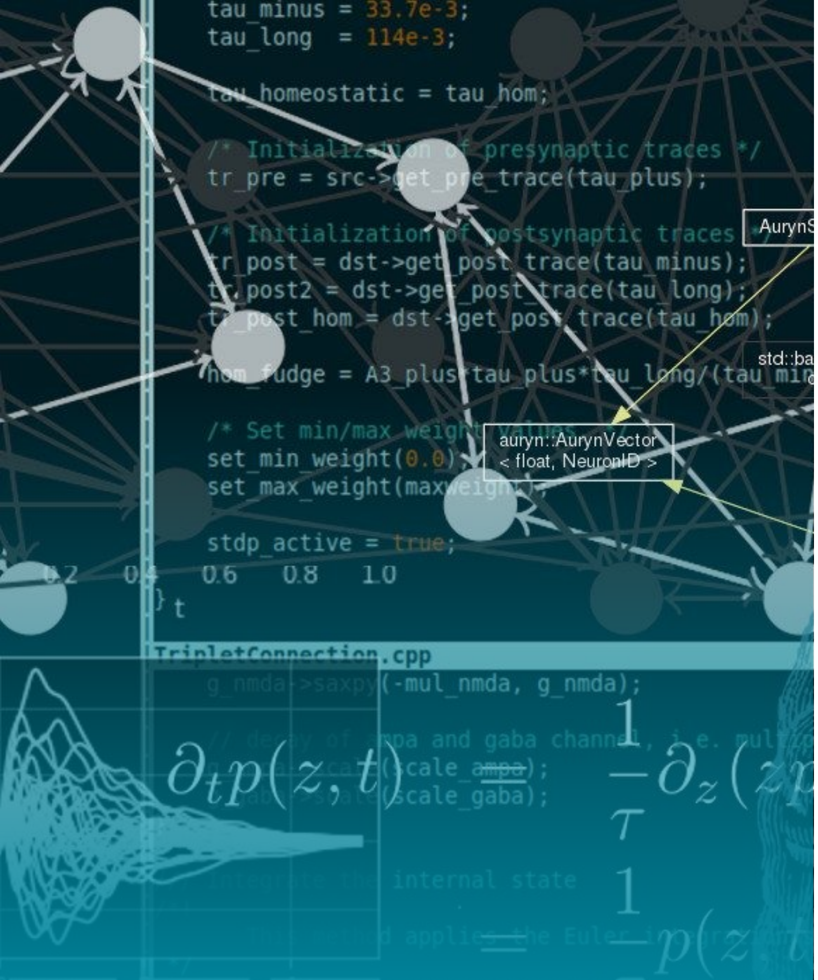


FMI
Friedrich Miescher Institute for Biomedical Research
Affiliated Institute of the University of Basel

Learning algorithms for spiking and physical neural networks

Friedemann Zenke
Computational Neuroscience @ FMI
www.zenkelab.org

University of Basel

Affiliated with Novartis Biomedical Research

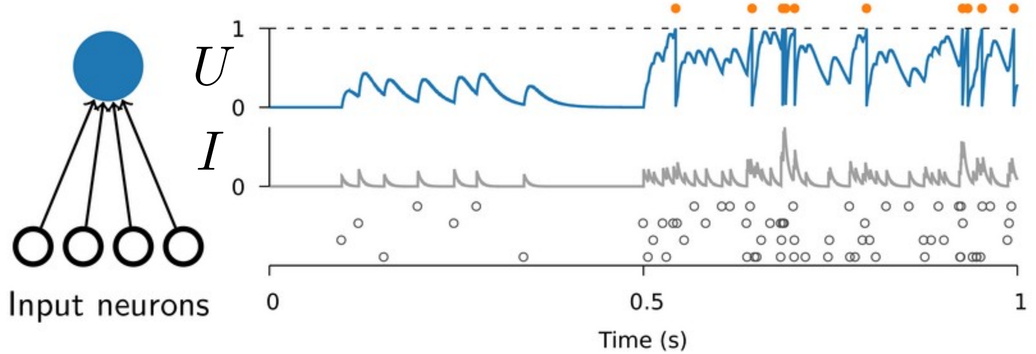
Today: Overview recent work on learning algorithms

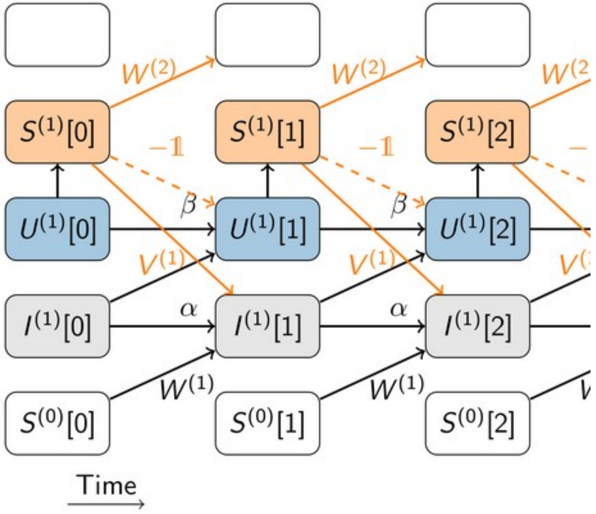
- Surrogate gradients for spiking neural networks
  - Getting spiking neural networks to do something interesting (recap)
    Neftci, Mostafa, and Zenke (2019) *IEEE SPM*
  - Sidestepping device mismatch on analog neuromorphic substrates
    Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). *PNAS*
- Resurrecting local learning rules (beyond backprop)
  - Training noisy substrates with holomorphic equilibrium propagation
    Laborieux and Zenke (2022) *Neurips*
  - Online self-supervised learning with local learning rules
    Halvagal and Zenke (2023) *Nature Neuroscience*

How do we get spiking neural networks to do something interesting?

Through end-to-end training!

Recap: Training spiking networks end-to-end





Neftci, Mostafa, & Zenke (2019)

- Spiking neurons & networks are RNNs
- They have implicit and explicit recurrence
- Known training procedures for networks **with hidden units**
  - Backpropagation-through time (BPTT)
  - Real-time recurrent learning (RTRL)

$$S_i^{(1)}[n] = \boxed{\Theta\left(U_i^{(1)}[n] - \vartheta\right)}$$

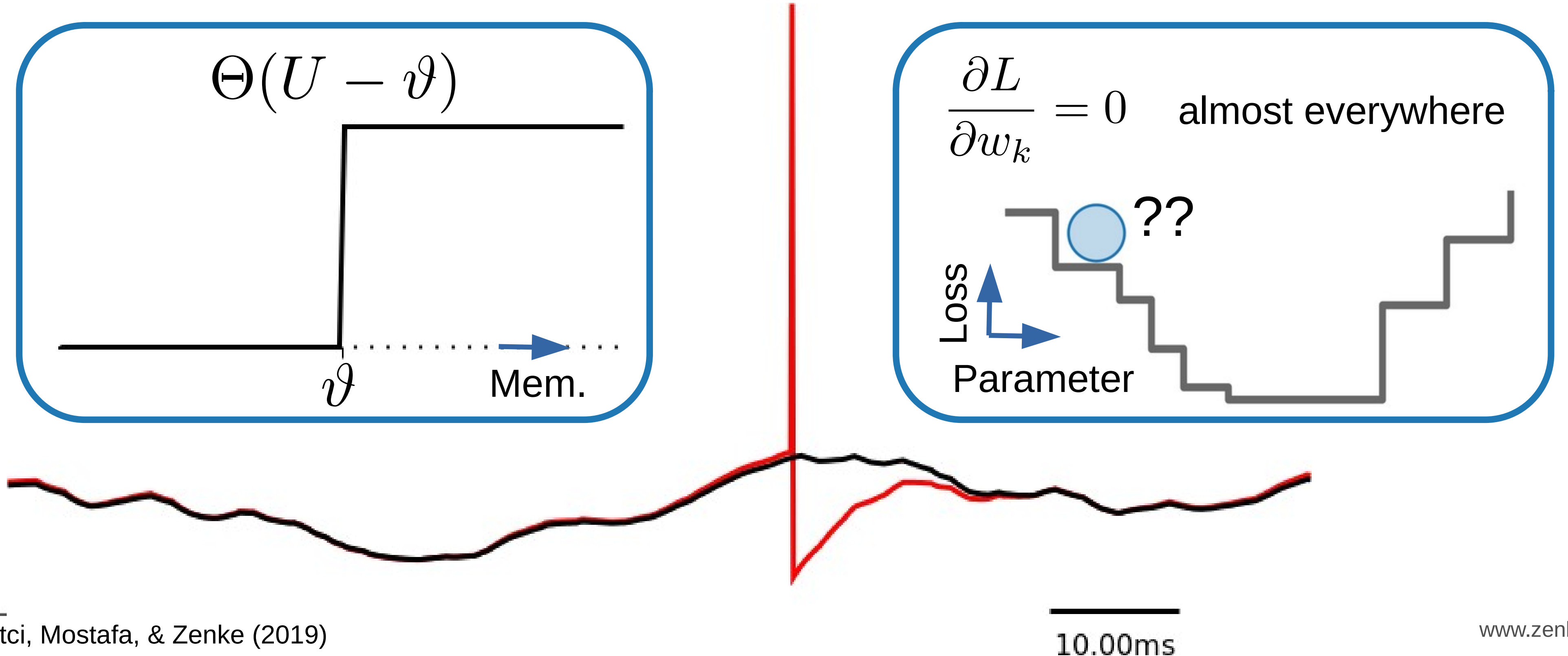
Problem

$$U_i^{(1)}[n+1] = \beta U_i^{(1)}[n] + I_i^{(1)}[n] - S_i[n]$$

$$I_i^{(1)}[n+1] = \underbrace{\alpha I_i^{(1)}[n]}_{\text{exp. current decay}} + \underbrace{\sum_j W_{ij} S_j^{(0)}[n]}_{\text{feed-forward input}}$$

Forward Euler integration

Problem: The derivative of a spike train is zero almost everywhere

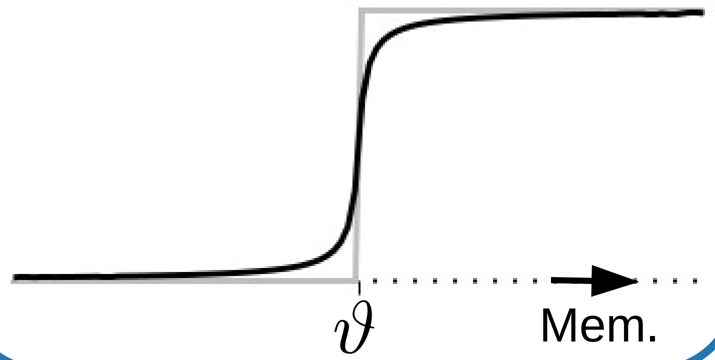


Neftci, Mostafa, & Zenke (2019)

