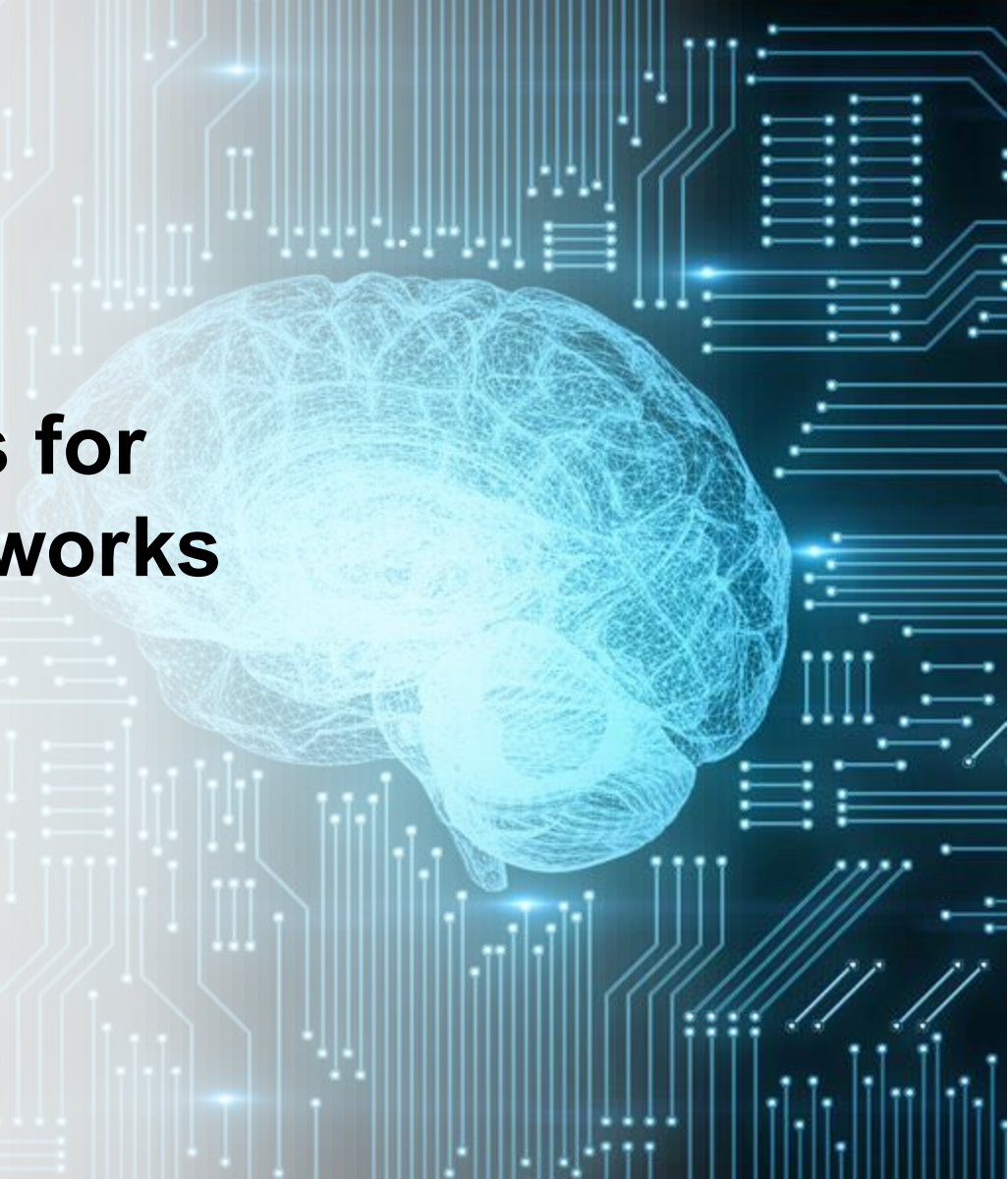
MANCHESTER 1824
The University of Manchester



An Overview: Tools for Spiking Neural Networks

Oliver Rhodes

15/1/2026

oliver.rhodes@manchester.ac.uk

MANCHESTER 1824
The University of Manchester

Overview

- Why model spiking neural networks?
- Building blocks of computational neuroscience
- Review of software simulators and applications

MANCHESTER 1824
The University of Manchester

The Brain

Process information
from a range of stimuli

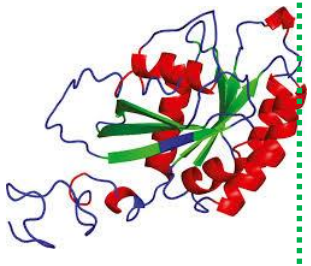
Learn and
memorise new
information

Energy
Efficient
(~20W)

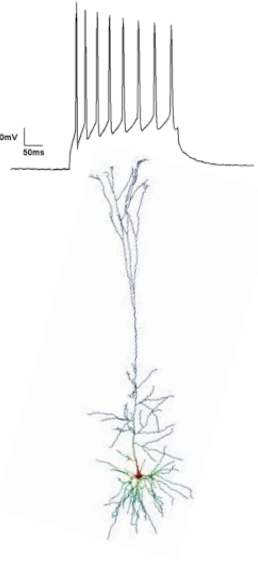
Last a lifetime

MANCHESTER 1824
The University of Manchester

Neural Modelling: Scales

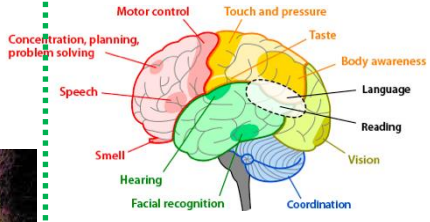


Neuromolecular Dynamics



Neurons

Neural Networks



Brain Regions

Functional Behaviour

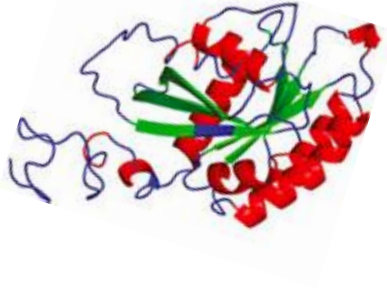
MANCHESTER 1824
The University of Manchester

Modelling Neurons and Circuits

- High performance computation and memory
- Robust and energy efficient

- Top-down: bridge scales from functional behaviour
  - Learning and Memory
  - Brain disease
  - Motor function

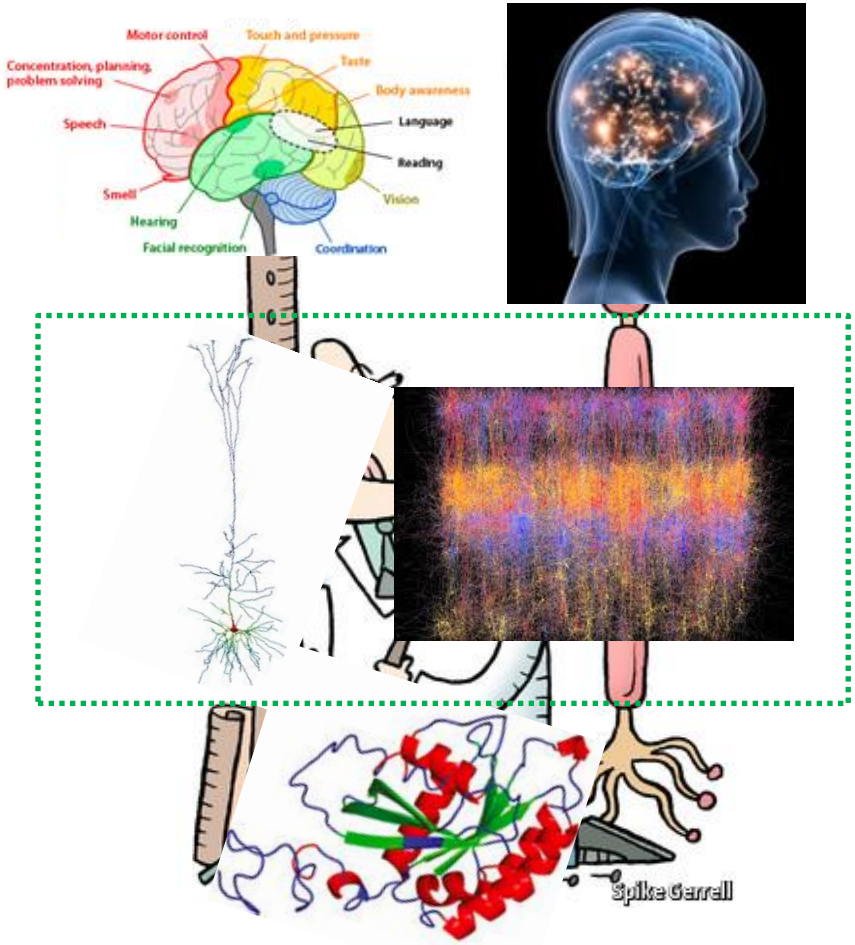
- Bottom-up: explore origins of neural dynamics and function
  - Test hypotheses infeasible in vivo
  - Treat neural disorders
  - Understand intelligence



