Spike-based Analog Computing

with BrainScaleS

Johannes Schemmel

Electronic Vision(s) Group
Kirchhoff Institute for Physics
Heidelberg University, Germany

Electronic Vision(s)

Kirchhoff Institute for Physics, Heidelberg University

Founded 1995 by Prof. Karlheinz Meier (†2018)

1995 HDR vision sensors

1996 analog image processing

2000 Perceptron based analog neural networks: EVOOPT and HAGEN

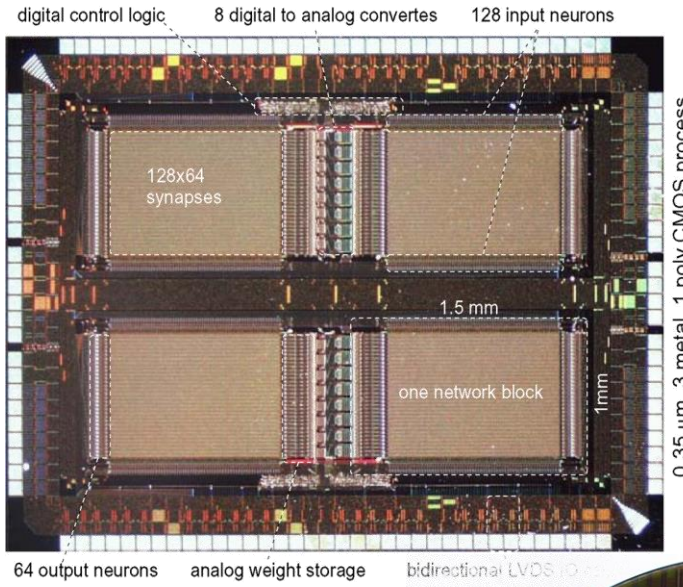
2003 first concepts for spike based analog neural networks

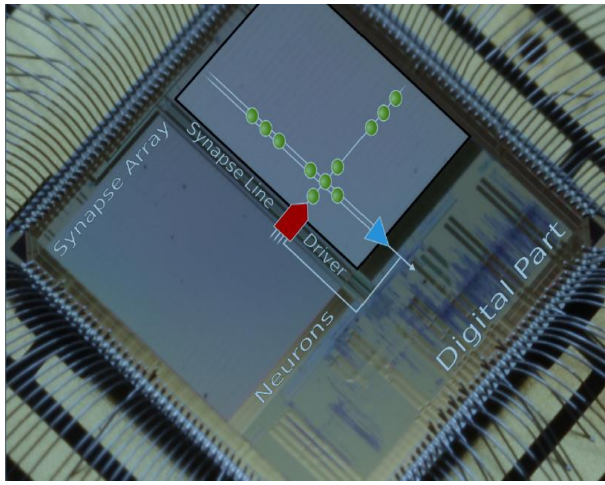
2004 first accelerated analog neural network chip with short and long term plasticity: Spikey

2010 first 20cm wafer-scale neuromorphic system: BrainScaleS

HAGEN: Perceptron-based neuromorphic chip introducing:

- accelerated operation
- mixed-signal Kernels



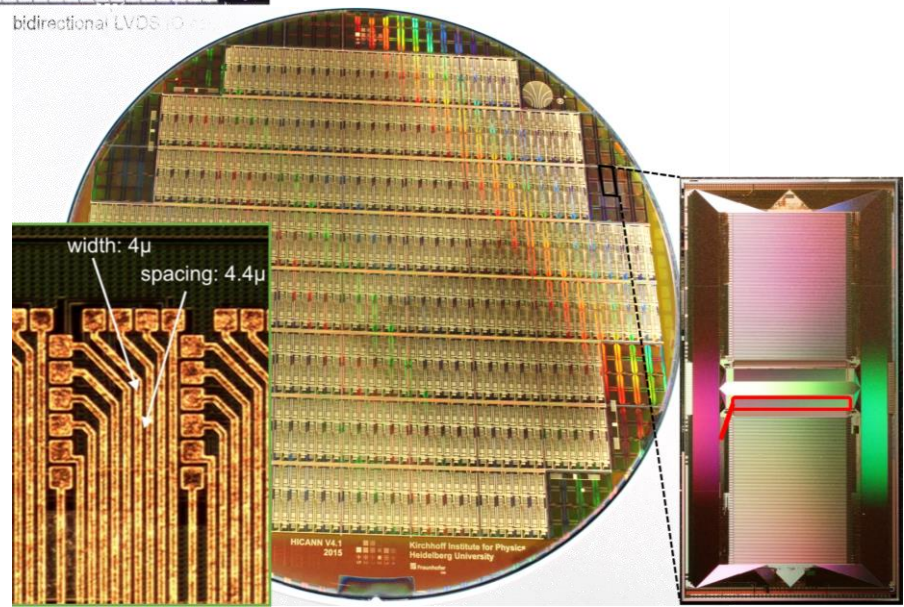


Spikey: spike-based neuromorphic chip introducing:

- Spike-Time-Dependent-Plasticity
- analog parameter storage for calibratable physical model

BrainScaleS: wafer-scale neuromorphic system introducing:

- high synaptic fan-in up-to 14k
- wafer-scale spike communication



Computers are becoming more brain-like : **brain-inspired computing**

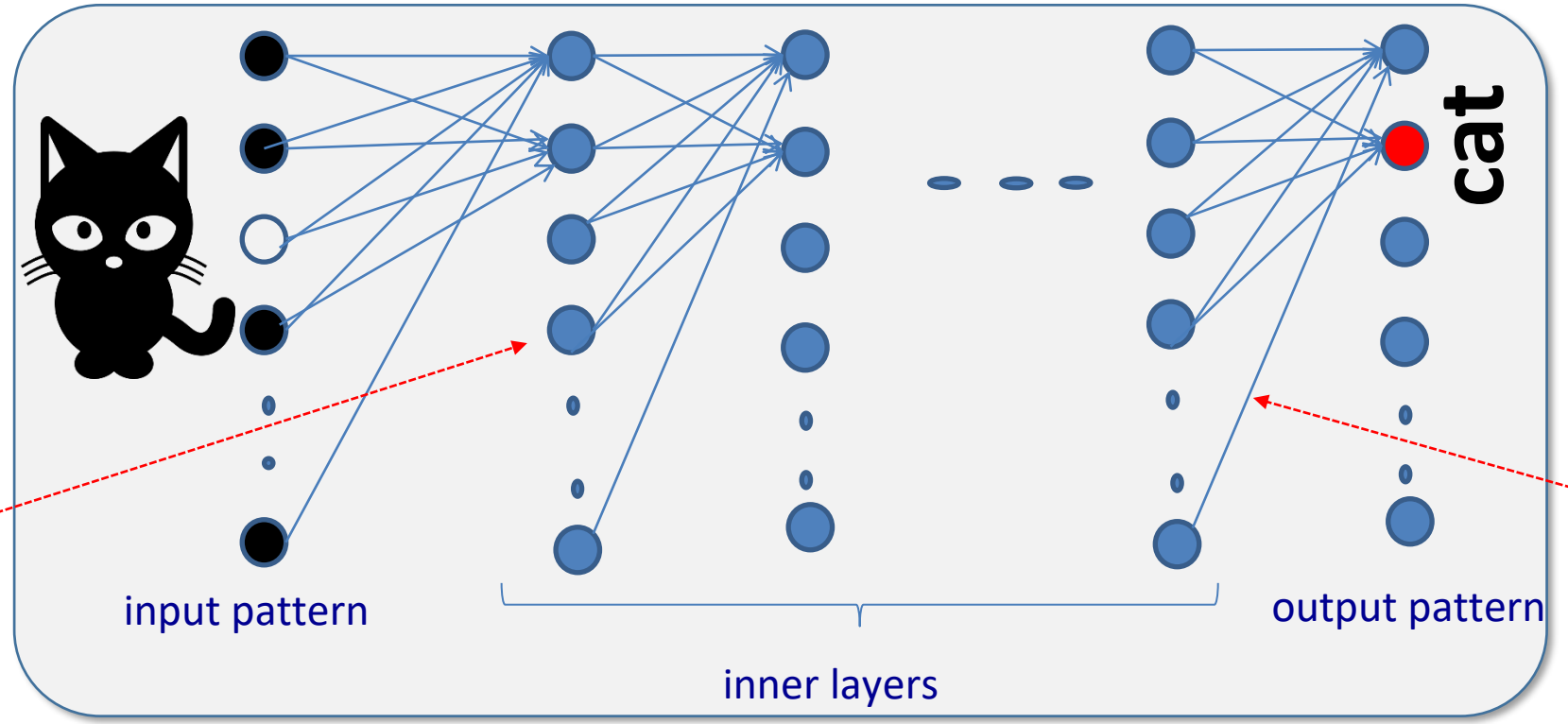
- one year training
- energy consumption: 500 kW →182500 kWh (36500 €)

Brain Inspired Computing :

overcoming the Von Neumann bottleneck using artificial neural networks

conceptually, no separation between memory content (weights) and compute units (neurons) :

→ in-memory computing



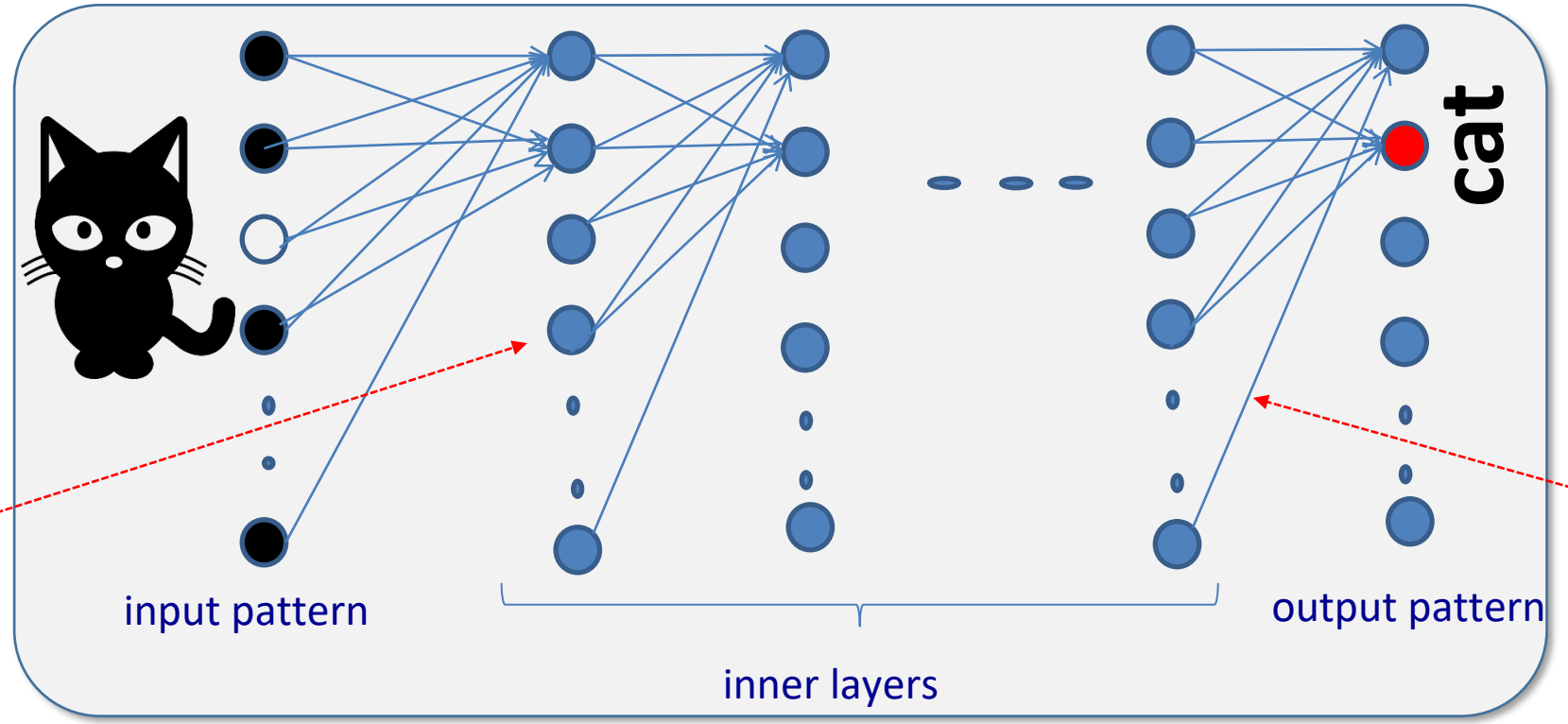
compute units :
“neurons“

“synaptic“
connections :
weights ω

Supervised training possible using stochastic gradient descent

- stochastic gradient descent minimizes the loss → difference between ground truth („cat“) and network output („maybe a cat“)

  with regards to training data set → set of different animal images
- automatic gradient calculation possible using backprop algorithm
- gradient estimation techniques exist for event(spike)-based networks



compute units : “neurons“

“synaptic“ connections : weights ω

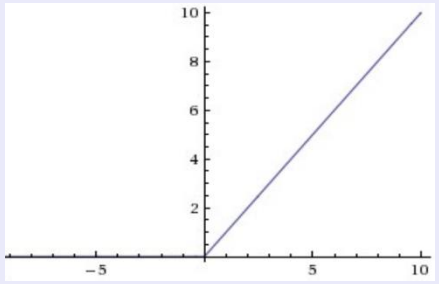
Perceptron model (biology of 1950)