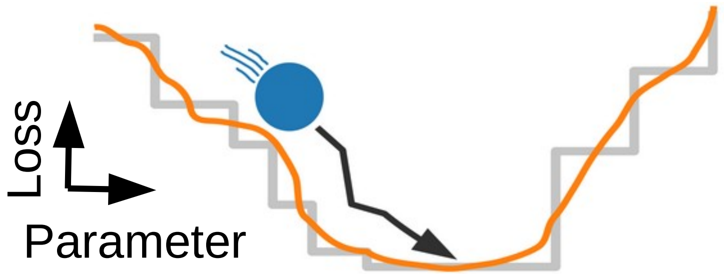
Solution: Surrogate gradient

Bohte (2011), Esser et al. (2015), Bellec et al. (2018),
Shrestha & Orchard (2018), Zenke & Ganguli (2018), ...
In ML: “Straight-through estimators” Bengio et al. (2013)

$$\Theta(U(t) - \vartheta) \approx \sigma(U(t))$$



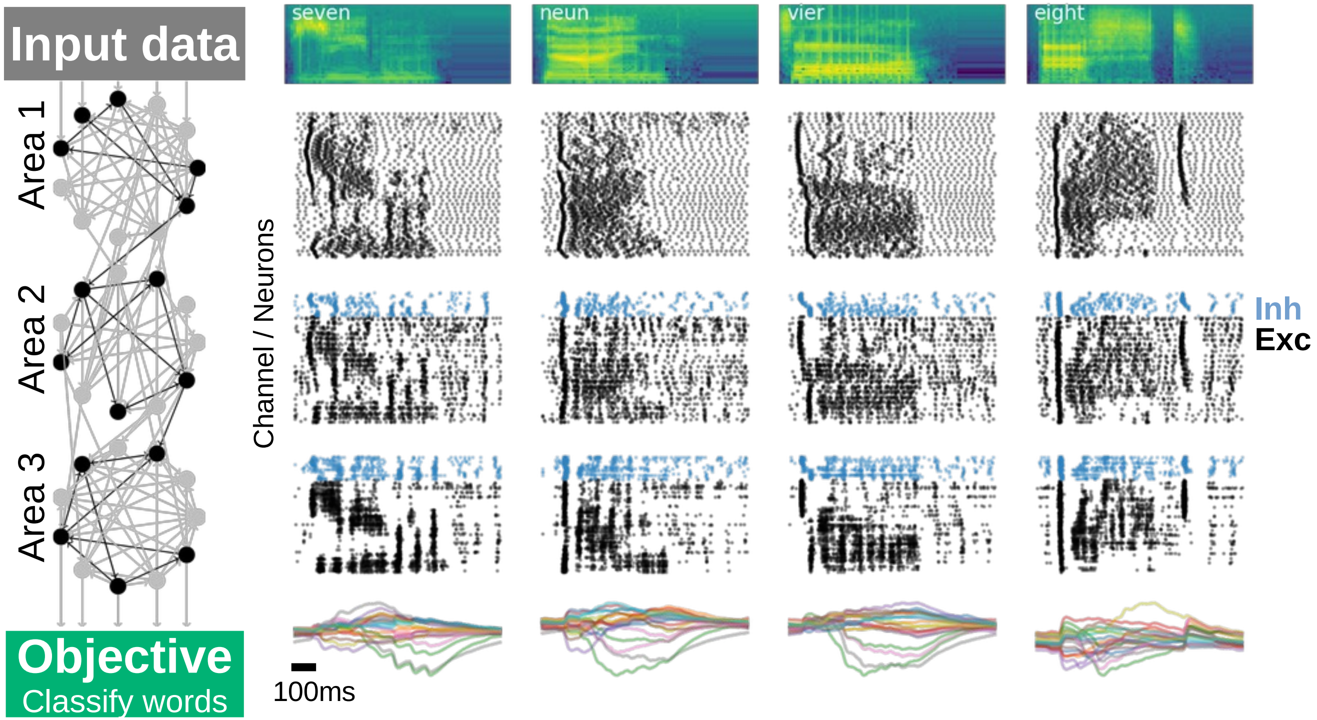
$$\frac{\partial L}{\partial w_k} \neq 0$$



Neftci, Mostafa, & Zenke (2019)

Surrogate gradients

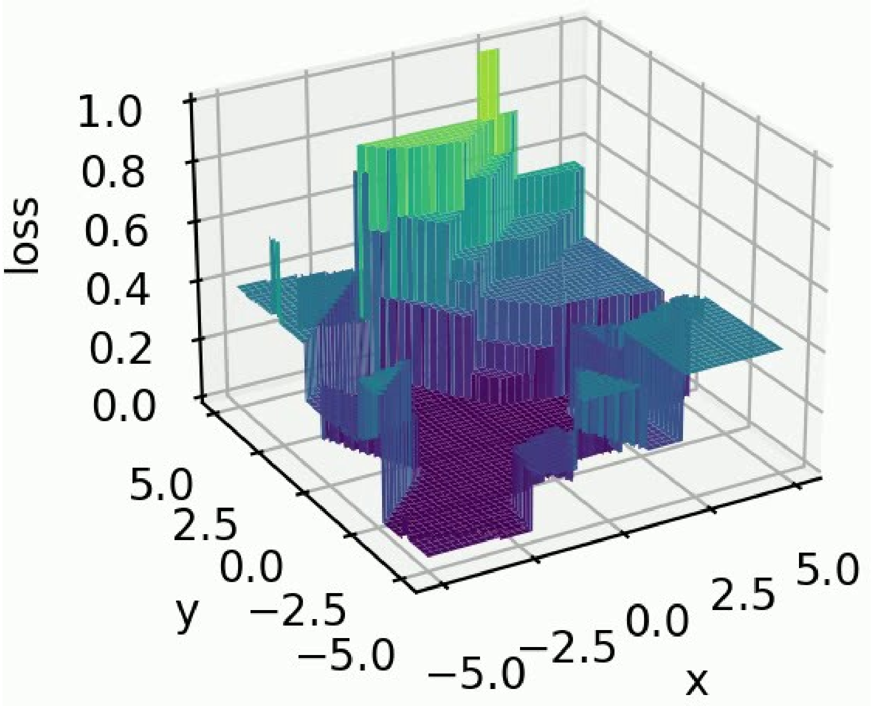
No assumption about rate or time coding required.

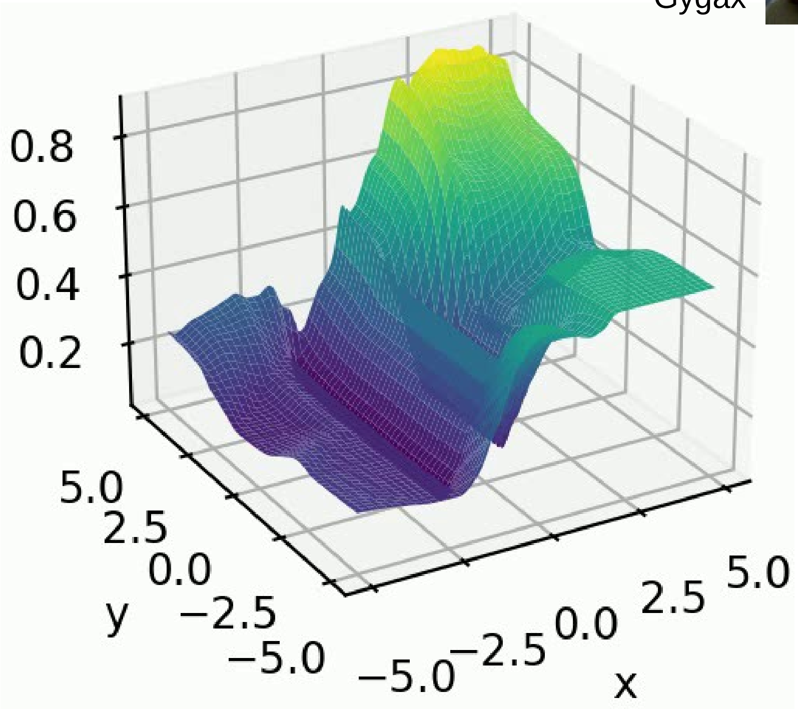


Cramer, Stradmann, Schemmel, and Zenke (2020)

Loss landscape of a spiking net (2D projection)

Julia
Gygax

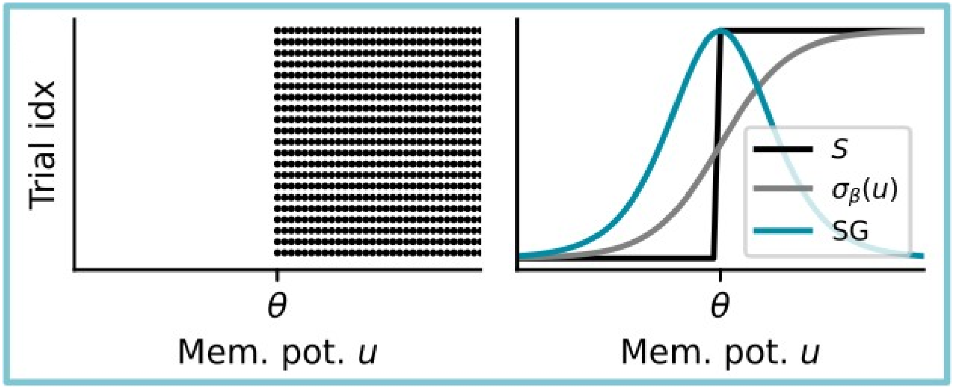


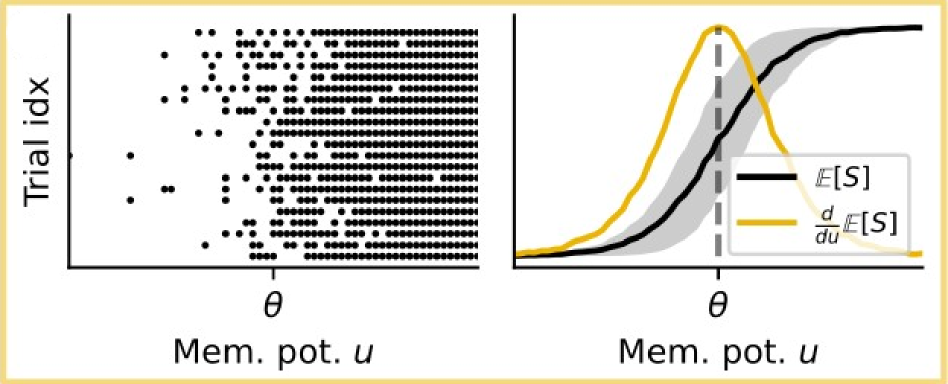


Integrated surrogate gradient

Problem: Surrogate gradients are a heuristic and lack theory.

Surrogate gradients are related to well-defined gradients in expectation in single stochastic neurons





In spiking neurons: Pfister, Toyoizumi, Barber & Gerstner (2006), Gardner, Sporea & Grüning (2015)

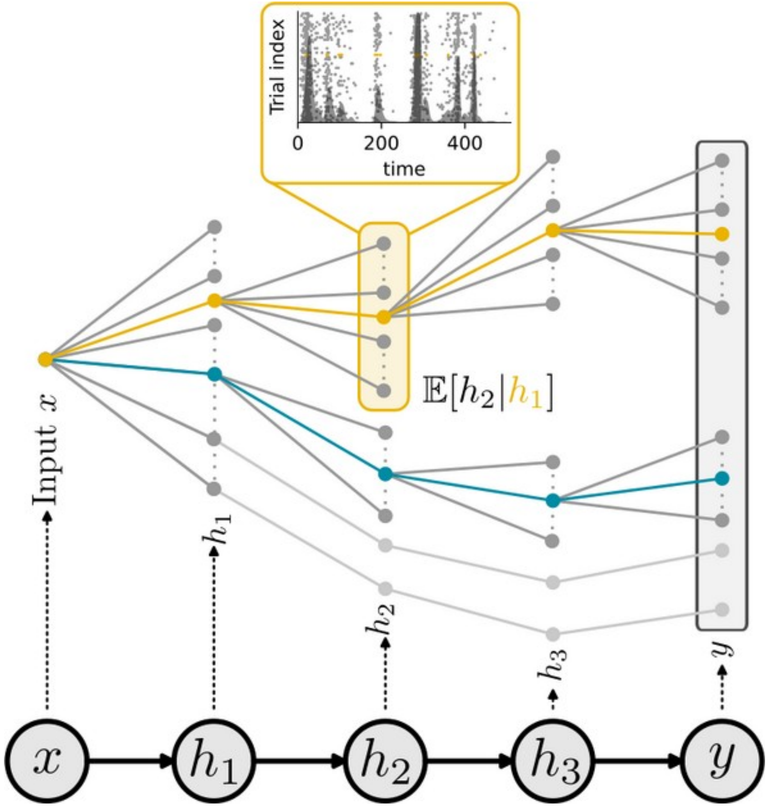
But does not work in multi-layer networks because it breaks the chain rule:

$$\mathbb{E}\left[\frac{\partial}{\partial p_y}y\frac{\partial}{\partial h_2}p_y\frac{\partial}{\partial p_2}h_2\dots\frac{\partial}{\partial w_1}p_1\right]\neq\mathbb{E}\left[\frac{\partial}{\partial p_y}y\right]\mathbb{E}\left[\frac{\partial}{\partial h_2}p_y\right]\mathbb{E}\left[\frac{\partial}{\partial p_2}h_2\right]\dots\mathbb{E}\left[\frac{\partial}{\partial w_1}p_1\right]$$

Gygax & Zenke (2024) arXiv

Stochastic Automatic Differentiation provides missing theoretical foundation for surrogate gradients

Julia Gygax



- Finite differences? Does not scale, high variance → not an option
- "Stochastic automatic differentiation"
  Arya, Schauer, Schäfer, Rackauckas, 2022. *Neurips*

$$\frac{\partial}{\partial w_1}\mathbb{E}[y] \approx \frac{\partial}{\partial h_2}\mathbb{E}[y|h_2^*]\frac{\partial}{\partial h_1}\mathbb{E}[h_2|h_1^*]\frac{\partial}{\partial w_1}\mathbb{E}[h_1|x]$$

- Surrogate gradients fall out of this framework

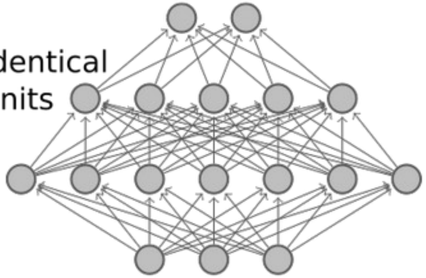
How do we perform efficient inference with (spiking) neural networks?

