TEXAS
INC LAB
Cockrell School of Engineering

Leveraging Physical Properties of 2D Materials for Neuromorphic Computing

Prof. Jean Anne Incorvia
Dept. of Electrical and Computer Engineering
The University of Texas at Austin

utinclab.com
incorvia@austin.utexas.edu

TEXAS
INC
LAB
Cockrell School of Engineering

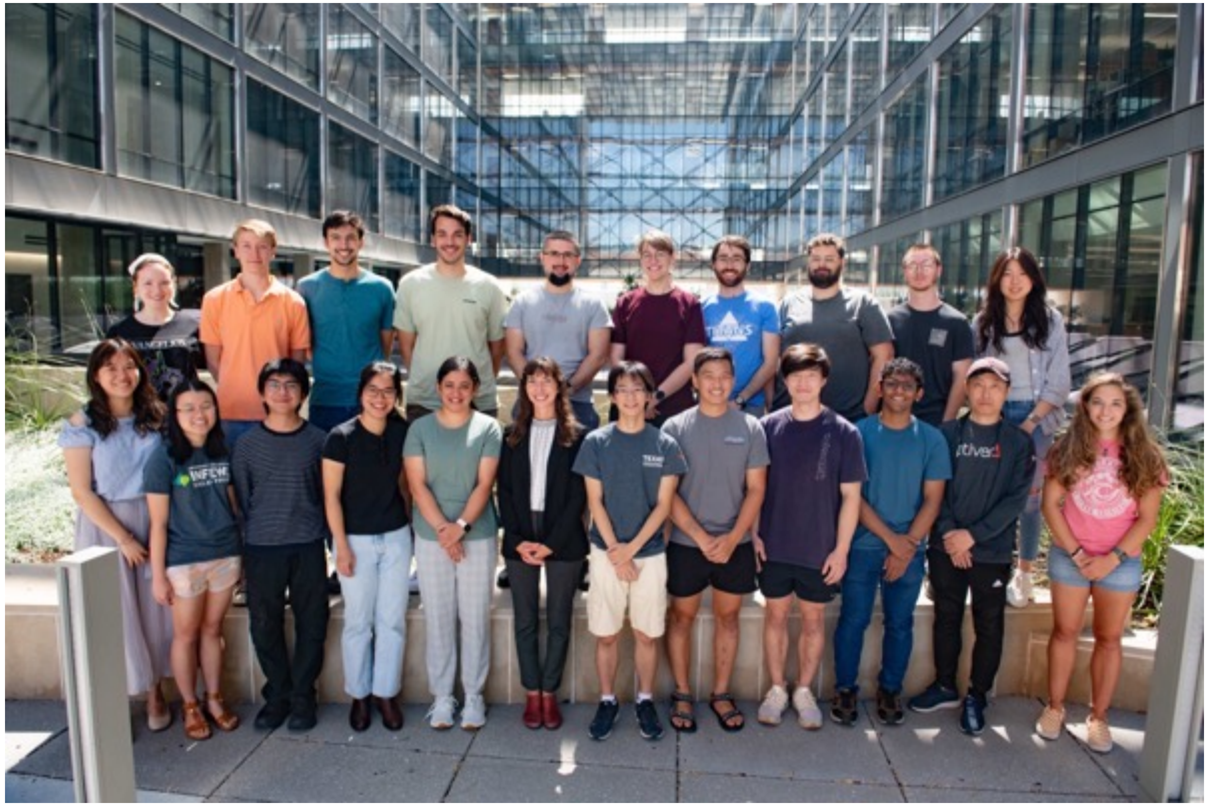
Acknowledgements

Sam Liu

Dr. Dmitry Kireev
UMass Amherst

Prof. Deji Akinwande
UT Austin

Harrison Jin

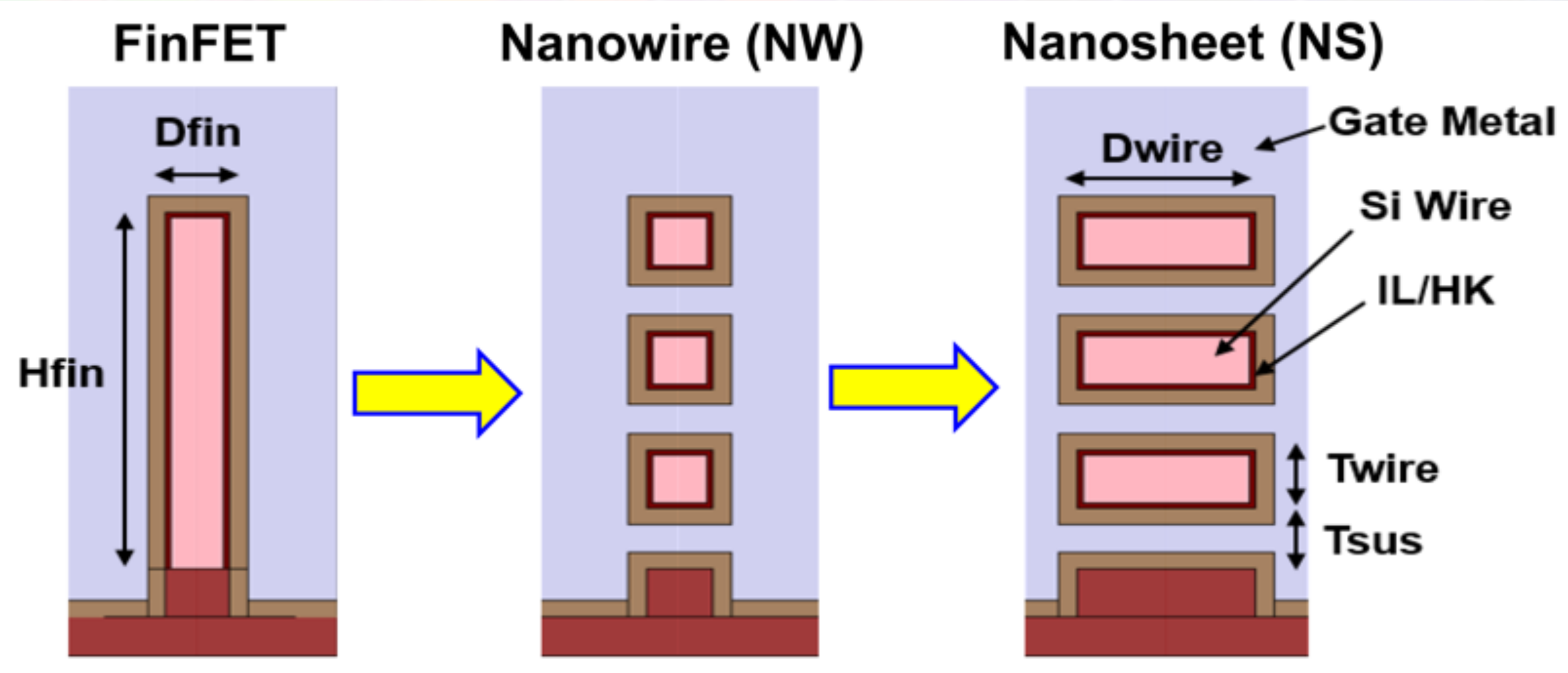


www.utinclab.com

OUTLINE

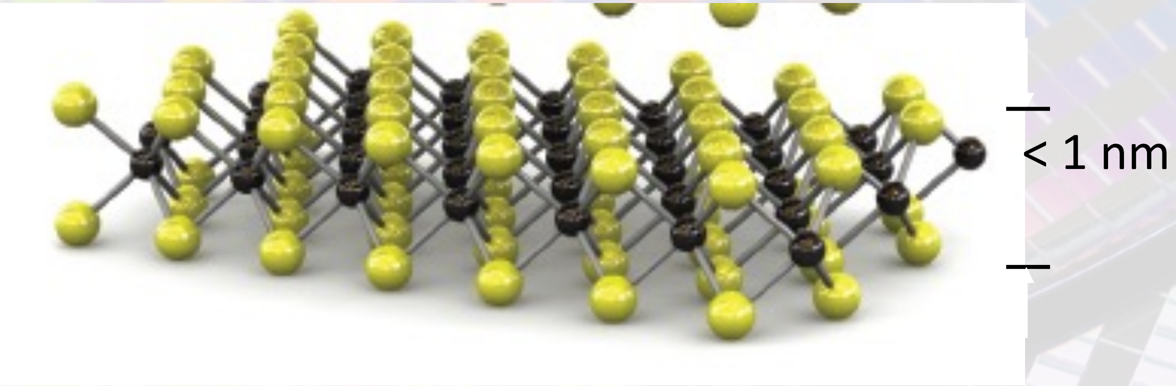
1. Motivation for using 2D materials for computing with physical systems
2. Graphene/Nafion transistors as artificial synapses
3. Graphene/Nafion transistors as artificial dendrites
4. Conclusions & Outlook