MANCHESTER 1824
The University of Manchester

Modelling Challenges

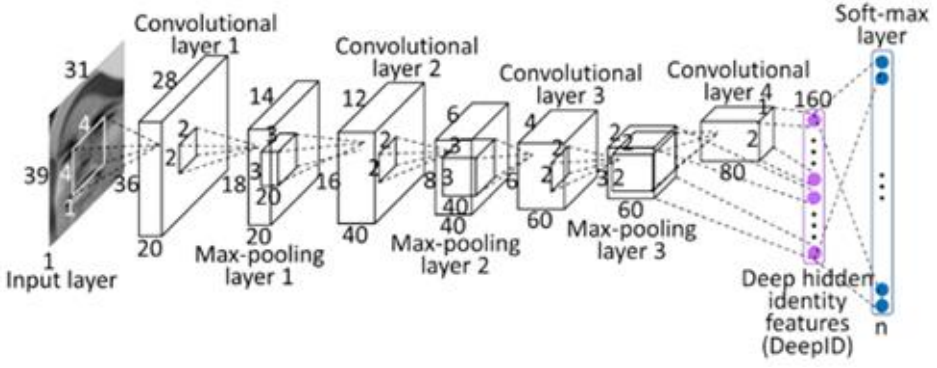
- Large scale
  - 100 billion neurons
  - 100 trillion synapses
- Circuit connectivity is a computational challenge
  - Unlike traditional physics problems (large variations between neurons)
  - Massively parallel
  - Long-range connectivity
- Neuromorphic Computing

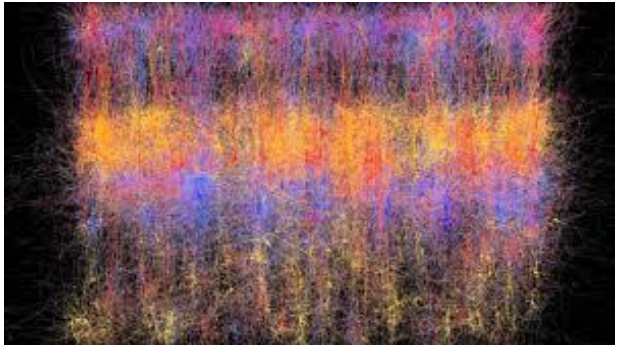


https://www.economist.com/science-and-technology/... 03/the-machine-of-a-new-soul

MANCHESTER 1824
The University of Manchester

Neural Networks





Artificial Neural Networks

- Dense convolution kernels
- Abstract neurons, continuously active, feed-forward connections
- Training: error-backpropagation + SGD
- Synchronous, shared memory

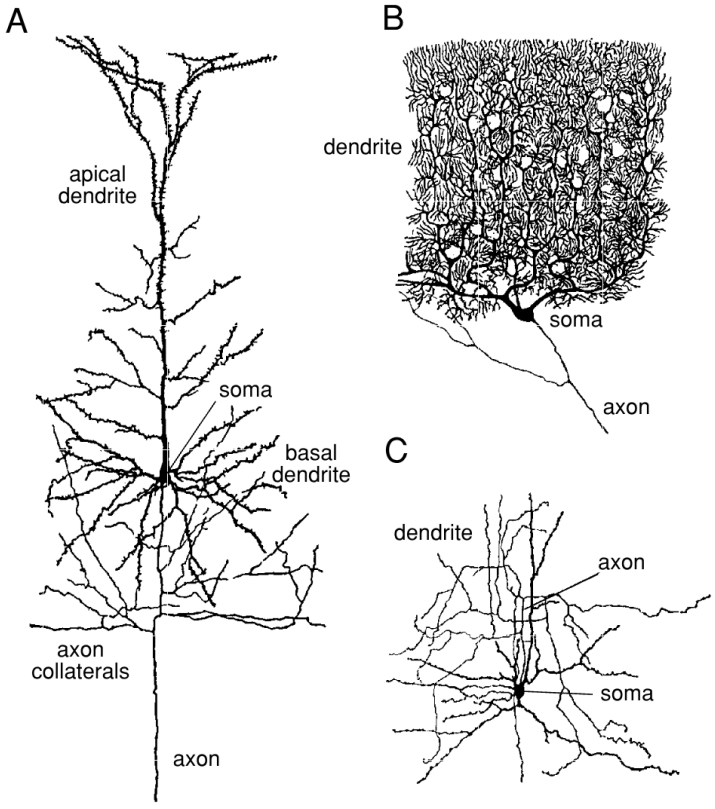
Biological Neural Networks

- Spiking neurons
- Two-dimensional structure
- Sparse connectivity/activations (event-based)
- Asynchronous, no shared memory

MANCHESTER 1824
The University of Manchester

Biological Neurons

- Neural structure has been studied for many years, first through the lens of a microscope
- There are many diverse types of neuron in the brain, and variations both within and across different brain regions
- Most neurons exhibit the 3 basic functional components:
  - ***Soma*** – central processing unit of the neuron, where dendritic inputs come together and are combined
  - ***Axon*** – the branch like projection which communicates output signals from the soma to other neurons
  - ***Dendrites*** – branches where incoming connections from other neurons are made

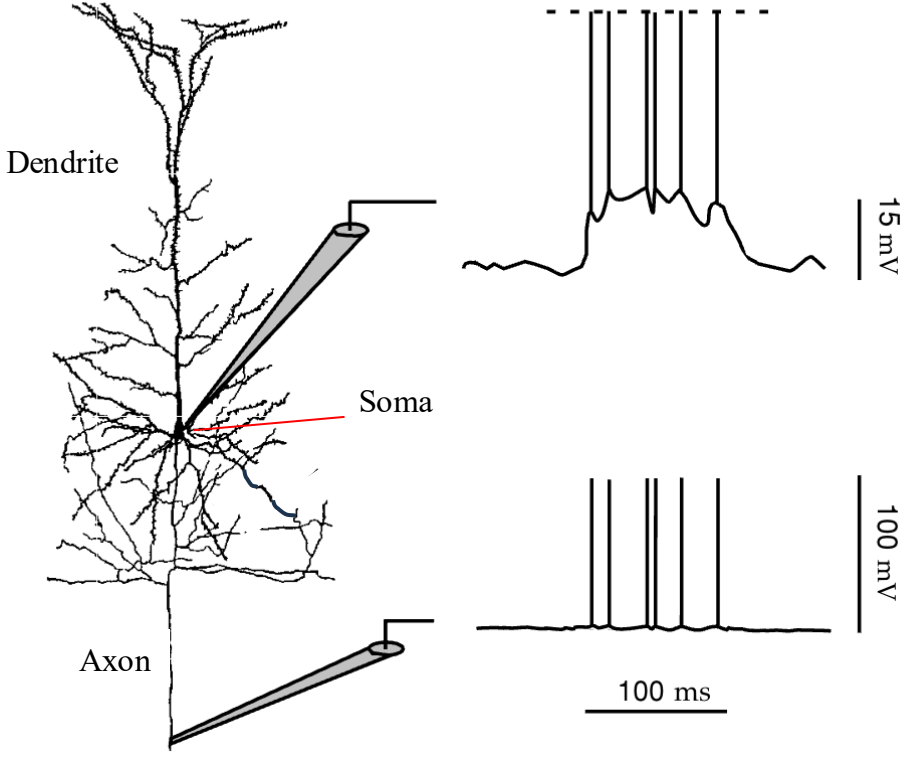


Walter, F. et al (2016). *Computation by Time.* https://doi.org/10.1007/s11063-015-9478-6

MANCHESTER 1824
The University of Manchester

Action Potential & ‘Spiking’

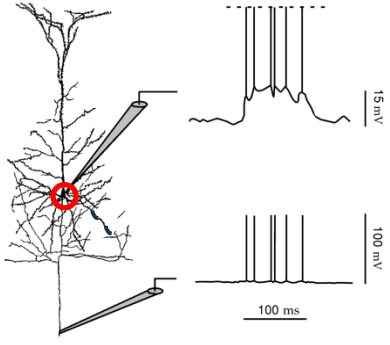
- Neural dynamics are governed by their internal electrical properties and overall structure
- The cell membrane is a relatively strong insulator, allowing substantial potential differences to build up between parts of the neuron, and relative to surroundings.
- Ion concentrations within the cell lead to rapid depolarization, and a pulse of voltage (approx. 100 mV), driven by incoming neural activity
- This pulse travels along the axon to synaptic connections with downstream neurons, enabling ‘computing’ within the system