With ultra-low power neuromorphic hardware!

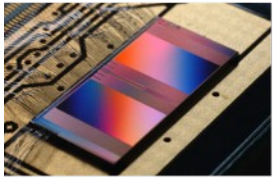
(use the device physics)

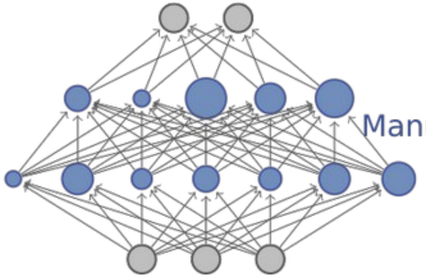
Problem: Device mismatch

Software implementation

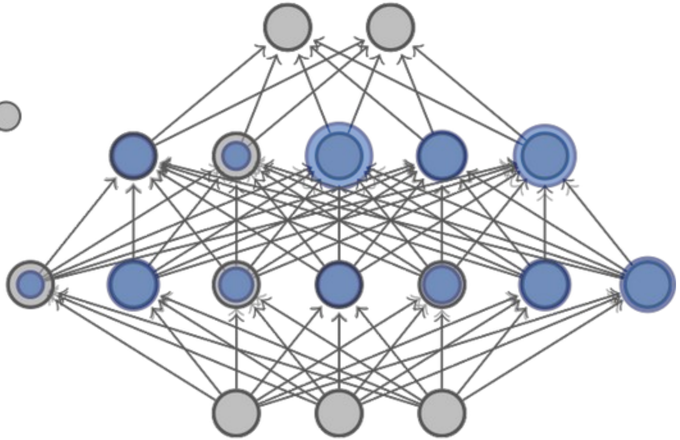


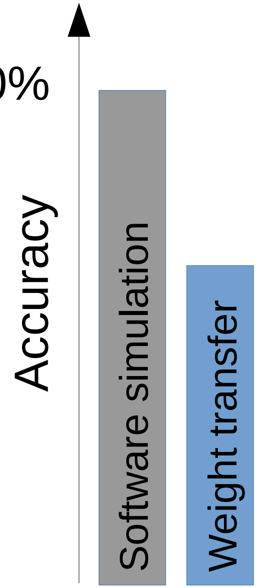
Hardware implementation





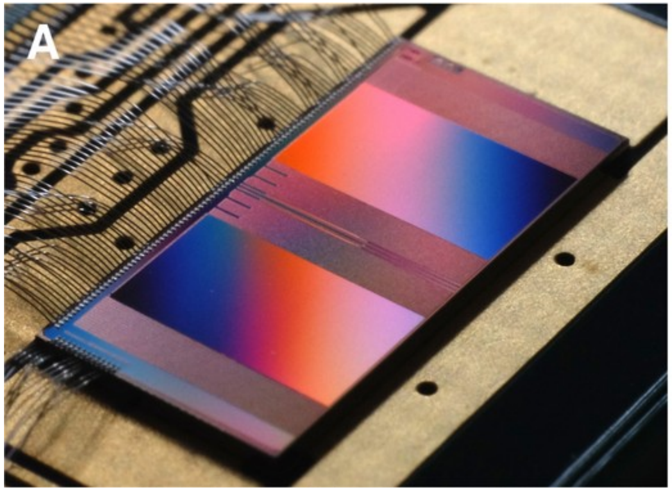
Mismatch!

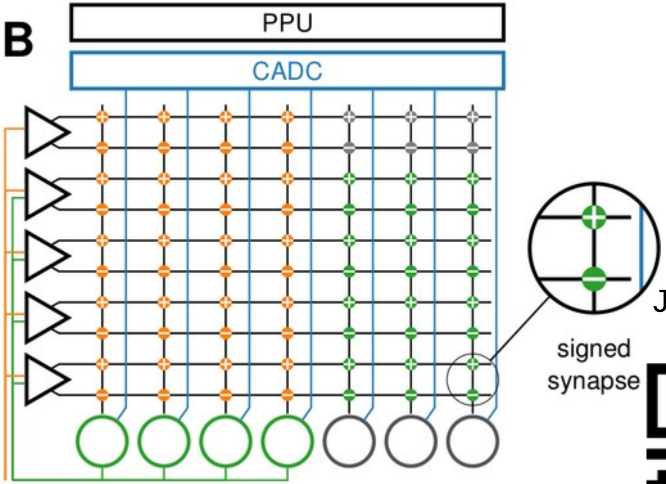




→ costly calibration

To study this question we used the BrainScaleS-2 **analog neuromorphic hardware** system





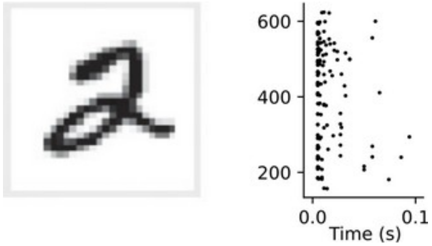
Johannes Schemmel
Uni Heidelberg

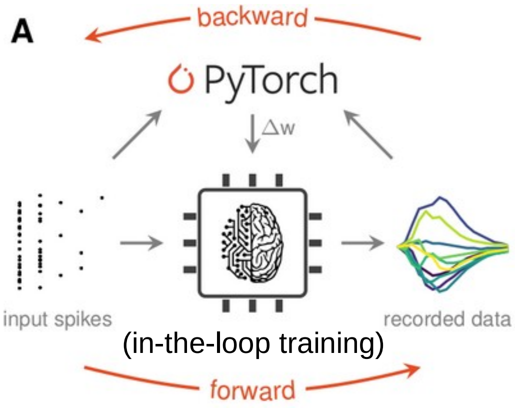
Benjamin Cramer

Sebastian Billaudelle

Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). PNAS

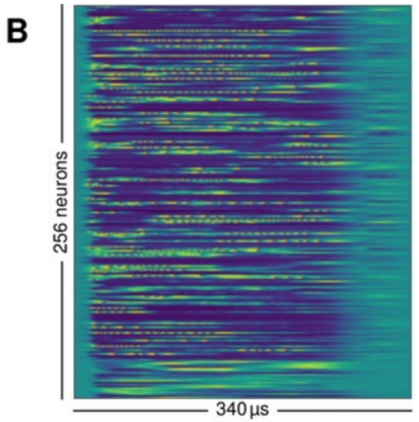
In-the-loop surrogate gradient training
Forward-pass on chip and backward pass in software



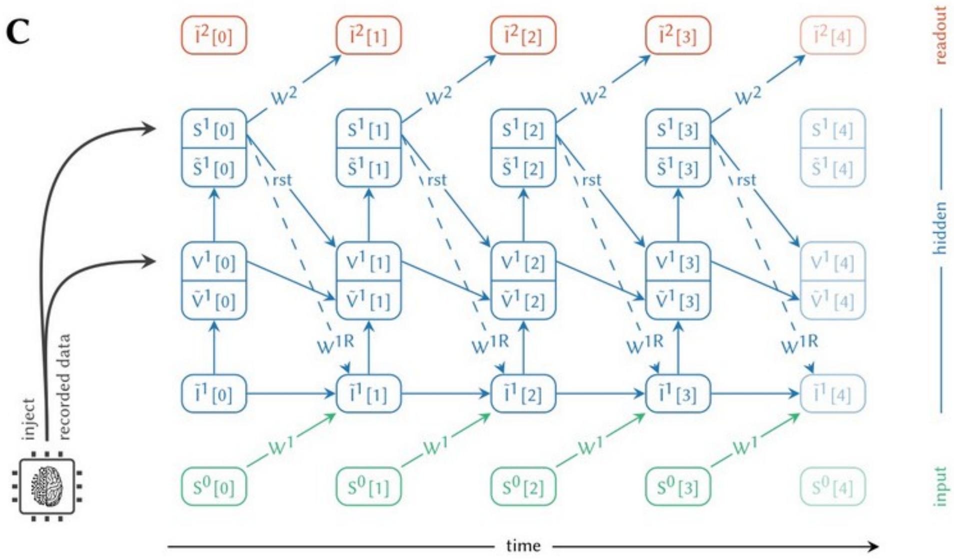


1) Forward pass on chip

2) Measure on-chip analog voltage traces



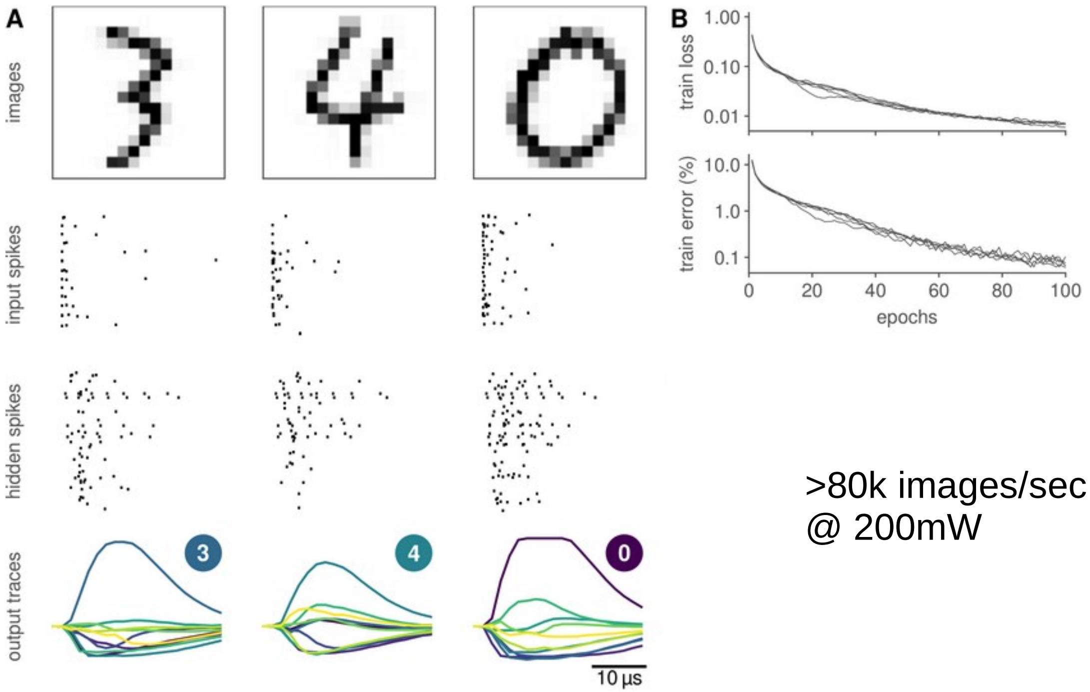
3) Inject voltage into computational graph



4) Compute surrogate gradients → update weights

Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). PNAS

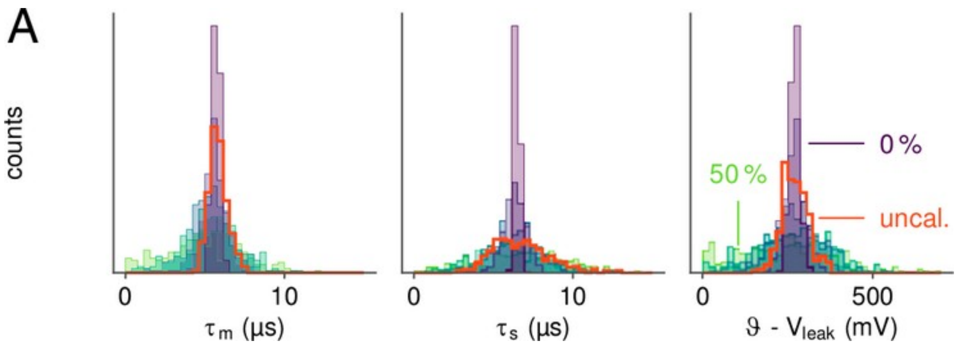
Functional spiking neural networks trained on BrainScaleS-2 **analog neuromorphic hardware**

