Natural Science of *Artificial* Intelligence for Trustworthy and Energy-Efficient AI

Hidenori Tanaka
Physics & Informatics Lab, NTT Research, Inc.
Center for Brain Science, Harvard University

“Physics of Intelligence” is a new frontier in physics!

Industrial revolutions give birth to new physics

History: Steam Engines & Thermodynamics


1712: The first commercially successful steam engine

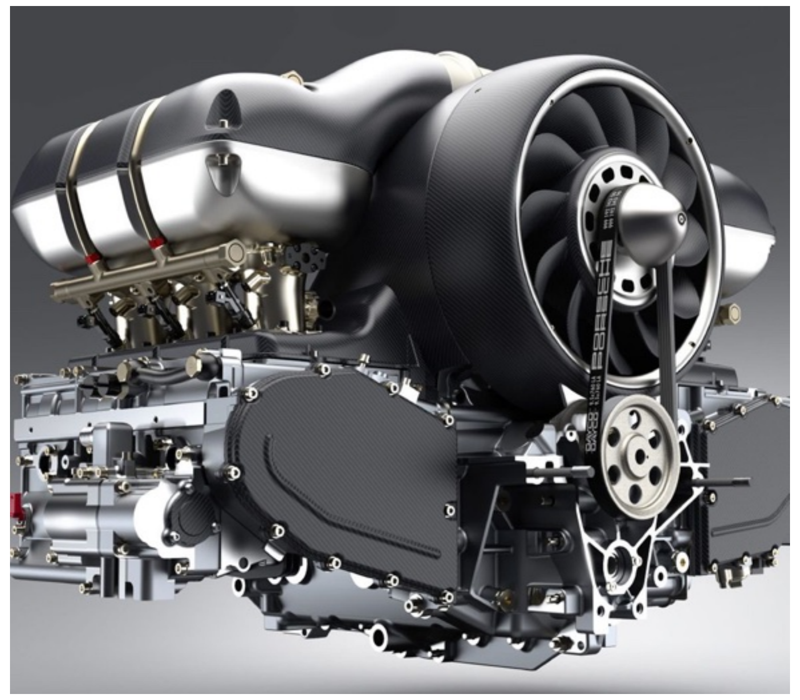
ChatGPT moment??


1776: Industrial revolution triggered by Watt’s steam engine


1824: The birth of thermodynamics By Sadi Carnot, Military Engineer

Scientifically deep and practically impactful questions open up new physics

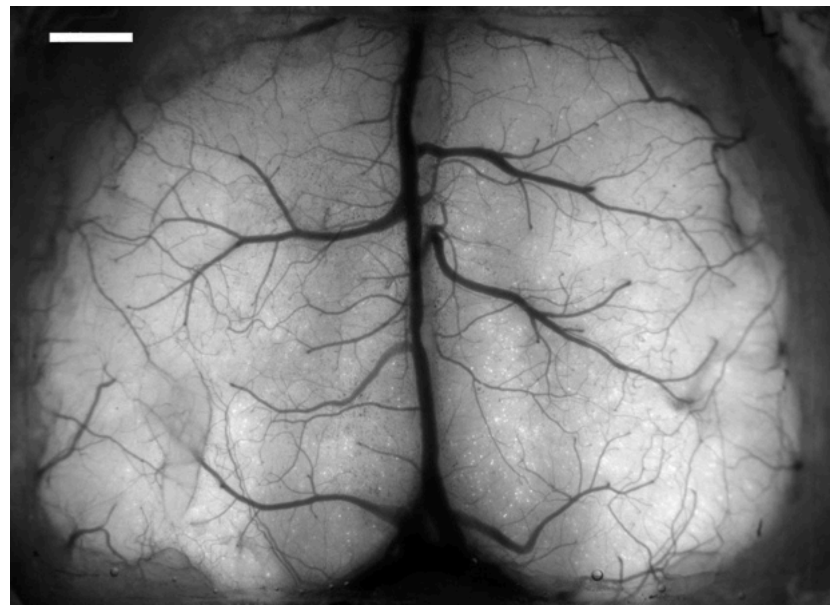
Engines & Thermodynamics



Electrical Engineering & Solid-state physics

Chemical Engineering & Soft matter physics

Physics of Biological/Artificial Neural Networks



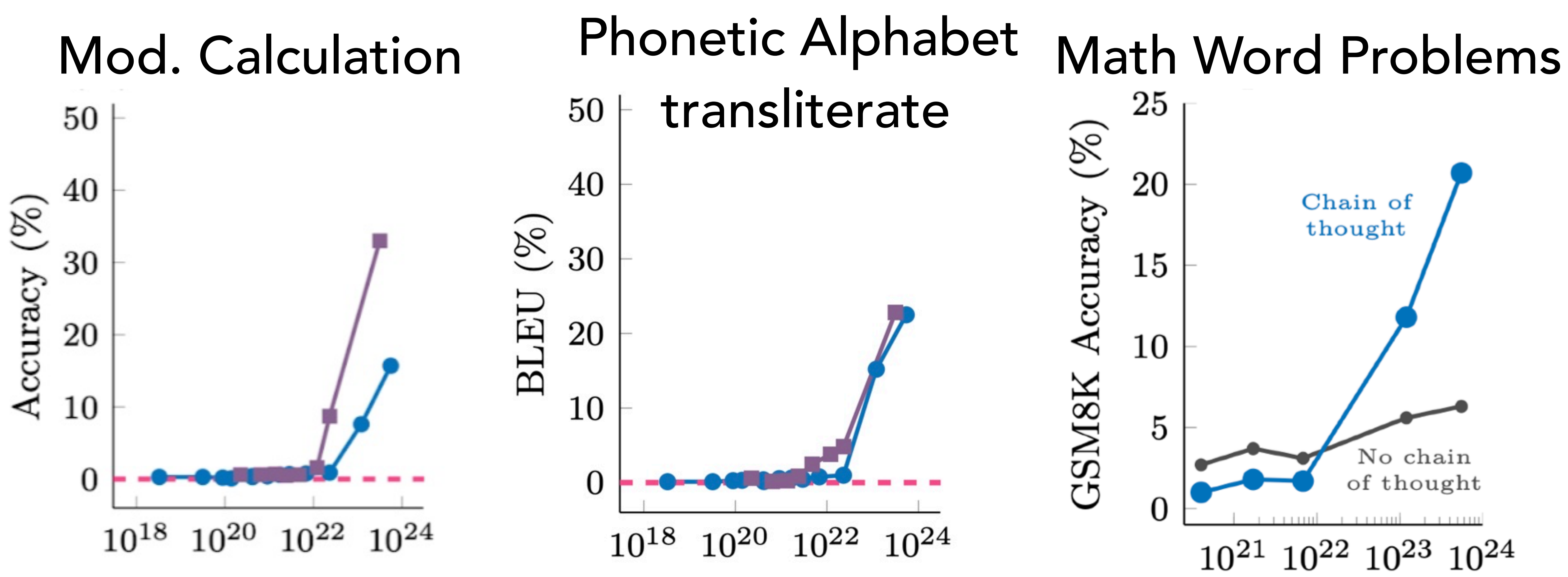
recording ~10,000 neurons

Kim, …, Schnitzer Cell Reports2016

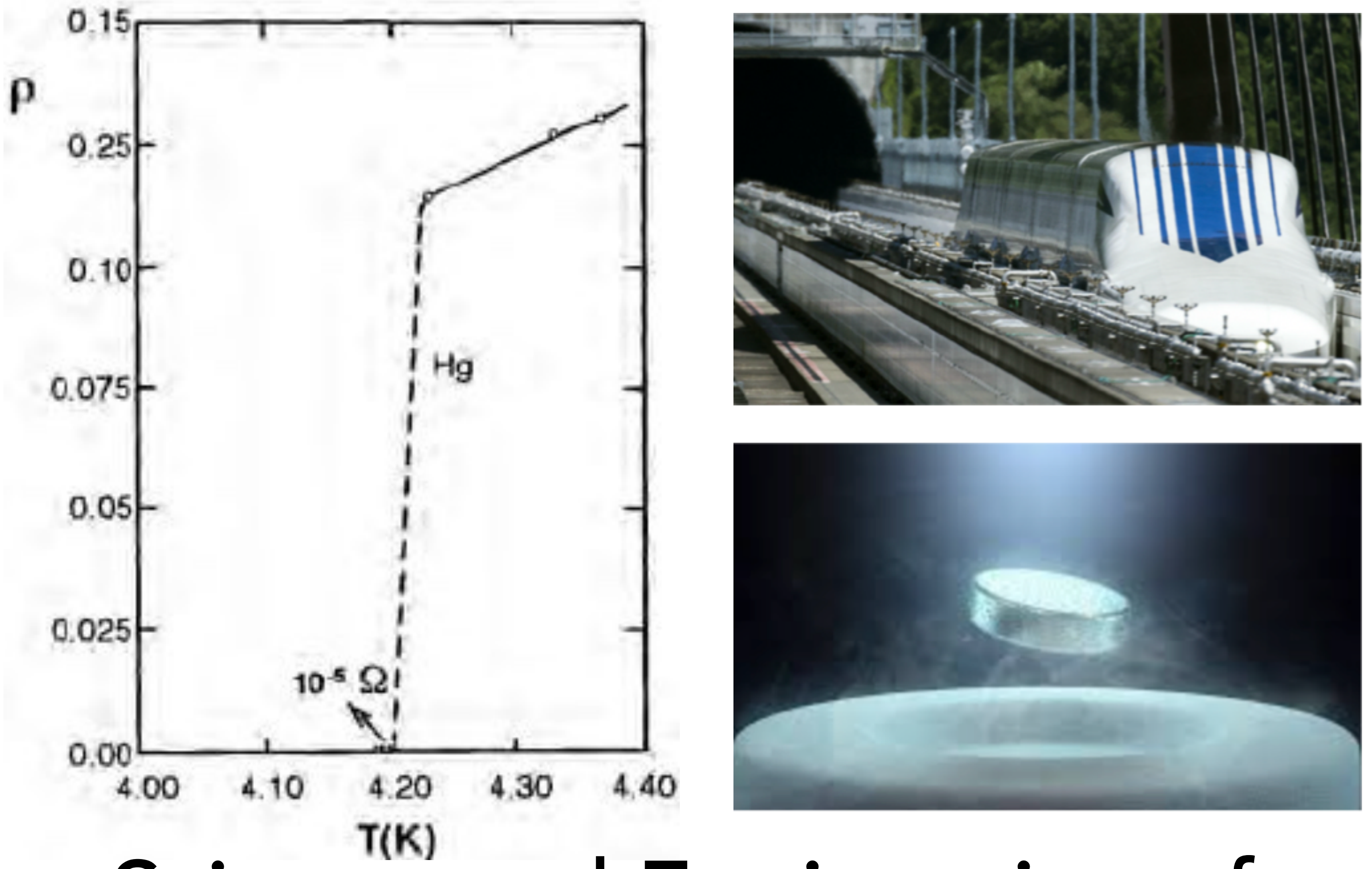
Can there be“Natural” Science of “Artificial” Intelligence?

Conventional Paradigm:
- A computer precisely executes human-defined algorithms.
- Theoretical Computer Science: Constructing a rigorous mathematical theory of convergence and error, etc.

Paradigm of Deep Learning: Engineering with Emergent Abilities
- Artificial organism with ~100 billion parameters trained on ~trillions of words
- Emergence of capabilities with the scaling of data, model, and compute.
- Empirical characterization and theoretical modeling of emergent phenomena, akin to physics.



Model scale (Training FLOPs)



Science and Engineering of Superconductivity

Modern AI systems are high-dimensional, nonlinear, and stochastic dynamical systems with rich emergent phenomena.

Computing and Learning as Physical Processes

1. Can generative AI (diffusion models) imagine? If so, how?

"Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task”
NeurIPS 2023
M. Okawa*, E.S. Lubana*, R.P. Dick, ***H. Tanaka****

2. Learning as physical dynamics:

"Noether’s Learning Dynamics: Role of Symmetry Breaking in Neural Networks” *NeurIPS 2021*
H. Tanaka, D. Kunin

“Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics” *ICLR 2021*
D. Kunin*, J. Sagastuy, S. Ganguli, D.L.K Yamins, H. Tanaka*

Computing and Learning as Physical Processes

1. Can generative AI (diffusion models) imagine? If so, how?

   "Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task"
   NeurIPS 2023
   M. Okawa*, E.S. Lubana*, R.P. Dick, ***H. Tanaka****



2. Learning as physical dynamics:

   "Noether's Learning Dynamics: Role of Symmetry Breaking in Neural Networks" *NeurIPS 2021*
   H. Tanaka, D. Kunin

   "Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics" *ICLR 2021*
   D. Kunin*, J. Sagastuy, S. Ganguli, D.L.K Yamins, H. Tanaka*

Abstraction and generalization is a cornerstone of natural intelligence!

But it's no longer unique to the brain with the rise of artificial neural networks.

