Sussex, Monday 3 July 2023

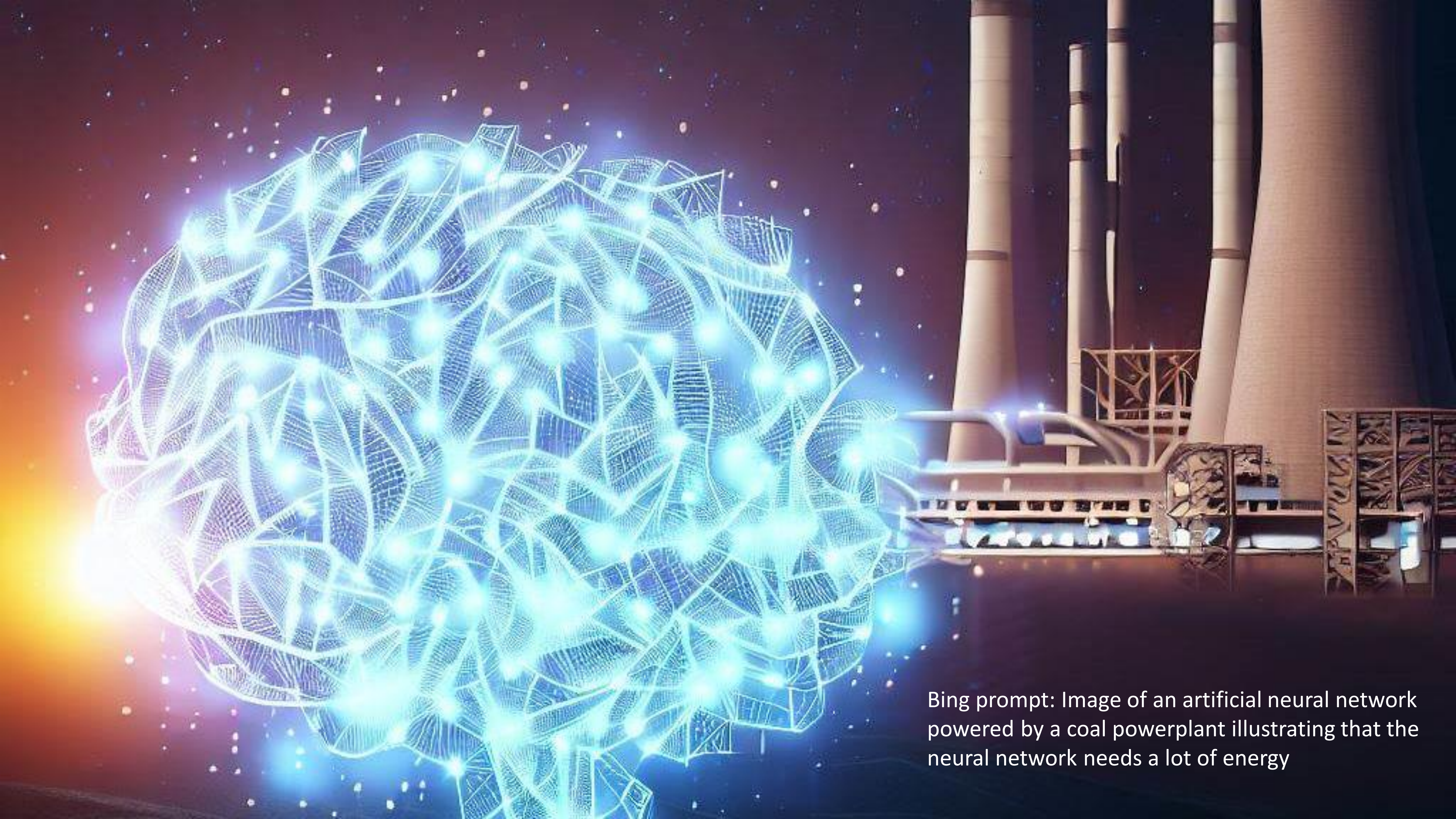
# Can we learn from the brain to make AI more energy efficient?

