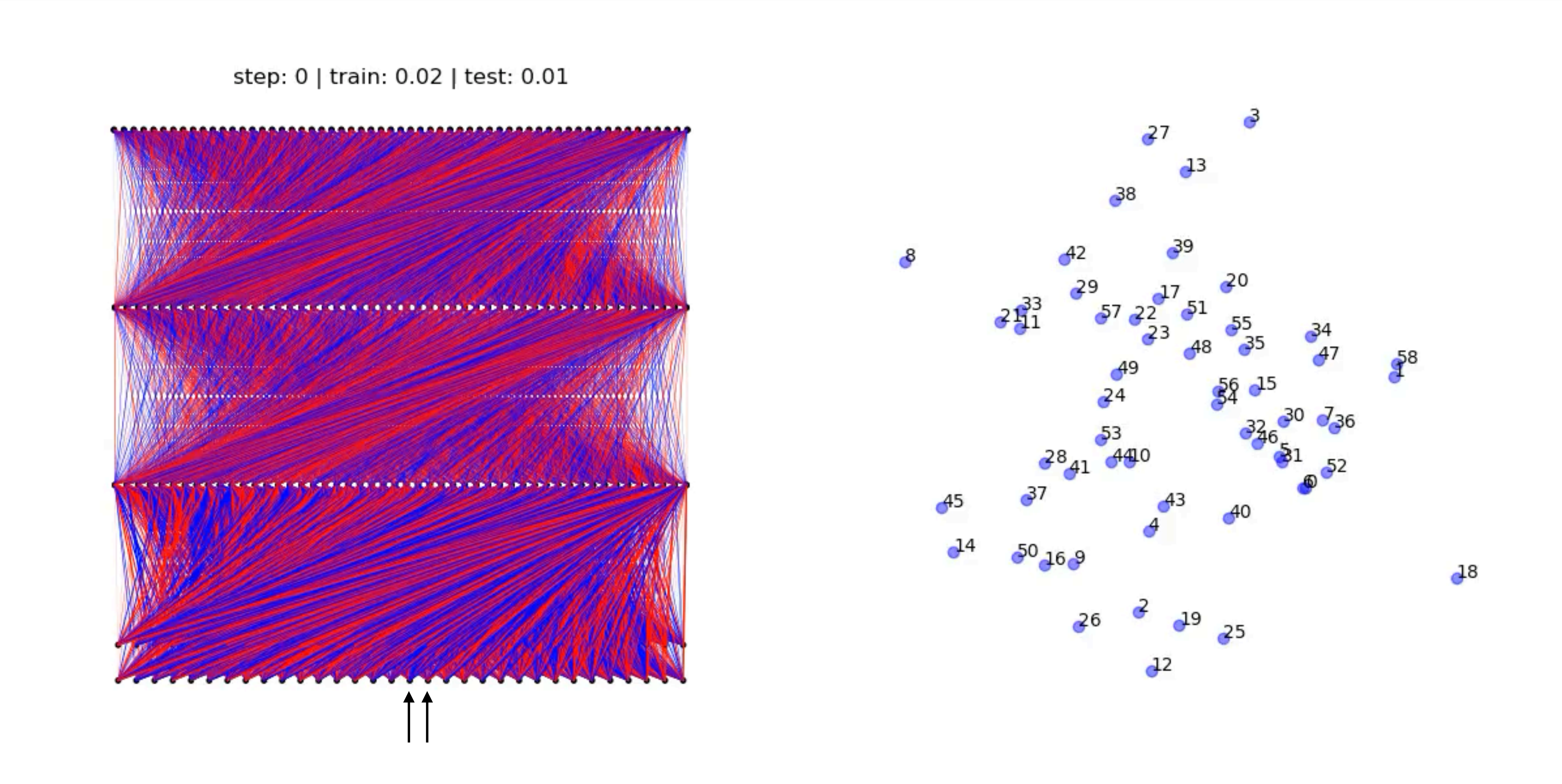
Intelligence from hunger

—*How do representations, modularity, quantisation emerge from limited resources?*

Ziming Liu (刘子鸣), MIT & IAIFI, advisor: Max Tegmark

@Tiktok, June 6, 2023

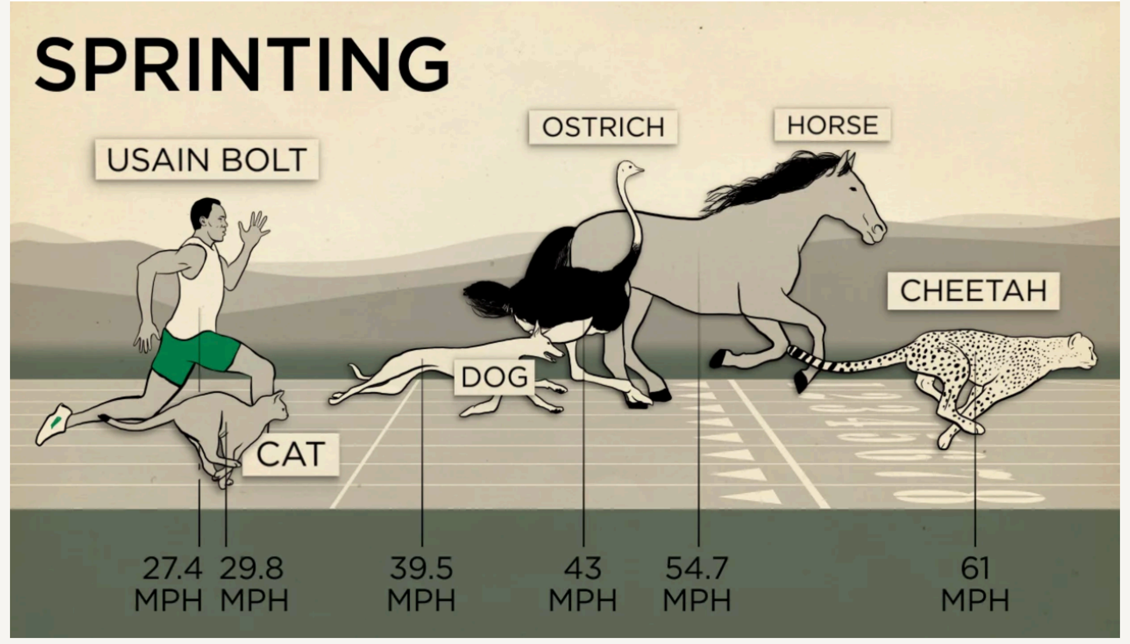


Evolution: Intelligence from hunger/danger

Predator

Climate

Physical limitations



Use Spears etc.

Build houses etc.

Build cars etc.

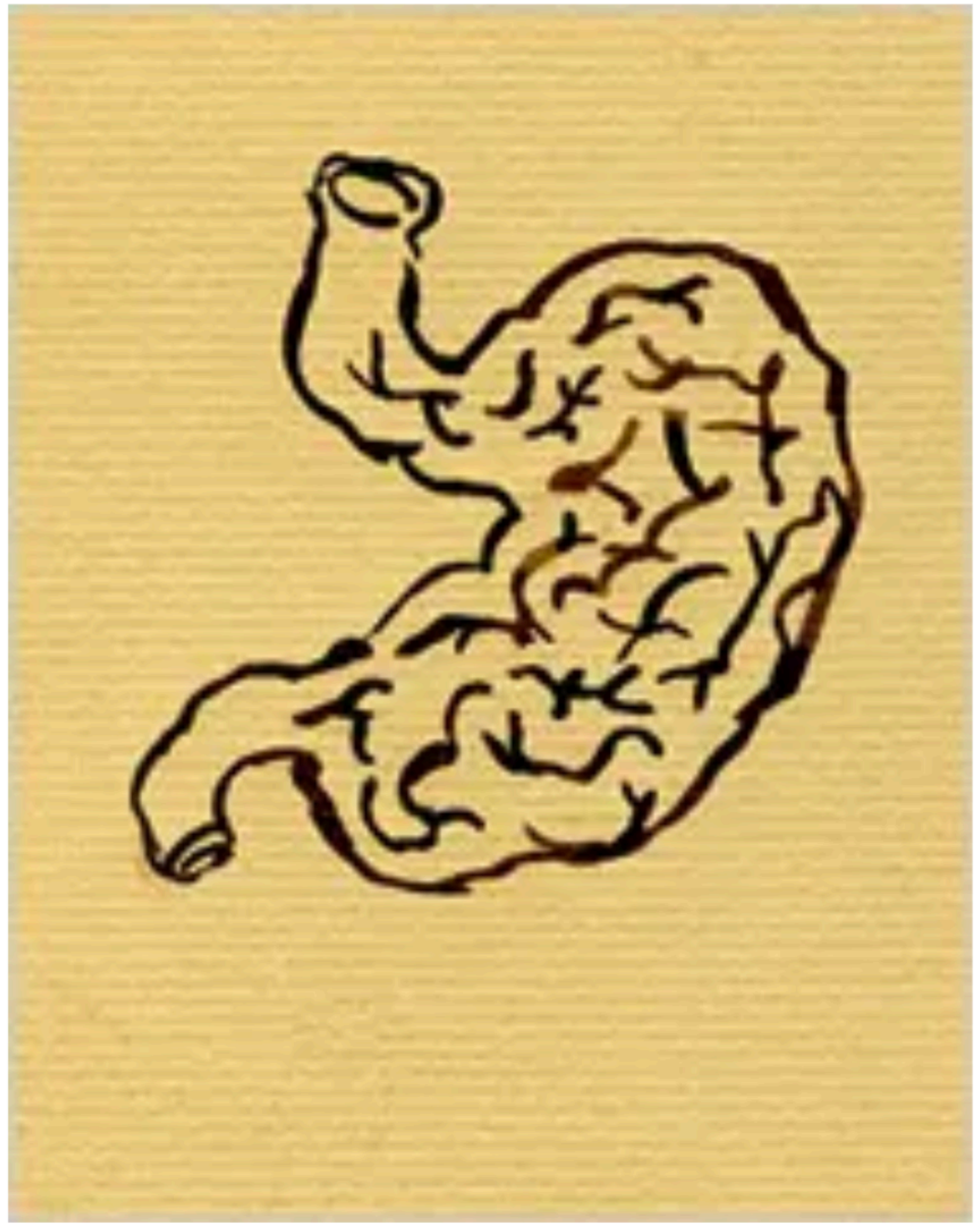
Empty-stomach intelligence: hunger keeps brains sharper

This is your brain on an empty stomach

Medicine@Yale, 2006 - May June

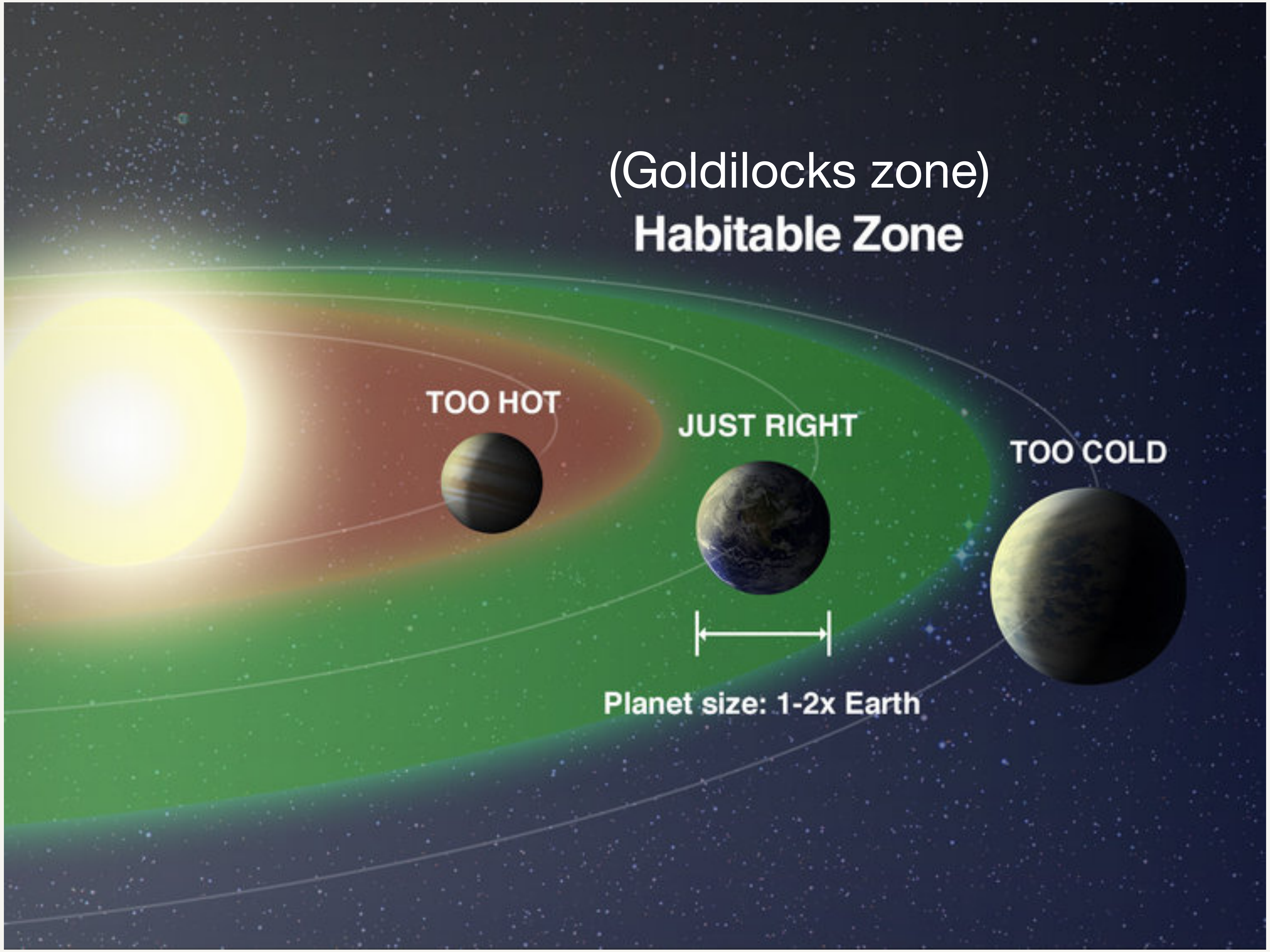
Cutting calories can definitely make you trimmer, and may help you live longer. Now a new Yale study suggests that dieting might also keep you mentally sharper.

Blood levels of a gut hormone called ghrelin (rhymes with “melon”) rise when the stomach is empty, flooding the brain’s eating control center and stimulating neurons that govern appetite. When Tamas L. Horvath, D.V.M., Ph.D., chair and associate professor of comparative medicine, and colleagues injected mice with ghrelin, the hormone rapidly altered circuits in the hippocampus, a brain region that is crucial to learning and memory. Ghrelin-treated mice were significantly better at learning and remembering their way around a maze.



Dietrich et al, 2012. "AgRP neurons regulate development of dopamine neuronal plasticity and non-food associated behaviour”

Goldilocks zone for lives



Representations

Representation learning

Resources

Too few resources

Just right

Too many resources

Representation learning

Resources

Too few resources

Just right

Too many resources

↓

Cannot learn anything

↓

No representation Learning

Representation learning

Resources

Too few resources

↓

Cannot learn anything

↓

No representation Learning

Just right

Too many resources

↓

Memorize everything

↓

No representation Learning

Representation learning

Resources

Too few resources

↓

Cannot learn anything

↓

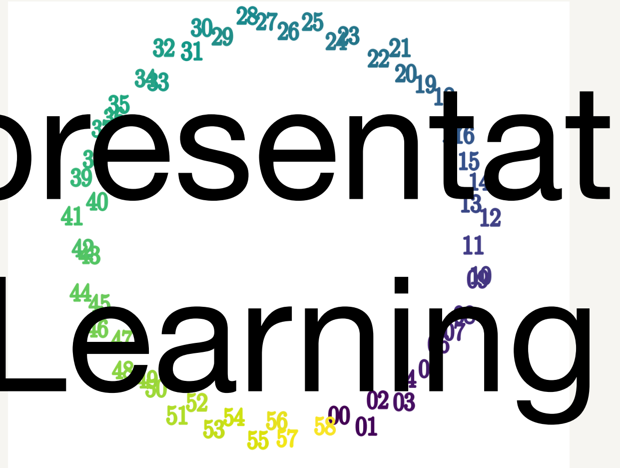
No representation Learning

Just right

↓

Search for clever ways for computation

↓



Representation Learning

Too many resources

↓

Memorize everything

↓

No representation Learning

Setup: Algorithmic datasets

$$\boxed{a} \circ \boxed{b} = \boxed{c}$$

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Setup: Algorithmic datasets

Split the table into train & val datasets

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

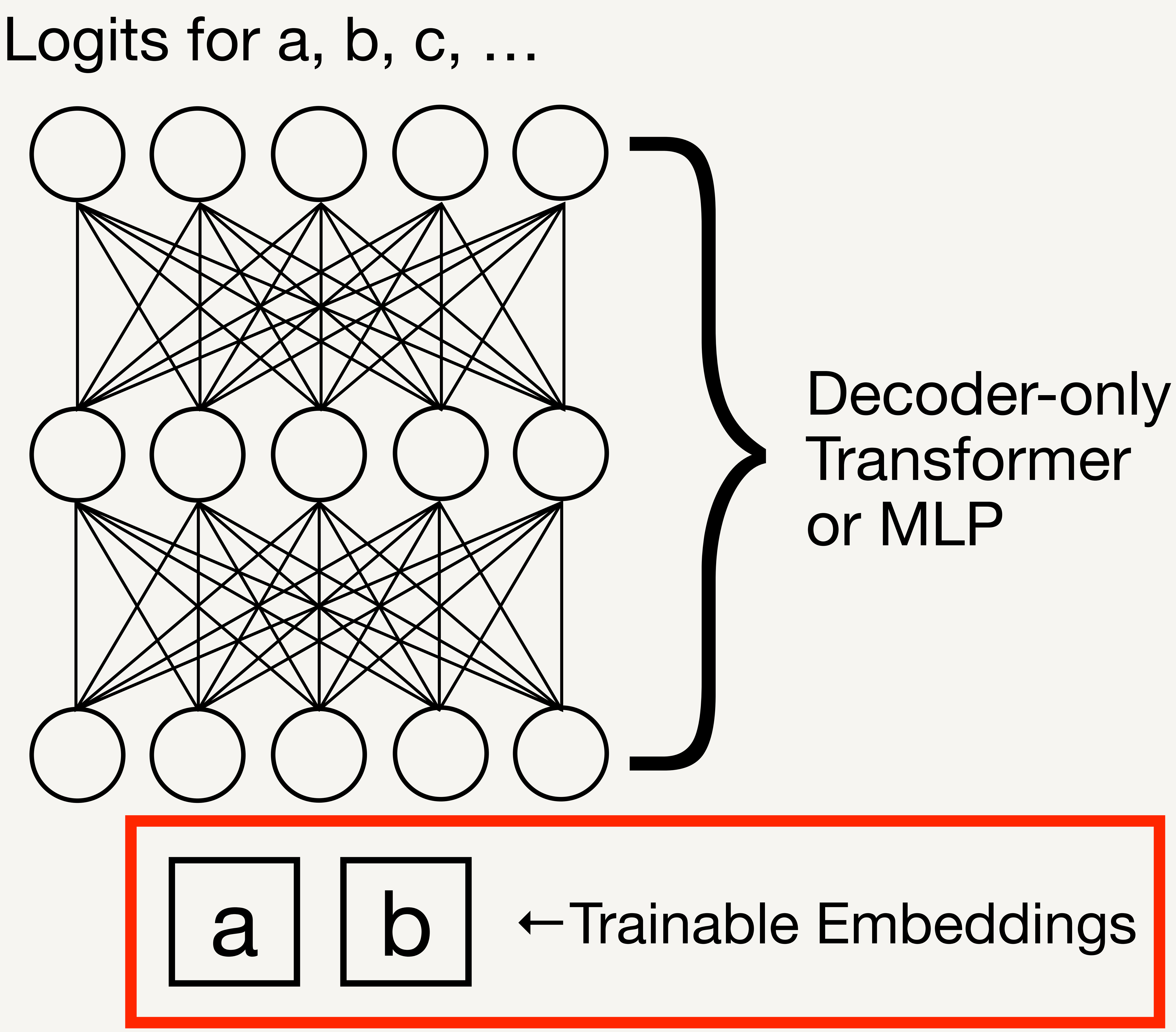
From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Setup: Algorithmic datasets