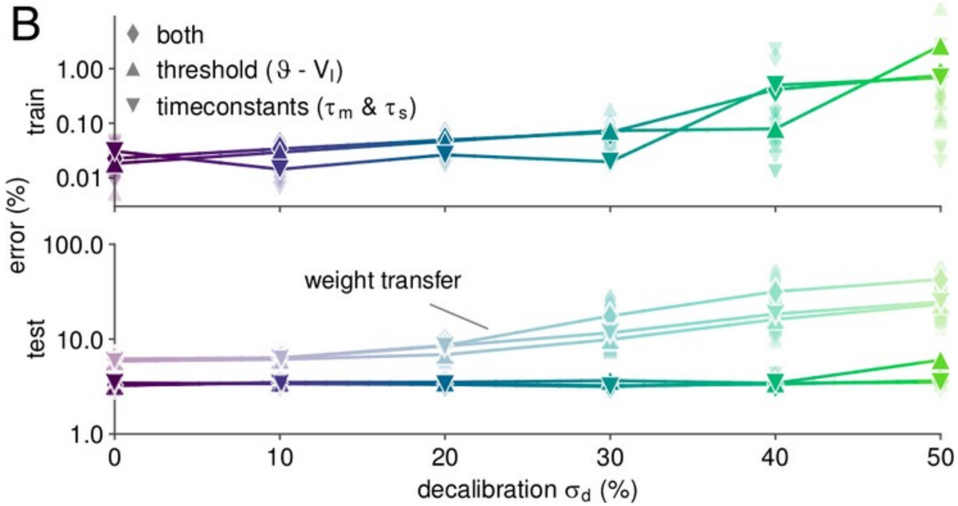


>80k images/sec
@ 200mW

Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). PNAS

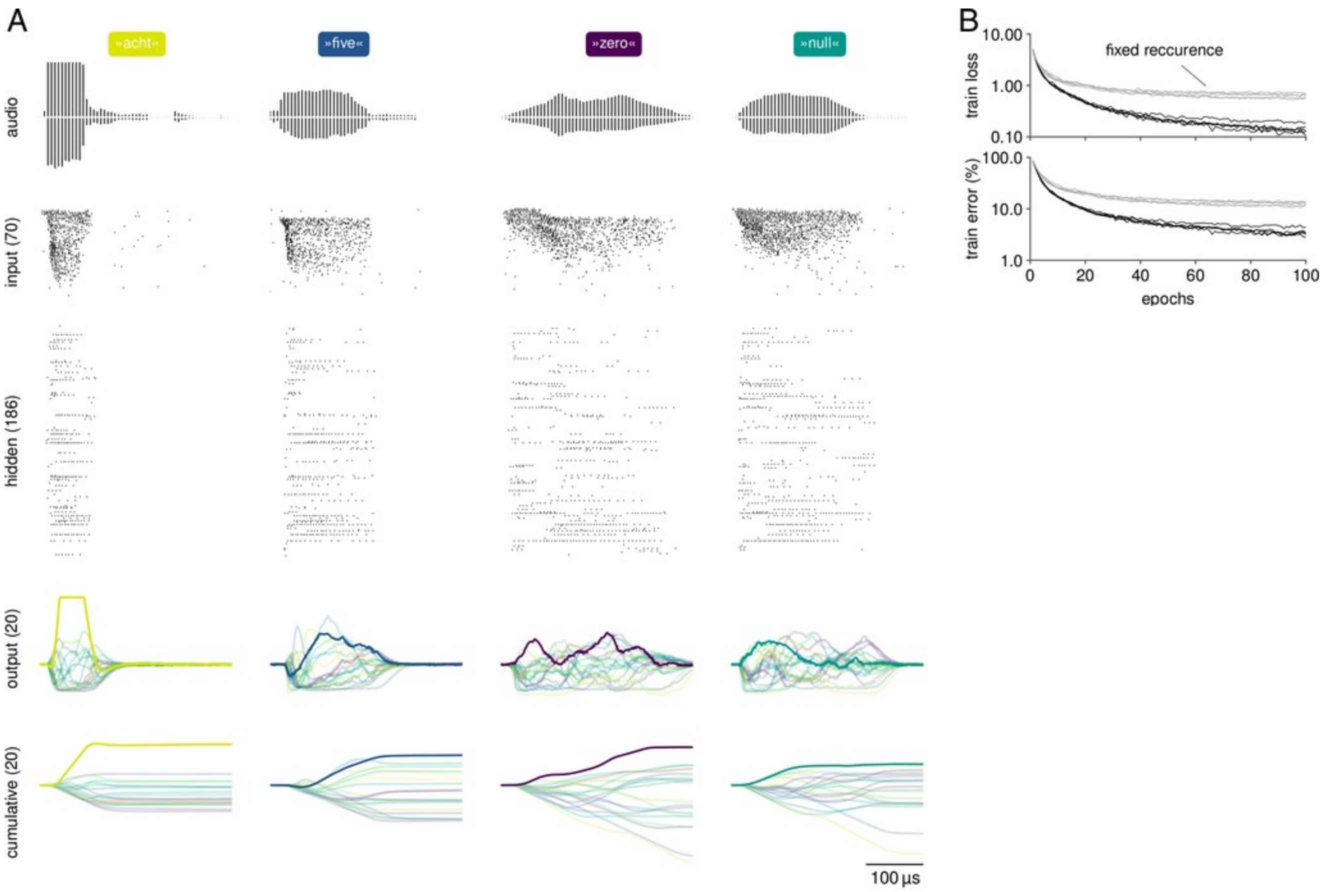
Surrogate gradient learning self-calibrates the **analog neuromorphic substrate**





Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). PNAS

Speech classification and keyword spotting (SHD)



Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). PNAS

Summary: Voltage aware surrogate gradients can self-calibrate analog neuromorphic substrates

Johannes Schemmel
Uni Heidelberg

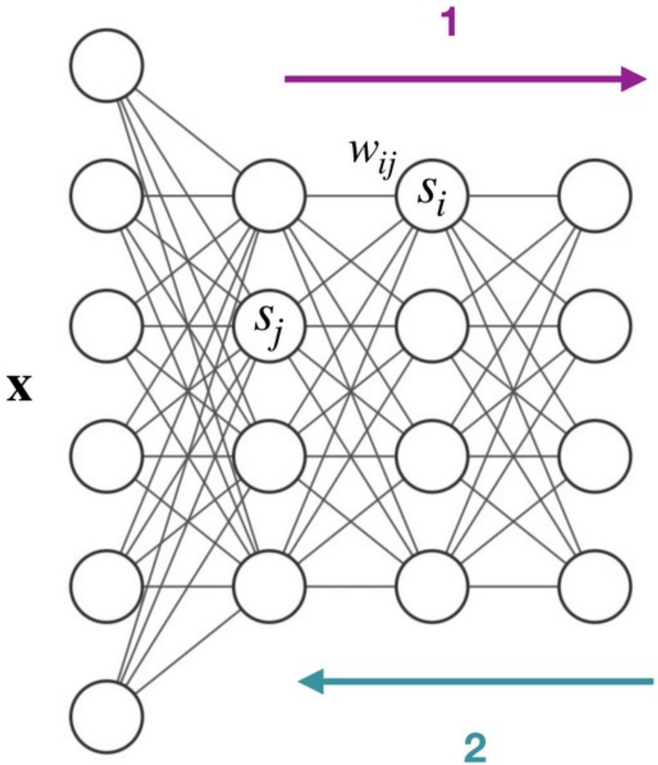
Benjamin Cramer

Sebastian Billaudelle

Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). *PNAS*

Still, training was done offline and used backprop.

Backprop is difficult to implement on neuromorphic systems



Rumelhart et al. *Nature 1986*

1 Nonlinear computation

$$s_i = \sigma\left(\Sigma_j w_{ij} s_j + b_i\right)$$

$\mathscr{L}$ loss function quantifying good or bad

2 Error Backpropagation (BP)

$$\Delta w_{ij} \propto -\frac{\mathrm{d}\mathscr{L}}{\mathrm{d}w_{ij}} = -\delta_i \sigma(s_j)$$

non local

Question
How to train noisy physical networks without backprop?

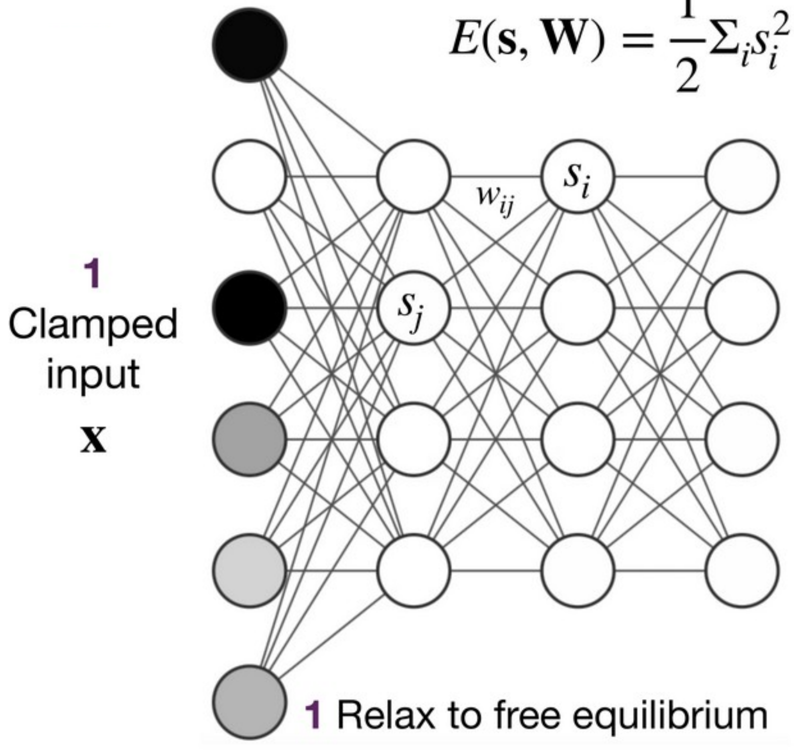
Axel Laborieux

Laborieux and Zenke (2022) *Neurips*

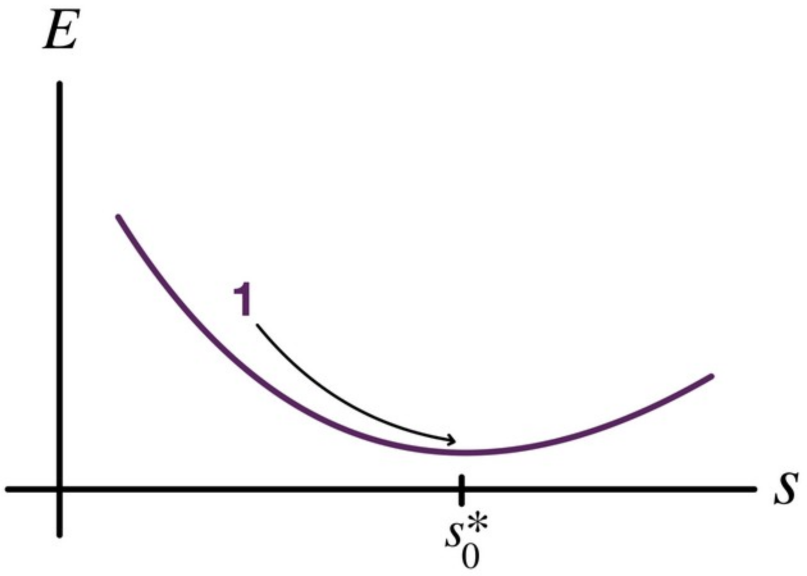
Holomorphic Equilibrium Propagation Computes Exact Gradients Through Finite Size Oscillations

Equilibrium Propagation (EP) is an alternative

$$E(\mathbf{s}, \mathbf{W}) = \frac{1}{2}\Sigma_i s_i^2 - \Sigma_{i<j} w_{ij}\sigma(s_i)\sigma(s_j)$$



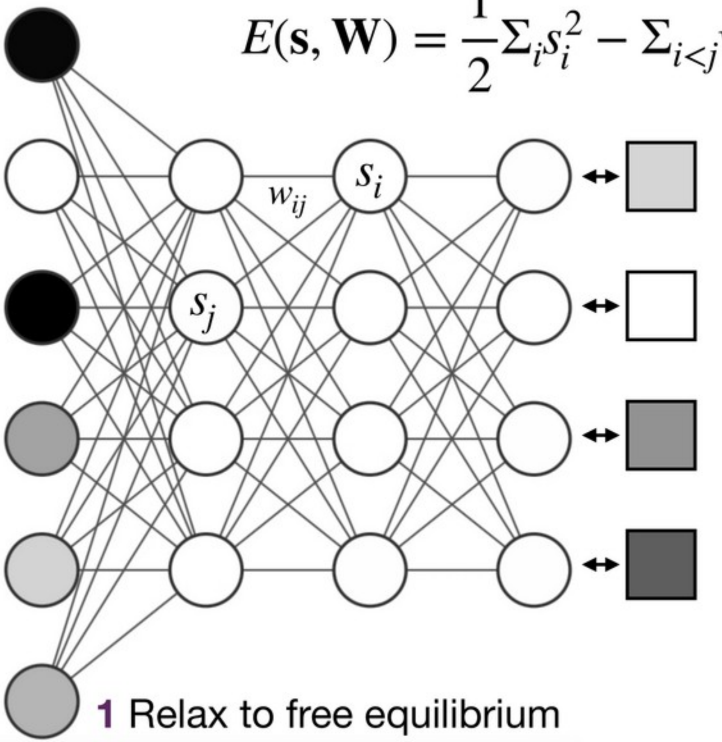
1 Free output



Scellier & Bengio *Frontiers in neuroscience* 2017

$$E(\mathbf{s}, \mathbf{W}) = \frac{1}{2}\Sigma_i s_i^2 - \Sigma_{i<j} w_{ij}\sigma(s_i)\sigma(s_j)$$

