Fast generative models on an accelerated neuromorphic physical model system

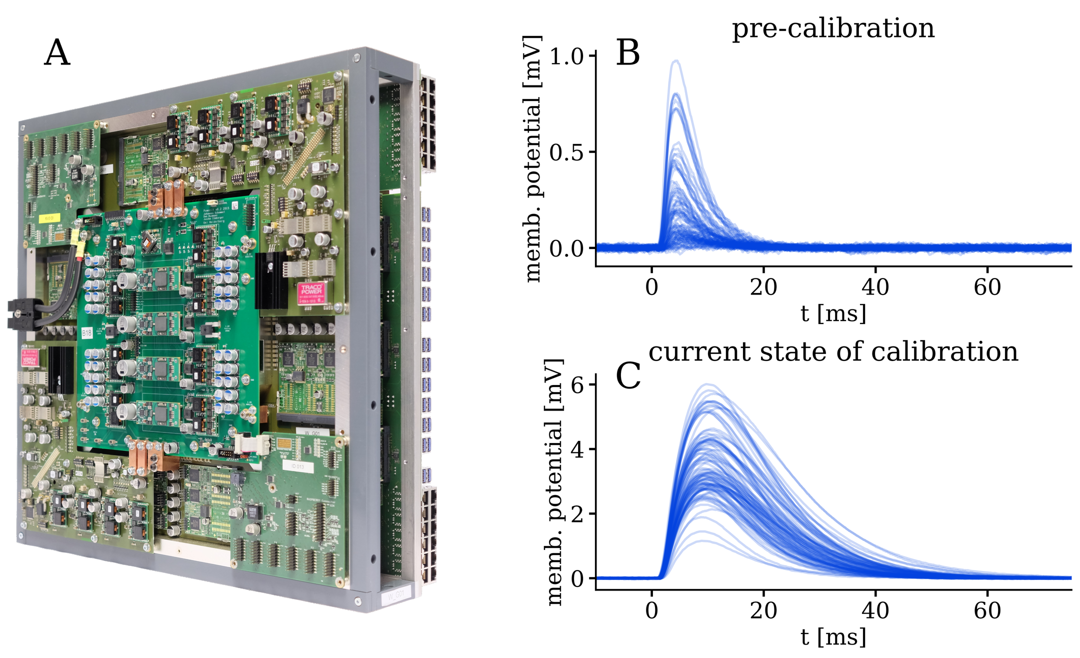
A. F. Kungl[1,2], A. Baumbach[1,2], S. Schmitt[1], J. Klähn[1], N. Gürtler[1], P. Müller[1], D. Dold[1,2], A. Kugele[1], L. Leng[1,2], E. Müller[1], C. Koke[1], M. Kleider[1], C. Mauch[1], O. J. Breitwieser[1], M. Güttler[1], D. Husmann[1], K. Husmann[1], A. Hartel[1], V. Karasenko[1], J. Ilmberger[1],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

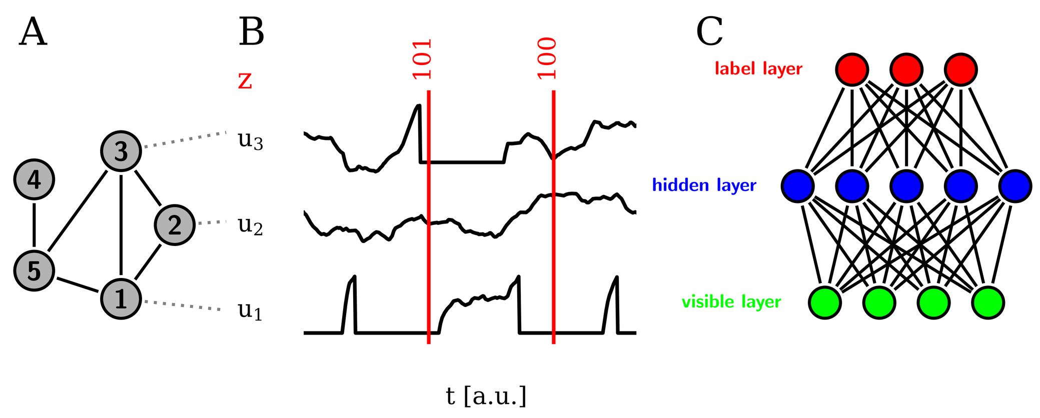
On a single module of the **BrainScaleS [1] analog neuromorphic hardware** (Fig-A) the physical model of **200k neurons and 40 million synapses** is implemented using CMOS technology. The system follows the principle of **physical modeling**: it uses the dynamics of the underlying substrate to implement computation.