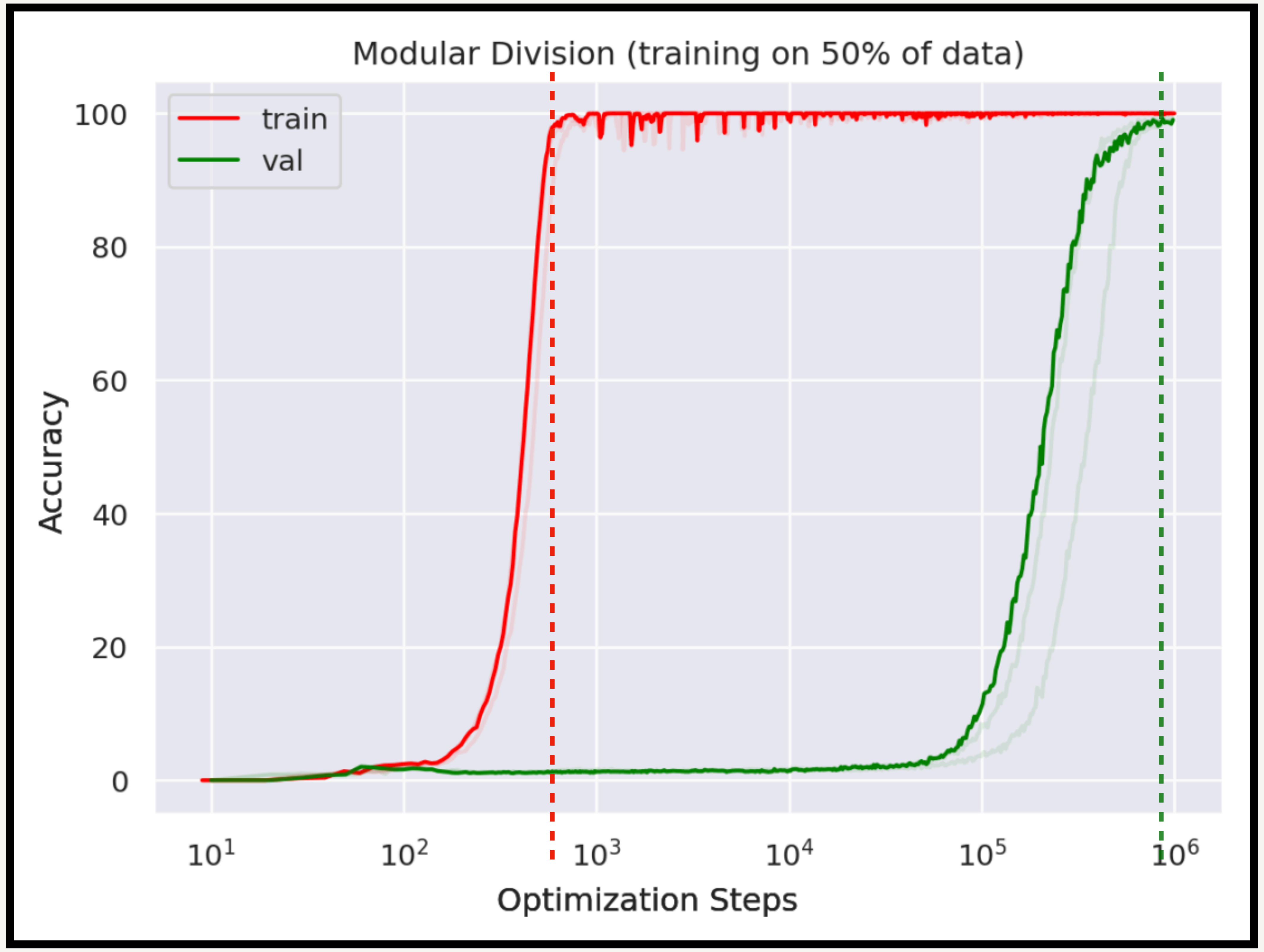
Task: learn a binary operation

$$a + b \mod p = c$$

$$12 + 23 \mod 59 = 35$$



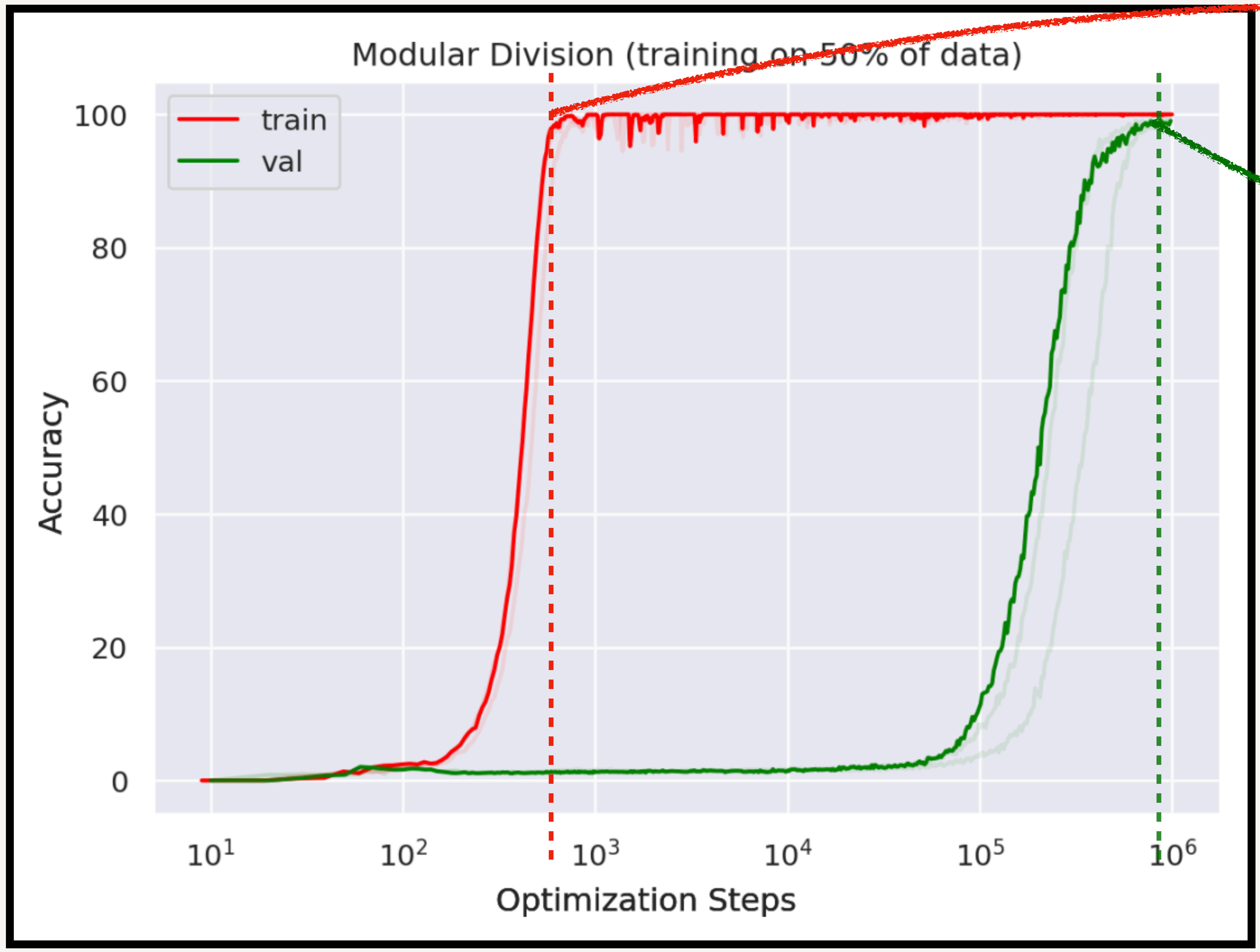
Grokking

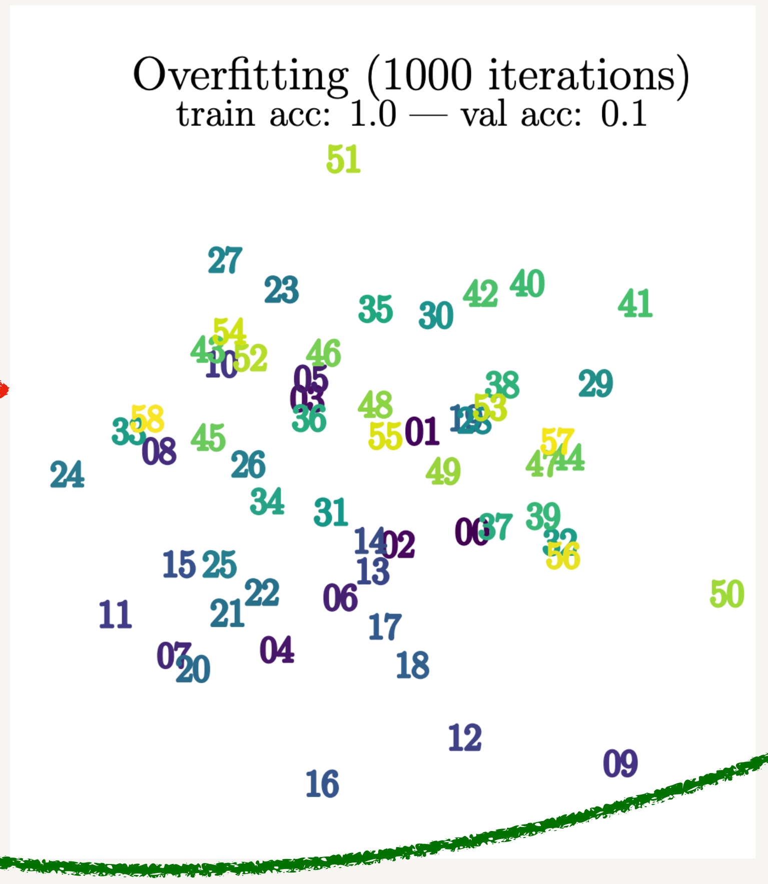


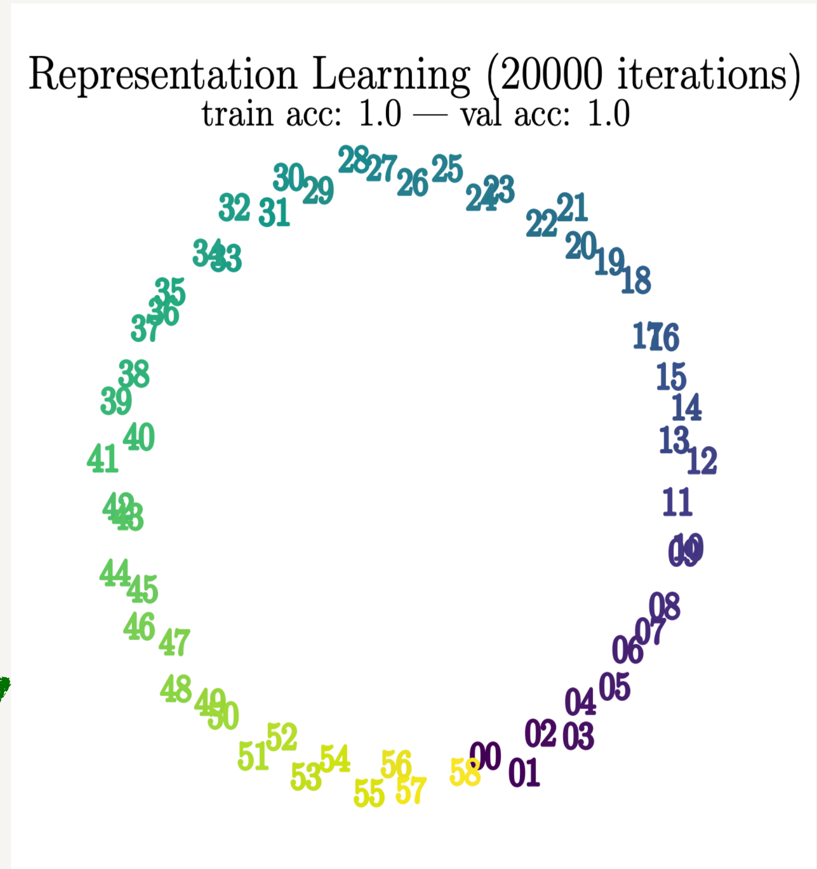
Power et al , “Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets”

Grokking

Representation is key for generalisation!



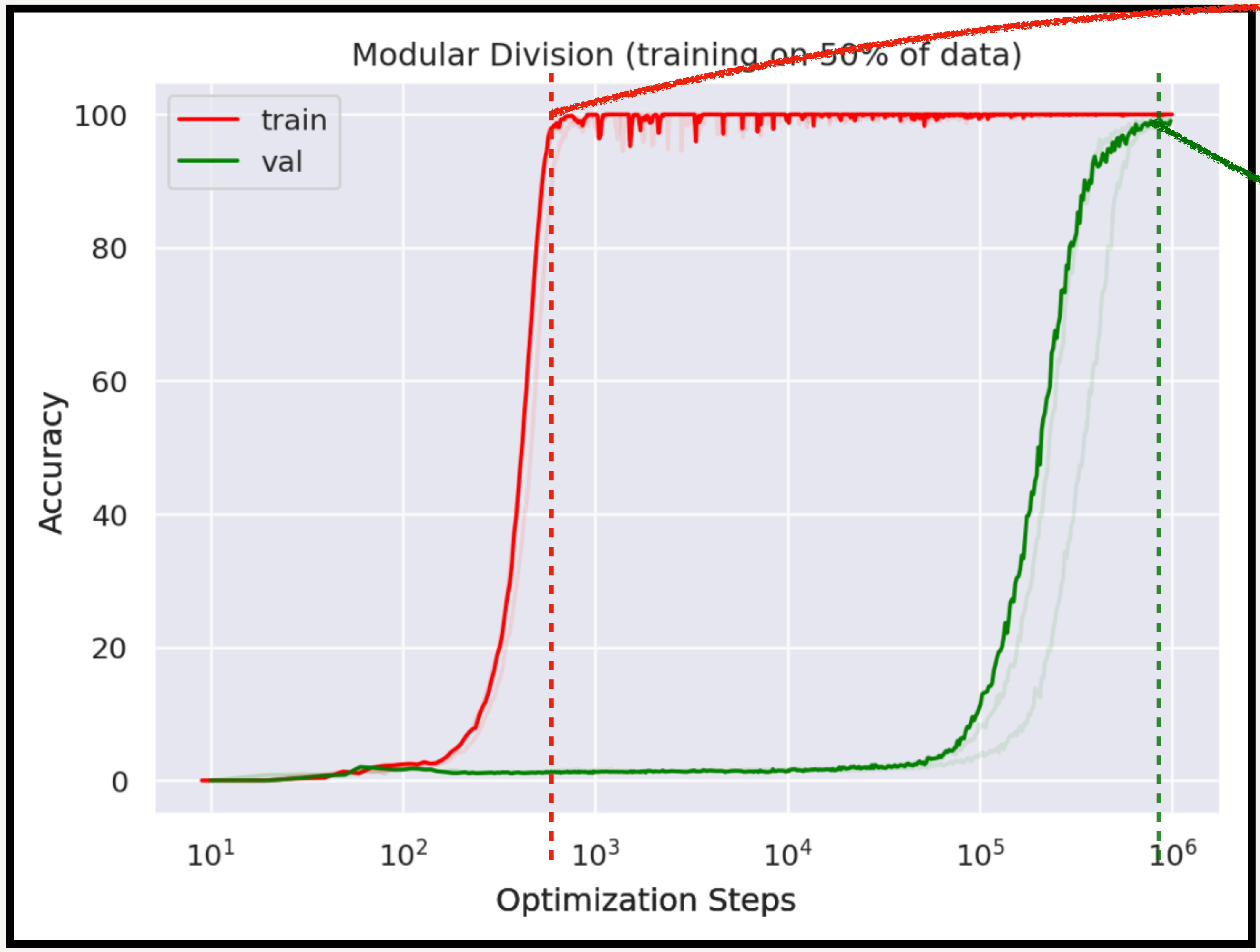




Liu et al , "Towards understanding grokking: An effective theory of representation learning"

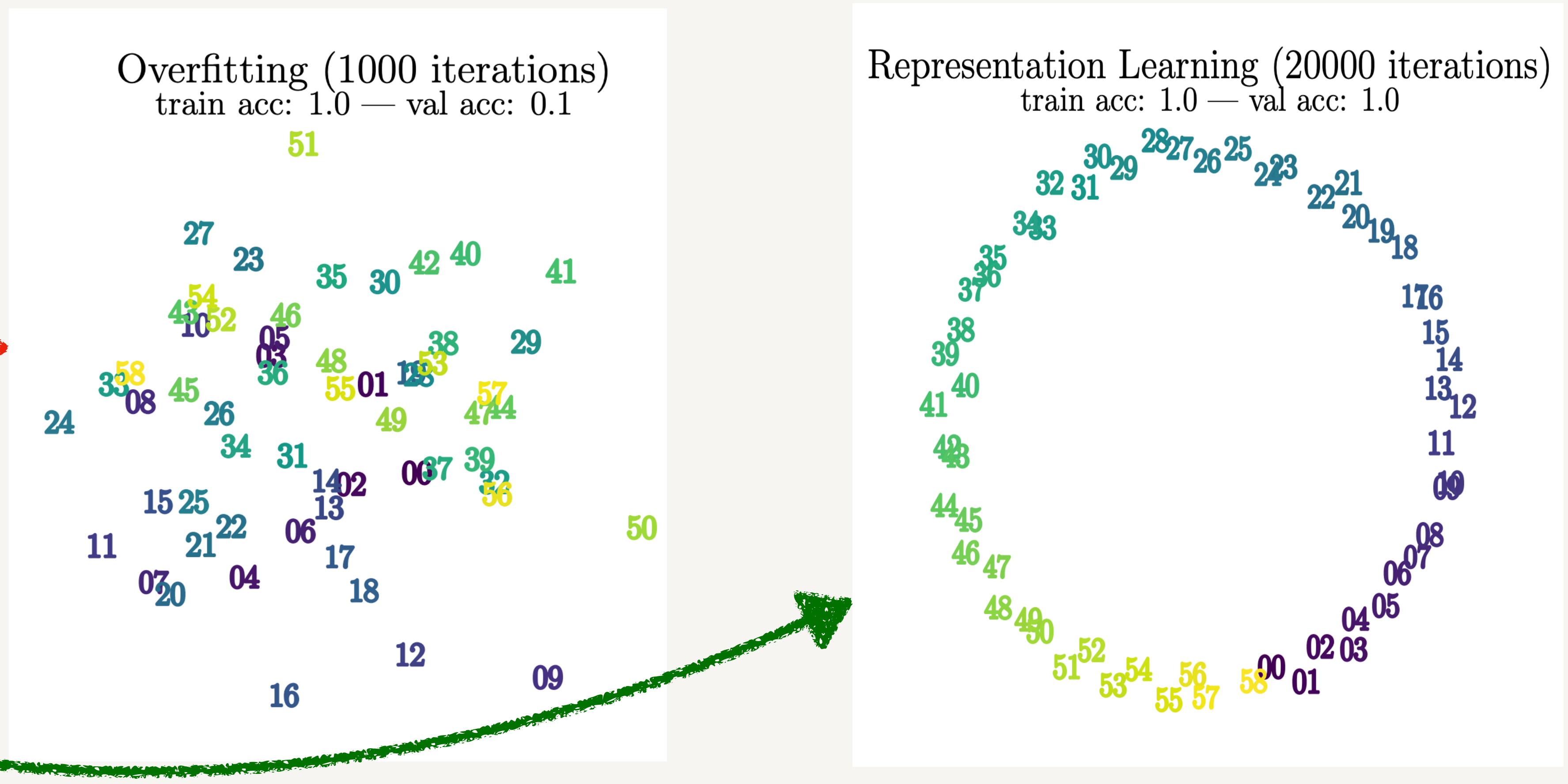
Power et al , "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets"

Grokking



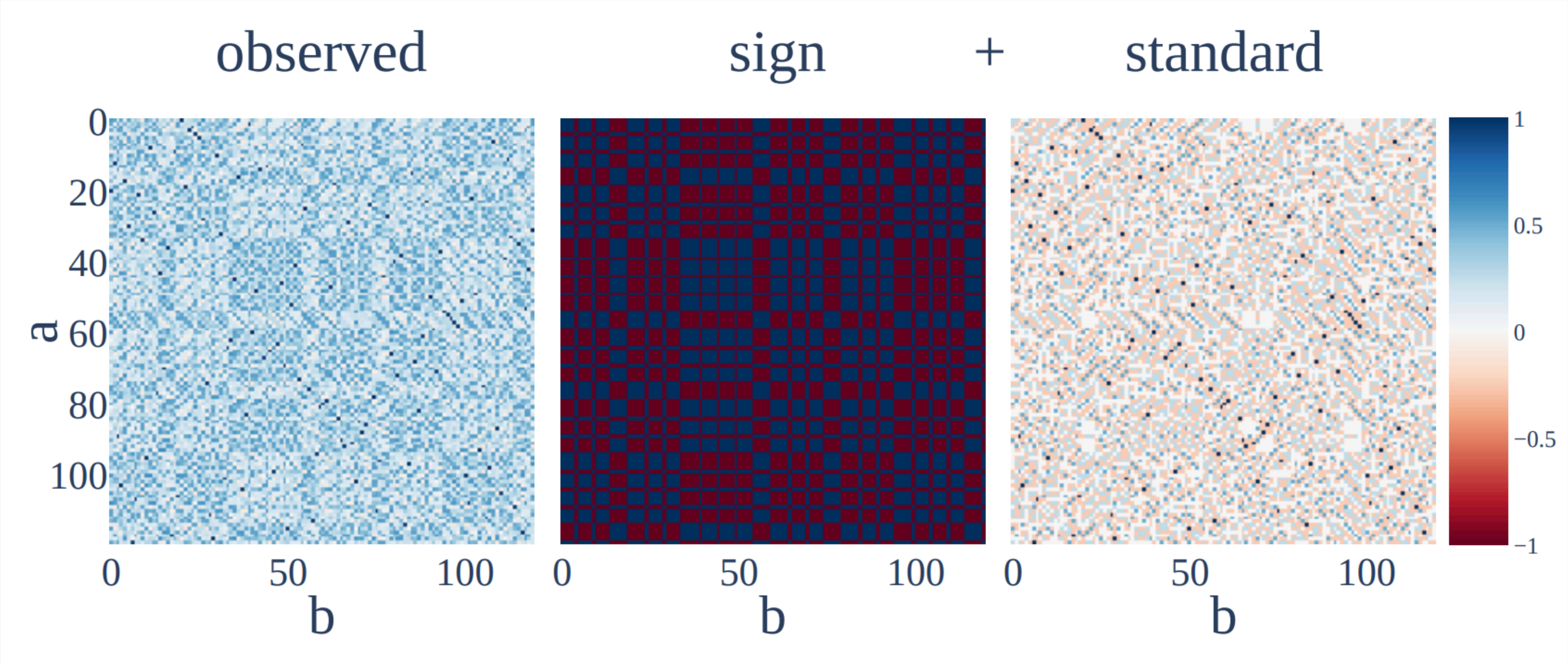
Power et al , "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets"

Representation is key for generalisation!



Liu et al , "Towards understanding grokking: An effective theory of representation learning"

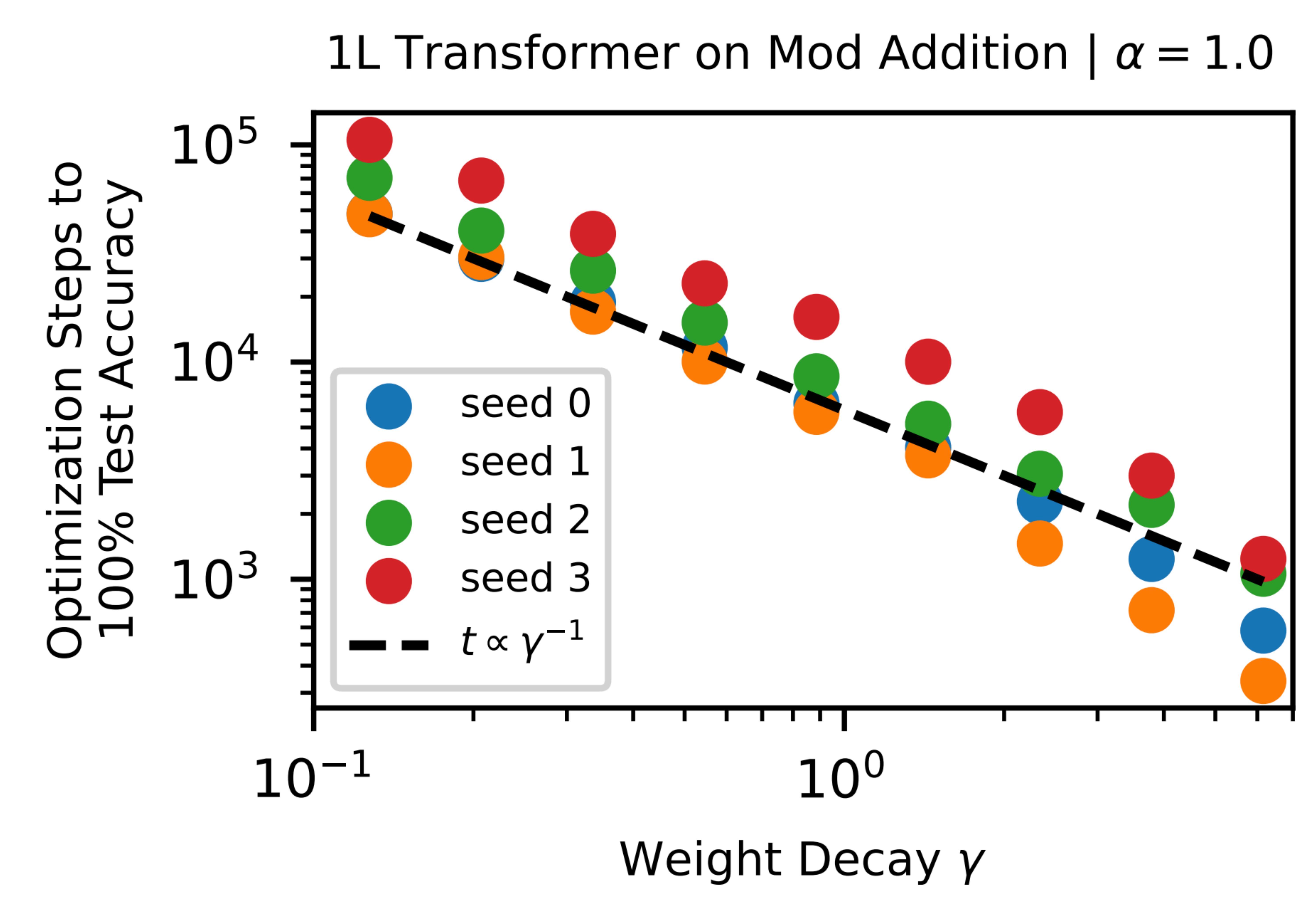
For general groups, learned representations are group representations.



Chughtai, Chan & Nanda , "A Toy Model of Universality: Reverse Engineering How Networks Learn Group Operations"

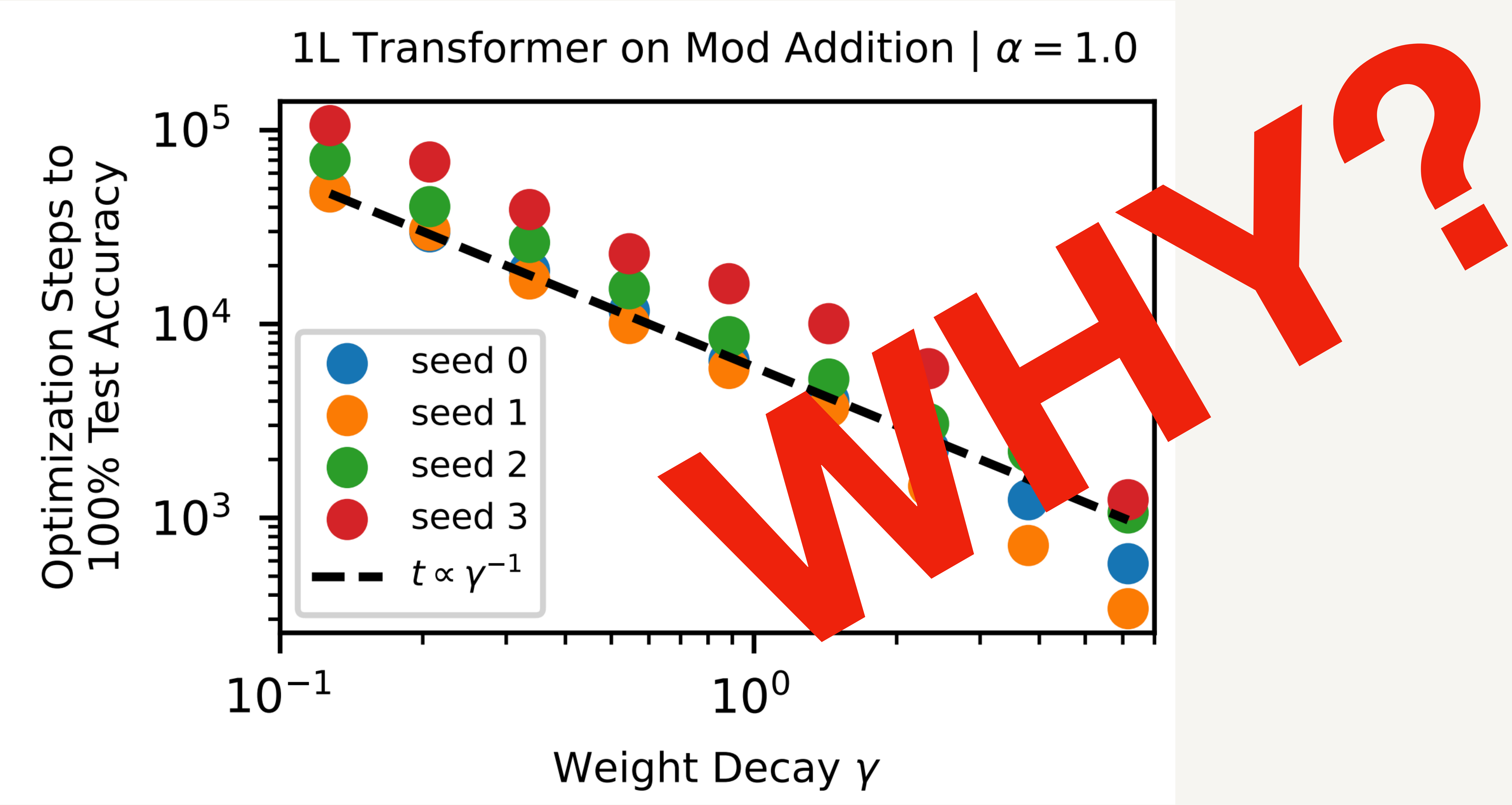
Q: Under which conditions can representations emerge, hence generalisation happens?

Larger weight decay => faster generalisation



Liu, Michaud & Tegmark "Omnigrok: Grokking Beyond Algorithmic Data"

Larger weight decay => faster generalisation

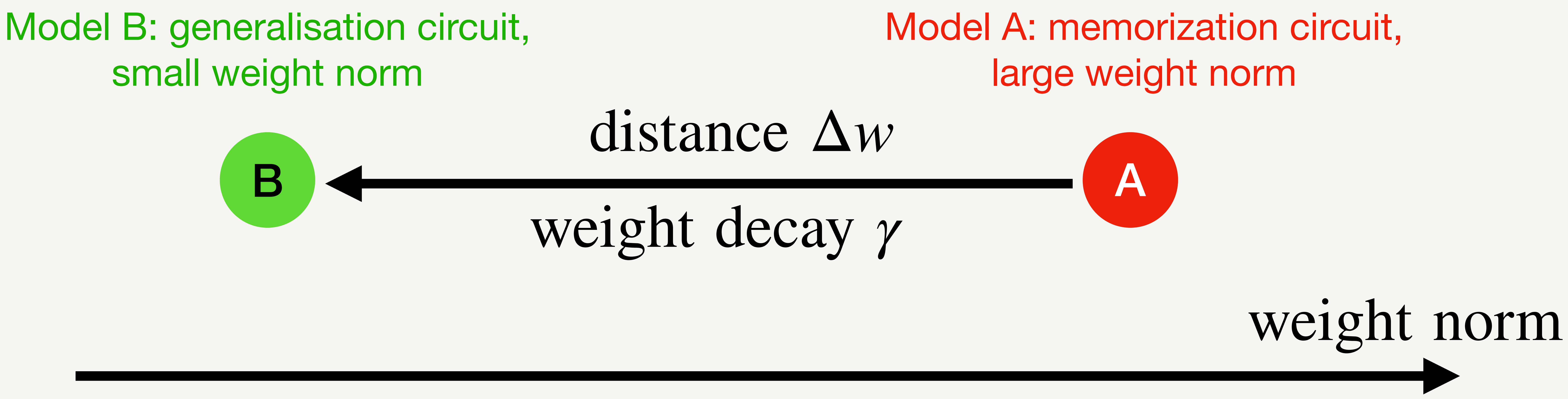


Liu, Michaud & Tegmark "Omnigrok: Grokking Beyond Algorithmic Data"

