Self-sustained probabilistic computing on spike-based neuromorphic systems

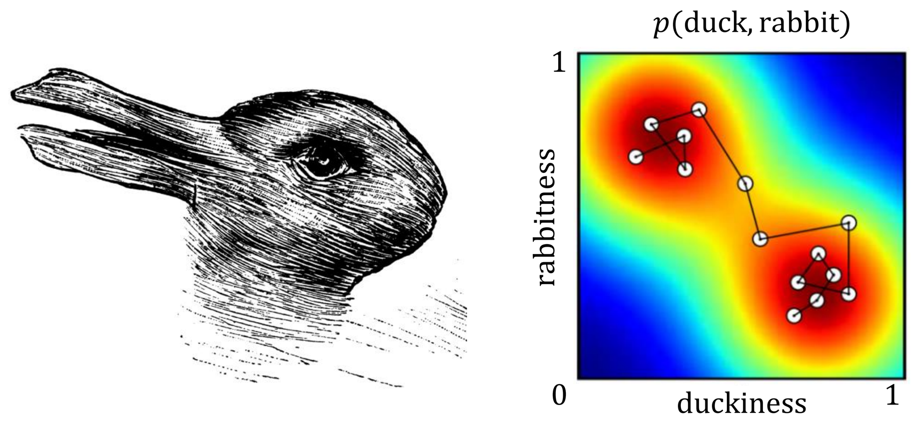
A. F. Kungl[1,2], D. Dold[1,2], A. Baumbach[1,2], S. Schmitt[1], J. Klähn[1], N. Gürtler[1], P. Müller[1], A. Kugele[1], L. Leng[1,2], E. Müller[1], C. Koke[1], M. Kleider[1], C. Mauch[1], O. J. Breitwieser[1], M. Güttler[1], D. Husmann[1], K. Husmann[1], A. Hartel[1], V. Karasenko[1], J. Ilmberger[1], A. Grübl[1], J. Schemmel[1], K. Meier[1], M. A. Petrovici[1,2]

[1]Heidelberg University, Kirchhoff-Institute for Physics; [2]University of Bern, Institute for Physiology

The Manfred Stärk Foundation

1 The Bayesian brain

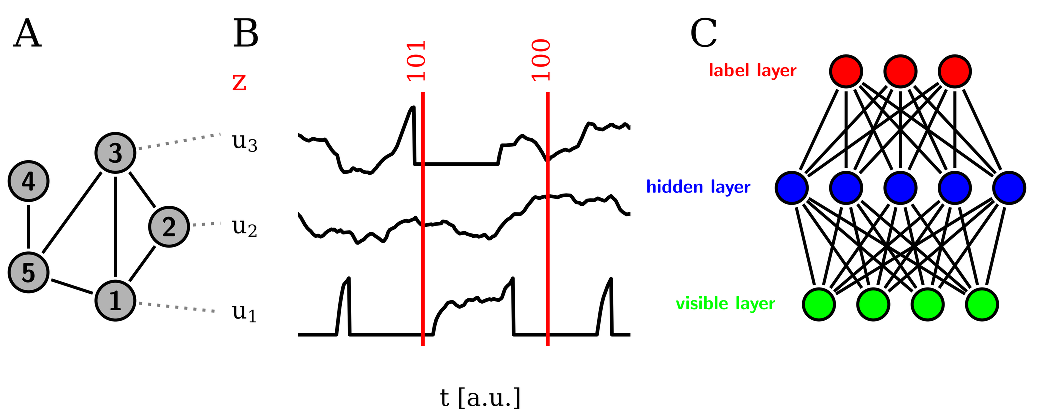
The noisy behavior of cortical neurons might be a hallmark of an underlying stochastic computation scheme. Such a scheme would enable the brain to cope with ambiguous inputs and offers an explanation for behavioral effects like bistable images (duck/rabbit) as sampling from different modes of a posterior distribution. The **neural sampling hypothesis** [1] proposes that some cortical areas implement **sampling-based Bayesian inference**.



These models are of particular interest for physical model systems, which face similar challenges as the brain. **We review two works [2, 3], in which we deployed stochastic spiking networks as robust and flexible models on analog neuromorphic hardware.**

2 Sampling with spikes

For applications on spiking hardware, we require models that explicitly treat and use spiking neural networks.



