Scaling Laws and Tradeoffs in Recurrent Networks of Expressive Neurons

Aaron Spieler[1,2,*] **Georg Martius**[1,3] **Anna Levina**[1,2]

[1]University of Tübingen [2]Max Planck Institute for Biological Cybernetics, Tübingen [3]Max Planck Institute for Intelligent Systems, Tübingen * aaron.spieler@tuebingen.mpg.de

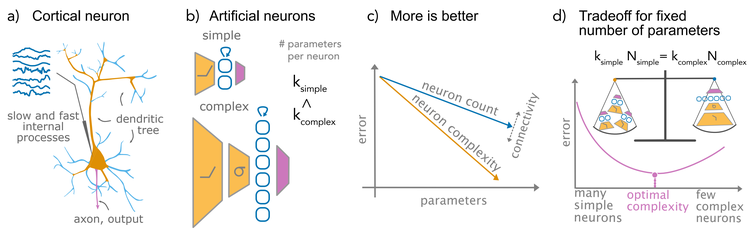
Paper

Code

Question

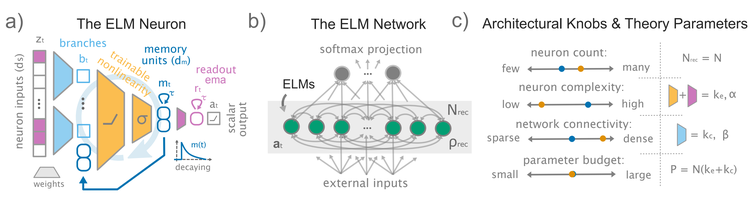
How should a recurrent network split a fixed parameter budget between the number of neurons, per-neuron complexity and connectivity?

1) Cortical neuron complexity motivates the question



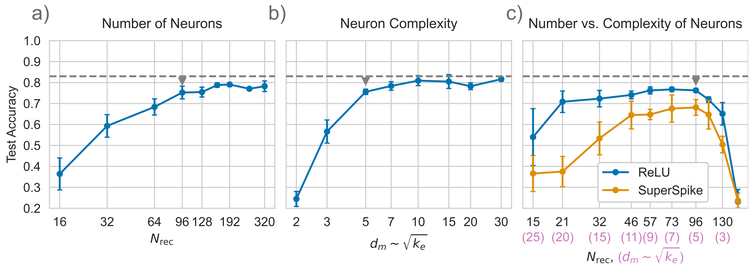
- Cortical neurons are **complex** [1]; machine learning mostly builds with **simple** units.
- Budget $P = N(k_e + k_c)$: N neurons, complexity k_e = internal parameters, connectivity k_c.

2) The ELM Network makes each axis tunable



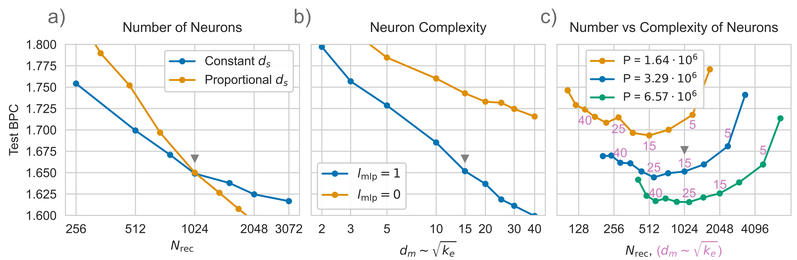
- **Architecture:** a wide recurrent layer of scalar-output Expressive Leaky Memory (ELM) neurons [2], each itself recurrent, integrating inputs nonlinearly over multiple timescales.
- **Scaling axes:** N, $k_e \sim d_m^2$, k_c; stable training across three orders of magnitude.

3) A clear optimum emerges under a fixed budget



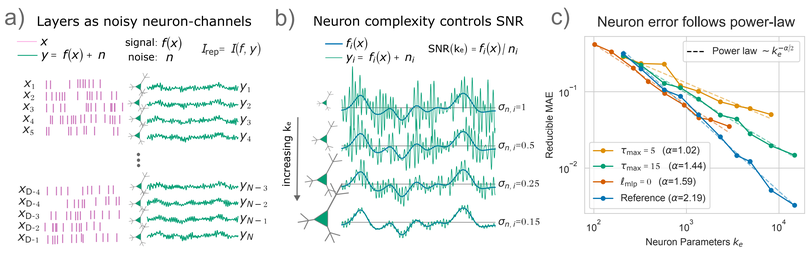
- **a, b)** On SHD-Adding [2, 3], more neurons and more complex neurons improve accuracy.
- **c)** Under a fixed parameter budget, the two compete: the **optimum lies between the extremes**, and sharpens with binary-spiking activation (SuperSpike).