What we know in elementary school …

The time to travel from city A to city B is $t = \dfrac{d}{v} \propto v^{-1}$

In the grokking case

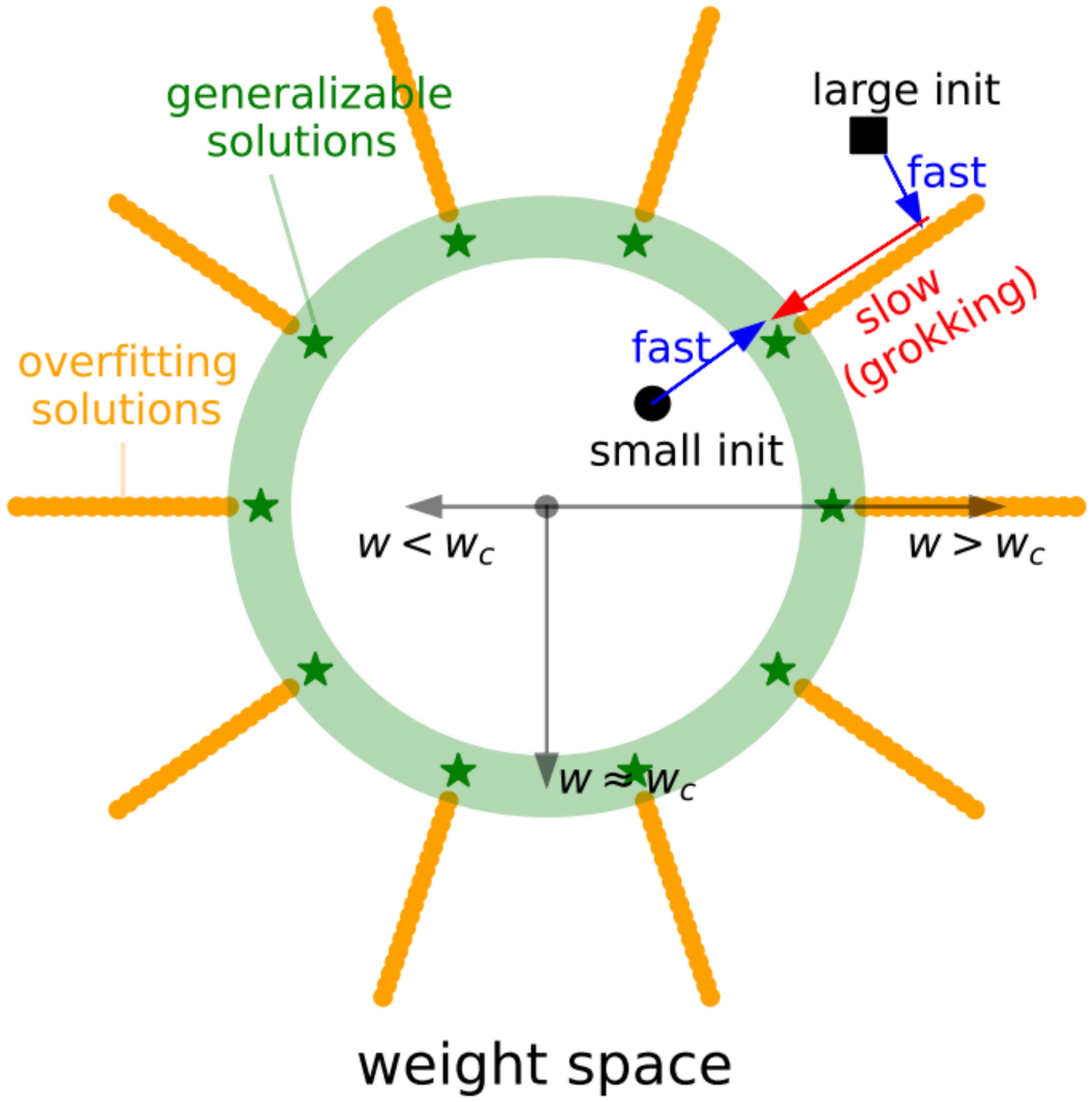


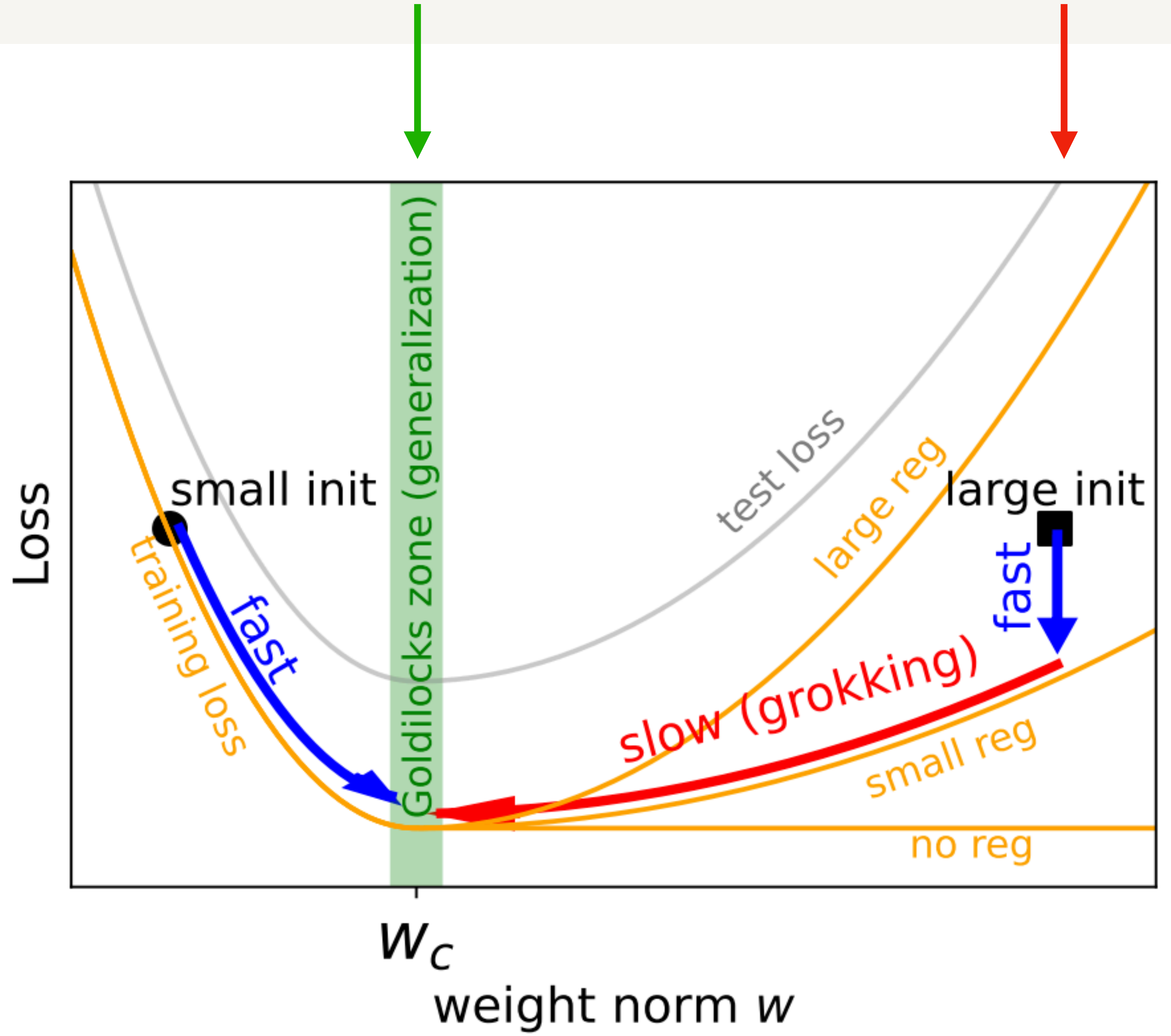
The time to travel from model A to model B is $t = \dfrac{\Delta \log w}{\gamma} \propto \gamma^{-1}$

Weight norm & LU mechanism

Model B: generalisation, small weight norm

Model A: memorization, large weight norm

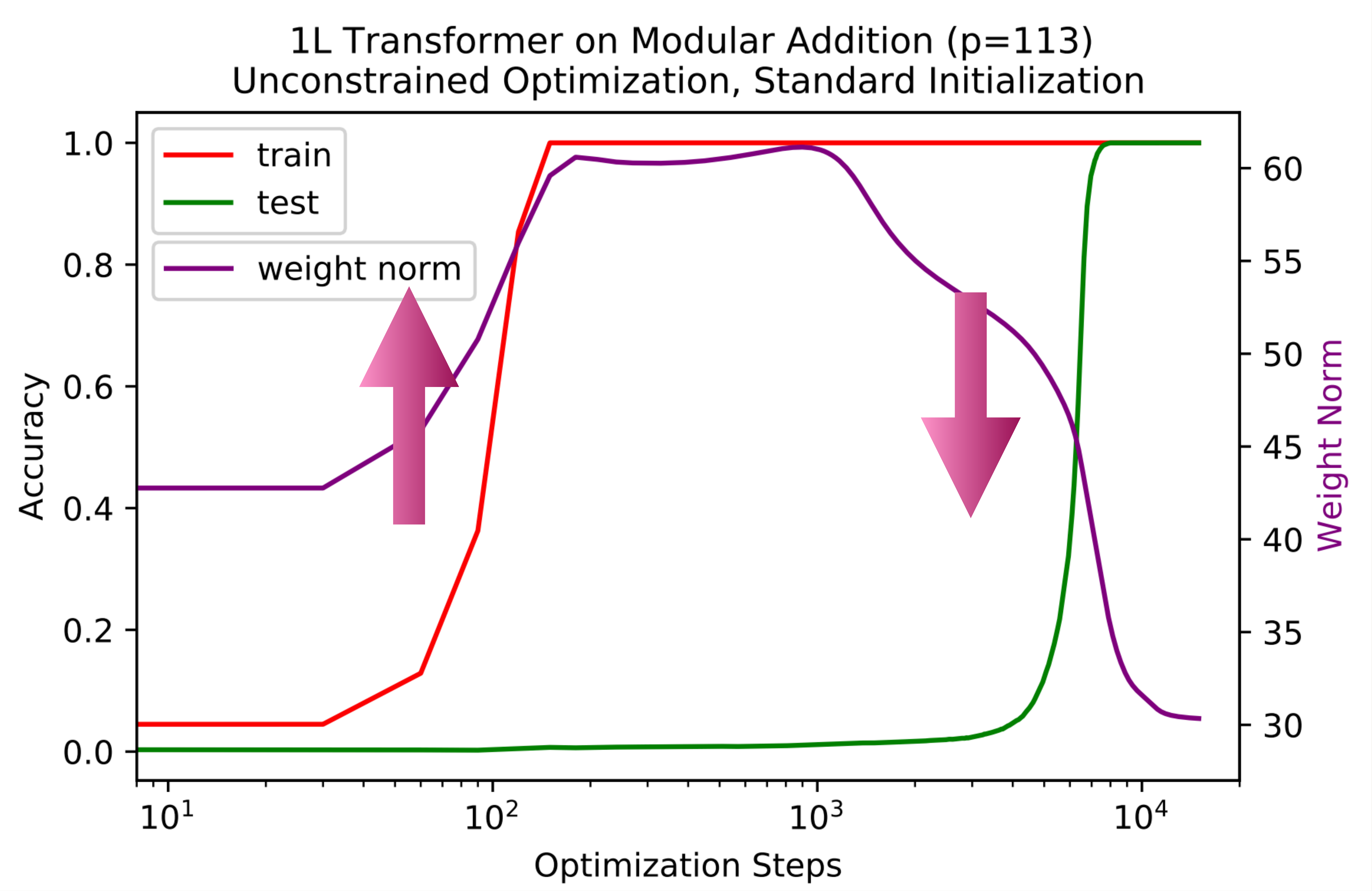




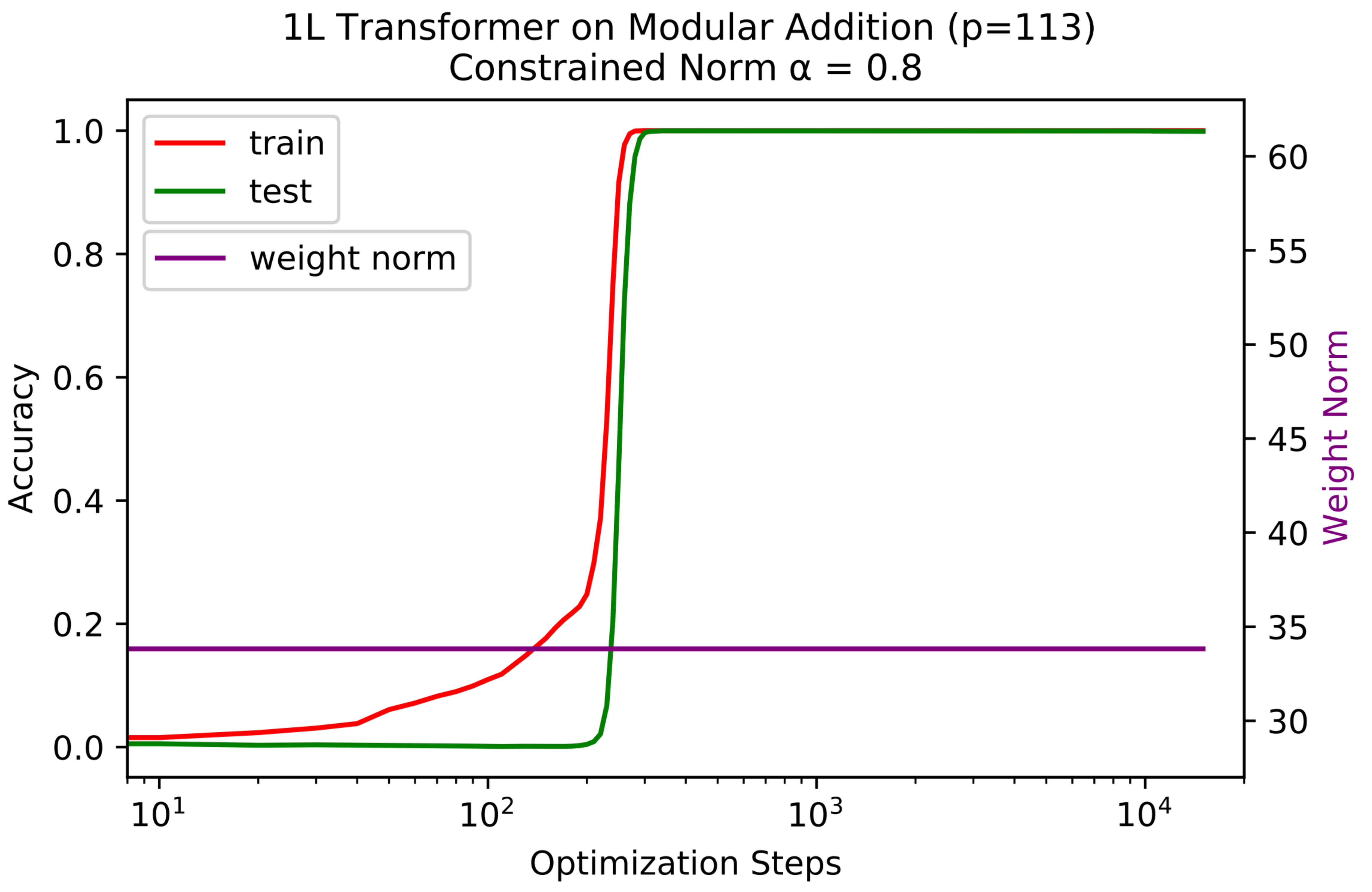
Liu, Michaud & Tegmark "Omnigrok: Grokking Beyond Algorithmic Data"

Eliminate grokking by constraining weight norm

Weight norm increases (overfitting), then decreases (generalisation)



Constraining weight norm eliminates grokking

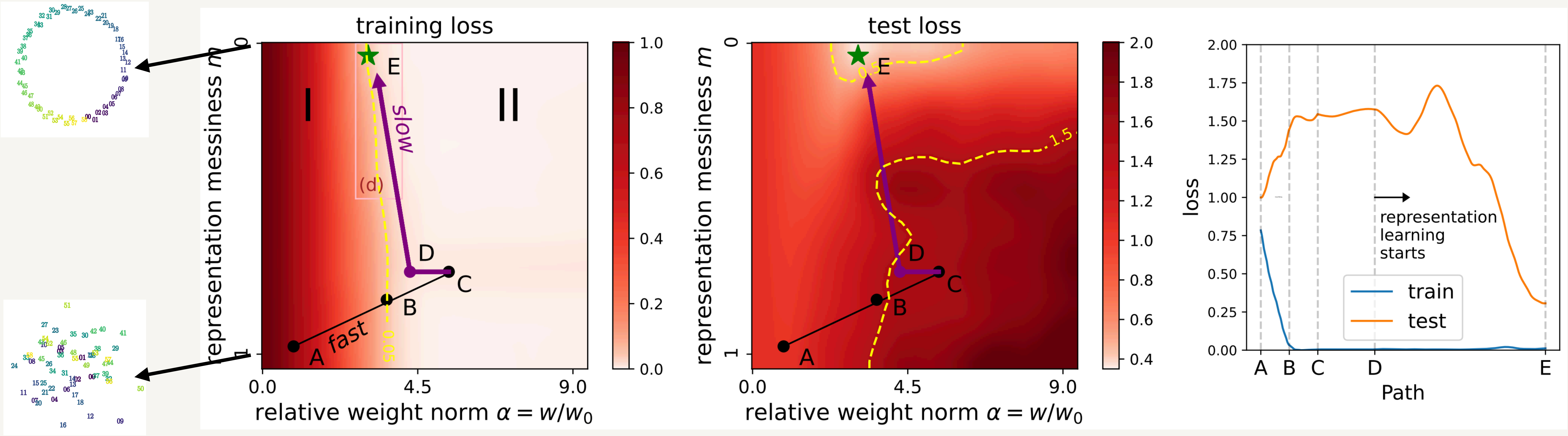


Liu, Michaud & Tegmark "Omnigrok: Grokking Beyond Algorithmic Data"

Weight norm and representation learning

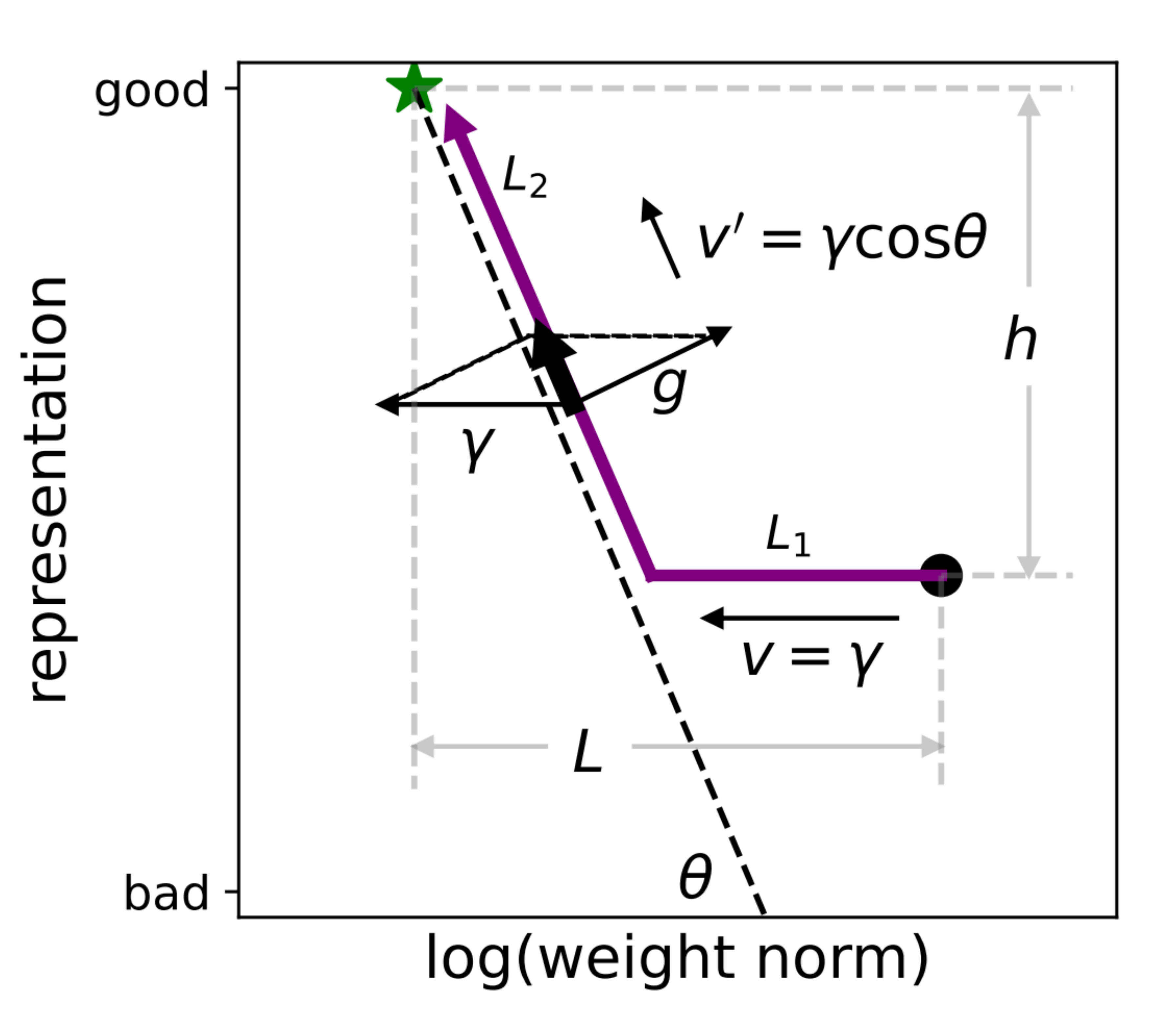
Q: Why does weight norm increase at first (despite weight decay)?

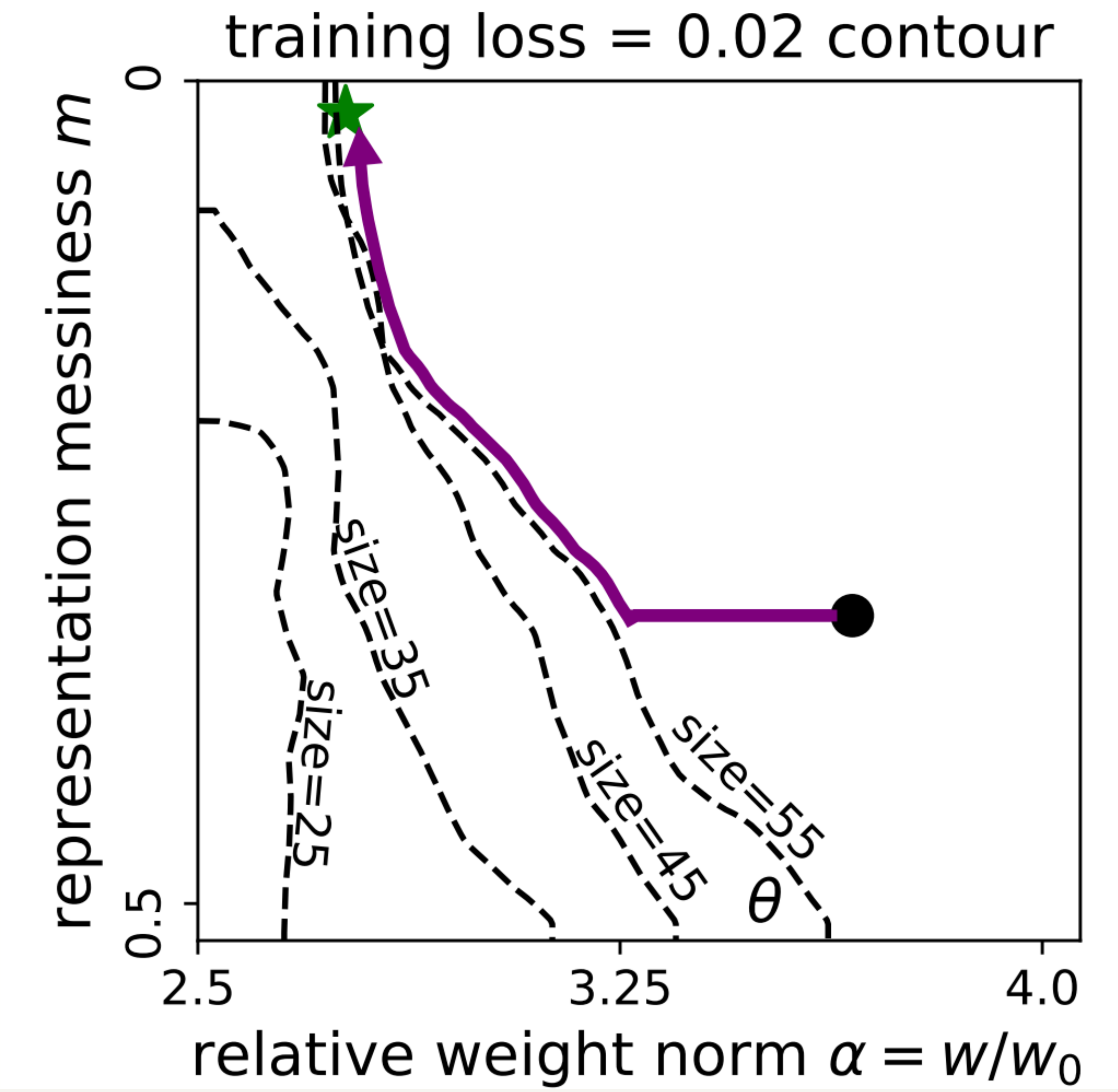
A: Again, we need to bring representation back into the whole picture!