Transistors roadmap



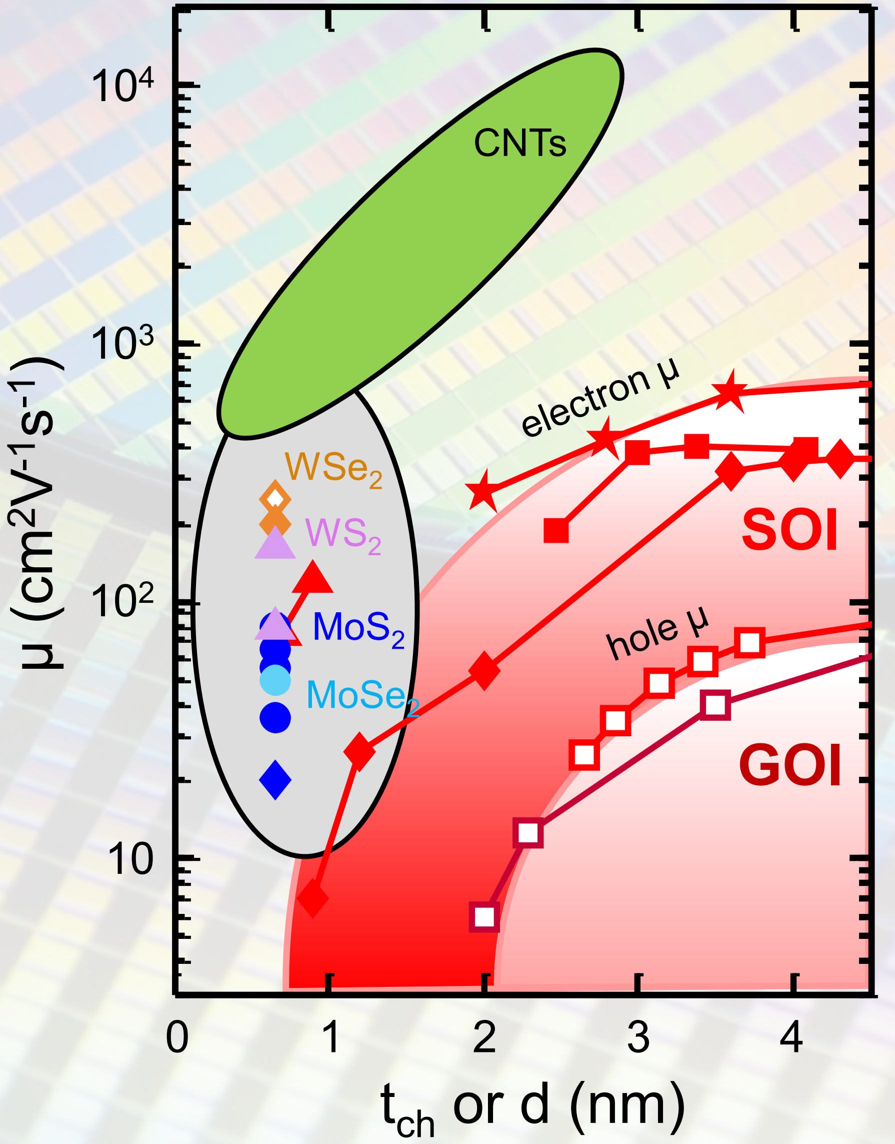
Source: N. Loubet (IBM), Symp. VLSI Technology short course SC1-1 (2020)

2D materials: on the roadmap



- Naturally atomically thin
- Van der Walls bonds at surface
- Conducting, semiconducting, insulating, etc. with bandgap tunable by layer number

2D materials: on the roadmap



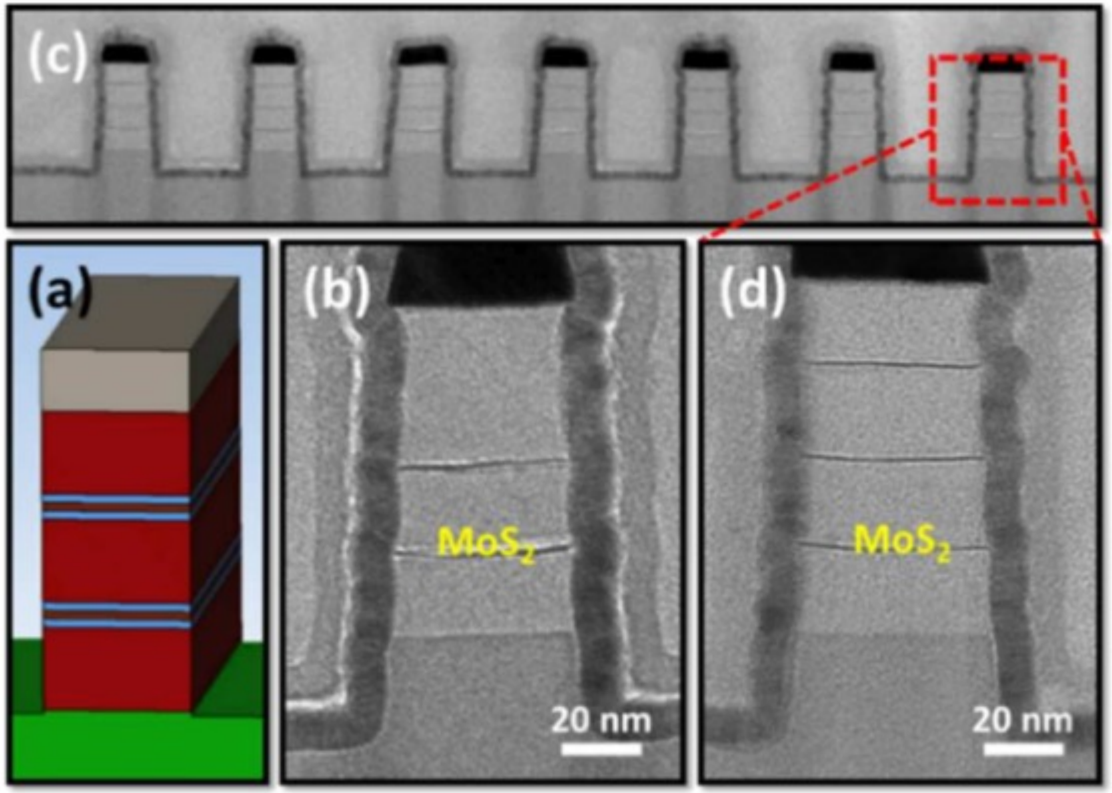


Fig. 7. Module 2 of fin patterning. (a) sketch of the structure. TEM cross sections of (b) 2-tier monolayer MoS_2 (c) and (d) 3-tier monolayer MoS2 fin structure.

Sources: K. Uchida, A. Kis, IBM, E. Pop (C. English, Nano Lett '16)
Y. -Y. Chung et al., 2022 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2022, pp. 34.5.1-34.5.4, doi: 10.1109/IEDM45625.2022.10019563.

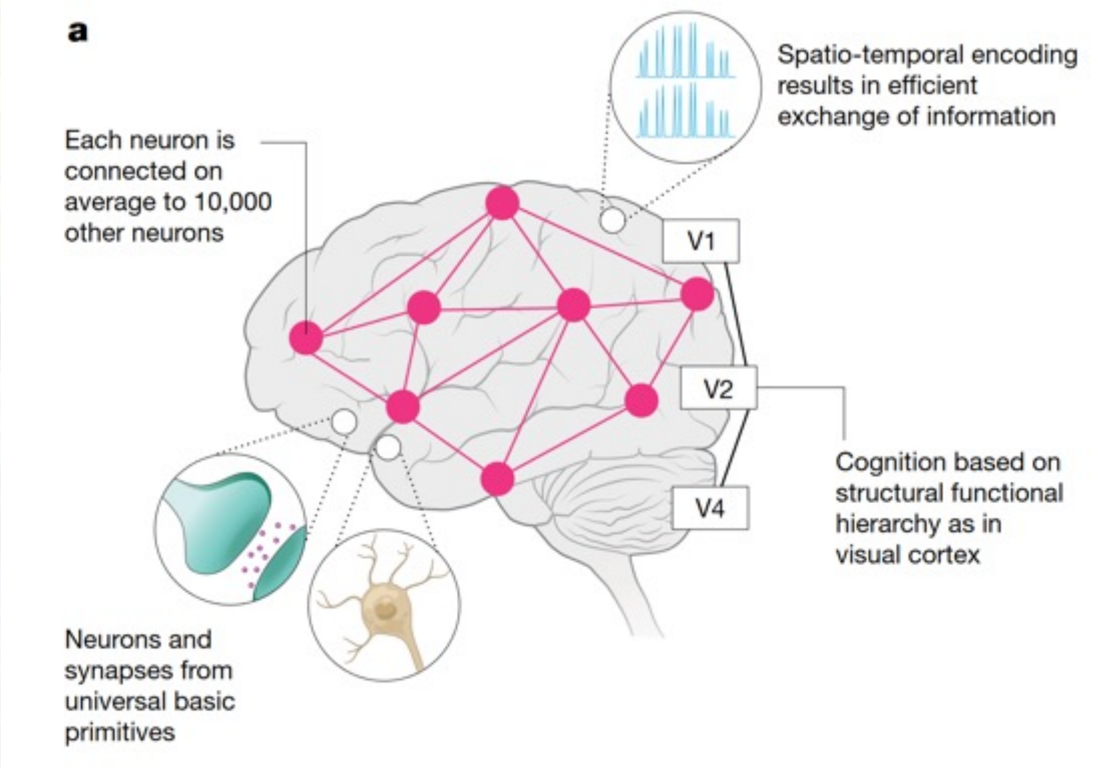
2D materials provide

1. Naturally scaled
2. Flexible
3. New physics that emerges from electron confinement and/or building heterostructures: ambipolarity, high mobility, gateable effects, highly tunable interfaces
4. Non-toxic when interacting with biological systems such as cells (e.g. graphene[1])

[1]Fabbro, A. et al. Graphene-based interfaces do not alter target nerve cells. ACS Nano 10, 615–623 (2016).