1 & **2** Clamped input $\mathbf{x}$



1 Free output

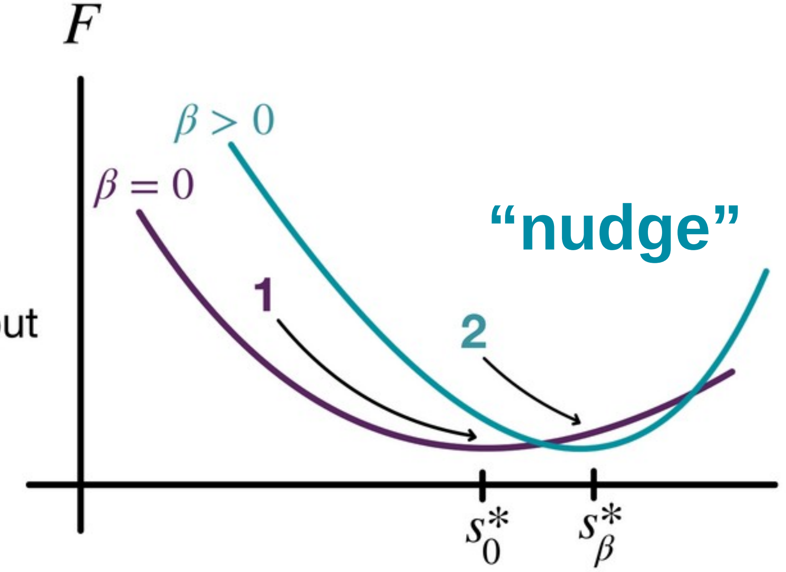
2 **Nudged** output

1 Relax to free equilibrium

2 Relax to nudged equilibrium

Scellier & Bengio *Frontiers in neuroscience* 2017

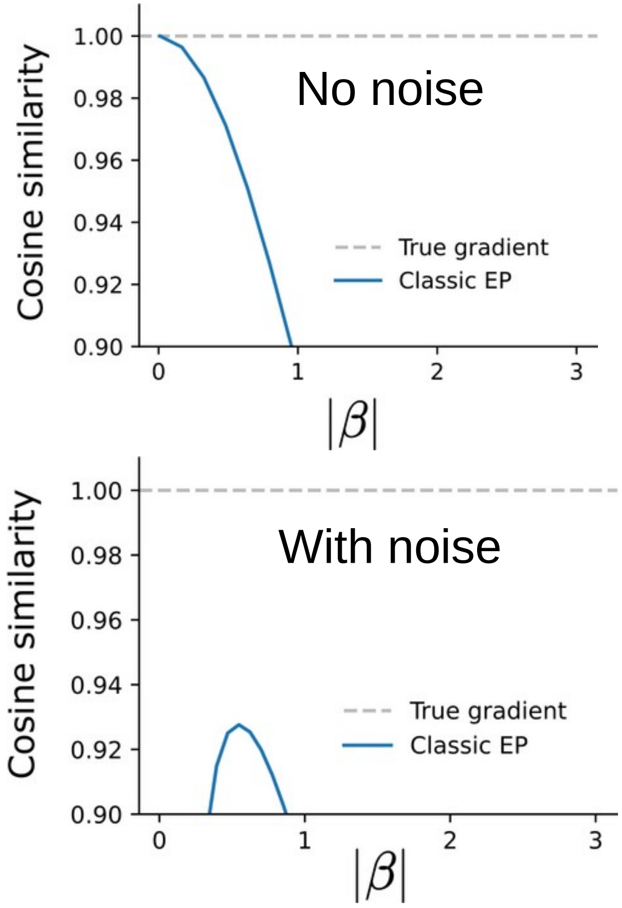
$$F(\mathbf{s}, \mathbf{W}) = E(\mathbf{s}, \mathbf{W}) + \beta\mathscr{L}$$

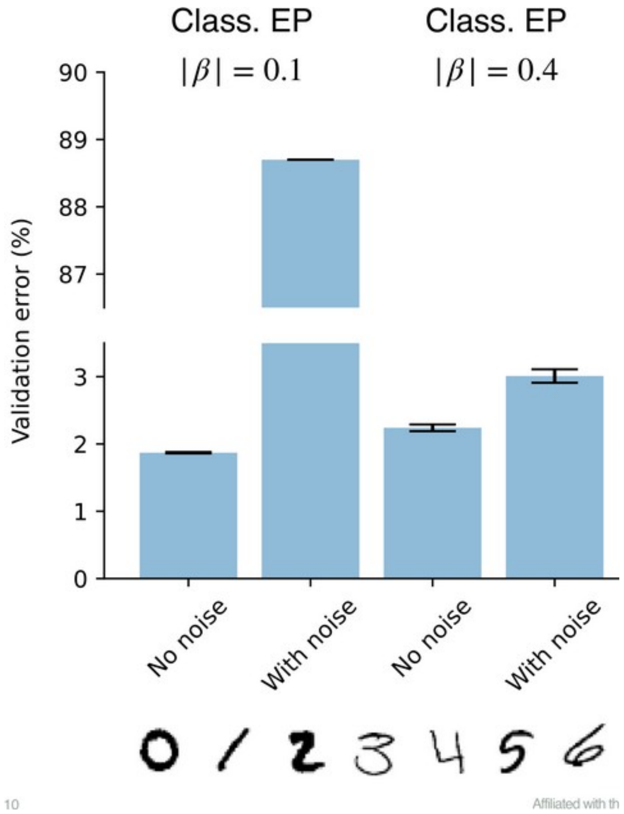


Local learning rule

$$\Delta w_{ij} \propto \frac{\sigma(s^*_{\beta,i})\sigma(s^*_{\beta,j}) - \sigma(s^*_{0,i})\sigma(s^*_{0,j})}{\beta} \rightarrow -\frac{\mathrm{d}\mathscr{L}}{\mathrm{d}w_{ij}} \text{ when } \beta \rightarrow 0$$

Classic Equilibrium Propagation is noise sensitive





Laborieux and Zenke (2022) *Neurips*

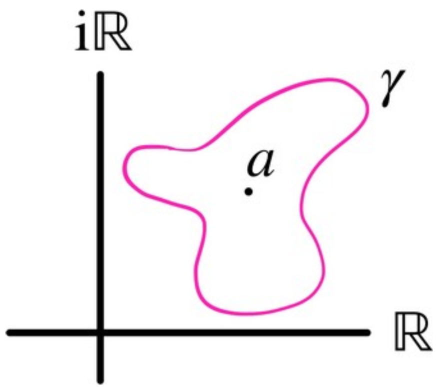
Complex analysis: Derivatives can be expressed as integrals

- Complex differentiability:

$$f'(a) := \lim_{z \to a} \frac{f(z) - f(a)}{z - a}$$

'holomorphic'

- Cauchy integral: $f'(a) = \frac{1}{2i\pi} \oint_\gamma \frac{f(z)}{(z-a)^2} dz$



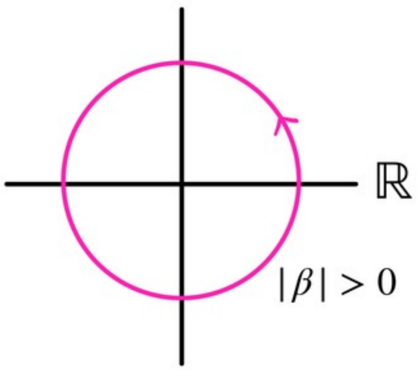
Laborieux and Zenke (2022) *Neurips*

Integration over one oscillation yields local learning rule

Replace derivative by a Cauchy integral

$$-\frac{\mathrm{d}\mathscr{L}}{\mathrm{d}\mathbf{W}}\bigg|_{\mathbf{W}_0} = \frac{\mathrm{d}}{\mathrm{d}\beta}\bigg|_{\beta=0}\left(\sigma(\mathbf{s}^*_\beta)\sigma(\mathbf{s}^*_\beta)^\top\right) = \frac{1}{2i\pi}\oint_\gamma \frac{\sigma(\mathbf{s}^*_\beta)\sigma(\mathbf{s}^*_\beta)^\top}{\beta^2}\mathrm{d}\beta$$

$\beta \in \mathbb{C}$



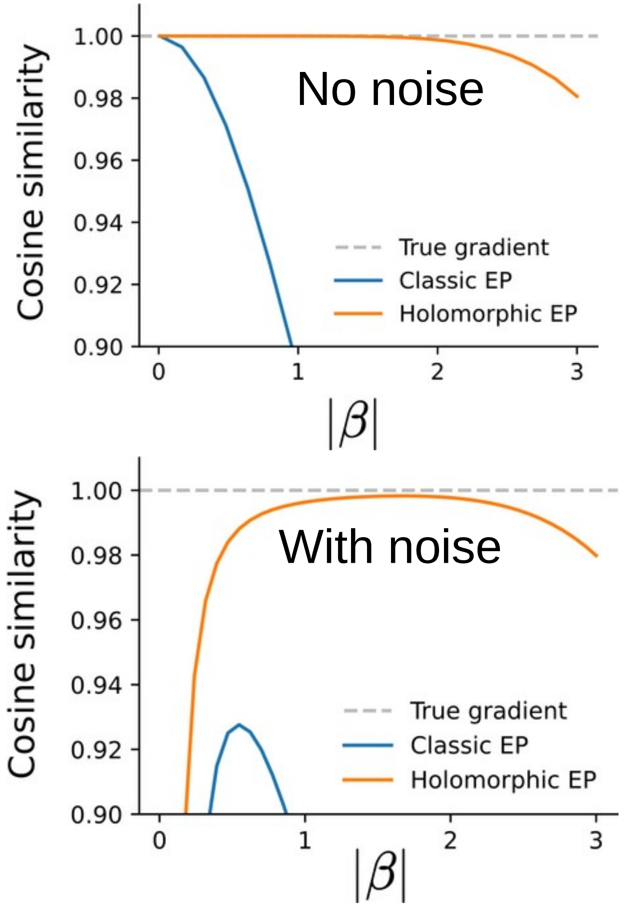
We choose the path: $t \in [0, T_{\mathrm{osc}}] \mapsto \beta(t) = |\beta| e^{2i\pi t/T_{\mathrm{osc}}}$

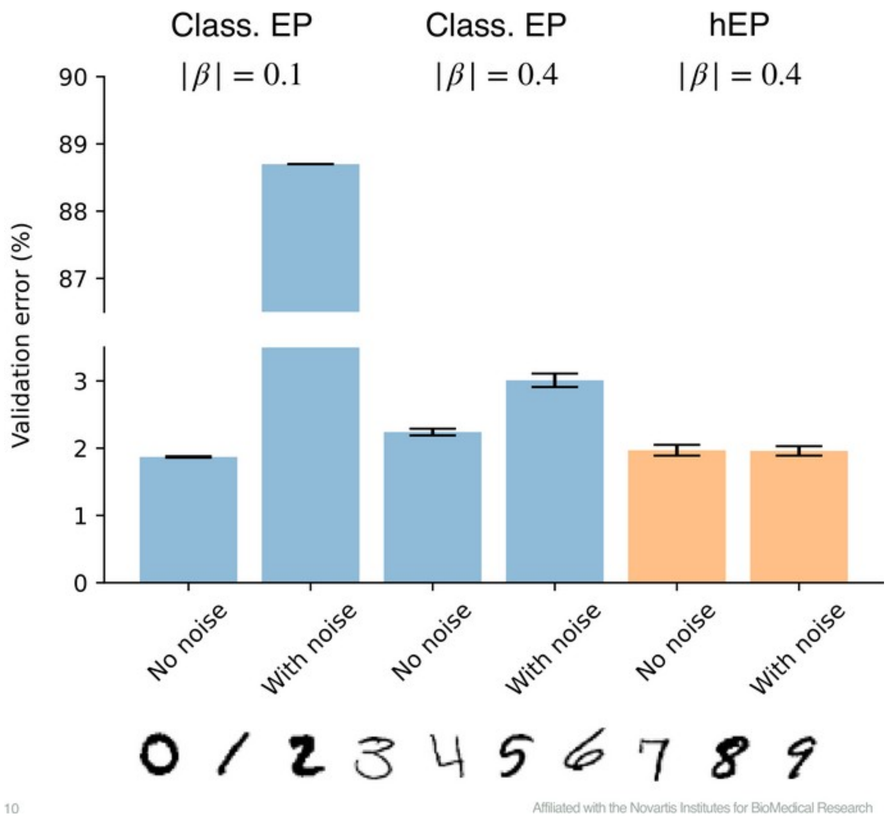
$$-\frac{\mathrm{d}\mathscr{L}}{\mathrm{d}\mathbf{W}}\bigg|_{\mathbf{W}_0} = \frac{1}{T_{\mathrm{osc}}|\beta|}\int_0^{T_{\mathrm{osc}}} \sigma(\mathbf{s}^*_{\beta(t)})\sigma(\mathbf{s}^*_{\beta(t)})^\top e^{-2i\pi t/T_{\mathrm{osc}}}\mathrm{d}t$$