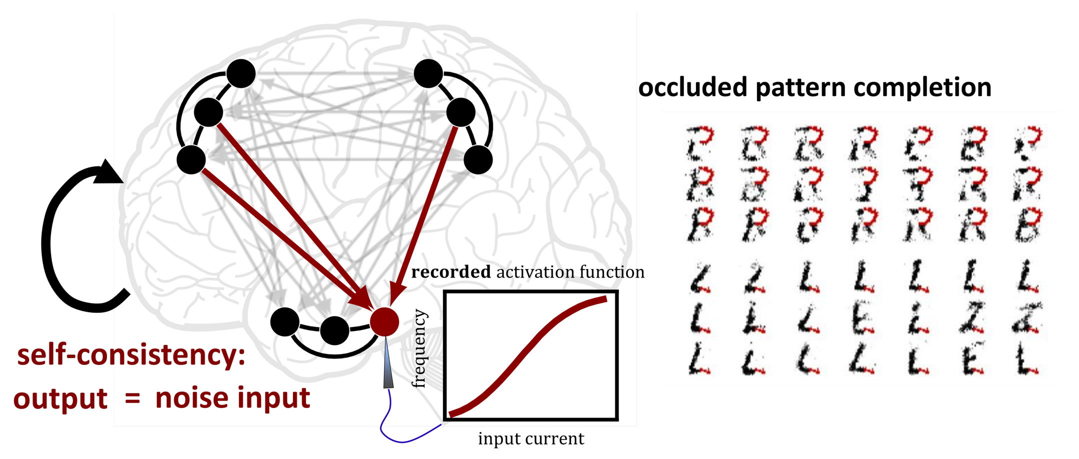
In the **LIF sampling** framework [4] a single neuron describes a binary random variable based on its spiking behavior (Fig-A-B). Immediately after a spike the neuron is in the *on-state* and otherwise in the *off-state*. The network approximately samples from a Boltzmann distribution over binary random variables z_i:

$$p(z) \sim \frac{1}{Z} \exp\left(\frac{1}{2}\sum_{ij} W_{ij} z_i z_j + \sum_i b_i z_i\right) . \qquad (1)$$

This framework establishes an imporant connection to Boltzmann machines. For practical applications we use a **hierarchical sampling network** (Fig-C) inspired by restricted Boltzmann machines [5].

3 Deterministic sampling

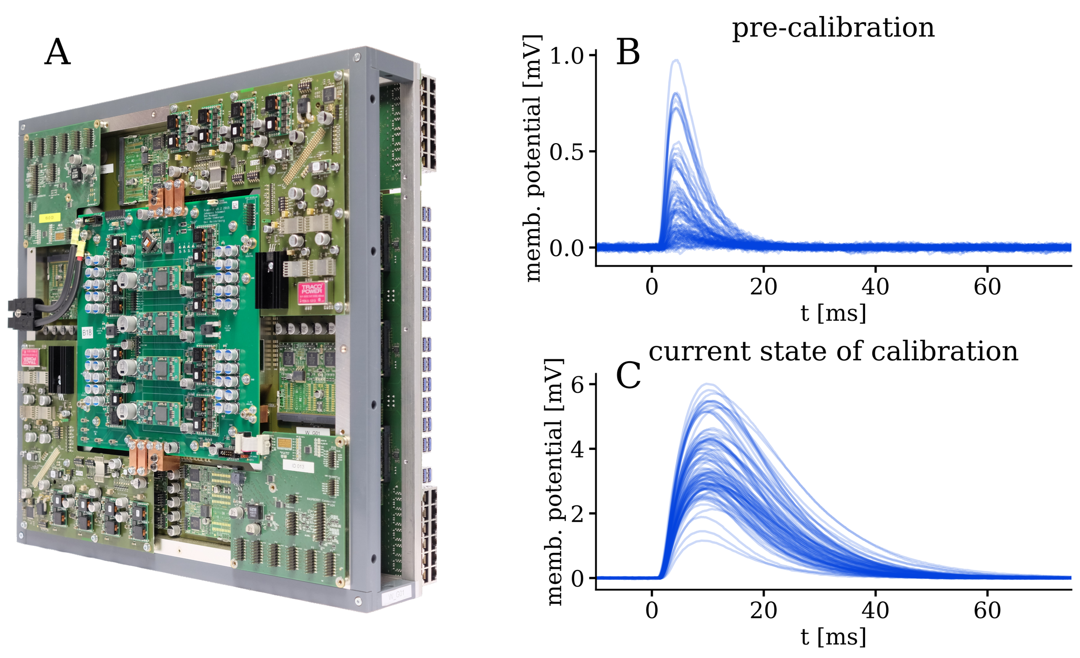
In most models of cortical networks, temporal variability is introduced using explicit white noise sources. This is, however, problematic because i) the **background activity of other brain areas is not necessarily white noise** and ii) a neuromorphic implementation would require dedicated, uncorrelated noise sources for every neuron.



We found [2] that **an ensemble of dynamically fully deterministic, but functionally probabilistic networks** can learn a connectivity pattern that enables probabilistic computation with a degree of precision that matches the one attainable with idealized, perfectly stochastic components (Fig-left). The key element of this construction is self-consistency, i.e., all input activity seen by a neuron is the result of output activity of other neurons that fulfill a functional role in their respective subnetworks (Fig-right).