Spiking neural networks

- **Spatio-temporal encoding** of neuron spikes allows expressive, low energy computation
- Analog SNNs on biological timescales (ms-s), can act as potential bio-interfaces



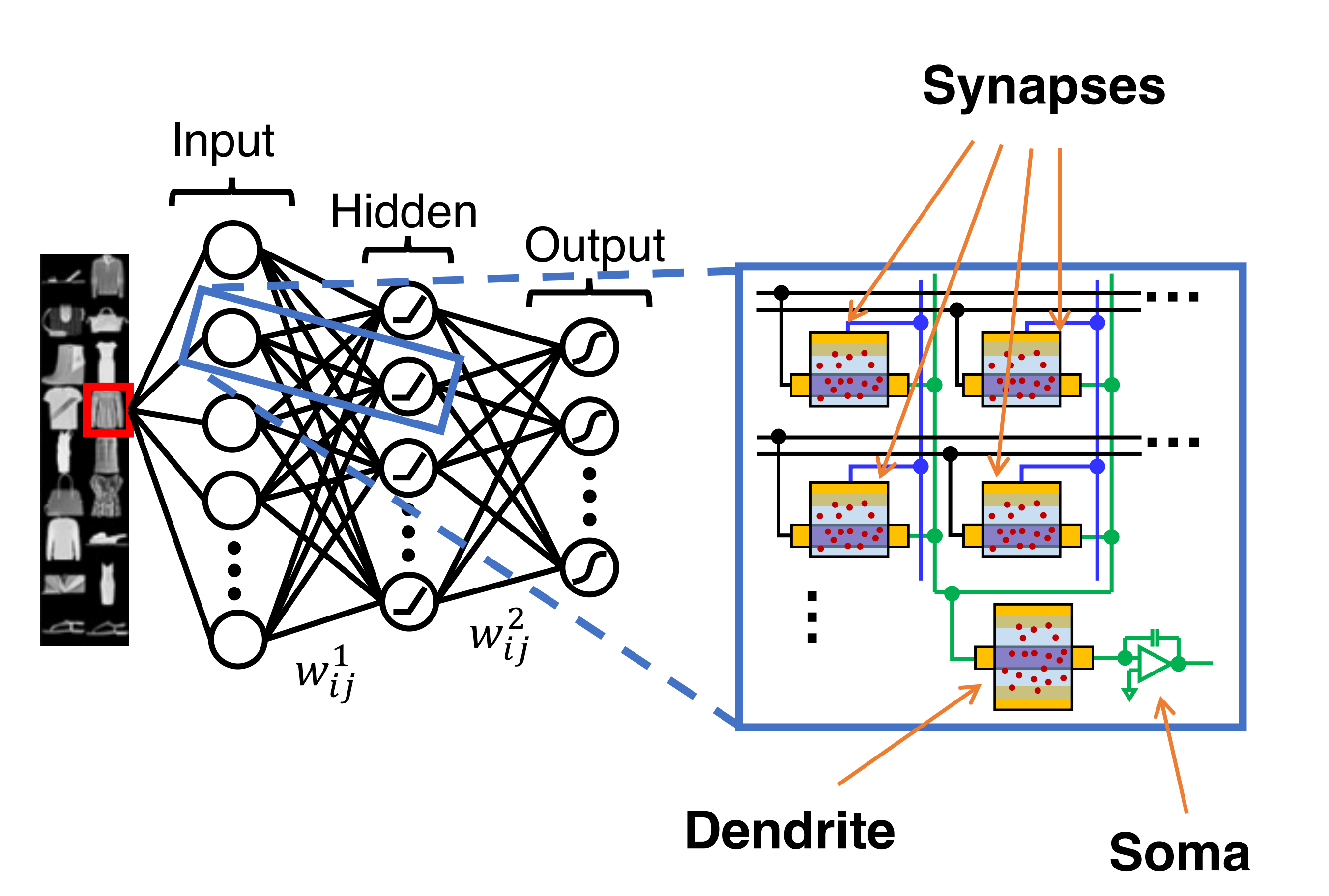
Roy, K., Jaiswal, A., & Panda, P. (2019). Towards spike-based machine intelligence with neuromorphic computing. *Nature*, *575*(7784), 607–617.

Spiking neural network building blocks

- **Synapses** connect neurons together and act as **memory**
- **Neurons** integrate incoming signals and activate

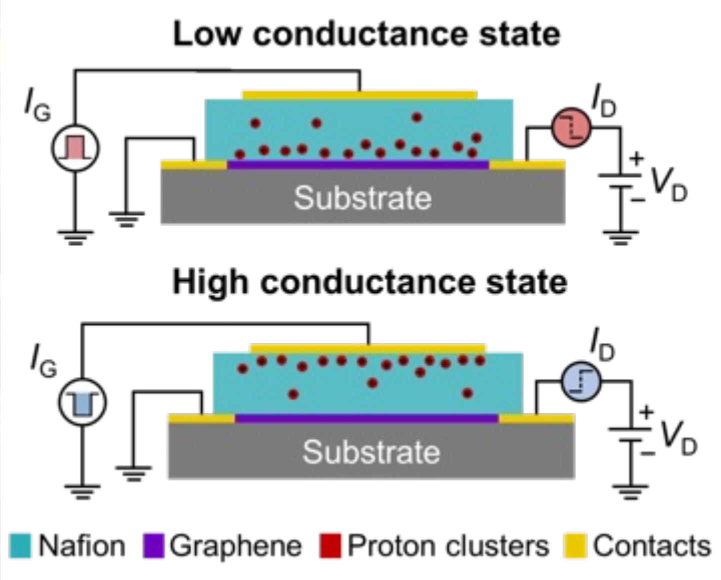
→ **Dendrites** represent internal states

→ **Soma** sets the firing threshold



Graphene/Nafion transistor provides high tunability for emulating NN building blocks

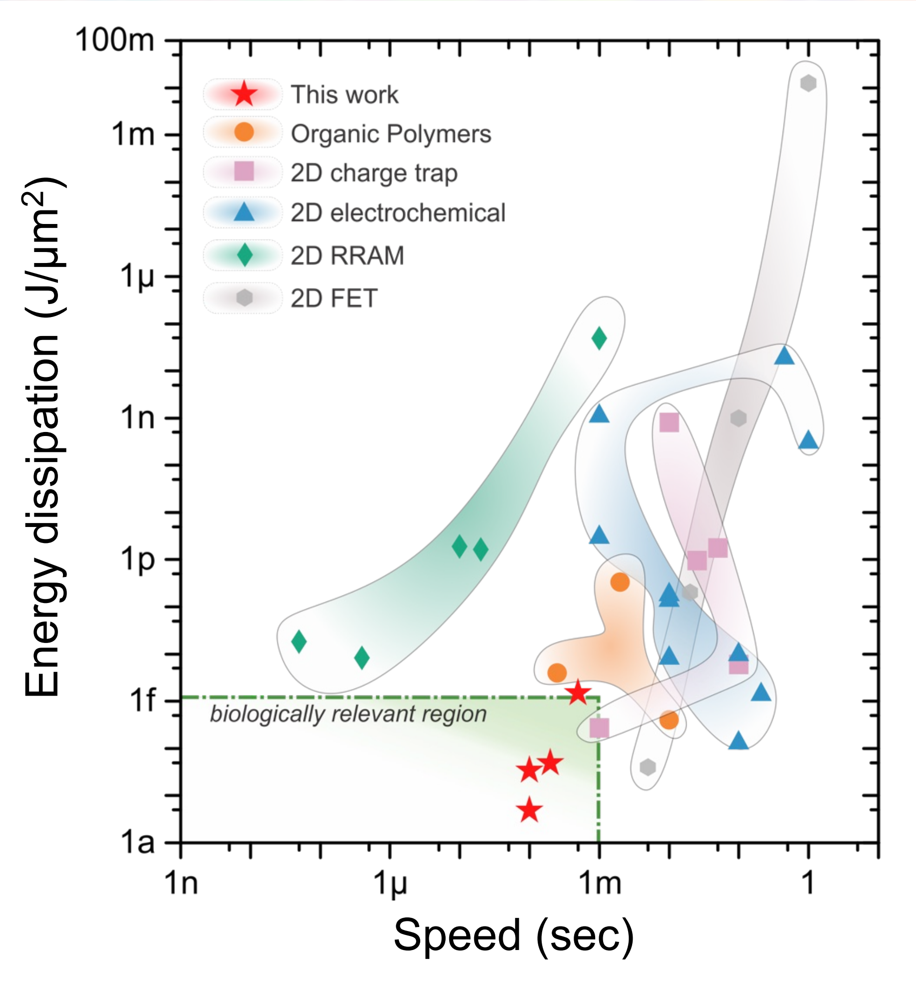
- Combination of **few-layer graphene** channel and **Nafion-117** ionic membrane
- Gate current I_G changes electronic state in device, causing conductance change in graphene channel