Machine Learning

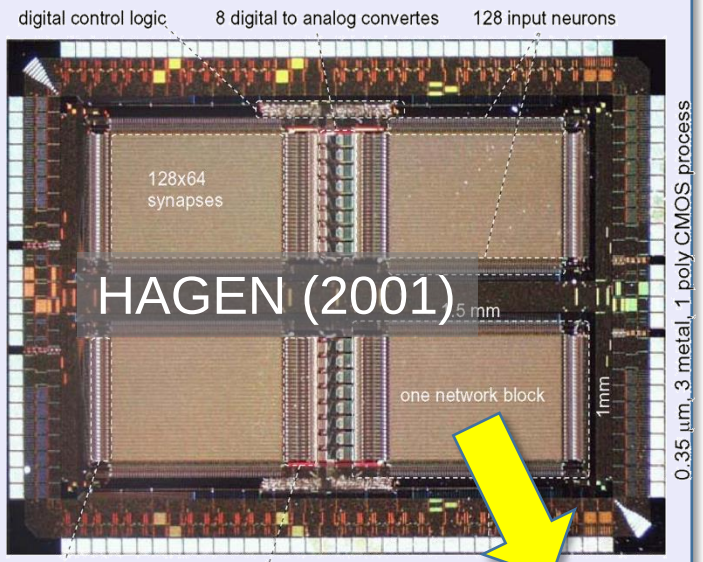
- vector-matrix multiplication

$$f\left(\sum_i w_i x_i + b\right)$$

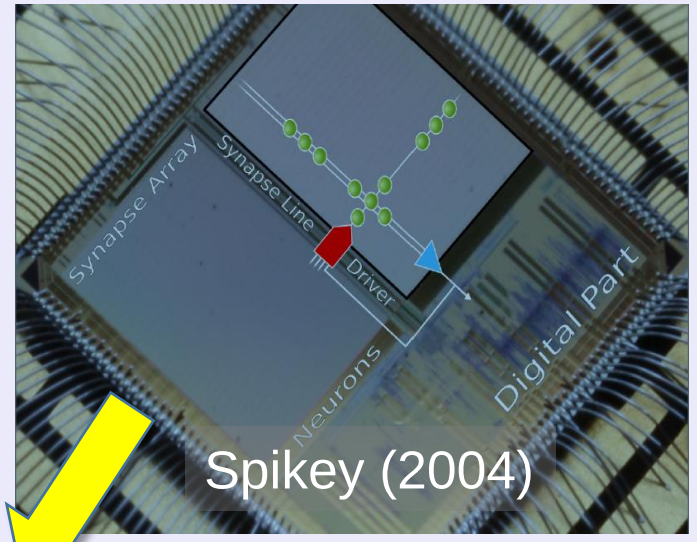
- simple non-linear activation function f (i.e. ReLU):



- Stochastic gradient descent using backpropagation

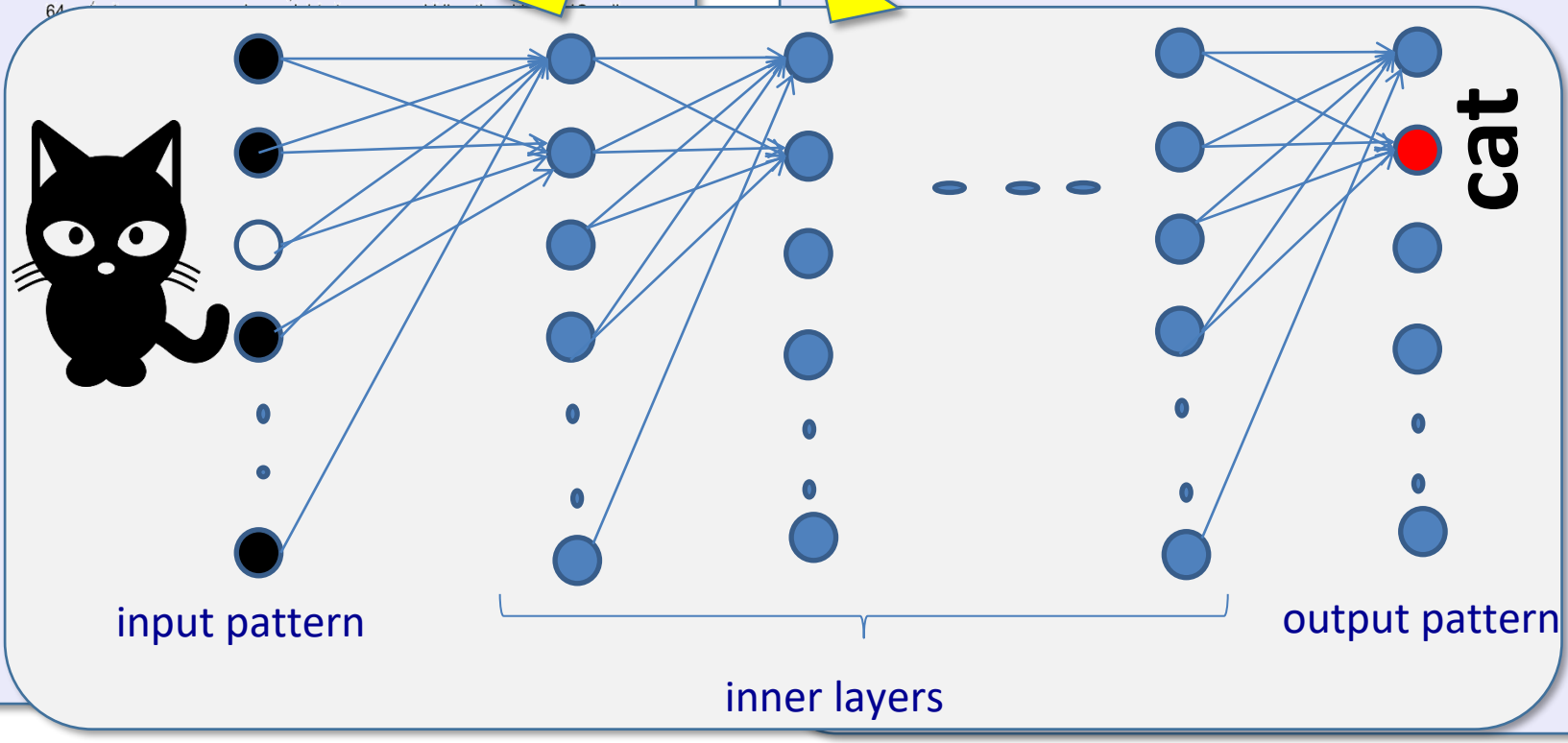


Spike-based model (current biology)

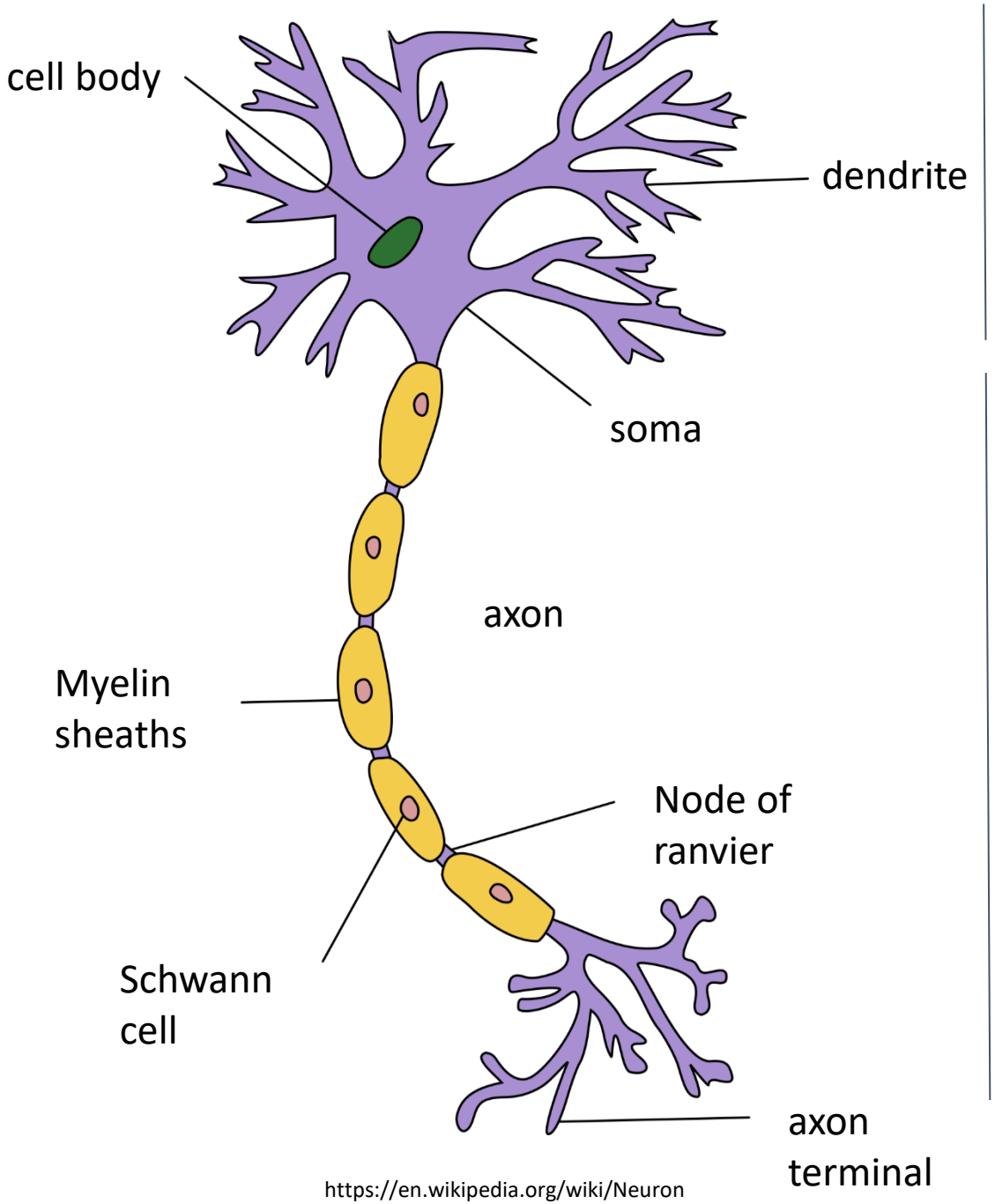


Neuromorphic Computing

- time-continuous dynamical system
- always recurrent
- vector-matrix multiplication
- complex non-linearities
- binary neuron output
- allows to model biological learning mechanisms
- can be trained with backprop-through-time

