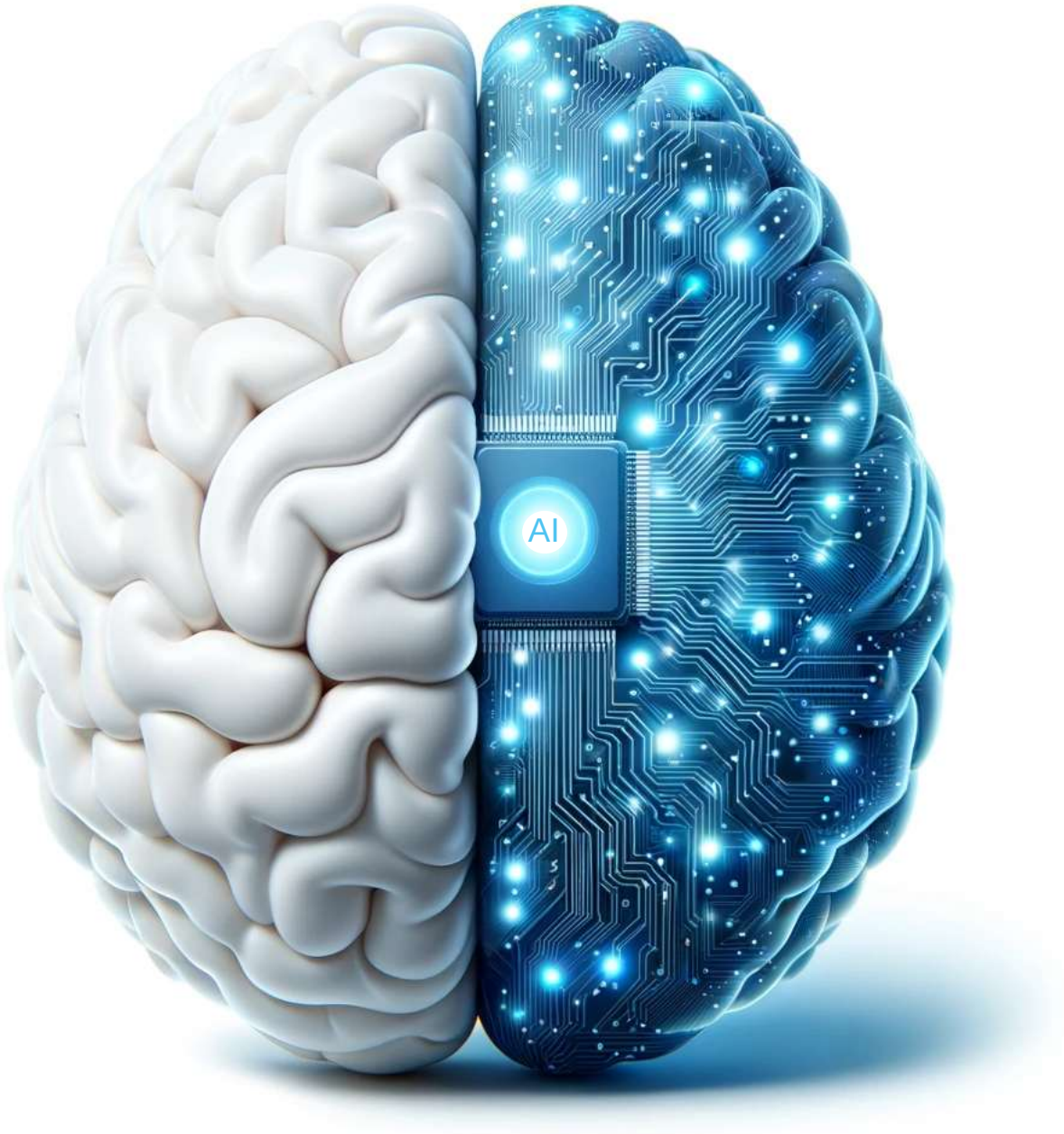
SpiNNcloud

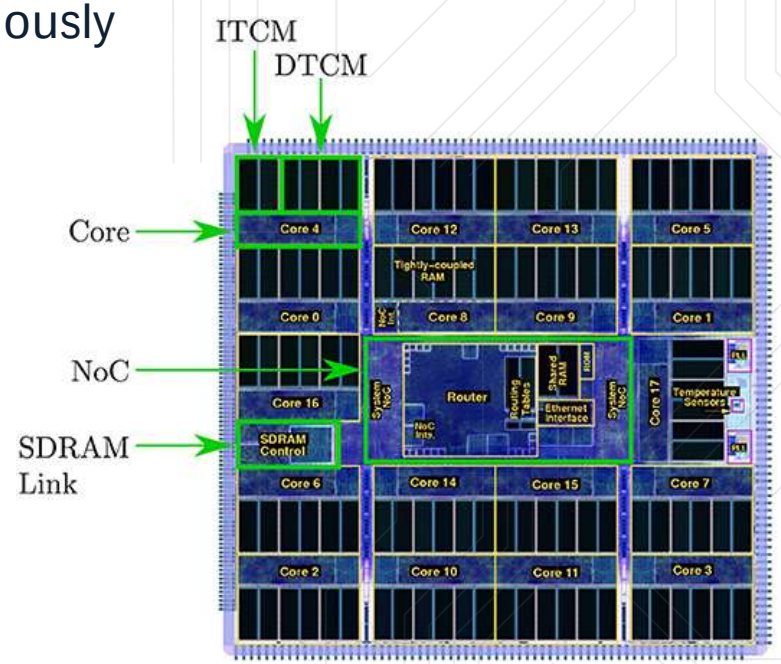
Brain-inspired computing: Systems for the next generation of AI

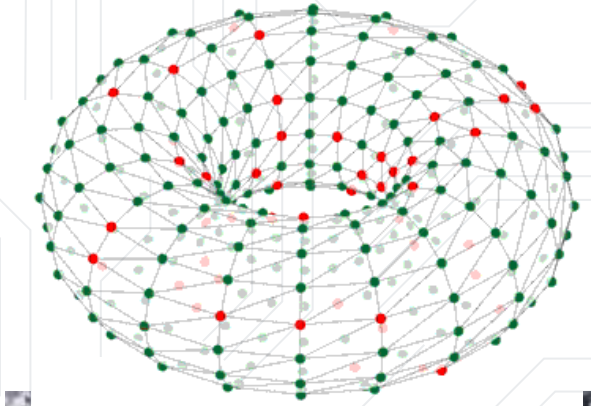


Matthias Lohrmann
CTO

Redefining Brain-inspiration: SpiNNaker (1)