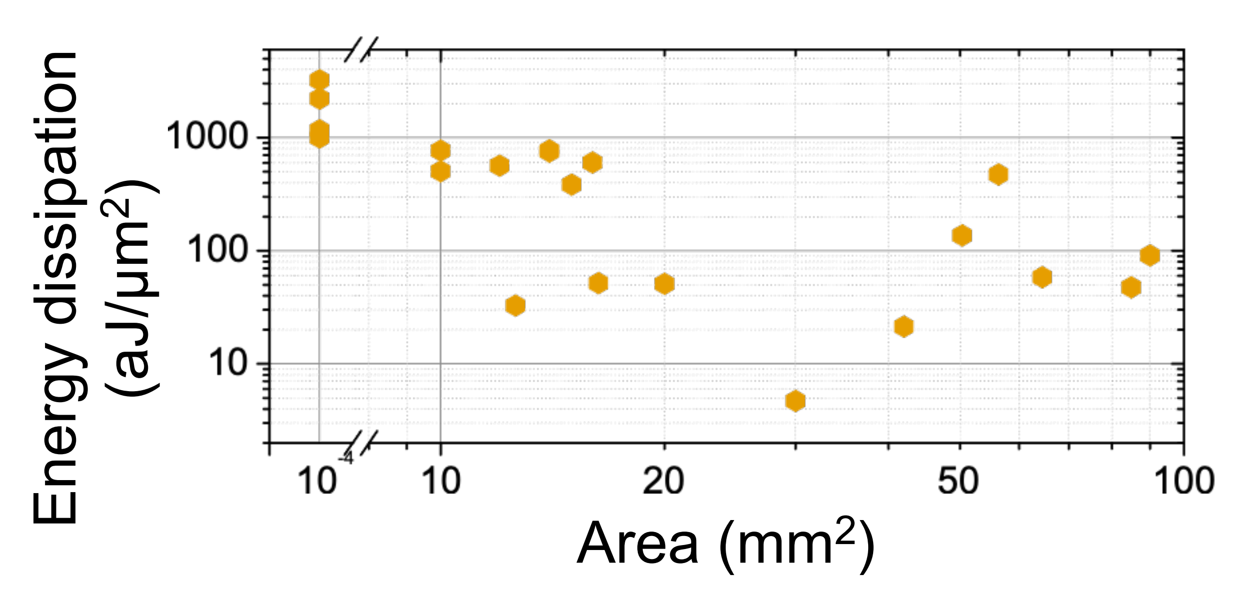


Kireev, D., Liu, S., Jin, H., Patrick Xiao, T., Bennett, C. H., Akinwande, D., & Incorvia, J. A. C. Metaplastic and energy-efficient biocompatible graphene artificial synaptic transistors for enhanced accuracy neuromorphic computing. *Nature Communications*, *13*(1), 4386 (2022).

Graphene/Nafion in class of ionic neuromorphic devices

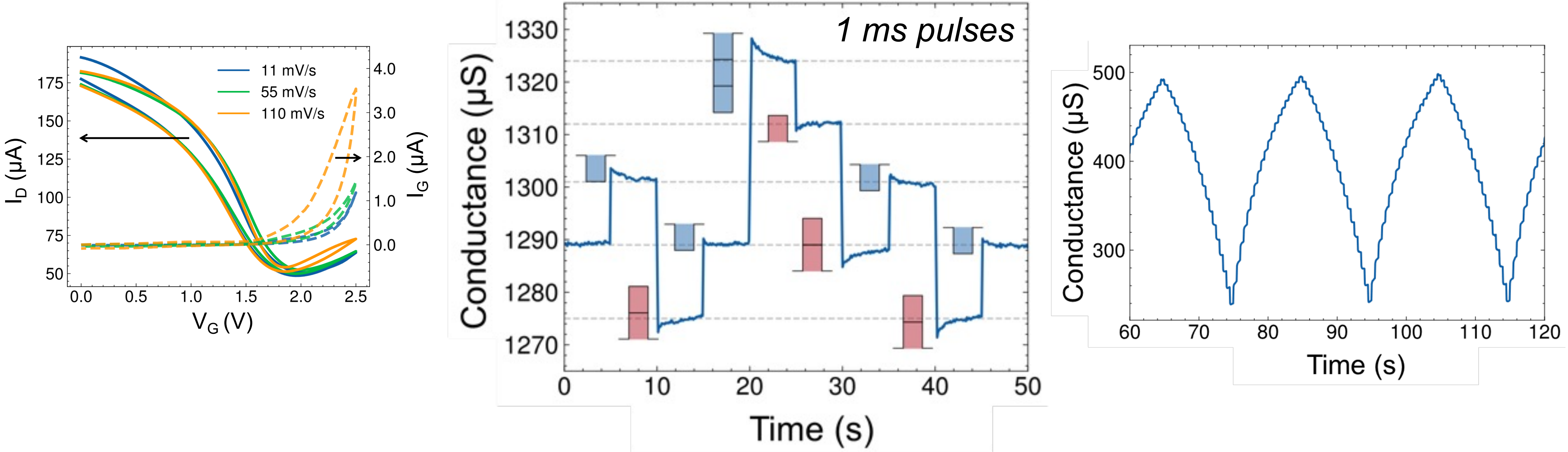
- Shows low energy per weight update
- Made from bio-compatible materials



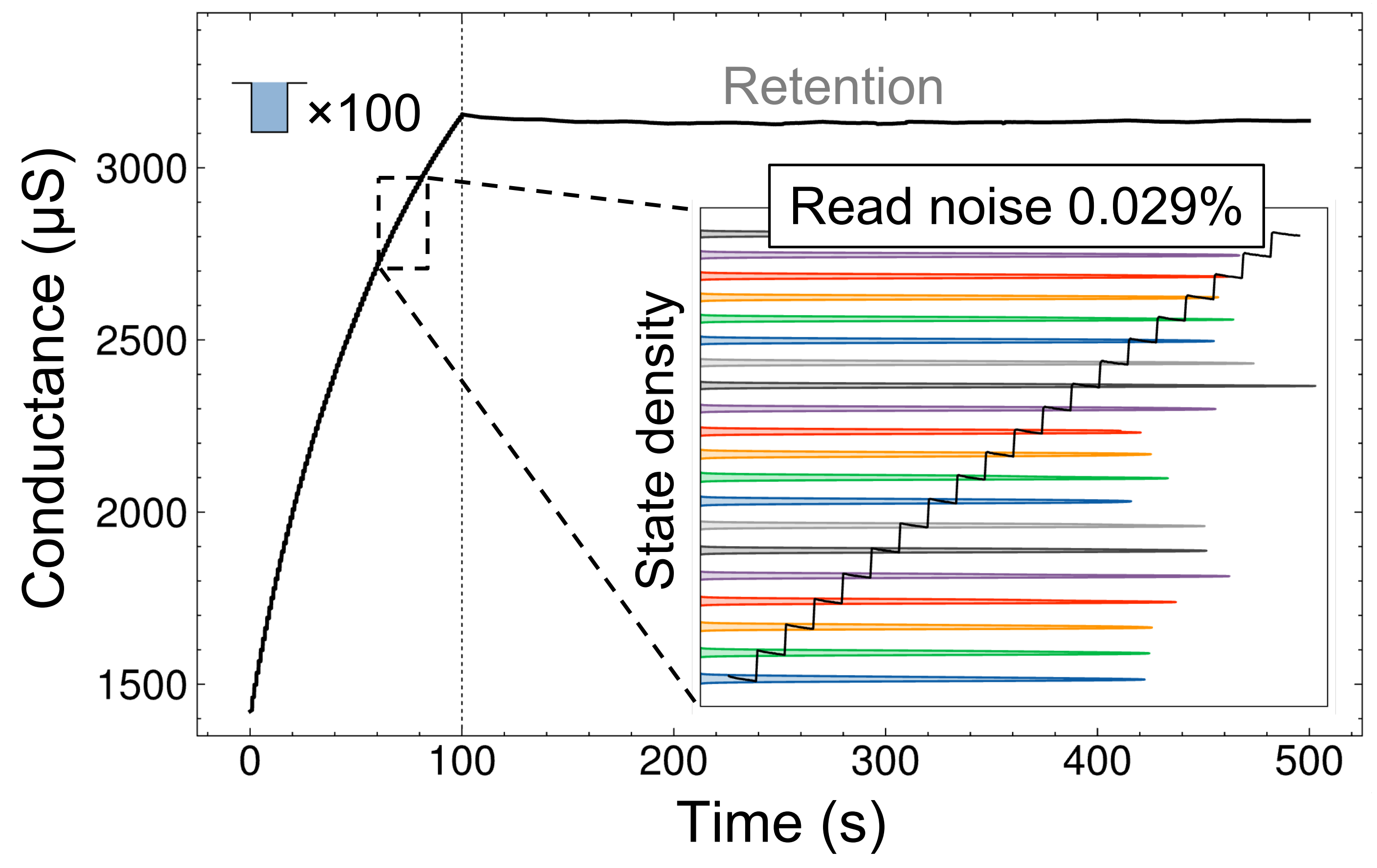


Kireev, D., Liu, S., Jin, H., Patrick Xiao, T., Bennett, C. H., Akinwande, D., & Incorvia, J. A. C. *Nature Communications*, *13*(1), 4386 2022).