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

2 Sampling as robust coding

According to the **neural sampling hypothesis** [2] certain cortical areas implement **sampling based probabilistic inference**. These models are of particular interest for physical model systems as the brain faces similar challenges.



In the **LIF sampling** framework [3] a single neuron describes a binary random variable based on its spiking behavior (Fig-A-B). The network approximately samples from a Boltzmann distribution over binary random variables:

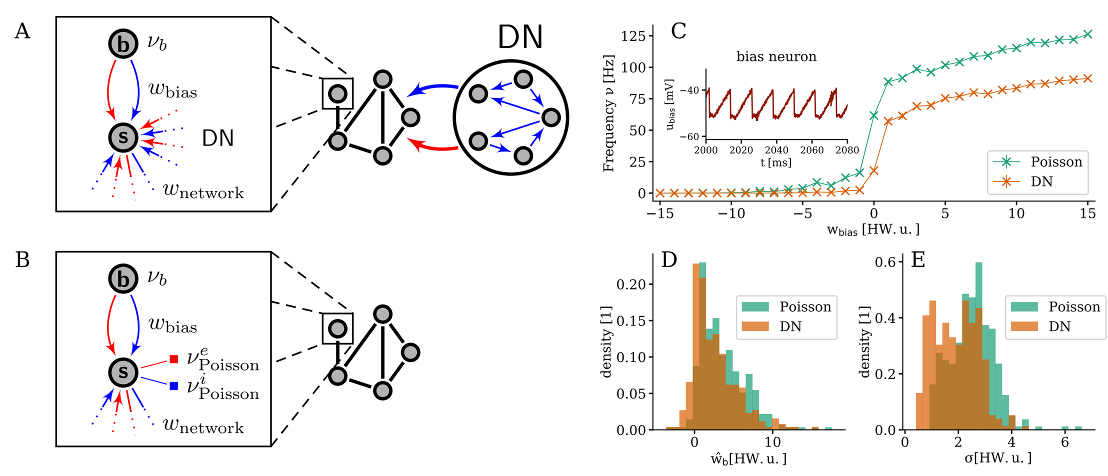
$$p(z) = \frac{1}{Z} \exp\left(\frac{1}{2}\sum_{ij} W_{ij} z_i z_j + \sum_i b_i z_i\right)$$

$$Z = \sum_{z \in \{0,1\}^N} \exp\left(\frac{1}{2}\sum_{ij} W_{ij} z_i z_j + \sum_i b_i z_i\right)$$

For practical applications we use a **hierarchical sampling network** (Fig-C) inspired by restricted Boltzmann machines [4].

3 From theory to hardware

We adapted the LIF sampling theory to the characteristics of the hardware. A random inhibition-dominated **decorrelation network (DN)** provides the necessary stochasticity to the sampling network (Fig-A). This way we realized a fully **self-contained sampler** on the accelerated substrate. As a comparison we implemented the setup with external Poisson noise (Fig-B).



The hardware constraints lead to distortions in the activation function from the ideal sigmoid shape (Fig-C-E); but this can be accounted for by in-the-loop training.