Gradient = first Fourier coefficient of <u>nonlinear</u> neural oscillations

Laborieux and Zenke (2022) *Neurips*

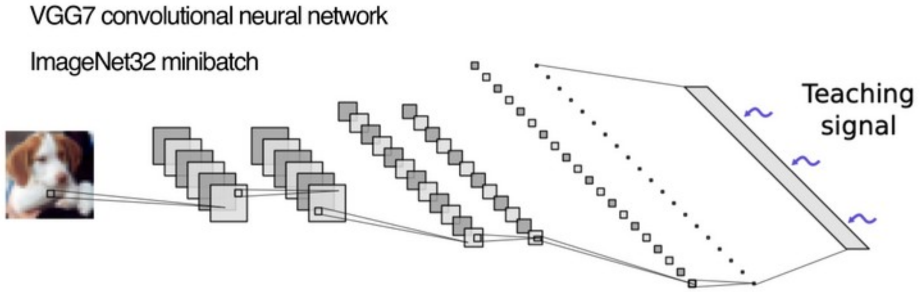
Holomorphic Equilibrium Propagation is robust

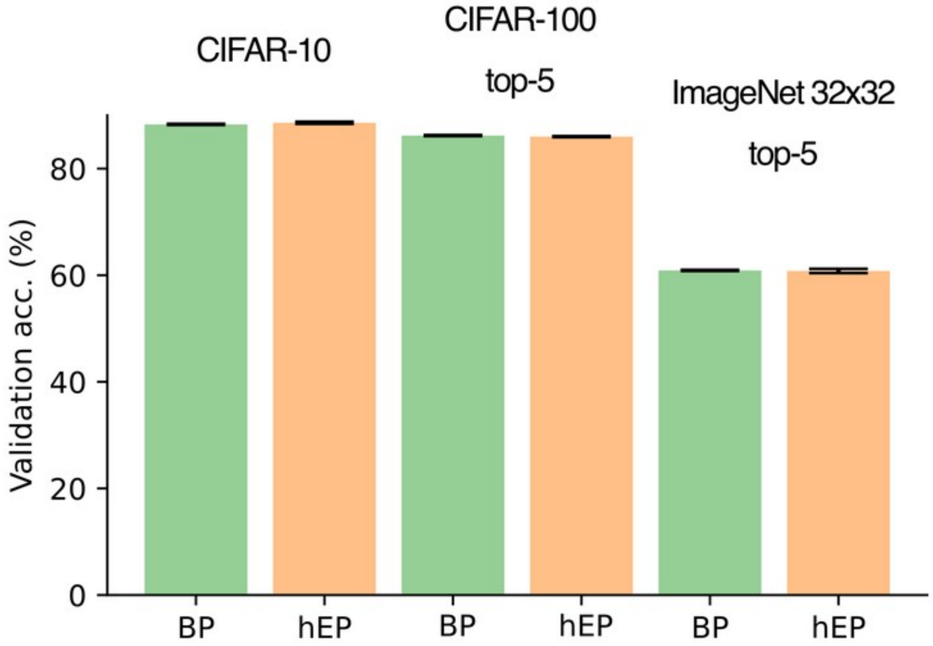




Laborieux and Zenke (2022) *Neurips*

Holomorphic EP scales to ImageNet
(thanks to larger teaching amplitudes)





Laborieux and Zenke (2022) *Neurips*

Summary: Holomorphic equilibrium propagation allows computing exact gradients on noisy physical systems (without backprop)

Axel Laborieux

Laborieux and Zenke (2022) *Neurips*

Holomorphic Equilibrium Propagation Computes Exact Gradients Through Finite Size Oscillations

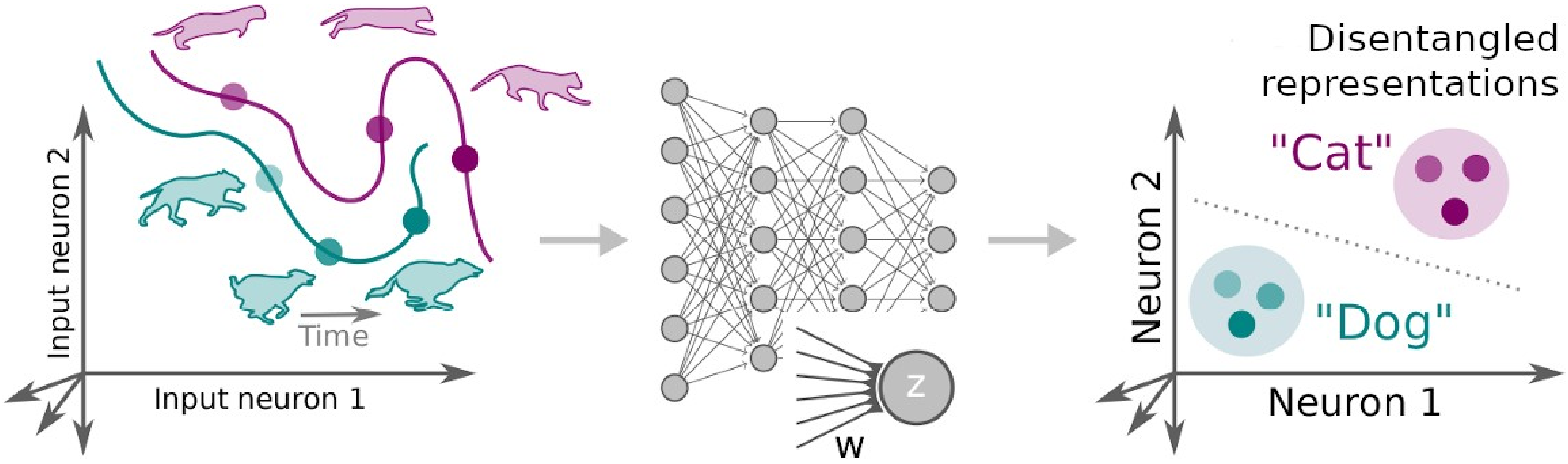
Follow-up paper dealing with weight asymmetry:

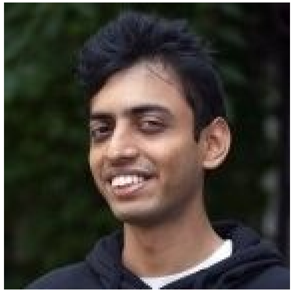
Laborieux and Zenke (2024) *accepted at ICLR*

Improving equilibrium propagation without weight symmetry through Jacobian homeostasis

Ongoing work: Make it real

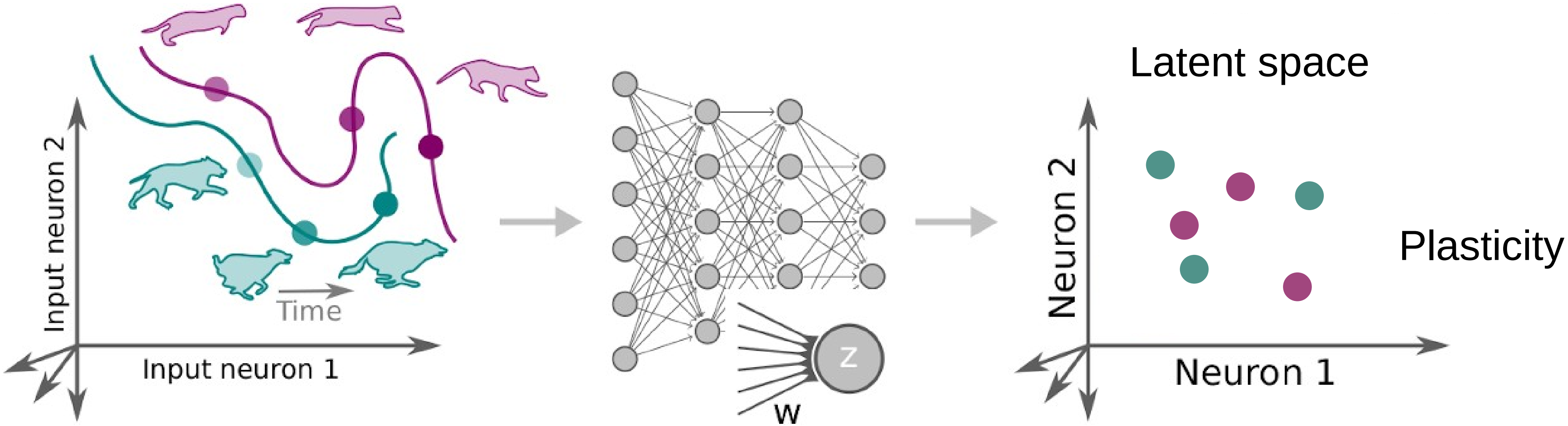
Latent Predictive Learning
Online self-supervised learning with local rules

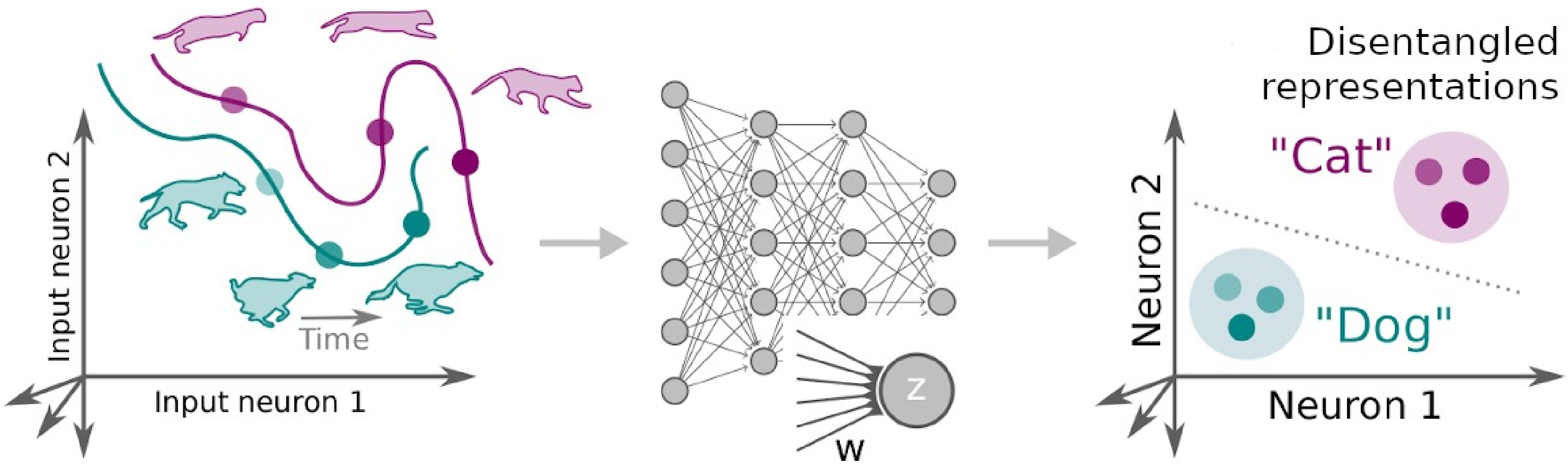




Manu S. Halvagal

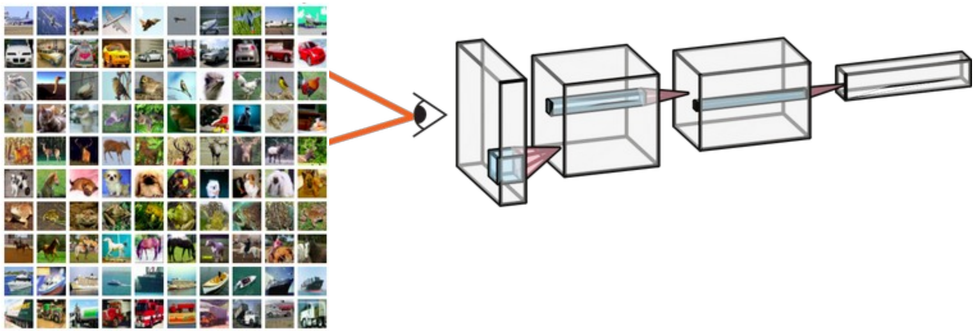
Halvagal and Zenke (2023) *Nat Neurosci*