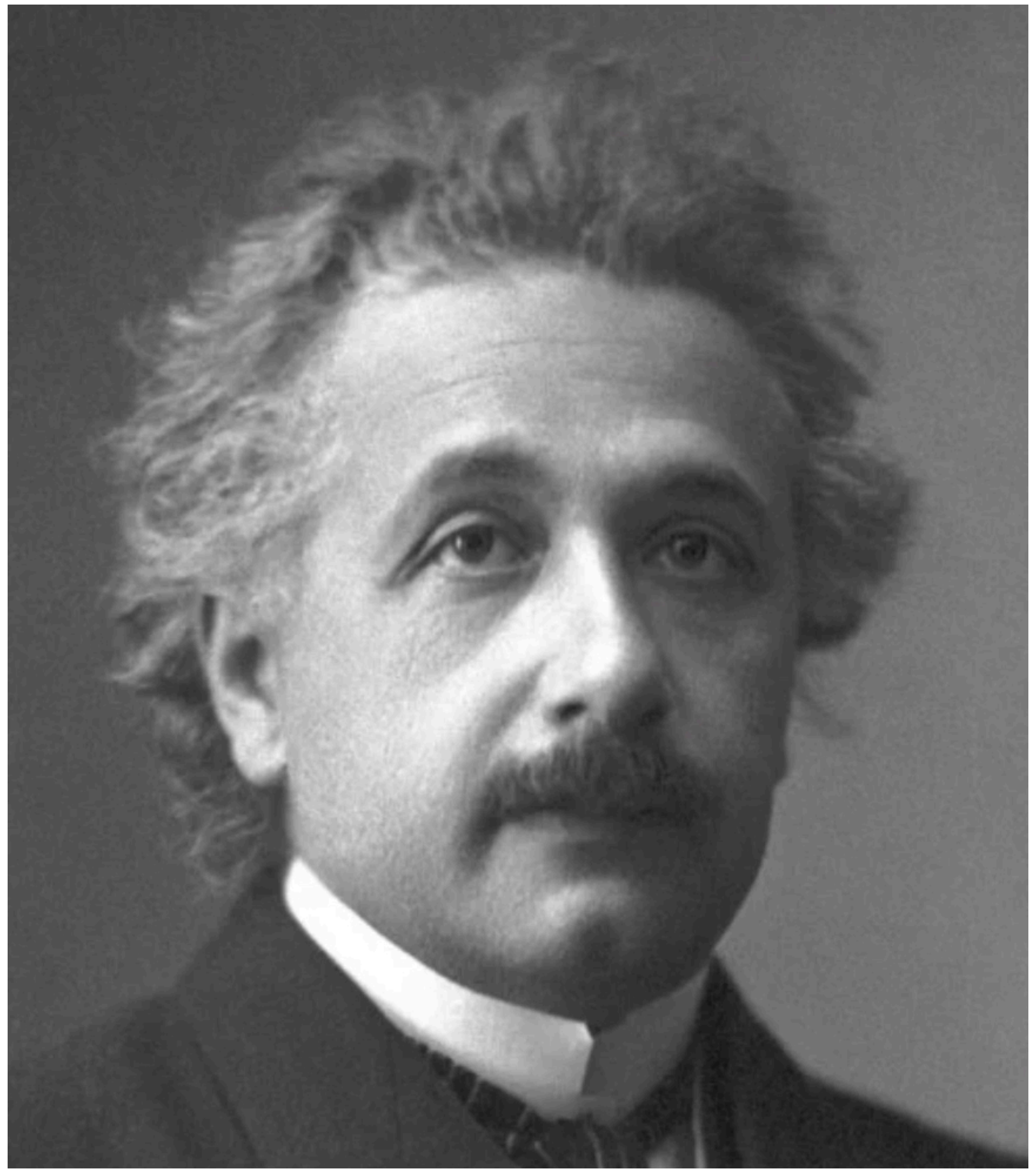
Q. Is there a 'universal' mechanism that governs intelligence?
If so, where does it come from?

Thesis: Universal mechanisms of intelligence emerge from shared evolutionary pressures (task) and experiences (data) within the physical world!

Let's build an interdisciplinary "Science of Natural and Artificial Intelligence", bridging physics, neuroscience, psychology, and computer science!

Concept Learning and Compositional Generalization

Babies play with the world to construct a causal predictive model.
This involves: (i) learning concepts, (ii) understanding their relationships,
and (iii) making predictions and conducting experiments to refine their model.



Artificial networks show ‘**sparks**’ of concept learning and generalization

✓

“A panda skiing with an iguana holding hands in Aspen.”

✗

“a panda through the lens of a magnifying glass.”

✗

“a small light ball and a large heavy ball balanced on a seesaw.”

Artificial networks even has trouble composing "shape" and "color"!

✓ ✓ ✗

"White-colored lizard" "Green-colored lizard"

"Blue-colored lizard" "Magenta-colored lizard"

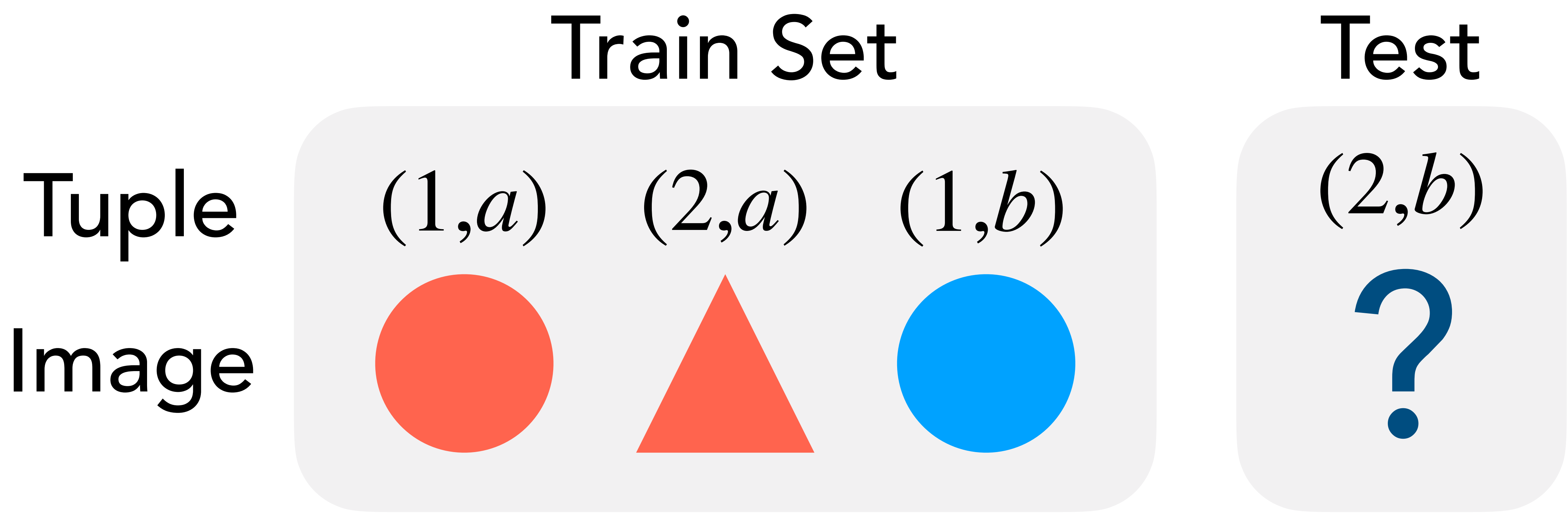
Stable Diffusion Model, as of July 2023

Q: Can artificial networks compose shape, size, and color concepts in novel ways?
If so, how does this capability emerge?

Our Approach: A Simple Task Requiring Compositional Generalization

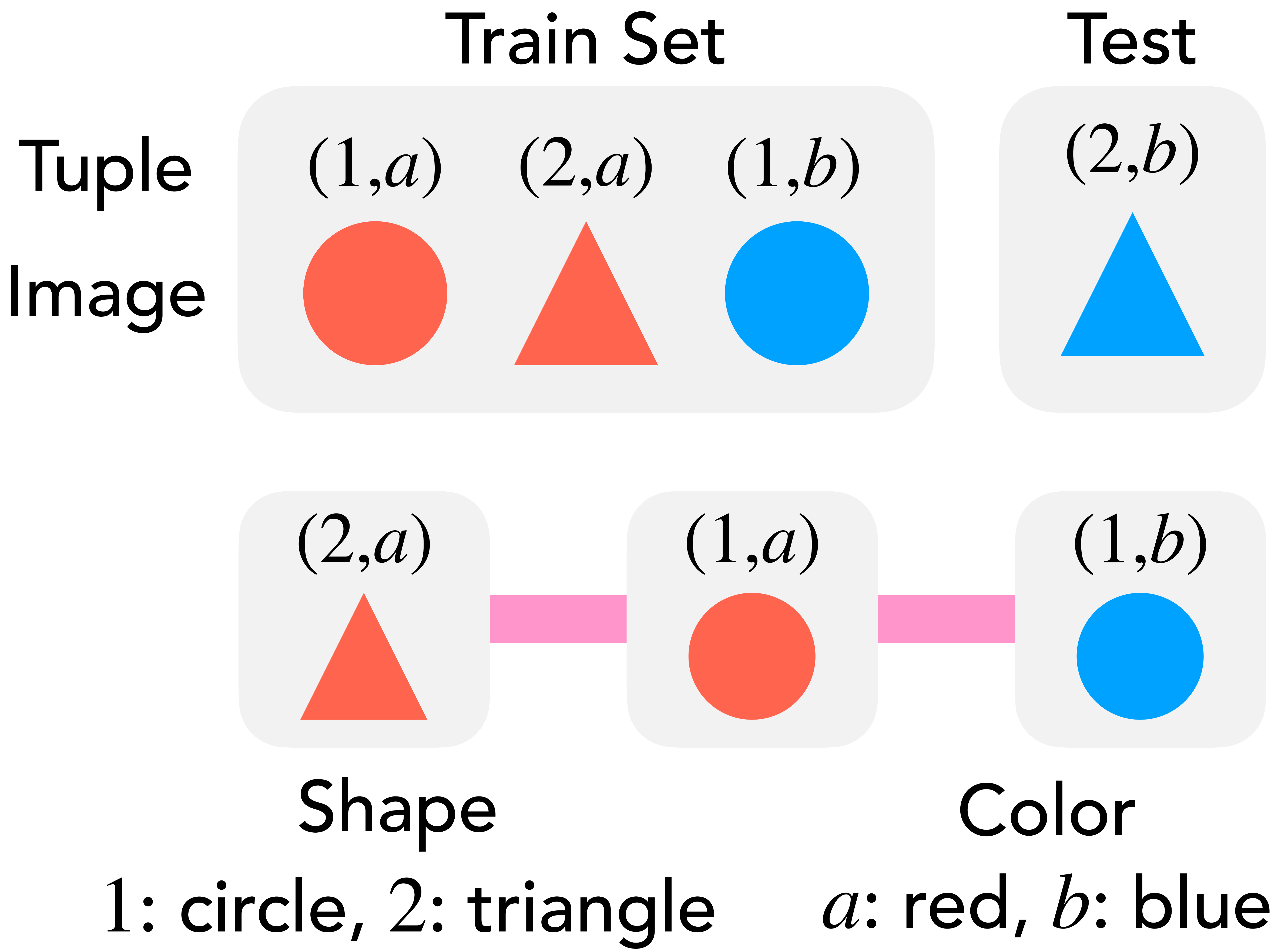
Our Approach: A Simple Task Requiring Compositional Generalization

Q. "Generate an object corresponding to (2,b)"



Our Approach: A Simple Task Requiring Compositional Generalization

Q. “Generate an object corresponding to (2,b)”



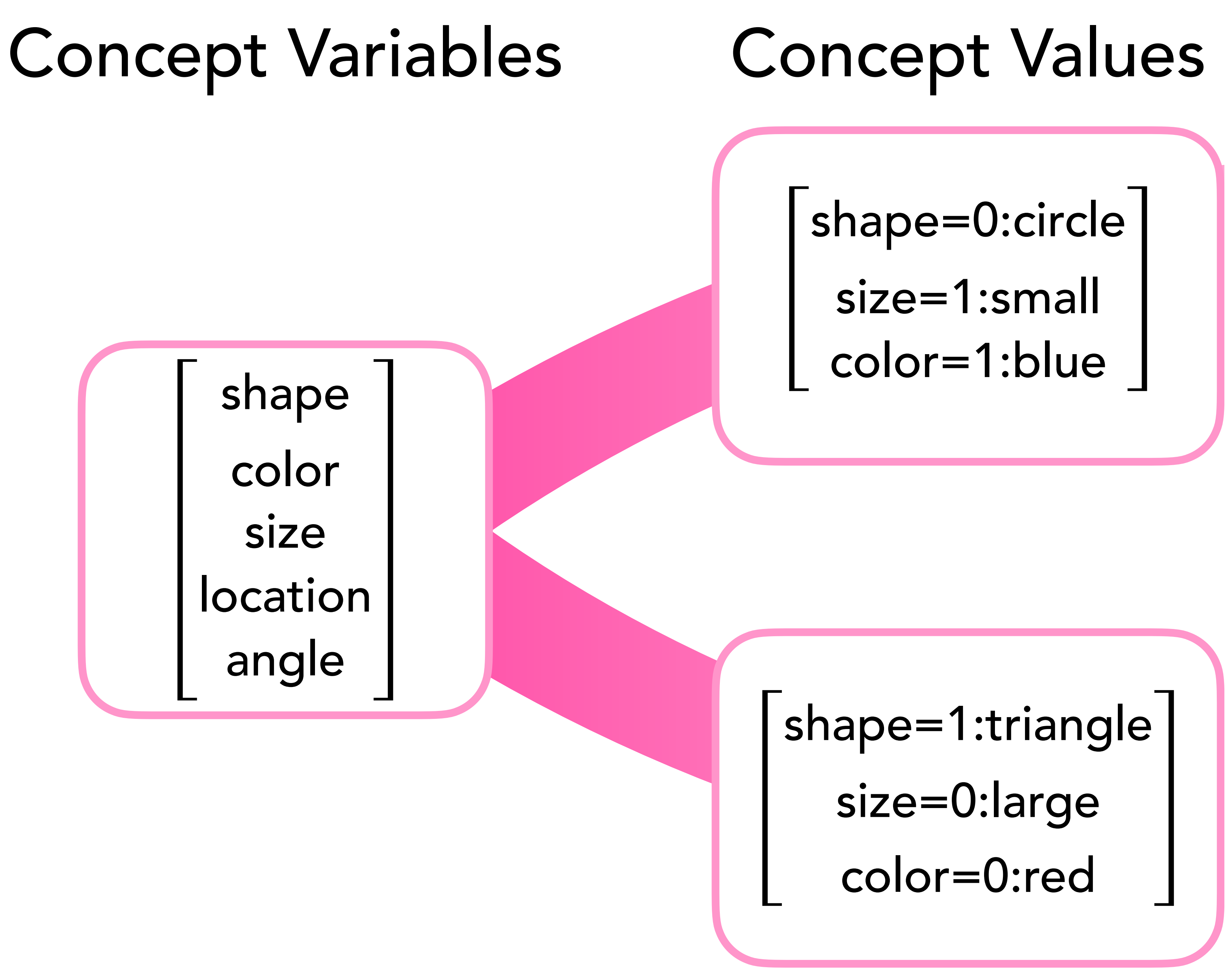
Concept Graph: A Novel Model of Compositional Structures