- Do spiking neurons digitally & asynchronously
- 18 ARM 968 processors / chip
- 1 router / chip with world model
- 1M core supercomputer built



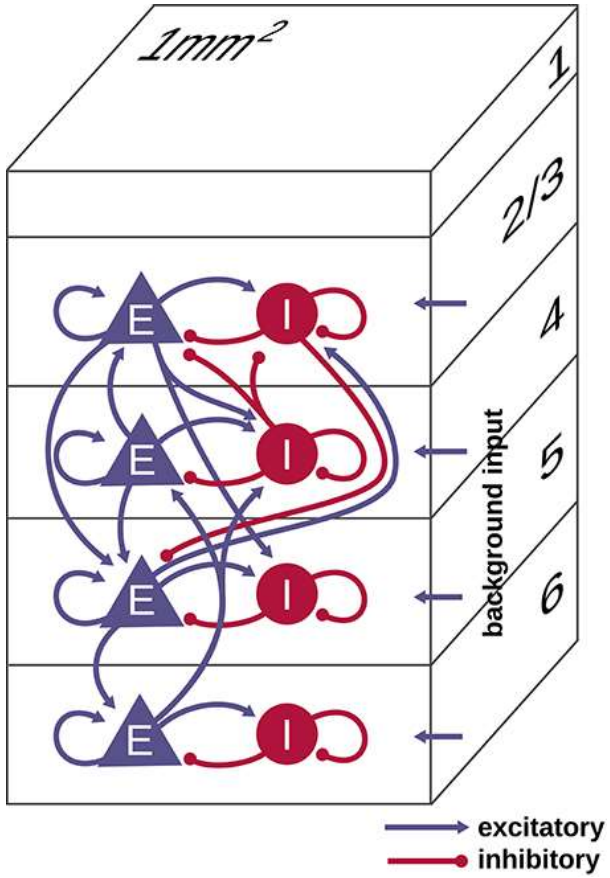


http://apt.cs.manchester.ac.uk/projects/SpiNNaker/SpiNNchip/
http://apt.cs.manchester.ac.uk/projects/SpiNNaker/project/Access/

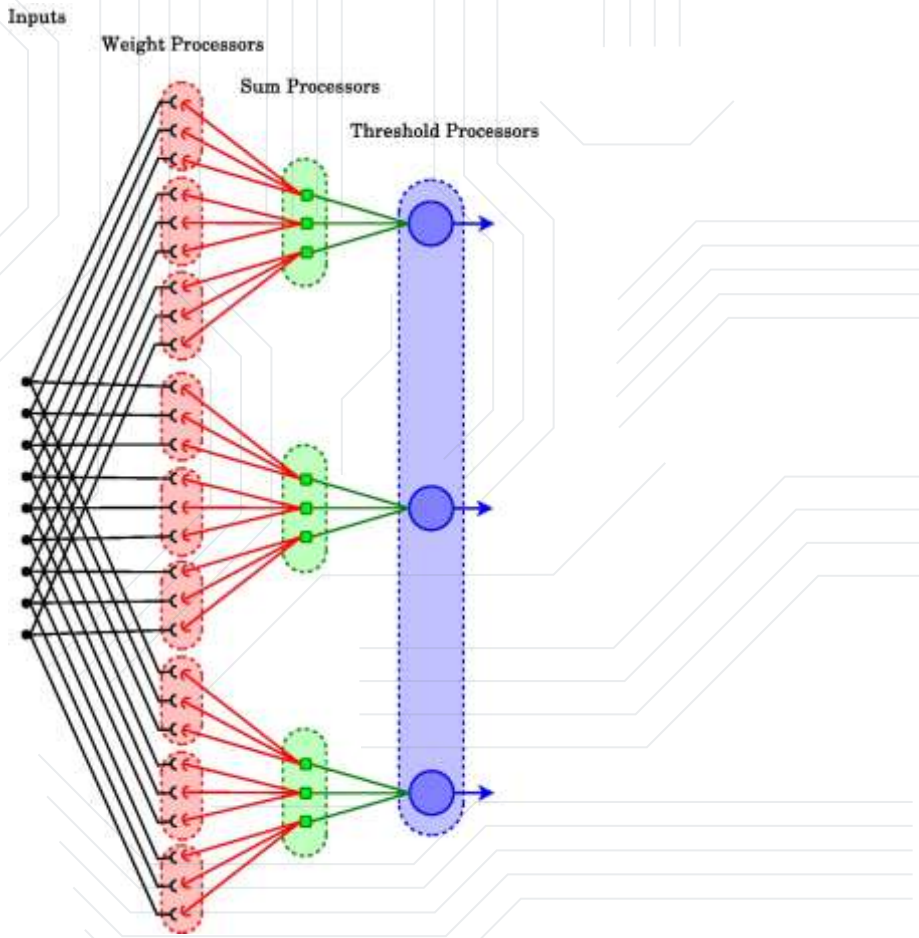
MANCHESTER 1824

SpiNNcloud

SpiNNaker 1



- Cortical Brain Model



- Multilayer Perceptron

Learnings for AI

Granularity

- MIMD
- hybrid systems
- exploit sparsity

Sparsity

- activation
- weights
- representation

Functionality

- memory dominates area
- compute is for free

Unique Hybrid Microchip SpiNNaker2

- **Faster** than NVIDIA's A100 in brain models
- Consumes 1/10th less power than GPUs
- Enables more-than-DNN AI systems