Prof Thomas Nowotny

School of Engineering and Informatics

University of Sussex

Bing prompt: A Spiking Neural Network

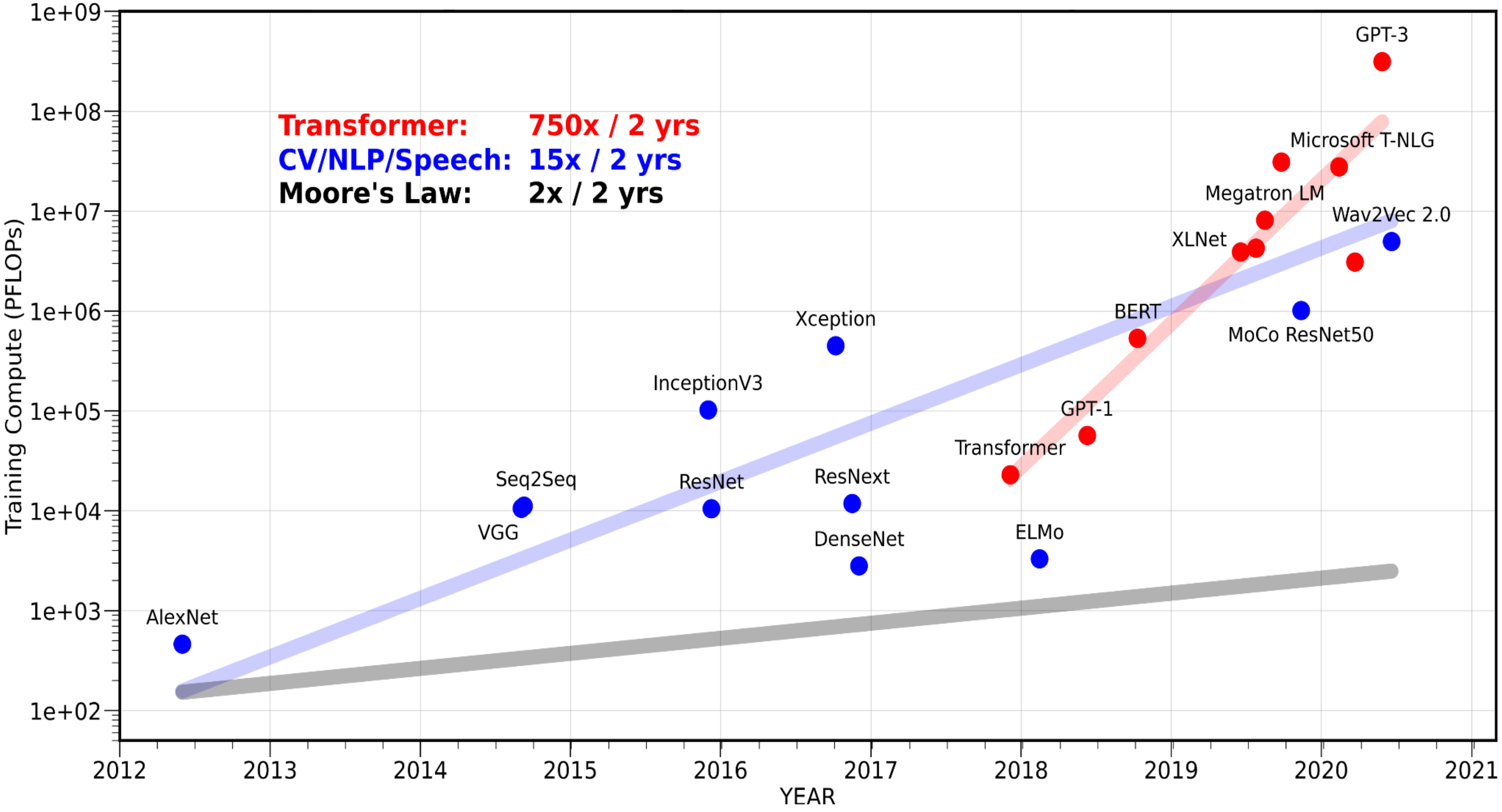


Bing prompt: Image of an artificial neural network powered by a coal powerplant illustrating that the neural network needs a lot of energy

Bing prompt: Car driving to the moon, using a lot of petrol

# Modern AI Resource Trends

Sussex AI




Gholami, A., Yao, Z., Kim, S., Mahoney, M. W., & Keutzer, K. (2021) AI and Memory Wall. *RiseLab Medium* Blog Post, University of Califonia Berkeley, 2021, March 29

# Brain Circuits vs Artificial Neural Networks

Sussex AI

- Electro-chemical “wetware”
- Evolved embodied
- Sparse communication with “spikes”

- Metal-oxide hardware
- Built for machine learning
- Dense communication of floating point numbers

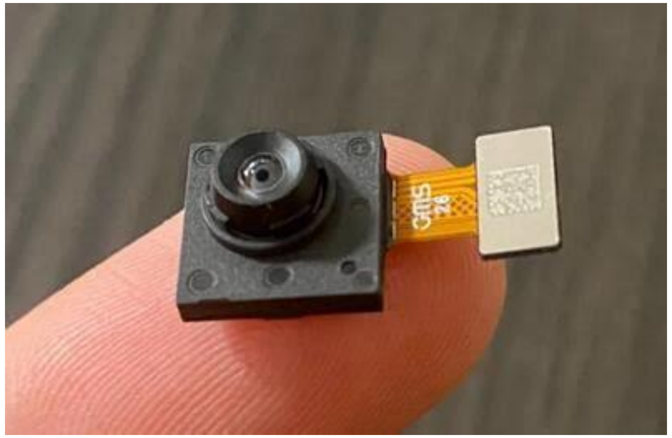
# Neuromorphic computing

SpiNNaker
(Manchester/
Dresden)

BrainScaleS
(Heidelberg)


Intel Loihi
(Intel Corporation)

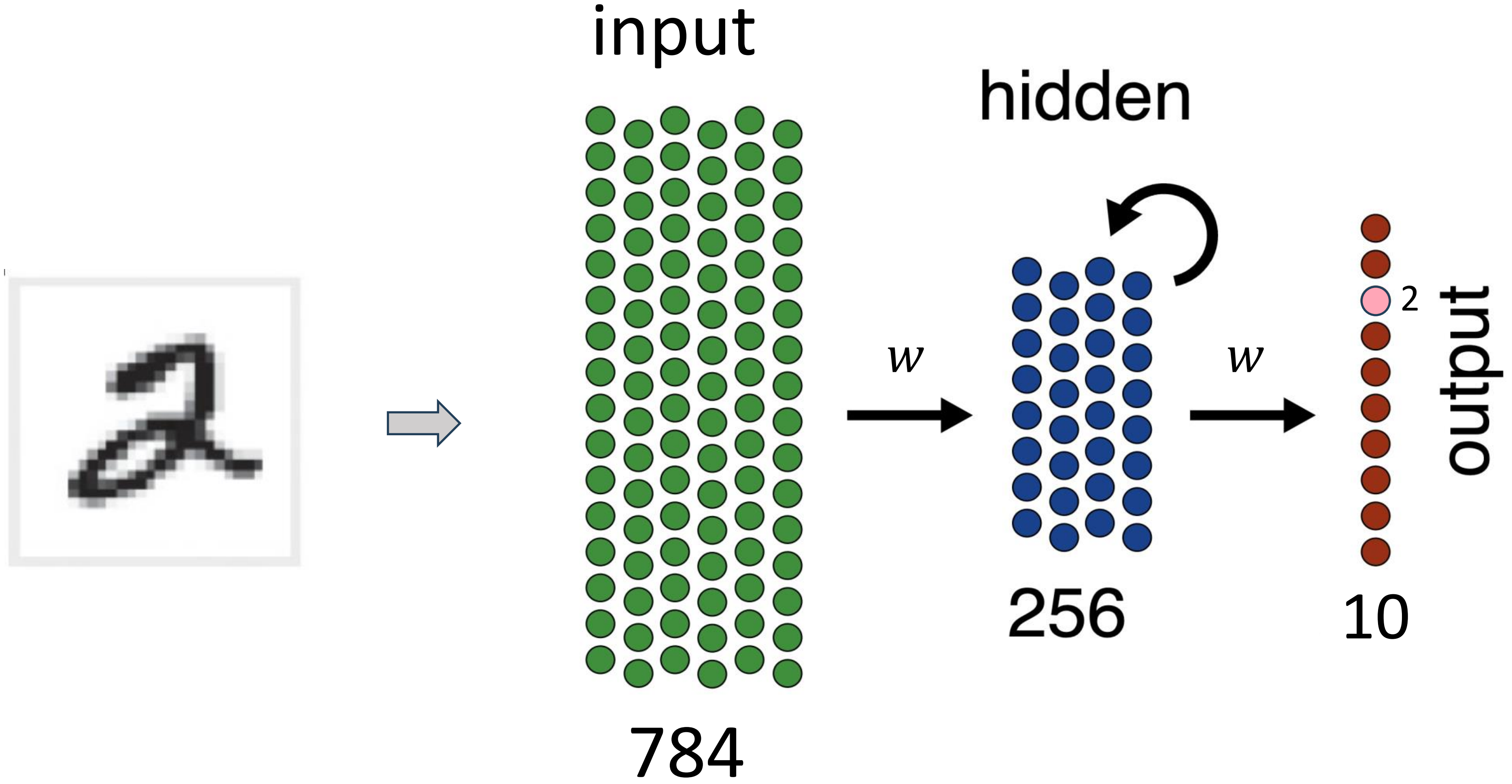
Synsense
Speck



IBM synapse
(IBM)