Concept Variables

shape
color
size
location
angle

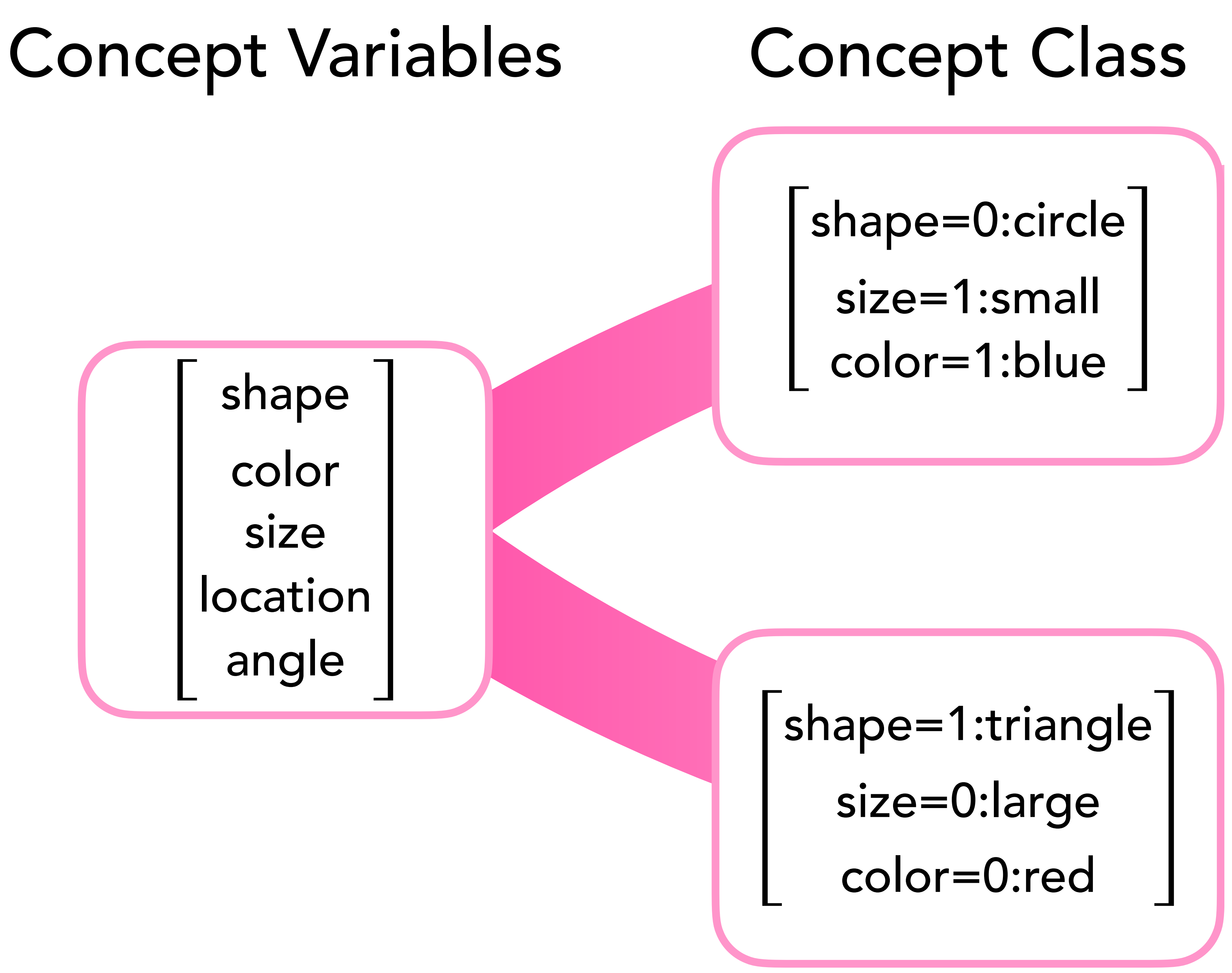
Definition 1. ***(Concept Variables.)*** *Let $V = \{v_1, v_2, \ldots, v_n\}$ be a set of n concept variables, where each v represents a specific property of an object.*

Concept Graph: A Novel Model of Compositional Structures



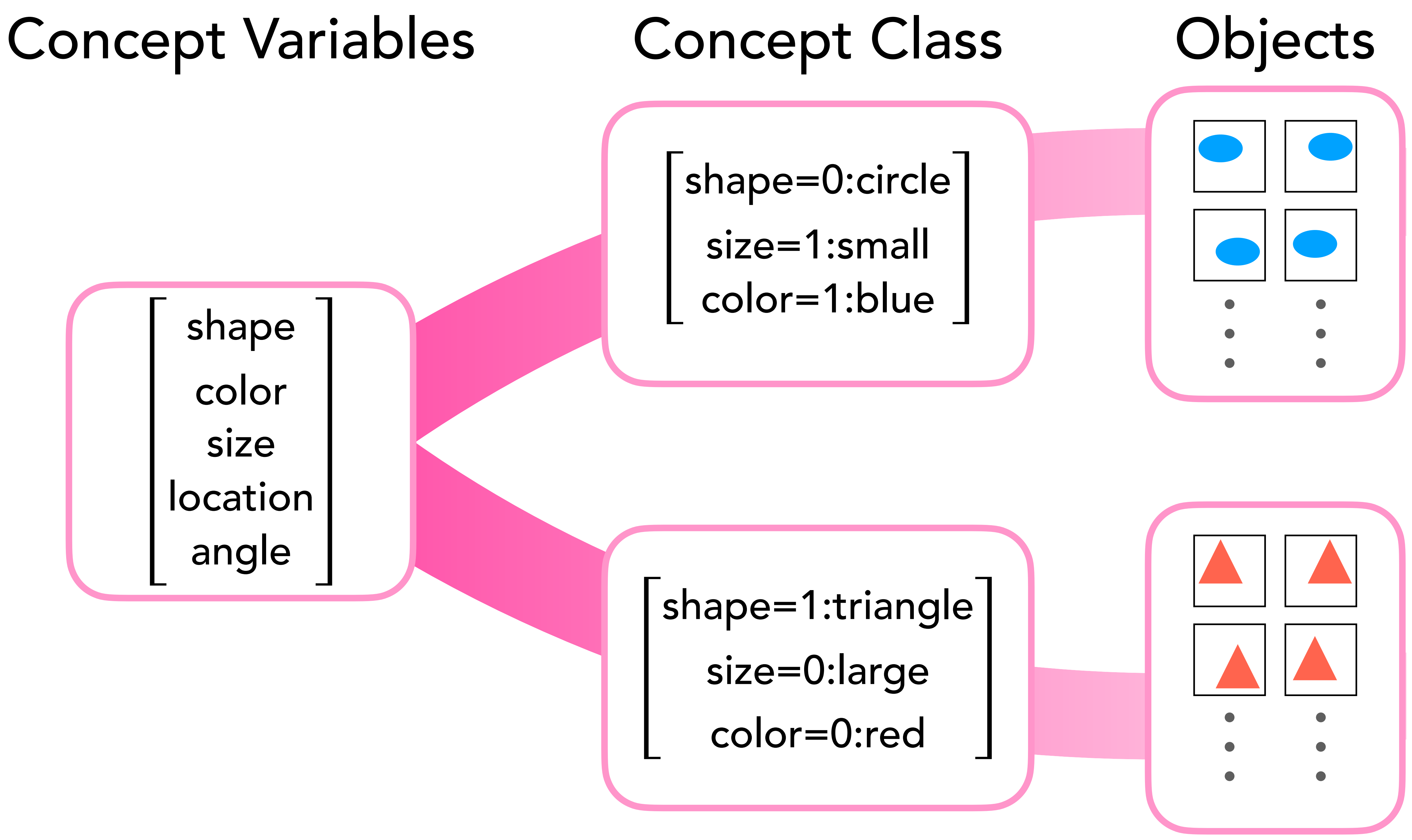
Definition 2. *(Concept Values.)* *For each concept variable $v_i \in V$, let $C_i = \{c_{i1}, c_{i2}, \ldots, c_{ik_i}\}$ be the set of k_i possible values that v_i can take. Each element of the set C_i is called a concept value.*

Concept Graph: A Novel Model of Compositional Structures

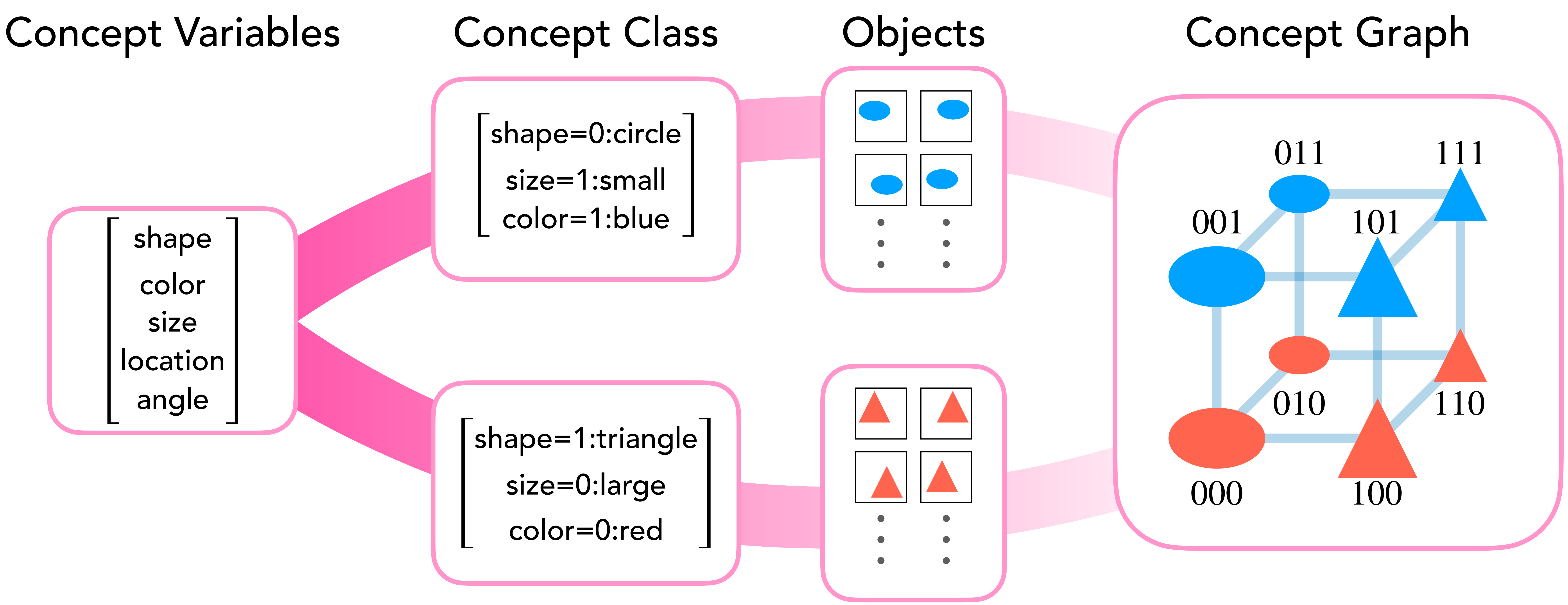


Definition 3. ***(Concept Class.)*** A concept class C is an ordered tuple $(v_1 = c_1, v_2 = c_2, \ldots, v_p = c_p)$, where each $c_i \in C_i$ is a concept value corresponding to the concept variable v_i. If an object x belongs to concept class C, then $v_i(x) = c_i \, \forall i \in 1, \ldots, p$.

Concept Graph: A Novel Model of Compositional Structures



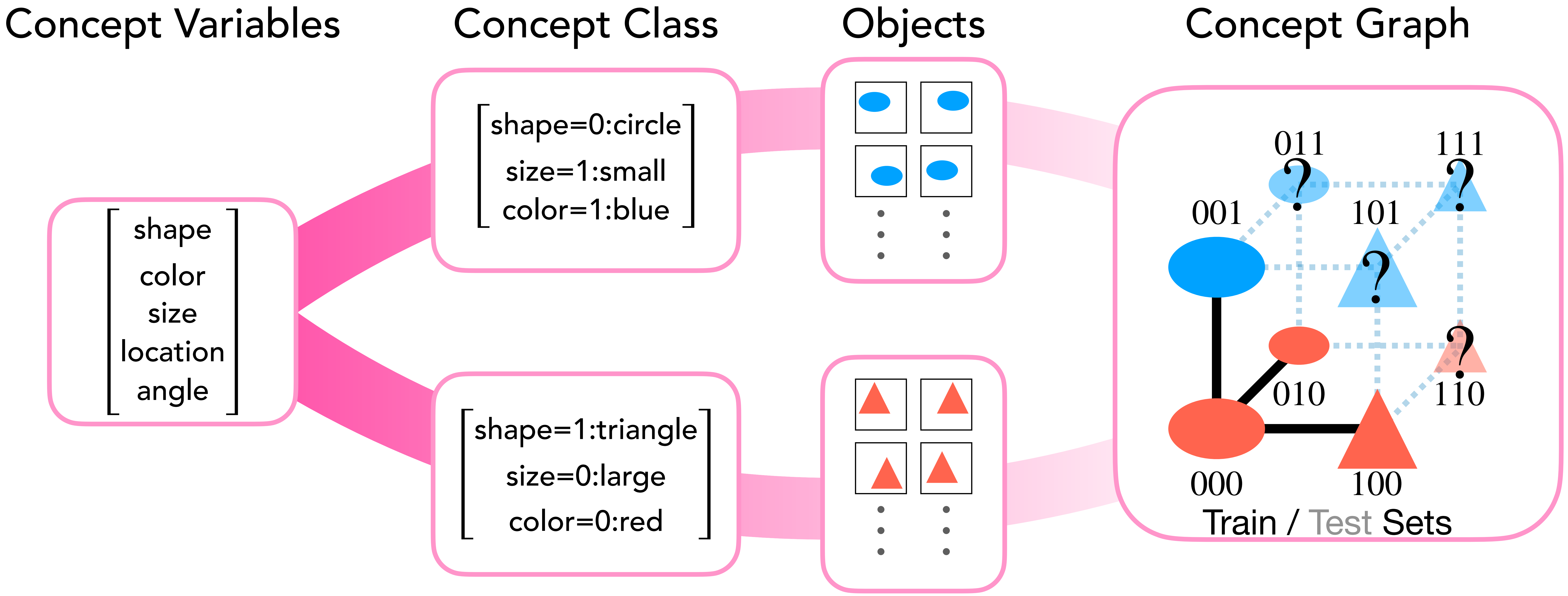
Concept Graph: A Novel Model of Compositional Structures