IP is backed by multiple international patents

SpiNNcloud

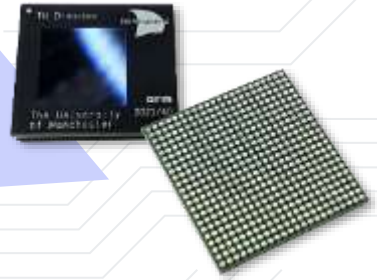
Unique Hybrid
Microchip SpiNNaker2

SpiNNcloud

Cloud Board

SpiNNode

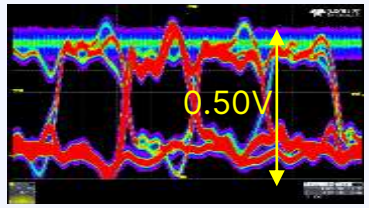
SpiNNaker2



IP is backed by multiple international patents

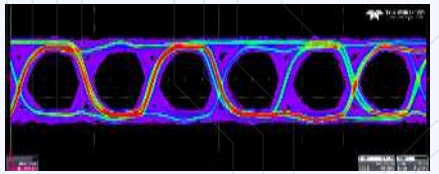
SpiNNcloud

The SpiNNaker2 Architecture



6x6 Serial Chip-2-Chip Links

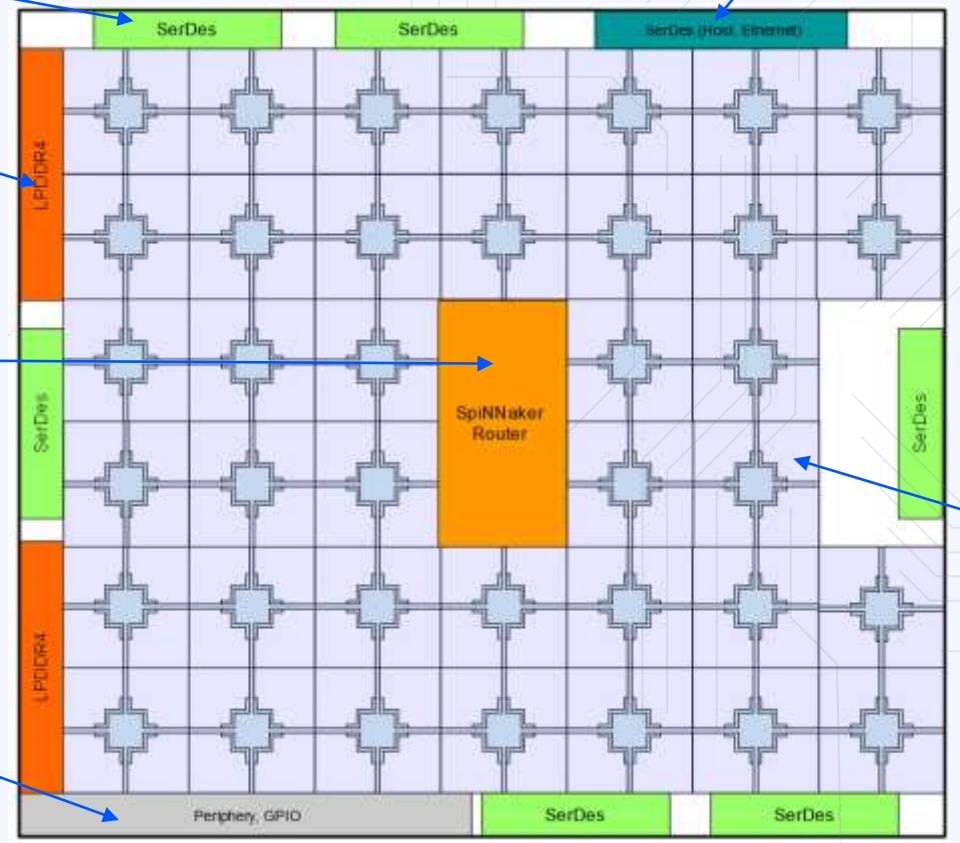
SerDes
- Board2Board coms
- Gbit Ethernet