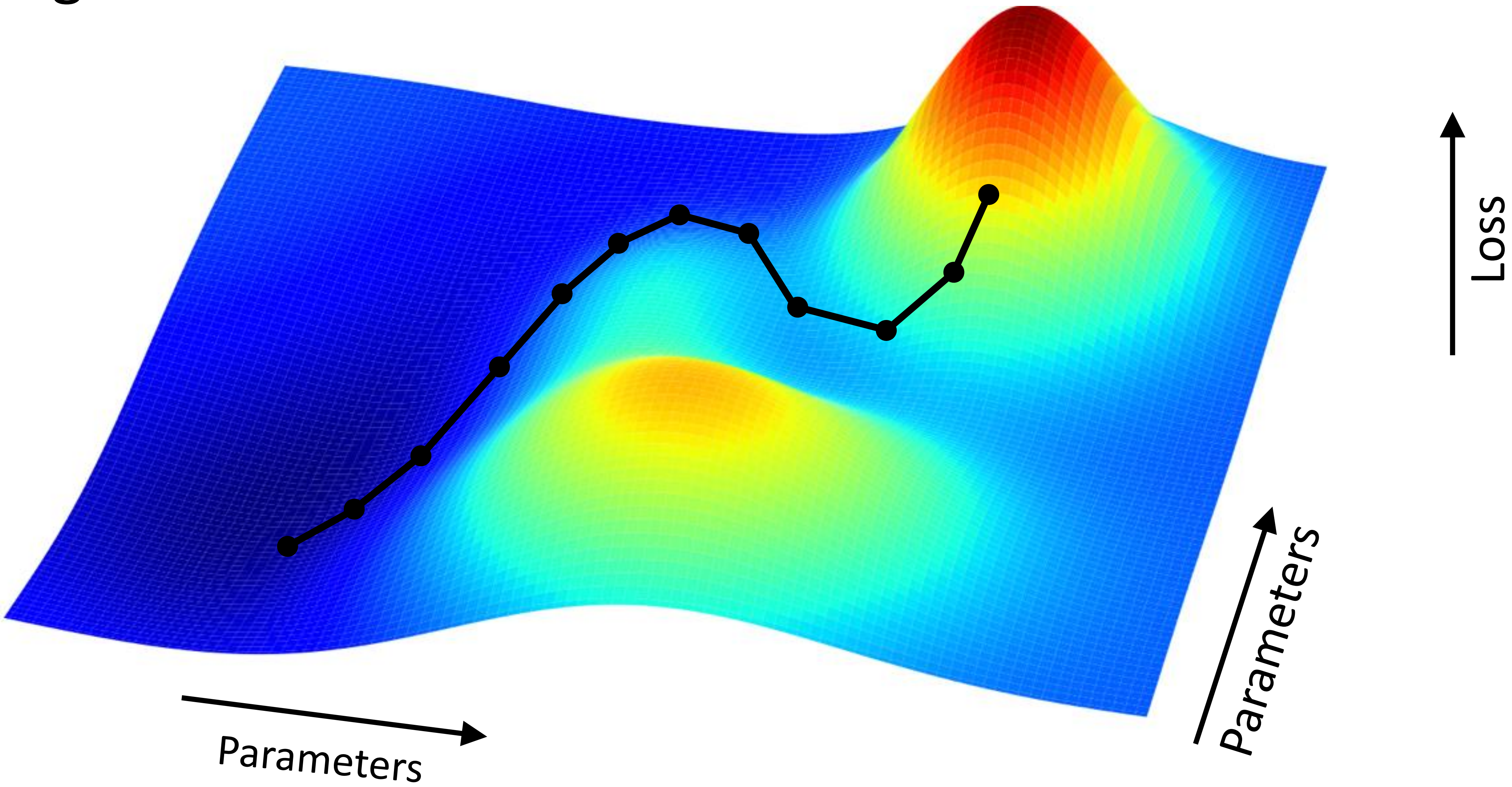
# General classification problem

Sussex AI



# Training: Gradient descent

Sussex AI



Sussex AI

www.nature.com/scientificreports

frontiers

ORIGINAL RESEARCH

IOP Publishing Neuromorph. Comput. Eng. 2 (2022) 024002 https://doi.org/10.1088/2634-4386/ac5ac5

NEUROMORPHIC
Computing and Engineering

CrossMark

OPEN ACCESS

RECEIVED
22 December 2021

REVISED
9 February 2022

ACCEPTED FOR PUBLICATION
4 March 2022

PUBLISHED
25 March 2022



PAPER

# mlGeNN: accelerating SNN inference using GPU-enabled neural networks

James Paul Turner[1,*], James C Knight[1], Ajay Subramanian[2] and Thomas Nowotny[1]

1 Centre for Computational Neuroscience and Robotics, School of Engineering and Informatics, University of Sussex, Brighton, United Kingdom
2 Department of Psychology, New York University, New York, NY 10003, United States of America
* Author to whom any correspondence should be addressed.

E-mail: J.P.Turner@sussex.ac.uk

Keywords: machine learning, spiking neural networks, GPU, ANN to SNN conversion, convolutional neural networks, GeNN, Re