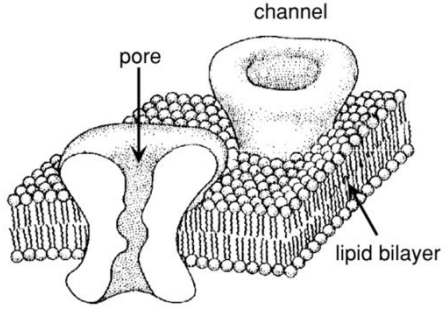


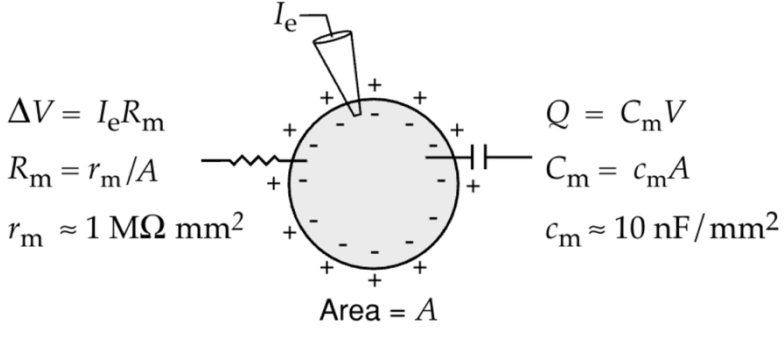
Dayan and Abbott. *Theoretical Neuroscience*. MIT Press. 2001

Cell Membrane

- The cell membrane potential difference can be modelled, e.g. considering soma as point location.
- Potential differences vary in time as different concentrations of ions build up, with a resting state supported through ion pump activity
- The membrane conductance depends on the type and density of ion channels embedded in the cell membrane. Captured via:
  - Membrane capacitance C_m determines how the membrane potential V and excess internal charge Q are related.
  - Membrane resistance R_m determines the size of the membrane potential deviation V caused by a small current – e.g I_e entering through an electrode







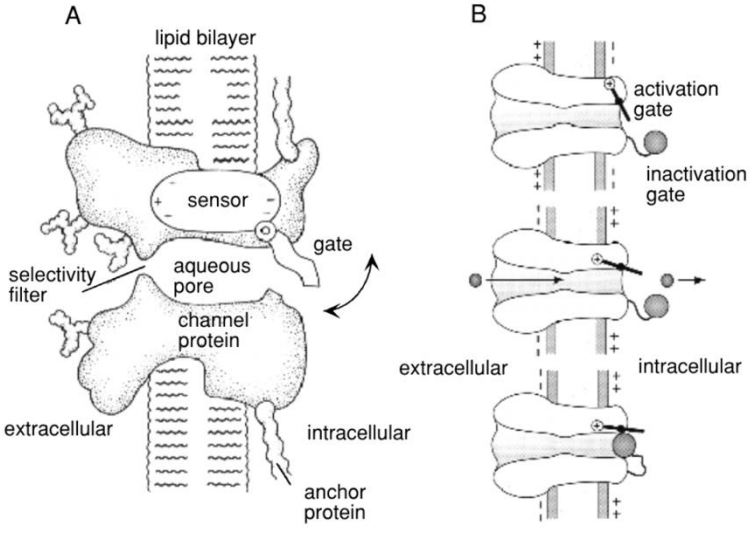
$$C_m \frac{dV}{dt} = \frac{dQ}{dt} = i_m = \sum_i g_i (V - E_i)$$

Dayan and Abbott. *Theoretical Neuroscience*. MIT Press. 2001

MANCHESTER 1824
The University of Manchester

Ion Channels

- Transport of charge across the membrane typically occurs through voltage-dependent ion channels, to drive membrane dynamics
- Selectively allow specific ions through which in turn polarize/depolarize the membrane: Sodium (Na^+), Potassium(K^+), Calcium (Ca^+) etc
- Voltage-gated, with separate dynamics for opening and closing of the channel
- Typically modelled according to the system of equations shown right, calibrated with data from experiments



$$\tau_x(V)\frac{dx}{dt} = x_\infty(V) - x$$

$$\tau_x(V) = \frac{1}{\alpha_x(u) + \beta_x(V)}$$

$$x_\infty(u) = \frac{\alpha_x(V)}{\alpha_x(V) + \beta_n(V)}$$

Dayan and Abbott. *Theoretical Neuroscience*. MIT Press. 2001

MANCHESTER 1824
The University of Manchester

Hodgkin-Huxley Model

- The Hodgekin-Huxley model for single-compartment neurons is able to describe the generation of action potentials.

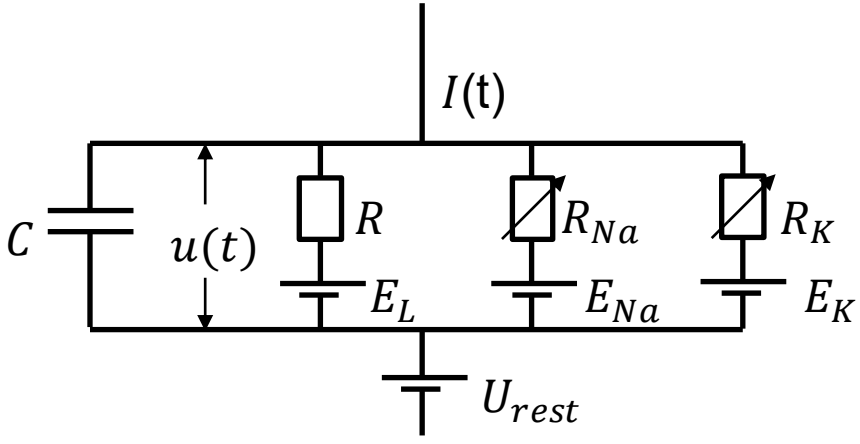
- It was first configured by experimental data, fit to recordings from the axon of a giant squid. This gave rise to the following mathematical description:

$$C\frac{dU}{dt} = -\sum_k i_m + I_{syn}(t)$$

$$i_m = g_L(u - E_L) + g_{Na}m^3h(u - E_{Na}) + g_K n^4(u - E_K)$$

- Individual ion channels described by gating variables m, h (Na), and n (K).

- Enabled accurate modelling of membrane potential during action potential generation



$$\tau_x(V)\frac{dx}{dt} = x_\infty(V) - x$$

$$\tau_x(V) = \frac{1}{\alpha_x(V) + \beta_x(V)}$$

$$x_\infty(V) = \frac{\alpha_x(V)}{\alpha_x(V) + \beta_n(V)}$$

MANCHESTER 1824
The University of Manchester

Hodgkin-Huxley Model

- The Hodgekin-Huxley model for single-compartment neurons is able to describe accurately the generation of action potentials.

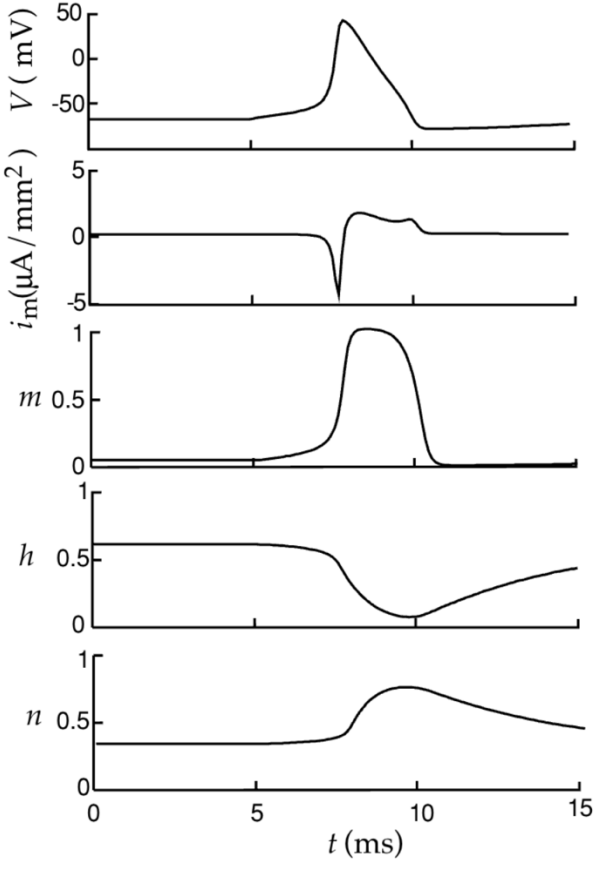
- It was first configured by experimental data, fit to recordings from the axon of a giant squid. This gave rise to the following mathematical description:

$$C\frac{dV}{dt} = -\sum_k i_m + I_{syn}(t)$$

$$i_m = g_L(V - E_L) + g_{Na}m^3h(V - E_{Na}) + g_K n^4(V - E_K)$$

- Individual ion channels described by gating variables m, h (Na), and n (K).