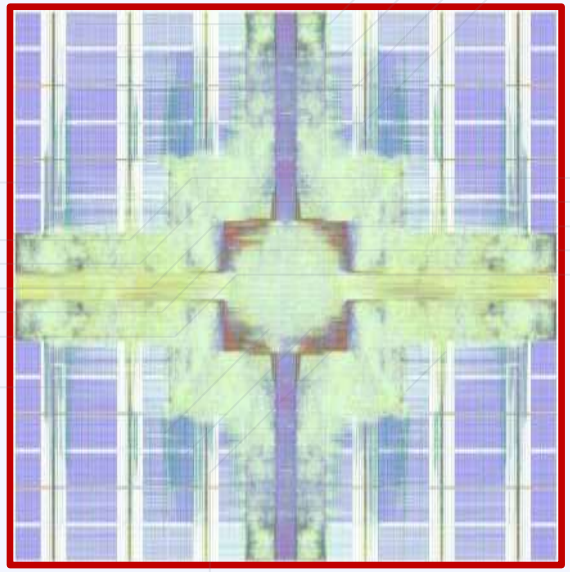
2x LPDDR4 Memory Interfaces

SpiNNaker Packet Router
- Lightweight event packet communication fabric

Periphery
- Flexible GPIO, QSPI, I2C (Master + Slave), JTAG



152 ARM M4F cores + accelerators



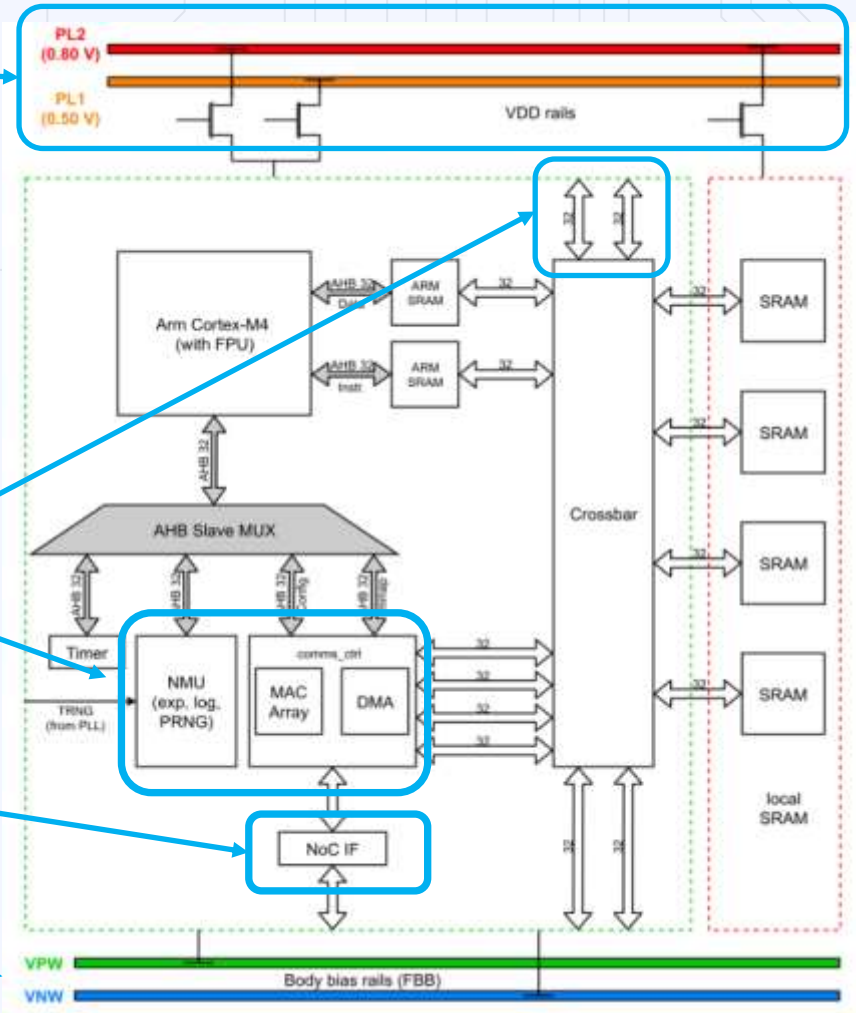
The SpiNNaker2 Architecture

Dynamic Power Management
- DVFS and PSO

Accelerators
- MAC + DMA
- Exp / Log
- Random numbers (PRNG, TRNG)

Communication
- Neighbour PE memory sharing (synchronous)
- On- and off-chip memory access
- Event & data handling

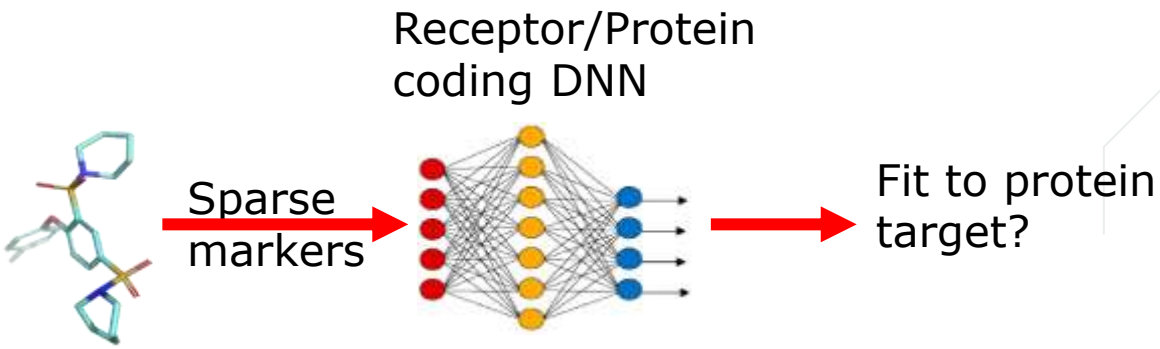
ABX Adaptive Body Biasing

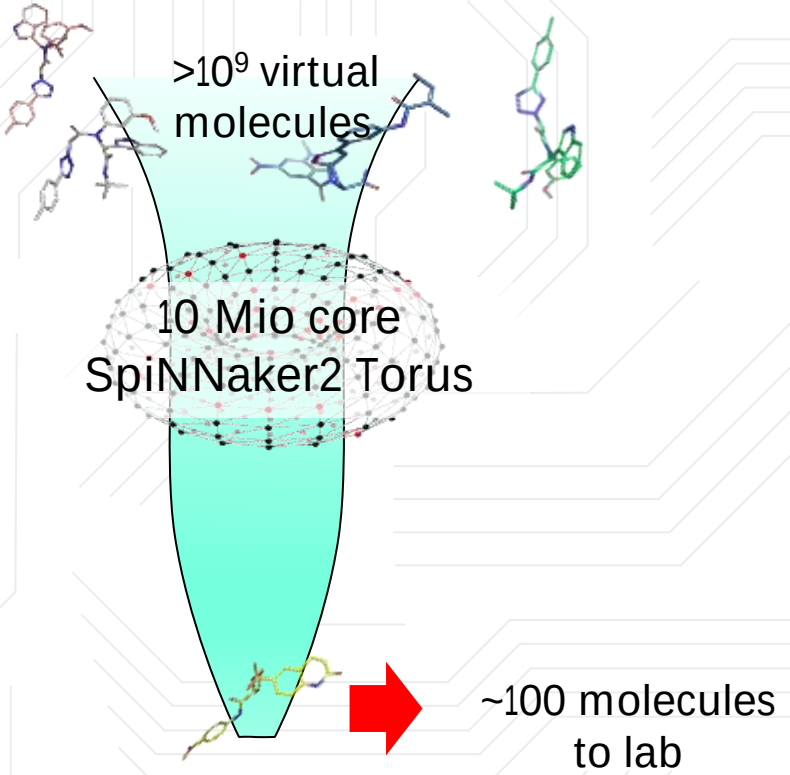


SpiNNcloud

Hybrid AI/non-AI Compute: Drug Discovery

- Target: 10^9 molecule scanning in O(hours) instead of O(weeks)
- Densely couple CPU + GPU for speed up





Granularity !