Definition 4. ***(Concept Distance.)*** Given two concept classes $C^{(1)} = (c_1^{(1)}, c_2^{(1)}, \ldots, c_n^{(1)})$ and $C^{(2)} = (c_1^{(2)}, c_2^{(2)}, \ldots, c_n^{(2)})$, the concept distance $d(C^{(1)}, C^{(2)})$ is defined as the number of elements that differ between the two concept classes:

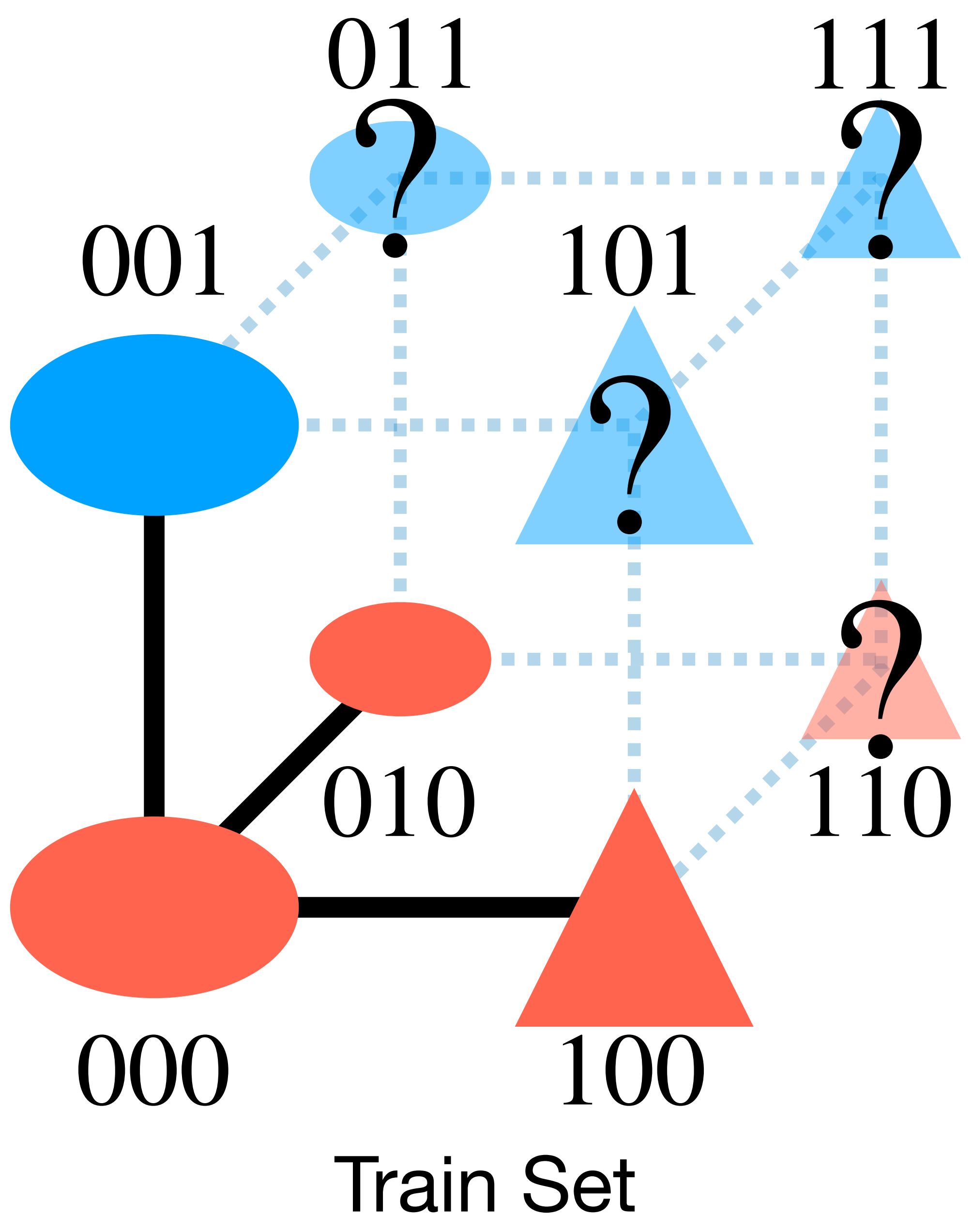
$$d(C^{(1)}, C^{(2)}) = \sum_{i=1}^{n} I(c_i^{(1)}, c_i^{(2)})$$

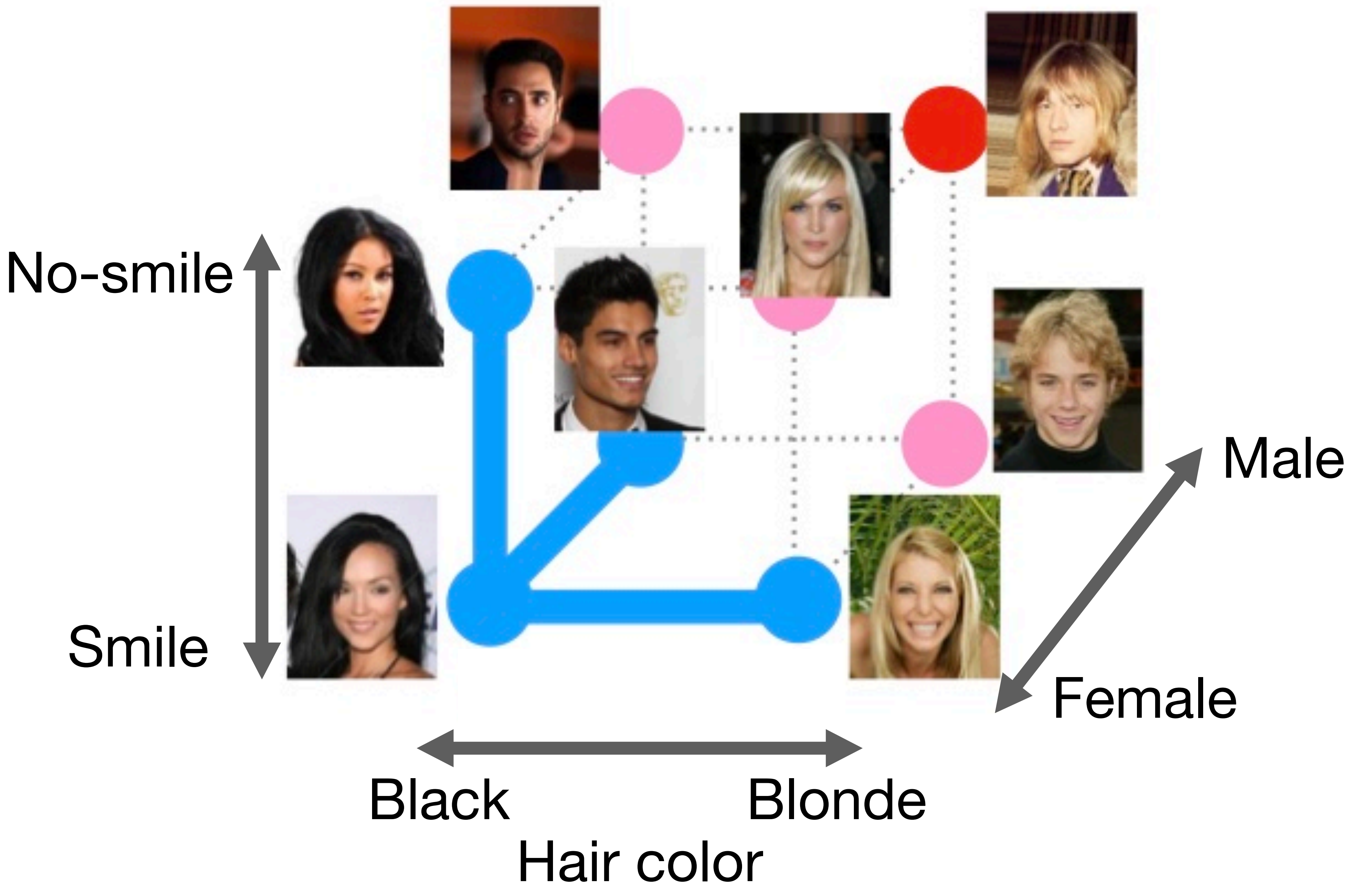
Concept Graph: A Novel Model of Compositional Structures



Definition 5. ***(Compositional Generalization.)*** Consider a model trained to generate samples from concept classes $\hat{C} = (C_1, C_2, \ldots, C_n)$. We say the model *compositionally generalizes* if it can generate samples from a class $\tilde{C}$ such that $d(\tilde{C}, C) \geq 1 \, \forall C \in \hat{C}$.

How do compositional structures shape neural learning and computation?



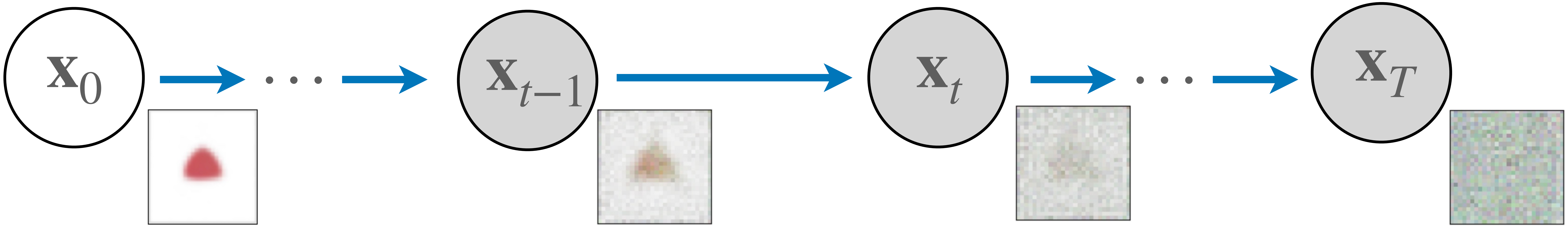


Q1. Can a “diffusion model” generalize to concept classes it has *never* seen in the training set?

Q2. If so, in what order does the diffusion model generalize?

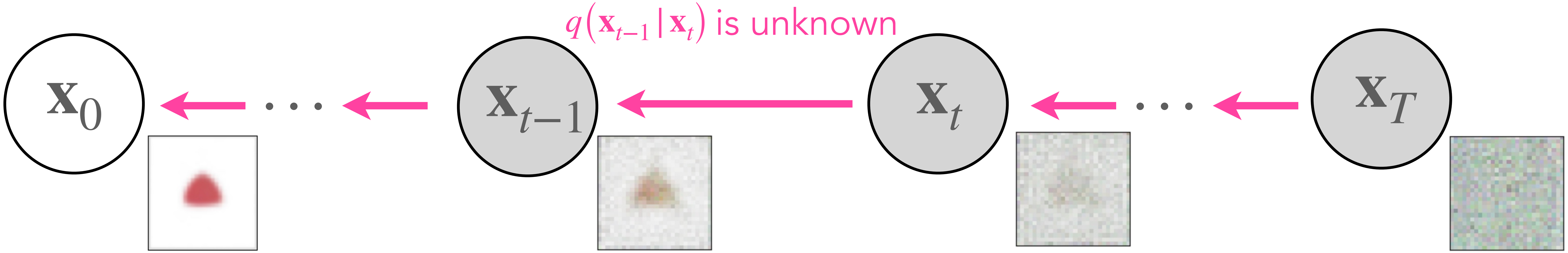
Diffusion model: Neural network model for image generation

Step 1. Forward Diffusion: Take an image $\mathbf{x}_0$ and keep adding Gaussian noise.



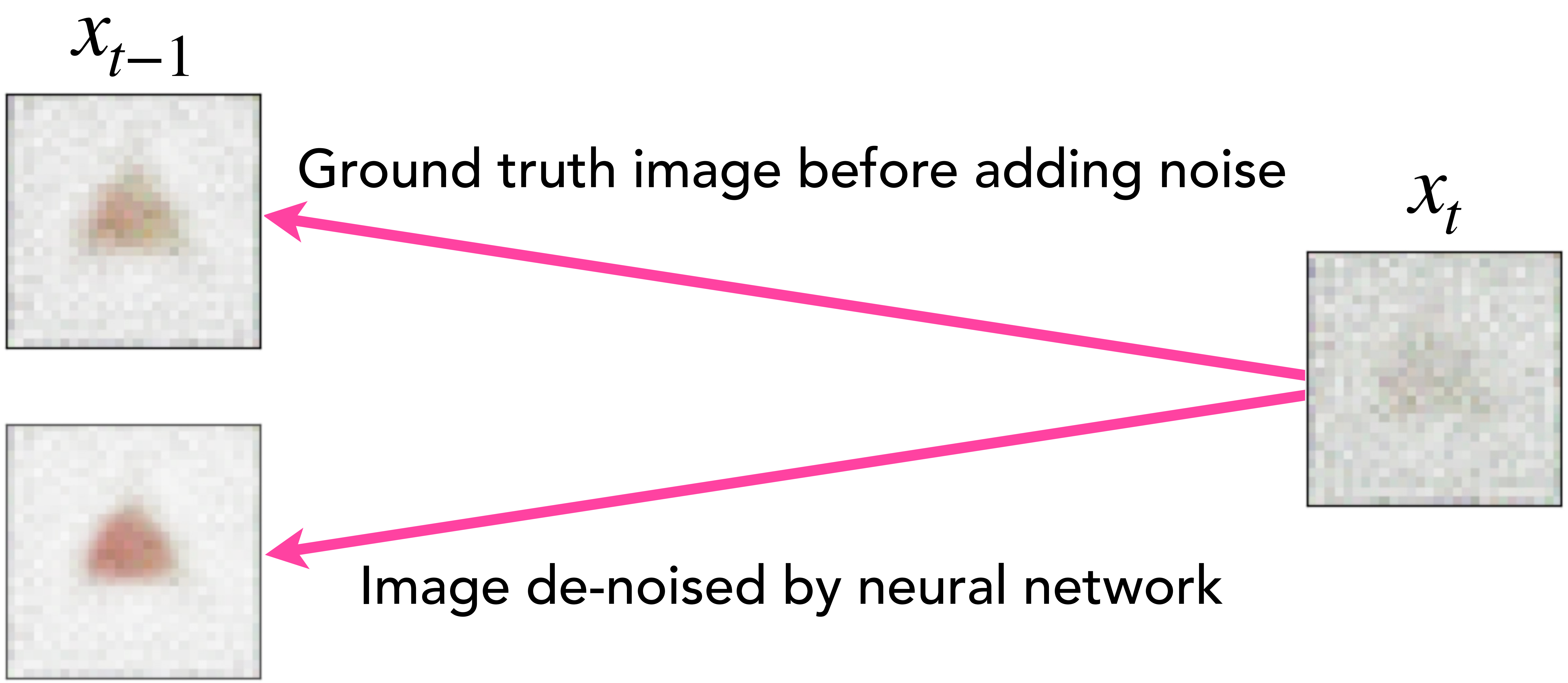
$$x_t = x_{t-1} + \xi, \text{ where } \xi \sim \mathcal{N}(0, I)$$

Step 2. Learning Reverse Process: Learn non-linear mapping to de-noise image from x_t to x_{t-1}.



