Explaining Neural Spike Activity for Simulated Bio-plausible Network through Deep Sequence Learning

Shruti R. Kulkarni, Anika Tabassum, Seung-Hwan Lim, Oak Ridge National Laboratory

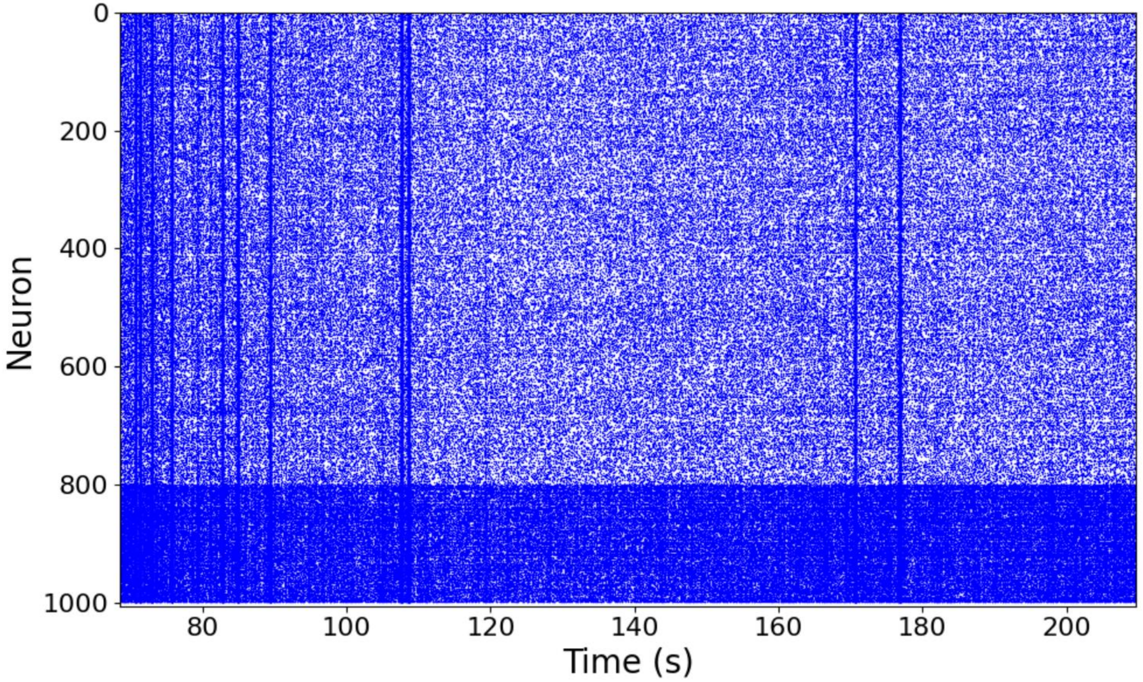
Catherine D. Schuman, University of Tennessee Knoxville, TN, USA

Bradley H. Theilman, Fred Rothganger, Felix Wang, James B. Aimone, Sandia National Laboratories, Albuquerque, NM, USA

Neuroscience Inspired Neural Network Simulation

- Modeling and Simulation integral for neuroscience research
- Simulations carried out by solving the dynamics of bio-plausible networks of neurons and synapses
- **STACS** – Simulation tool for Asynchronous Cortical Stream enables large scale HPC simulation of neuroscience model
- Analyzing the simulation data - How do we interpret the large-scale results?
  - Employ deep learning models to process temporal data
  - Scalable deep learning models along with neural simulations

- Reservoir network of excitatory and inhibitory neurons (80-20)
- Izhikevich neurons and plastic synapses
- Simulation carried for 180 seconds with spike timing dependent plasticity



- Felix Wang, Simulation Tool for Asynchronous Cortical Streams (STACS): Interfacing with Spiking Neural Networks, Procedia Computer Science, Volume 61, 2015, Pages 322-327.
- F. Wang, et al., "Scaling neural simulations in STACS", IOP Neuromorphic Computing and Engineering, April 2024.

Neural Data Representation

- Reservoir network of E-I spiking neurons
  - 800 excitatory neurons
  - 200 inhibitory neurons
- Simulated by applying random thalamic noise input
  - Output spikes at an average rate of 10 Hz per neuron
- Spike trains for neuron i at times t^f:

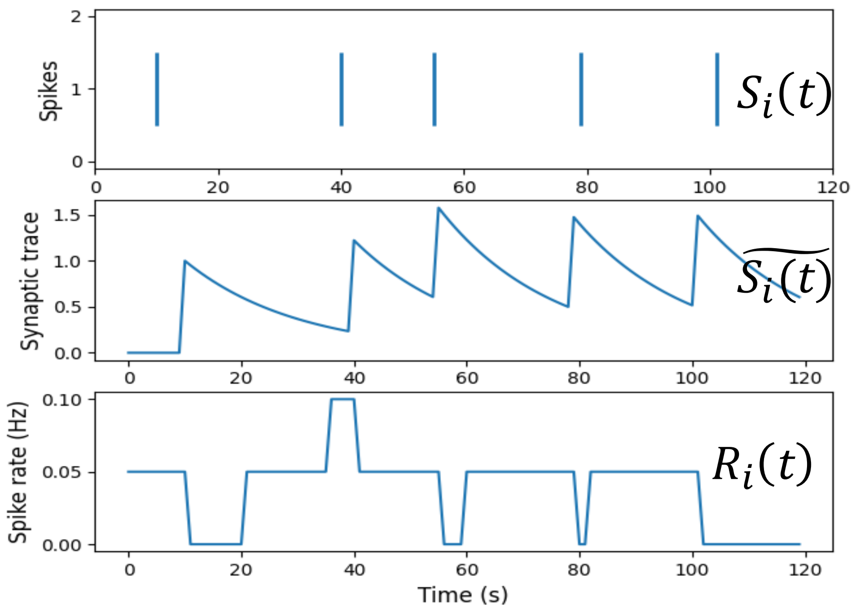
$$S_i(t) = \sum_{t^f} \delta(t - t^f)$$

- Smoothened spike signals using kernel $\alpha(t) = e^{-\frac{t}{\tau}}.\mathrm{H}(t - t^f)$:

$$\widetilde{S_i(t)} = S_i(t) * \alpha(t)$$

- Running average spike rates over T_w:

$$R_i(t) = \sum S_i(t\colon t + \mathrm{T_w})/\mathrm{T_w}$$



- Causal components:
  - Combines underlying structure and network activity
  - Adjacency matrices over a temporal window τ
  - Ω for a spike traveling from neuron i to j*:

$$\Omega_{i,j} = W_{i,j} \times e^{(t_j^f - t_i^f)/\tau} \times H(t_j^f - t_i^f - \delta_{i,j})$$

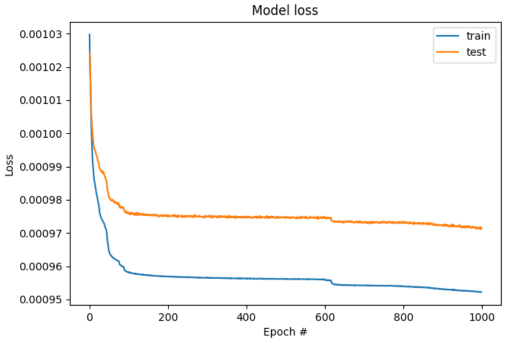
* Theilman, B. H., Wang, F., Rothganger, F., & Aimone, J. B. (2023). Decomposing spiking neural networks with Graphical Neural Activity Threads. *arXiv preprint arXiv:2306.16684*.

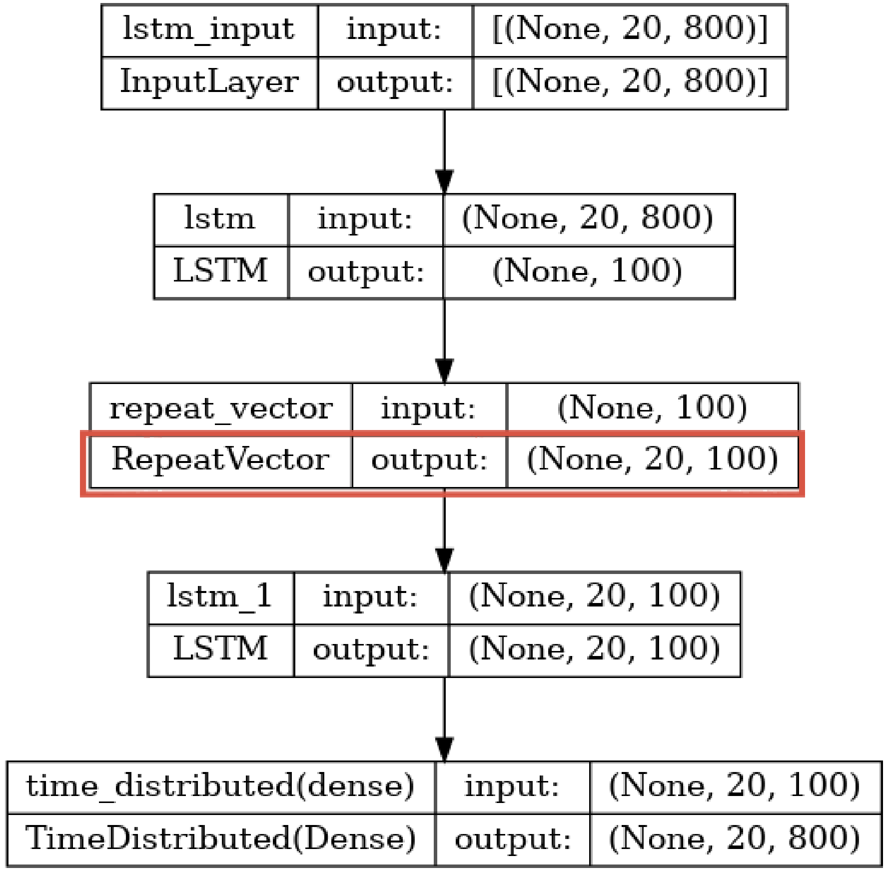
Research goals

- Can deep learning models:
  - be alternative to reduced order representation?
  - Help predict spike sequences in an SNN?
  - Explain the transition of spike activity over time?
- Our approach:
  - Capture a reduced order representation of neuron states
  - Capture attention of each neuron for the SNN spike prediction activity