Efficient inference of large language models

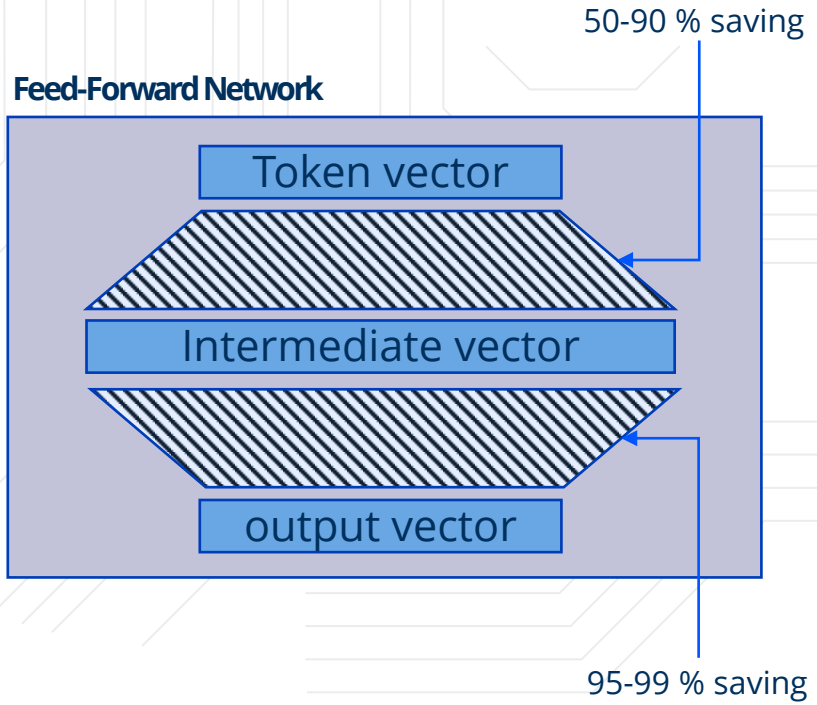
Efficient inference = Parameter Sparsity × Activation Sparsity

50 % savings (one-shot i.e. no retraining)

Frantar & Alistarh
"SparseGPT: Massive Language Models Can be Accurately Pruned in One-Shot"

75-95 % savings

Zhu et al.
"SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks"



Sparsity !

SpiNNcloud

Particle Swarm Evolutionary Algorithms

- Island-based, distributed genetic algorithms for optimization
- Combine the robust local search performance the global exploration power of PSO (particle-swarm optimization)

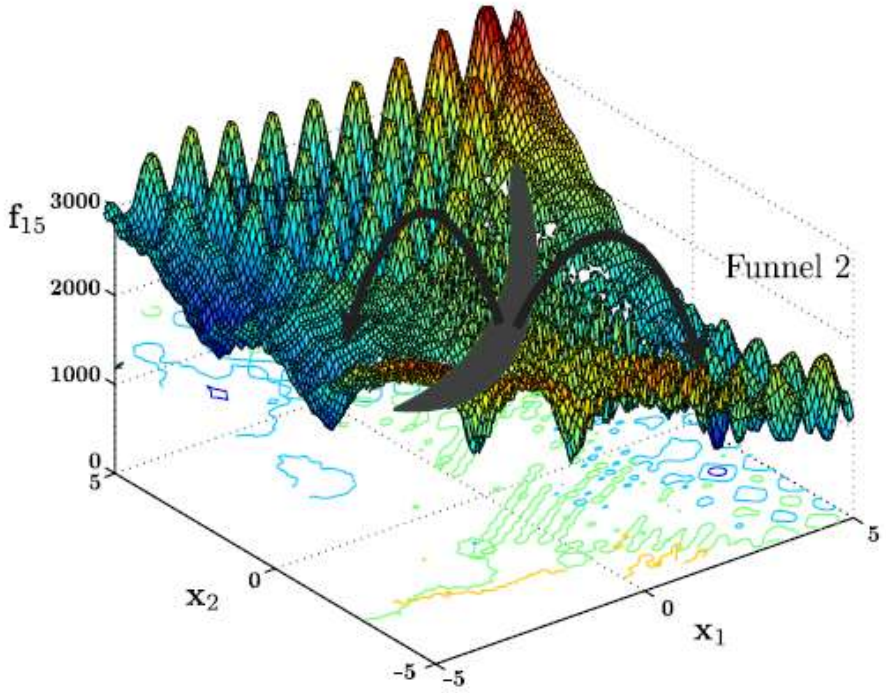
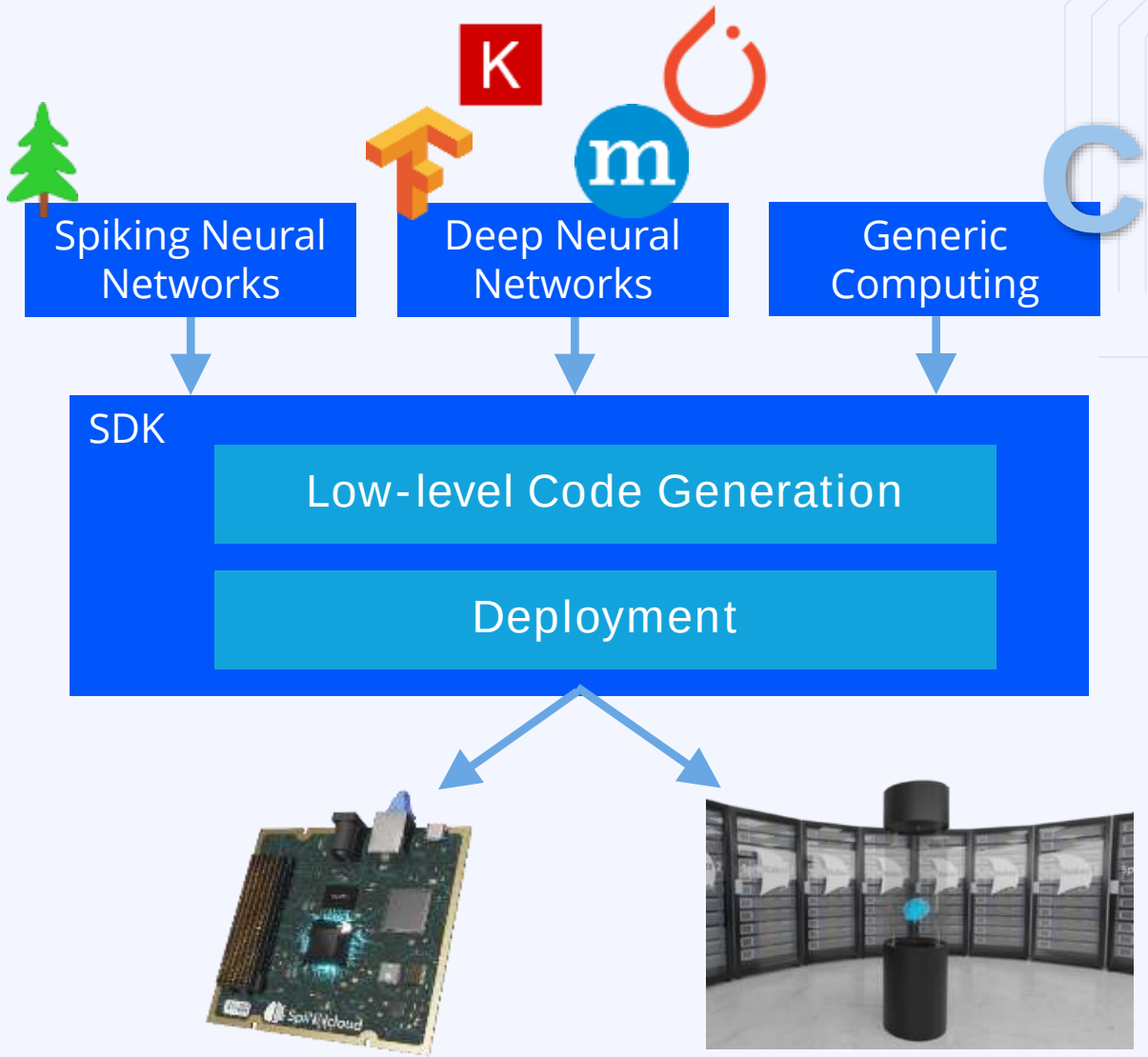