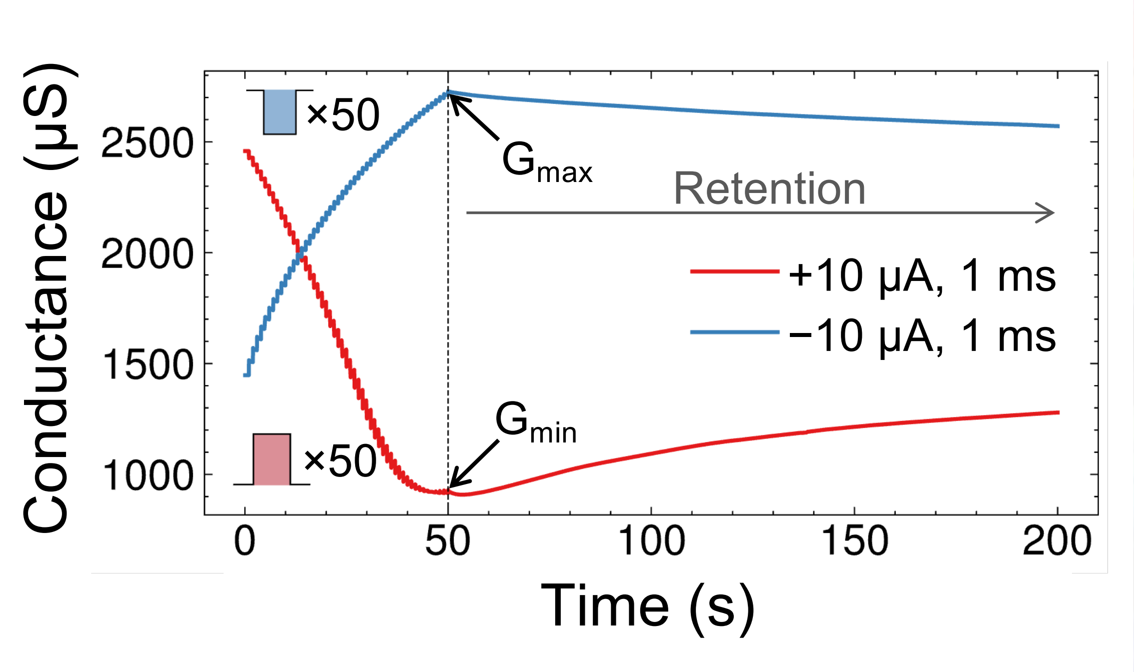
Emulation of synaptic operation

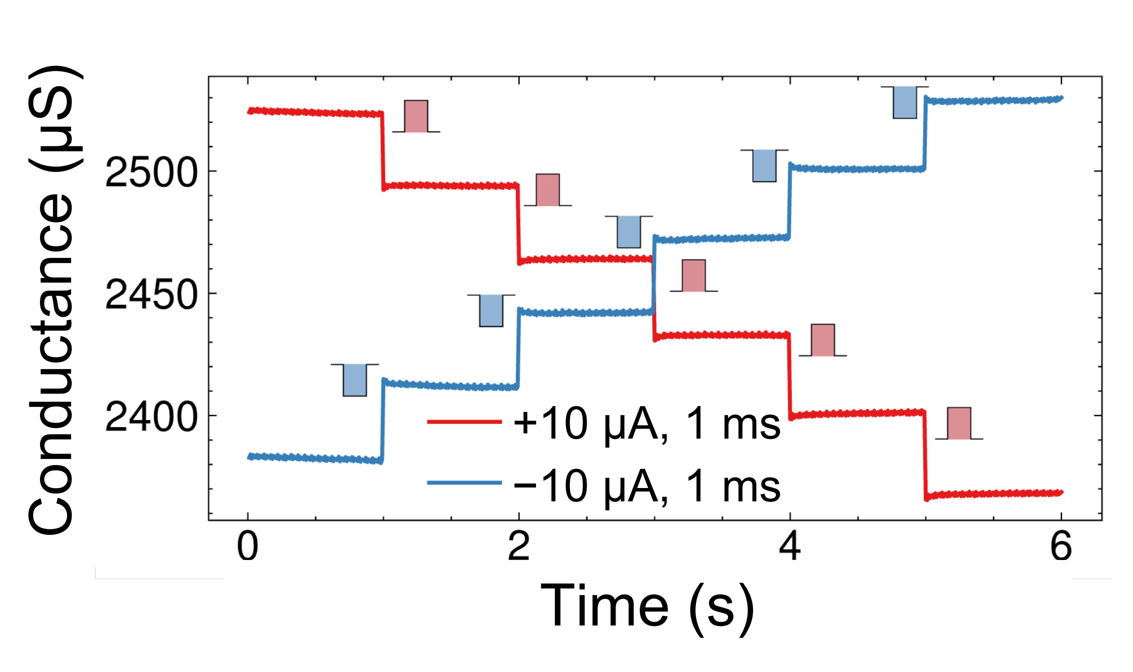


High state density achievable



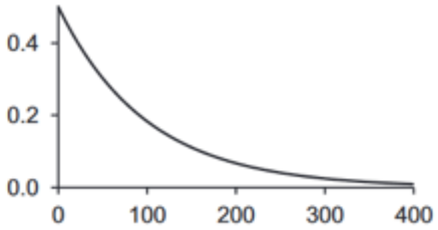
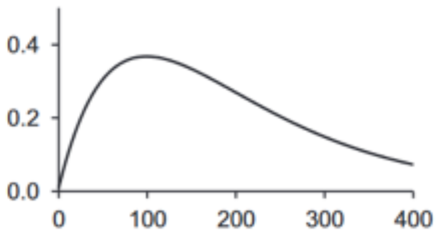
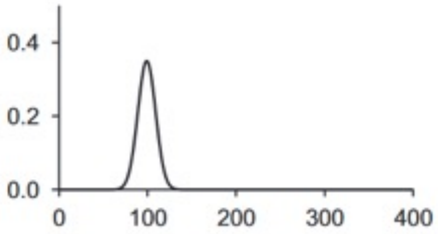
Retention sufficient for up to ~s timescale operation





Biological dendrites show tunable response function in time to incoming spikes

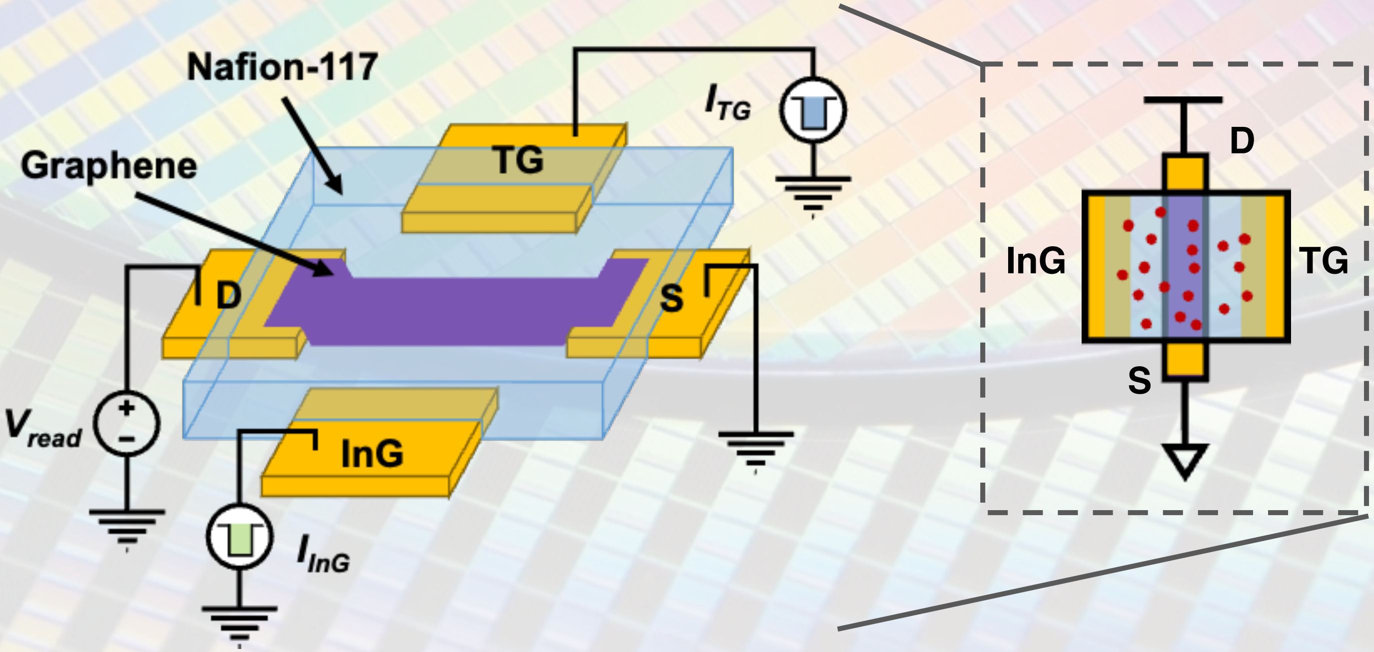
Table 1 | Typical synaptic kernels in mathematical and graphical form.