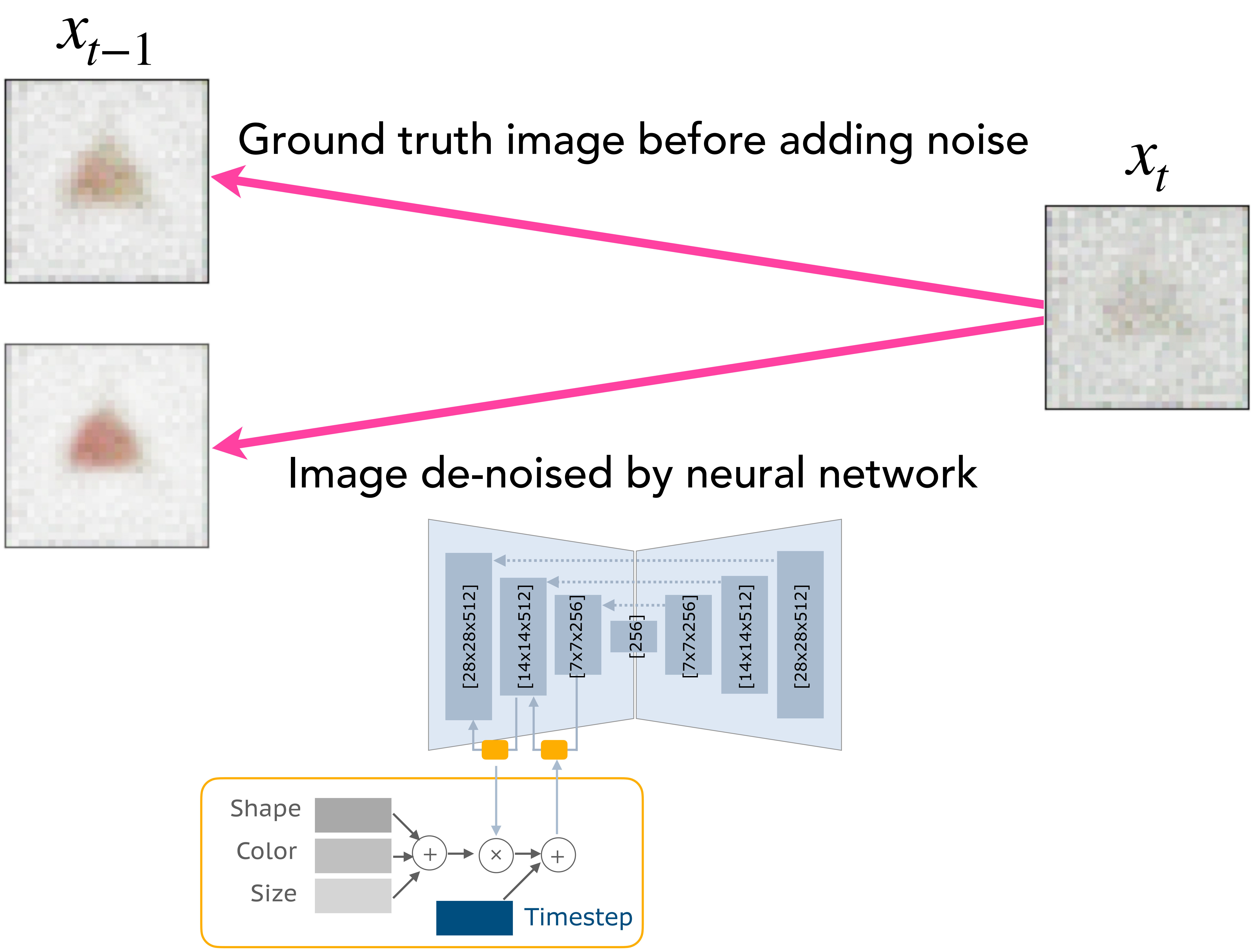
"Deep Unsupervised Learning using Nonequilibrium Thermodynamics"
J. Sohl-Dickstein, E.A. Weiss, N. Maheswaranathan, S. Ganguli *ICML (2015)*

Diffusion model: Neural network model for image generation



$$f(x_t; \Theta, \text{shape}, \text{size}, \text{color})$$

Diffusion model: Neural network model for image generation



Diffusion model: Neural network model for image generation

x_{t-1}

Ground truth image before adding noise

x_t

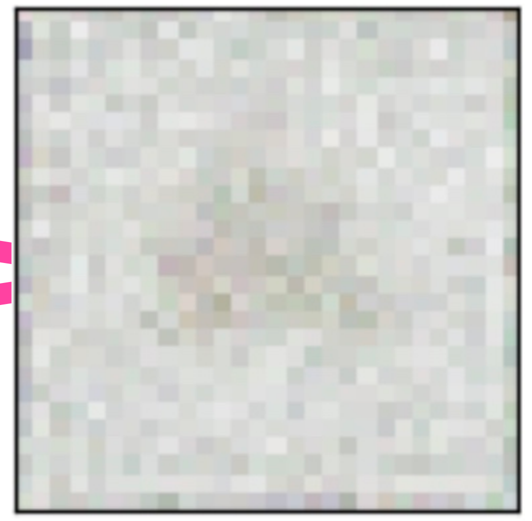
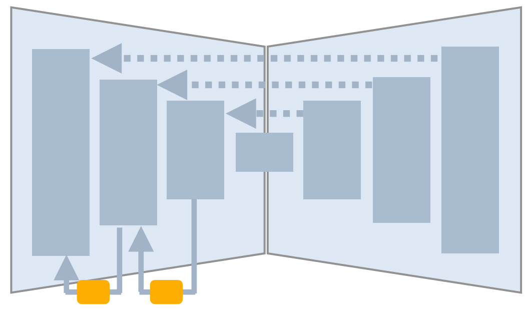
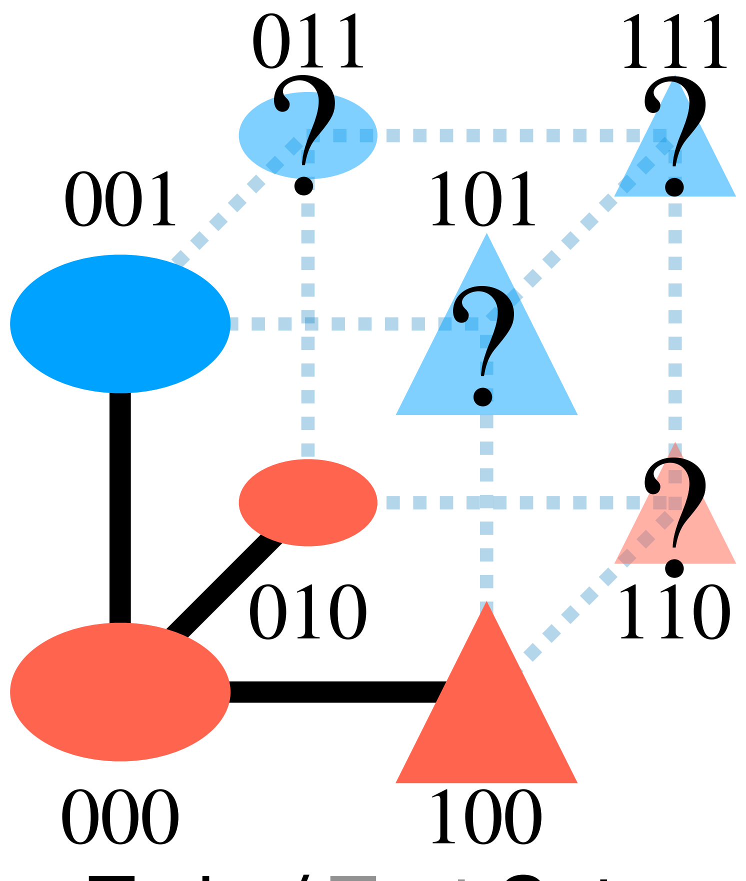


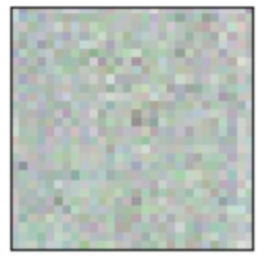
Image de-noised by neural network

$f(x_t; \Theta, \text{shape, size, color})$ 

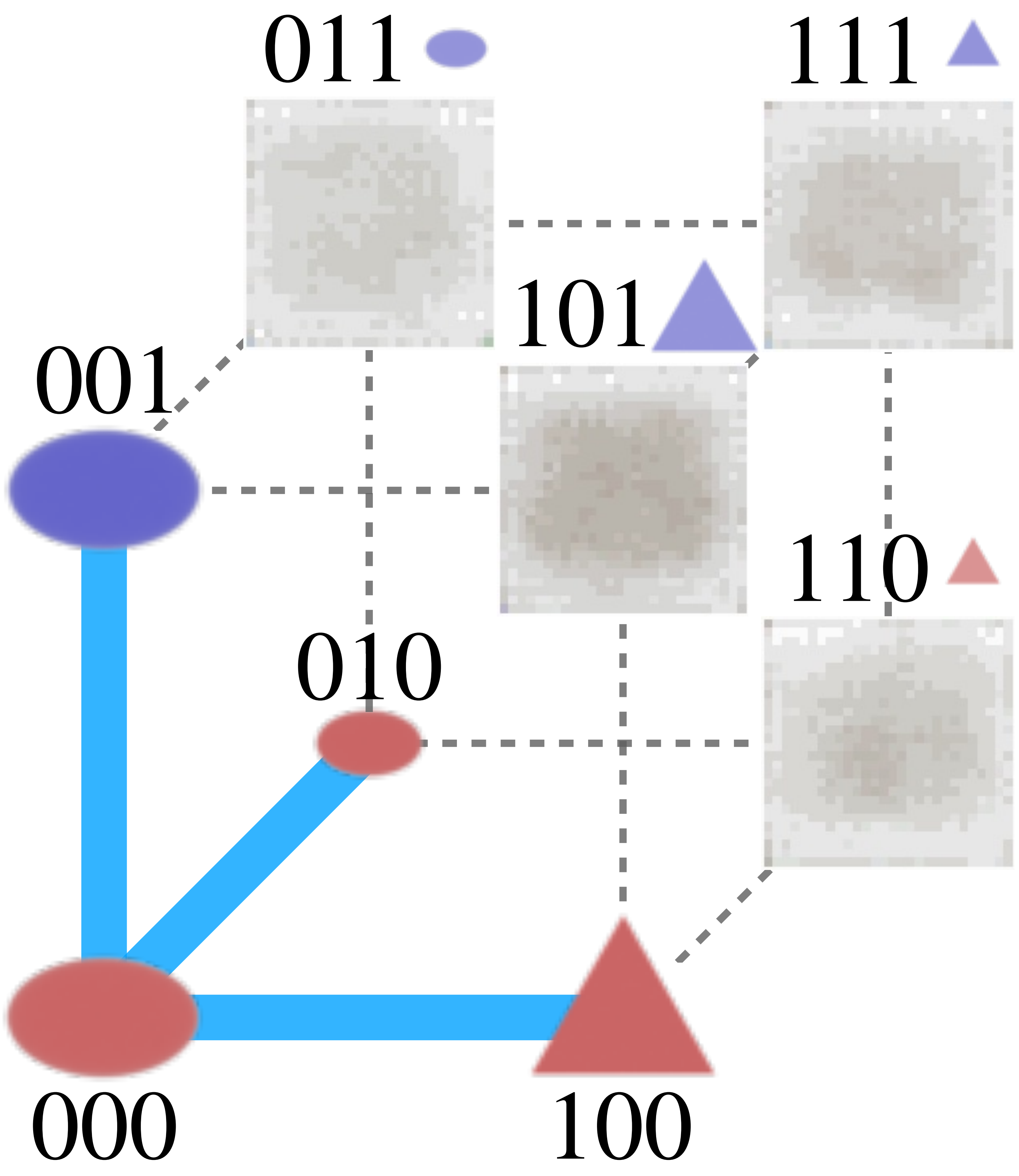
$\text{argmin}_\Theta \, | \, x_{t-1}$  $- f(x_t; \Theta, \text{shape, size, color})$  $|^2$



Step 3. Generate image from noise using the learned function with optimized parameters (Θ)