Fig. 1. Two-dimensional version of the highly dispersive function f15 from the CEC benchmark test suite [17]. The global topology is a double funnel separated by the central ridge region (in gray). The global and several local minima are contained in funnel 1, several deep local minima in funnel 2. This topology is hard since a search heuristic can be trapped in the broad funnel 2.

Functionality!

SpiNNcloud

Unified Software Stack

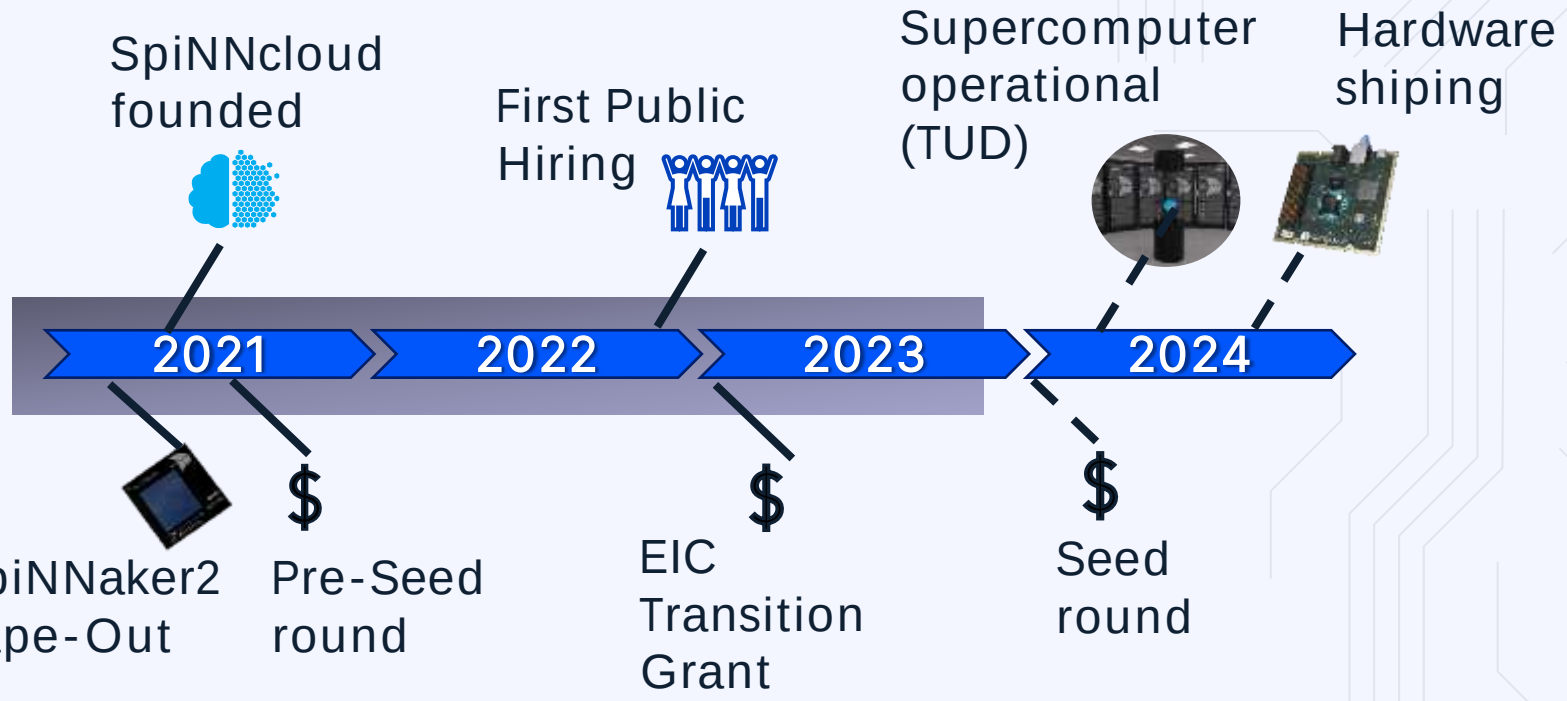


- Simple development + deployment with high flexibility
- Full suite available in Summer 2024

```
from spinnaker2 import snn, hardware

neuron_params = {
        "threshold":1.,
        "alpha_decay":0.9,
        }

stim = snn.Population(
        size=10,
        neuron_model="spike_list",
        params={0:[1,2,3], 5:[20,30]},
        name="stim")

pop1 = snn.Population(
        size=20,
        neuron_model="lif",
        params=neuron_params,
        name="pop1")
```

SpiNNcloud

About SpiNNcloud



Looking for

- Follow investors for seed round
- Compiler engineering talent
- Proof-of-concept partners

SpiNNcloud

Thank You!