Long Short Term Memory (LSTM) Autoencoder

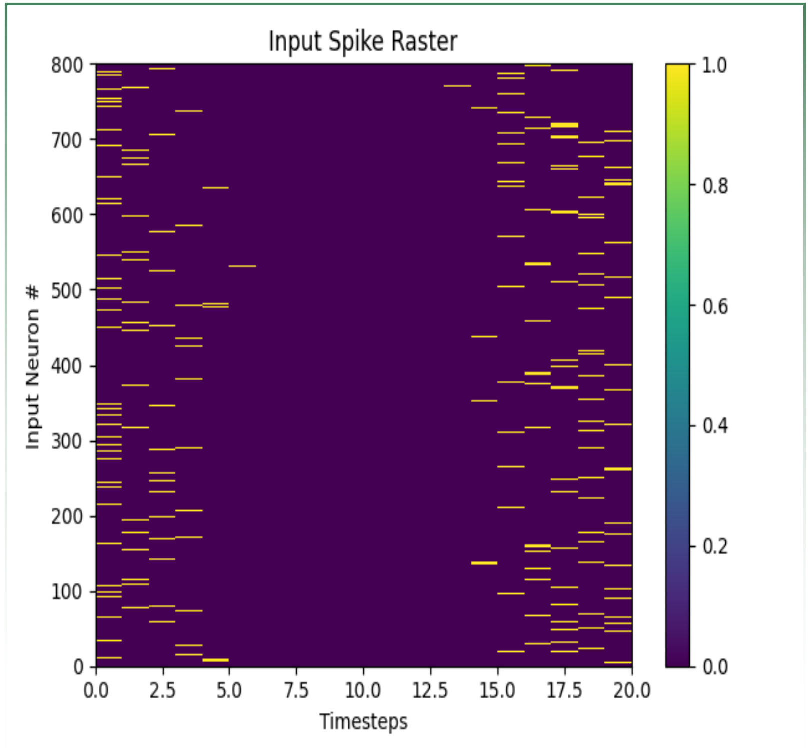
- Use LSTM Autoencoder model for sequence reconstruction
- SNN simulation spikes divided into sequences of 20 timesteps
- Use Spike traces ($\widetilde{S(t)}$) as inputs and targets
- Encoding layer captures a low order encoding with 100 units
- Model presented trained with 6K sequences for 1000 epochs