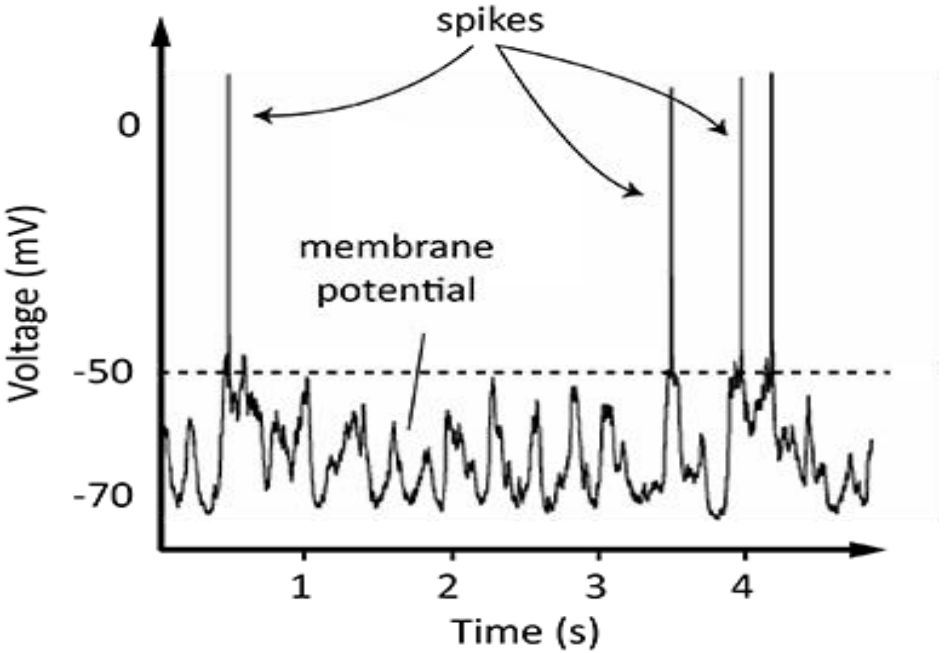


Basic neuron structure



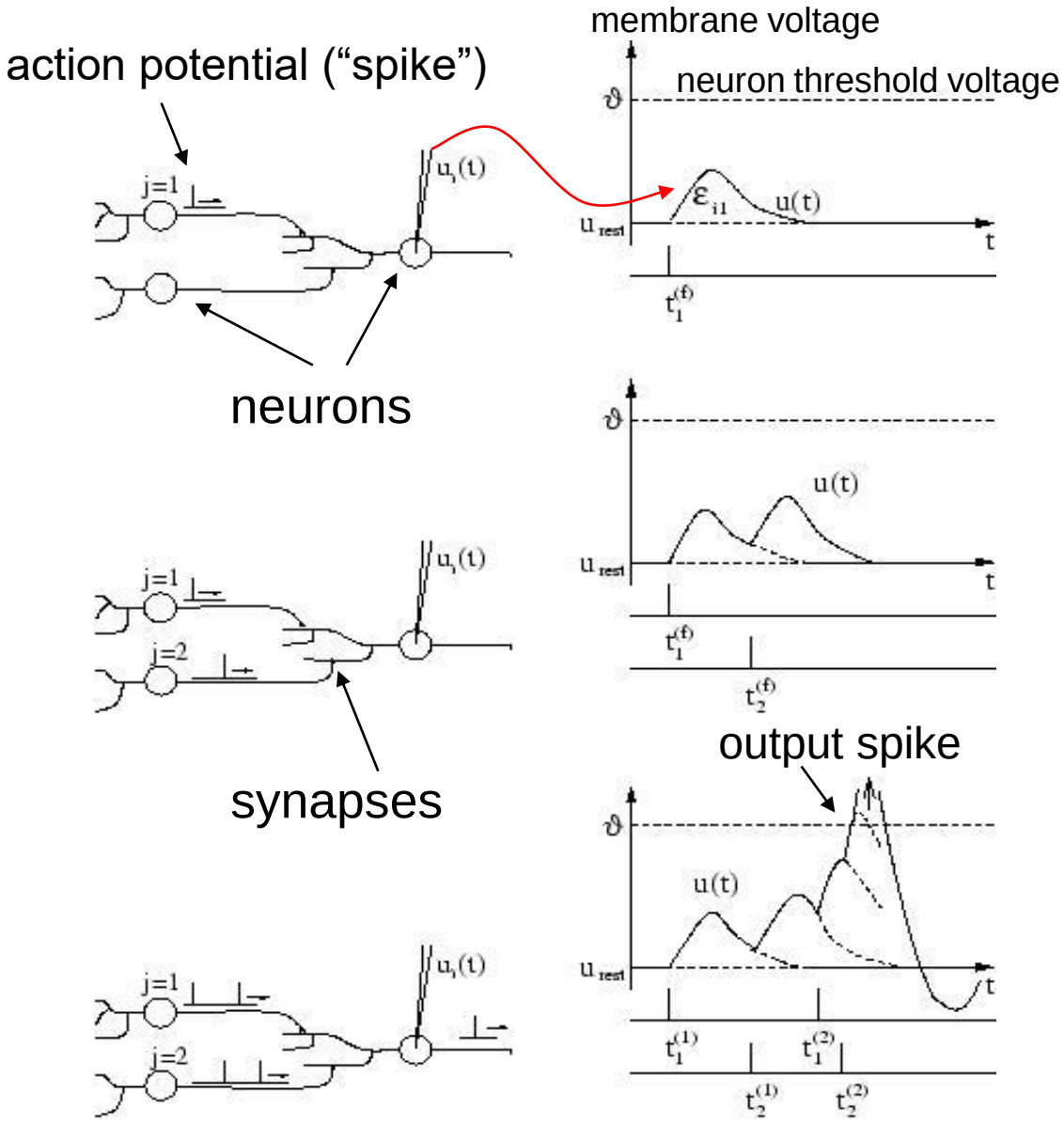
https://en.wikipedia.org/wiki/Neuron

- high fan-in and fan-out :
  10k to > 100k
- axon length can be > 1m
- complex internal state
  → reduced to membrane voltage of soma for simple neuron models



https://qbi.uq.edu.au/brain-basics/brain/brain-physiology/action-potentials-and-synapses

Principles of spike-based neural communication



- neurons integrate over space and time
- temporal correlation is important
- mixed-signal system: action potential (digital) ↔ membrane voltage (analog)
- fault tolerant
- low power consumption → 100 Billion neurons: 20 Watts

Dimensions of neuromorphic computing

analog NMC in Heidelberg

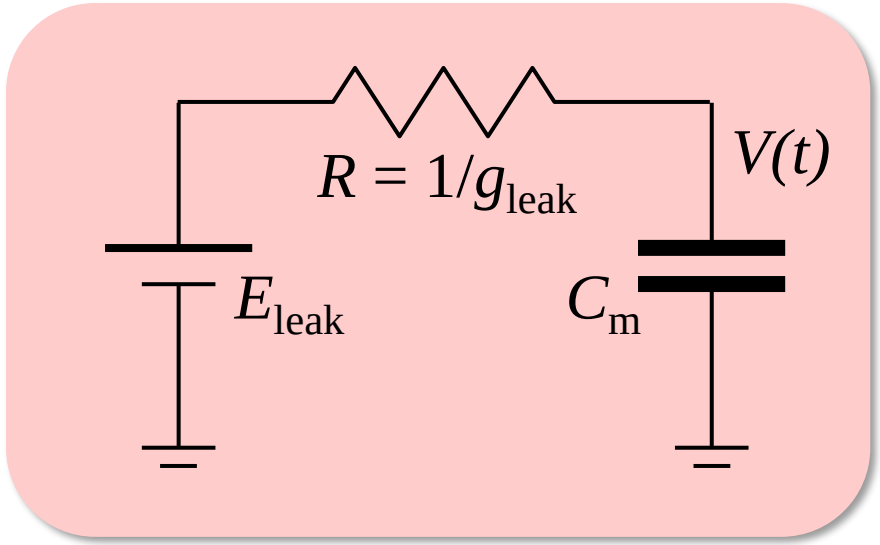
conceptual dimensions :	event based	←→ rate based
	continuous time, c. valued	←→ discrete time, d. valued
	approximate, noisy, stochastic comp.	←→ exact computations
	structured neurons	←→ point neurons
	non-linear dentrides	←→ linear dentrides
	plastic	←→ static
technological dimensions :	analog	←→ digital
	electrical	←→ optical
	standard CMOS	←→ novel devices
	fully programmable	←→ fixed structure
	in-memory computing	←→ von Neumann computing
	constant speed	←→ variable speed
	(real time or accelerated)	(best effort)
application dimensions :	research	←→ commercial
	energy, size, cost constrained	←→ energy, size, cost agnostic
	brain emulation	←→ machine learning, AI
	needs to adapt	←→ fixed function

Fundamental architecture for analog neuromorphic computing :

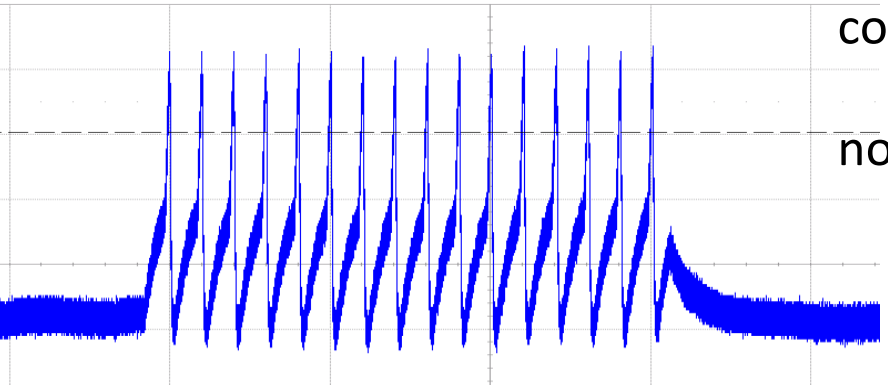
- Consider a simple physical model for the neuron's cell membrane potential V:

$$C_{\mathrm{m}} \frac{dV}{dt} = g_{\mathrm{leak}} \left(E_{\mathrm{leak}} - V \right)$$

- representing model parameters as physical quantities : **voltage, current, charge**



$$\frac{dV}{dt}_{bio} \ll \frac{dV}{dt}_{VLSI} \rightarrow \text{accelerated neuron model}$$

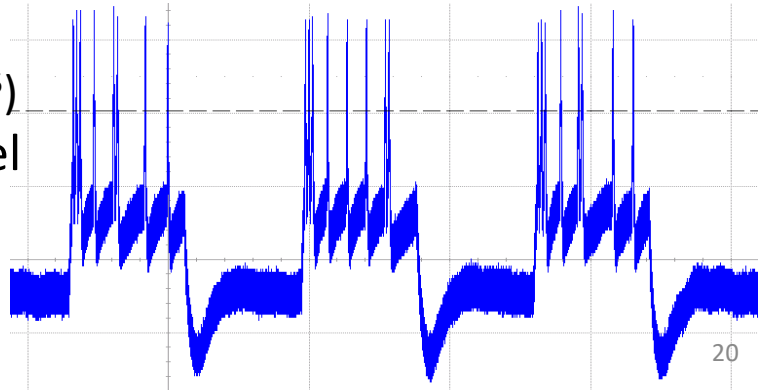


continuous time

- fixed acceleration factor (we use 10^3 to 10^5)