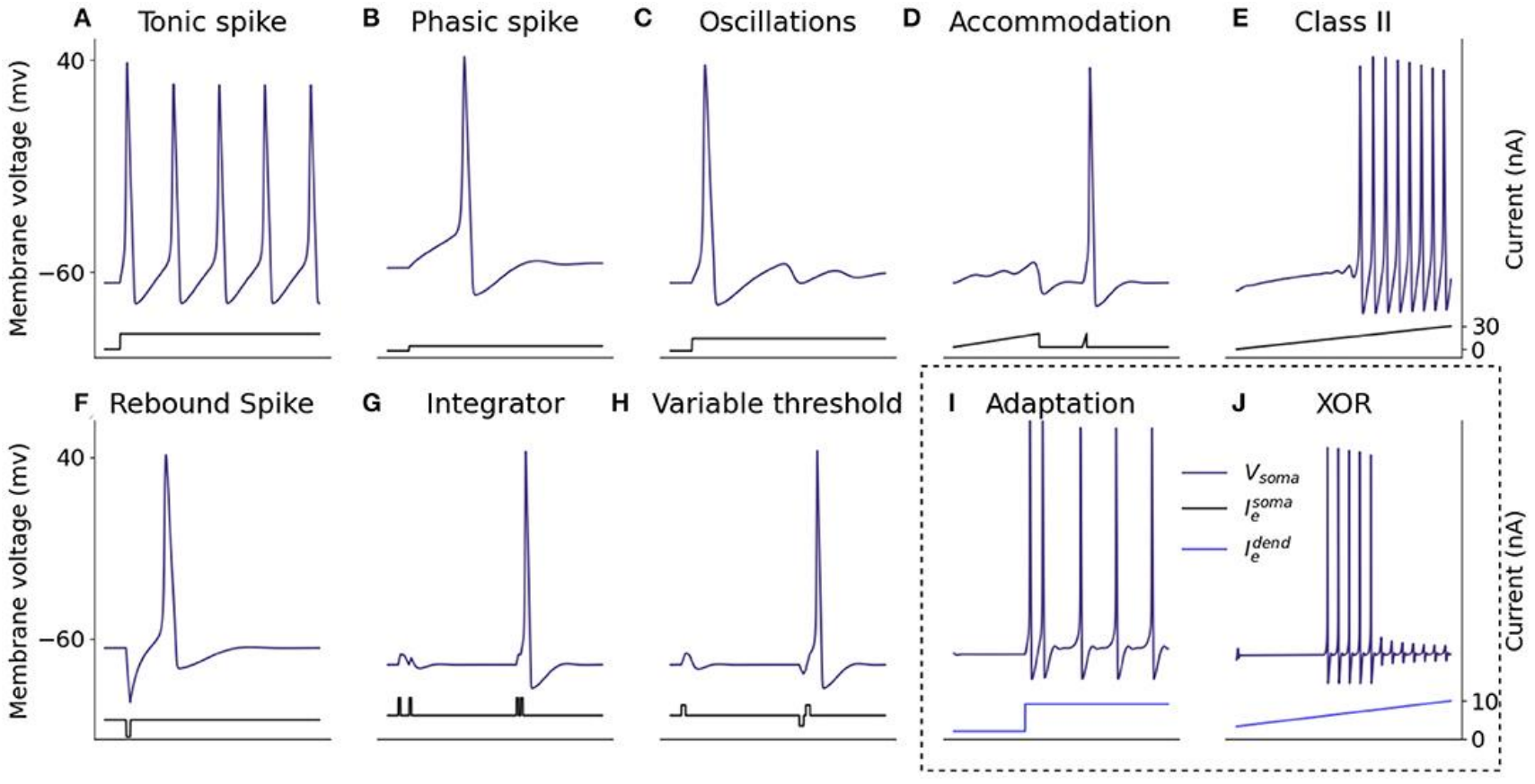
- Enabled accurate modelling of membrane potential during action potential generation



https://www.nobelprize.org/prizes/medicine/1963/summary/

MANCHESTER 1824
The University of Manchester

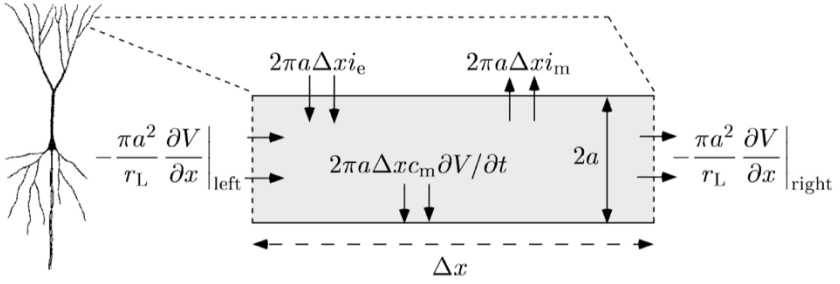
Single HH Neuron Firing Patterns



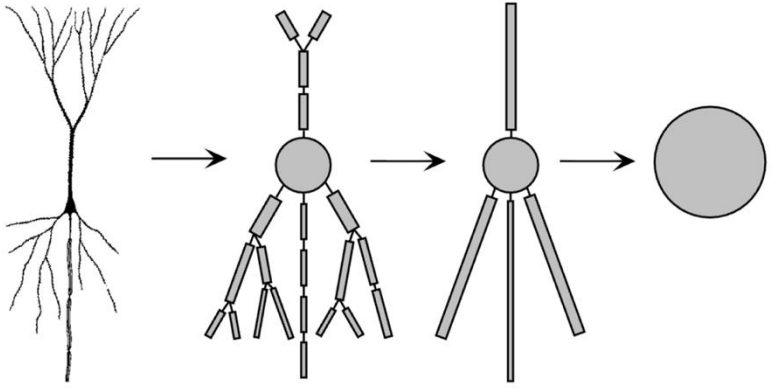
Ward and Rhodes. *Beyond LIF Neurons on Neuromorphic Hardware*. 2022. https://doi.org/10.3389/fnins.2022.881598

Morphology

- Neuronal signals vary significantly as they travel down the thin projections of dendrites and axons.
- Introduces spatial variation to the potential (rather than considering it at a single point), capture mathematically via the **cable equation.**
- Boundary conditions applied through continuity of $V(x,t)$ at nodes (branches/joins between segments),along with charge conservation, enable solution
- Due to the complexities of neuronal membrane currents, the cable equation is typically solved numerically using multi-compartment techniques.



$$c_m \frac{\partial V}{\partial t} = \frac{1}{2ar_L}\frac{\partial}{\partial x}\left(a^2\frac{\partial V}{\partial x}\right) - i_m + i_e$$

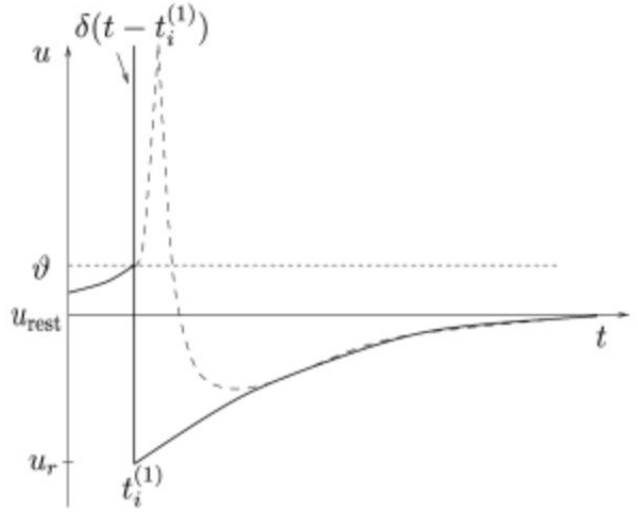


Dayan and Abbott. *Theoretical Neuroscience*. MIT Press. 2001

MANCHESTER 1824
The University of Manchester

Leaky Integrate & Fire

- Action potentials are thought to carry minimal information in their shape, rather it's their presence (or absence) that carries information
- Large amount of computational effort is required to solve the conductance based models.
- Researchers looked to develop simpler models that could capture certain properties of neurons, such as average firing rates at lower computational cost, to enable modelling of networks of neurons.