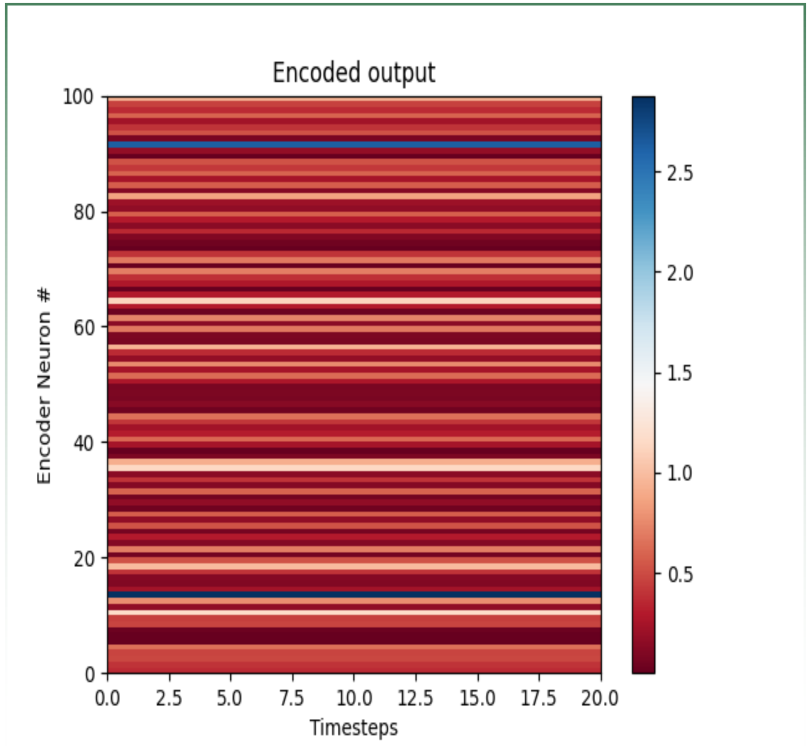


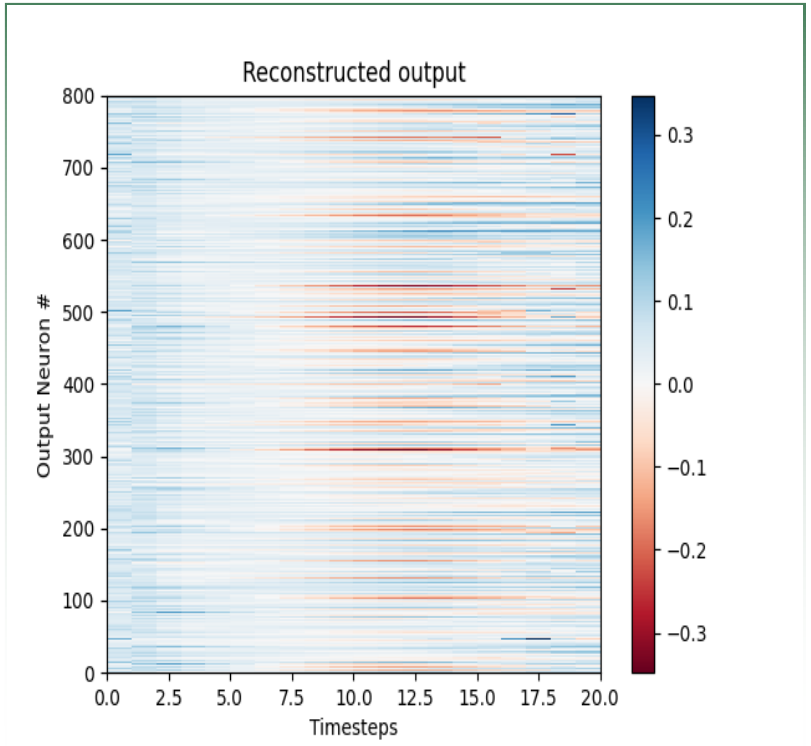


Oak Ridge National Laboratory

Sample outputs from trained LSTM – AE model





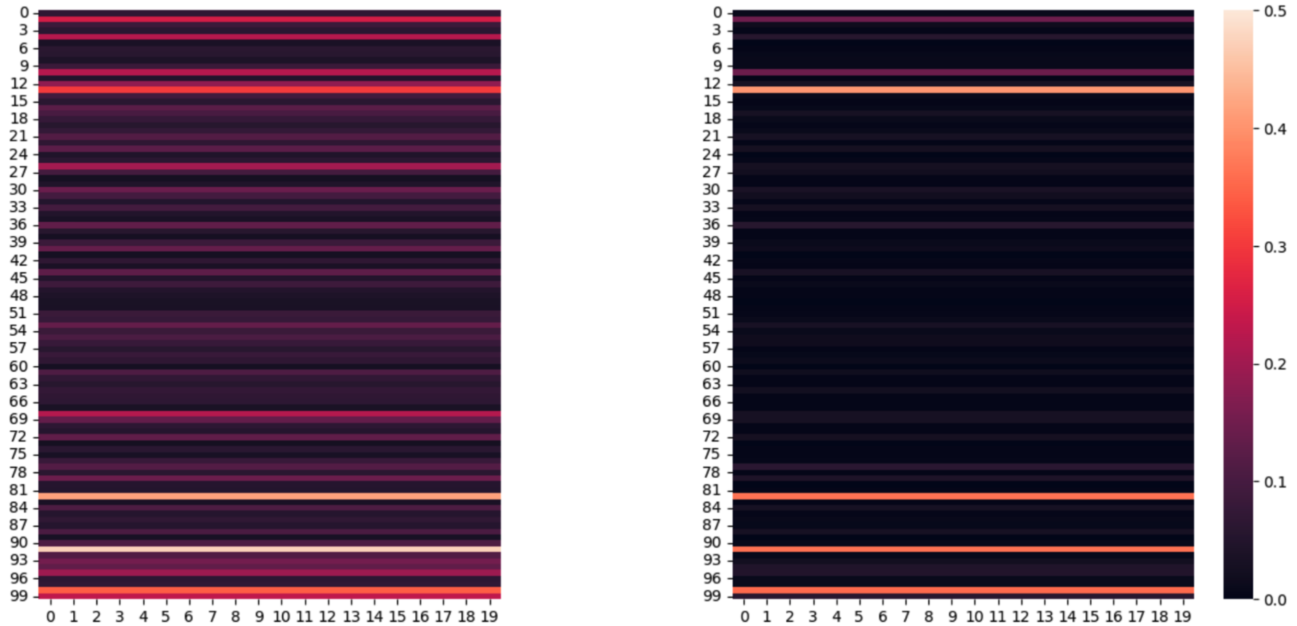


Reconstructed output tracks the temporal regions of high spikes as seen in the input sample.