Liu, Michaud & Tegmark "Omnigrok: Grokking Beyond Algorithmic Data"

Weight norm and representation learning





$$t = \frac{L + h\tan\theta}{\gamma}$$

data size $\uparrow \rightarrow \theta \downarrow \rightarrow t \downarrow$

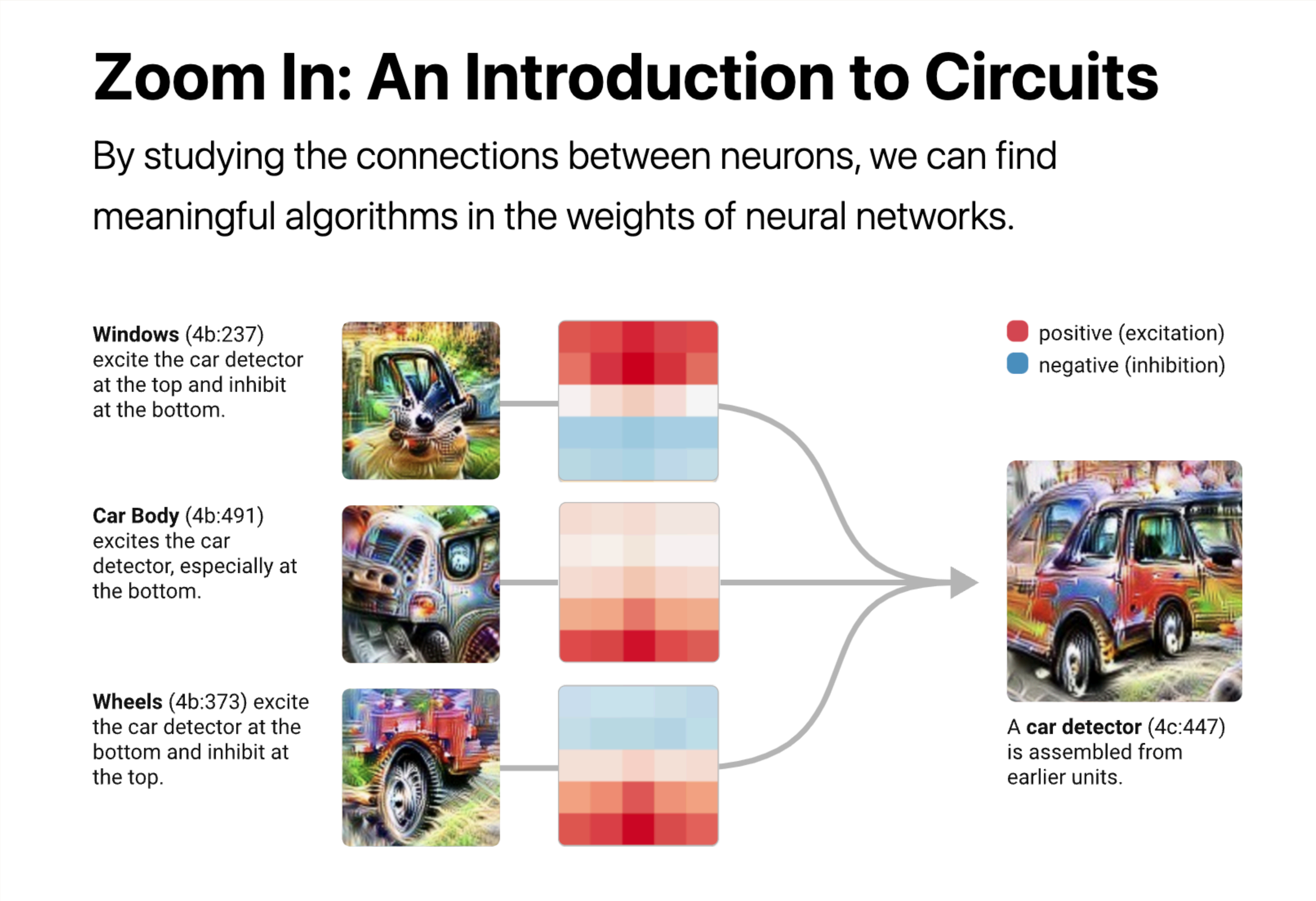
$\gamma \uparrow \rightarrow t \downarrow$

Liu, Michaud & Tegmark "Omnigrok: Grokking Beyond Algorithmic Data"

Modularity

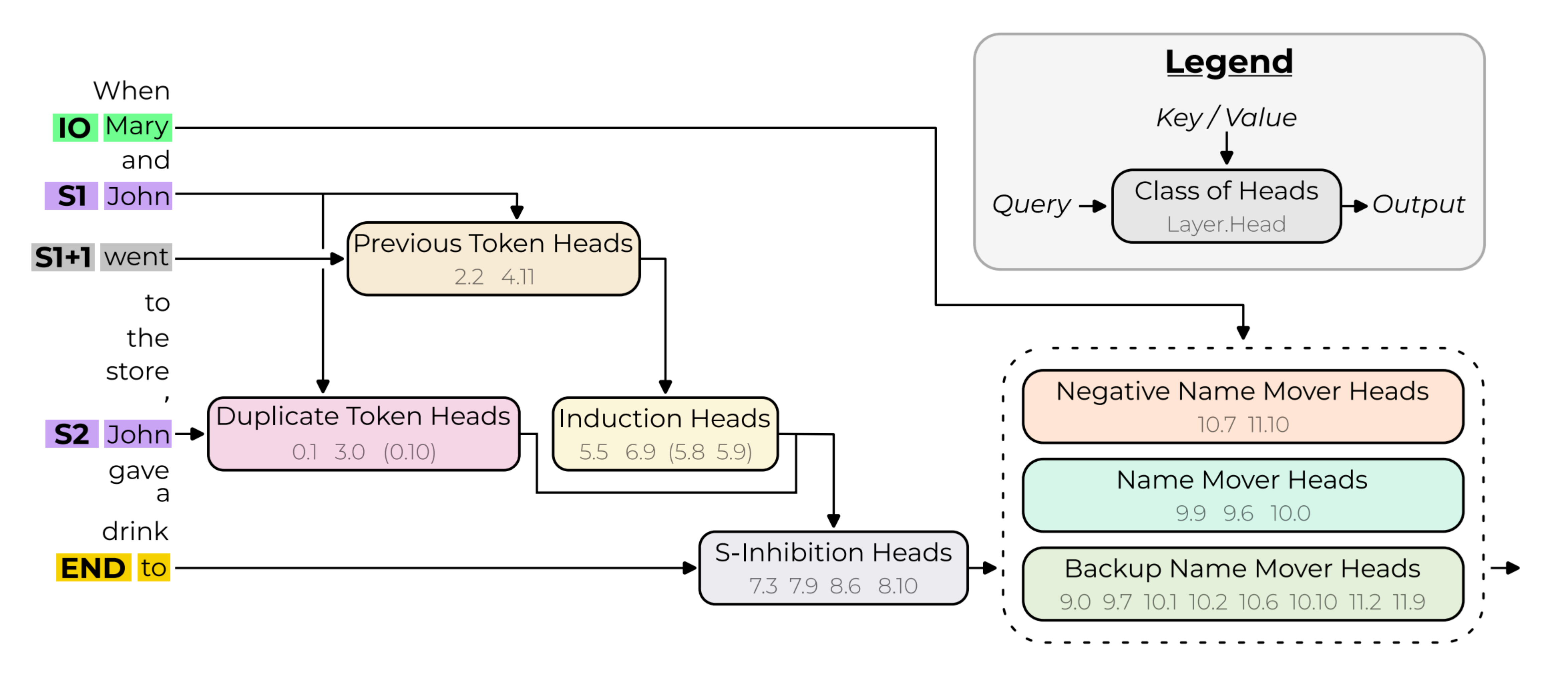
Neural Network Modularity/Circuitry

Vision



Olah et al., "Zoom in: An Introduction to Circuits"

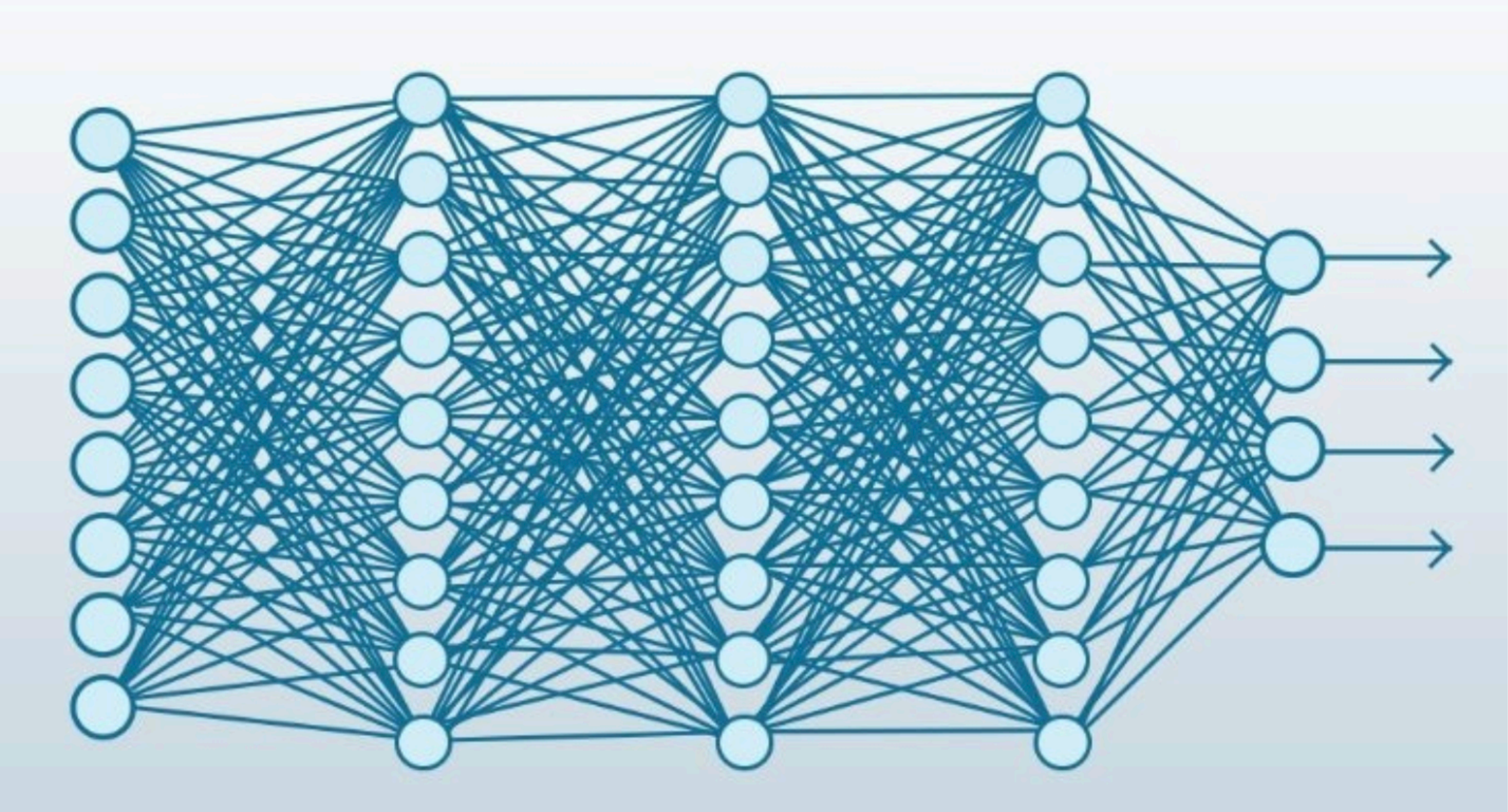
Language



Wang et al., "Interpretability in the wild: A circuit for indirect object identification in GPT-2 small"

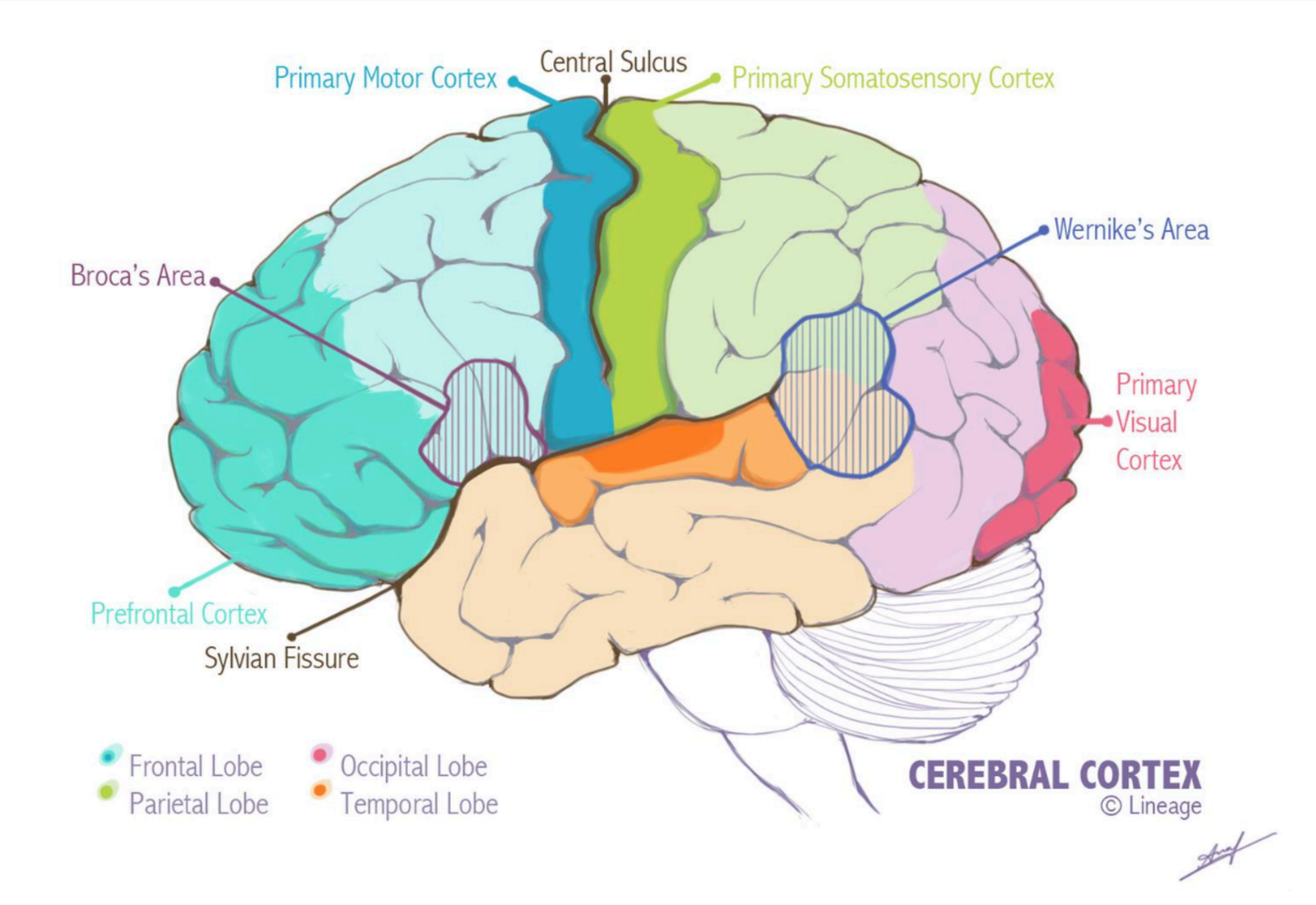
Neural networks vs brains

Neural networks



???

Brains

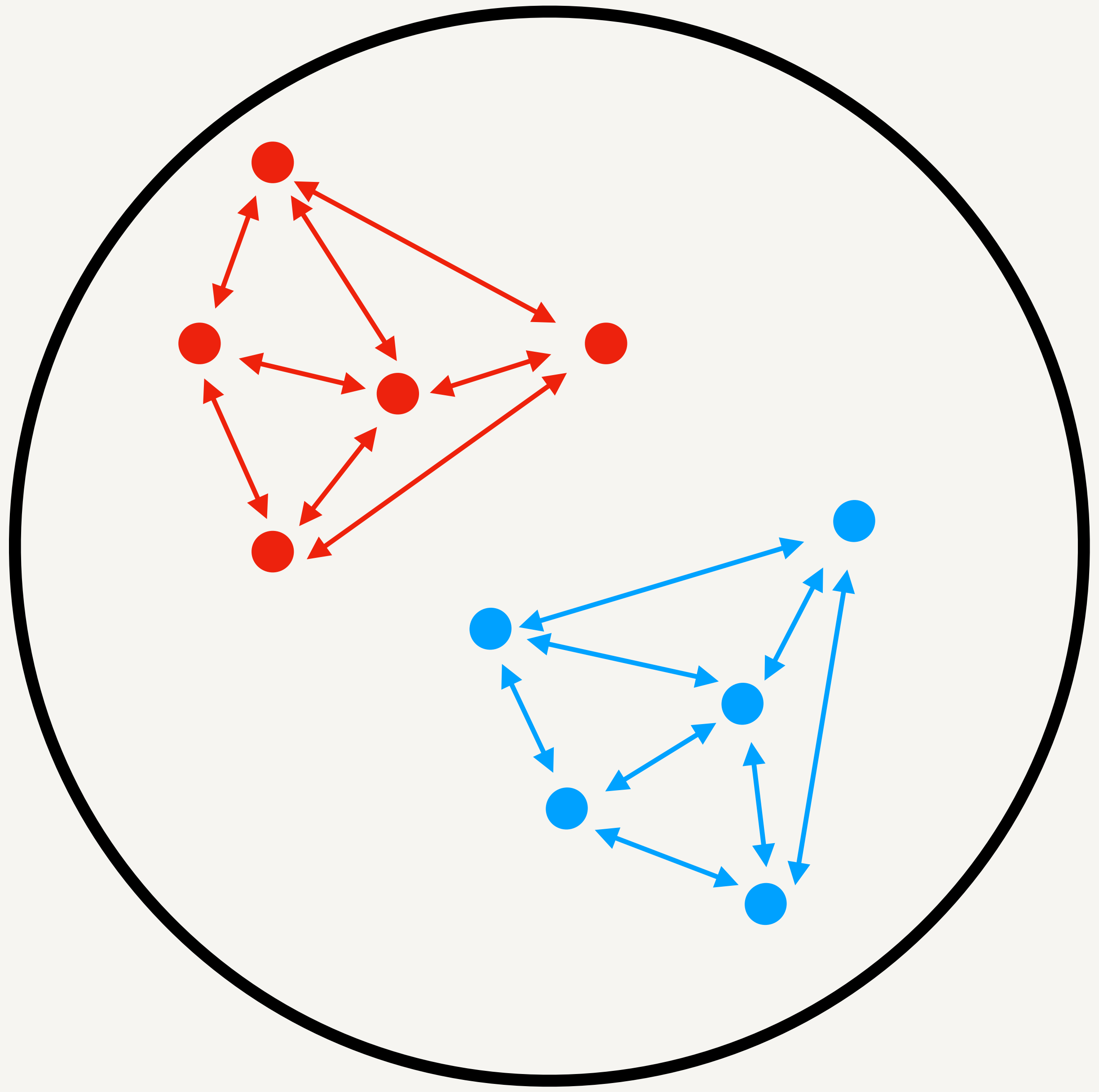


But, there’s a key difference between brains and neural networks …

Modular brains have survival advantages, but modular NNs don't

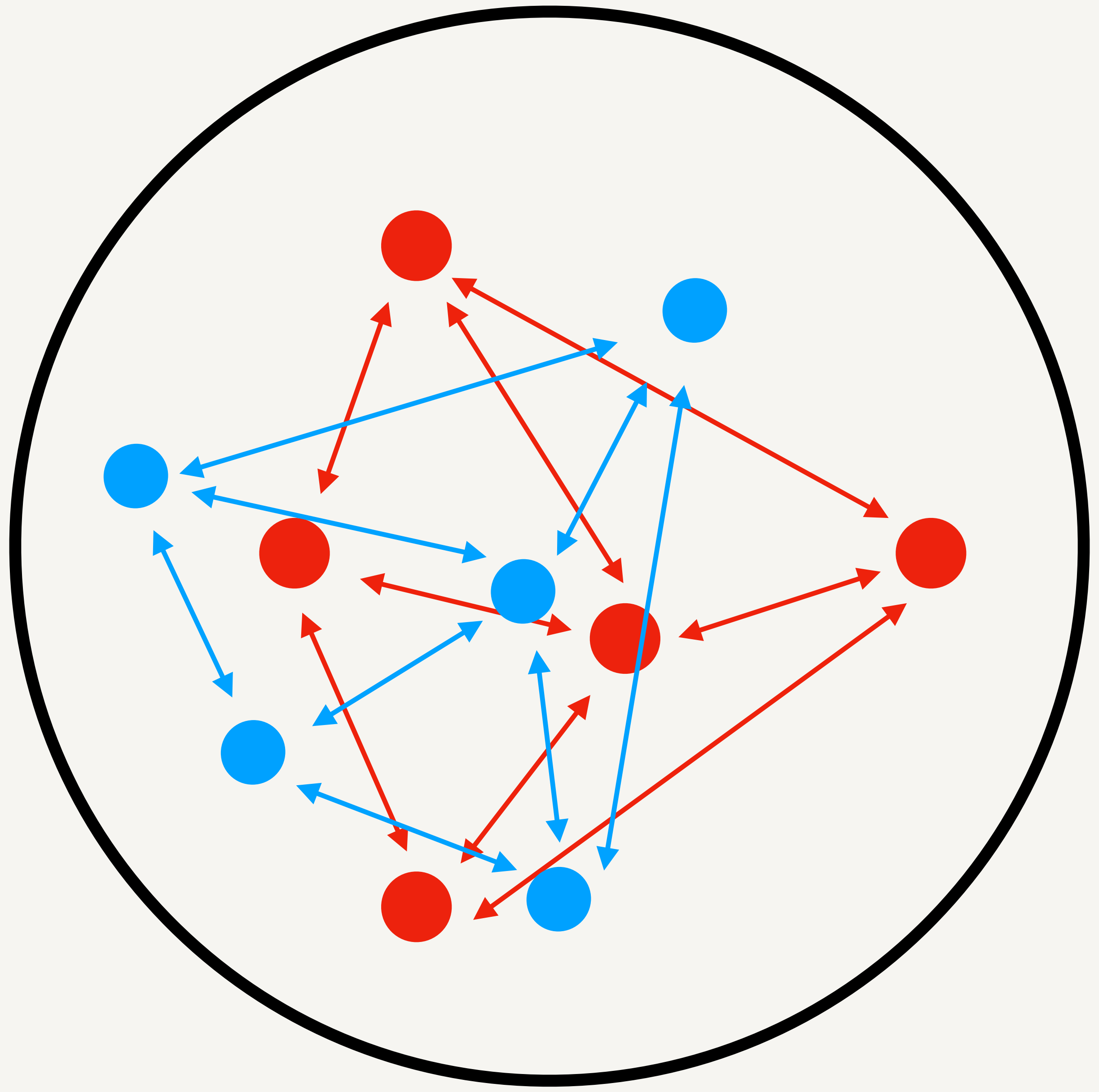
When humans deal with a specific task …

Modular brains



Relevant neurons are local
Shorter neuron connections
React faster
More likely to survive

Non-Modular brains



Relevant neurons are non-local
Longer neuron connections
React slower
Less likely to survive

Q: Do modular neural networks have “survival advantages”?

A: No! Because there is no (explicit) incentive for artificial neural networks to become modular if it only cares about prediction.

Q: What training techniques can induce modularity in otherwise non-modular networks?

A: Need to introduce “locality” and **limit resources (hunger)**!

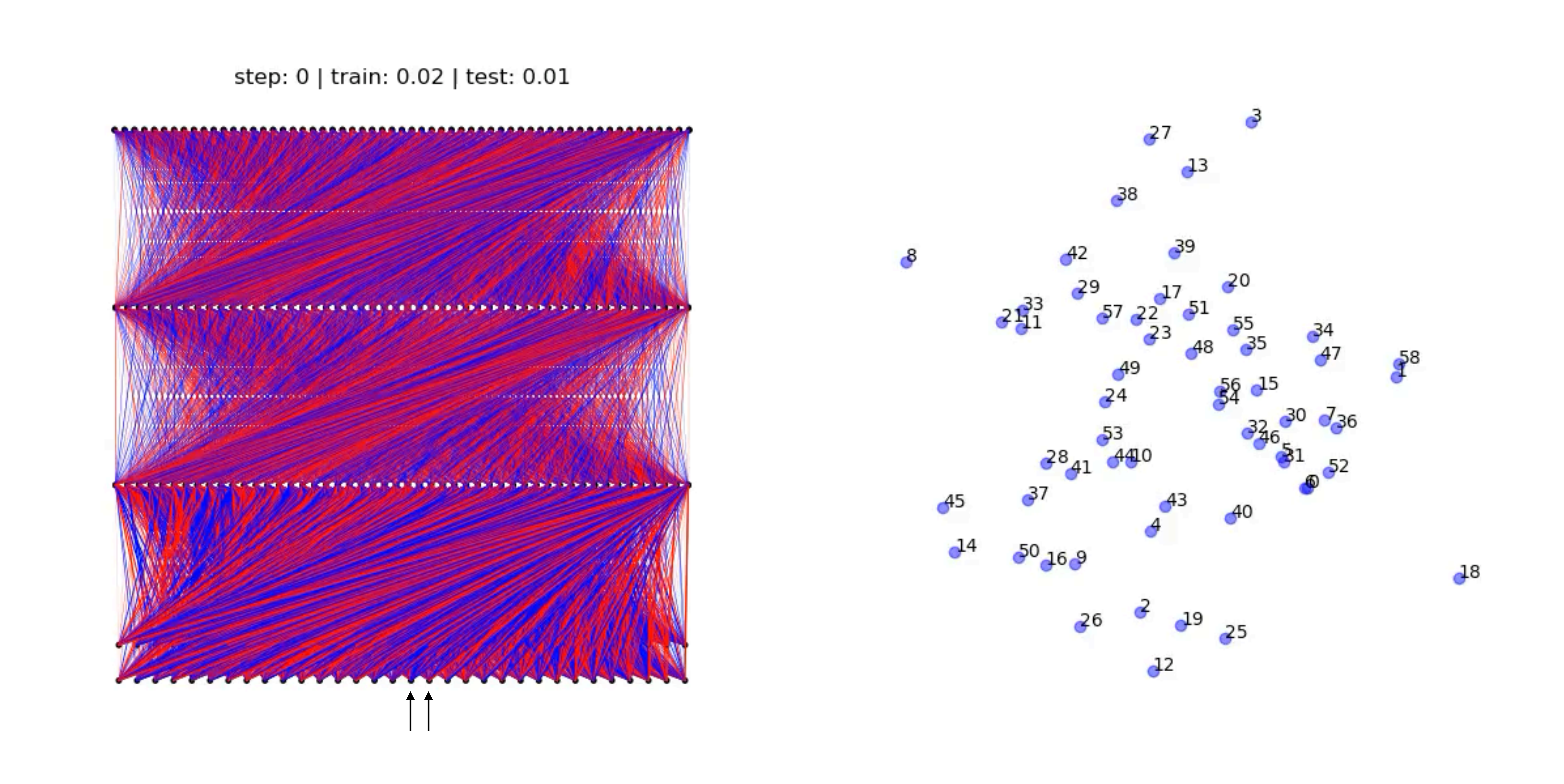
Liu, Gan & Tegmark “Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability” https://arxiv.org/abs/2305.08746

Brain-inspired modular training (BIMT)

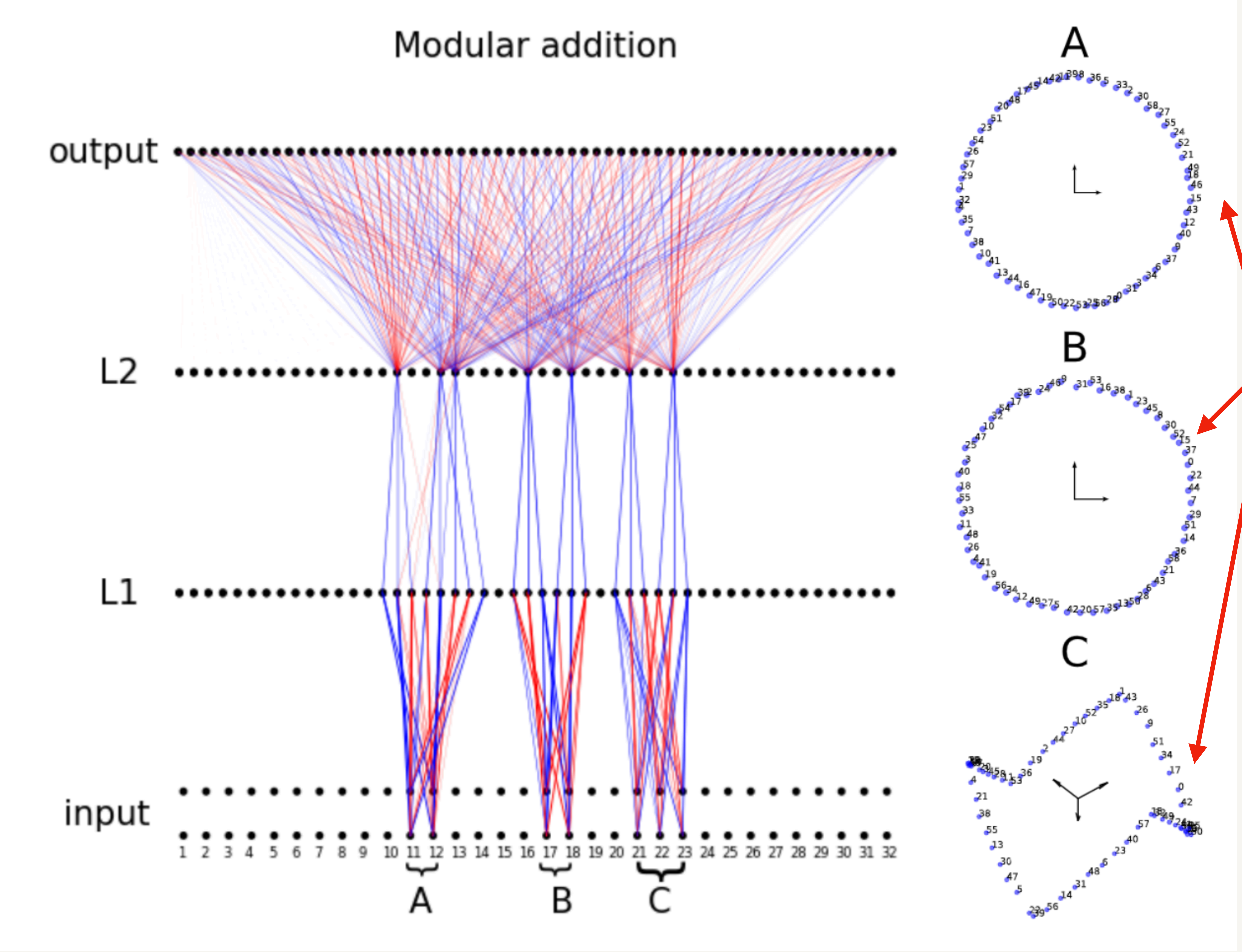
Liu, Gan & Tegmark "Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability"
https://arxiv.org/abs/2305.08746