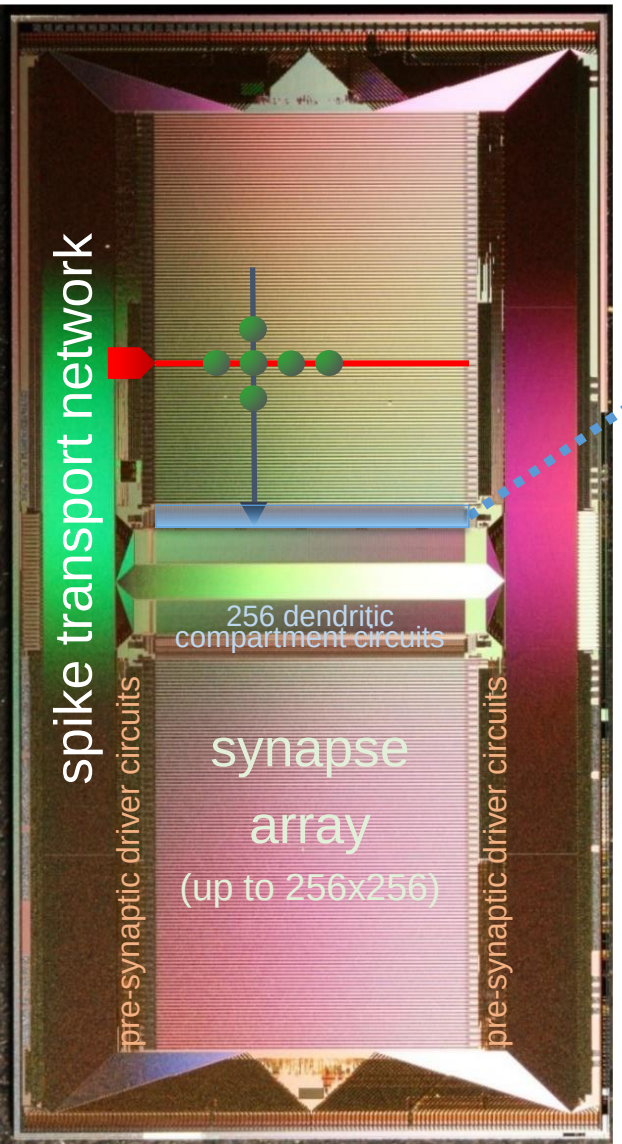
no multiplexing of components storing model variables

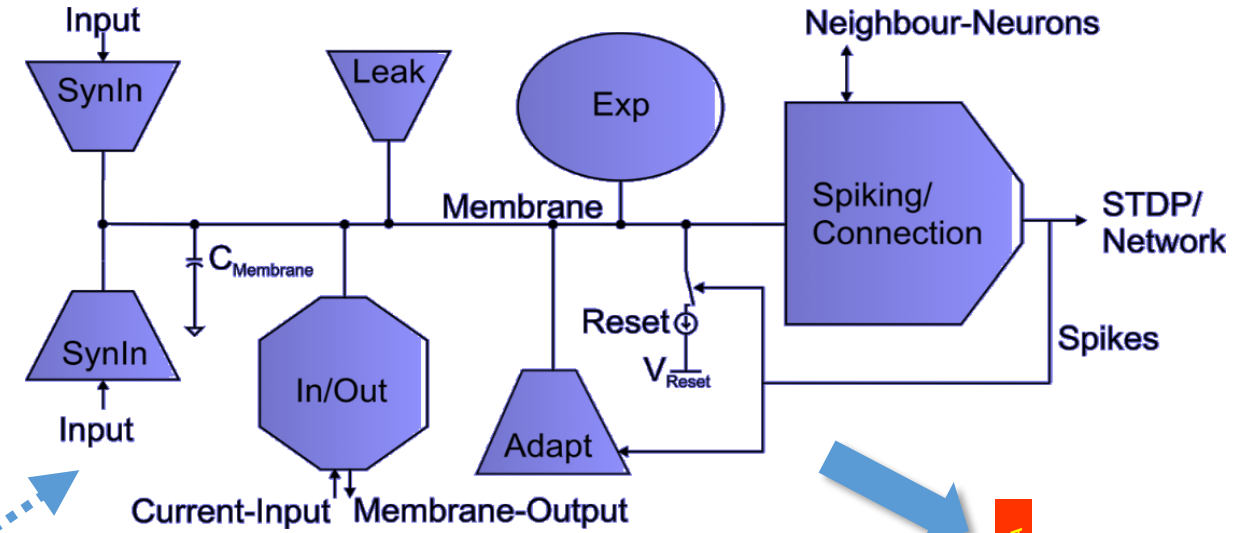
- each neuron has its membrane capacitor
- each synapse has a physical realization



Realization of a physical neural network : BrainScaleS spiking neurons built from parameterized dendritic compartments

photograph of the BrainScaleS 1 neuromorphic chip

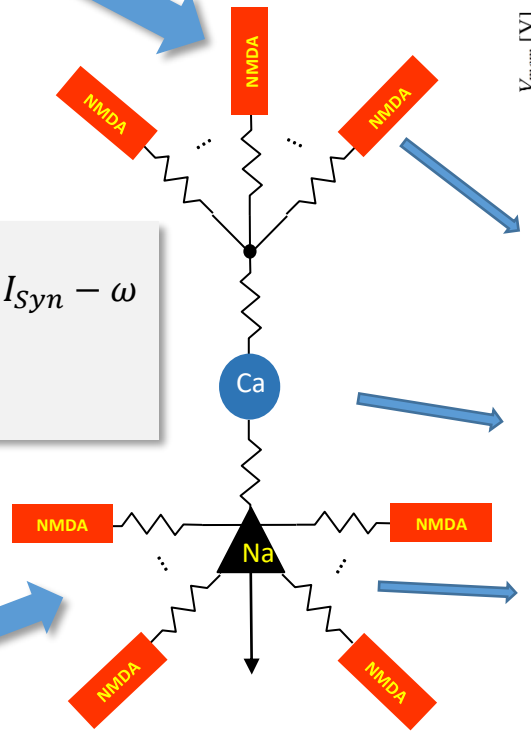


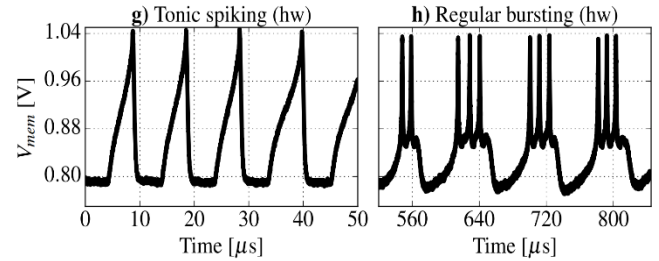


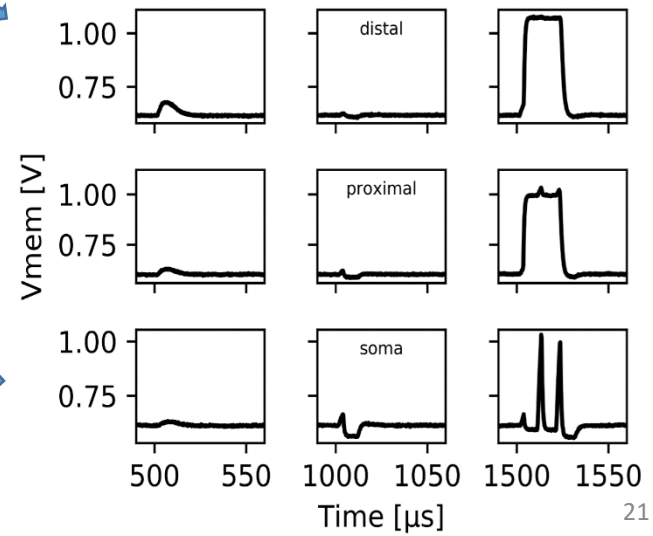
- modular structure
- Adaptive Exponential I&F model
- full set of ion-channel circuits for each compartment
- 24 calibration parameters per compartment

$$C\frac{dV}{dt} = -g_L(V - E_L) + g_L\Delta_T \exp\left(\frac{V - V_T}{\Delta_T}\right) + I_{Syn} - \omega$$

$$\tau_\omega \frac{d\omega}{dt} = a(V - E_L) - \omega$$

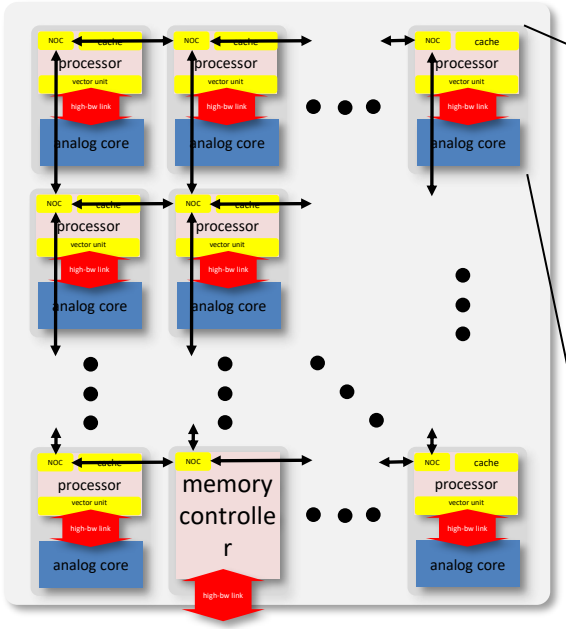




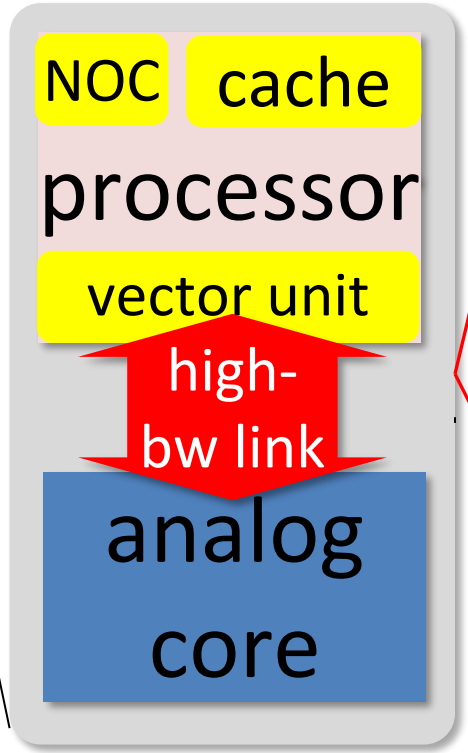


complex neurons can be build by connecting individual compartments

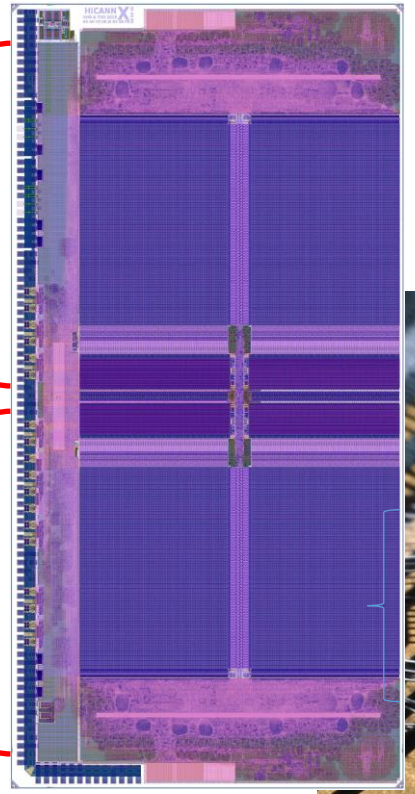
BrainScaleS architecture: analog neuromorphic core as coprocessor