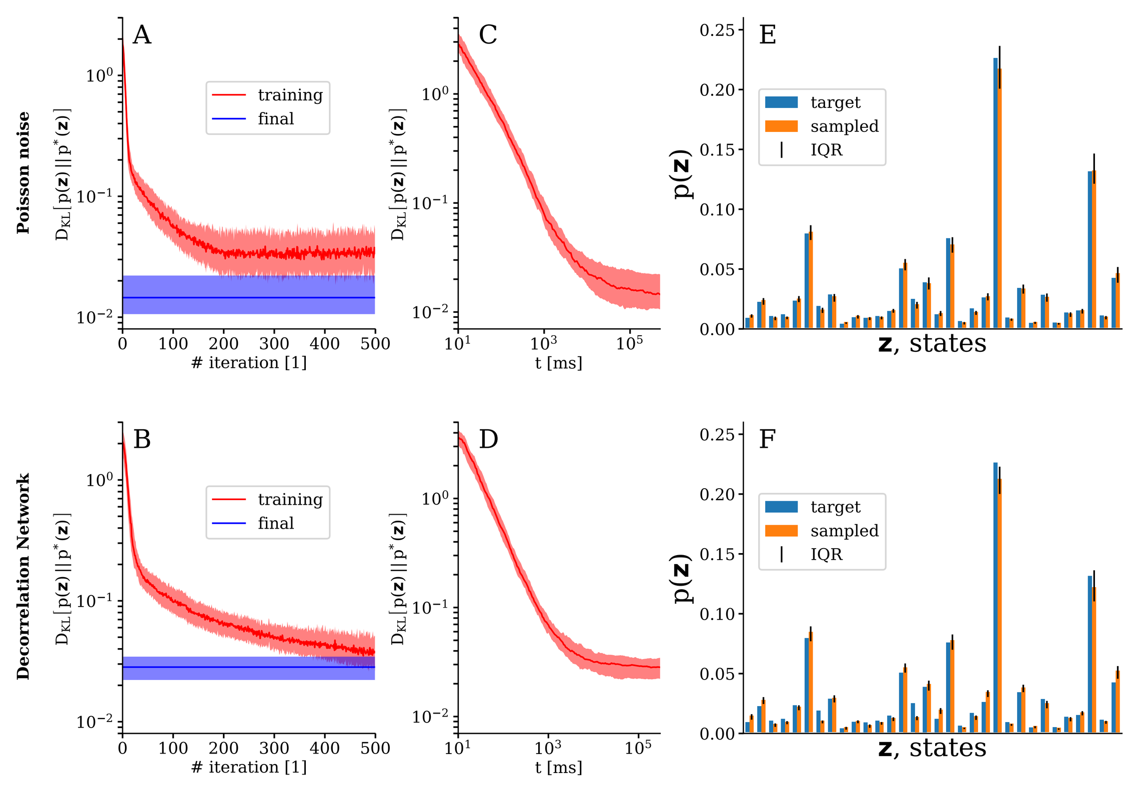
4 In-the-loop training

We trained a 5 neuron network to **sample from a target probability distribution**. We used the **wake-sleep algorithm** [5]:

$$\begin{aligned} \Delta b_i &= \eta(\langle z_i \rangle_{data} - \langle z_i \rangle_{model}) \\ \Delta W_{ij} &= \eta(\langle z_i z_j \rangle_{data} - \langle z_i z_j \rangle_{model}) \end{aligned} \tag{1}$$

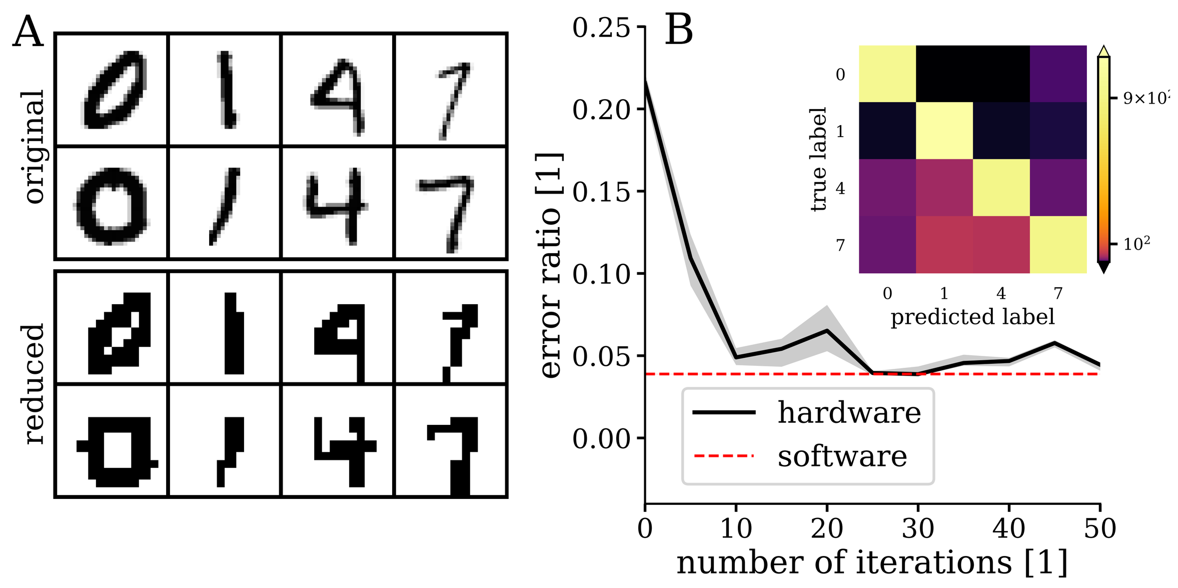
where the data phase is given by the target distribution.