Modular addition

blue/red stands for positive/negative weights



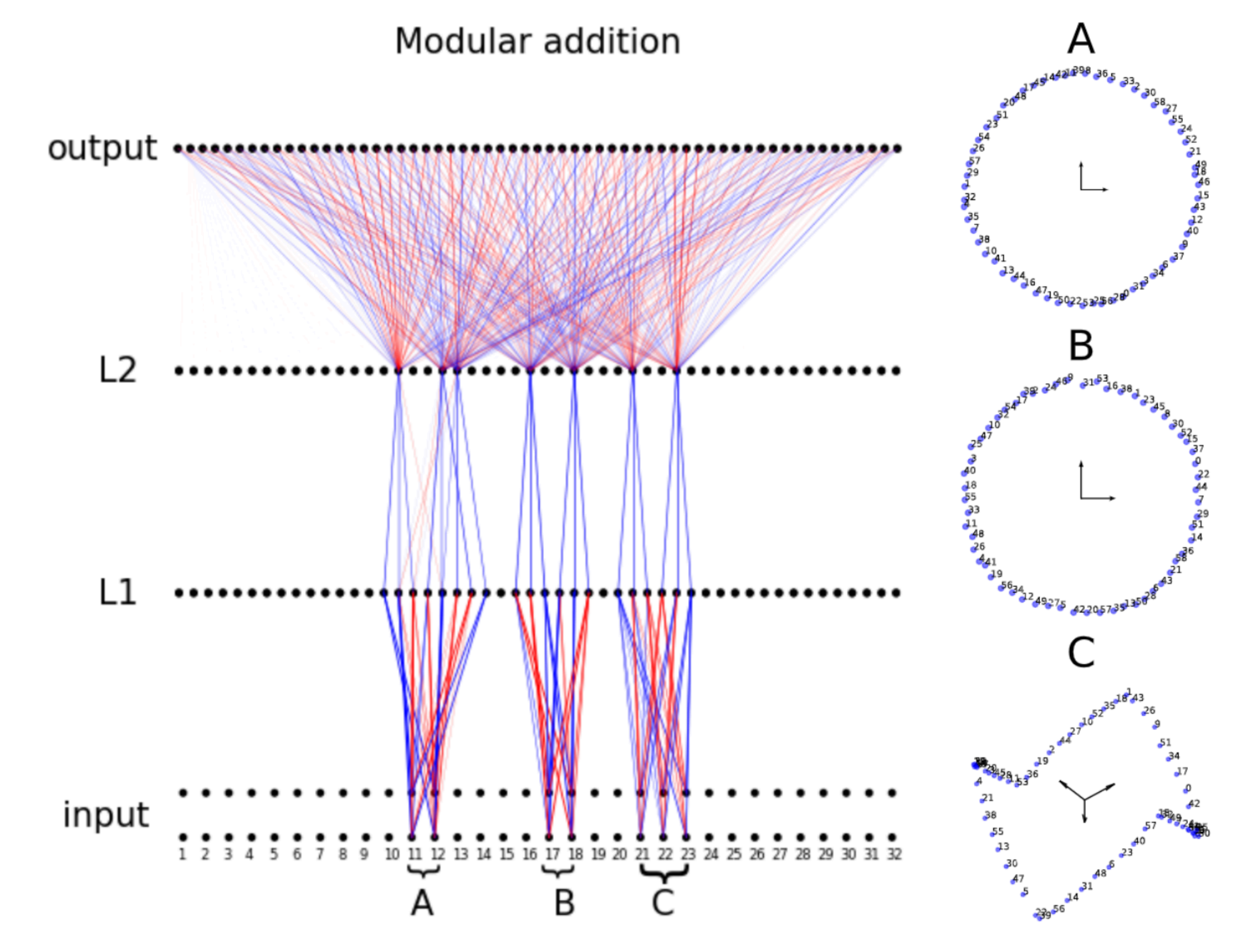
Representations emerge on privileged bases



Representations emerge on privileged bases

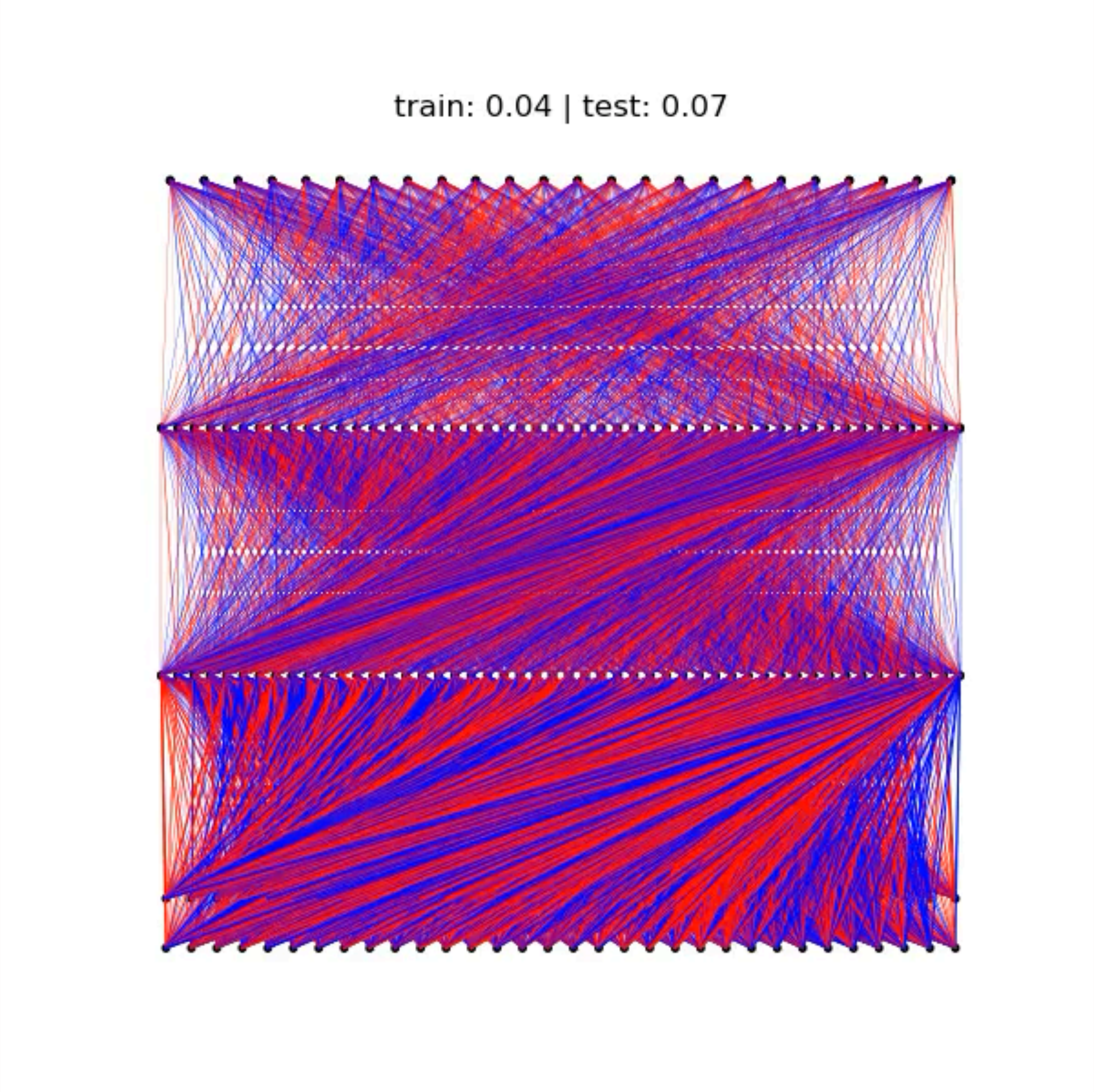
No need to
search for directions or
do PCA!

Voting mechanism

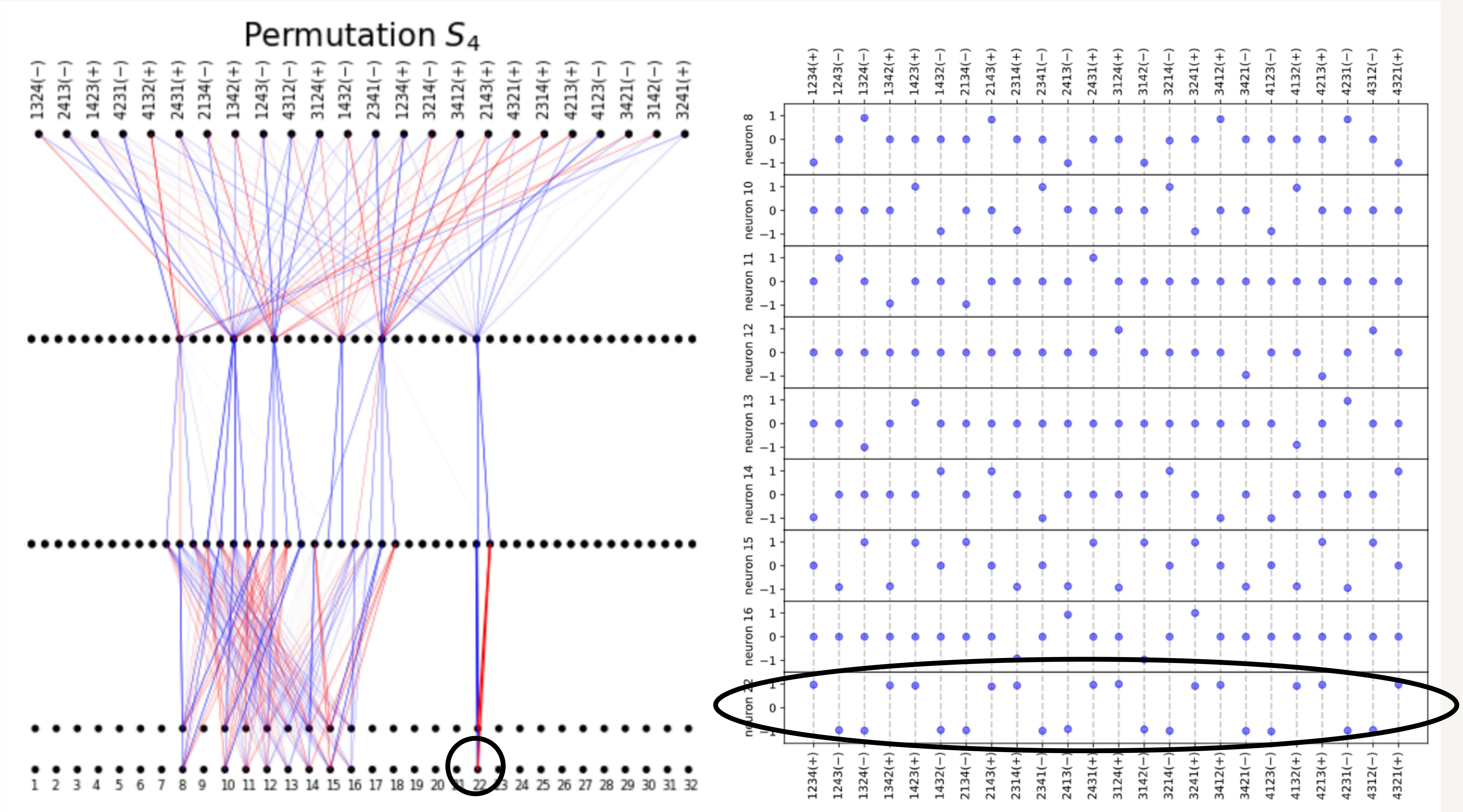


Knockout	Accuracy
None	100.00%
A	15.25%
B	29.33%
C	33.67%
A, B	3.39%
A, C	5.08%
B, C	10.28%
A, B, C	1.69%
A10	47.11%
A11	46.51%
B16	50.47%
B17	51.42%
C20	77.10%
C21	73.60%
C22	78.17%
All but A, B, C	100.00%

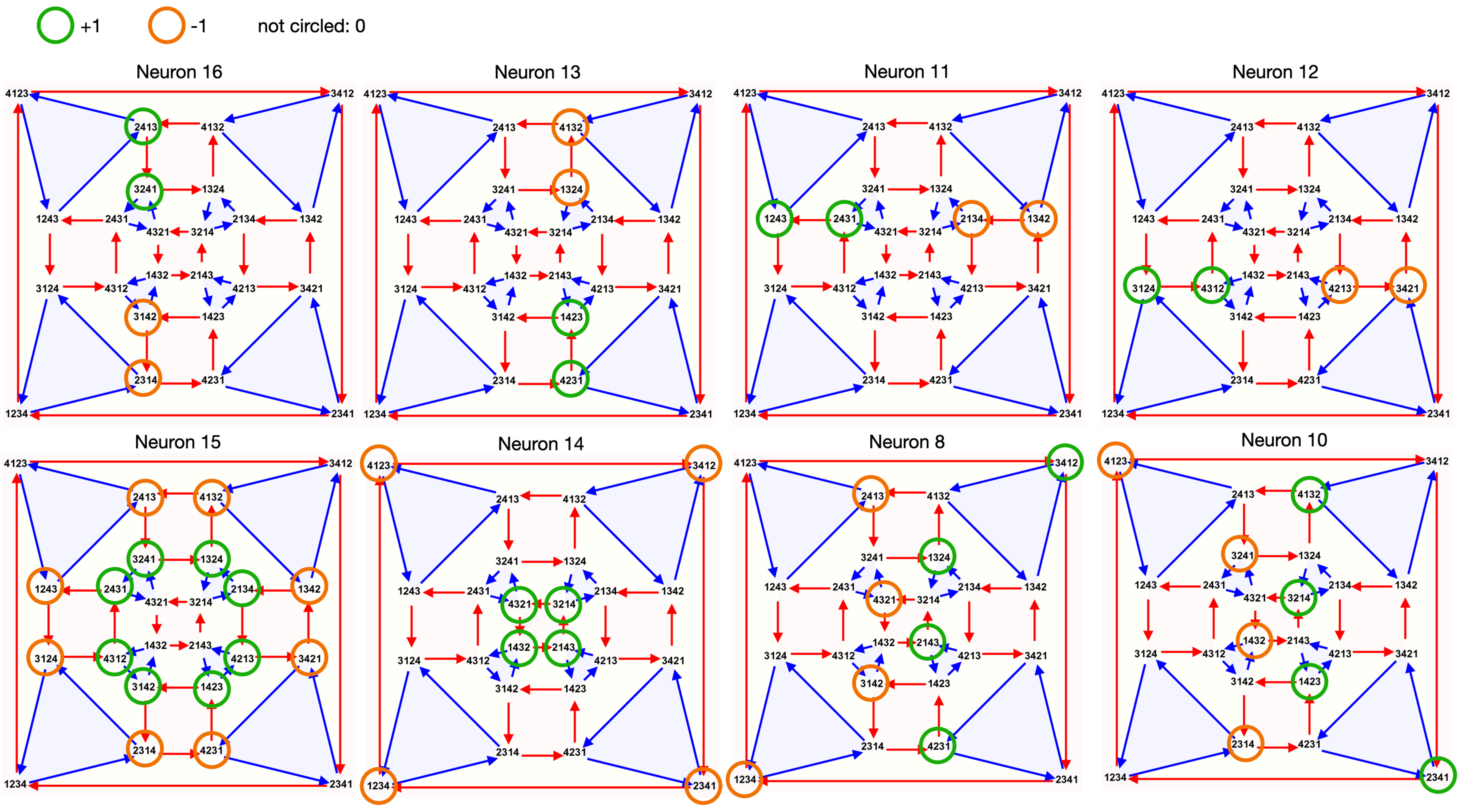
Permutation S4



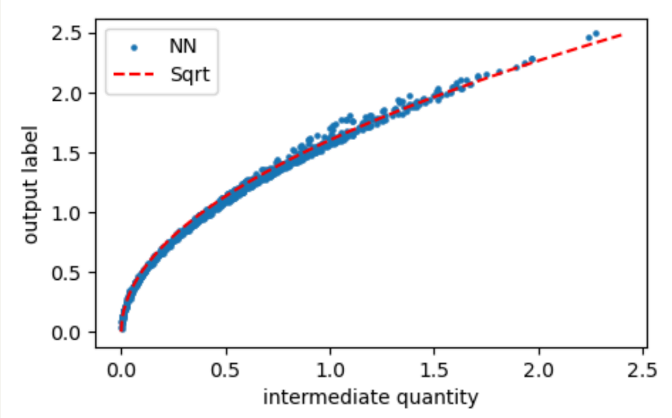
Permutation S4



Visualising neurons with Cayley graphs



Symbolic formulas

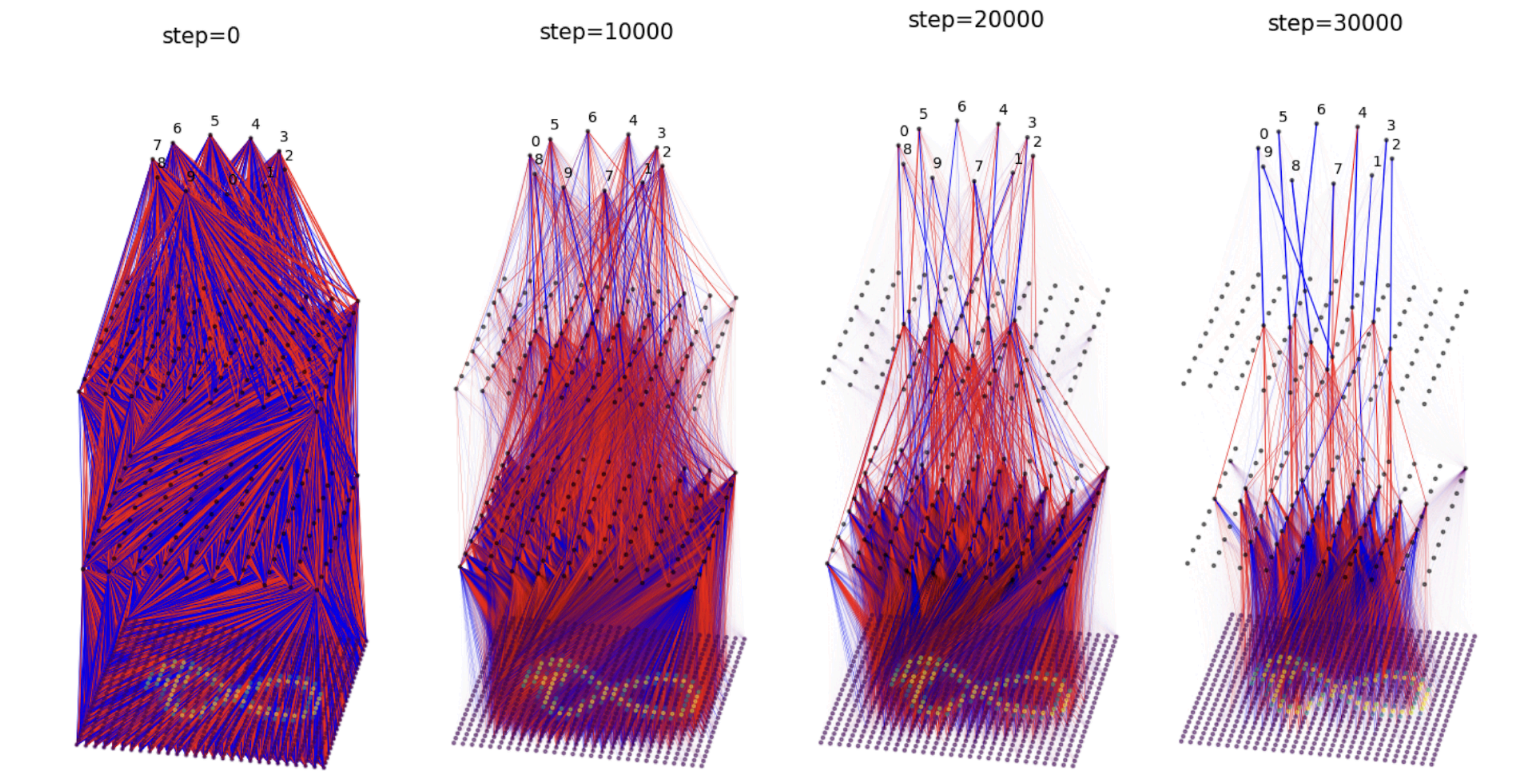


(a) independence

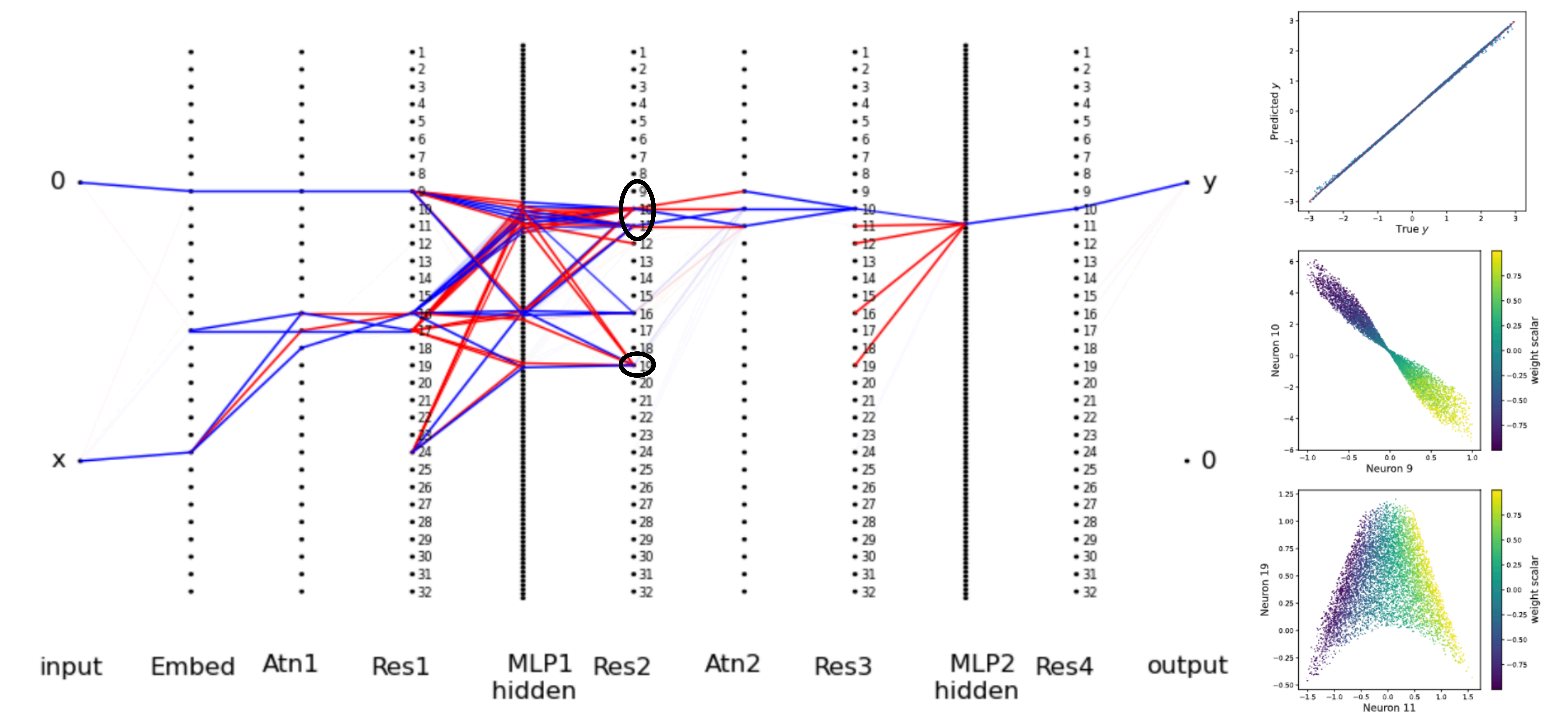
(b) feature sharing

(c) compositionality

MNIST



Transformer + in-context linear regression



Task from Akyurek et al, "Which learning algorithm is in-context learning? Investigations with linear models"

Plot twist: LLM

Neural Scaling Laws (NSL)

arXiv > cs > arXiv:2001.08361

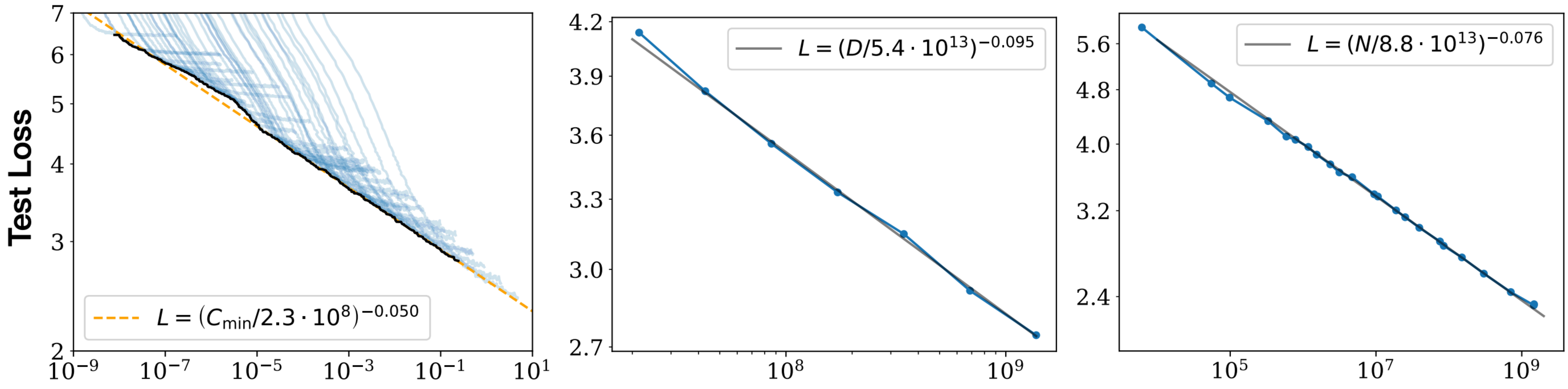
Search...
Help | Advanced

Computer Science > Machine Learning

[Submitted on 23 Jan 2020]

Scaling Laws for Neural Language Models

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei



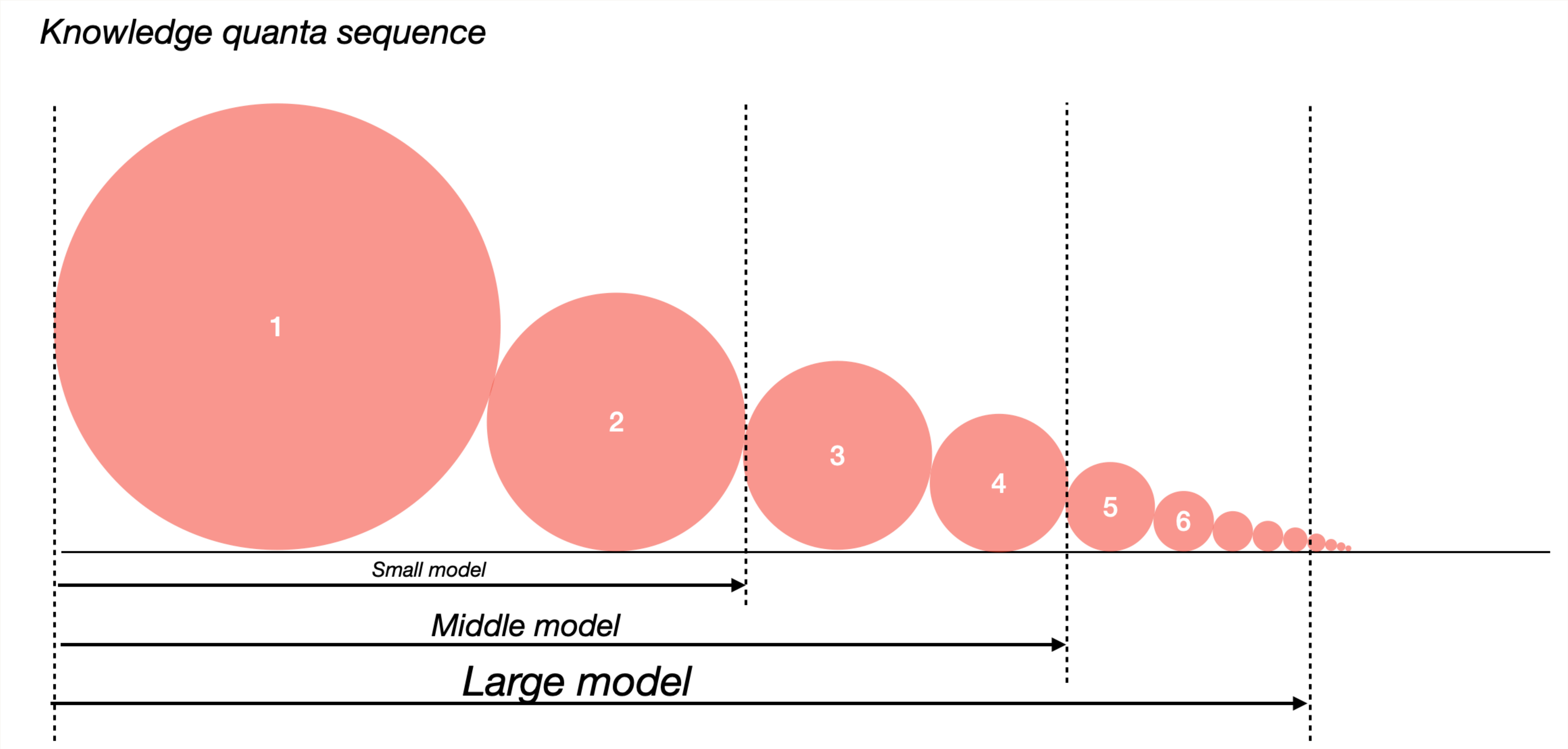
Compute
PF-days, non-embedding

Dataset Size
tokens

Parameters
non-embedding

LLM seem to contradict intelligence from hunger?