Embedded output static over the 20 timestep interval

Encoder outputs from LSTM-AE

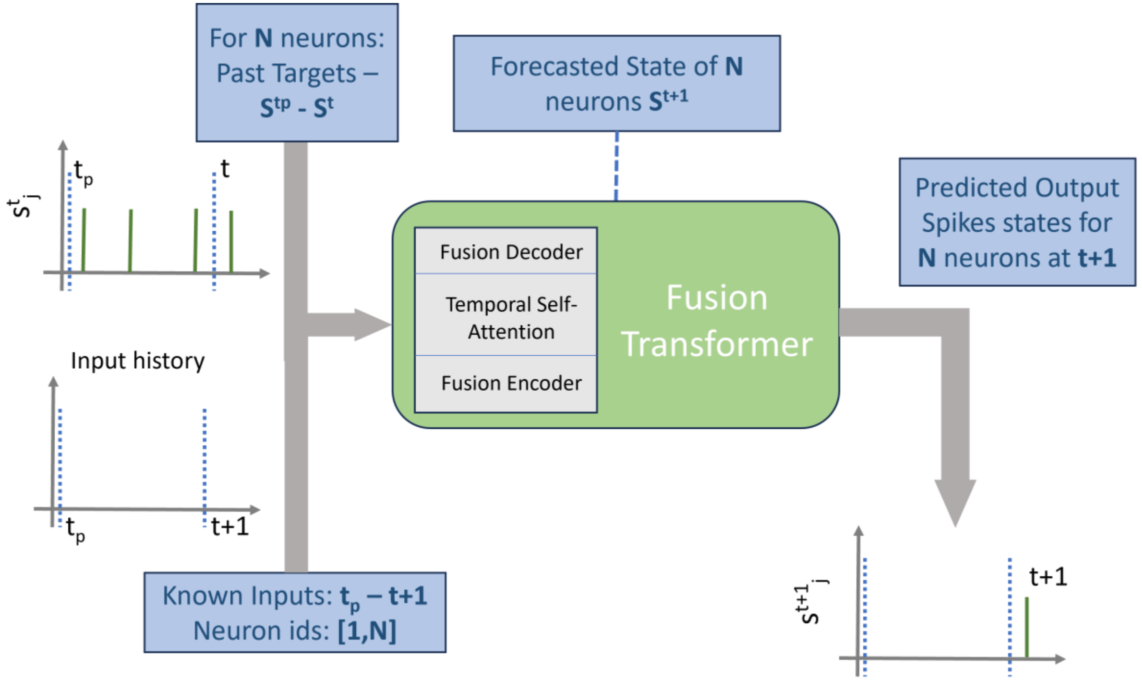
- Mean and variance of the encoded outputs across all test samples
- Overall low variation for a 20 timestep sequence
- LSTM captures the compressed representation of the sparse SNN activity, but not the temporal variations