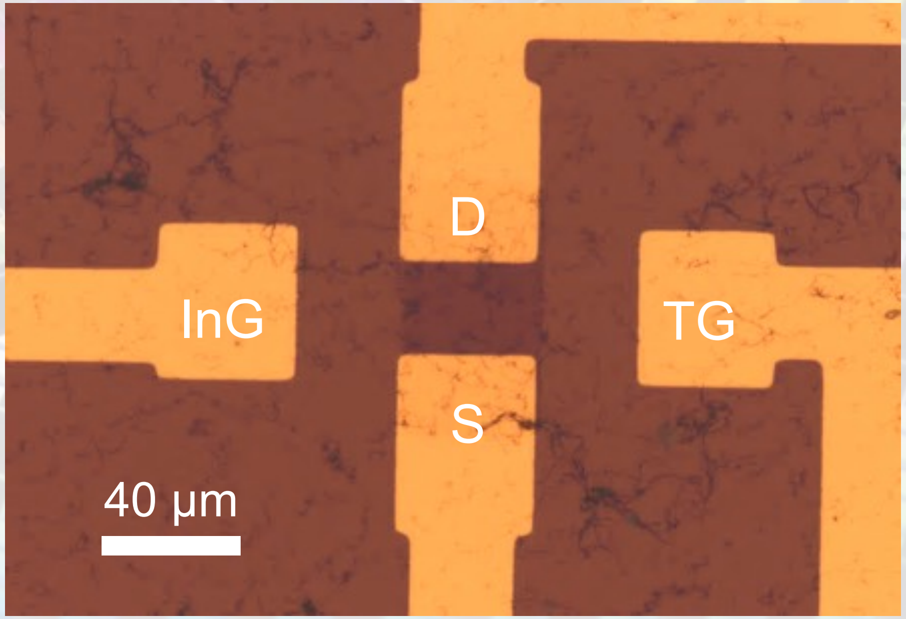
	Kernel type	Mathematical expression for filter response	Typical function (Spike at $t = 0$)
Leaking	Stable recurrent connection (leaky integration) with non-linear leak	$\mathcal{F}_j(\cdot) = \frac{1}{1+(\mathcal{F}(t-\tau))^2}\int_{t_0}^{t}\sum_{i=1}^{L} w_{ji}^{(1)} x_{i,t}\,dt$	
Alpha	Alpha function followed by compressive non-linearity	$\mathcal{F}_j(\cdot) = \tanh\left(\left[\sum_{i=1}^{L} w_{ji}^{(1)} x_{i,t}\right]\frac{t}{\tau}e^{-\frac{t}{\tau}}\right)$	
Gaussian	Synaptic or dendritic delay with Gaussian function, followed by compressive non-linearity	for $t \geq \Delta T$: $\mathcal{F}_j(\cdot) = \tanh\left(\left[\sum_{i=1}^{L} w_{ji}^{(1)} x_{i,t}\right]\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(t-\Delta T)^2}{2\sigma^2}}\right)$ for $t < \Delta T$: $\mathcal{F}_j(\cdot) = 0$	

Tapson, J. C., Cohen, G. K., Afshar, S., Stiefel, K. M., Buskila, Y., Wang, R. M., Hamilton, T. J., & van Schaik, A. (2013). Synthesis of neural networks for spatio-temporal spike pattern recognition and processing. *Frontiers in Neuroscience*, 7.

Dual gate operation for tunable dendritic response

InG = Input Gate
TG = Tuning Gate



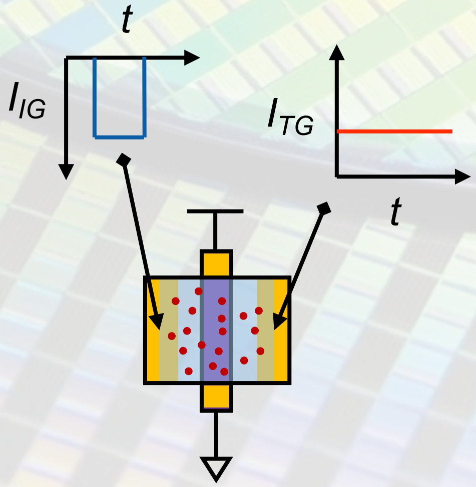


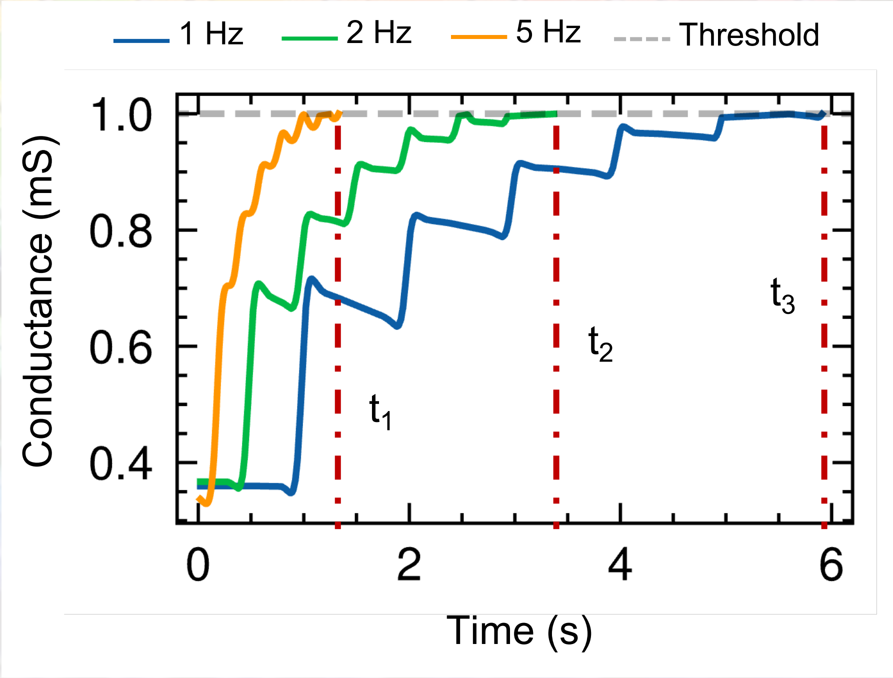
S. Liu, D. Akinwande, D. Kireev, and J. A. C. Incorvia. "Graphene-Based Artificial Dendrites for Bio-Inspired Learning in Spiking Neuromorphic Systems." *Under review. ArXiv 2310.02364.*

Leaking response

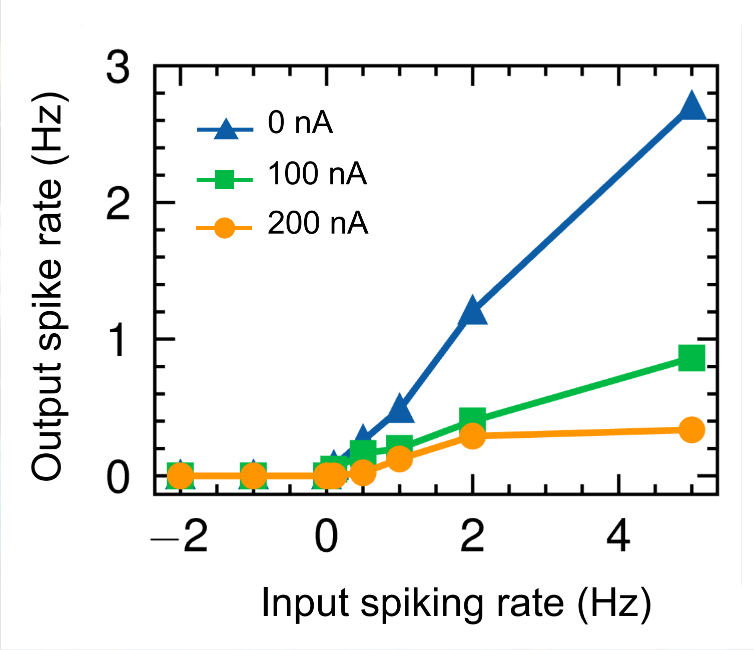
-10 µA, 1 ms pulse to InG as the spiking input signals at various frequencies
100 nA DC to TG

Rectified Linear Unit (ReLU): threshold and slope mediated by tuning gate amplitude