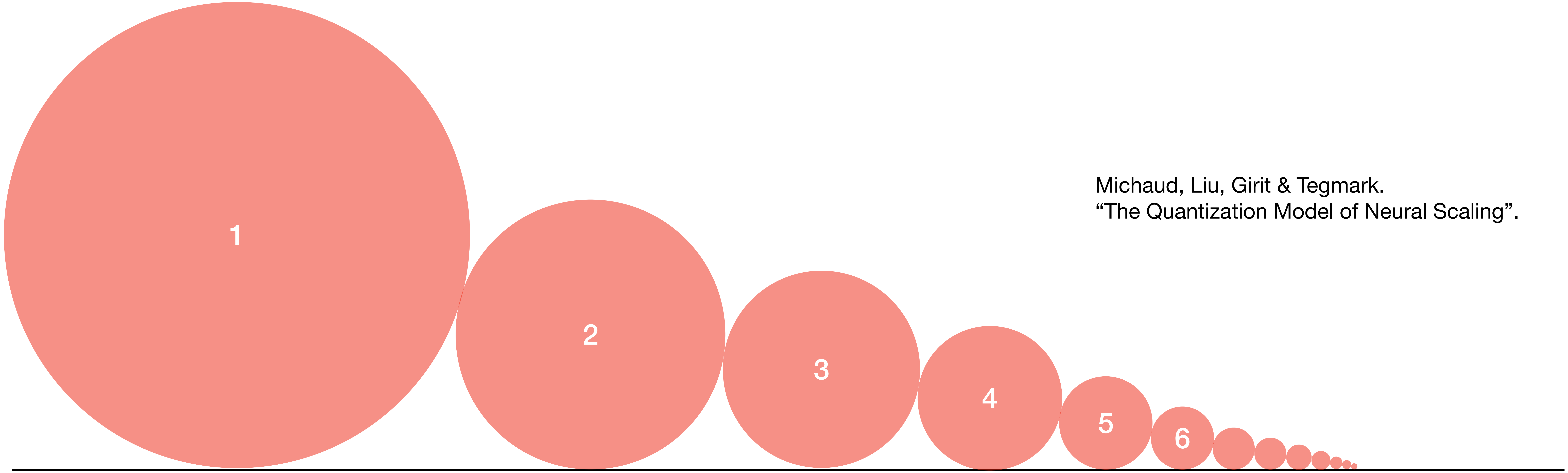
My understanding: Current LLM are still underfitting (too hungry).



Michaud, Liu, Girit & Tegmark.
"The Quantization Model of Neural Scaling".

Quantization Hypothesis

Knowledge quanta sequence

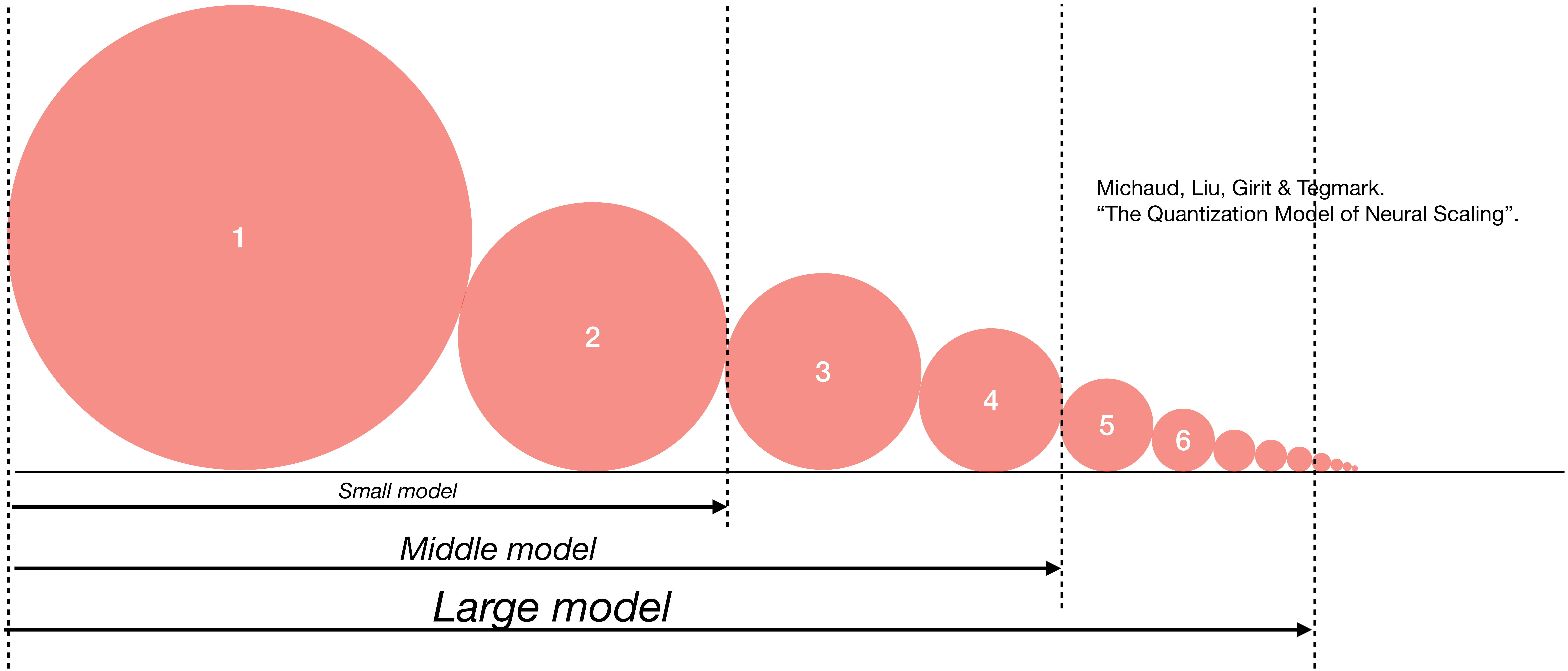


Michaud, Liu, Girit & Tegmark.
"The Quantization Model of Neural Scaling".

Size = Frequency (Importance)

Quantization Hypothesis

Knowledge quanta sequence



Michaud, Liu, Girit & Tegmark.
"The Quantization Model of Neural Scaling".

Quantization Hypothesis

Michaud, Liu, Girit & Tegmark.
"The Quantization Model of Neural Scaling".

In this paper, we conjecture the Quantization Hypothesis:

QH1 Many natural prediction problems involve a discrete set of computations which are natural to learn and instrumental for reducing loss. We call these "quanta". Model performance is determined by *which* quanta have been learned.

QH2 Some abilities are more useful for reducing loss than others, leading to a natural ordering of the quanta. We call the ordered quanta the **Q Sequence**. Optimally trained networks should therefore learn the quanta in that order. The effect of scaling is to learn *more* of the quanta in the Q Sequence, so scaling performance is simply determined by *how many* quanta are successfully learned.

QH3 The frequencies at which the quanta are used for prediction drop off as a power law.

Theory

We model the Quantization Hypothesis as follows. Let $\mathbf{q}$ denote a bit string whose k^{th} bit $q_k = 1$ if the k^{th} quantum in the Q Sequence has been learned, and $q_k = 0$ otherwise. QH1 implies that the mean loss L is simply a function of $\mathbf{q}$. QH2 implies that when $n \equiv \sum_k q_k$ quanta have been learned, we have $q_k = 1$ for $k \leq n$. Let L_n denote the mean loss in this case.

From QH3, we have that the k^{th} quantum benefits prediction on a randomly chosen sample with probability

$$p_k = \frac{1}{\zeta(\alpha+1)} k^{-(\alpha+1)} \propto k^{-(\alpha+1)} \tag{1}$$

for a Zipf power law $\alpha > 0$, where $\zeta(s) \equiv \sum_{k=1}^{\infty} k^{-s}$. Let us also assume that learning the k^{th} quantum reduces average loss from b_k before it is learned to a_k after it is learned on the samples where it is utilized.

If a_k and b_k are k-independent ($a_k = a$, $b_k = b$), then a model that has learned the first n quanta will have expected loss

$$\begin{aligned} L_n &= \sum_{k=1}^{n} a p_k + \sum_{k=n+1}^{\infty} b p_k = \sum_{k=1}^{\infty} a p_k + \sum_{k=n+1}^{\infty} (b-a) p_k \\ &\approx a + \frac{b-a}{\zeta(\alpha+1)} \int_n^{\infty} k^{-(\alpha+1)} dk = a + \frac{b-a}{\alpha\zeta(\alpha+1)} n^{-\alpha}. \end{aligned} \tag{2}$$

In other words, $L_\infty = a$ and $(L_n - L_\infty) \propto n^{-\alpha}$ is a power law.

Parameter scaling

Parameter scaling: In networks of finite size, only finitely many quanta can be learned – network capacity is a bottleneck. If we assume that all quanta require the same capacity of C network parameters, and we have a network with N total parameters, roughly $n = N/C$ elements in the Q Sequence can be learned. We therefore expect loss to depend on the number of model parameters N like so:

$$L(N) = L_{N/C} \approx \frac{1}{\alpha\zeta(\alpha+1)} \left(\frac{N}{C}\right)^{-\alpha} \propto N^{-\alpha}. \quad (3)$$

Data scaling (multi-epoch)

Data scaling (multi-epoch): For data scaling, we assume that a threshold of τ examples utilizing quantum k are needed in the training set in order for quantum k to be learned. τ can perhaps be thought of as the minimum number of examples on average requiring quantum k needed to uniquely specify its computation. Assuming network capacity is not a bottleneck, how many quanta will be learned? If we have a training set of D samples, then it will contain roughly Dp_1 samples utilizing quantum 1, Dp_2 samples utilizing quantum 2, and so on. If $p_k = \frac{1}{\zeta(\alpha+1)}k^{-(\alpha+1)}$, the last quantum n learned in the Q Sequence will then roughly be n such that $D\frac{1}{\zeta(\alpha+1)}n^{-(\alpha+1)} = \tau$ and so $n = (D/\tau\zeta(\alpha+1))^{1/(1+\alpha)}$. Under this model of how the training set size D influences which quanta are learned, we would therefore expect data scaling:

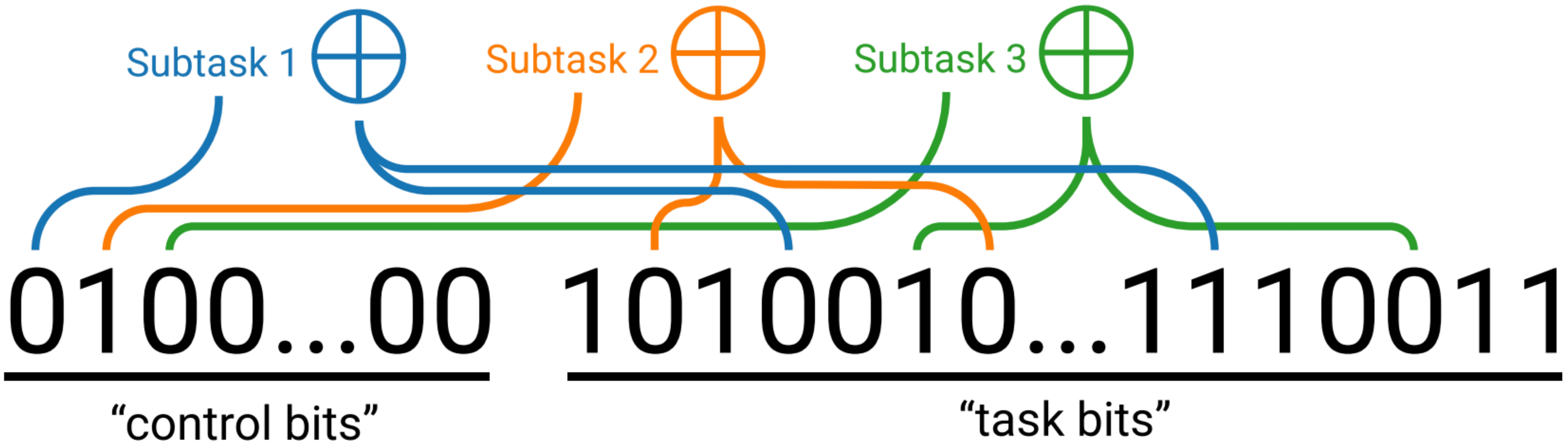
$$L(D) = L_{(D/\tau\zeta(\alpha+1))^{1/(1+\alpha)}} \approx \frac{1}{\alpha\zeta(\alpha+1)}\left(\frac{D}{\tau\zeta(\alpha+1)}\right)^{-\frac{\alpha}{\alpha+1}} \propto D^{-\frac{\alpha}{\alpha+1}}. \tag{4}$$

Data scaling (single-epoch)

Data scaling (single-epoch): In multi-epoch training, the information contained in the training dataset can bottleneck which quanta are learned. However, the rate of convergence of SGD can also bottleneck performance. For single-epoch training, a greater number of training samples allows one to train for longer. Assume that batches are large and that they contain effectively perfect gradient information. If quanta each reduce mean loss by an amount given by a power law, then the gradients incentivizing each quantum to form may also roughly follow a power law in magnitude. We might therefore expect that the number of training steps S to learn quantum k to be inversely proportional to use frequency p_k (more commonly useful quanta have larger gradients and are learned faster). Therefore if the first quantum requires T steps to be learned, then quantum n will require $Tn^{\alpha+1}$ steps to converge. As a function of the number of training steps S, the number of quanta learned is therefore $n = (S/T)^{1/(\alpha+1)}$, and so:

$$L(S) = L_{(S/T)^{1/(\alpha+1)}} \approx \frac{1}{\alpha\zeta(\alpha+1)}\left(\frac{S}{T}\right)^{-\frac{\alpha}{\alpha+1}} \propto S^{-\frac{\alpha}{\alpha+1}}. \tag{5}$$

Toy example: Multitask sparse parity

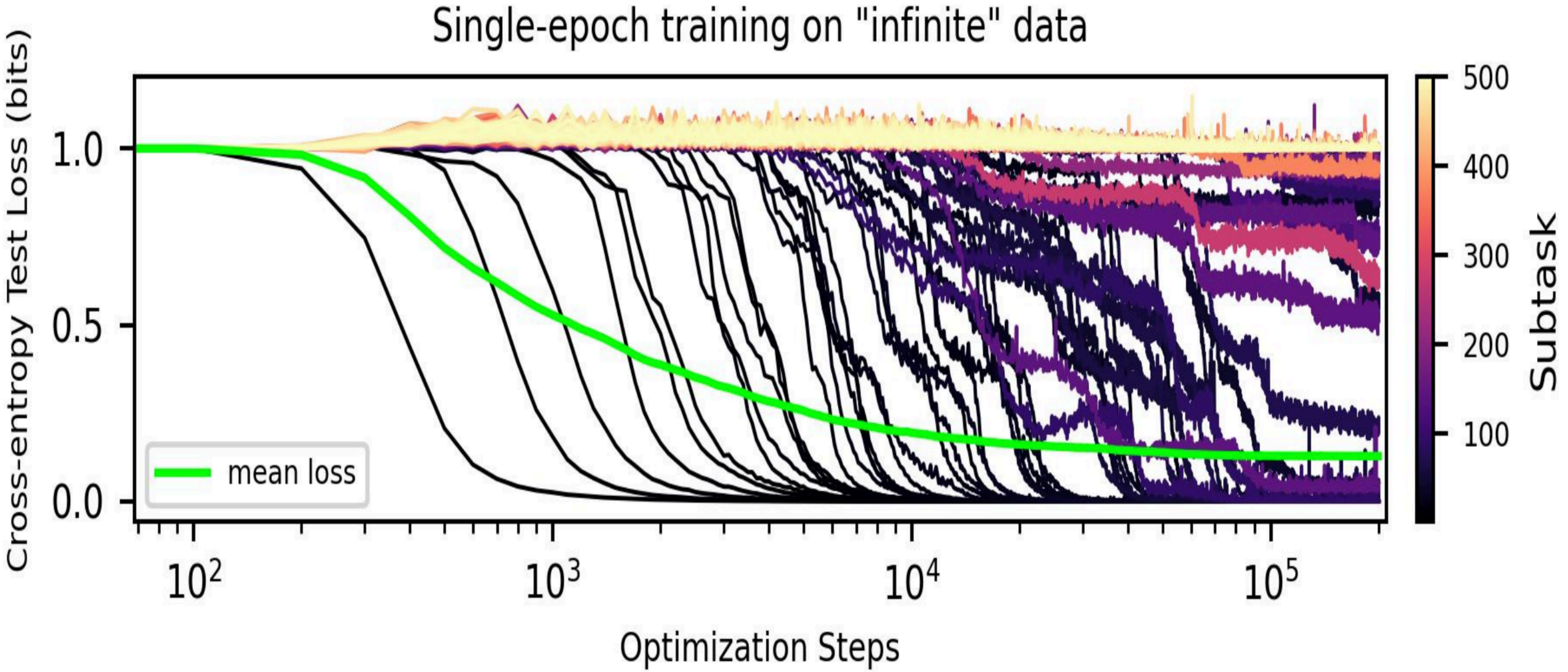


Toy example: dynamics

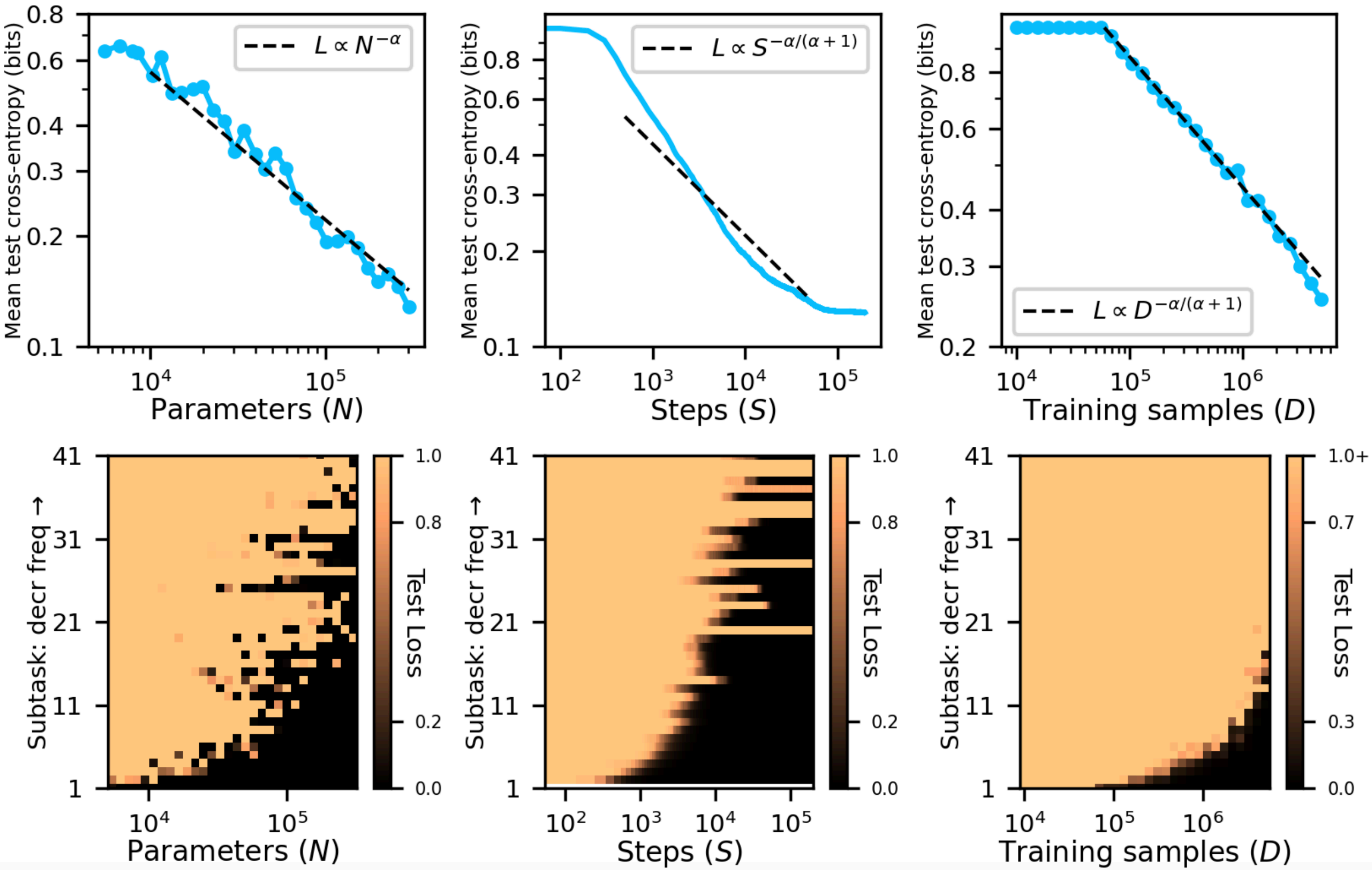
Individual task loss = grokking
Total loss = scaling law

A Quantization Model of Neural Scaling

arXiv: 2303.13506



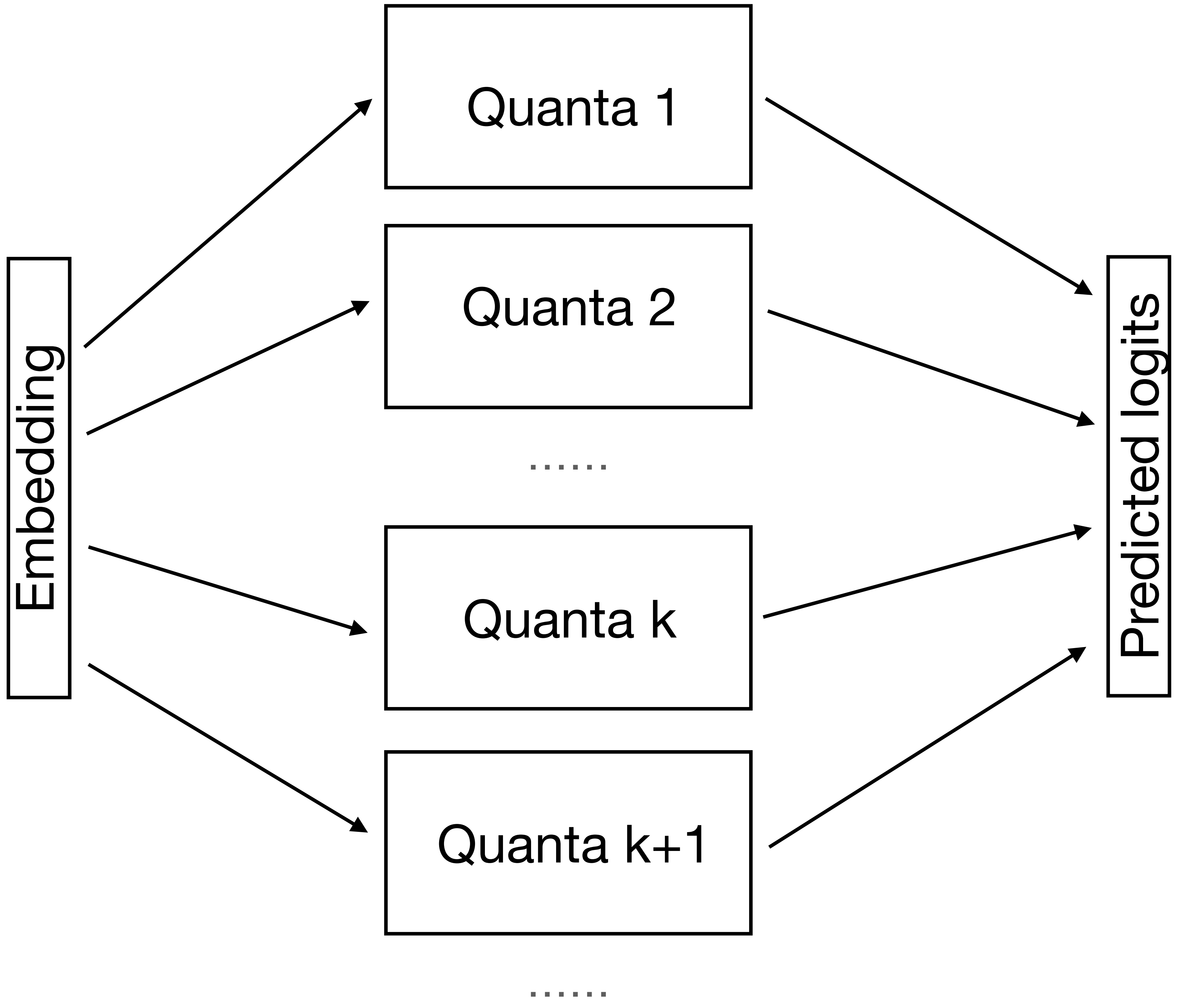
Toy example: scaling



Language Model

We now study how scaling curves for large language models decompose. For our experiments, we use the "Pythia" model sequence from Eleuther (EleutherAI 2023). These are decoder-only transformers of varying size trained on the same data in the same order – approximately 300 billion tokens of the train set of The Pile (Gao et al. 2020). Eleuther released 143 checkpoints for these models, spaced 1000 optimization steps apart. We can therefore study scaling w.r.t. model parameters N and training steps S. We evaluate the first seven models in the sequence, which range from 19m to 6.4b non-embedding parameters, on approximately 10 million tokens from the test set of The Pile. We record cross-entropy loss on every token. With this collection of loss values, we are able to study how neural scaling decomposes – rather than looking just at how mean test loss changes with scale, we can see how the distribution over losses changes with scale.