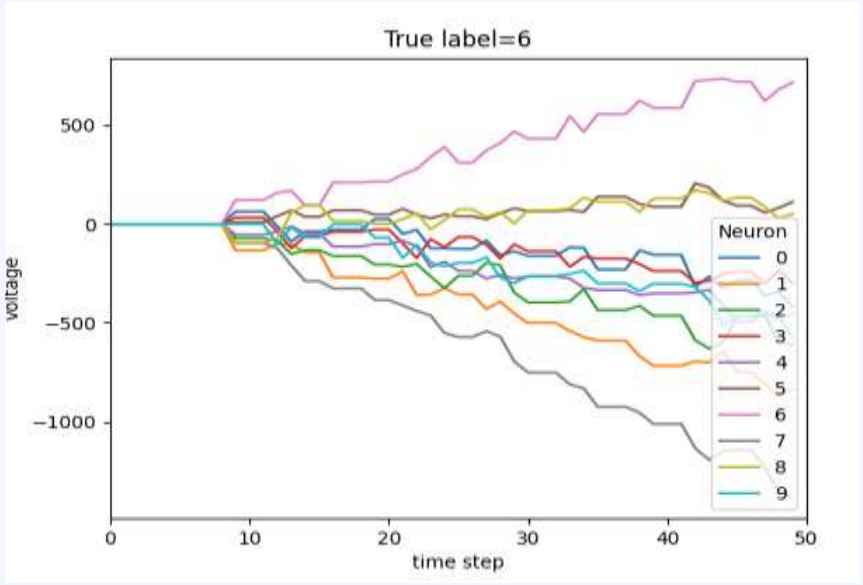
SpiNNcloud

Contact: matthias.lohrmann@spinncloud.com

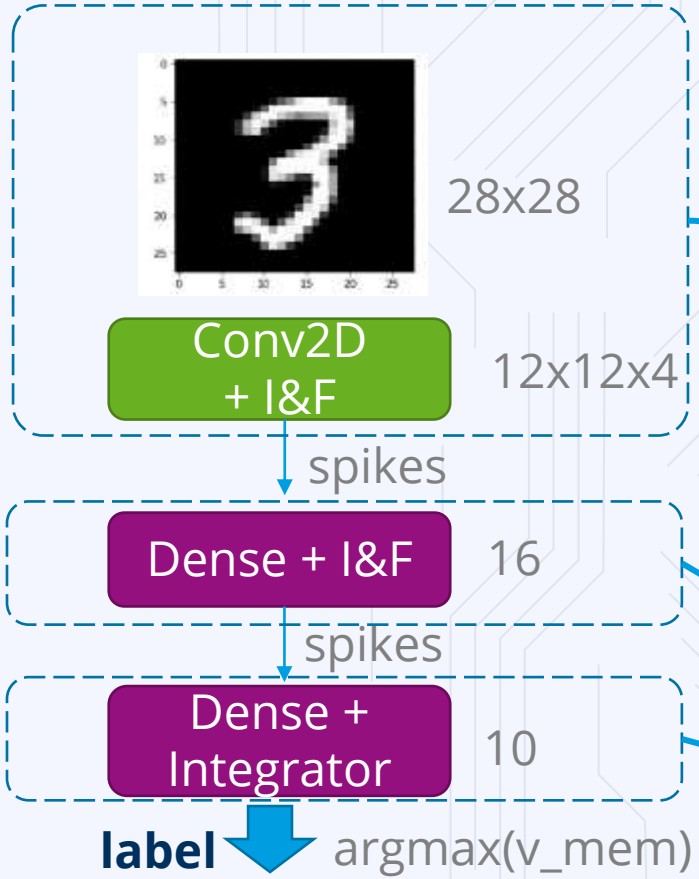
Unified Software Stack

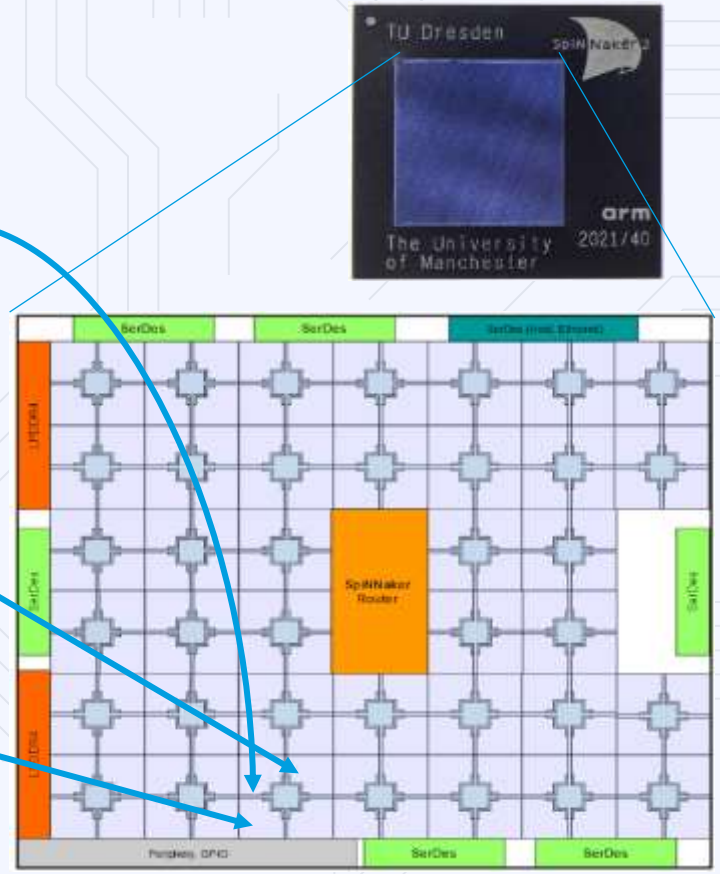
conv2d_if_neuron_rate

lif_neuron



Accuracy identical to DNN





SpiNNcloud

Particle Swarm Evolutionary Algorithms

- Island-based, distributed genetic algorithms for optimization

- Combine the robust local search performance the global exploration power of PSO (particle-swarm optimization)