Varying InG input spiking frequency

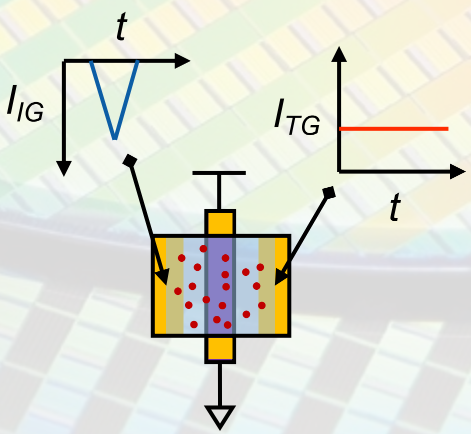


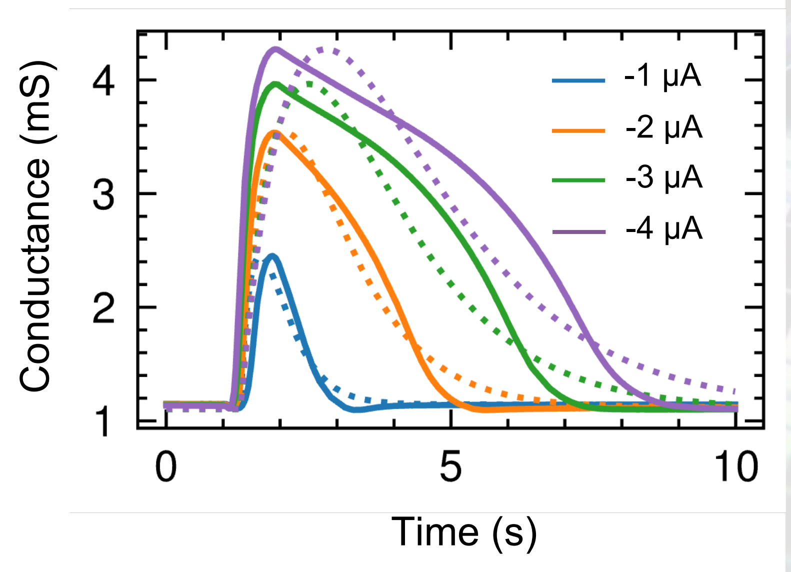
Varying TG amplitude

Alpha response

Triangular current pulses, 500 ms half-max pulse duration, to InG

Varying InG pulse amplitude with TG DC = 200 nA



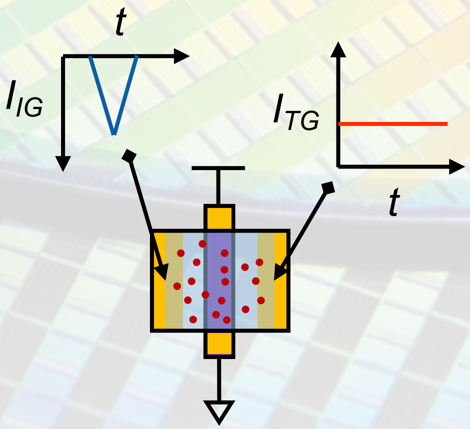


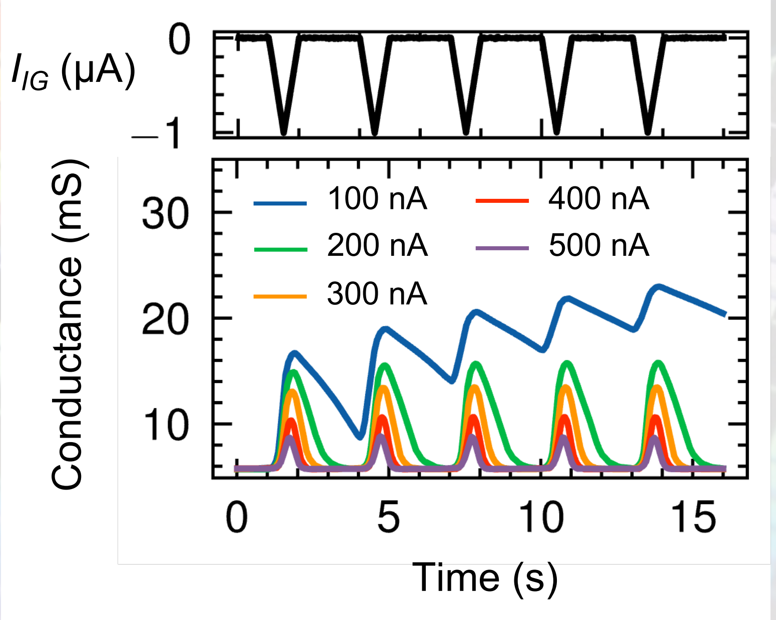
Solid = data
Dotted = ideal alpha response

Gaussian response

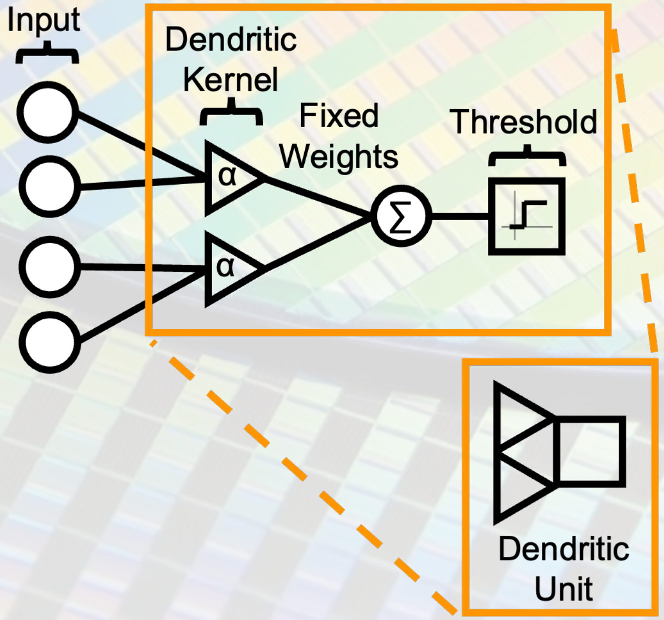
Triangular current pulses, 500 ms half-max pulse duration, to InG

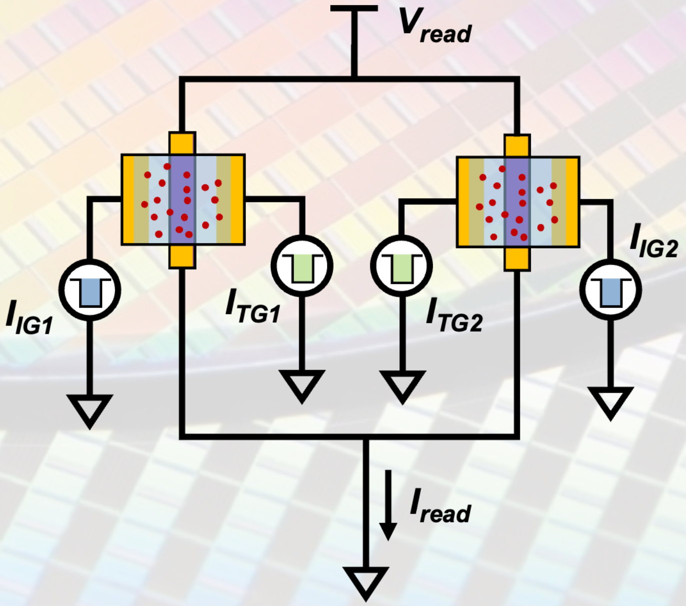
InG pulses with fixed -1 μA amplitude, varying TG DC amplitude