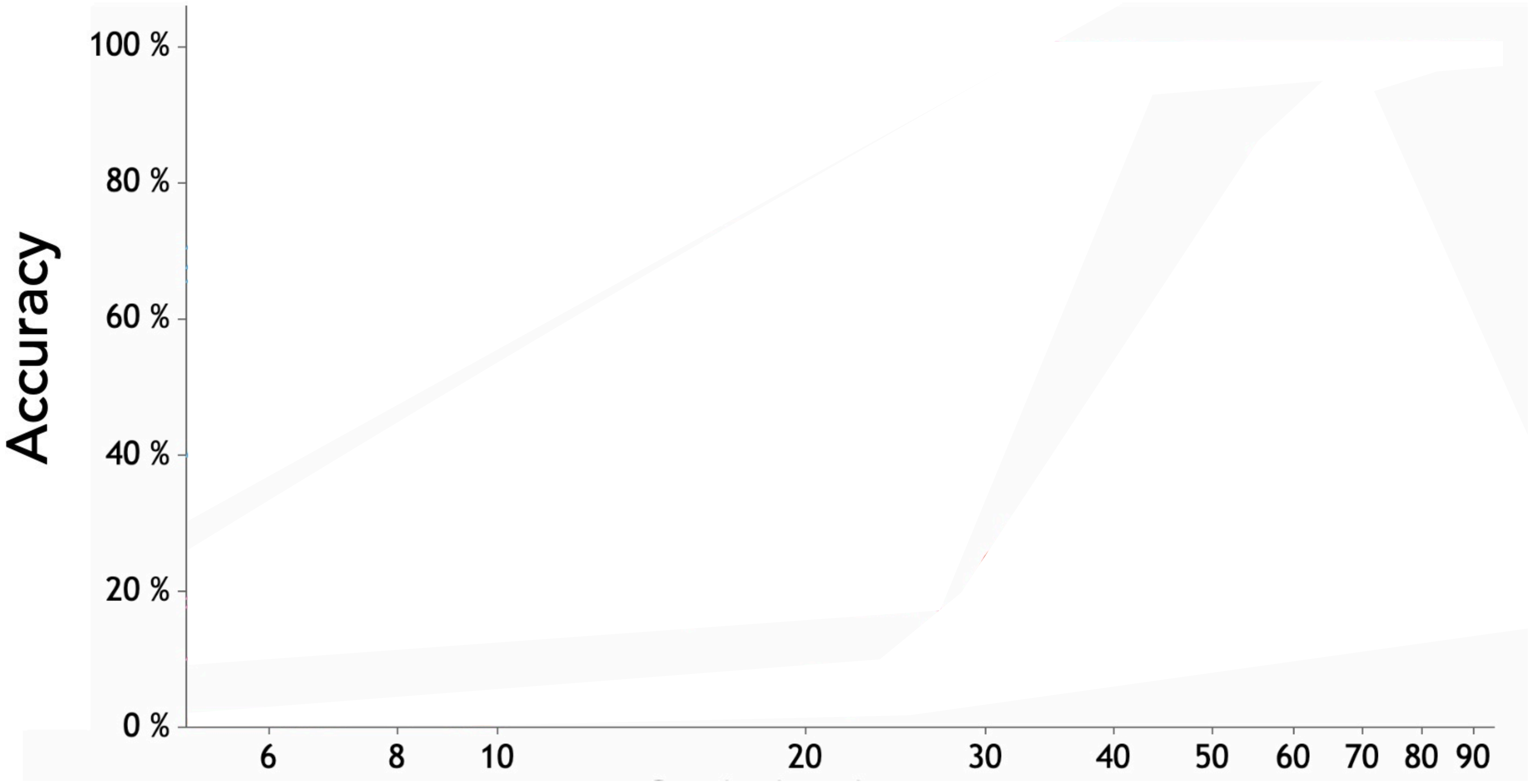
$f \circ f \circ \ldots f \circ f($  $; \Theta, \text{triangle, small, blue}) =$ 

Can a neural network learn concepts and compose them in new ways?

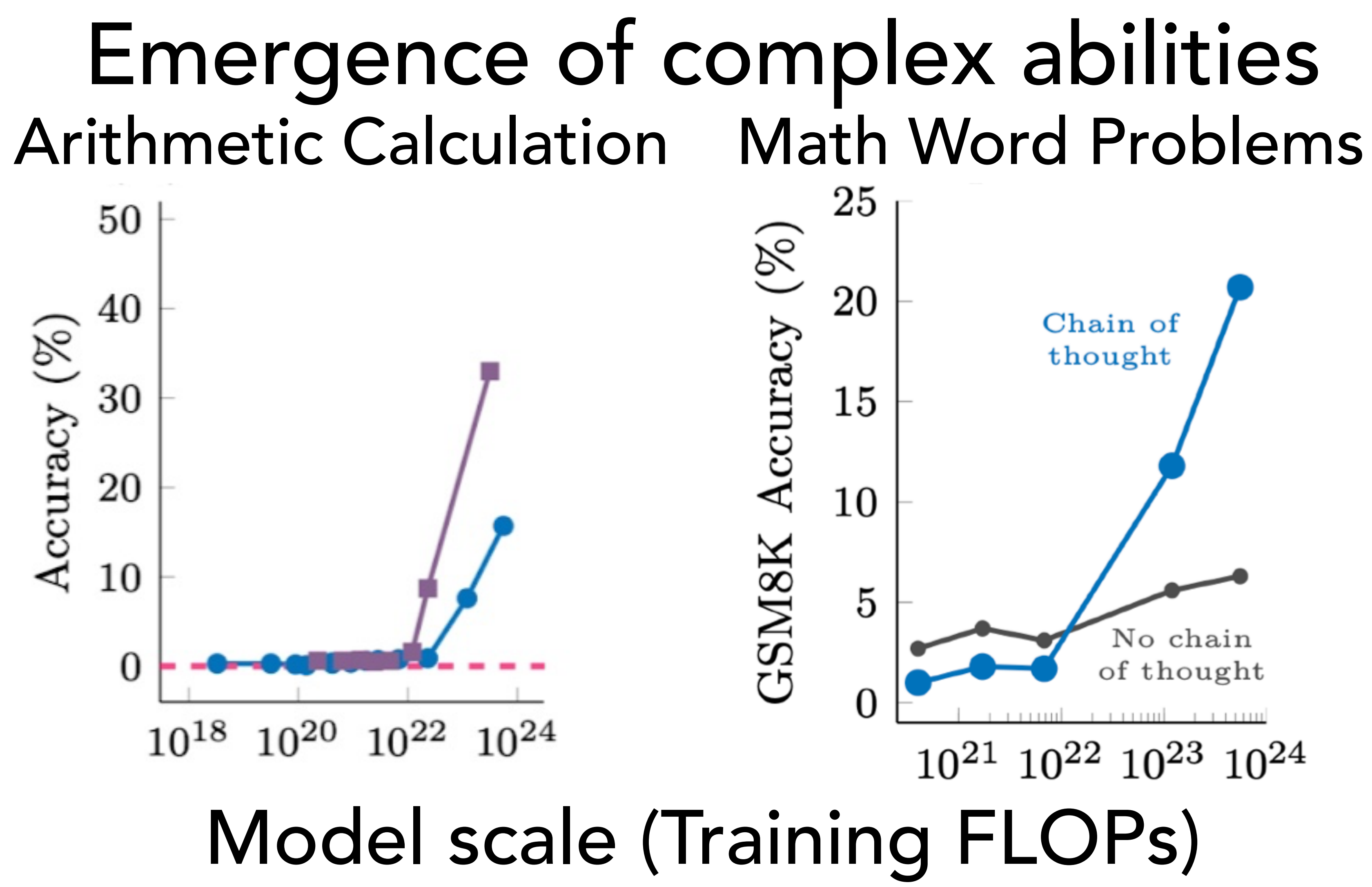


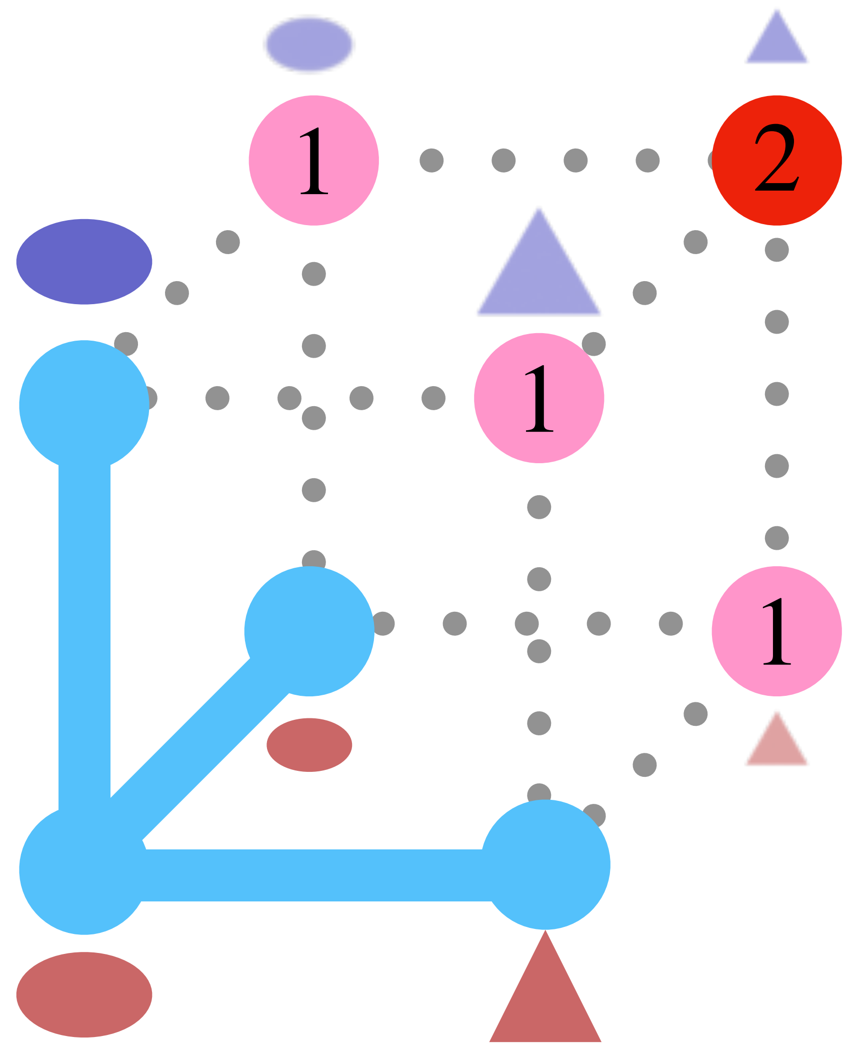
Generated outputs of the diffusion model as a function of optimization steps

Concept distance governs the order of generalization



Optimization time (epochs)

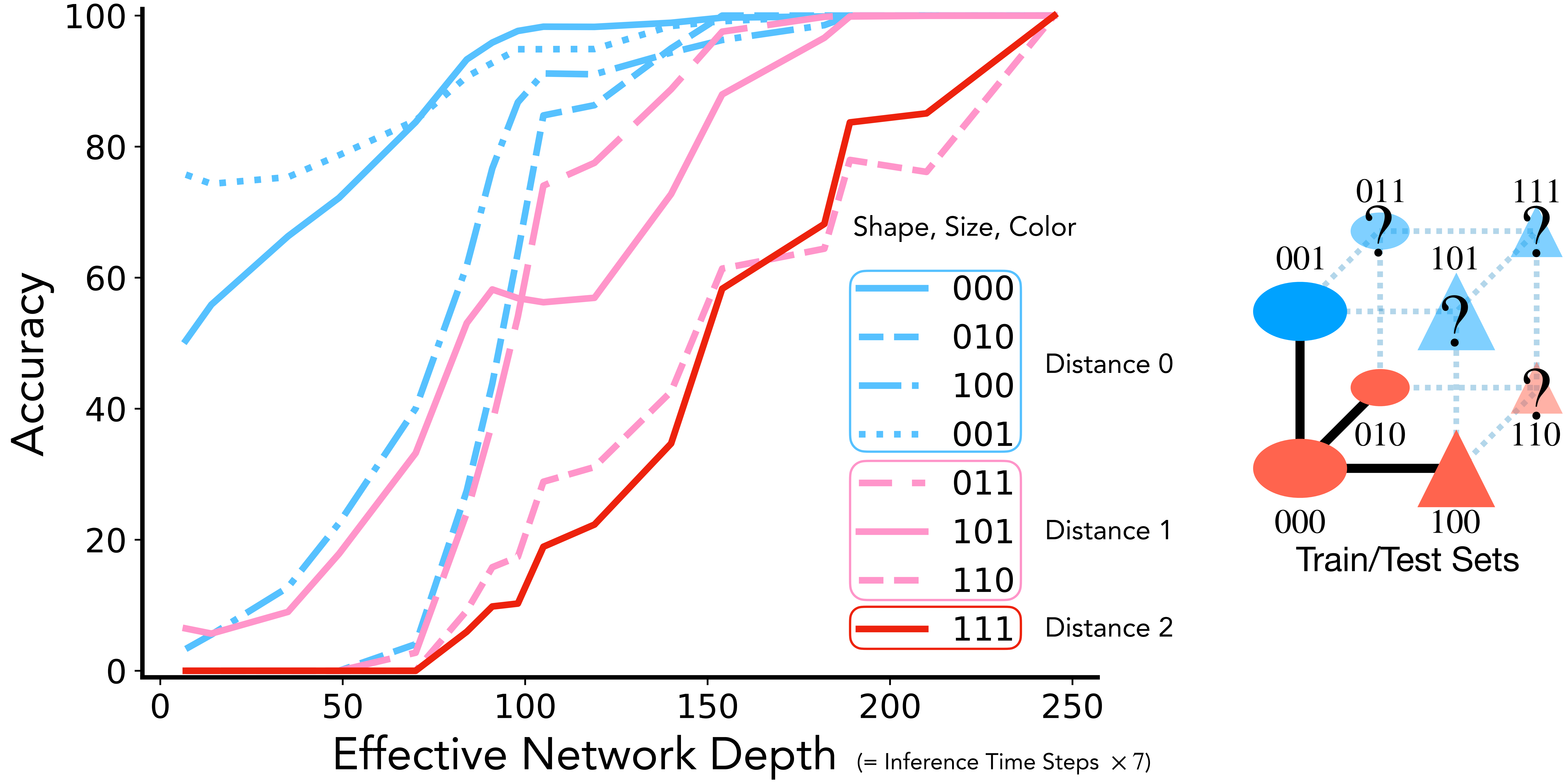




Training data

Accuracy: Train linear probes for each concept and measure accuracy as product of probability of correctness of concept (a usual metric in Disentanglement literature)