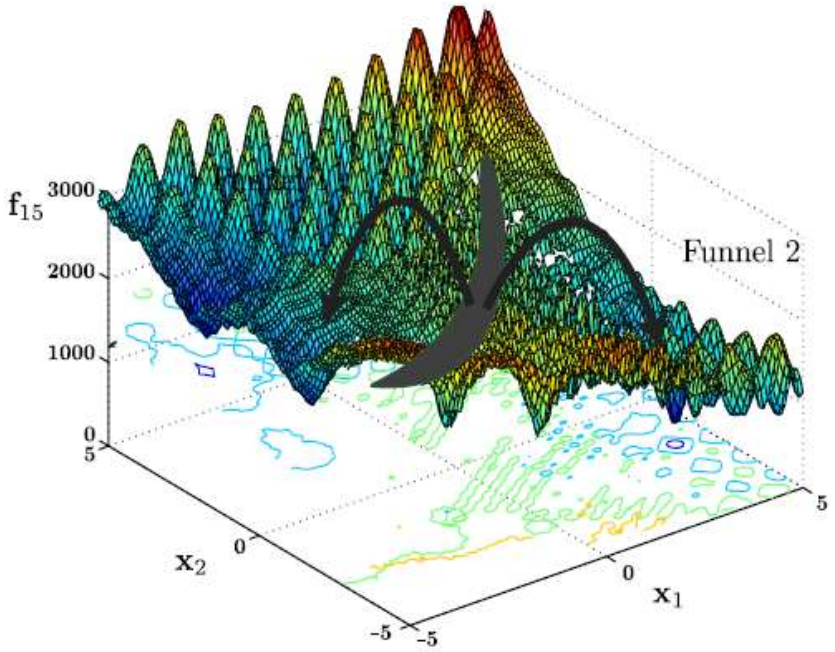
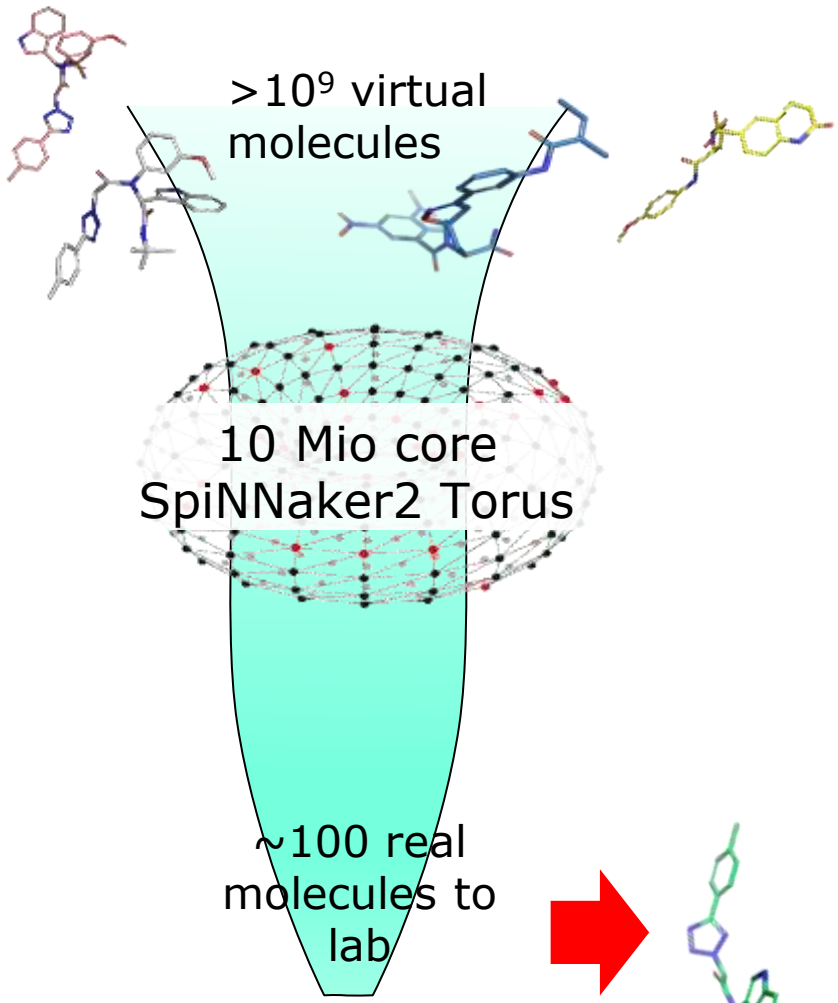
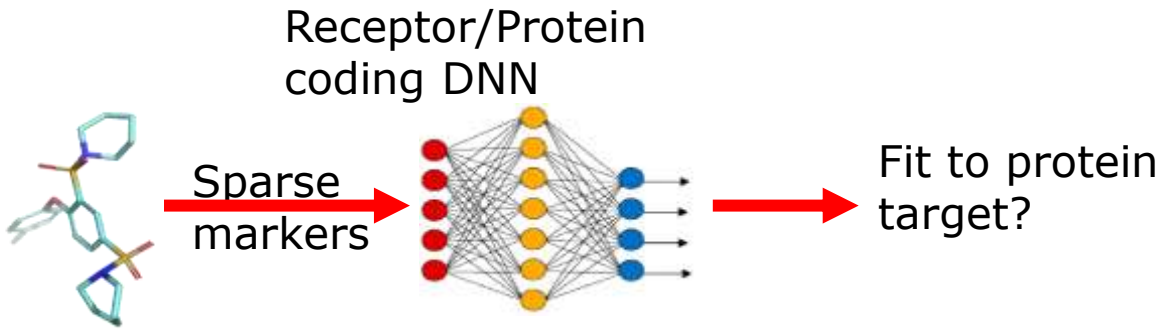


Fig. 1. Two-dimensional version of the highly dispersive function f15 from the CEC benchmark test suite [17]. The global topology is a double funnel separated by the central ridge region (in gray). The global and several local minima are contained in funnel 1, several deep local minima in funnel 2. This topology is hard since a search heuristic can be trapped in the broad funnel 2.

SpiNNcloud

Sbalzarini and Müller: Particle Swarm CMA Evolution Strategy for the Optimization of Multi-Funnel Landscapes, 2009 IEEE Congress on Evolutionary Computation





SpiNNcloud

Del Alamo, Sala, McHaourab, **Meiler**: "Sampling alternative conformational states of transporters and receptors with AlphaFold2"; *Elife*; **2022** || Brown, Mendenhall, Geanes, **Meiler**; "General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps"; *J Chem Inf Model*; **2021** ||