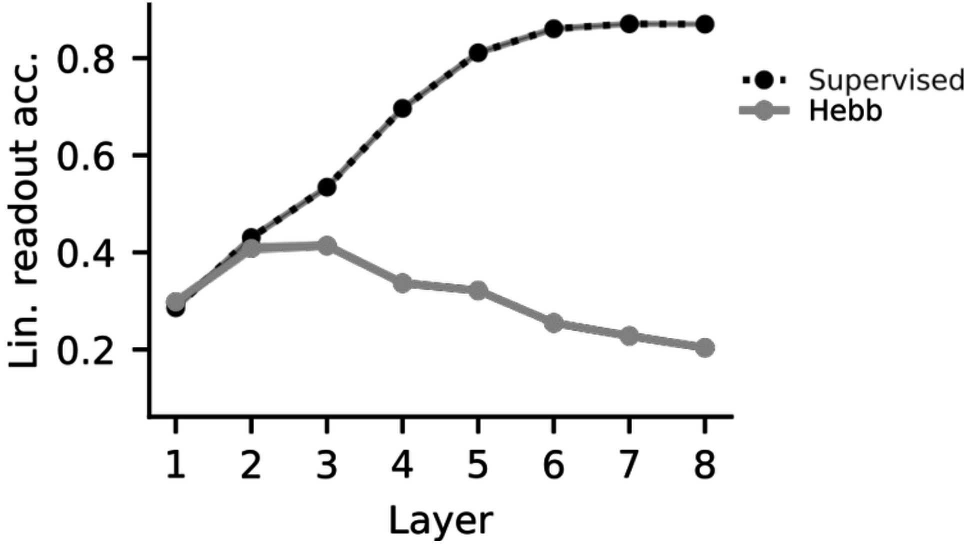

Latent space
Plasticity
Input neuron 2
Input neuron 1
Time
z
w
Neuron 2
Neuron 1



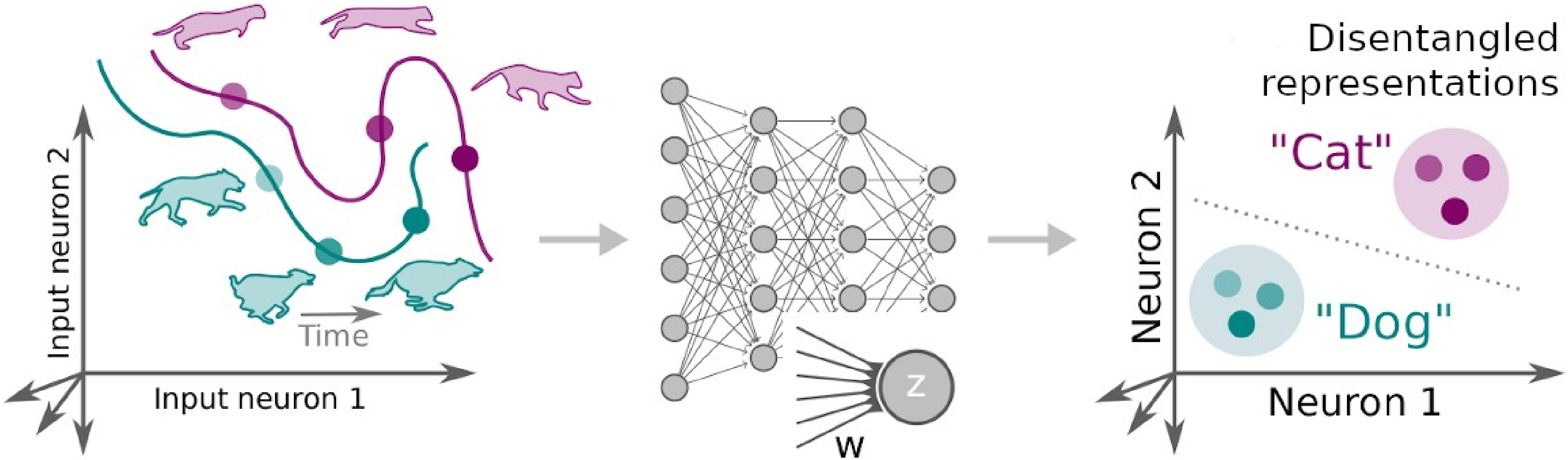
Input neuron 2
Input neuron 1
Time
w
z
Disentangled representations
"Cat"
"Dog"
Neuron 2
Neuron 1

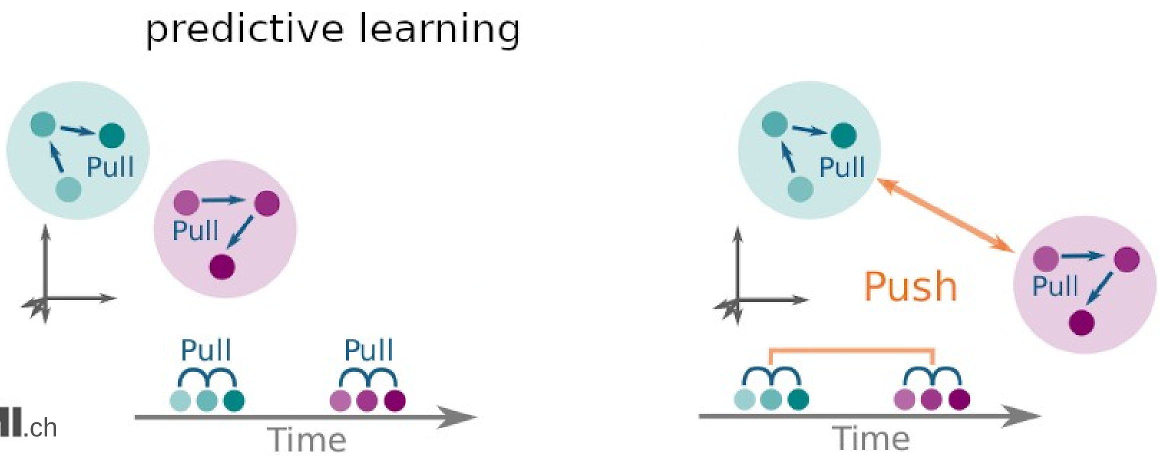
Problem: Hebbian plasticity, a bio-inspired local learning rule, does not learn *good* representations in deep nets



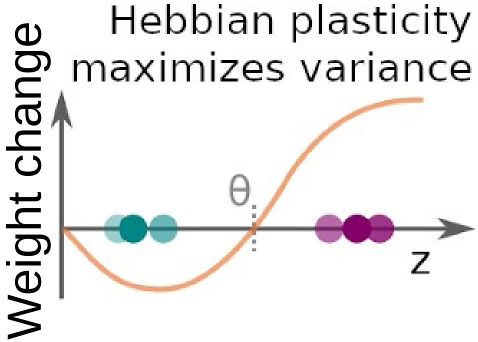


Idea: Optimize for latent space prediction





Oja (1982):



Combining latent space prediction and Hebbian plasticity yields local learning rule

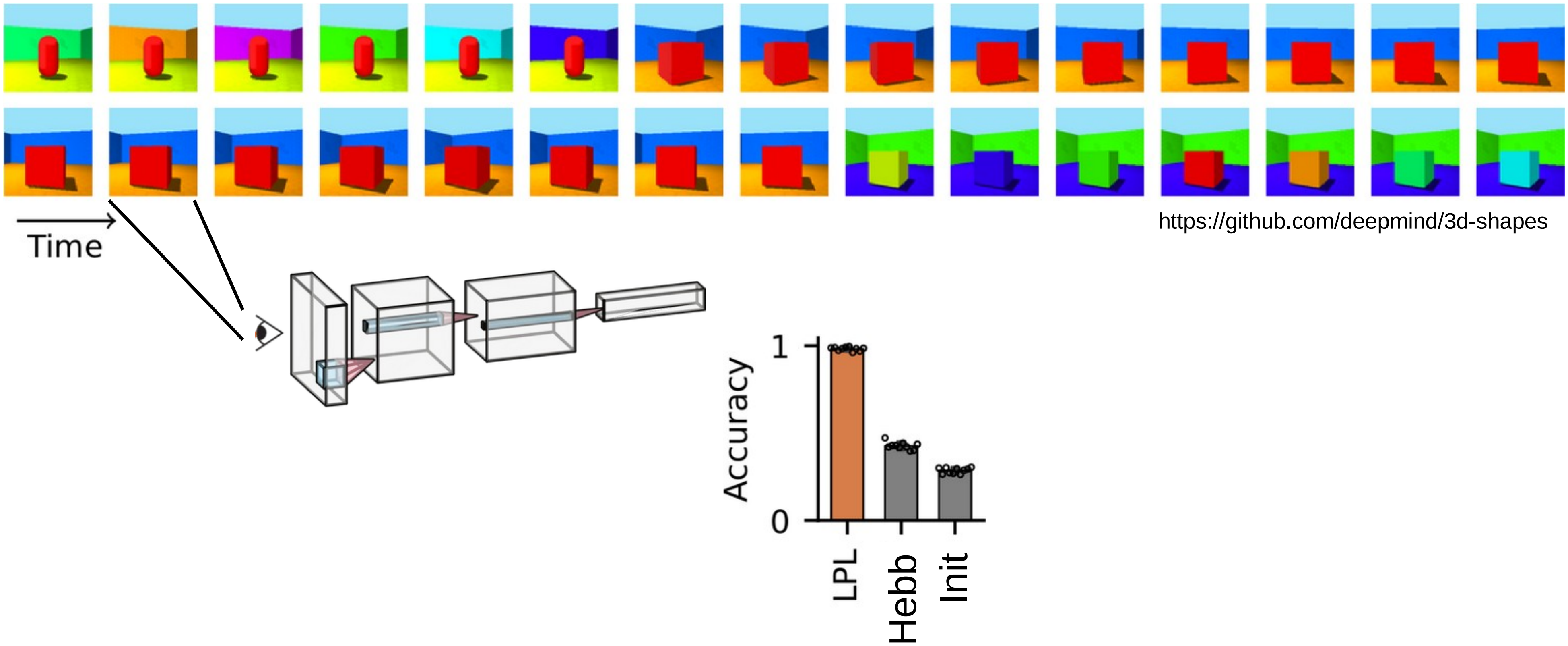
Pull **Push**

$$\mathcal{L} = \mathcal{L}_{\mathrm{Pred.}} + \mathcal{L}_{\mathrm{Hebb}}$$

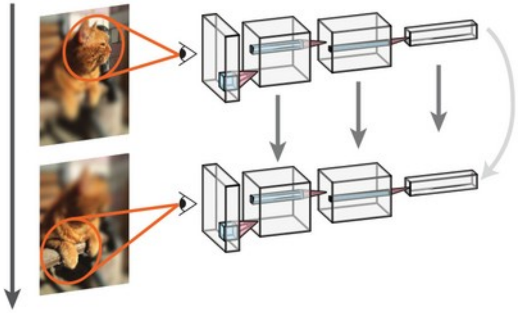
Resulting learning rule is local!