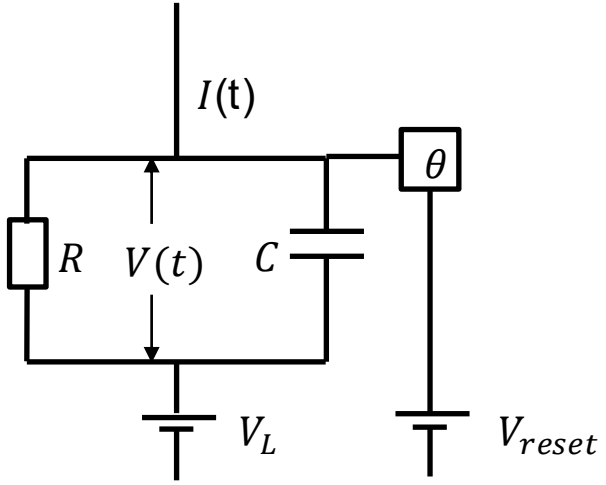


In formal models of spiking neurons the shape of an action potential (dashed line) is usually replaced by a δ pulse (vertical line). The negative overshoot (spike-afterpotential) after the pulse is replaced by a 'reset' of the membrane potential to the value ur. The pulse is triggered by the threshold crossing at $t(1)i$.

MANCHESTER 1824
The University of Manchester

Leaky Integrate & Fire

- Action potentials are thought to carry minimal information in their shape, rather it's their presence (or absence) that carries information
- Large amount of computational effort is required to solve the conductance based models.
- Researchers looked to develop simpler models that could capture certain properties of neurons, such as average firing rates at lower computational cost, to enable modelling of networks of neurons.
- LIF neuron model combines all membrane currents into a single leakage term (V_L and τ_m), and applies hard threshold V_θ to generate spikes



$$\frac{dV}{dt} = -\frac{V - \left(V_L + R_m I(t)\right)}{\tau_m},$$

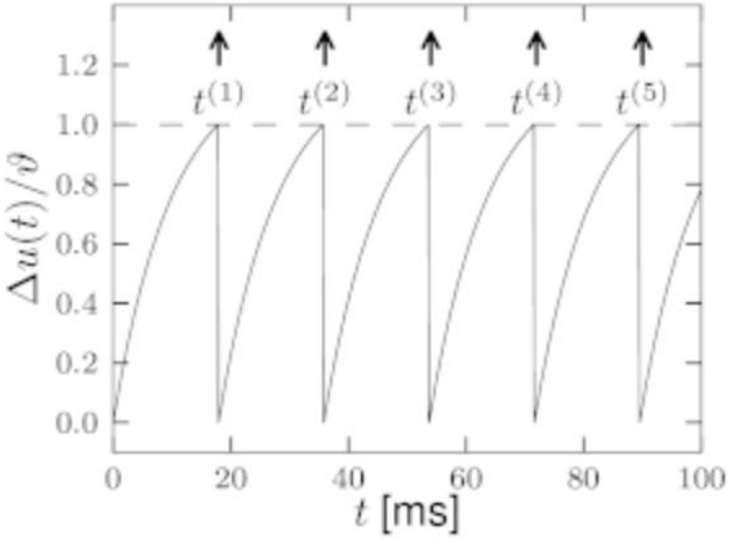
$$\text{if } V > V_\theta, V = V_{reset}$$

Gerstner, W., et al. Neuronal Dynamics. https://neuronaldynamics.epfl.ch/index.html

MANCHESTER 1824
The University of Manchester

Leaky Integrate & Fire

- Action potentials are thought to carry minimal information in their shape, rather it's their presence (or absence) that carries information
- Large amount of computational effort is required to solve the conductance based models.
- Researchers looked to develop simpler models that could capture certain properties of neurons, such as average firing rates at lower computational cost, to enable modelling of networks of neurons.
- LIF neuron model combines all membrane currents into a single leakage term (V_L and τ_m), and applies hard threshold V_θ to generate spikes



$$\frac{dV}{dt} = -\frac{V - \left(V_L + R_m I(t)\right)}{\tau_m},$$

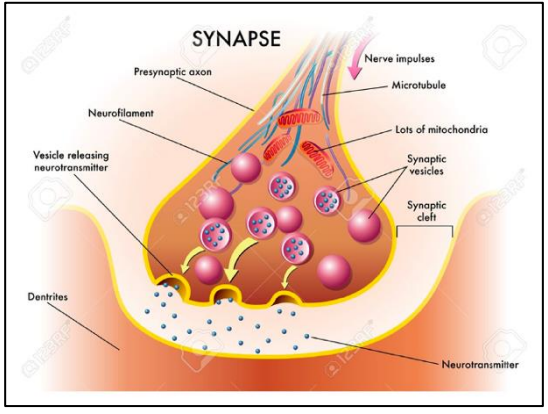
$$\text{if } V > V_\theta, V = V_{reset}$$

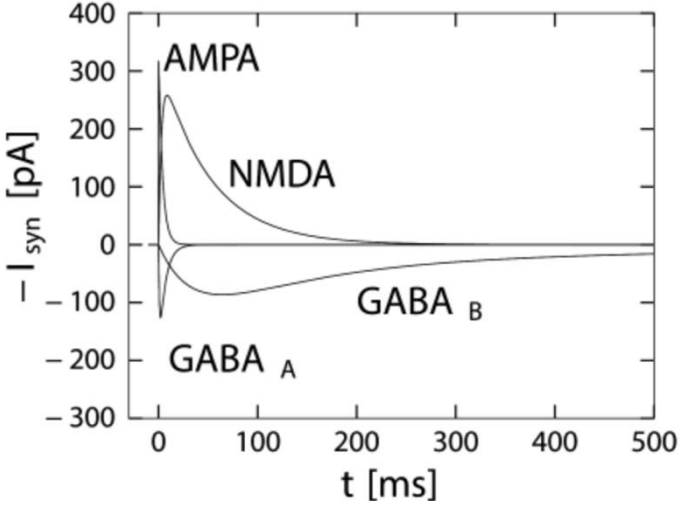
Gerstner, W., et al. Neuronal Dynamics. https://neuronaldynamics.epfl.ch/index.html

MANCHESTER 1824
The University of Manchester

Synapses

- Spikes are modelled as pulses of current, arriving at a neuron via a synapse.
- Neurons receive spikes through their connections to many other neurons in the brain, each via an individual synapse.
- The action potential arriving at the synapse causes release of neurotransmitters into the synaptic cleft, which in turn are taken up by the post-synaptic neuron.
- Range of timescales for different neurotransmitters (e.g. AMPA and NMDA, or GABA A/B). Often abstracted as impulse response with exponential decay. Sub threshold response to combinations of incoming spikes is generally abstracted as being linear, enabling weighted summing of incoming spikes.





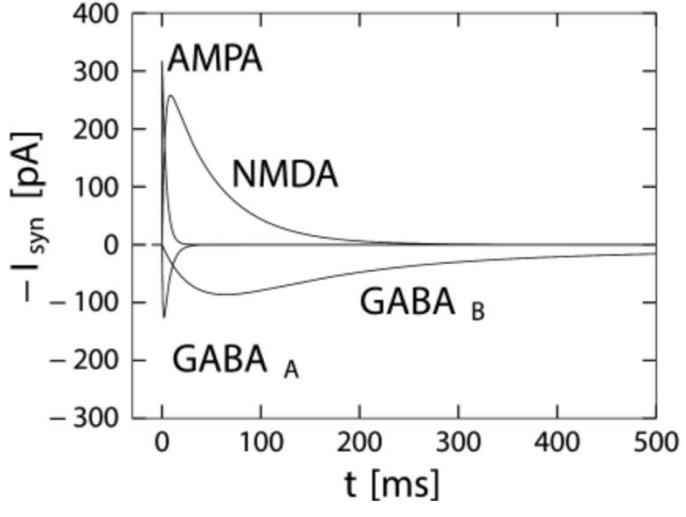
MANCHESTER 1824
The University of Manchester

Synaptic processing

- Action potentials propagate by activating ion channels as they travel down an axon from one neuron to another – can be modelled via the Cable Equation (but expensive)

$$\frac{dI_{syn}}{dt} = -\frac{I_{syn}}{\tau_{syn}} + \sum_j w\delta(t - t^j)$$

- The junction between two neurons is known as a synapse
- Can be excitatory or inhibitory (positive or negative), but a single neuron typically only makes one type of connection to downstream neurons (Dale's principle)
- Take non-zero time to travel from one neuron to another: known as *synaptic delay*
- Can be modelled more simply using an ODE to capture the impulse current response to an incoming spike, abstracting away the low-level ion transport



MANCHESTER 1824
The University of Manchester

LIF with Current-based Synapses

- Putting together the LIF equation with the synapse dynamics ODE of the previous slide, defines the current-based LIF neuron
- Coupled set of first-order differential equations, efficiently solved using numerical techniques such as Euler's method. Accurate for small Δt
- Event-based, suitable for AER implementation, meaning all synaptic information can be stored/processed at the post-synaptic processing node (e.g. core in HPC system). Synaptic delays handled through post-synaptic memory buffers
- Targets real-time execution wall-clock Δt = simulation Δt