(a) Mean (b) Variance

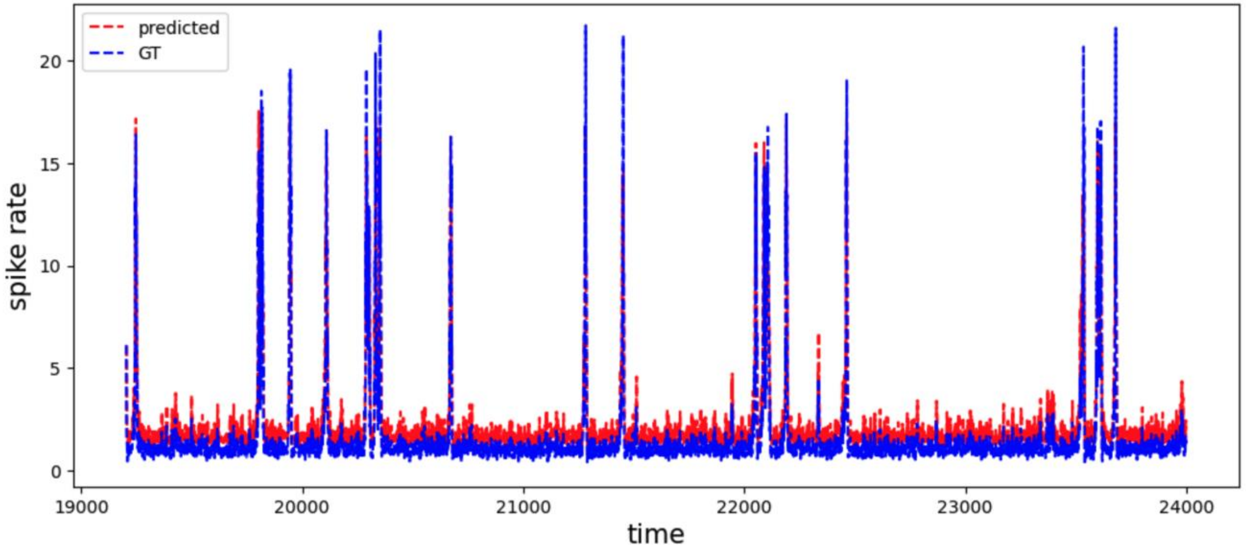
Temporal Fusion Transformer (TFT) Model

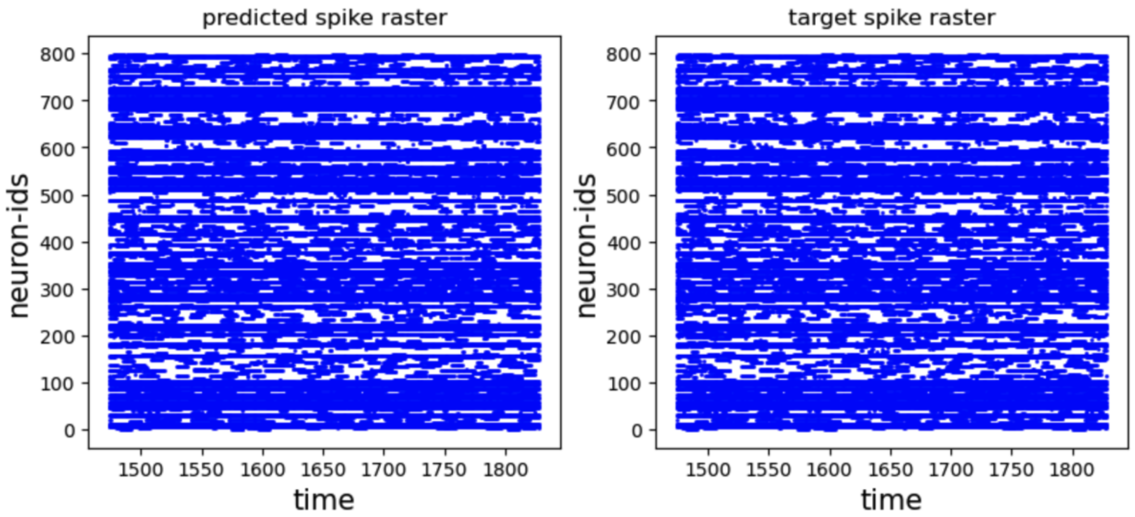
- TFT - Aggregation of LSTM embedding plus attention mechanism
- Problem to be solved – multi-variate (neuronal spike rates) time series forecasting
- **Input history sequence:**
  - Multi-variate spike rate sequences $R^{t:t+t_m}$
- **Known Inputs:**
  - Time step values t
  - Neuron ids i
- **Target:** Predict spike rates for t_h timesteps in the future
- Analyze attention weights across different neurons