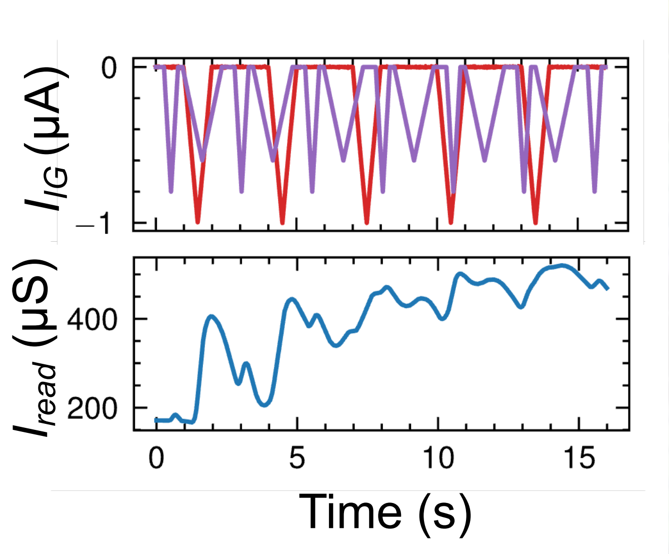




Multi-device behavior of alpha response

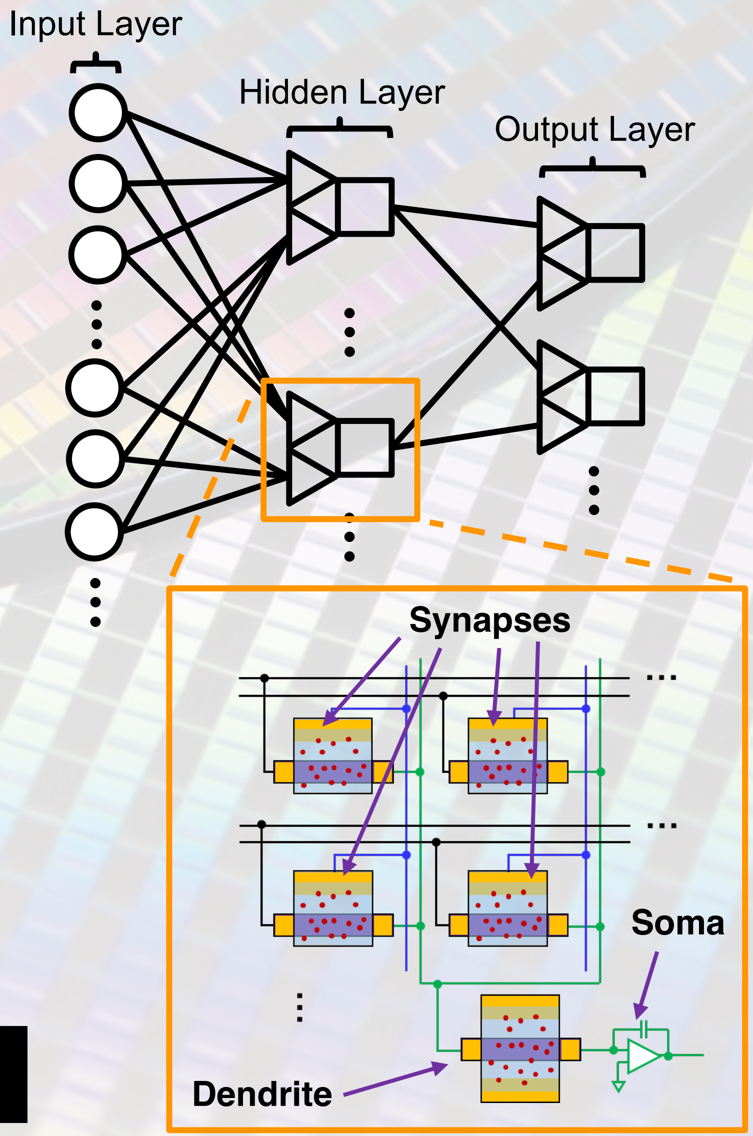






Neuromorphic classification performance

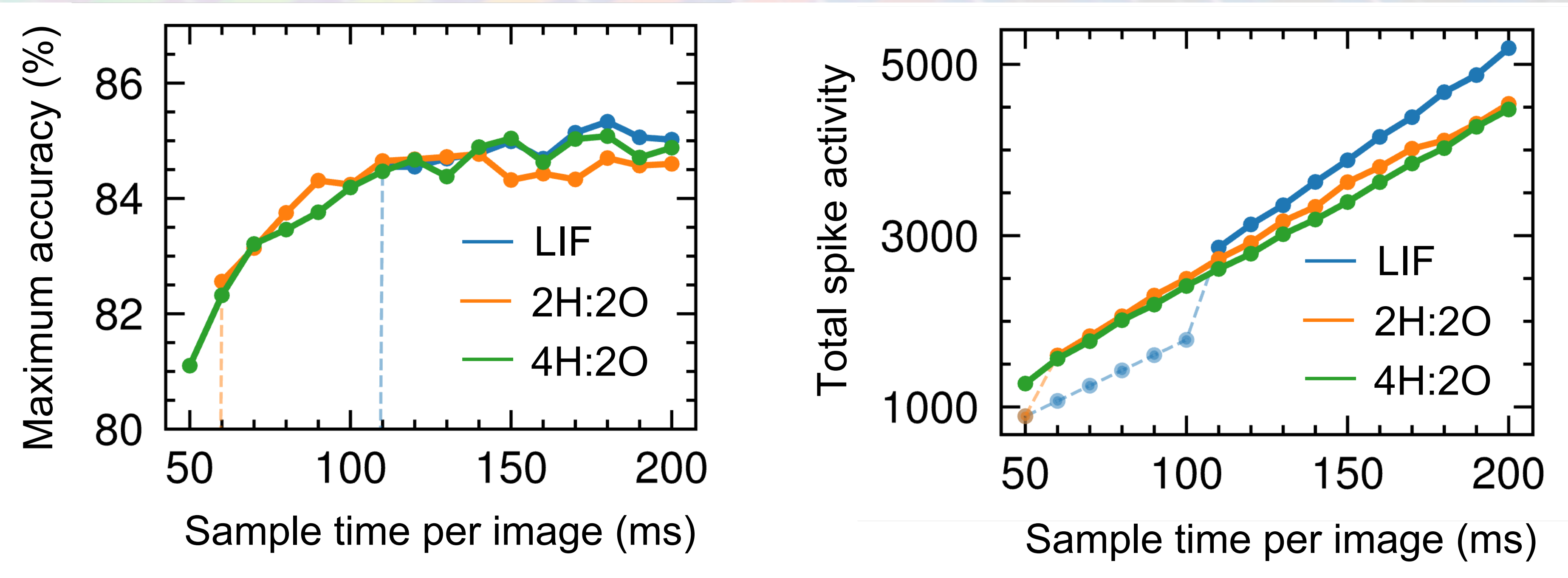
- Measured data input into simulated supervised SNN for both synapses and neurons/dendrites
- Fashion-MNIST classification benchmark task
- Training simulated with custom module using Norse and Pytorch
- Multilayer perceptron w/ one hidden layer of 200 units
- # of dendrites per dendritic unit varied in both the hidden layer (H) and output layer (O)



S. Liu, D. Akinwande, D. Kireev, and J. A. C. Incorvia. *Under review. ArXiv 2310.02364.*

Online learning performance vs. sampling time per image

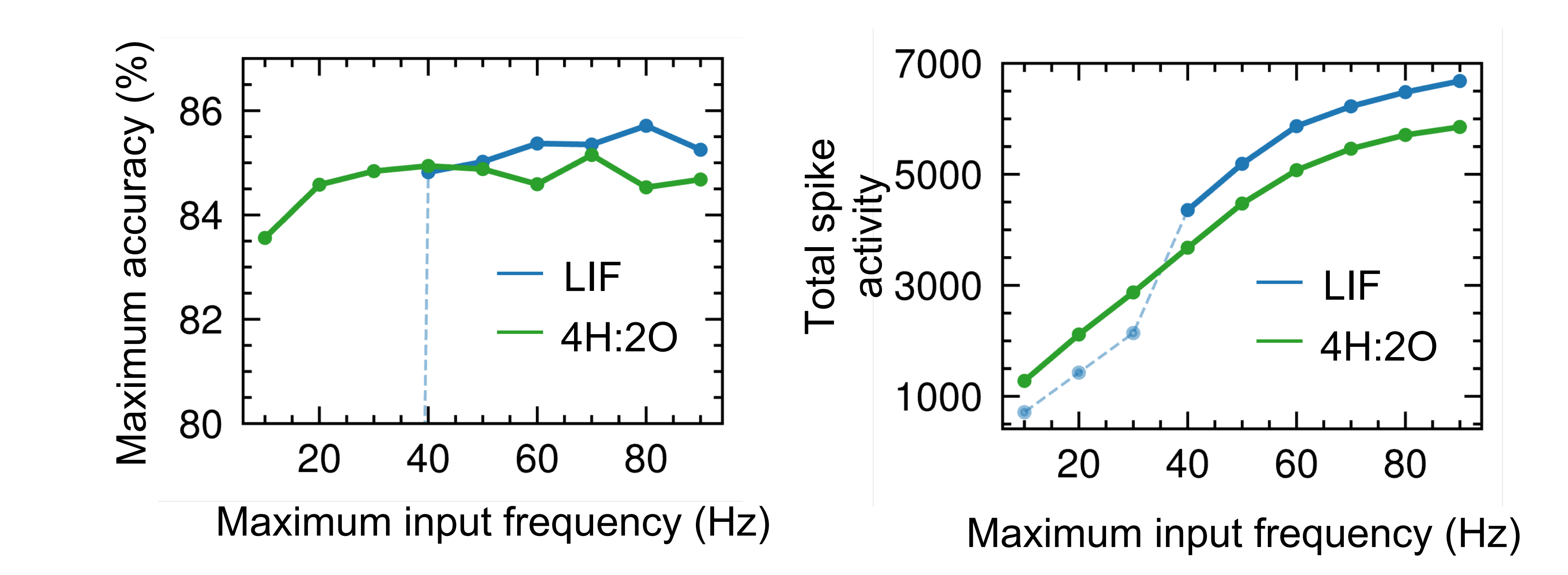
- Maximum accuracy after 20 epochs, with maximum input spiking frequency at 50 Hz



- LIF and 2H:2O have accuracy drop-off below a certain sampling time per image
- 4H:2O can be trained for full range of sampling time per image, showing enhanced stability even under short sampling times
- For the successfully trained sampling times per image, dendritic unit shows 9-15% reduced spiking activity, and therefore higher energy efficiency, compared to LIF

Online learning performance vs. maximum input frequency

- Maximum accuracy after 20 epochs, with sampling time/image at 200 ms



- Dendritic network can maintain accuracy at low spiking frequencies compared to LIF
- For the successfully trained spiking frequencies, dendritic network shows 15% lower spiking activity (higher energy efficiency) than LIF

Conclusions

1. 2D materials have rich physics and have both silicon-compatible and bio-compatible applications

2. Graphene/Nafion transistors are ionically-driven neuromorphic devices that show low energy per weight update and are made from bio-compatible materials

3. With a single gate, they operate as artificial synapses with good linearity and symmetry

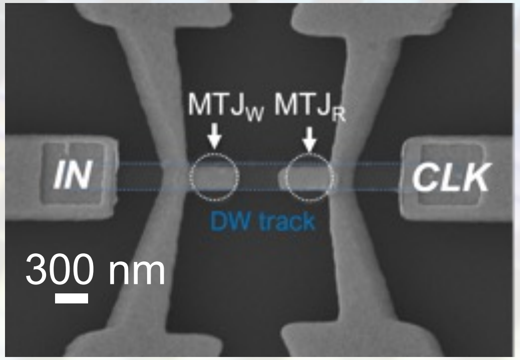
4. With a dual gate, they operate as artificial dendrites with tunable response function to input spikes

5. Neural network simulations show the graphene dendrites have significantly higher training stability at low power operation and lower spiking activity compared to an LIF benchmark

Additional projects: see students and collaborators

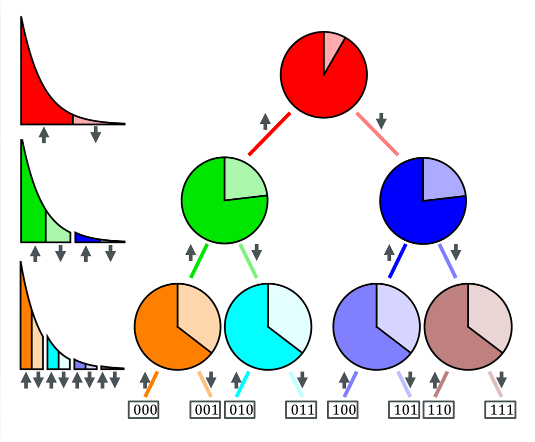
Can Cui

Neuromorphic Computing with Magnetic Domain Walls and Skyrmions



Dr. Suma Cardwell
Sandia National Labs

Tunable Random Number Generators using Magnetic Tunnel Junctions



Conclusions

1. 2D materials have rich physics and have both silicon-compatible and bio-compatible applications

2. Graphene/Nafion transistors are ionically-driven neuromorphic devices that show low energy per weight update and are made from bio-compatible materials

3. With a single gate, they operate as artificial synapses with good linearity and symmetry

4. With a dual gate, they operate as artificial dendrites with tunable response function to input spikes

5. Neural network simulations show the graphene dendrites have significantly higher training stability at low power operation and lower spiking activity compared to an LIF benchmark