- tile-based
- each tile combines analog and digital parts



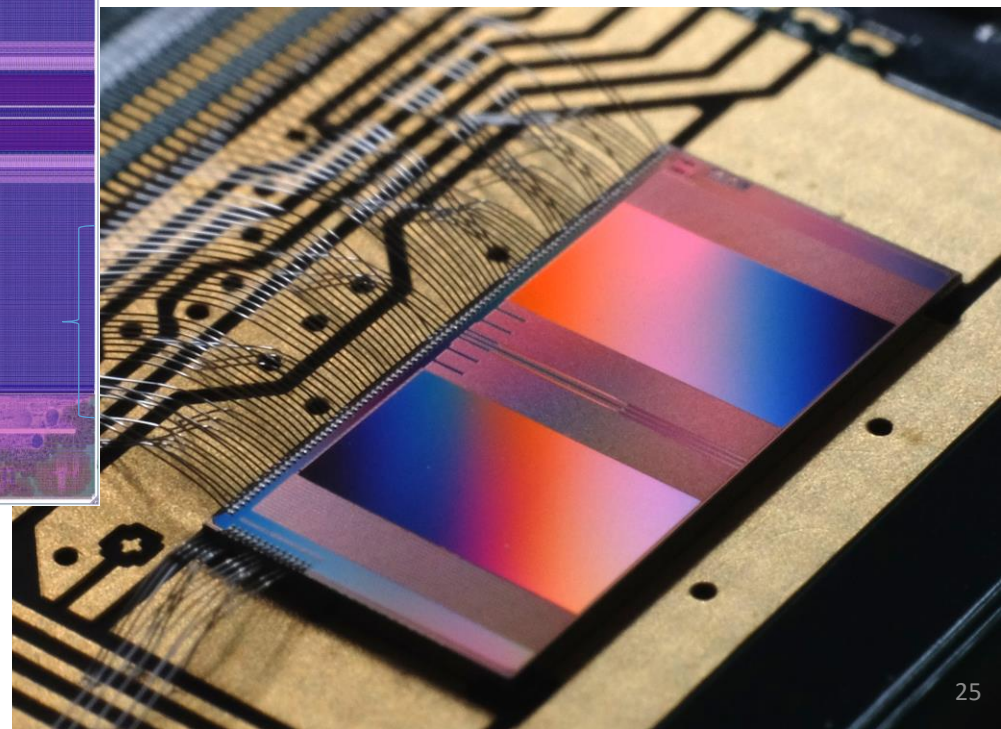
current two-tile ASIC:



4x8 mm^2

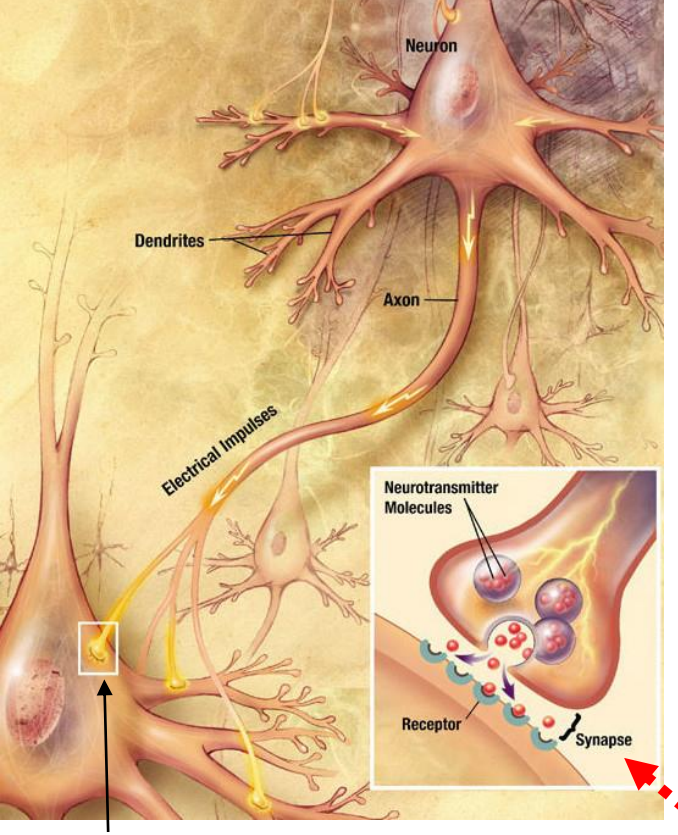
65nm CMOS

- on-chip training with complex learning rules
- learning capabilities scale with system size
- can cope with scaled-up speed of accelerated physical model

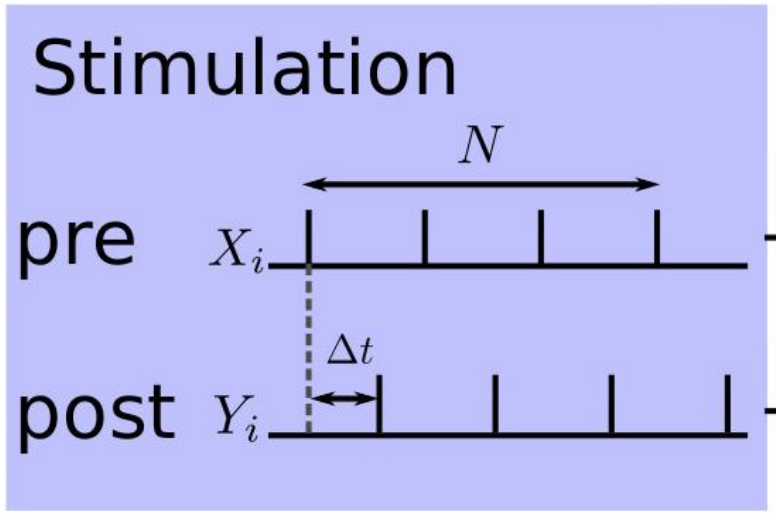


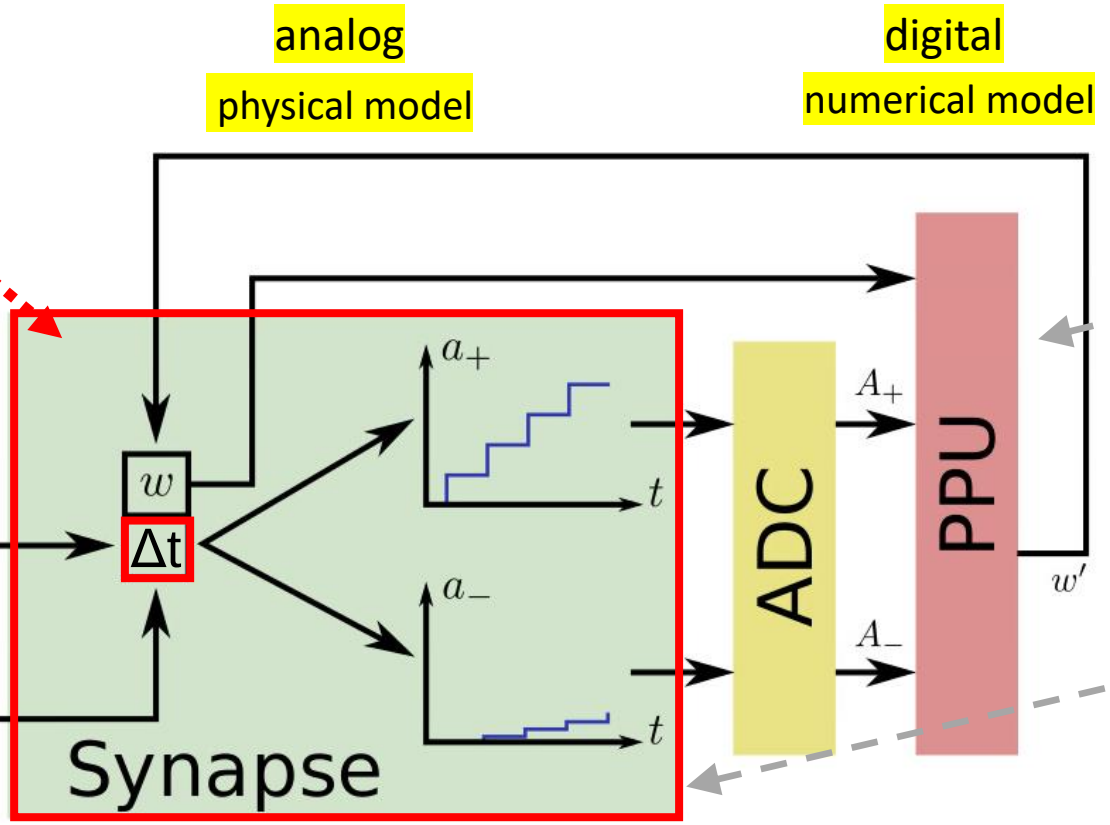
BrainScaleS-2: Hybrid Plasticity

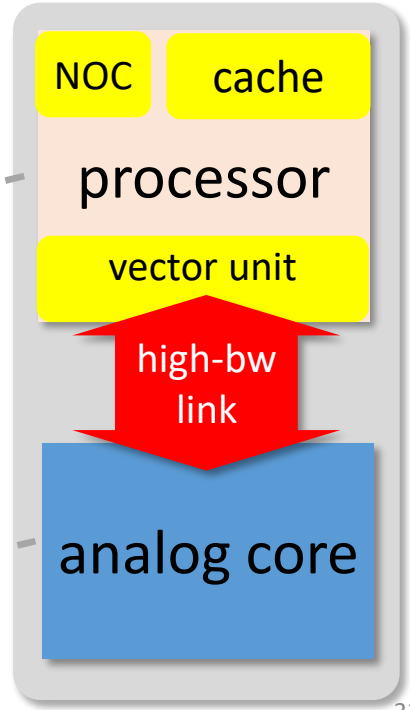
- analog correlation measurement in synapses
- A/D conversion by parallel ADC
- digital Plasticity Processing Units can access
  - synaptic weights (ω)
  - pre-synaptic connection → structural plasticity
  - neuron voltages and firing rates



plasticity takes place at the synapse





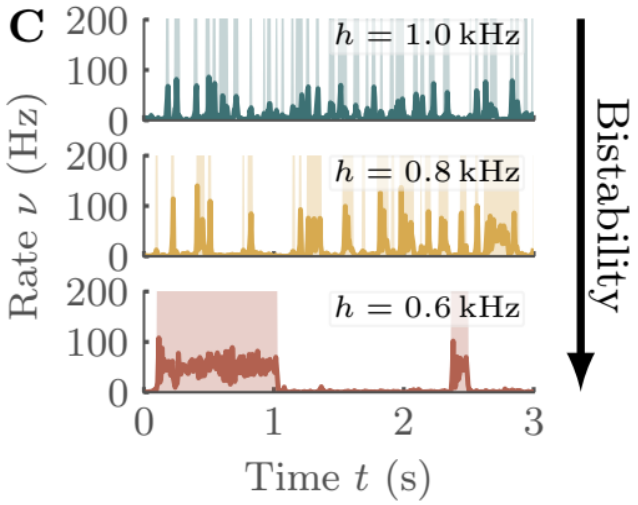


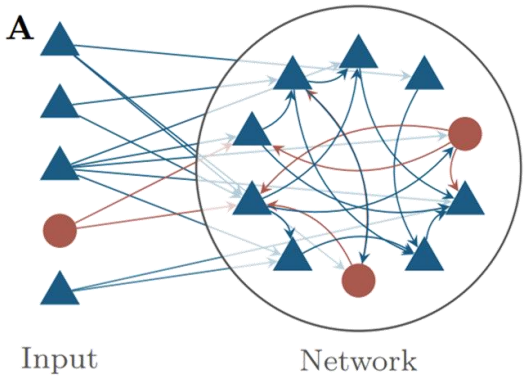
BrainScaleS is a substrate for different neuromorphic algorithms