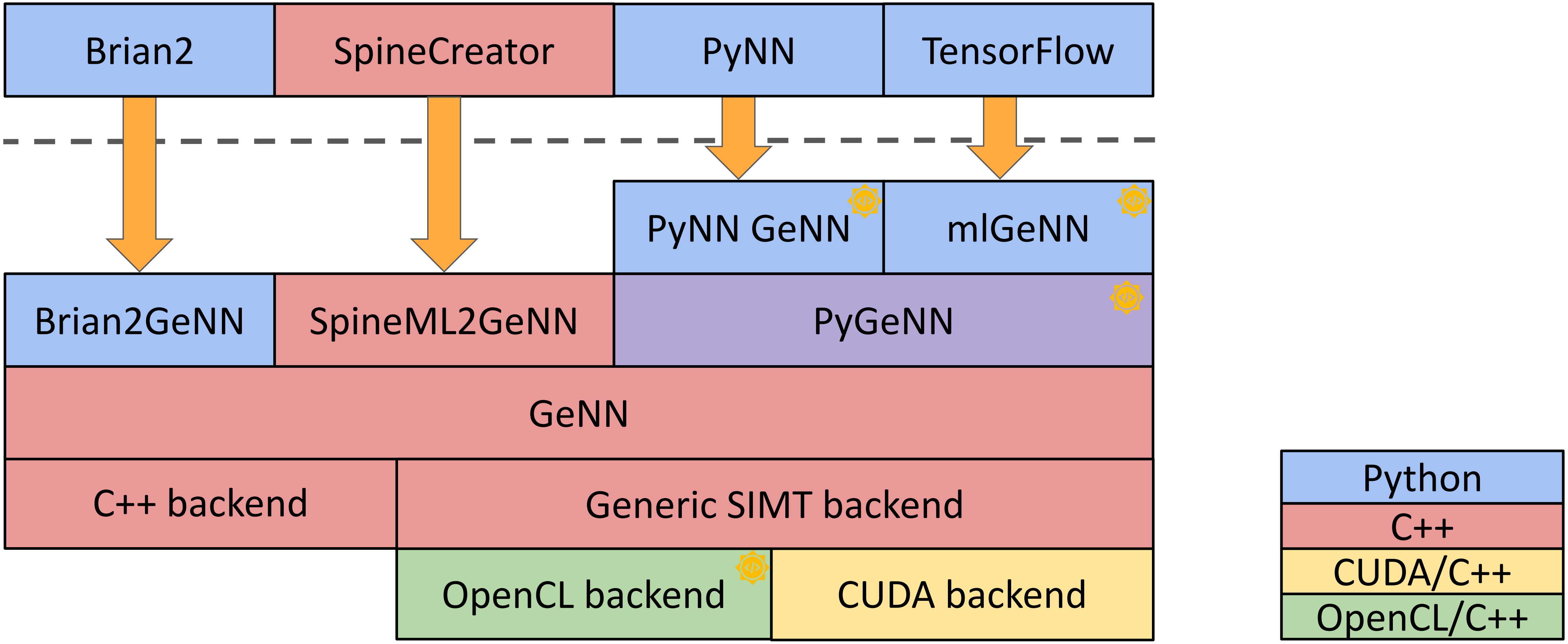
# GeNN (**G**PU **e**nhanced **N**eural **N**etworks)

Sussex AI

GeNN generates CUDA kernels and data transfer “convenience functions” based on a model definition provided by a user.

(Meta Compiler)