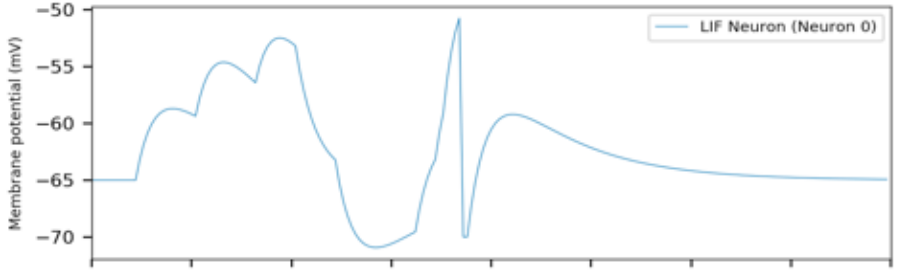
$$\frac{dV}{dt} = -\frac{V - \left(V_L + R_m(I_{syn}(t) + I_e)\right)}{\tau_m},$$

$$\text{if } V > V_\theta, V = V_{reset}$$

$$\frac{dI_{syn}}{dt} = -\frac{I_{syn}}{\tau_{syn}} + \sum_j w\,\delta\left(t - t^j\right)$$

$$V_{t+\Delta t} = V_t e^{-\frac{\Delta t}{\tau}} + (1 - e^{-\frac{\Delta t}{\tau}})(V_L + IR)$$

$$I_{syn,t+1} = I_{syn,t} e^{-\frac{\Delta t}{\tau_{syn}}} + \sum_i w_{ij}\delta(t - t_i)$$

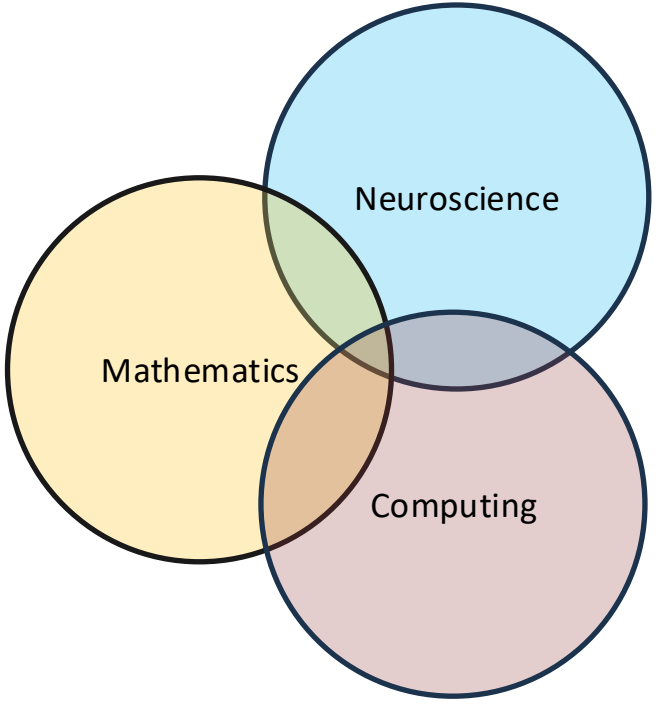


MANCHESTER 1824
The University of Manchester

Neural Simulators

Use computers to model complex neural systems:

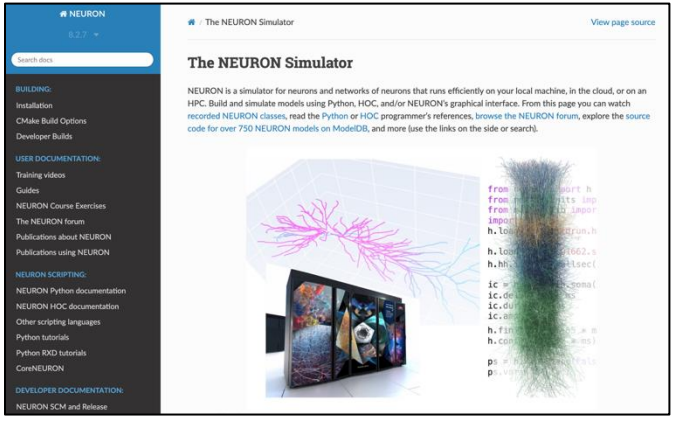
- Enable definition of neural network models, often in terms of individual neuron morphology, or in terms of populations of neurons making up a brain region (very different to a CNN!)
- Provide numerical integration schemes for solving the commonly encountered neural systems
- Map the numerical simulation models to computer hardware (e.g. CPUs, HPC, GPUs etc)
- Provide tools/APIs to record variables of interest throughout simulation, and efficiently store/visualise data in meaningful ways (e.g. post processing statistical metrics)



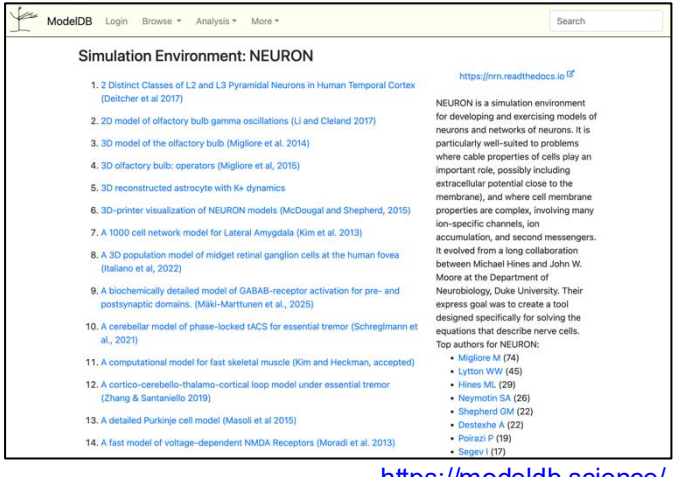
MANCHESTER 1824
The University of Manchester

NEURON

- Simulation environment developed by Yale University, designed to support neuronal models in which complex membrane properties and extended geometry play important roles.
- Models formulated around continuous cable sections which can be connected together to form any kind of branched cable.
- Model descriptions are compiled to enable membrane voltage and gating states to be computed efficiently using an implicit integration method optimized for branched structures.
- Python/HOC scripting language to pre/post process models, and active community with over 815 models hosted on ModelDB.



https://nrn.readthedocs.io/en/8.2.7/

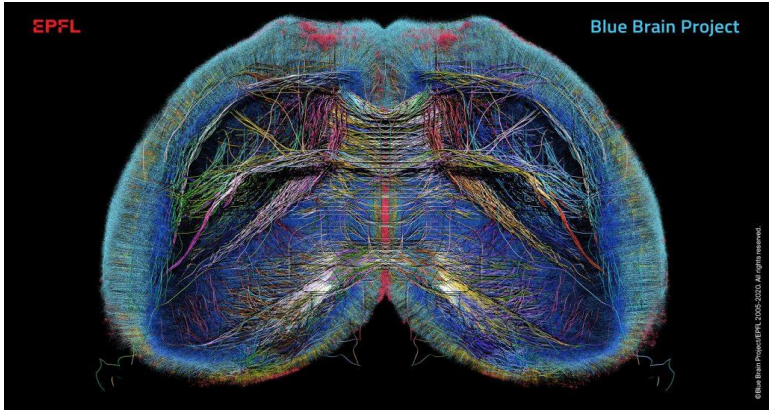


https://modeldb.science/

MANCHESTER 1824
The University of Manchester

Blue Brain Project

- Swiss brain research initiative aiming to develop a digital reconstruction and high-fidelity simulation of the rodent brain using experimentally recorded data.
- Blue Gene/Q supercomputer used to both create and execute models. For example, fitting models to experimental data recorded for individual neurons, and the simulate the systems of equations governing neural behaviour
- NEURON simulation engine provided the framework for integration and testing of the necessary, allowing complex morphologies and neural components to be integrated into the same model
- Later Blue Brain IV (and later 5) took over as simulation platform, comprised of a four-rack IBM BlueGene/Q system, 65 536 PowerPC A2 1.6 GHz cores for computing, providing a peak performance of 839 TFlops 65 TB of RAM 128TB, connected via 40GB/s A FDR Infiniband network. Together with 40-node x86 visualisation system.