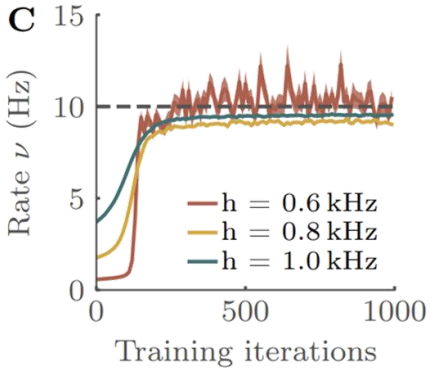
- accelerated emulation of networks of structured neurons with non-linear dendrites
  (Emulating dendritic computing paradigms on analog neuromorphic hardware, Jakob Kaiser et.al., Neuroscience, 2021)
- large parameter sweeps for network operation tuning
  (Autocorrelations in homeostatic spiking neural networks as a result of emergent bistable activity, J Zierenberg et. al., Bulletin of the American Physical Society, 2022/3/14,
  Control of criticality and computation in spiking neuromorphic networks with plasticity, B Cramer et.al., Nature communications, 2020/6/5)
- biology inspired learning experiments with programmed local plasticity
  (Structural plasticity on an accelerated analog neuromorphic hardware system, S Billaudelle et. al., Neural Networks, 2021/1/1)
- learning-to-learn sweeps of meta-parameters (Neuromorphic Hardware Learns to Learn, T Bohnstingl et.al., Front Neurosci.,2019)
- inference experiments for solving tasks using optimized network parameters generated by hardware-in-the-loop gradient-based training
  (Surrogate gradients for analog neuromorphic computing, B Cramer et.al., pnas.2109194119, 2022;
  Fast and energy-efficient neuromorphic deep learning with first-spike times, J Göltz et.al, Nature machine intelligence, 2021/9)
- applications of spiking neural networks for approximate computing
  (Spiking neuromorphic chip learns entangled quantum states, S Czischek, et. al. , SciPost Physics 12 (1), 039, 2022)
- parameter fitting to match experimental observations
- direct real-time coupling between in-vitro preparations in wet-labs and the BrainScaleS system
  → initially with HeiCINN in Heidelberg, but open for others
- repeated execution of a network and/or long operation to gather statistical information or for sampling from stochastic models
- interactive execution of small models with immediate visualization for educational purposes
  → girls' day, advanced lab course
- experimental platform for analogue computing research
  (Towards Addressing Noise and Static Variations of Analog Computations Using Efficient Retraining, B Klein et.al., ECML PKDD, 2021/9/13)
  → first industry collaboration shows promising results in the area of optical communication
  ("Best Student Paper Award" for Elias Arnold at SPPcom 2022, nominated for "Best Student Paper Award" at ECOC 2022)

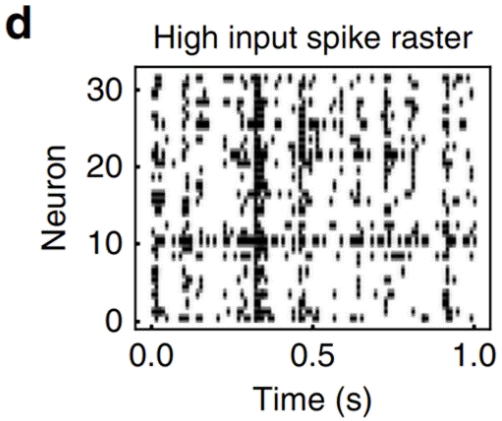
Emergent bistability in homeostatic regulated spiking networks

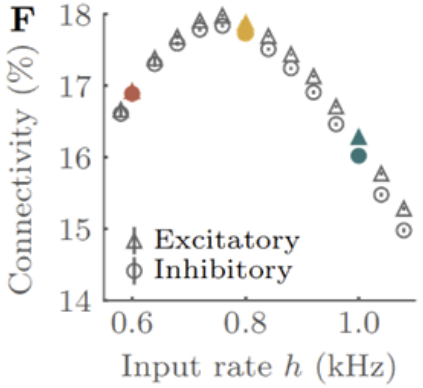
- recursive network of 512 LIF neurons, 20% inhibitory
  → hardware plasticity used to find homeostatic balance: $\Delta\omega_{ij}=\lambda(\nu^{*}- \nu_j)$
- reducing the strength of external inputs shifts network to bistable behaviour
  → alternating between high and low firing rates while mean stays stable
  → emerging bistability increases autocorrelation time
  → network compensates for a lack of input to preserve firing rate

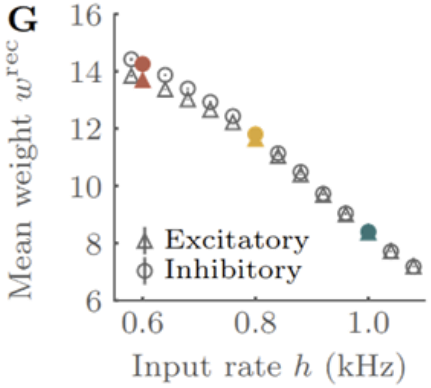


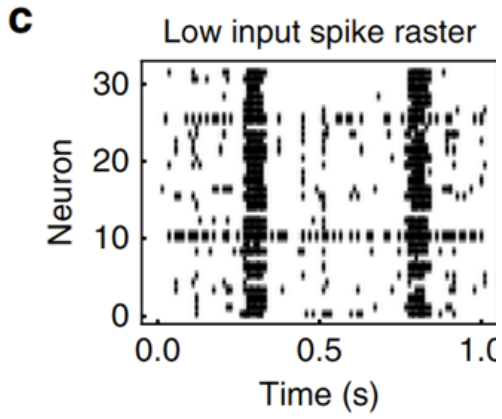








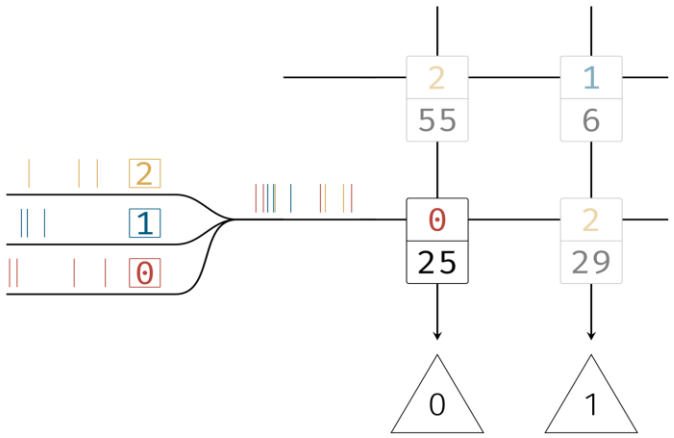


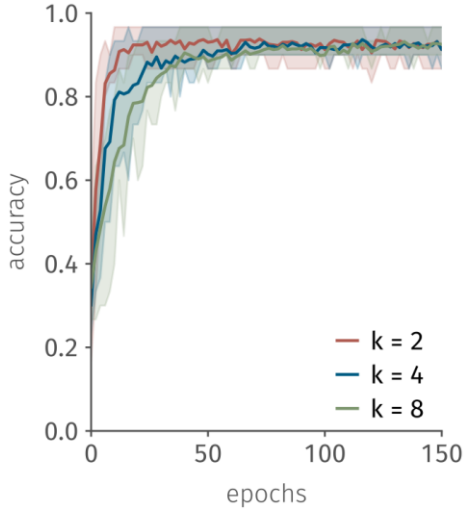


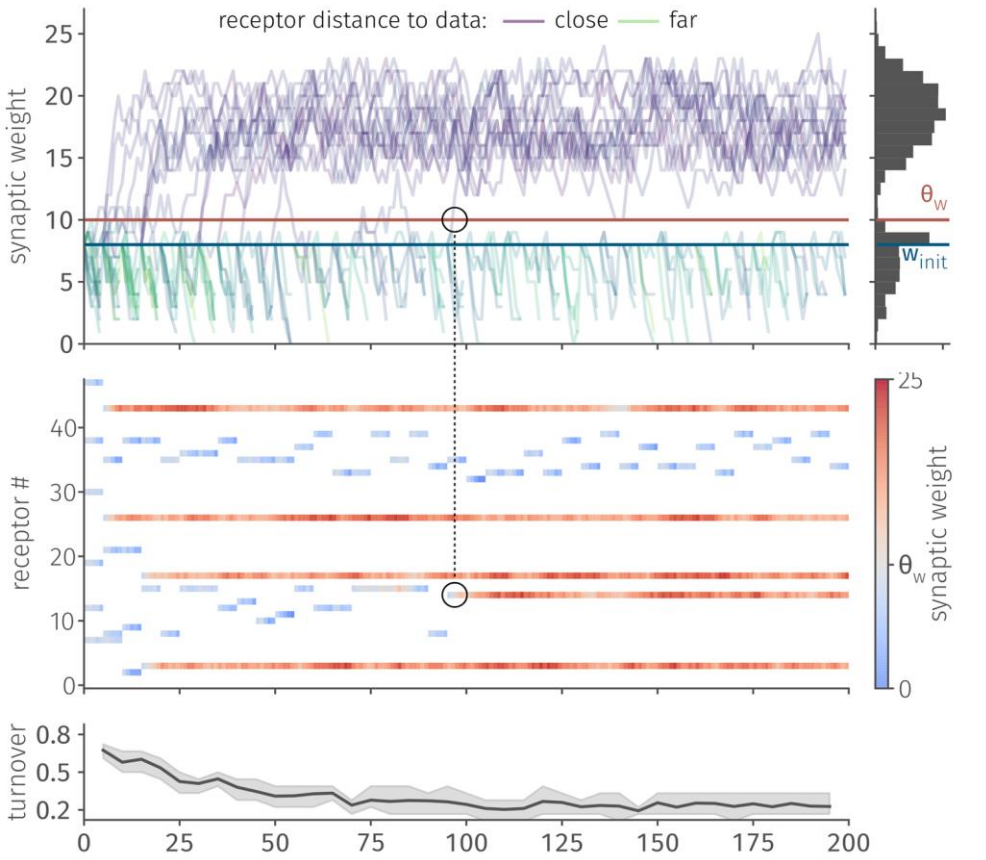
B.Cramer, …, V. Priesemann et.al. *"Control of criticality and computation in spiking neuromorphic networks with plasticity"*, Nat Commun 11, 2853 (2020). https://doi.org/10.1038/s41467-020-16548-3

B. Cramer, …, V. Priesemann and J. Zierenberg et al. *"Autocorrelations from emergent bistability in homeostatic spiking neural networks on neuromorphic hardware"* ArXiv 2208.08329v1