Quanta Discovery with Gradients (QDG)

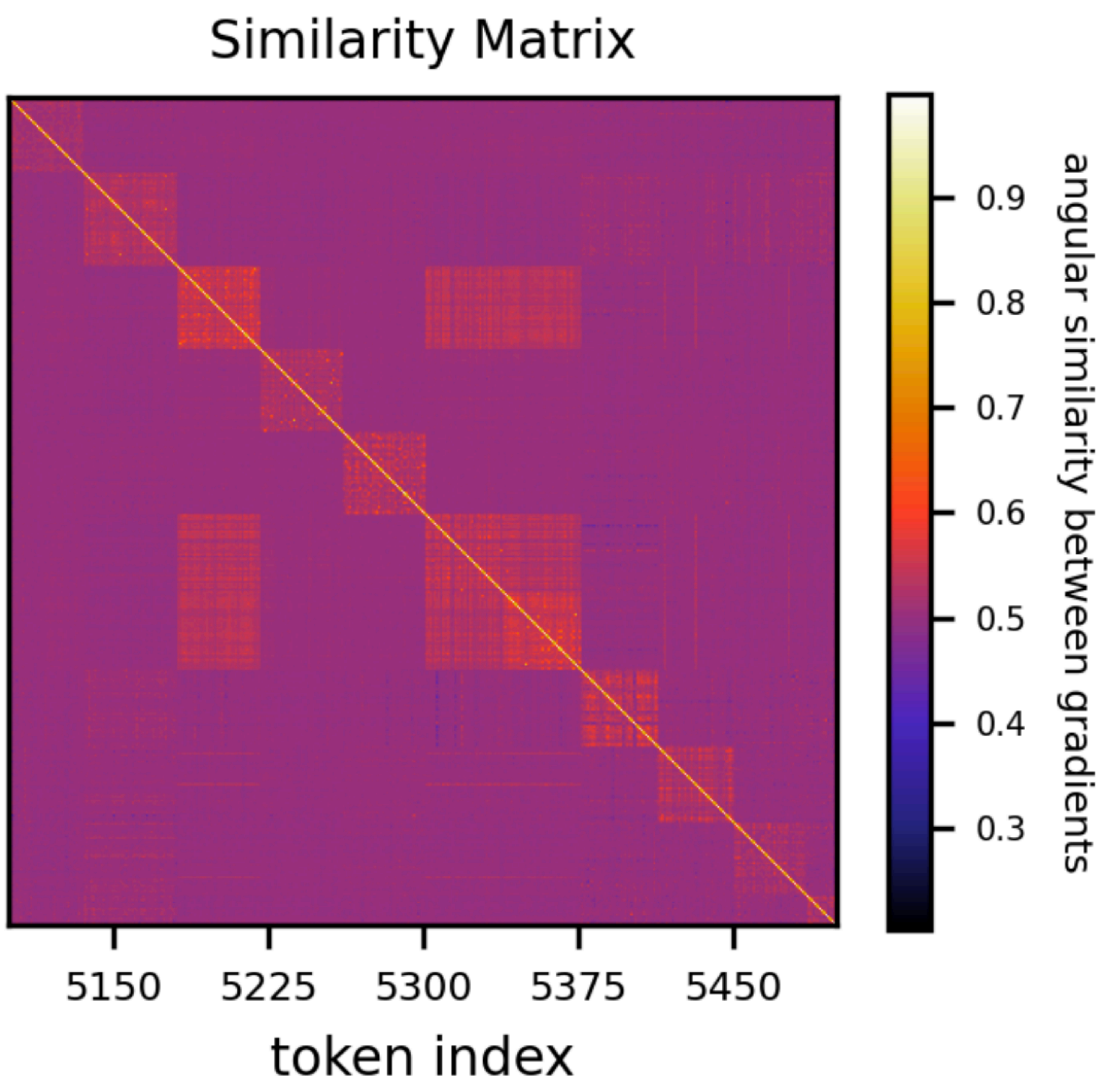


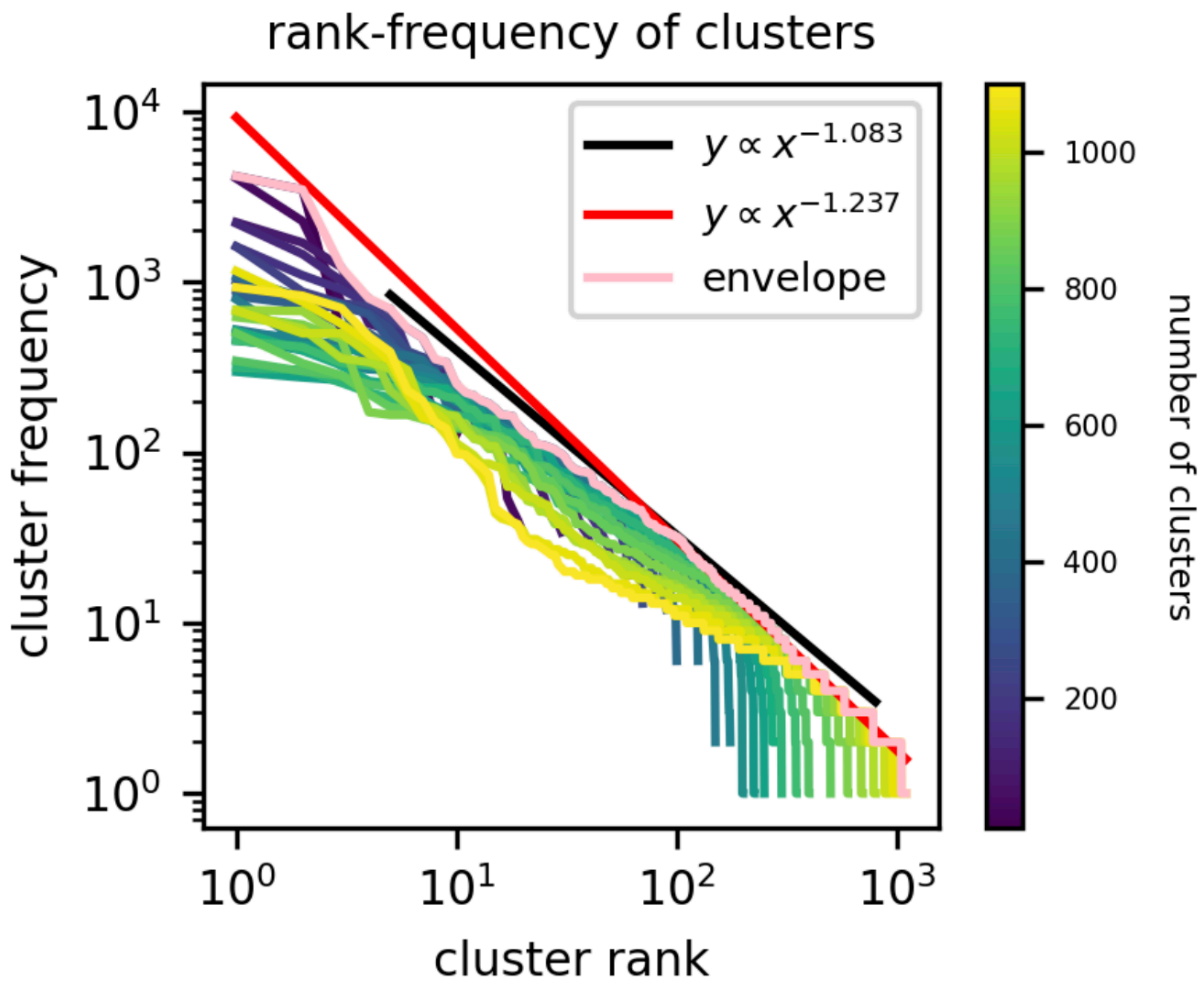
If two tokens belong to the same quanta,
their activations/gradients should align.

QDG main idea:

(1) Compute model gradients for tokens.
(2) Clustering gradients. Each cluster is a quanta.

QDG results


Similarity Matrix
0.9
0.8
0.7
0.6
0.5
0.4
0.3
angular similarity between gradients
5150
5225
5300
5375
5450
token index


rank-frequency of clusters
$y \propto x^{-1.083}$
$y \propto x^{-1.237}$
envelope
10^4
10^3
10^2
10^1
10^0
cluster frequency
10^0
10^1
10^2
10^3
cluster rank
1000
800
600
400
200
number of clusters

Knowledge quanta

“Quanta” of LLM capabilities auto-discovered in natural text

quantum for numerical sequence continuation (examples from cluster 50)

```
...ents his famous tonadas, a genre of the Venezuelan plains folk music.

Track listing
01- Mi Querencia (Simón Díaz)
02- Tonada De Luna Llena (Simón Díaz)
03- Sabana (José Salazar/Simón Díaz)
04- Caballo Viejo (Simón Díaz)
05- Todo Este Campo Es Mío (Simón Díaz)
06- La Pena Del Becerrero (Simón Díaz)
07
```

```
...sis supplied.) Appealing from that order, the city asserts (1)
plaintiffs have no standing or right to maintain the action; (2) that the
proposed road was in an undedicated part of the park; (3) that the
proposed road was an access road and not a through street or part of the
city's street system; (4
```

```
...
  4. _Introduction_
  5. Chapter 1: What Is Trust?
  6. Chapter 2: Trust Brings Rest
  7. Chapter 3: Who Can I Trust?
  8. Chapter 4: The Folly of Self-Reliance
  9. Chapter 5: Trust God and Do Good (Part 1)
  10. Chapter 6: Trust God and Do Good (Part 2)
  11. Chapter 7: At All Times
  12. Chapter 8
```

```
...gn of noncavitated lesion seen only when the tooth is dried; 2 =
visible noncavitated lesion seen when wet and dry; 3 = microcavitation in
enamel; 4 = noncavitated lesion extending into dentine seen as an
undermining shadow; 5 = small cavitated lesion with visible dentine: less
than 50% of surface; 6
```

```
      ...DynamicKey><Action>F1</Action><Label>F1</Label></DynamicKey>
         <DynamicKey><Action>F2</Action><Label>F2</Label></DynamicKey>
         <DynamicKey><Action>F3</Action><Label>F3</Label></DynamicKey>
         <DynamicKey><Action>F4</Action><Label>F4</Label></DynamicKey>
         <DynamicKey><Action>F5
```

```
...
    GetPrepareVoteMsg       = 0x07
    PrepareVotesMsg         = 0x08
    GetQCBlockListMsg       = 0x09
    QCBlockListMsg          = 0x0a
    GetLatestStatusMsg      = 0x0b
    LatestStatusMsg         = 0x0c
    PrepareBlockHashMsg     = 0x0d
    GetViewChangeMsg        = 0x0e
    PingMsg                 = 0x0f
```

quantum for predicting newlines to maintain text width (examples from cluster 100)

```
...C REGRESSION.
THE GOALS OF THIS VIDEO ARE
TO PERFORM QUADRATIC REGRESSION
ON THE TI84 GRAPHING CALCULATOR,
DETERMINE HOW WELL THE
REGRESSION MODEL FITS THE DATA,
AND THEN MAKE PREDICTIONS
USING THE REGRESSION EQUATION.
IN STATISTICS,
REGRESSION ANALYSIS INCLUDES
ANY TECHNIQUES USED FOR MODELING \n
```

```
...ump is free software: you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation, either version 3 of the License, or
# (at your option) any later version.
#
# creddump is distributed in the hope that it will be useful,\n
```

```
...  *
        Pursuant to 5TH CIR. R. 47.5, the court has determined
that this opinion should not be published and is not precedent
except under the limited circumstances set forth in 5TH CIR.\n
```

```
...
files (the
// "Software"), to deal in the Software without restriction, including
// without limitation the rights to use, copy, modify, merge, publish,
// distribute, sublicense, and/or sell copies of the Software, and to
permit
// persons to whom the Software is furnished to do so, subject to the\n
```

```
<!--
/**
 * Copyright (c) 2019, The Android Open Source Project
 *
 * Licensed under the Apache License, Version 2.0 (the "License");
 * you may not use this file except in compliance with the License.\n
```

```
...f maturity and an underdeveloped
sense of responsibility, leading to recklessness, impul-
sivity, and heedless risk-taking.... Second, children
are more vulnerable... to negative influences and
outside pressures, including from their family and
peers; they have limited contro[l] over their own envi-\n
```

Summary

* Representation
* Modularity
* Quantization

How does intelligence emerge under hunger?

A scientific theory of deep learning?

Contact

Email: zmliu@mit.edu, website: kindxiaoming.github.io

Thank you!

Max Tegmark

Eric J. Michaud

Eric Gan

Mike Williams

Niklas Nolte

Ouail Kitouni

Backup

Interpretability vs accuracy tradeoff

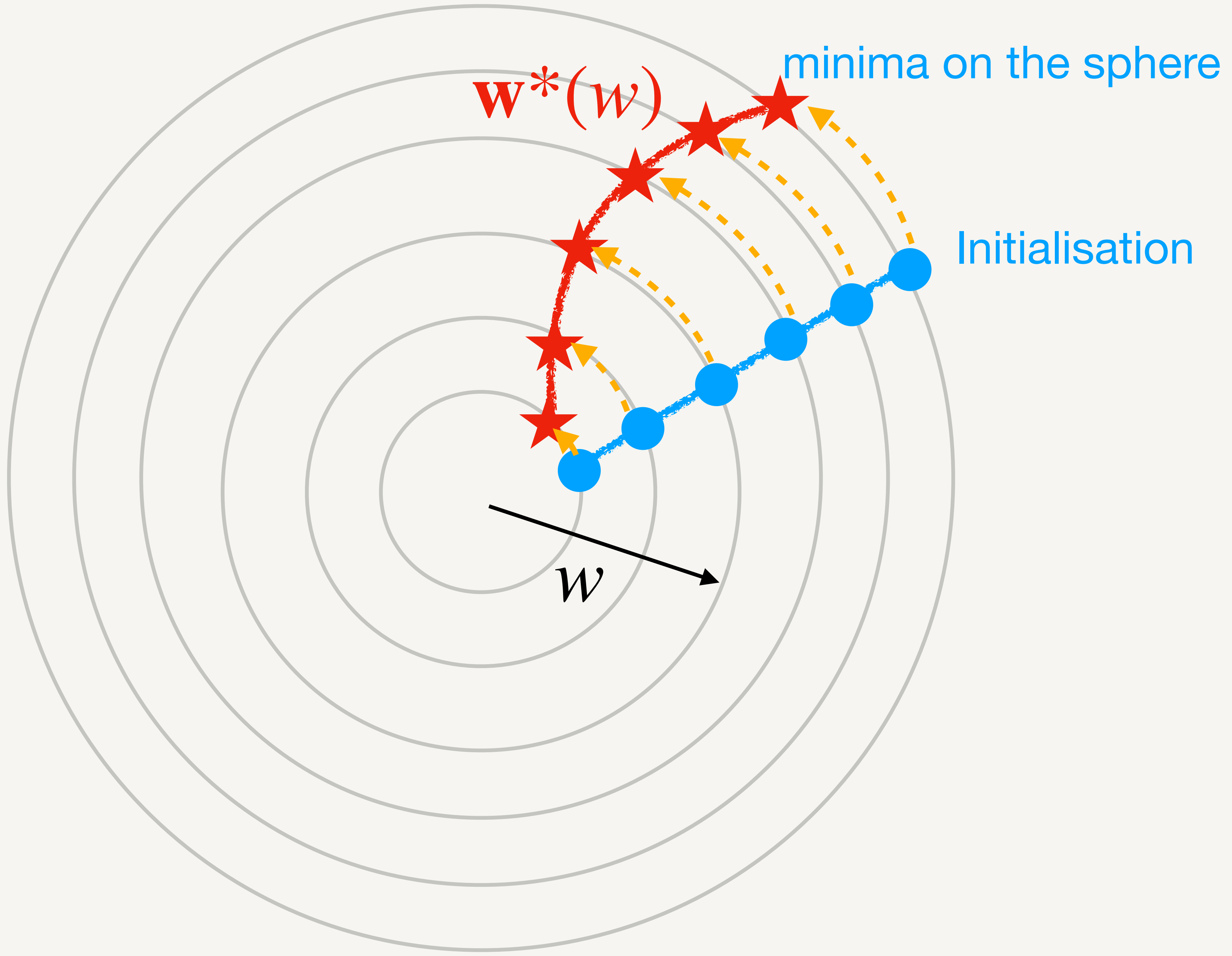
Table 1: BIMT achieves interpretability with no or modest performance drop

dataset	symbolic (a)	symbolic (b)	symbolic (c)	two moon	modular addition	permutation	in-context	MNIST
metric	loss	loss	loss	accuracy	accuracy	accuracy	loss	accuracy
without BIMT	5.8e-3	1.1e-5	1.2e-4	100.0%	100.0%	100.0%	7.2e-5	98.5%
with BIMT	7.4e-3	8.5e-5	1.3e-3	100.0%	100.0%	100.0%	1.8e-4	98.0%

Reduced 1D landscape

$$\tilde{f}(w) \equiv f(\mathbf{w}^*(w)), \quad \text{where } \mathbf{w}^*(w) \equiv \underset{||\mathbf{w}||_2 = w}{\text{argmin}}\; l_{\text{train}}(\mathbf{w})$$

Any quantity of interest, e.g., train/test loss/error.

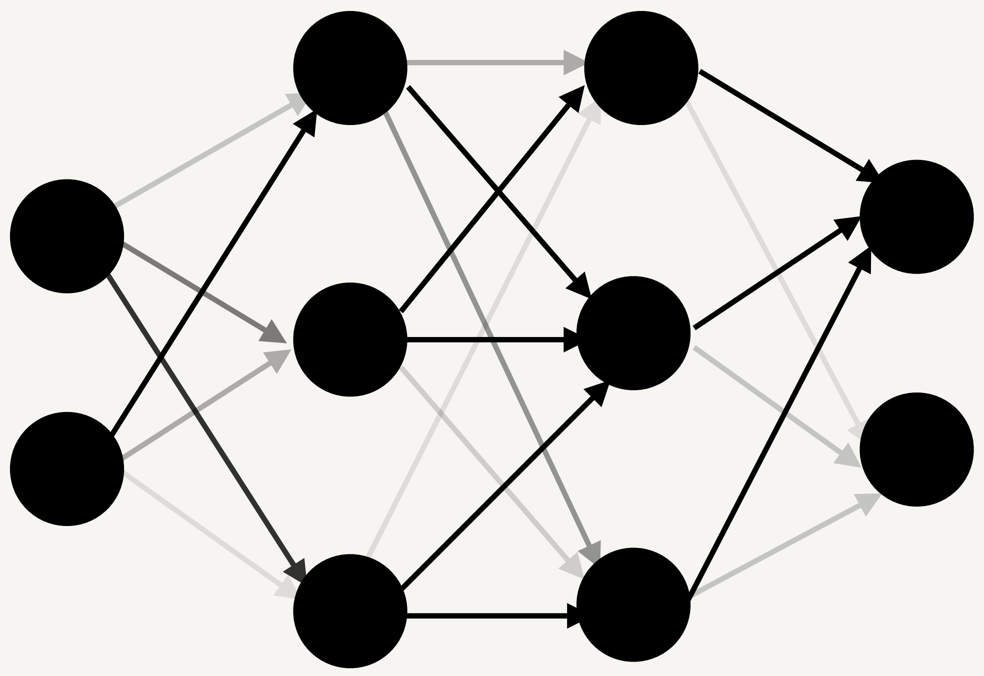


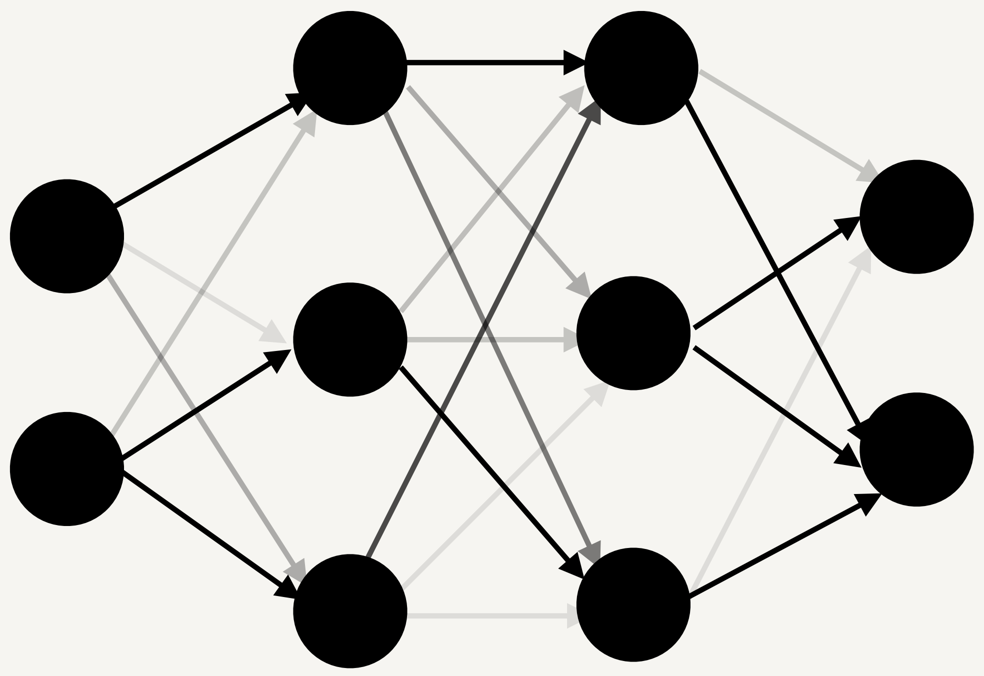
Toy: Teacher-student

Same architecture

Teacher network

Student network





Random seed: 0

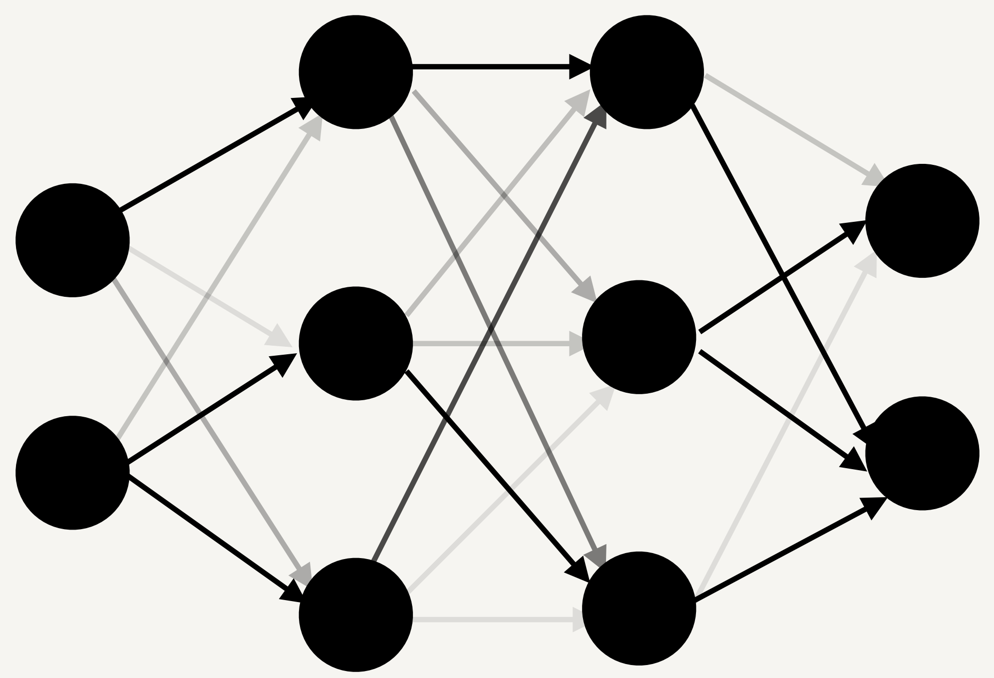
Standard initialisation

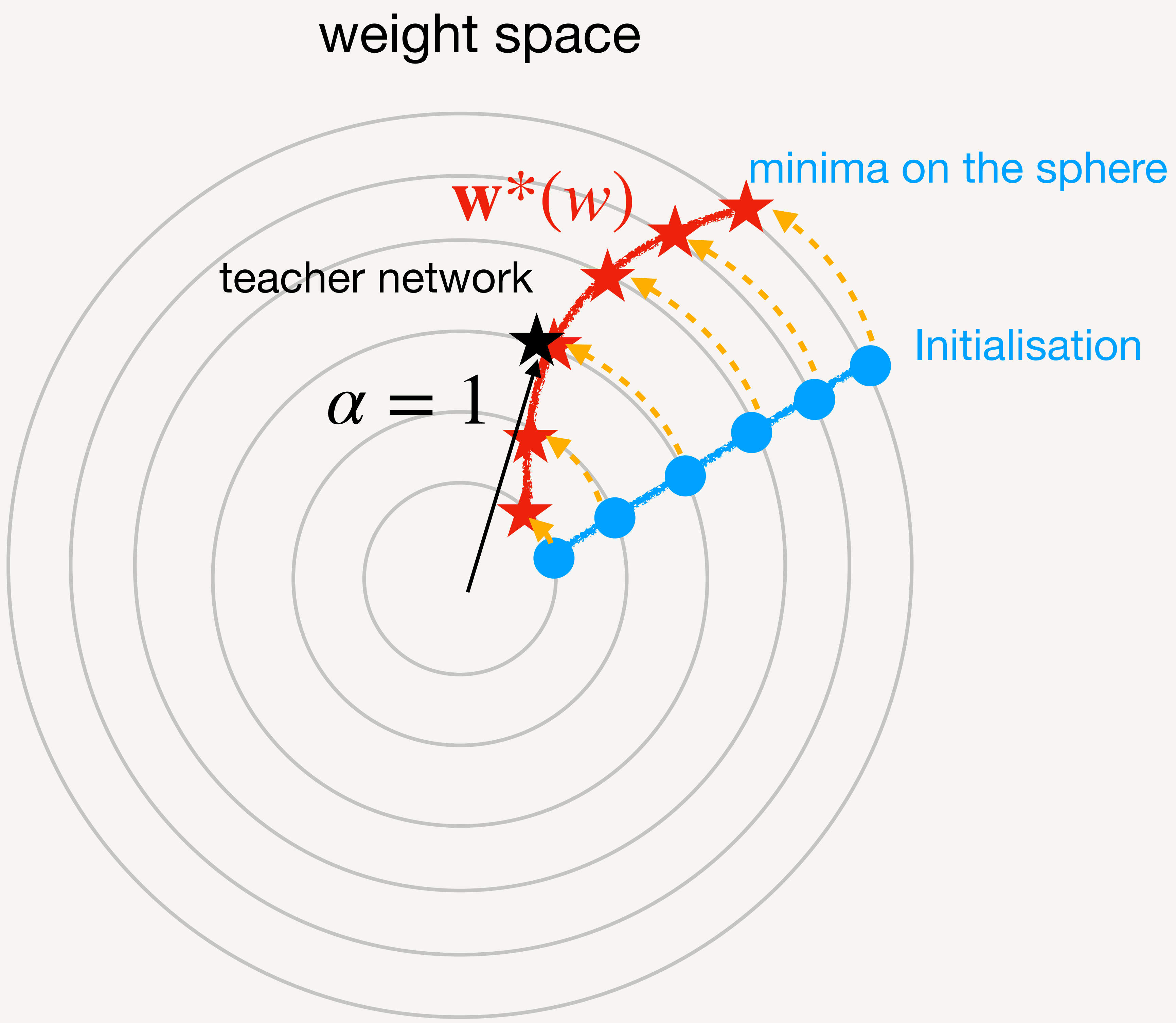
Random seed: 1

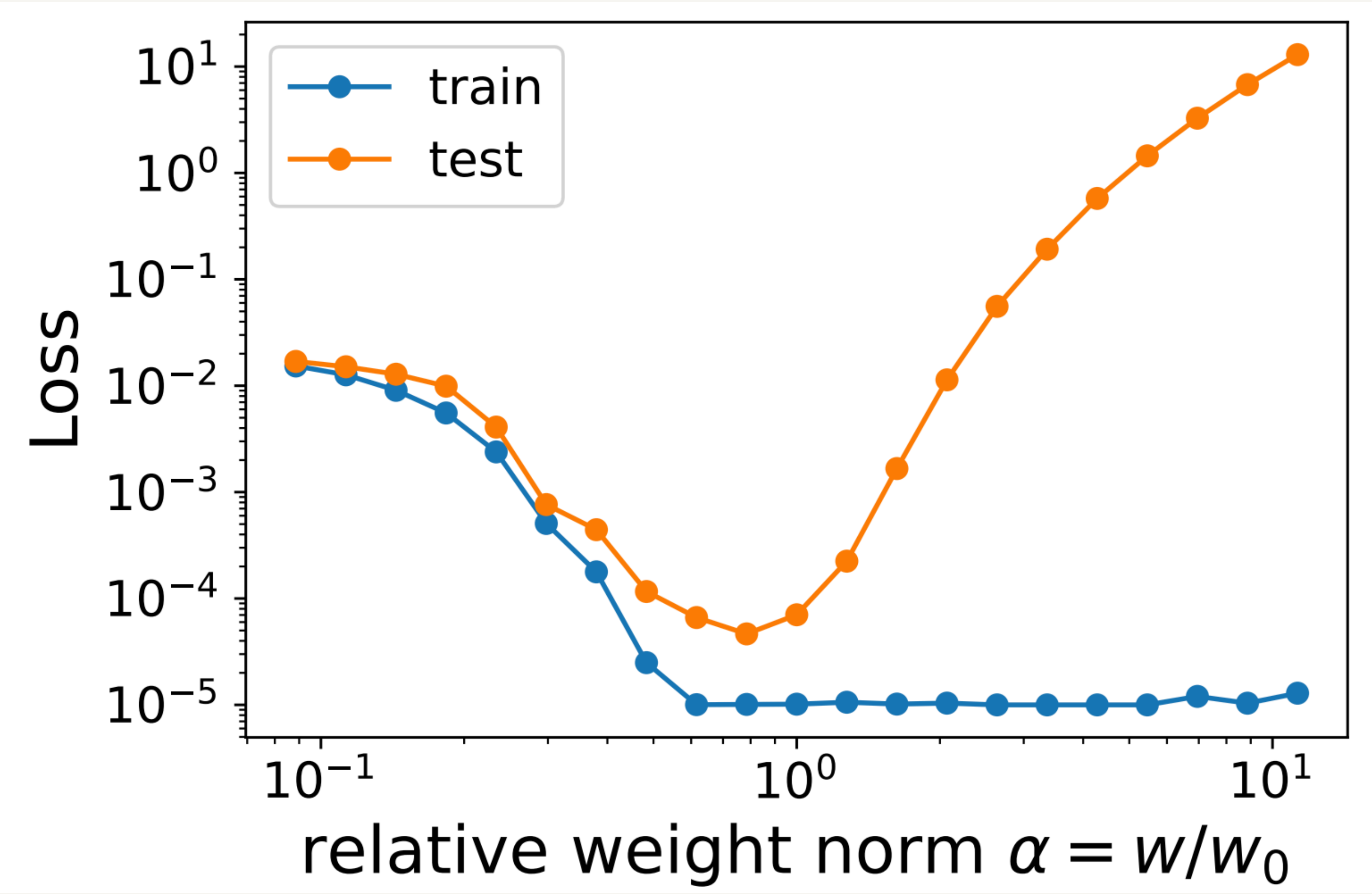
After standard initialization, multiply all weights by α