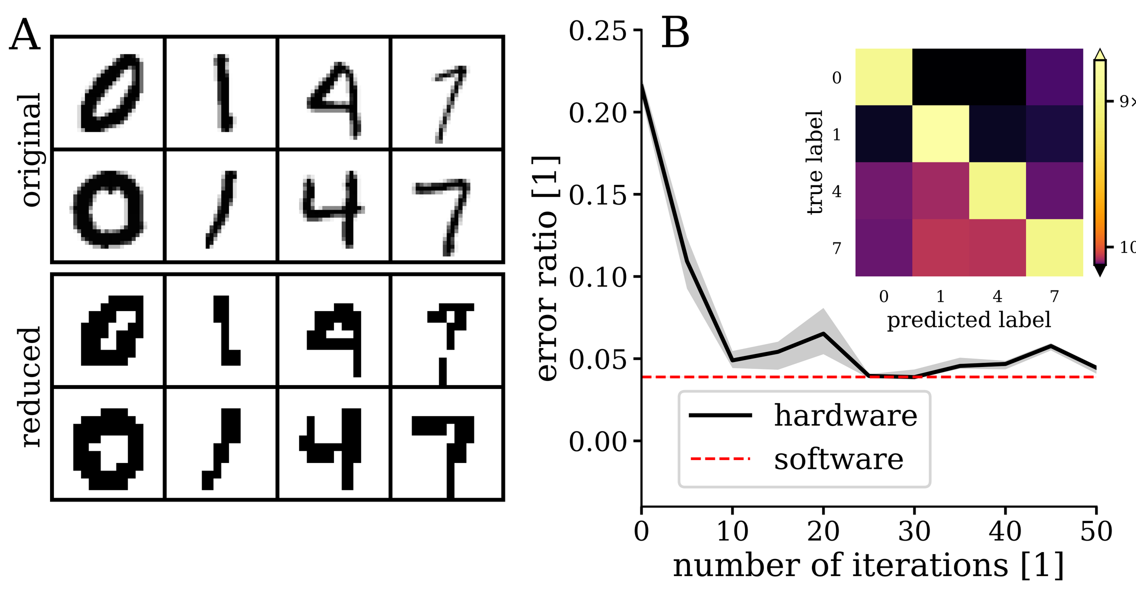
4 The BrainScaleS system

On a single module of the **BrainScaleS [6] analog neuromorphic hardware** (Fig-A) the physical model of **200k neurons and 40 million synapses** is implemented using CMOS technology. The system follows the principle of **physical modeling**: it uses the dynamics of the underlying substrate to implement computation.