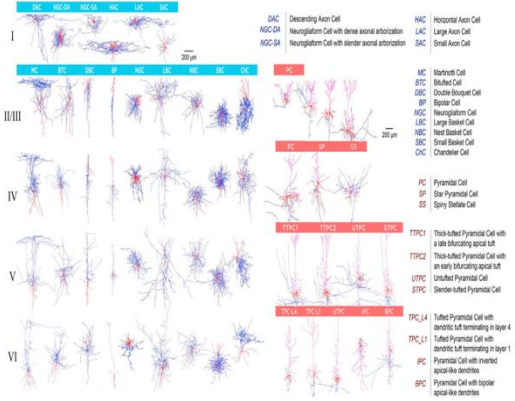


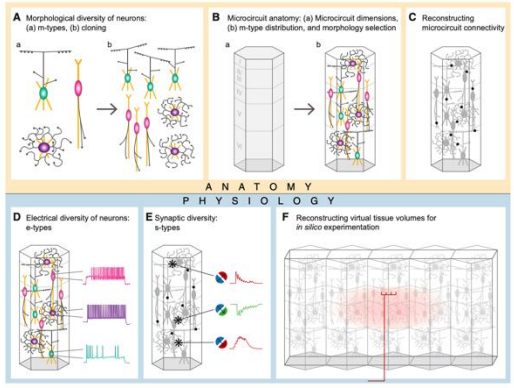
Markram, H. *The Blue Brain Project*. Nat Rev Neurosci. https://doi.org/10.1038/nrn1848

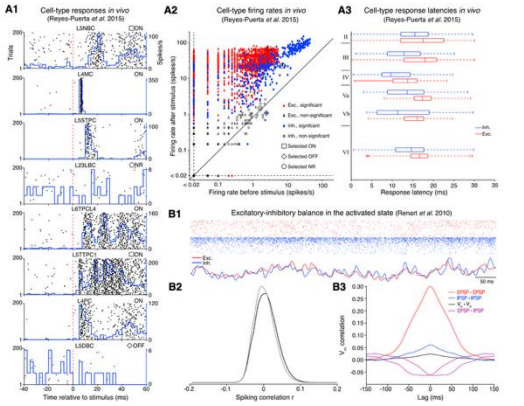
MANCHESTER 1824
The University of Manchester

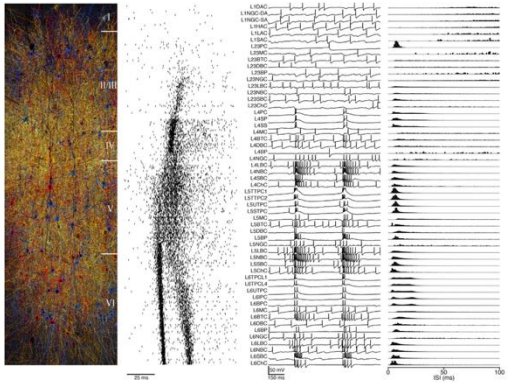
Blue Brain Project

- A key model to come out of the project captured $0.29\ \mathrm{mm}^3$ of rat cortex, containing:
  - 31000 neurons (each individually customised)
  - >250 neuron types
  - 37 million synapses
- The complete mathematical model contained:
  - Average 20000 differential equations per neuron
  - Solution of approx 8 billion differential equations for each second of simulated time.
- Simulations reproduce an array of in vitro and in vivo experiments **without** parameter tuning.
- Despite the impressive achievement, many limitations remain: plasticity, generalisation beyond a single rodent, short timescales, etc









Markram, H., et al. 2015 *Reconstruction and Simulation of Neocortical Microcircuitry*. Cell. https://doi.org/10.1016/j.cell.2015.09.029

MANCHESTER 1824
The University of Manchester

nest ::

JÜLICH Forschungszentrum

- The NEural Simulation Technology (NEST) Initiative, from Jülich Supercomputing Centre

- NEST is a simulator for spiking neural network models that focuses on the dynamics, size and structure of neural systems rather than on the exact morphology of individual neurons

- NEST is ideal for networks of spiking neurons of any size, for example:
  - Models of information processing e.g. in the visual or auditory cortex of mammals,
  - Models of network activity dynamics, e.g. laminar cortical networks or balanced random networks,
  - Models of learning and plasticity.

https://nest-simulator.org/conference

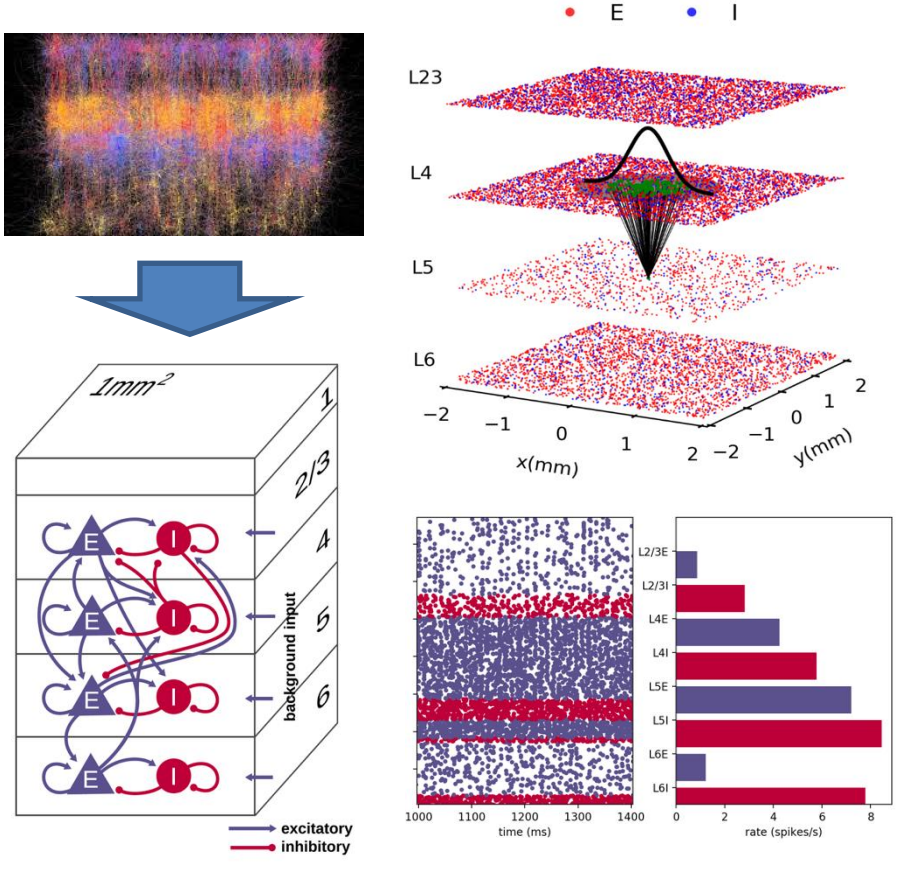
- Python API to programme models (PyNEST)
- Multithreading and MPI for parallelization and scaling to HPC
- Wide range of preconfigured models (with solvers), pre and post-processing routines

https://www.nest-simulator.org/

MANCHESTER 1824
The University of Manchester

P&D Cortical Microcircuit Model

- Model captures connectivity and activity representing early sensory cortex
- Model contains:
  - 77k neurons, organized in layered cortical structure (simple LIF neurons)
  - 300M synapses: recurrent, intra- and inter-layer connections.
- Reproduces biologically-observed neuronal population activity
  - Output activations (spike rates) ~1-10Hz
  - Sparse connectivity: groups of neurons have low probability of connection <1%
  - Individual neurons experience high fan-in (receiving over 10000 connections)

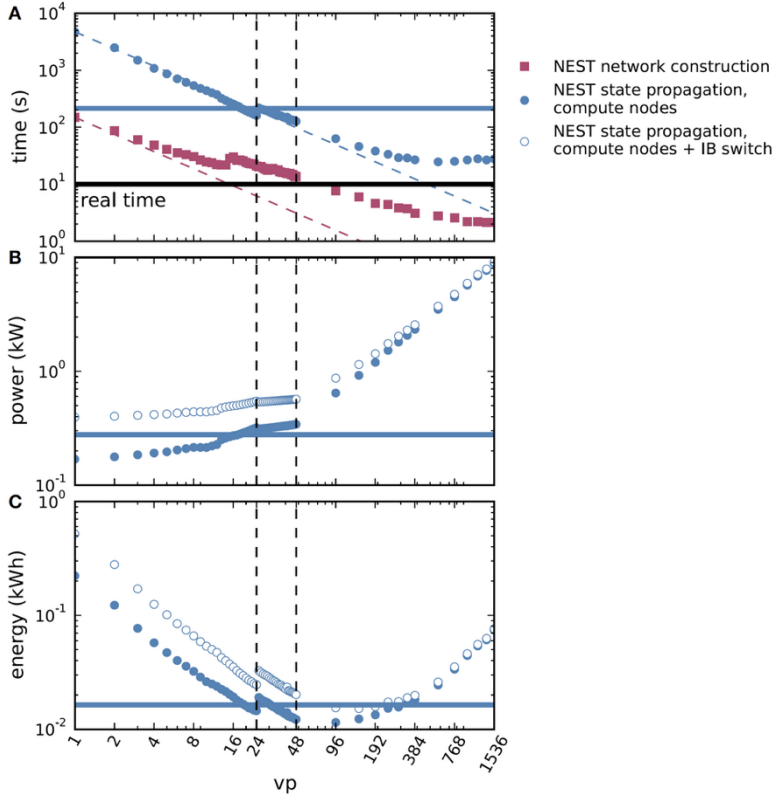


Potjans and Diesman. 2014. *The cell-type specific cortical microcircuit: relating structure and activity in a full-scale spiking network model*. 10.1093/cercor/bhs358