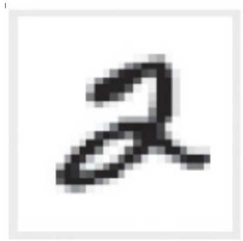
# GeNN ecosystem

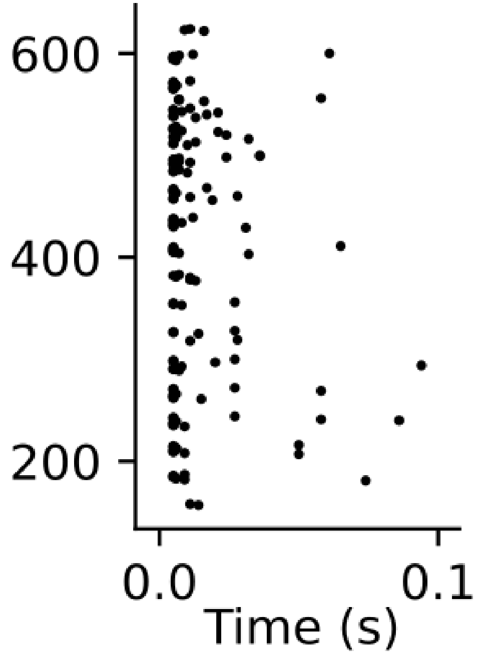


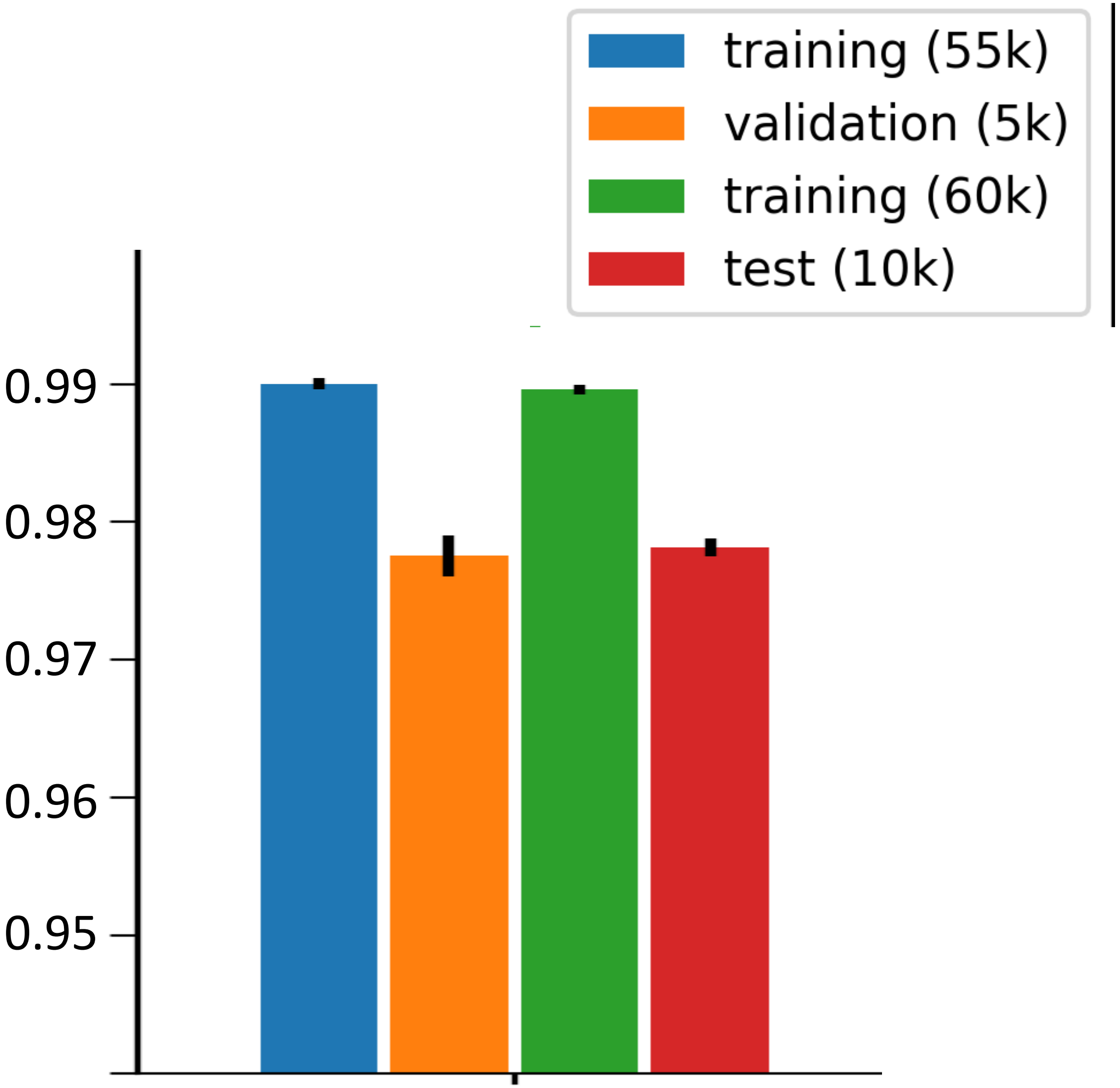

https://github.com/genn-team/genn/

# Handwritten digits with EventProp

Sussex AI



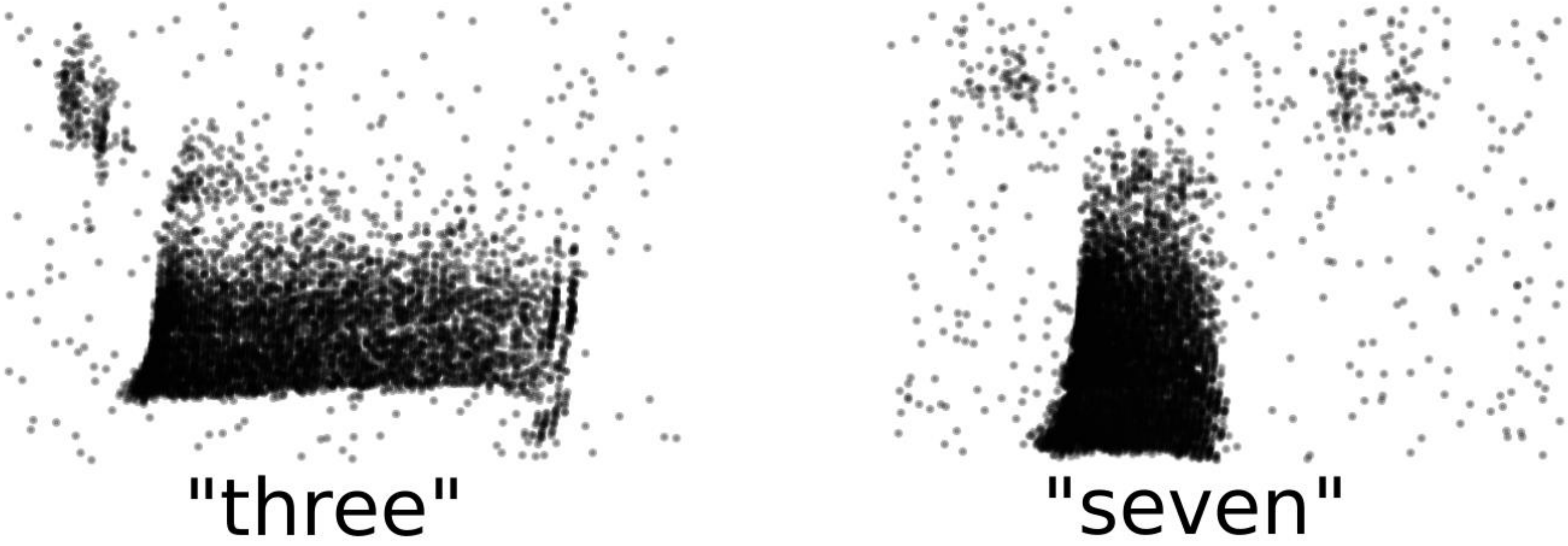



Trains to >97% in 27s on a PC with RTX 3090

Sussex AI