"Latent Predictive Learning LPL"

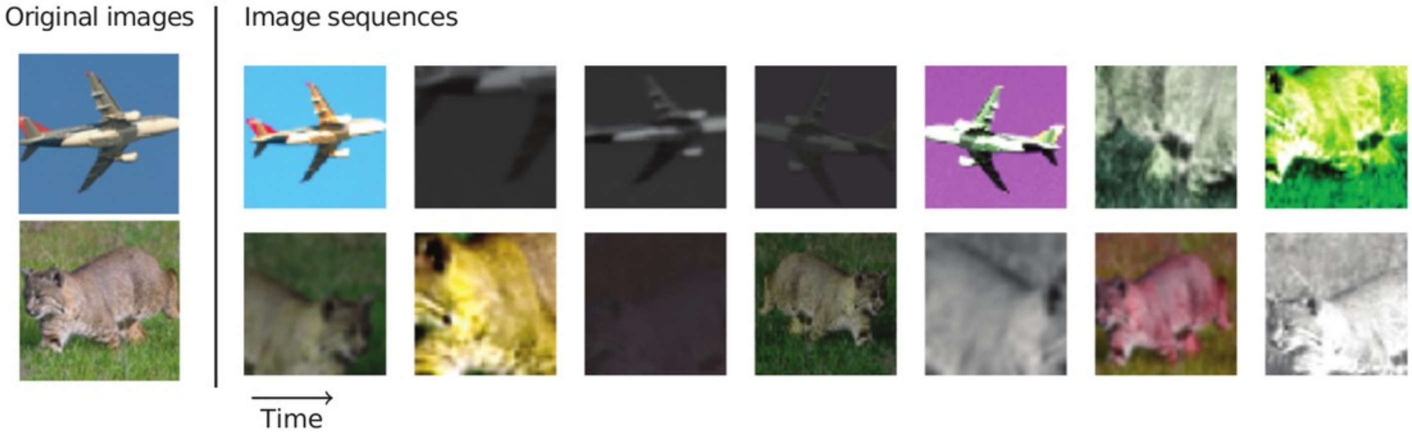
LPL disentangles objects from video data



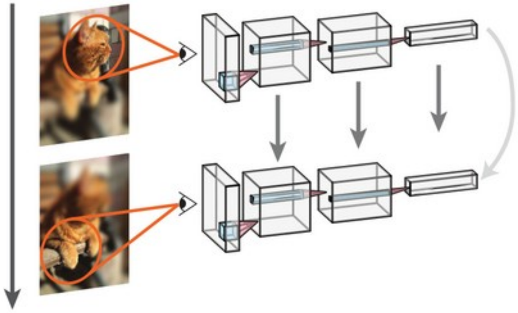
LPL learns invariant representations from augmented images



- VGG-11 model
- Neurons in all layers learn with **LPL**
- No backprop

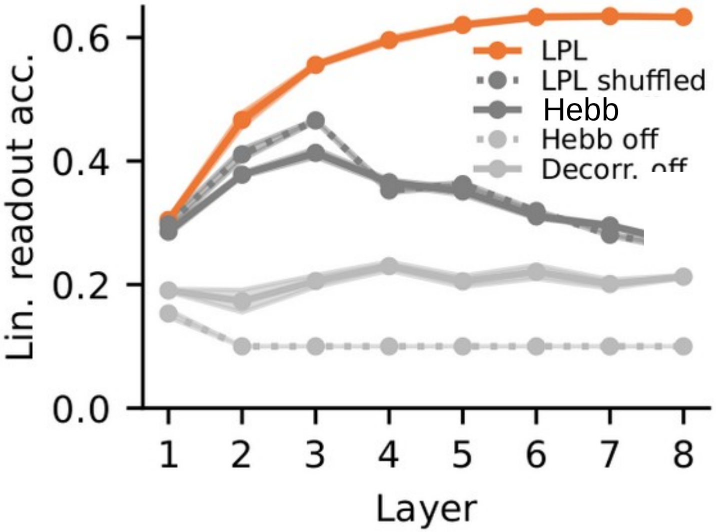


LPL learns invariant representations from augmented images



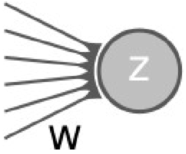
- VGG-11 model
- Neurons in all layers learn with **LPL**
- No backprop

After “watching” millions of image sequences …



LPL can be formulated as a local spiking learning rule

Based on SuperSpike: Zenke & Ganguli (2018)



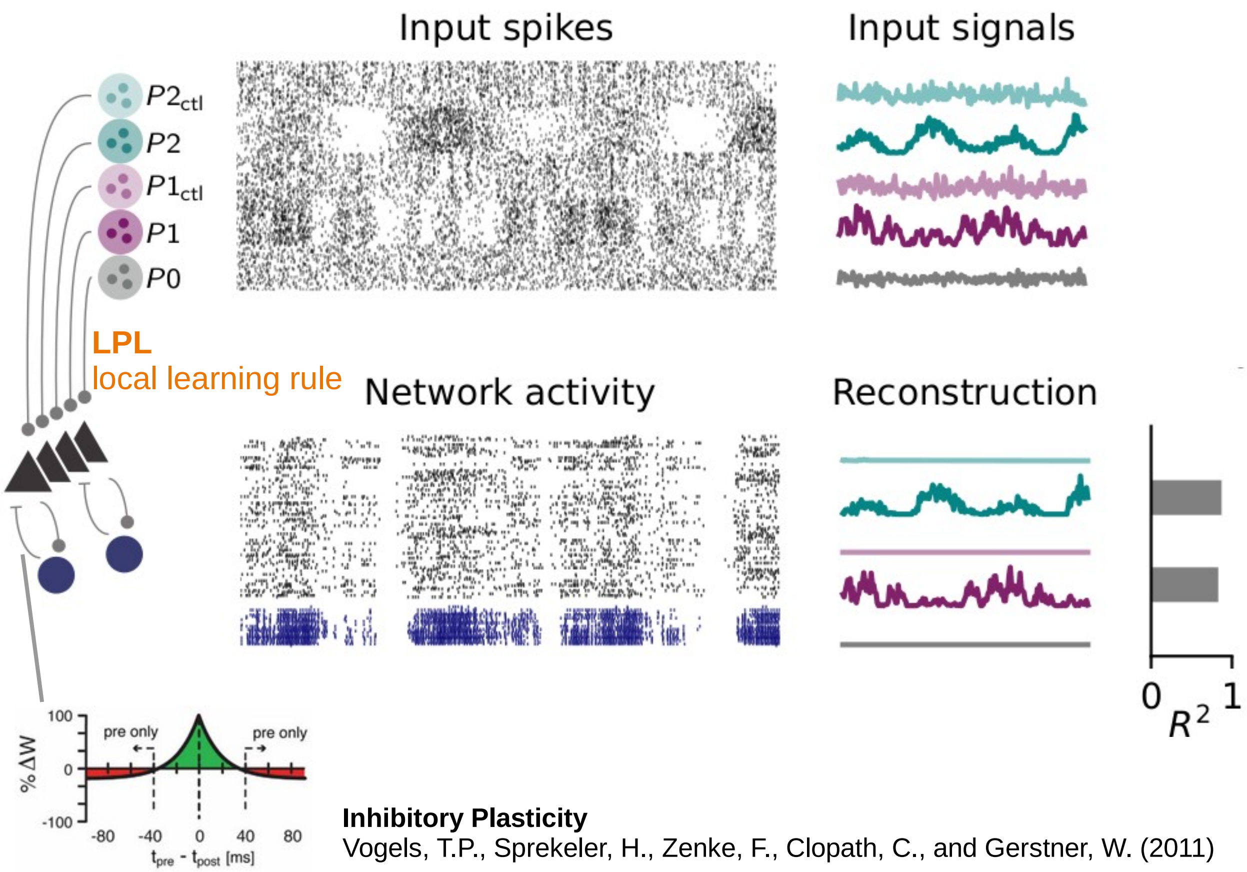
$$S_i(t)$$

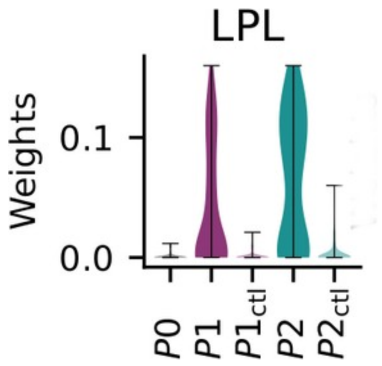
$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \mathcal{L}_{\mathrm{Hebb}}$$

+ inhibitory neurons & plasticity

$$\frac{\mathrm{d}w_{ij}}{\mathrm{d}t} = \eta\, \alpha * \Bigg(\underbrace{\epsilon * S_j(t)}_{\text{pre}} \underbrace{f'(U_i(t))}_{\text{post}} \Bigg) \Bigg[\alpha * \Bigg(\underbrace{-\left(S_i(t) - S_i(t-\Delta t)\right)}_{\text{predictive}} + \underbrace{\frac{\lambda}{\sigma_i^2 + \xi} \left(S_i(t) - \bar{S}_i(t)\right)}_{\text{Hebb}} \Bigg) \Bigg] + \eta \underbrace{\delta S_j(t)}_{\text{transmitter-triggered}}$$

LPL learns interesting features in streaming data

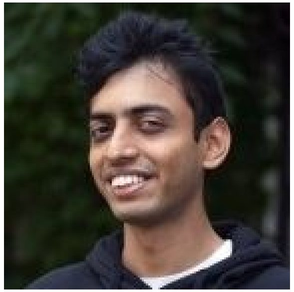




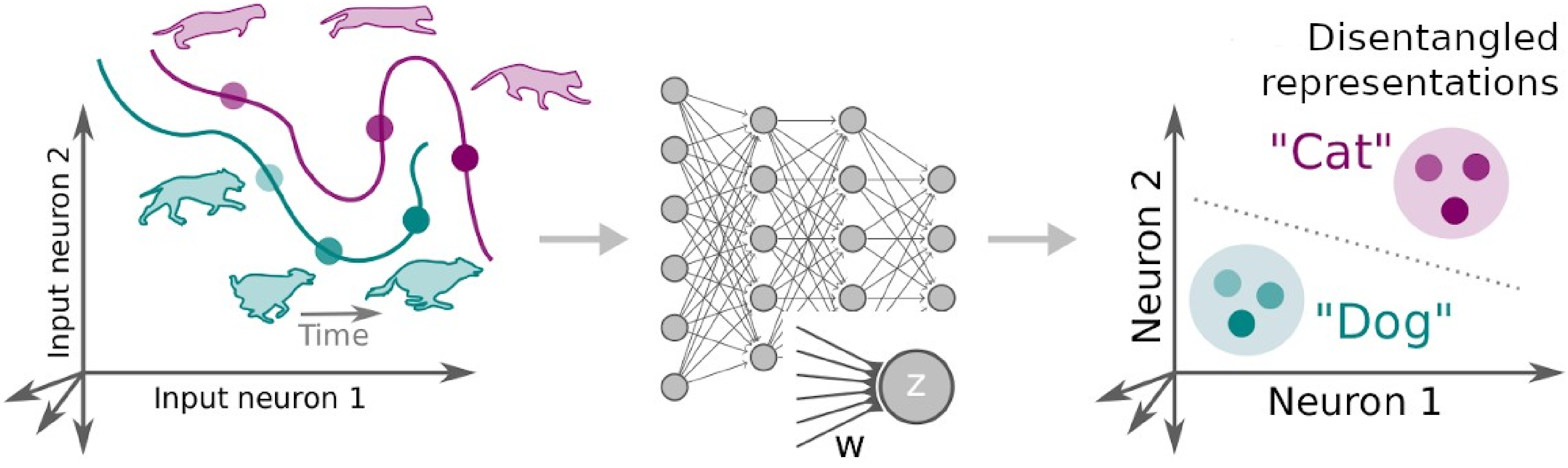
Inhibitory Plasticity
Vogels, T.P., Sprekeler, H., Zenke, F., Clopath, C., and Gerstner, W. (2011)

Latent Predictive Learning: Enables online learning with a local rule without supervision

(also works in spiking nets)



Manu S. Halvagal



Halvagal and Zenke (2023) *Nat Neurosci*

Summary

- Surrogate gradients allow training spiking neural networks end to end Neftci, Mostafa, and Zenke (2019) *IEEE SPM*

- Surrogate gradients can self-calibrate analog neuromorphic substrates and deal with device mismatch
Cramer, B., Billaudelle, S., Kanya, S., Leibfried, A., Grübl, A., Karasenko, V., Pehle, C., Schreiber, K., Stradmann, Y., Weis, J., Schemmel, J., and Zenke, F. (2022). *PNAS*

- Holomorphic Equilibrium Propagation allows training noisy analog substrates without Backprop
Laborieux and Zenke (2022) *Neurips*

- Latent predictive learning enables online learning without supervision
Halvagal and Zenke (2023) *Nature Neuroscience*

Way forward for online learning for Edge AI

- Need lightweight online learning rules:
  - Must be robust to noise and heterogeneity
  - No Backprop please!
  - Algorithms like holomorphic EP are promising, but must be practical (real-numbered)
  - Close the gap between self-supervised and supervised learning.
- Need joint efforts in algorithms and hardware development

Thanks!

The team https://zenkelab.org/team/

Axel Laborieux

Mattias Nilss`on

Manu Srinath Halvagal

Julian Rossbroich

Julia Gygax

Peter Buttaroni

Jeremias Seitz

Atena Mohammadi

Tengjun Liu

Aaron Spieler

INTERNATIONAL FMI PhD PROGRAM

Collaborators

Benjamin

Sebastian

Johannes Schemmel
Uni Heidelberg

Emre Neftci
FZ Julich

Andreas Luthi (FMI), Georg Keller (FMI), Rainer Friedrich (FMI),
Emre Neftci (FZ Julich), Johannes Schemmel (Uni Heidelberg),
Richard Naud (uOttawa), Blake Richards (Mila)

Funding

Swiss National Science Foundation