MANCHESTER 1824
The University of Manchester

Challenge for Conventional HPC

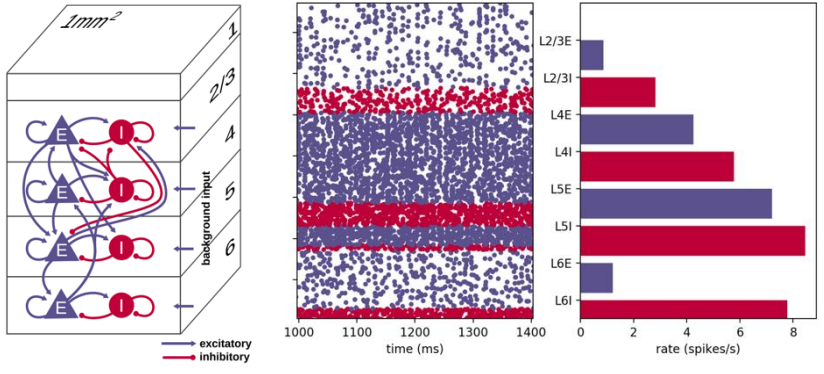
- Even simplified neurons are a computational challenge when combined into large-scale networks
- Performance is communication bound, with energy-to-solution and solution time increasing nonlinearly with SNN size
- Adding HPC compute nodes decreases performance – increases communication traffic
- Traditional MPI based HPC systems are inefficient when dealing with sparse activations and sparse connectivity



S.J. van Albada et al., "Performance comparison of the digital neuromorphic hardware SpiNNaker and the Neural network simulation software NEST for a full-scale cortical microcircuit model", Frontiers 2018. https://doi.org/10.3389/fnins.2018.00291

Cortical Microcircuit on SpiNNaker

- First real-time simulation of cortical microcircuit model: 77,169 neurons, 285M synapses, 2,901 cores
- 1 second of model time simulated in 1 second of wall-clock time (3x faster than GPUs, 2x faster than CPUs)
- Model validated against double-precision reference solution running on synchronous shared memory system
- 12 hour simulation performed, demonstrating feasibility for studying long-term neurological behaviour through models
- Low-power execution – energy per synaptic event of 0.6 microjoules



Configuration	Wall-Clock Time	Total Energy (kWh)	Synaptic Events	Energy per synaptic event (μJ/syn-event)
System Only	12 h	0.93	–	–
12 booted boards	12 h	4.93	–	–
Cortical microcircuit, DC input, real-time	12 h	6.59	–	–
Cortical microcircuit, Poisson input, real-time	12 h	7.11	–	–
Cortical microcircuit, DC input, real-time	10 s	0.001525*	9.135×10^9	0.601
Cortical microcircuit, Poisson input, real-time	10 s	0.001646*	9.343×10^9	0.628

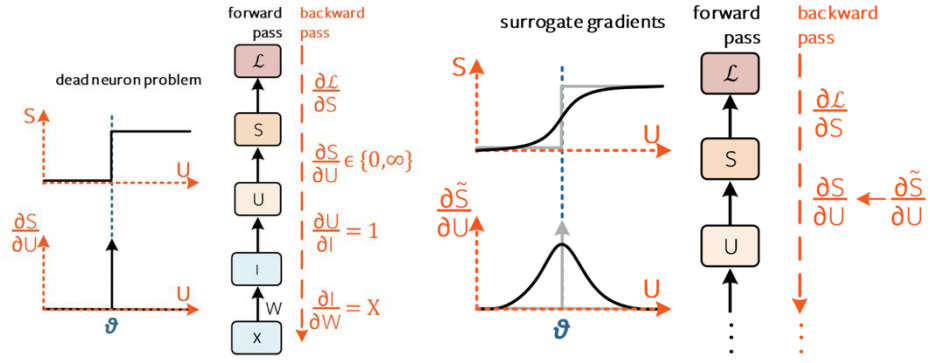
Table 1. Energy measurements from 24-board SpiNNaker system used to execute cortical microcircuit simulations (*quantity estimated from extended duration simulation).

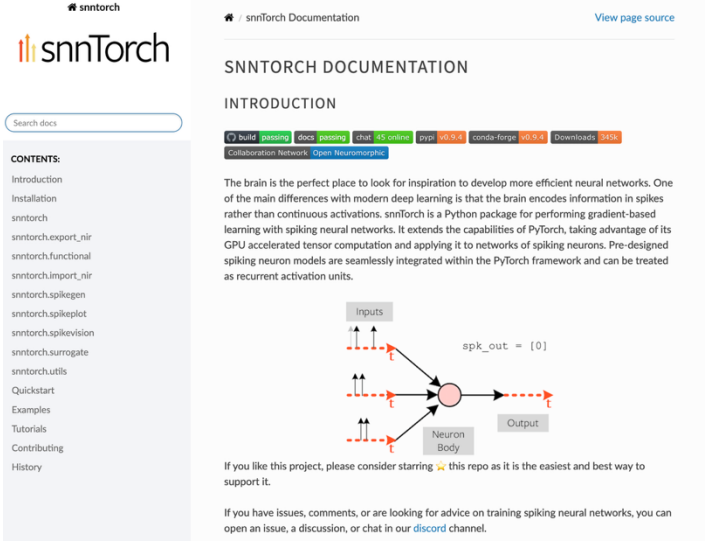
rates; B, coefficient of variation of inter-spike interval; C, correlation coefficient between binned spiketrains.

Rhodes, O. et al., "Real-time cortical simulation on neuromorphic hardware", R. Soc. A. 2020.
https://doi.org/10.1098/rsta.2019.0160

SNNTorch

- PyTorch extensions now enable SNN 'simulation'. Surrogate gradient techniques have opened up new possibilities to train SNNs
- Simple models which can be integrated numerically can fit the PyTorch graph (single compartment LIF variant, no synaptic delays, etc)
- Facilitates automatic differentiation, integration with PyTorch dataloaders and optimization algorithms
- Leverages GPU acceleration to enable batch training, parallelism, and training of large-scale SNN models.





Eshraghian et al. *Training Spiking Neural Networks Using Lessons from Deep Learning*. 2022. https://arxiv.org/abs/2109.12894

SpikeGPT: Generative Pre-trained Language Model with Spiking Neural Networks

Abstract

As the size of large language models continue to scale, so does the computational resources required to run them. Spiking Neural Networks (SNNs) have emerged as an energy-efficient approach to deep learning that leverage sparse and event-driven activations to reduce the computational overhead associated with model inference. While they have become competitive with non-spiking models on many computer vision tasks, SNNs have proven to be more challenging to train. As a result, their performance lags behind modern deep learning, and until now, SNNs have yet to succeed at language generation on large-scale datasets. In this paper, inspired by the Receptance Weighted Key Value (RWKV) language model, we successfully implement 'SpikeGPT', a generative language model with binary, event-driven spiking activation units. We train the proposed model on two model variants: 46M and 216M parameters. To the best of our knowledge, SpikeGPT is the largest backpropagation-trained SNN model when released, rendering it suitable for both the generation and comprehension of natural language. We achieve this by modifying the transformer block to replace multi-head self-attention to reduce quadratic computational complexity $\mathcal{O}(T^2)$ to linear complexity $\mathcal{O}(T)$ with increasing sequence length. Input tokens are instead streamed in sequentially to our attention mechanism (as with typical SNNs). Our experiments show that SpikeGPT remains competitive with non-spiking models on tested benchmarks, while maintaining 32.2× fewer operations when processed on neuromorphic hardware that can leverage sparse, event-driven activations. Our code implementation is available at https://github.com/ridgerchu/SpikeGPT.

1 Introduction

Artificial Neural Networks (ANNs) have recently achieved widespread, public-facing impact in Natural Language Processing (NLP), but with a significant computational and energy consumption burden across training and deployment

Rui-Jie Zhu, et al. 2024. *SpikeGPT: Generative Pre-trained Language Model wit*
Spiking Neural Networks. https://doi.org/10.48550/arXiv.2302.13939

MANCHESTER 1824
The University of Manchester

Summary

- Why model spiking neural networks?
- Building blocks of computational neuroscience
- Review of software simulators and applications