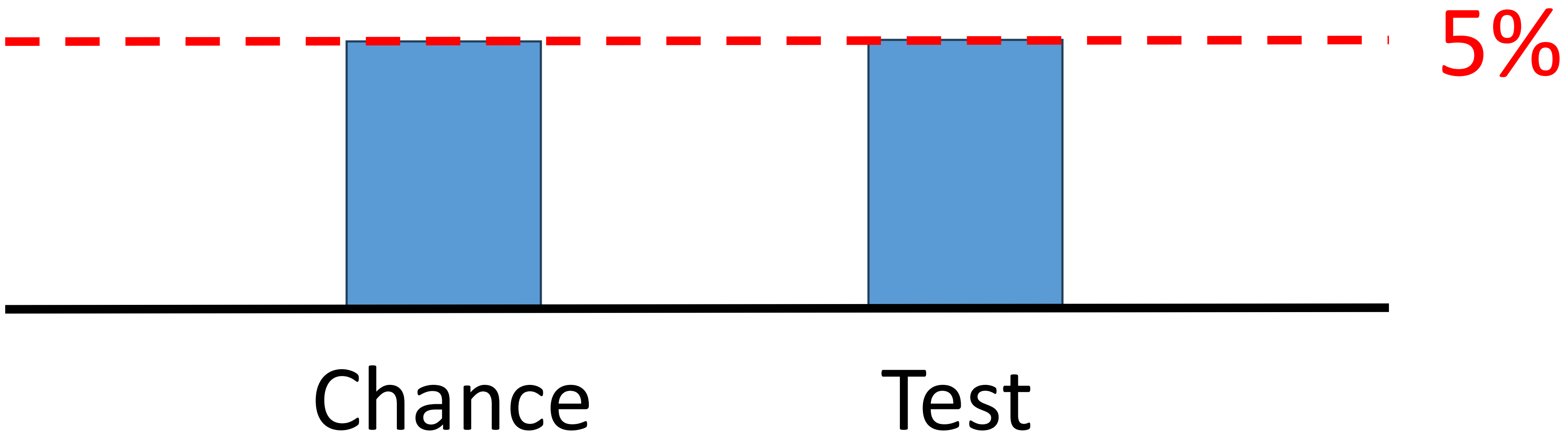
# Spiking Heidelberg Digits

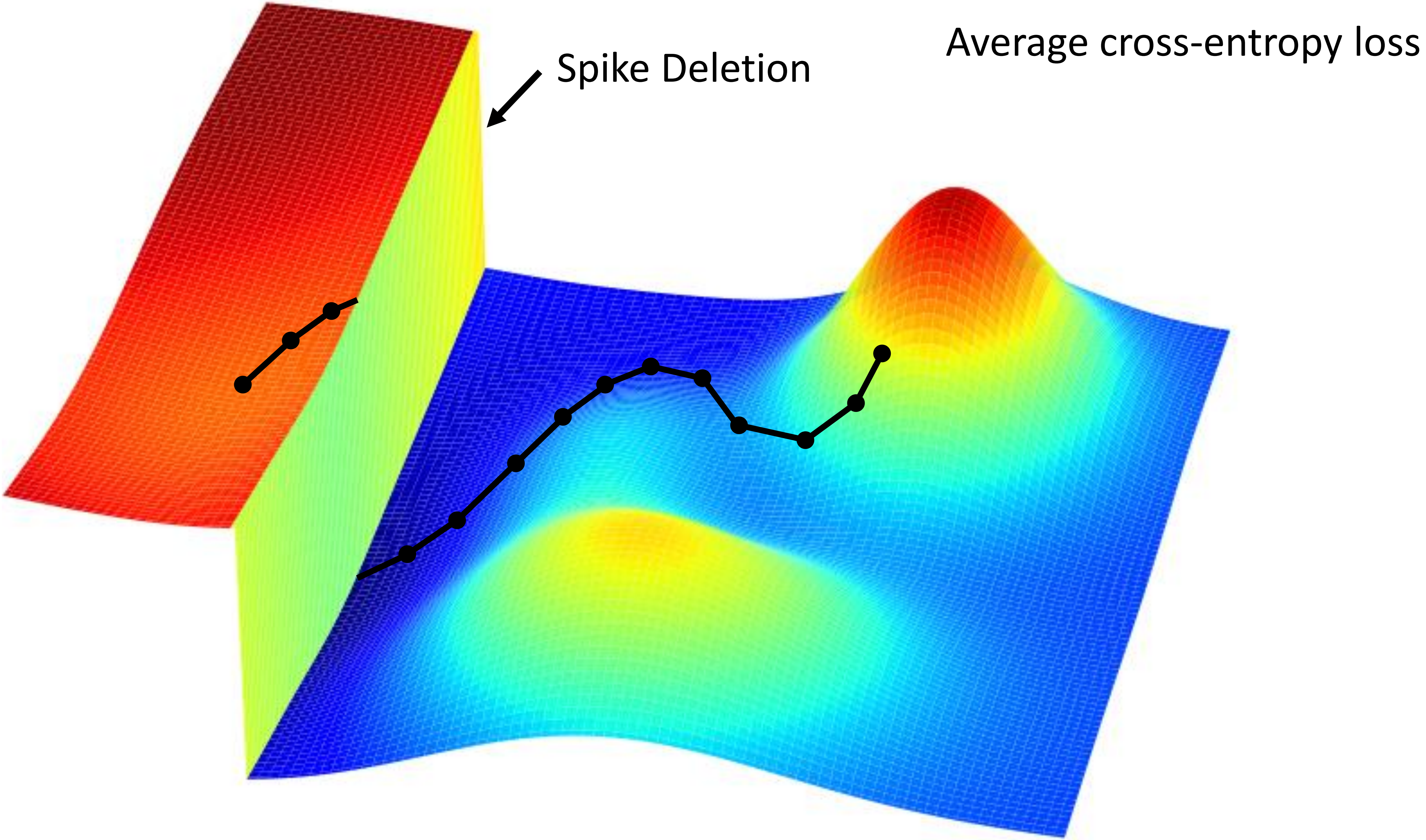



Cramer, B., Stradmann, Y., Schemmel, J., & Zenke, F. (2020). The Heidelberg Spiking Data Sets for the Systematic Evaluation of Spiking Neural Networks. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.3044364