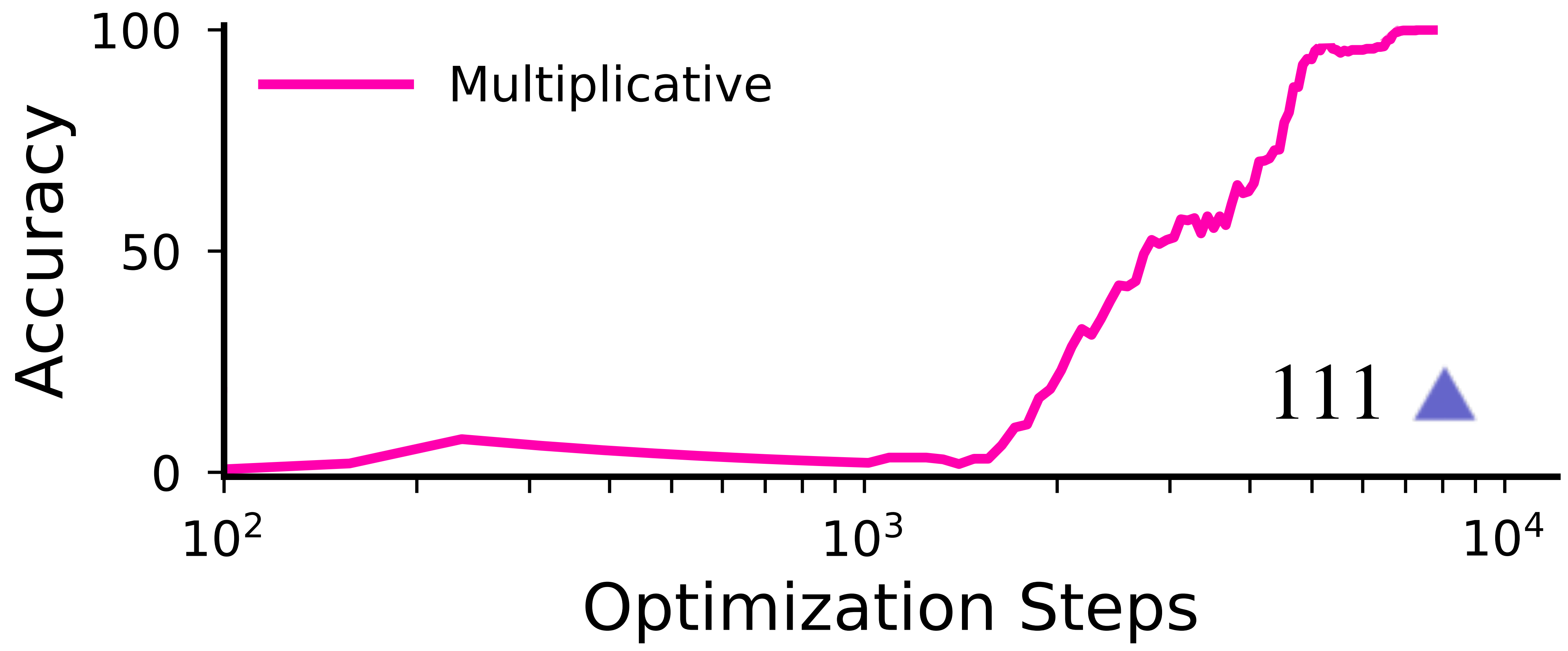
Why can the diffusion model generalize compositionally?
Effective "**network depth**"!

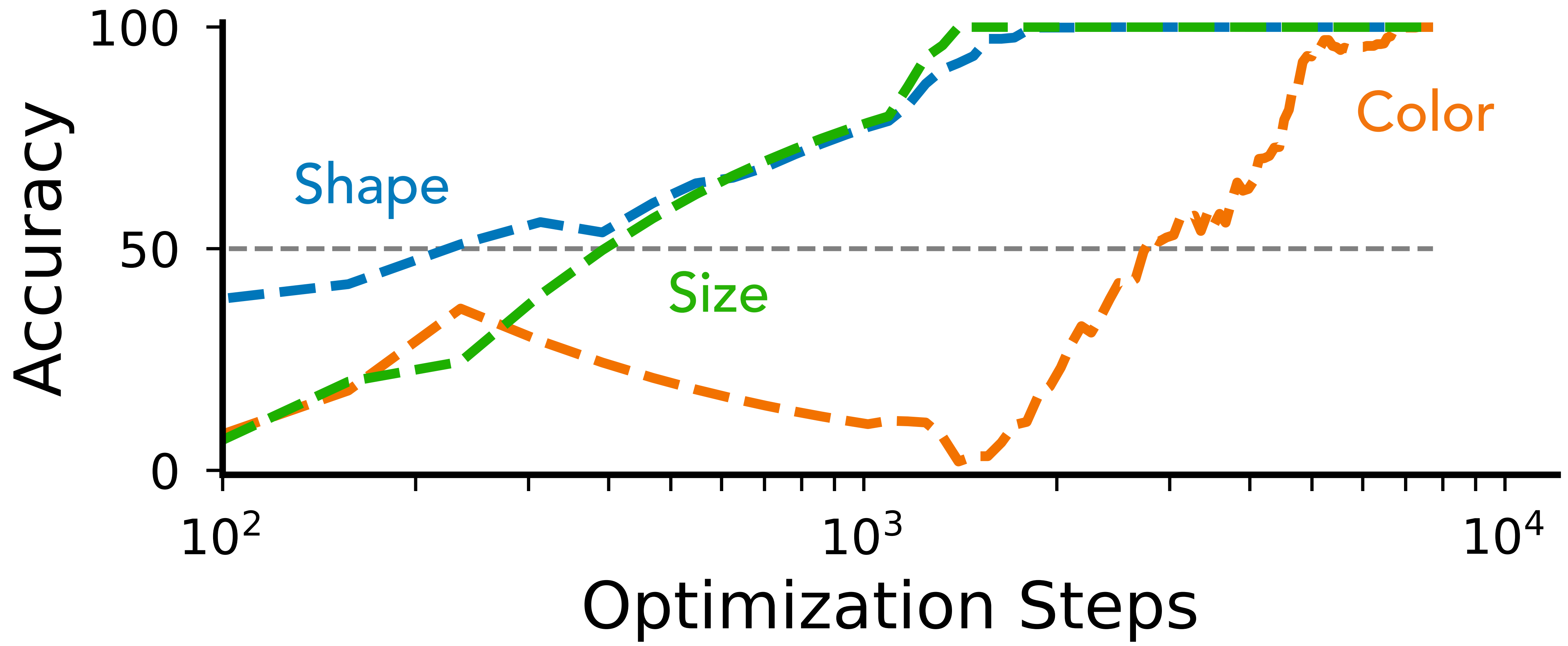


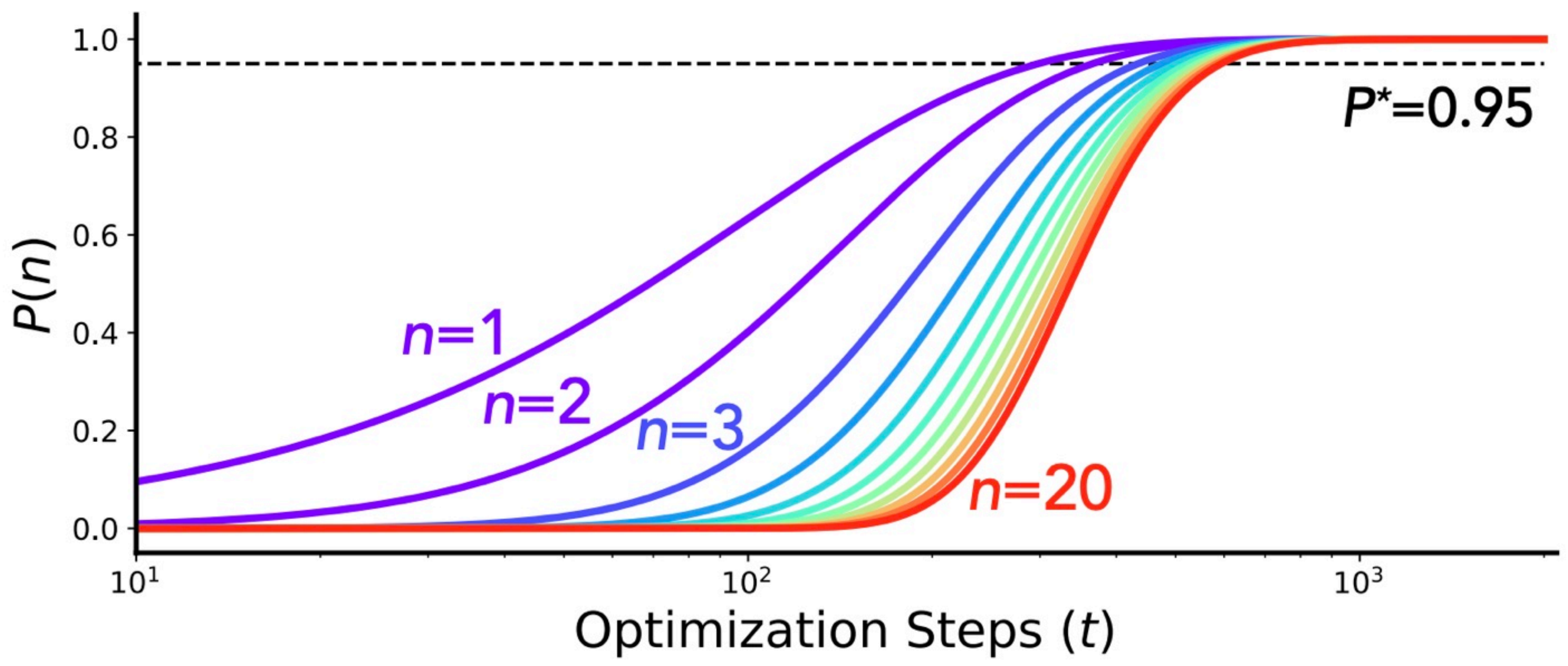
Compositionality underlies the emergence

Claim:

Capabilities that require composition of atomic abilities (skills) show emergent curves