We trained the networks in an **in-the-loop** manner: Parameter updates are calculated on the host computer with eq. (1) based on the spike times measured on hardware.

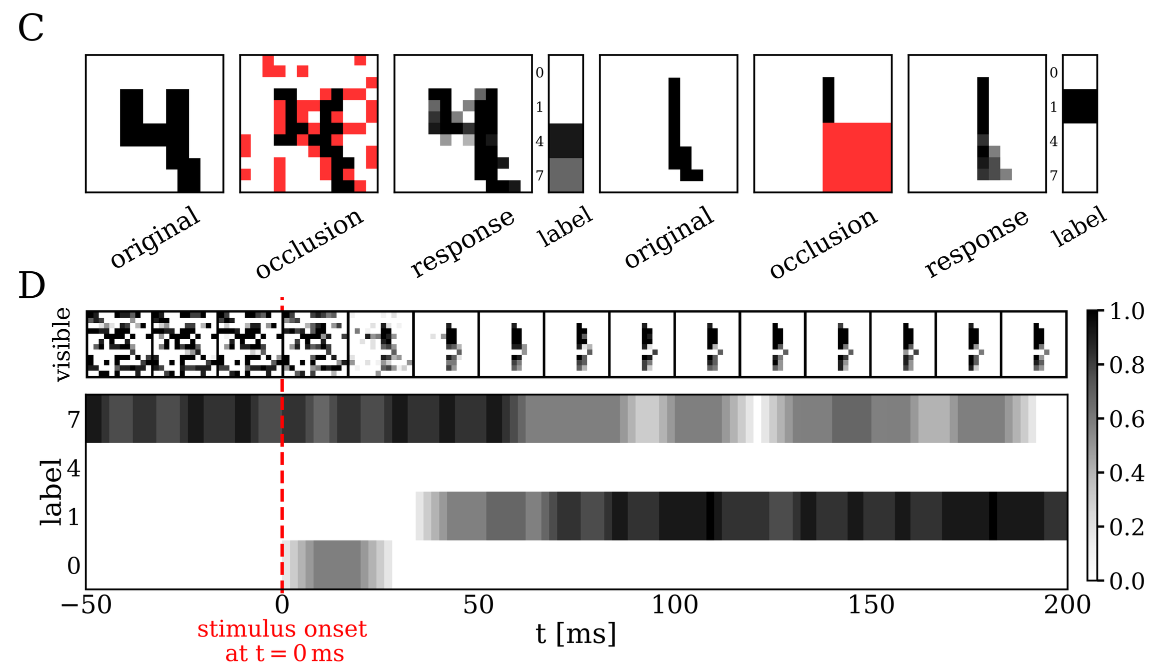


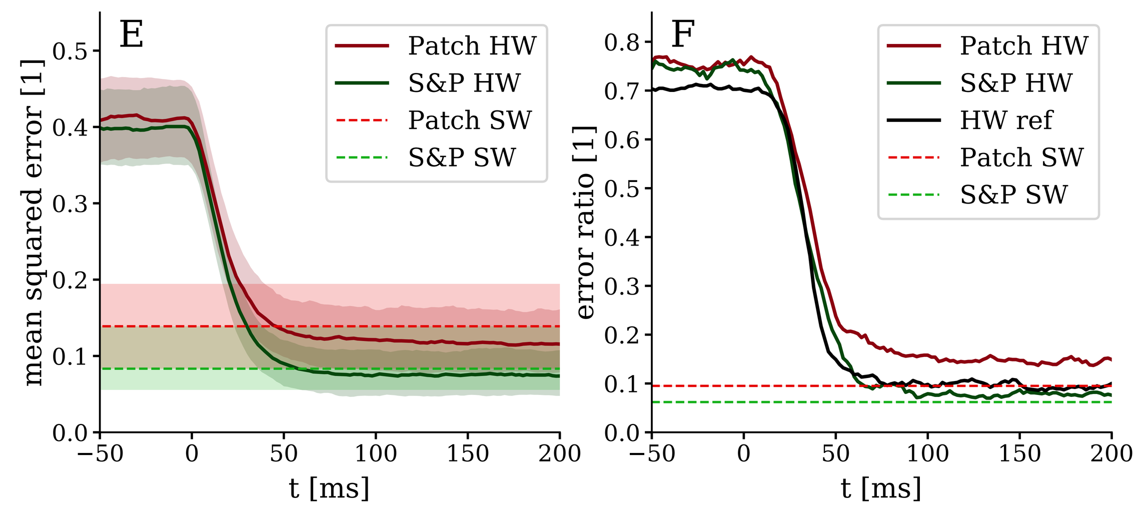
The **training reduced the Kullbeck-Leibler Divergence (DKL)** between the sampled and the target distribution (Fig-A-F). Due to the acceleration of the hardware **the full training scheme (including overhead) is two orders of magnitude faster than biological real-time.**

5 Application to handwritten digits

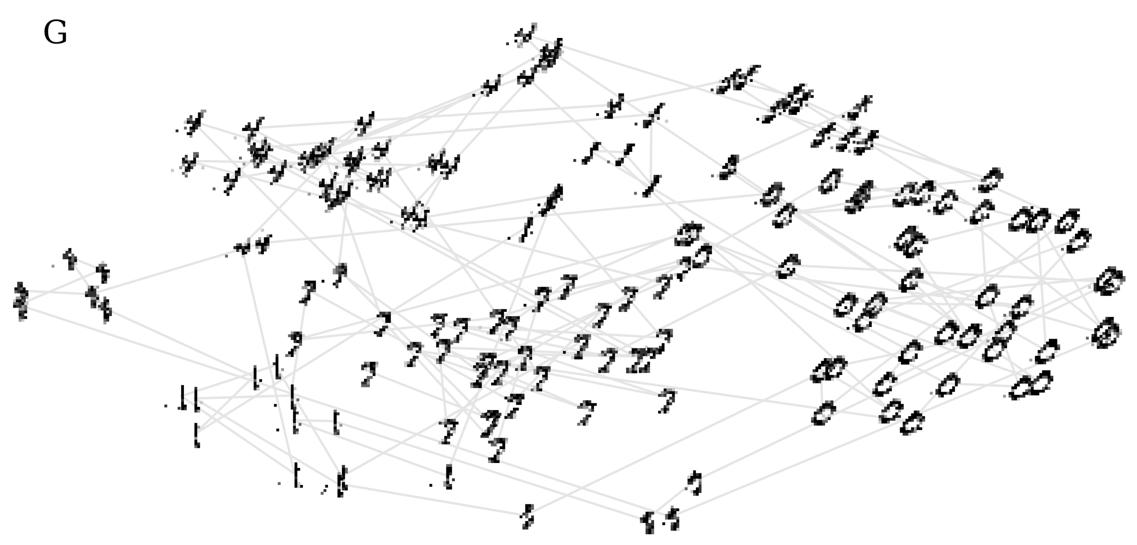


Using the LIF sampling framework we implemented a **restricted Boltzmann machine** (RBM) [4] on the BrainScaleS System. We evaluate the model on a reduced version of the MNIST dataset [6]. The original pictures were binarized, reduced to 12×12 pixels and the digits 0,1,4 and 7 were selected (Fig-A).





We use an on the host computer pretrained RBM 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. 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.

[2] 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.

[3] 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.

[4] 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.

[5] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," in *Readings in Computer Vision*, pp. 522–533, Elsevier, 1987.

[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

BrainScaleS guidebook

publication

fkungl@kip.uni-heidelberg.de