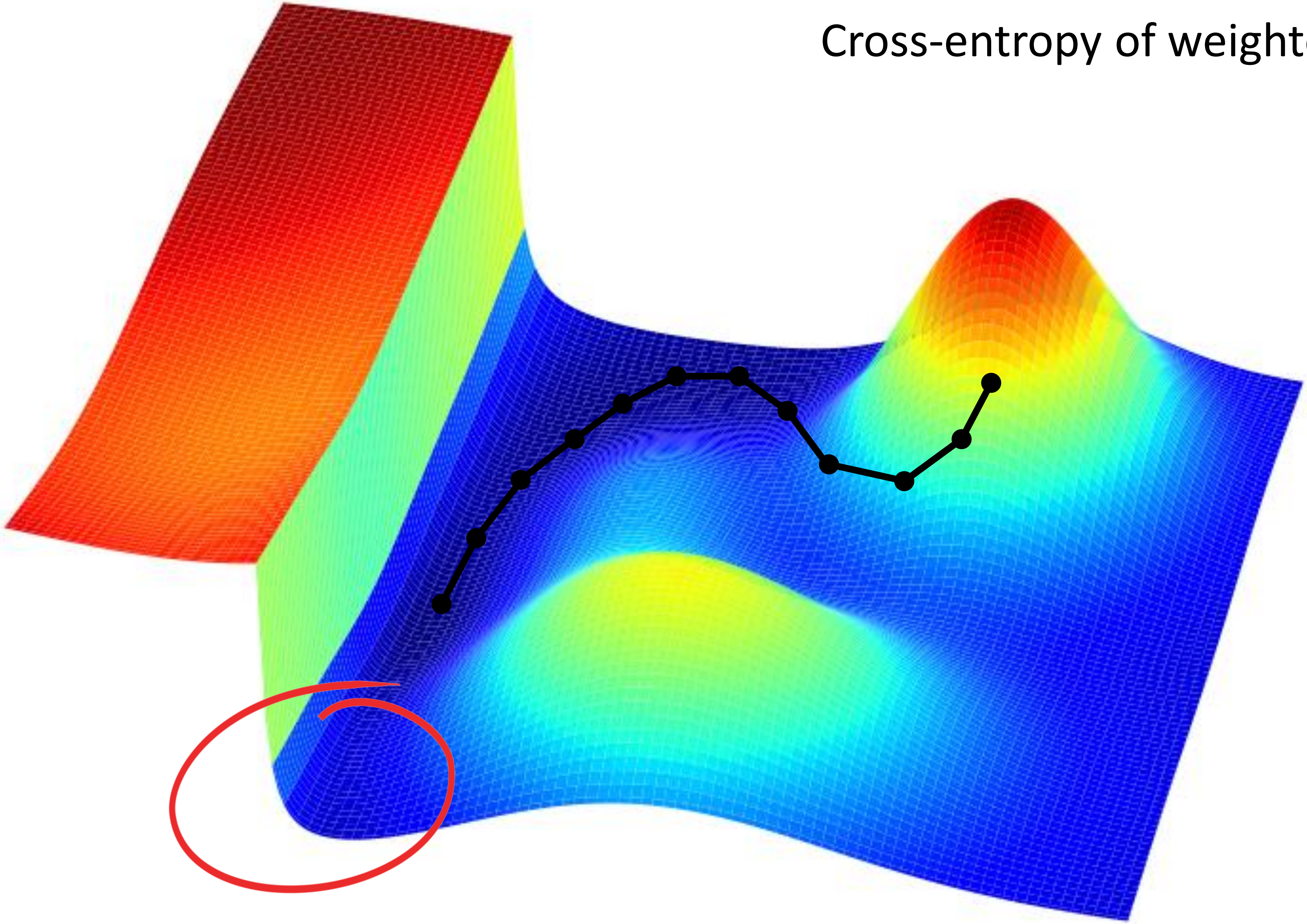
Sussex AI

# No learning!



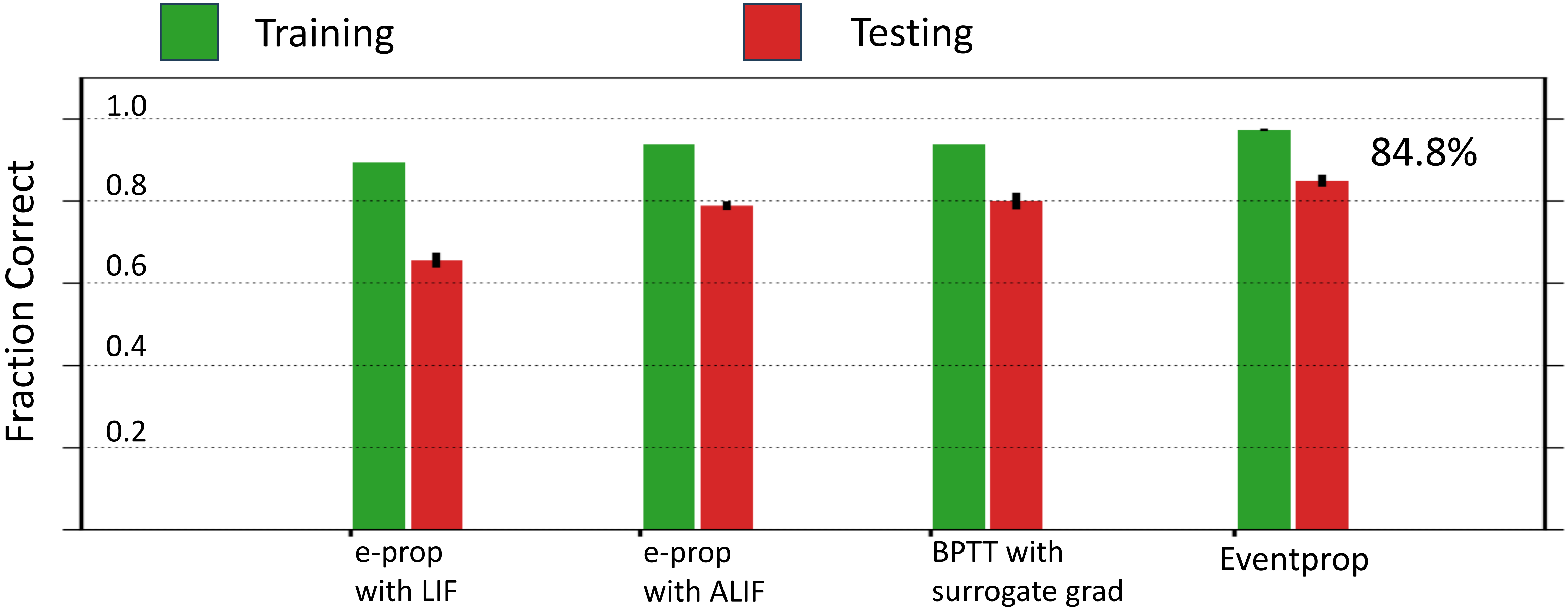

Average cross-entropy loss
Spike Deletion

# Cross-entropy of weighted average loss



# Spiking Heidelberg Digits

Sussex AI





1. Zenke F, Vogels T. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. Neural Comput 33:899-925 (2021)
2. J. C Knight, T. Nowotny. Efficient GPU training of LSNNs using eProp. NICE 2022. ACM, 8–10


EventProp
Spiking simulation
Gradient Descent
Augmentation

# Augmentation

Sussex AI

- Augmentation is used to make data sets larger to counter-act over-fitting to few examples
- Augmentation for event-based data is an open field with not much work
- Candidates:
  - Global random channel shift
  - Local random channel swap
  - Random time compression/dilation
  - Blending

# Parameter sweeps

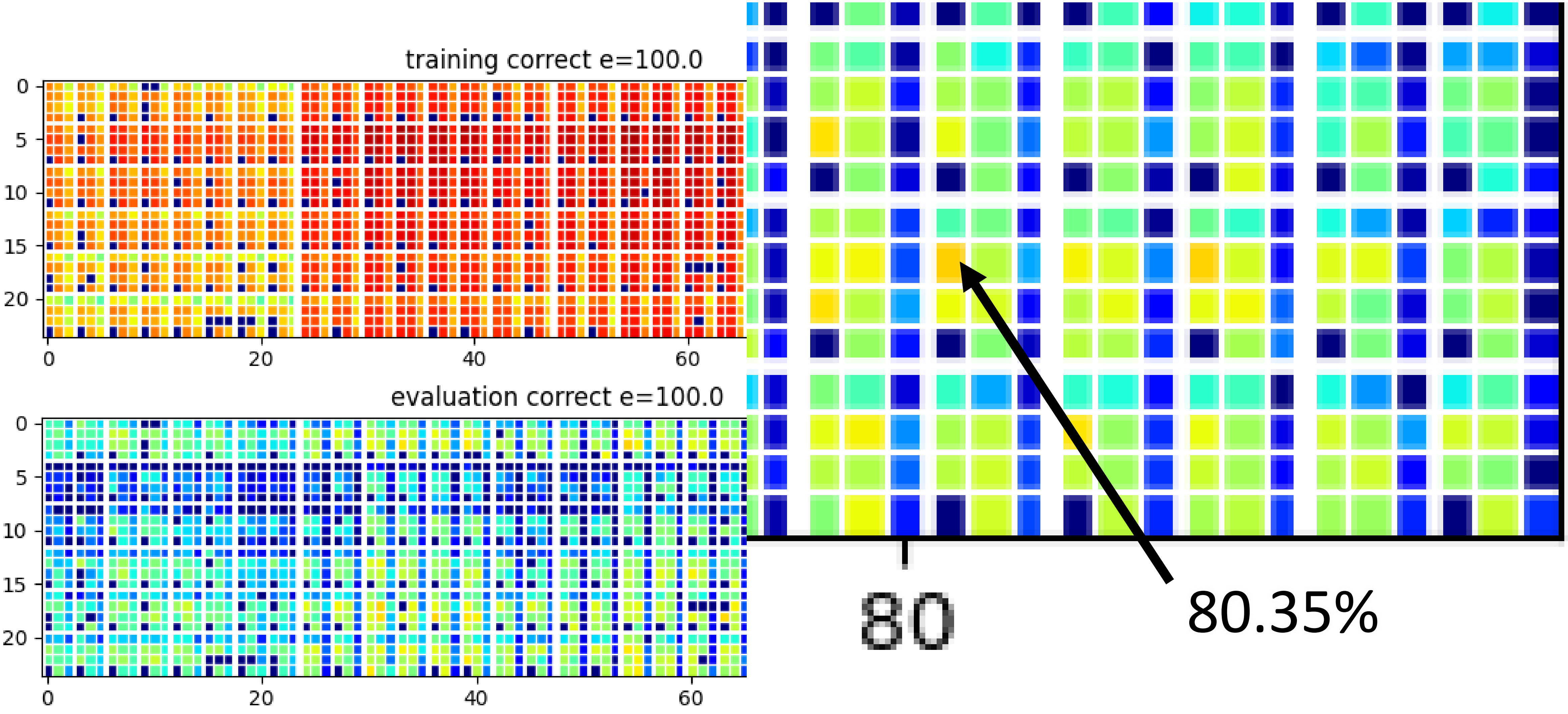
- Try parameters for the network and augmentations
- Best proxy for test set performance is leave-one-speaker-out (loso) cross-validation
- A full 10-fold loso cross-validation with 300 epochs per fold takes about **16 hours on an A100 GPU**