OAK RIDGE National Laboratory

TFT prediction results



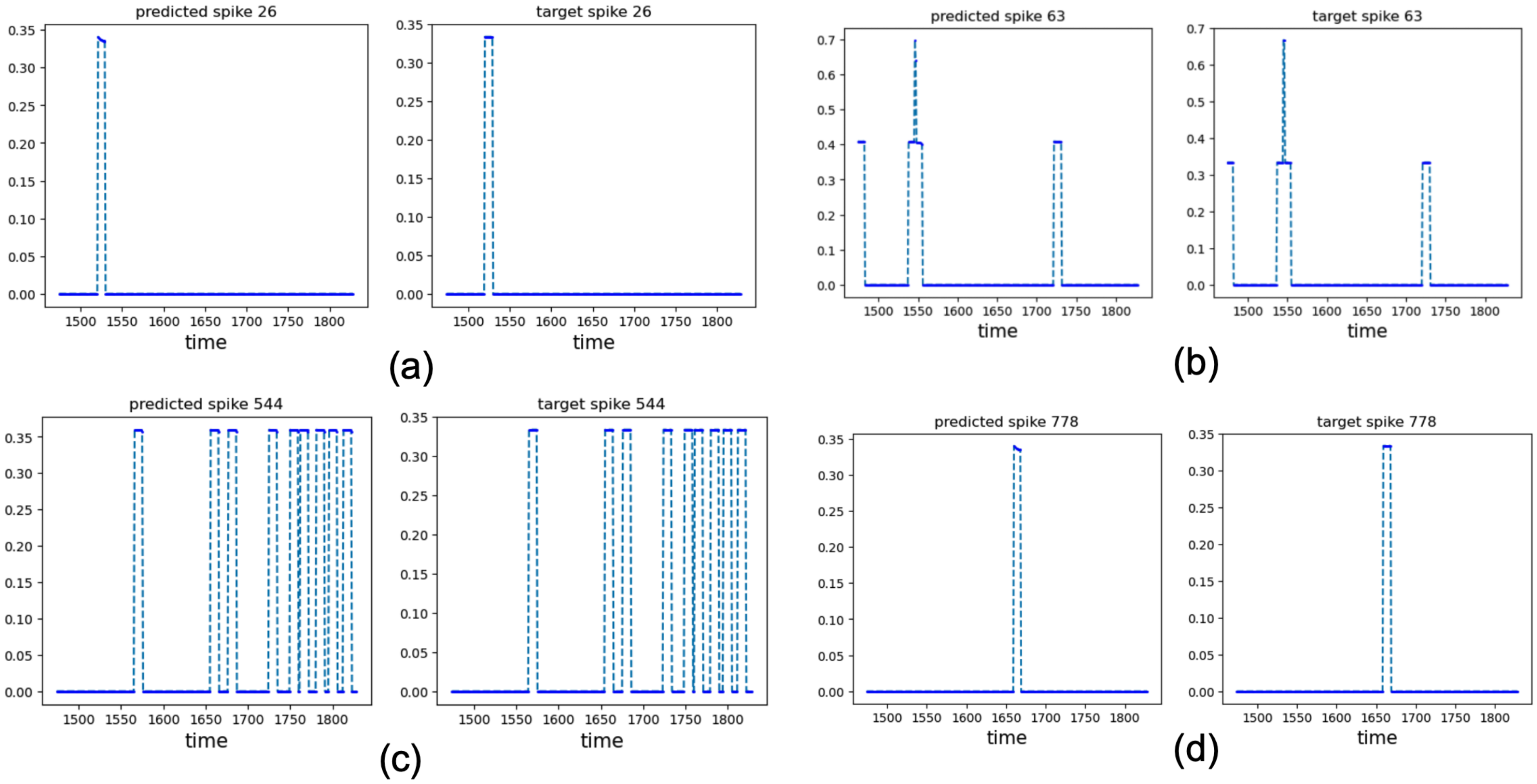


- **Task: Prediction over the average network spike rate sequence**
- TFT successfully predicts peak spike rates for $(t+1)^{th}$ timestep

- **Task: Predicting individual neuronal spike rates**
- TFT successfully captures the binarized spike rates for each neuron

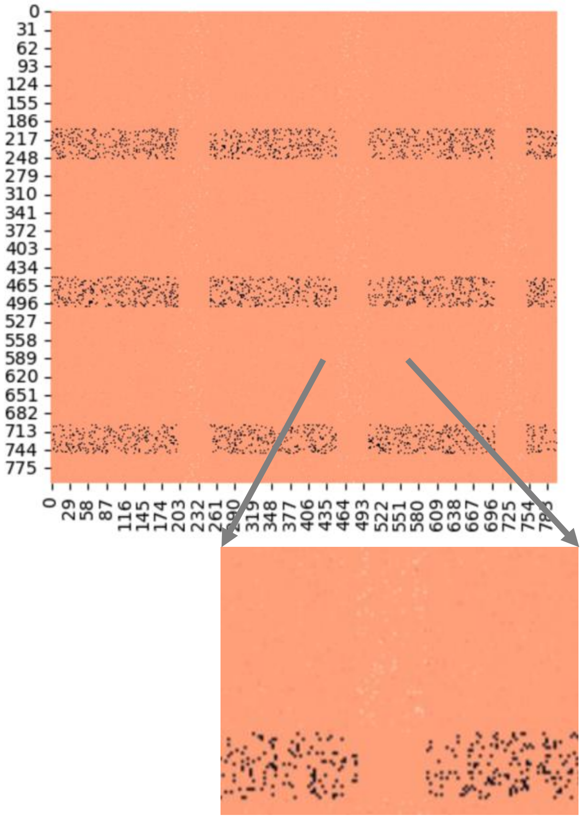
OAK RIDGE National Laboratory

Spike Sequence Prediction with TFT

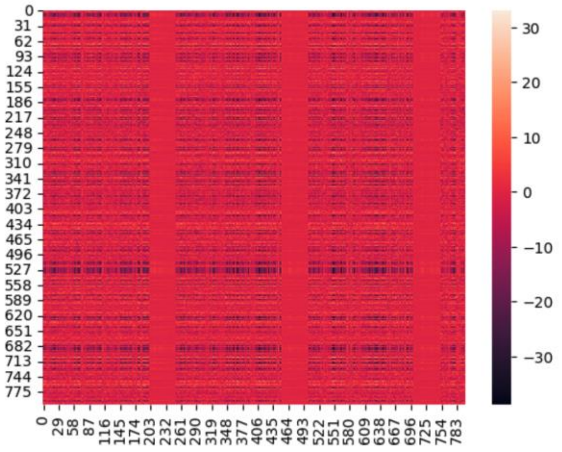


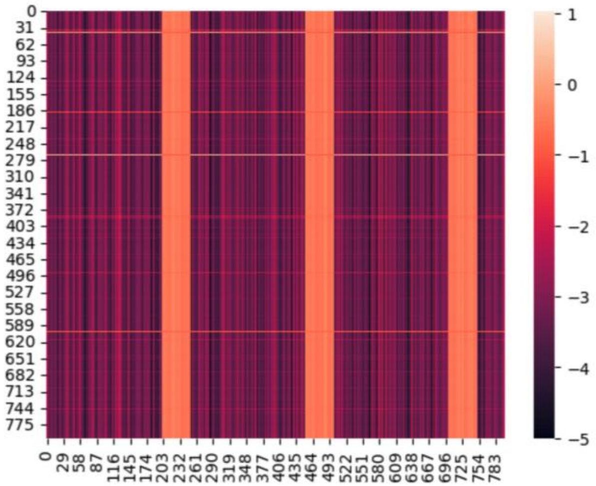
TFT successfully predicts the spike rates for successive timesteps for each neuron

