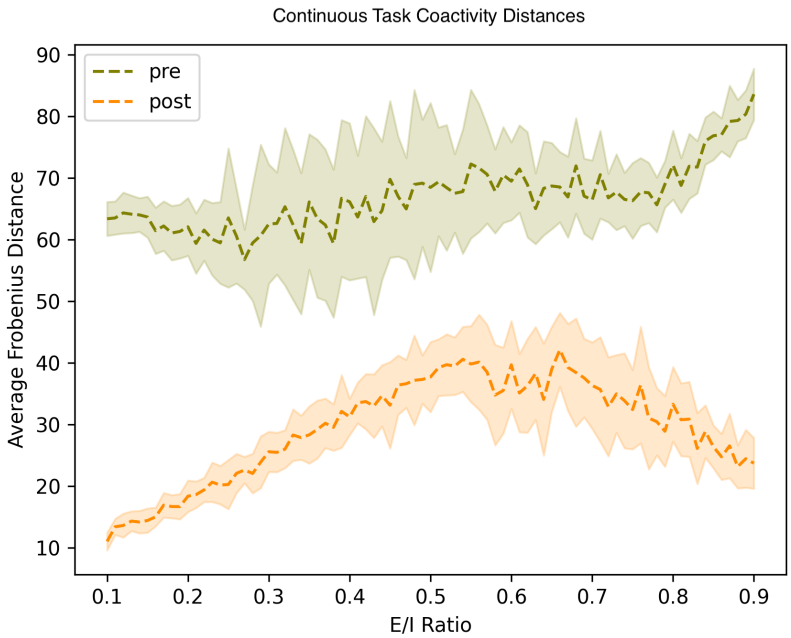

Figure 23: 2D representation of Frobenius difference trends between continuous task coactivity states for E/I network. Dotted line shows mean Frobenius distance from Figures 21 and 22, shading shows one standard deviation. Using distance as a proxy for the level of robustness of the system, we see that a 70% excitatory population maximizes the system's spread of solutions, confirming results from Figures 17, 18, 19, and 20.

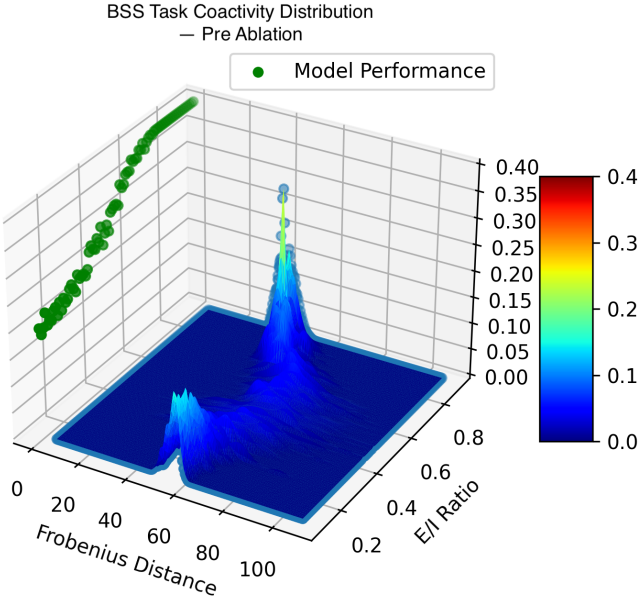

Figure 24: Frobenius distance of entire network coactivity for differing instantiations of a BSS with n = 4, m = 64. Warmer colors represent higher Frobenius distance densities. Green line emphasizes model performance across each E/I operating point. We see that performance peaks near an E/I ratio of 70% and remains constant thereafter.

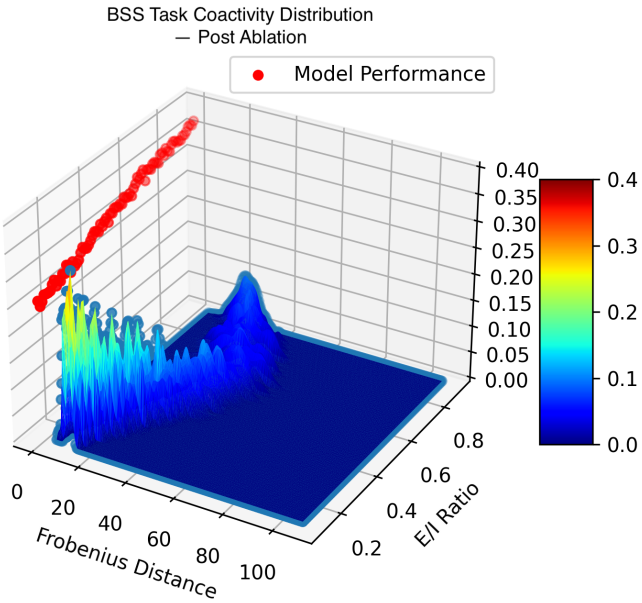

Figure 25: Frobenius distance distribution from BSS task coactivity following *90%* ablation.

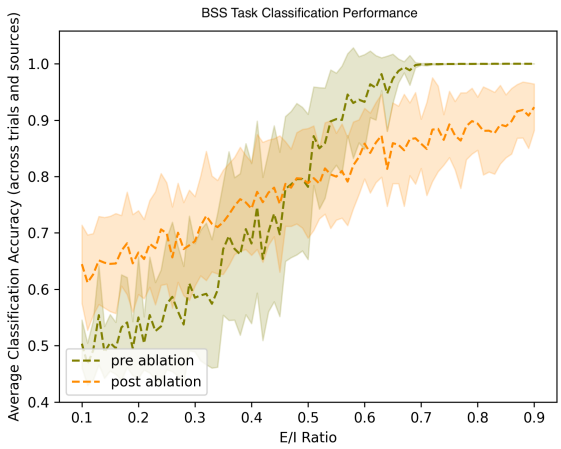

Figure 26: 2D representation of BSS task classification performance from Figures 24 and 25. Dotted green line is the same as the green performance line in Figure 23. Dotted orange line is the same as the red performance line in Figure 24. Shaded regions represent one standard deviation. We see that source classification performance hits a ceiling after a 70% E/I ratio.

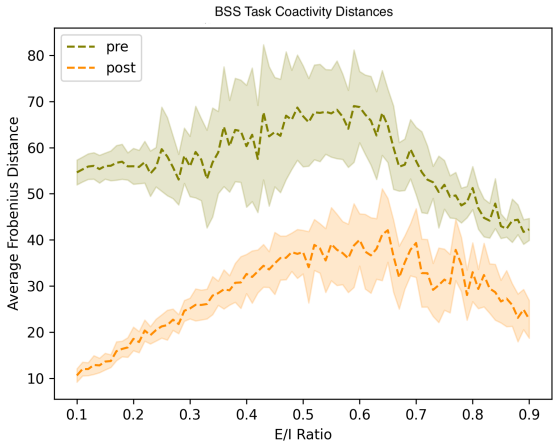

Figure 27: 2D representation of Frobenius difference trends between BSS task coactivity states for E/I network. Dotted line shows mean Frobenius distance from Figures 24 and 25, shading shows one standard deviation. Using distance as a proxy for the level of robustness of the system, we see that a 60% excitatory population maximizes the system's spread of solutions, showing that the results from the continuous task also map to the more complex BSS.