JADE
Tier 2 HPC


ORACLE Cloud
Research

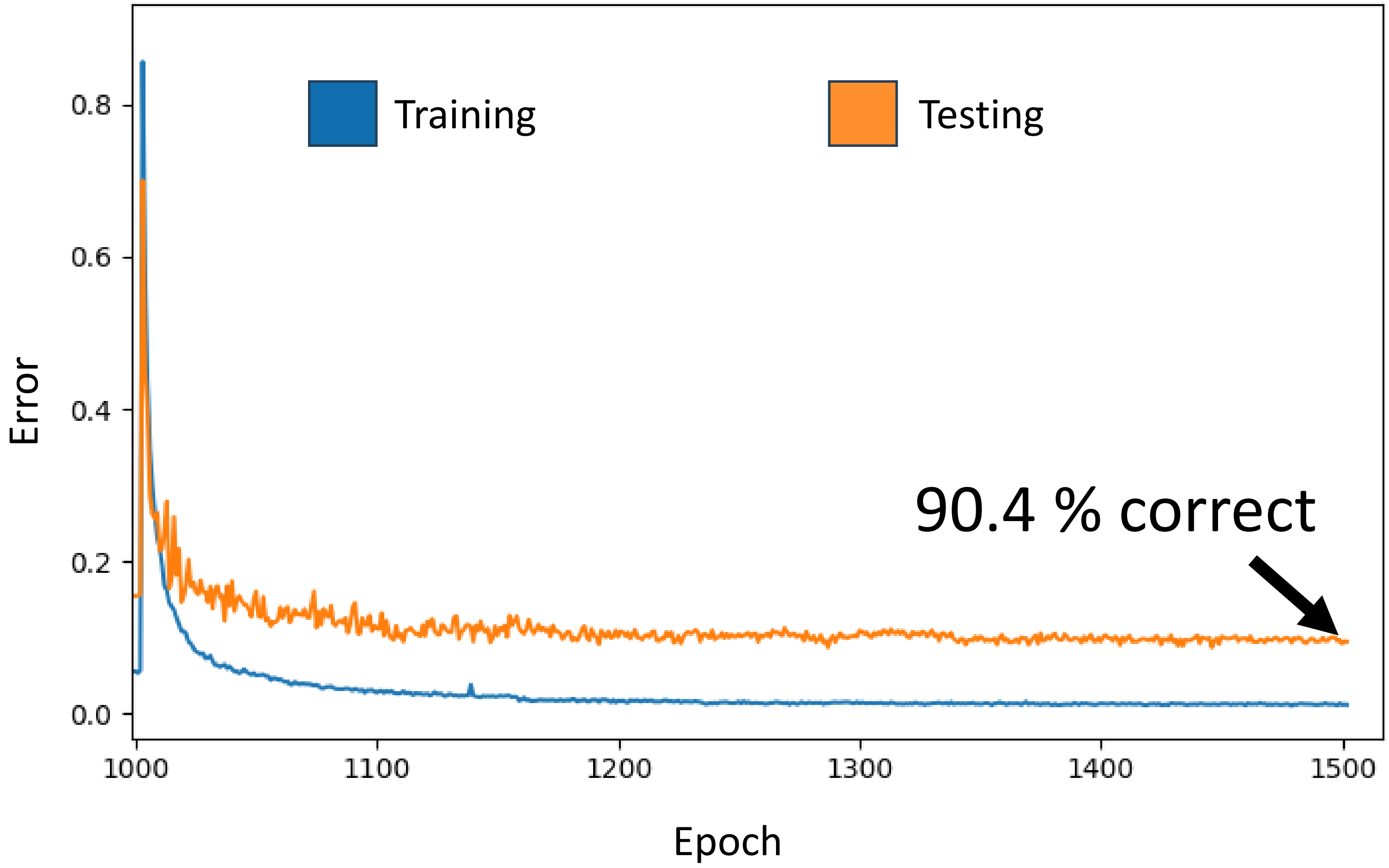
# Example Scan Results





# Test set performance





# Summary

Sussex AI

- JADE 2 (and other GPU clusters) help us investigate event-based neural network solutions competitive against SOTA

| Dataset | Method | Recurrent | Delays | # Params | Top1 Accuracy |
|---|---|---|---|---|---|
| **SHD** | EventProp-GeNN [31] | ✓ | ✗ | N/a | $84.80 \pm 1.5\%$ |
| | Cuba-LIF [7] | ✗ | ✗ | N/a | $87.80 \pm 1.1\%$ |
| | Adaptive SRNN [44] | ✓ | ✗ | N/a | 90.40% |
| | SNN with Delays [2] | ✗ | ✓ | 0.1M | 90.43% |
| | TA-SNN [43] | ✗ | ✗ | N/a | 91.08% |
| | STSC-SNN [46] | ✗ | ✗ | 2.1M | 92.36% |
| | Adaptive Delays [37] | ✗ | ✓ | 0.1M | 92.45% |
| | RadLIF [3] | ✓ | ✗ | 3.9M | 94.62% |
| | **Our work (2 hidden layers)** | ✗ | ✓ | **0.2M** | $\mathbf{95.07 \pm 0.24\%}$ |

← 90.4 % correct (Adaptive SRNN)

} Use test set for validation (RadLIF, Our work)

Hammouamri, I., Khalfaoui-Hassani, I., & Masquelier, T. (2023). Learning Delays in Spiking Neural Networks using Dilated Convolutions with Learnable Spacings. *arXiv preprint arXiv:2306.17670*.

# Acknowledgements

Sussex AI


- **Jamie Knight**
- **Other GeNN developers:**
  - Esin Yavuz
  - James Turner
  - Anton Komissarov
  - Fawad Ali, Obaid-Ur-Rehman, Muhammad Irfan Ali, University of Engineering and Technology, Taxila, PAKISTAN (OpenCL backend - internship)

The Leverhulme Trust

ROYAL ACADEMY OF ENGINEERING

NVIDIA

GREEN BRAIN

brains on BOARD

Human Brain Project

EPSRC Engineering and Physical Sciences Research Council

JADE Tier 2 HPC




ORACLE Cloud Research