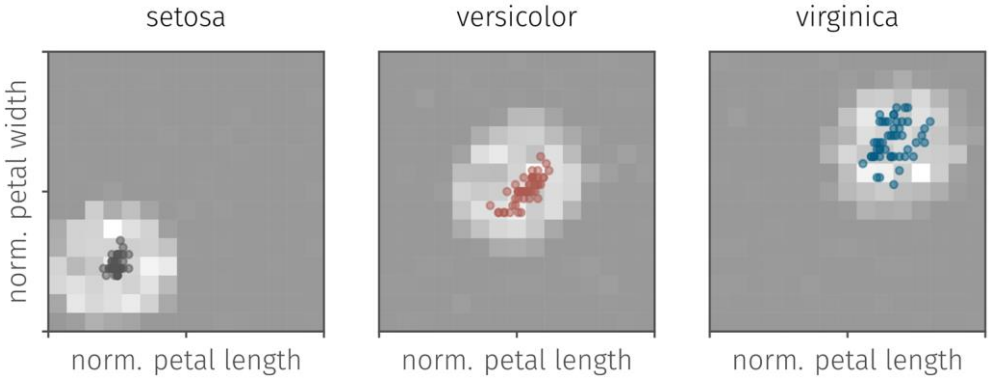
Structural plasticity on BrainScaleS-2







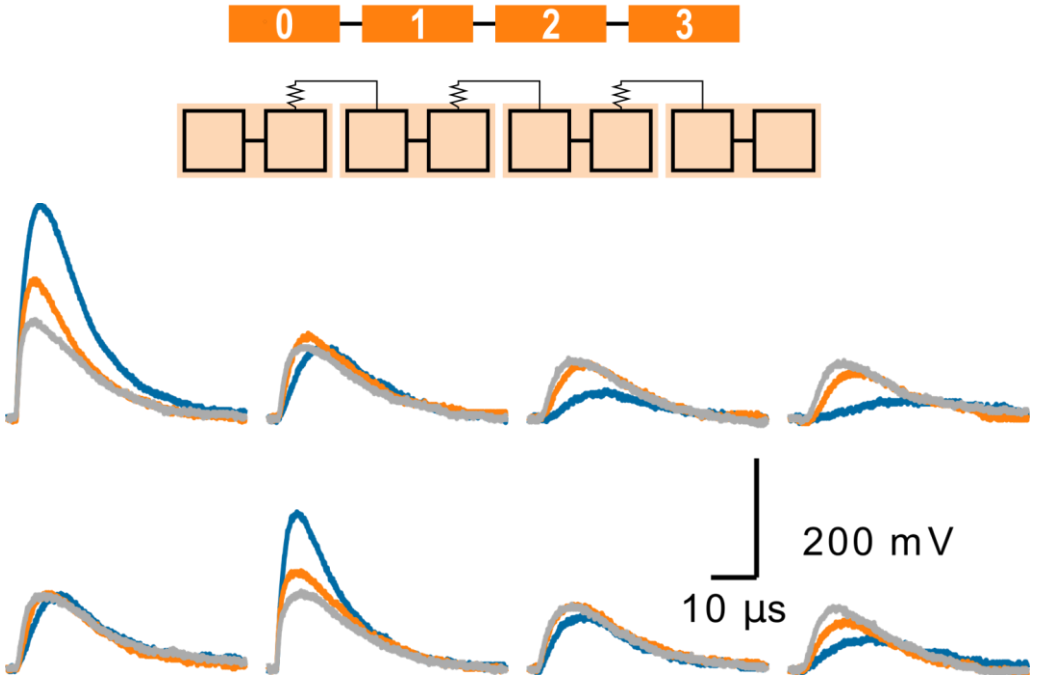
- On-chip structural plasticity
- Self-configuring receptive fields
- Efficient use of synaptic resources

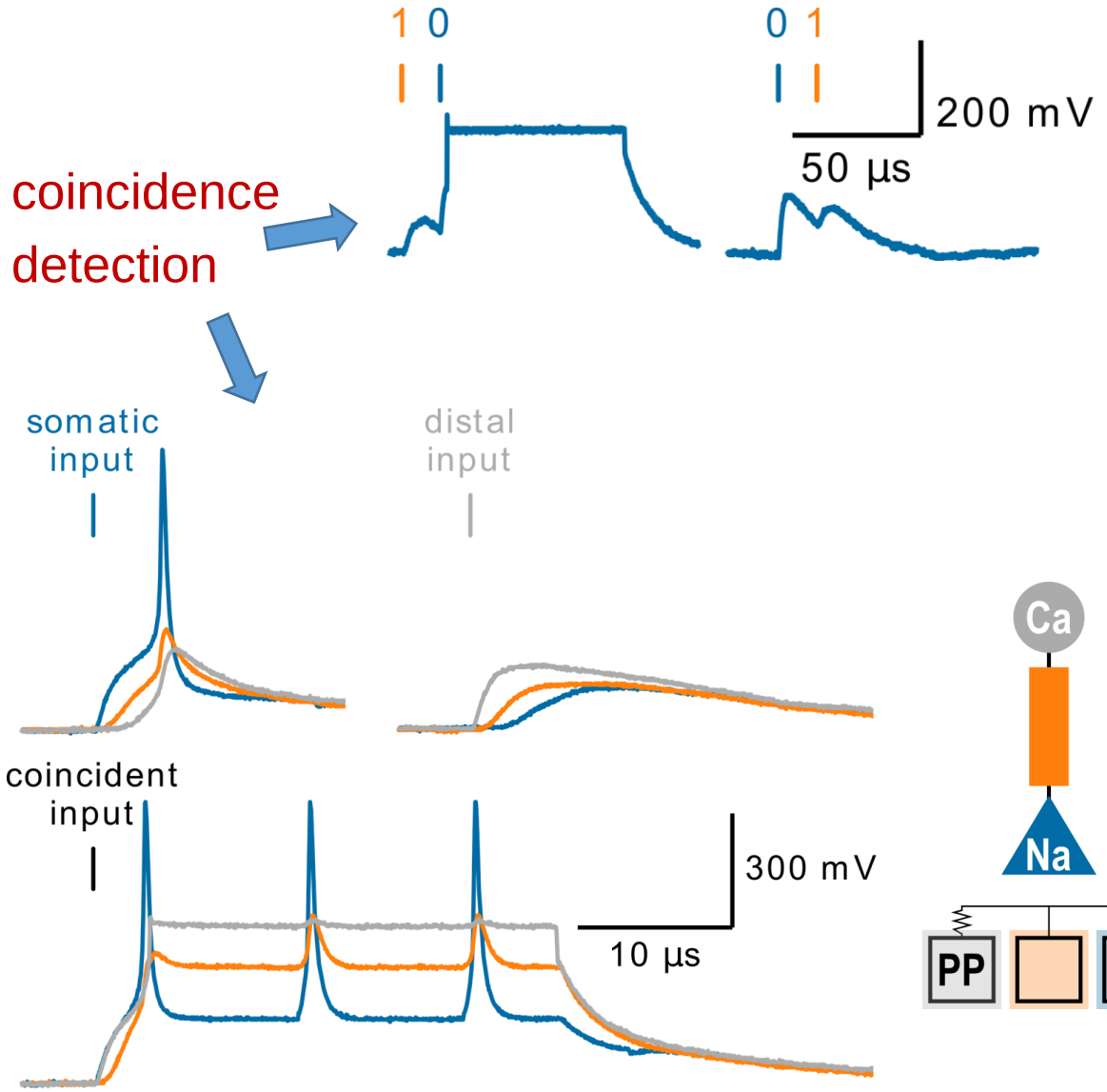


S. Billaudelle, B. Cramer, et al. "Structural plasticity on an accelerated analog neuromorphic hardware system." Neural Networks 133 (2021): 11-20.

Multi-compartment neurons on BrainScaleS-2

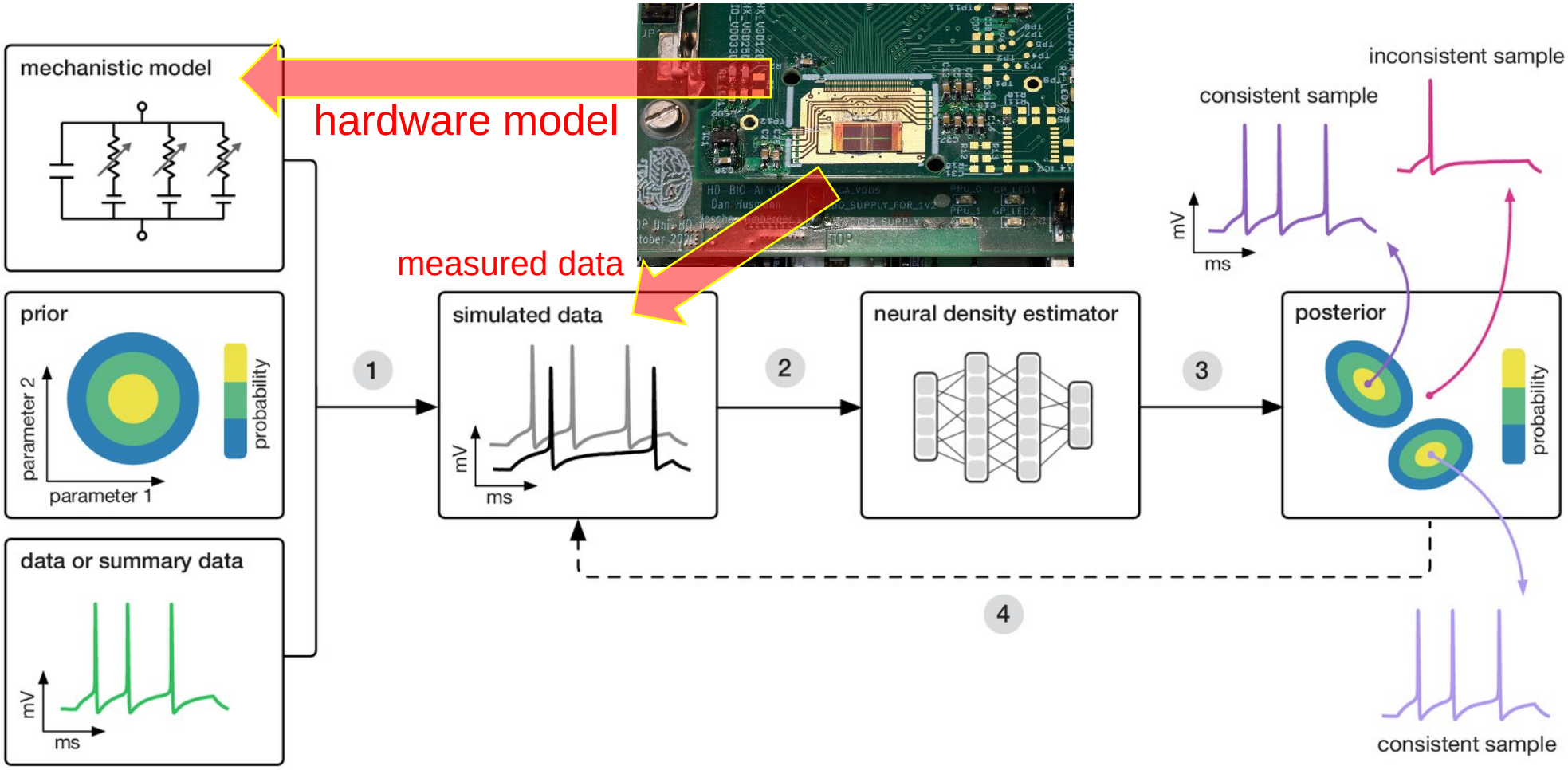
- User-defined morphologies
- Parameter adjustable for each compartment
- Dendritic spikes
- No additional energy consumption
- 1000x acceleration factor independent of neuron morphology





Emulating Dendritic Computing Paradigms on Analog Neuromorphic Hardware, J Kaiser et.al., Neuroscience, 2021