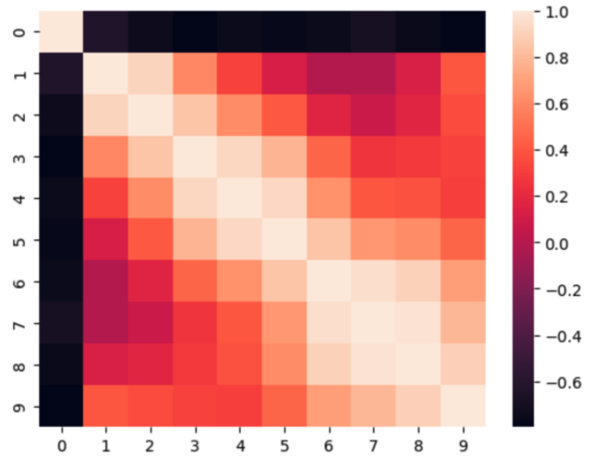
Analyzing attention weights



Ground Truth:
Network adjacency heat-map

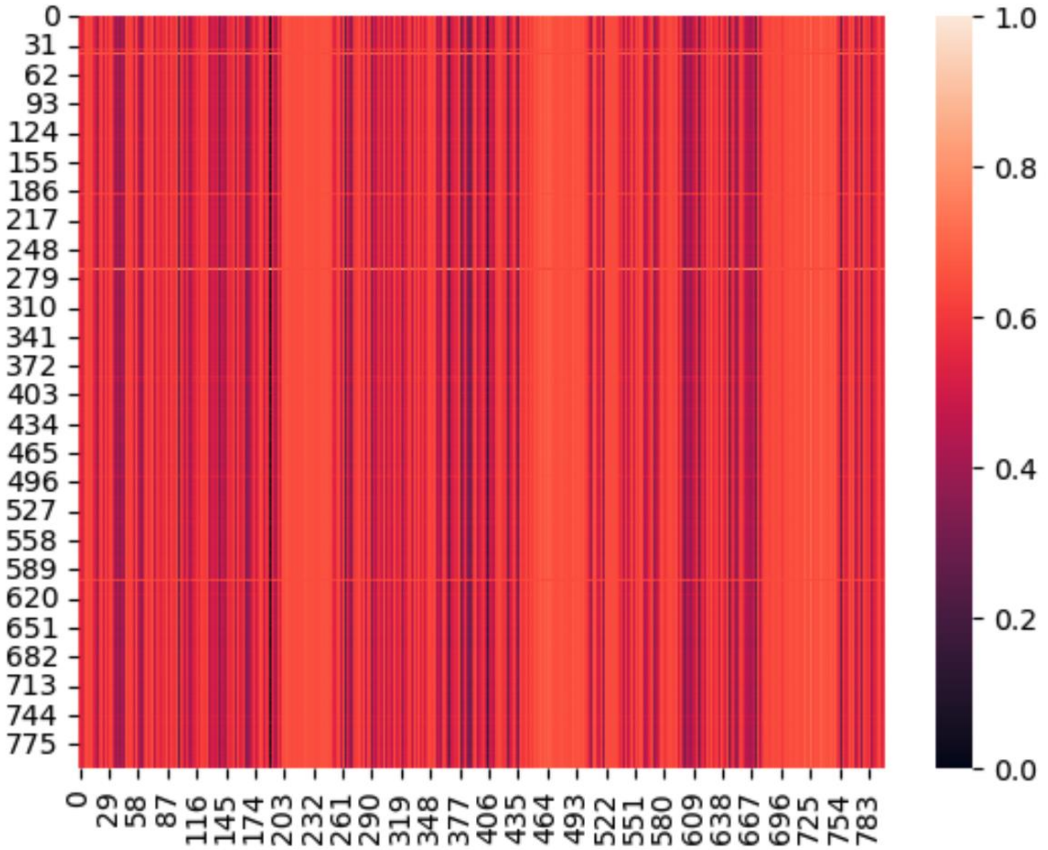






- **Self-attention** weights of TFT of size $N \times k \times m$
- **Correlation** among the mean attention weights for all test samples high for connected neurons in the network
- **TFT Attention Similarity:** High similarity among average connectivity weight and attention weights of all N neurons
- Spike activities among neurons have short-range temporal dependencies.

TFT Attention weights and Causality



- Computed causality adjacency matrices over entire simulation T
  - $N \times N$ sized matrix
- Dot product similarity of mean attention weights (for all test samples) with network's causality components ($\Omega_{i,j}$)
- High similarity for connected neurons

Summary and Conclusion

- Demonstrated two Deep Learning models – LSTM and transformers to analyze sparse neural data
- Application of DL models for performing underlying tasks along with representing the underlying architecture for sparse spike signals
- Models of LSTM and TFT are scalable on HPC, hence, could be applied to very large-scale neural simulation
- Future directions –
  – Developing reduced order cognitive learning models
  – Demonstrate scalability over large-scale simulations
  – Causal representation to interpret relation between different neuronal groups
  – Causal representation for codesigning energy-efficient SNN hardware

Acknowledgements

- U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR) program.

- U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA).

Thank you!

Questions?

Contact:
tabassuma@ornl.gov
lims1@ornl.gov
kulkarnisr@ornl.gov