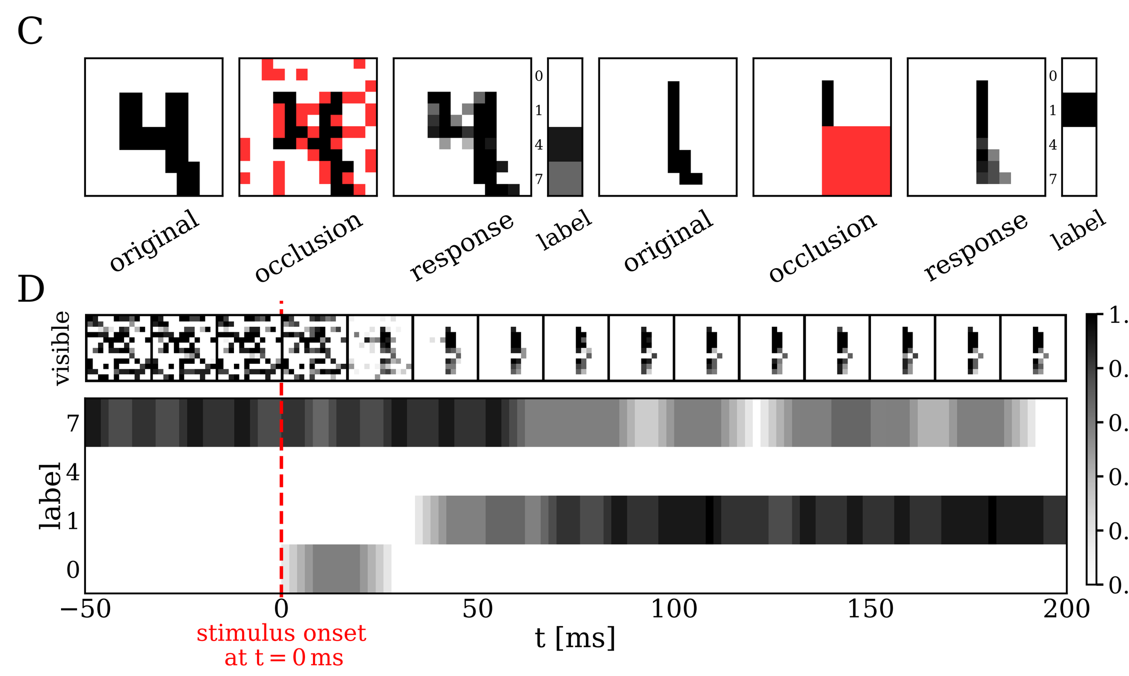


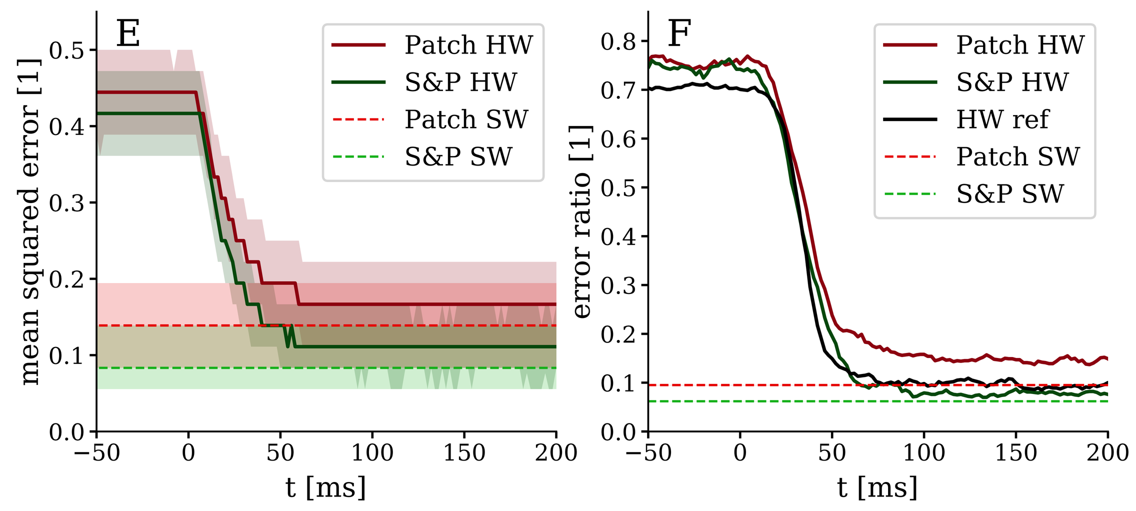
As such it can emulate networks of spiking neurons with **10^4-fold speed-up** compared to biological realtime, but suffers from considerable variability of neuron parameters (Fig-B-C). Hence, we require robust network dynamics and learning rules.

5 Use-case on neuromorphic hardware



Using the LIF sampling framework we implemented a **restricted Boltzmann machine** (RBM) [5] on the BrainScaleS system [3]. We evaluate the model on a reduced version of the MNIST dataset. The original pictures were binarized, reduced to 12×12 pixels and the digits 0,1,4 and 7 were selected (Fig-A).





We map an RBM that was pretrained on a computer to BrainScaleS and perform *in-the-loop training* to compensate for the model and substrate imperfections. The **classification** rate recovers software level performance after $O(10)$ training steps (Fig-B). The implemented model is able to **complete partially occluded images** while predicting the label correctly (Fig-C-F). Finally, it is able to **generate recognizable images** if the respective label is clamped (Fig-G).

References

[1] J. Fiser, P. Berkes, G. Orbán, and M. Lengyel, "Statistically optimal perception and learning: from behavior to neural representations," *Trends in cognitive sciences*, vol. 14, no. 3, pp. 119–130, 2010.
[2] D. Dold, I. Bytschok, A. F. Kungl, A. Baumbach, O. Breitwieser, W. Senn, J. Schemmel, K. Meier, and M. A. Petrovici, "Stochasticity from functionwhy the bayesian brain may need no noise," *Neural networks*, vol. 119, pp. 200–213, 2019.
[3] A. F. Kungl, S. Schmitt, J. Klähn, P. Müller, A. Baumbach, D. Dold, A. Kugele, E. Müller, C. Koke, M. Kleider, *et al.*, "Accelerated physical emulation of bayesian inference in spiking neural networks," *Frontiers in neuroscience*, vol. 13, p. 1201, 2019.
[4] M. A. Petrovici, J. Bill, I. Bytschok, J. Schemmel, and K. Meier, "Stochastic inference with spiking neurons in the high-conductance state," *Physical Review E*, vol. 94, no. 4, p. 042312, 2016.
[5] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, *Boltzmann machines: Constraint satisfaction networks that learn*. Carnegie-Mellon University, Department of Computer Science Pittsburgh, PA, 1984.
[6] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on*, pp. 1947–1950, IEEE, 2010.

fkungl@kip.uni-heidelberg.de