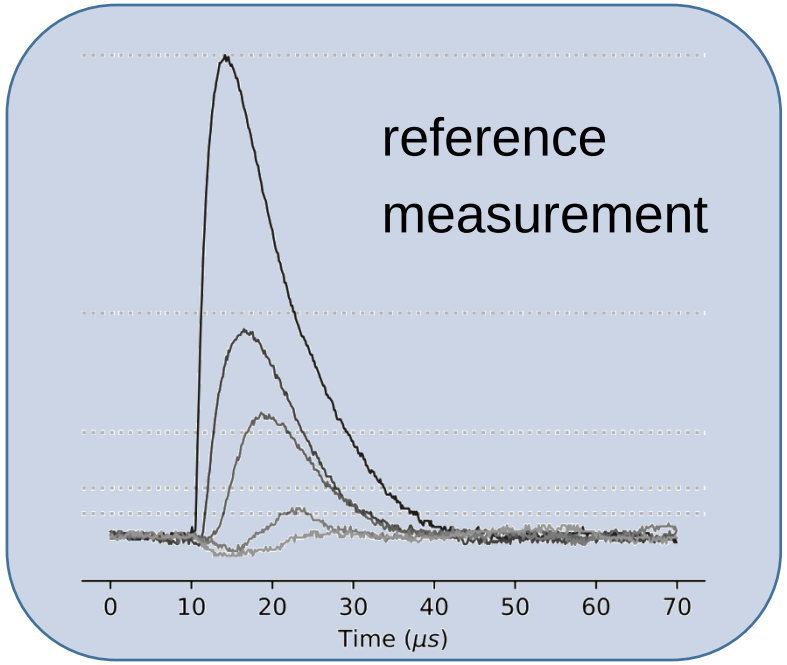
fitting BrainScaleS neurons to experimental data



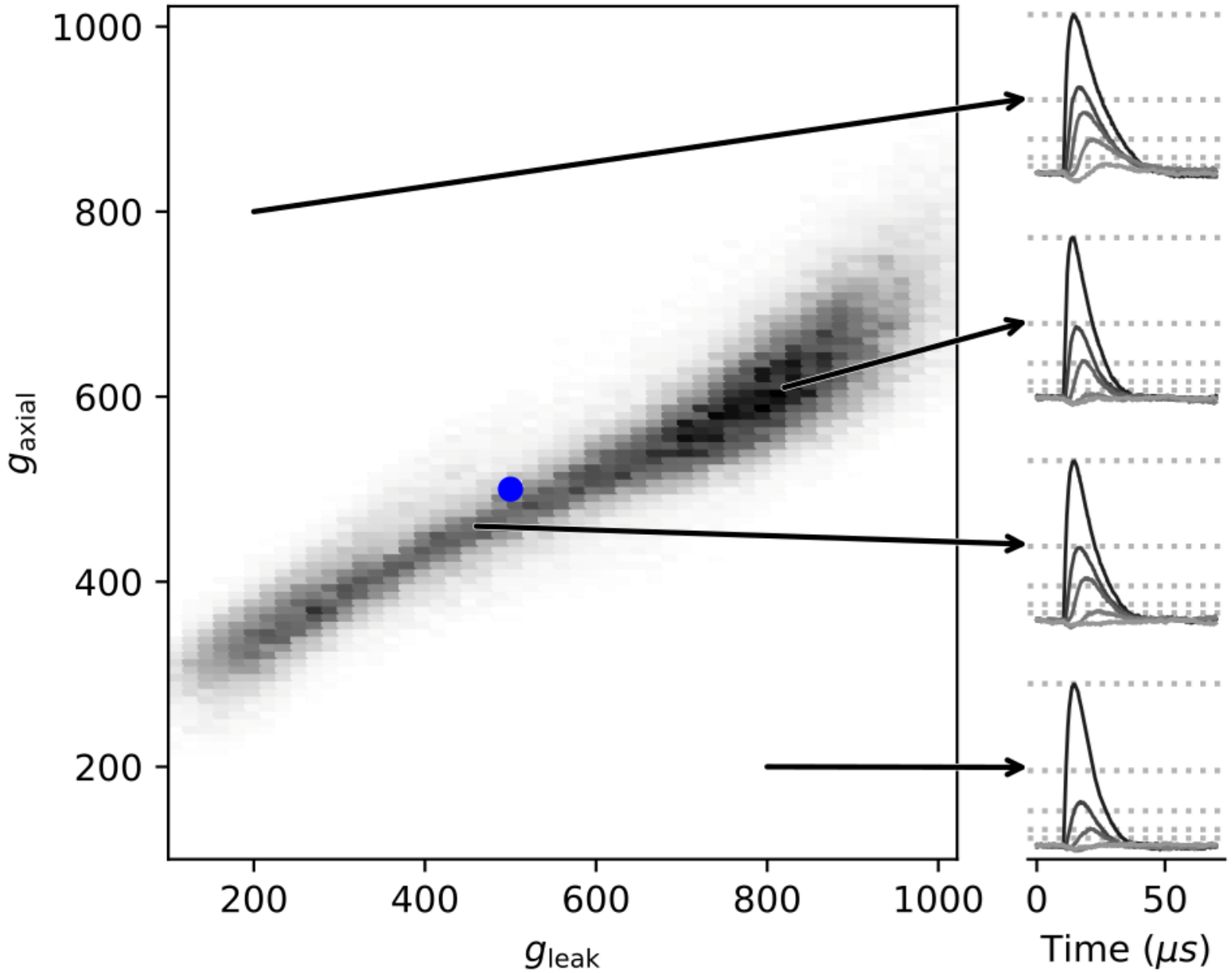
"Training deep neural density estimators to identify mechanistic models of neural dynamics", Pedro J Gonçalves et. al., eLife 2020;9:e56261 DOI: 10.7554/ELIFE.56261

early results from BSS-2 hardware

- chain of five dendritic compartments
- finding the correct parameter for leakage and inter-compartmental conductance

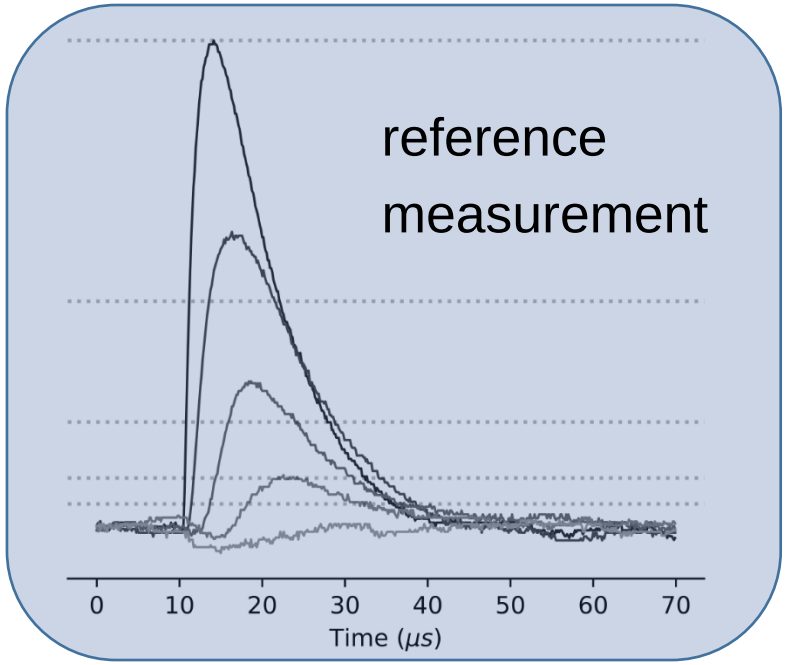


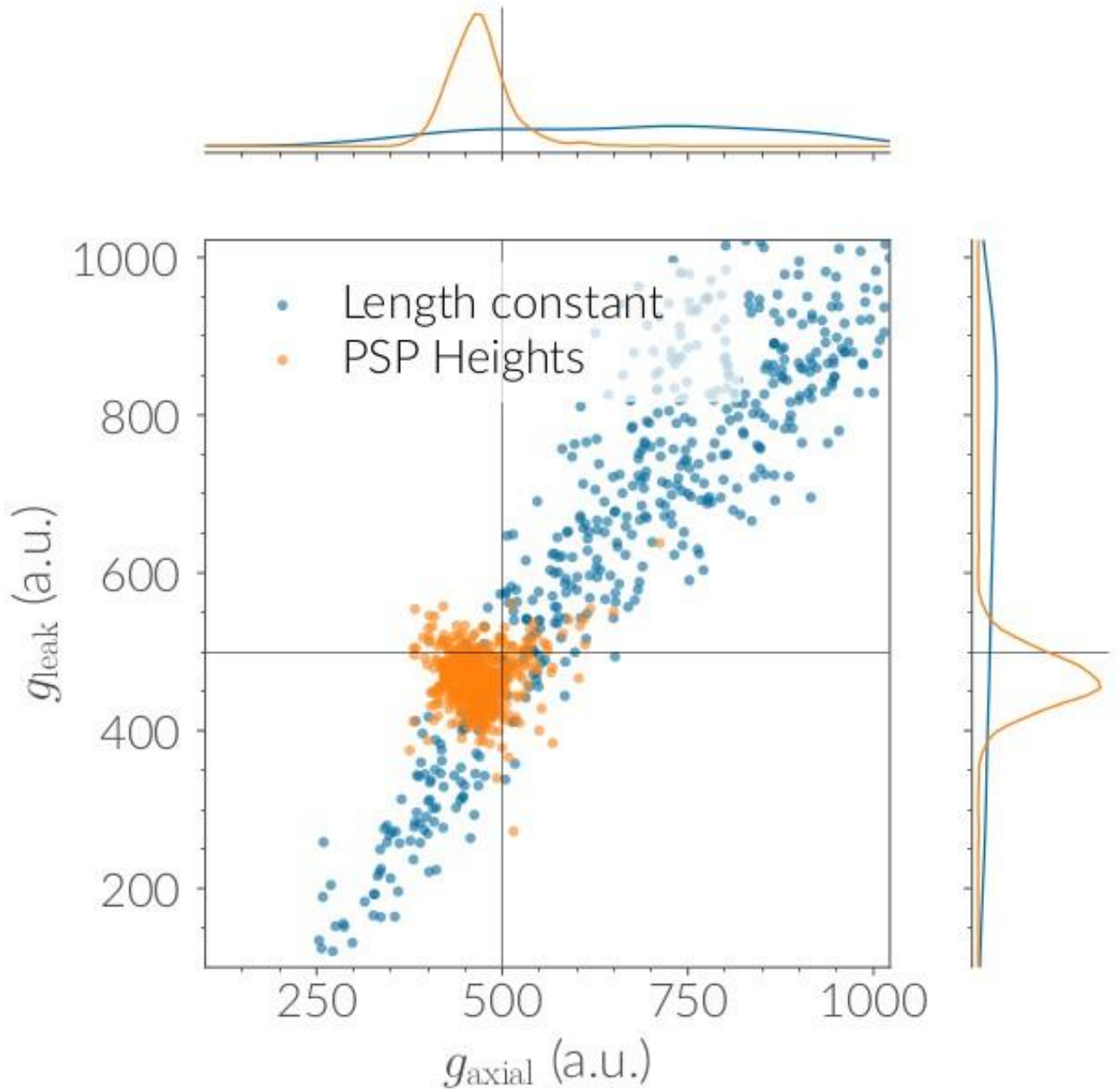
ongoing PhD thesis from Jakob Kaiser, in collaboration with Sebastian Schmitt, Tetzlaff lab, University Göttingen



fitting for absolute PSP heights

- fitting for different criteria possible
- with exact PSP heights as target, the likely values for g_{leak} and g_{axial} are clustered





ongoing PhD thesis from Jakob Kaiser, in collaboration with Sebastian Schmitt, Tetzlaff lab, University Göttingen

Summary & Outlook

- BrainScaleS-2 is a physical substrate with continuous time neuron dynamics
- supported neuromorphic algorithms :
  - deep networks (spiking and non-spiking, recurrent and feed-forward)
  - brain emulation with complex neuron models to fit experimental data
  - local learning with hardware support for a multitude of plasticity rules
    - hardware support for structural plasticity improves hardware utilization
- publicly available via the European EBRAINS service
  (not yet funded after the end of the Human Brain Project)
  - easy-to-use PyNN based API
  - high-level support available
- BrainScaleS architecture can be scaled to large-size multi-chip networks
  - already demonstrated with the BrainScaleS-1 wafer-scale prototype
  - will be part of the next major version of the BrainScaleS-2 ASIC