Teacher-student: Landscape

Student network

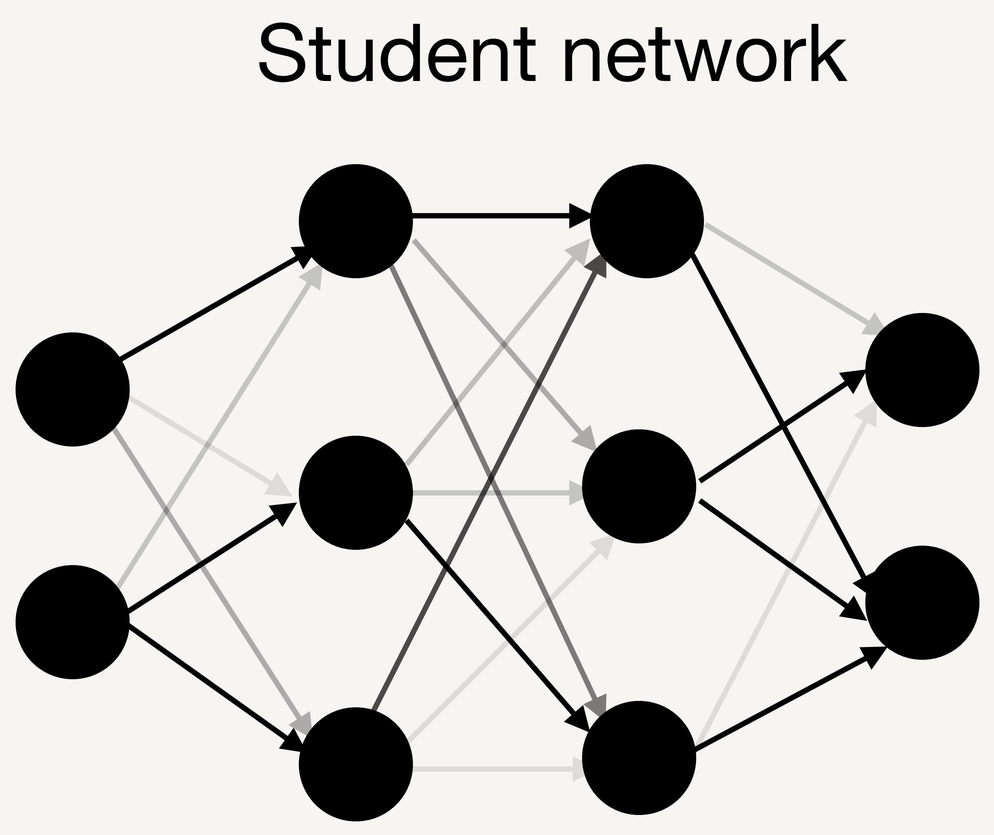


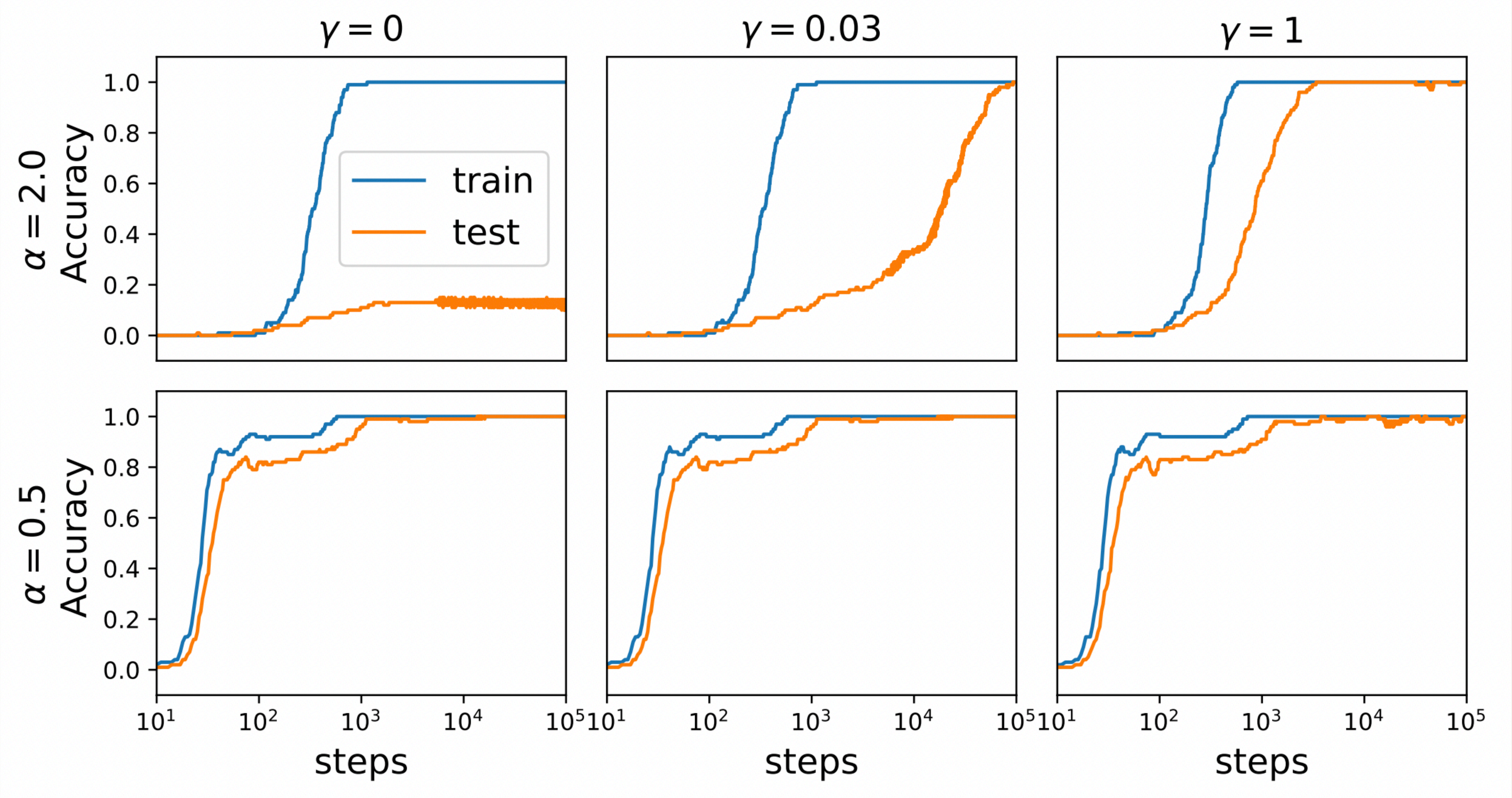




Teacher-student: Grokking

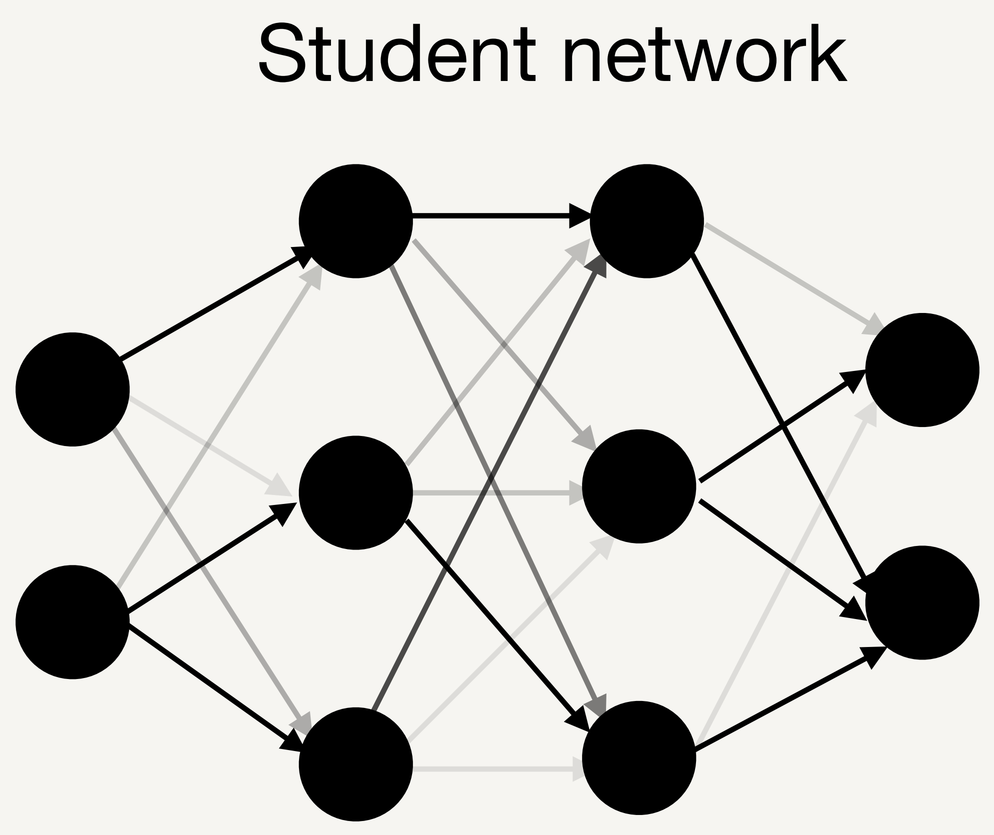
Note: weight norm is not constrained here.

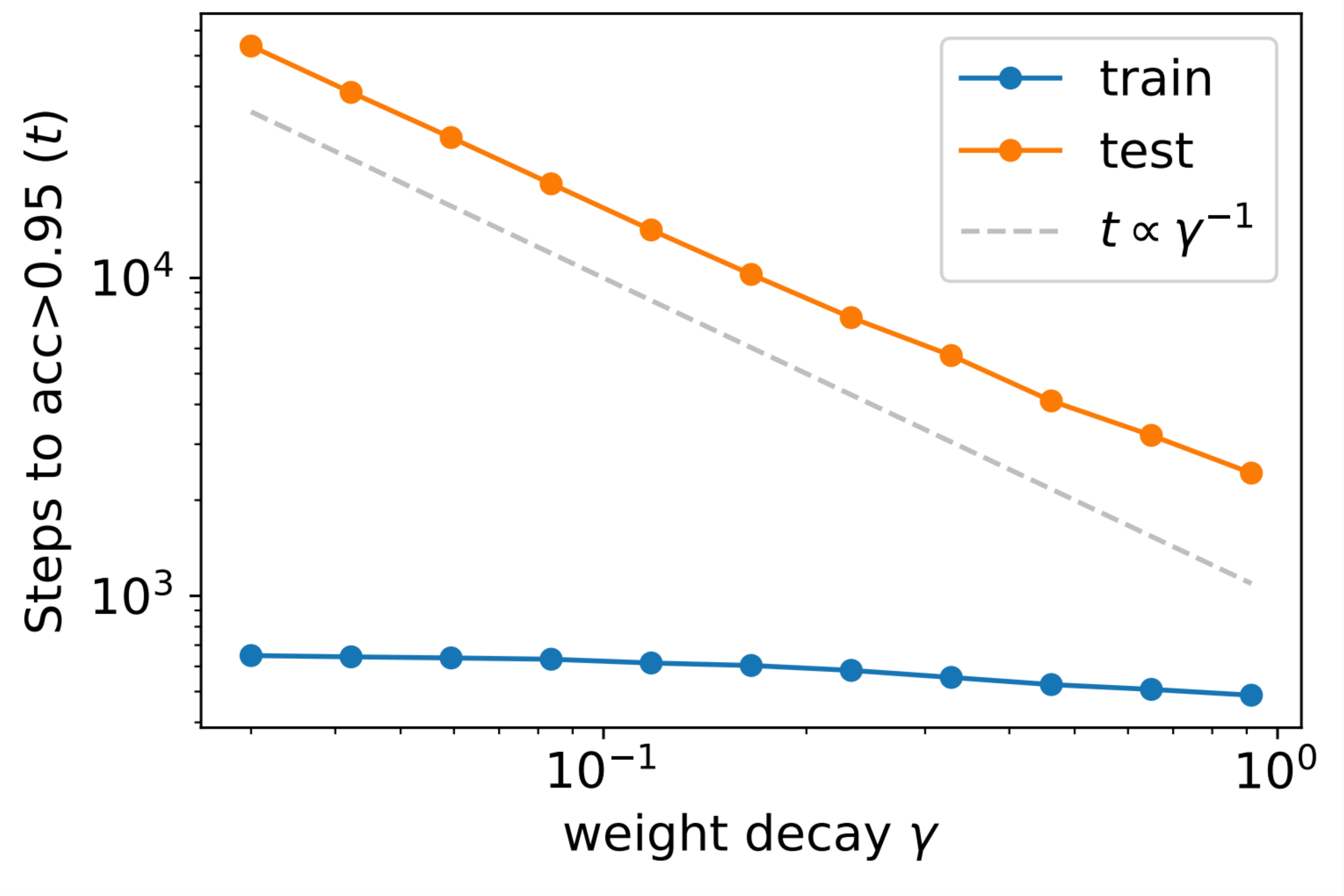




Teacher-student: Grokking

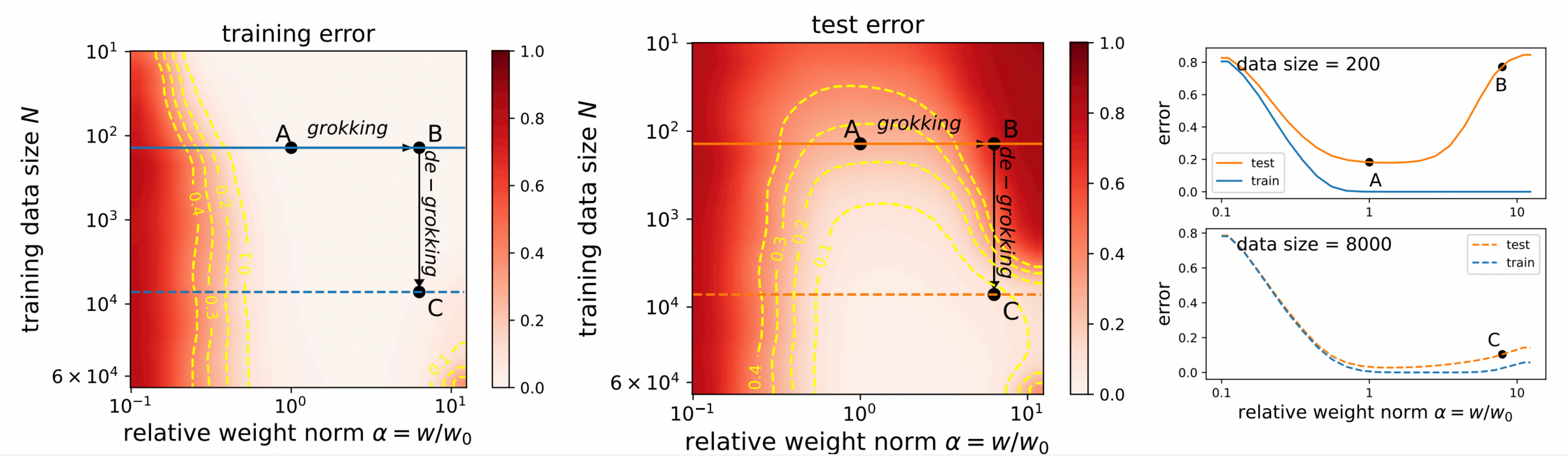
Note: weight norm is not constrained here.



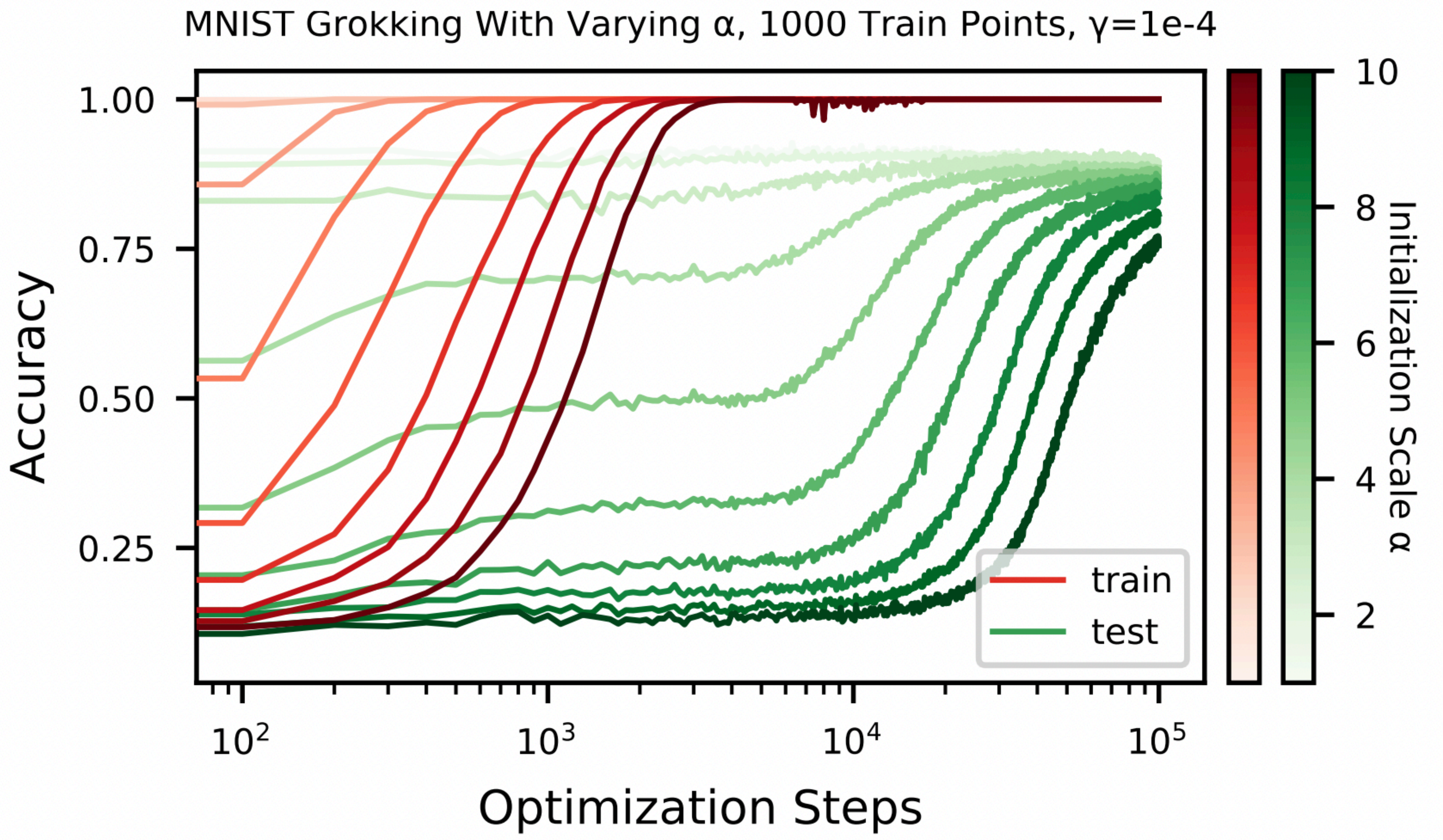


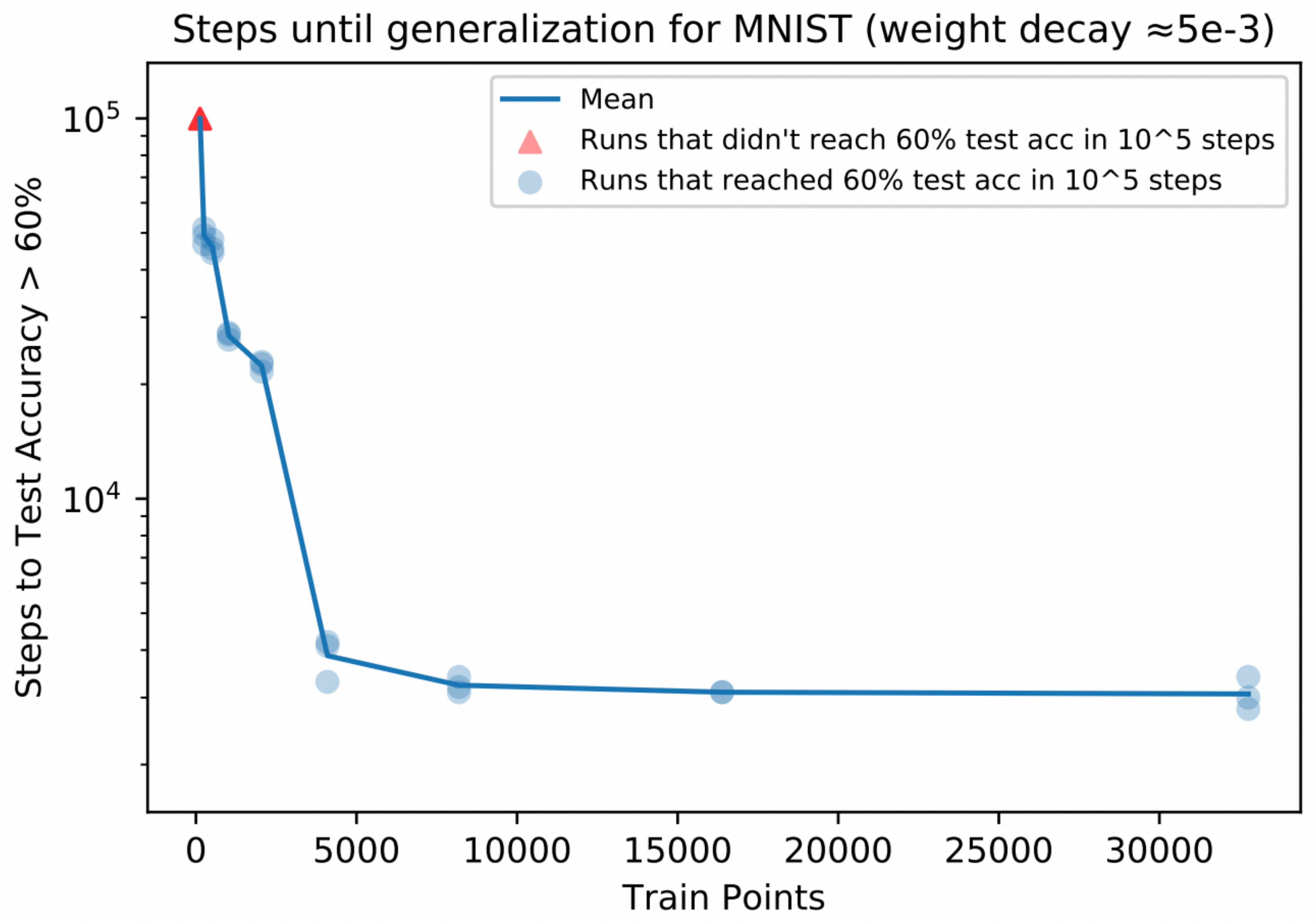
MNIST: landscape analysis

Model: MLP



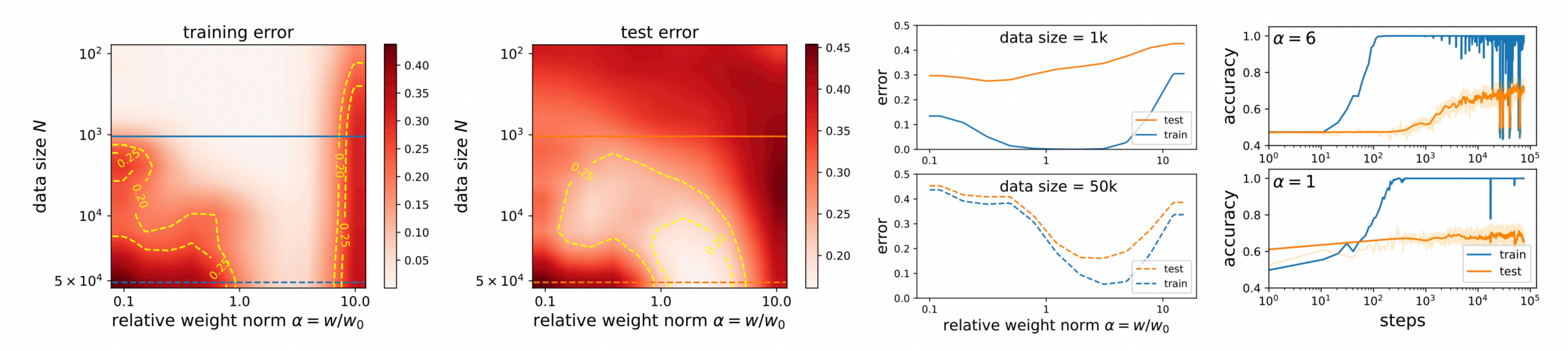
MNIST: Grokking





More datasets

IMDb (Sentiment Analysis) + LSTM



QM9 (Molecule) + Graph Convolutional Neural Network