4) Larger budgets favor both more *and* more complex neurons



- **a, b)** On character-level language modeling (Enwik8 [4]), the gains extend over at least another order of magnitude in neuron number and complexity.
- **c)** The optimum emerges robustly across three parameter budgets and **shifts with budget**: from $(N, d_m) = (512, 15)$ to $(1024, 25)$ when the budget is quadrupled.
- The optimum is robust to recurrence level, number of hidden layers and budget type (wall-time, FLOPs); similar tradeoffs appear in multi-head attention Transformers.
- Connectivity introduces an additional tradeoff; performance benefits especially from recurrent rather than only feedforward connections.

5) An information-theoretic model explains the optimum



- **a)** We model a neural layer as a set of **noisy neuron-channels** and quantify how much task-relevant information it carries [5]: the *effective representation information* I_{rep}.
- **b, c)** More complex neurons have a higher signal-to-noise ratio: fitted to a simulated cortical neuron [1], an ELM neuron's error decreases as a power law in its parameters k_e.
- For a fixed budget, complex neurons improve signal quality but fewer of them fit:

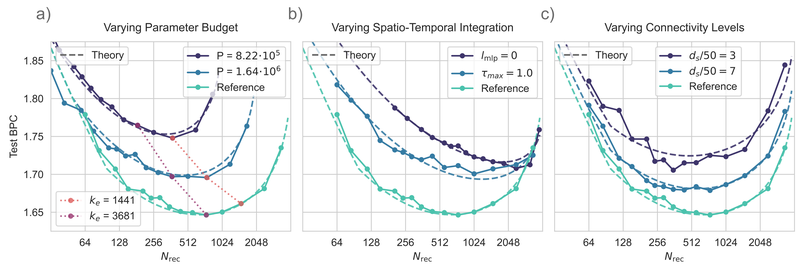
$$I_{\text{rep}}(k_e) = \frac{1}{2} \sum_{i=1}^{P/(k_e+k_c)} \log_2\left(1 + s(k_e)\, i^{-\beta}\right), \qquad s(k_e) = \min\left((\gamma k_e)^{\alpha},\ q_\infty^{-1}\right)$$

α **expressivity exponent**, set by the neuron's spatio-temporal integration

β **inter-neuron redundancy exponent**, shaped by connectivity

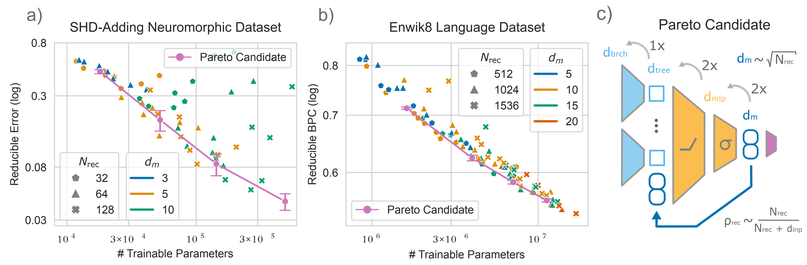
k_e, k_c neuron complexity, connectivity cost • s signal-to-noise ratio • γ conversion efficacy • q_∞ noise floor

6) The theory reproduces the tradeoffs and their shifts



- A **single joint fit** of l_{rep} (dashed) qualitatively reproduces all curves ($r = 0.98$).
- **Measured independently**, α and β still predict the curves at other budgets ($r = 0.98$).
- **a)** More budget favors more numerous and more complex neurons.
- **b)** Weaker spatio-temporal integration (lower α) favors wider layers.
- **c)** Reduced connectivity (higher β) favors fewer, more complex neurons.
- The optimum emerges from diminishing returns at two ends: **per-neuron signal-to-noise saturation** and **across-neuron redundancy**.

7) A simple joint scaling rule tracks the Pareto frontier



- **a, b)** A single scaling rule closely traces the empirical Pareto frontier on both datasets, across roughly three orders of magnitude in trainable parameters, yielding a scaling law [6].
- **c) Pareto candidate**: only N_{rec} is free, with $d_m \sim \sqrt{N_{\text{rec}}}$; per-neuron parameters and connectivity scale linearly with N_{rec}.

Take-home

- More neurons, more complex neurons and more connectivity each improve performance. Under a fixed budget, the best networks **balance** them.
- Larger budgets favor both **more and more complex neurons**.
- Our information-theoretic model explains the **optimum and its shifts**.
- The simple-unit default in ML is **not obviously optimal**.

KEY REFERENCES

[1] Beniaguev et al. (2021) *Neuron.*
[2] Spieler et al. (2024) *ICLR.*
[3] Cramer et al. (2020) *IEEE TNNLS.*
[4] Mahoney (2011) *Large Text Compression Benchmark.*
[5] Shannon (1948) *Bell Syst. Tech. J.*
[6] Kaplan et al. (2020) *arXiv:2001.08361.*

ACKNOWLEDGEMENTS

Funded by the BMFTR through the RealmNN project under the NeuroAI initiative (01GQ2508); supported by the BMBF through the Tübingen AI Center (01IS18039B). AL is a member of the Cluster of Excellence "Machine Learning: New Perspectives for Science", EXC 2064/1 (39072764). Compute: MPCDF Raven.

MAX PLANCK INSTITUTE FOR BIOLOGICAL CYBERNETICS

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

NI:KI INITIATIVE