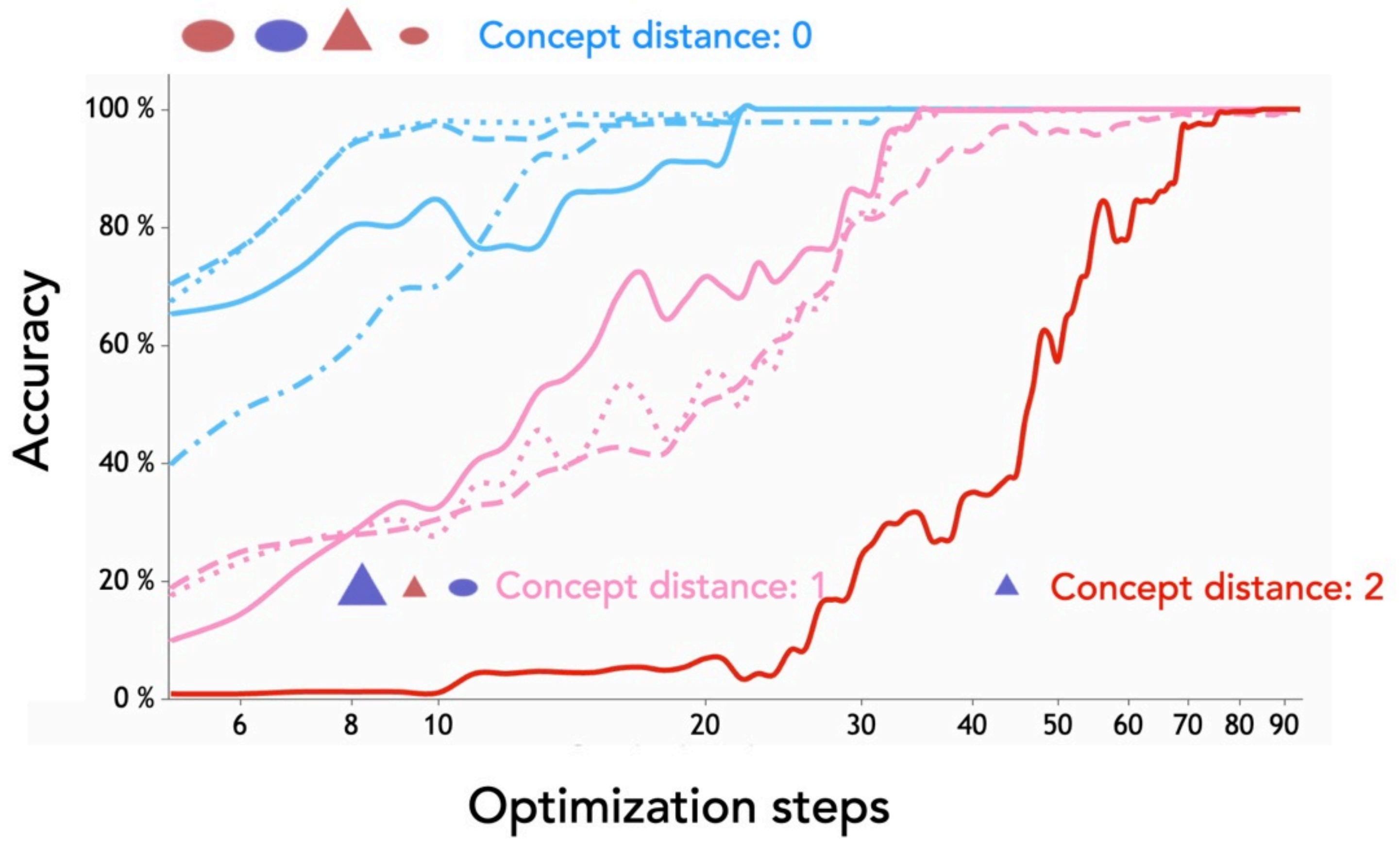




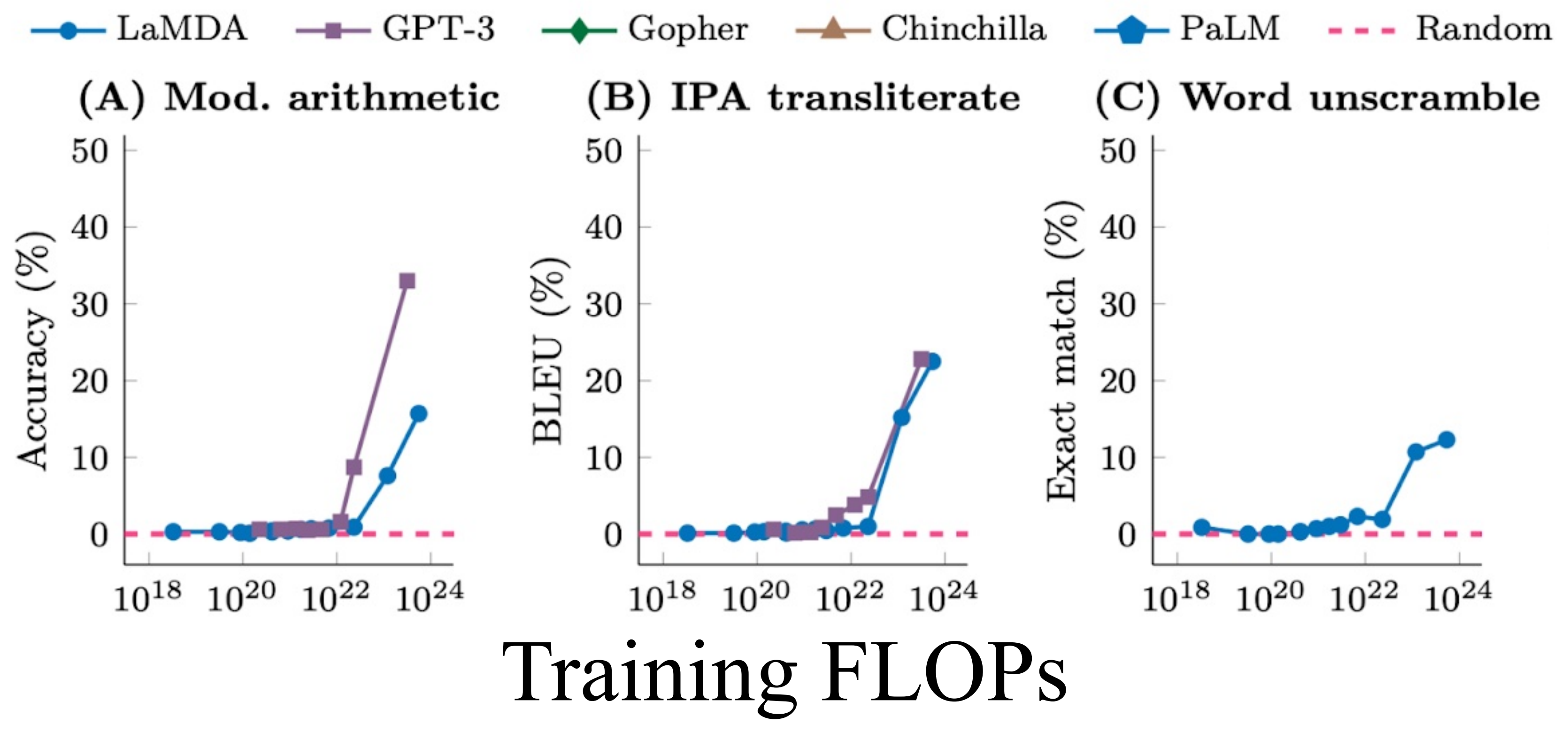
- There are n capabilities, each with a probability p of being learned in a given time step. (i.e., the dynamics of learning as a Bernoulli coin flip)
- The probability that the ability will be learned in t steps: $1 - (1 - p)^t$
- The probability that the compositional capability has been learned by time t is $P(n) = (1 - (1 - p)^t)^n$.
- The critical time t^* at which a compositional capability is learned: $$t^* = \frac{\log(1 - (P^*)^{1/n})}{\log(1 - p)}$$

The learning curve becomes sharper as the task becomes more compositional!

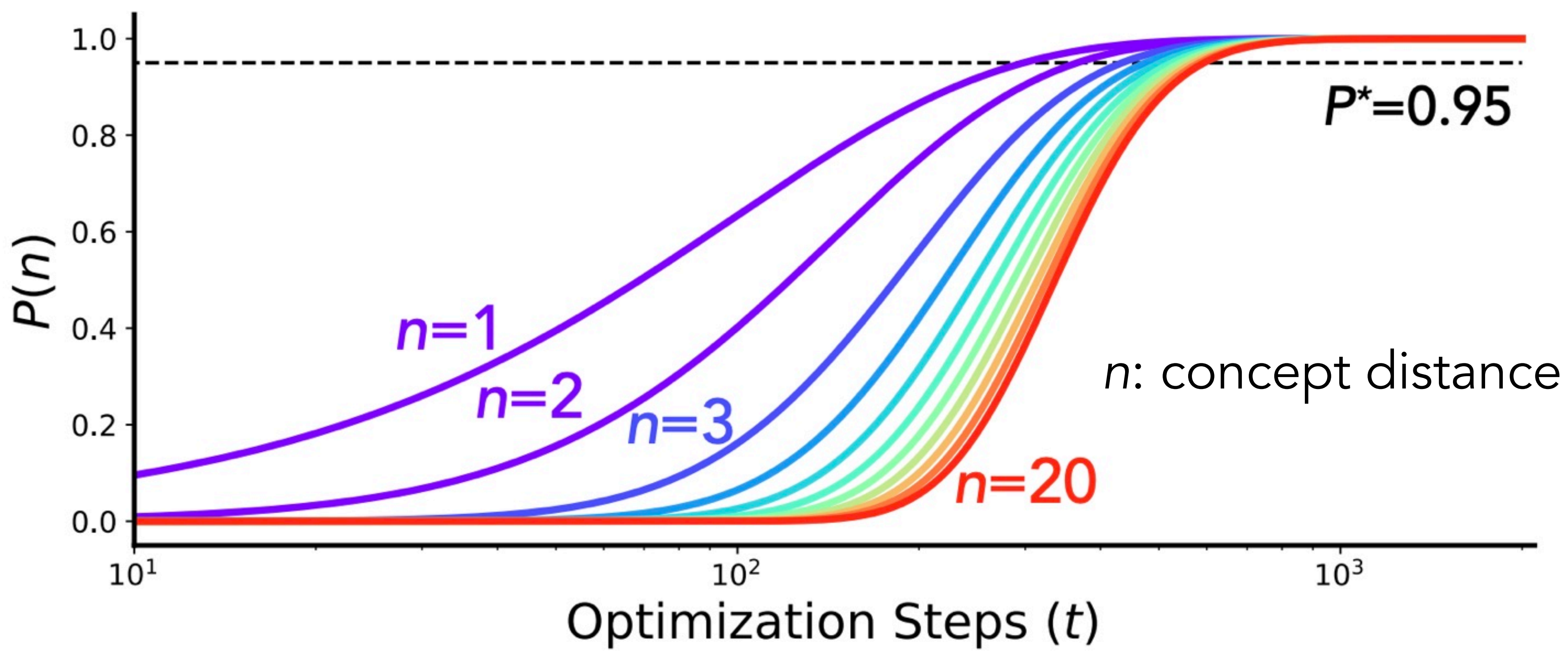
Compositionality underlies the emergence



Our experiment with diffusion models



Emergent abilities in large language models

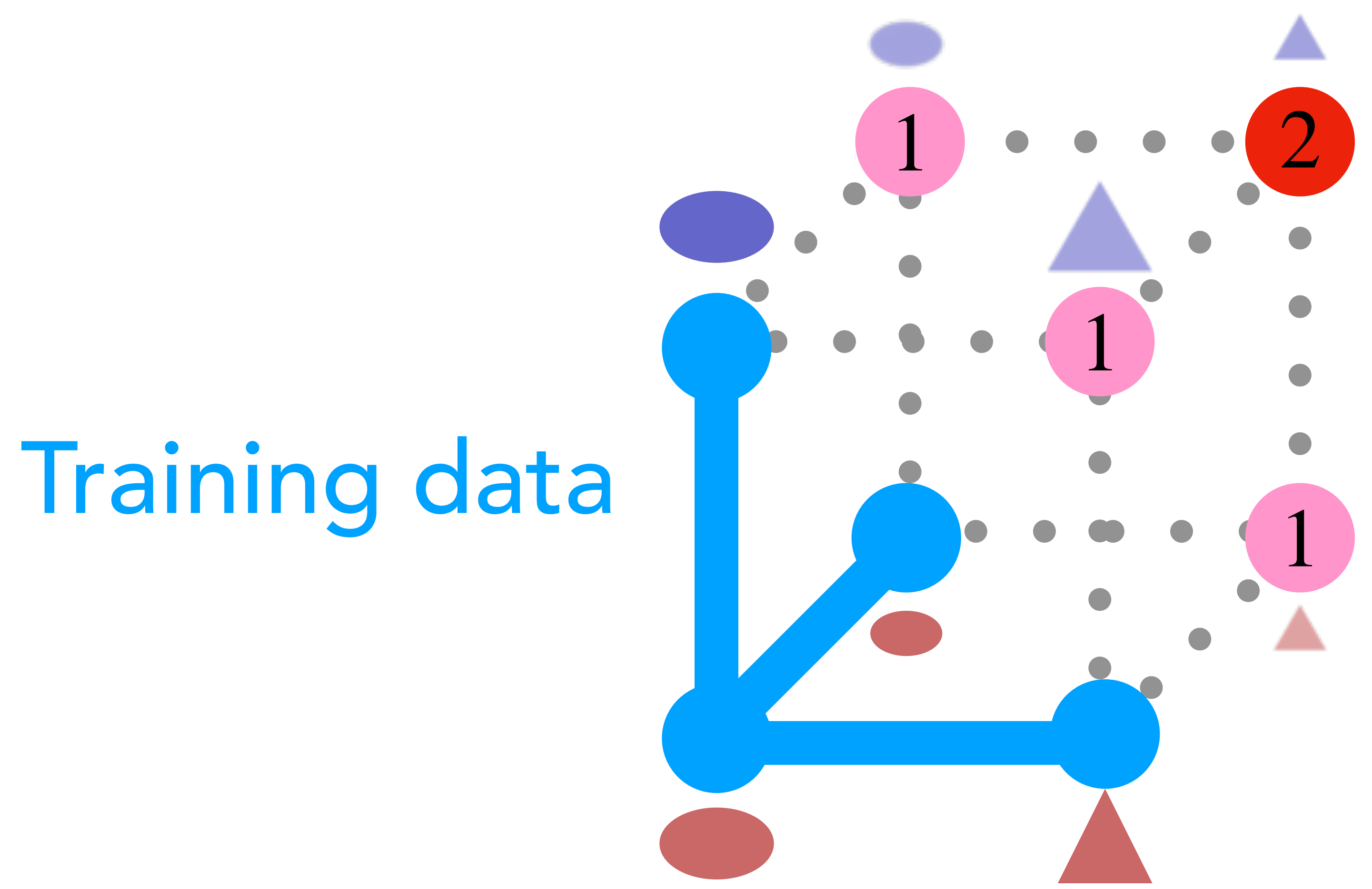


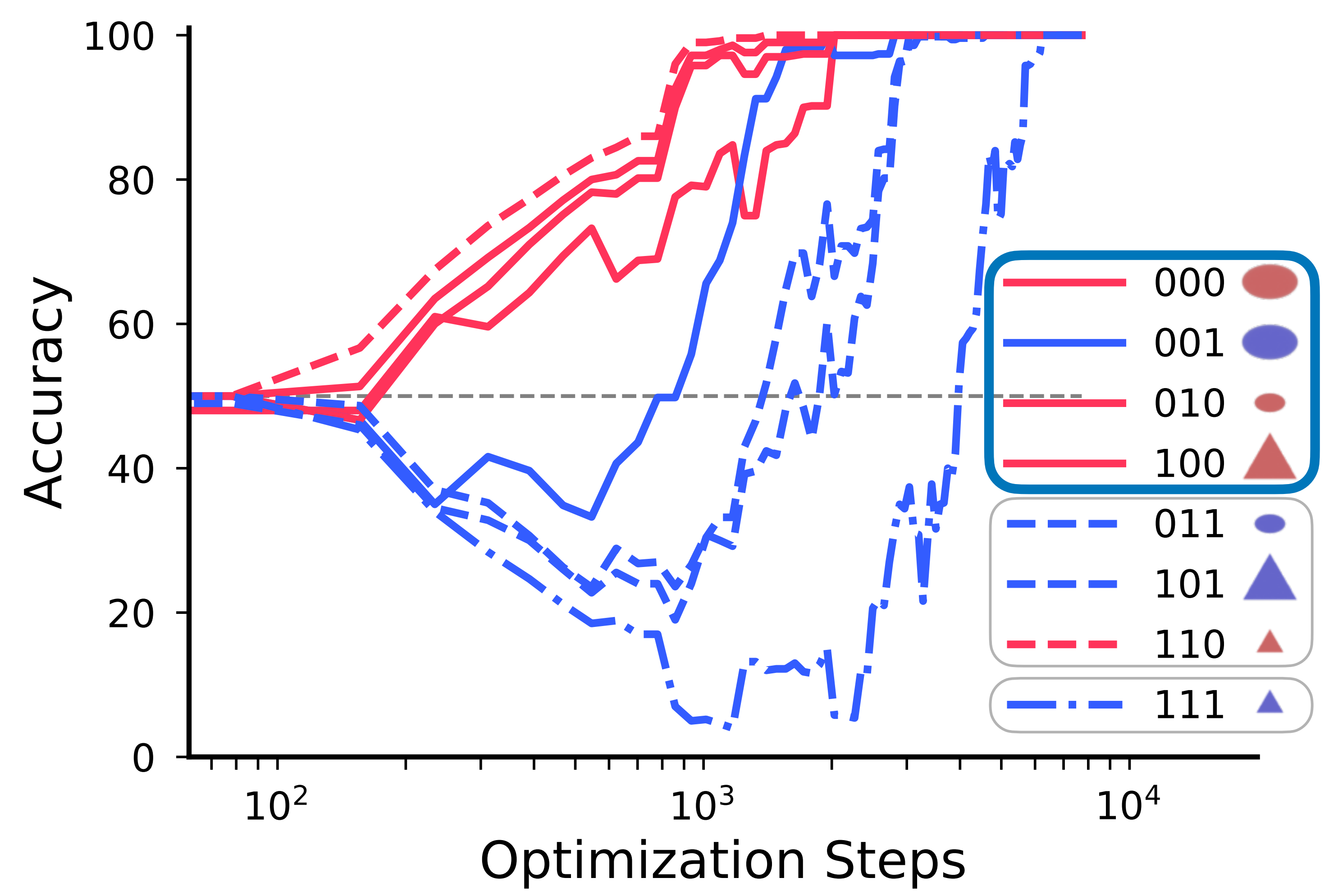
- There are n capabilities, each with a probability p of being learned in a given time step. (i.e., the dynamics of learning as a Bernoulli coin flip)
- The probability that the ability will be learned in t steps: $1-(1-p)^t$
- The probability that the compositional capability has been learned by time t is $P(n)=(1-(1-p)^t)^n$.
- The critical time t^* at which a compositional capability is learned:

$$t^* = \frac{\log(1-(P^*)^{1/n})}{\log(1-p)}$$

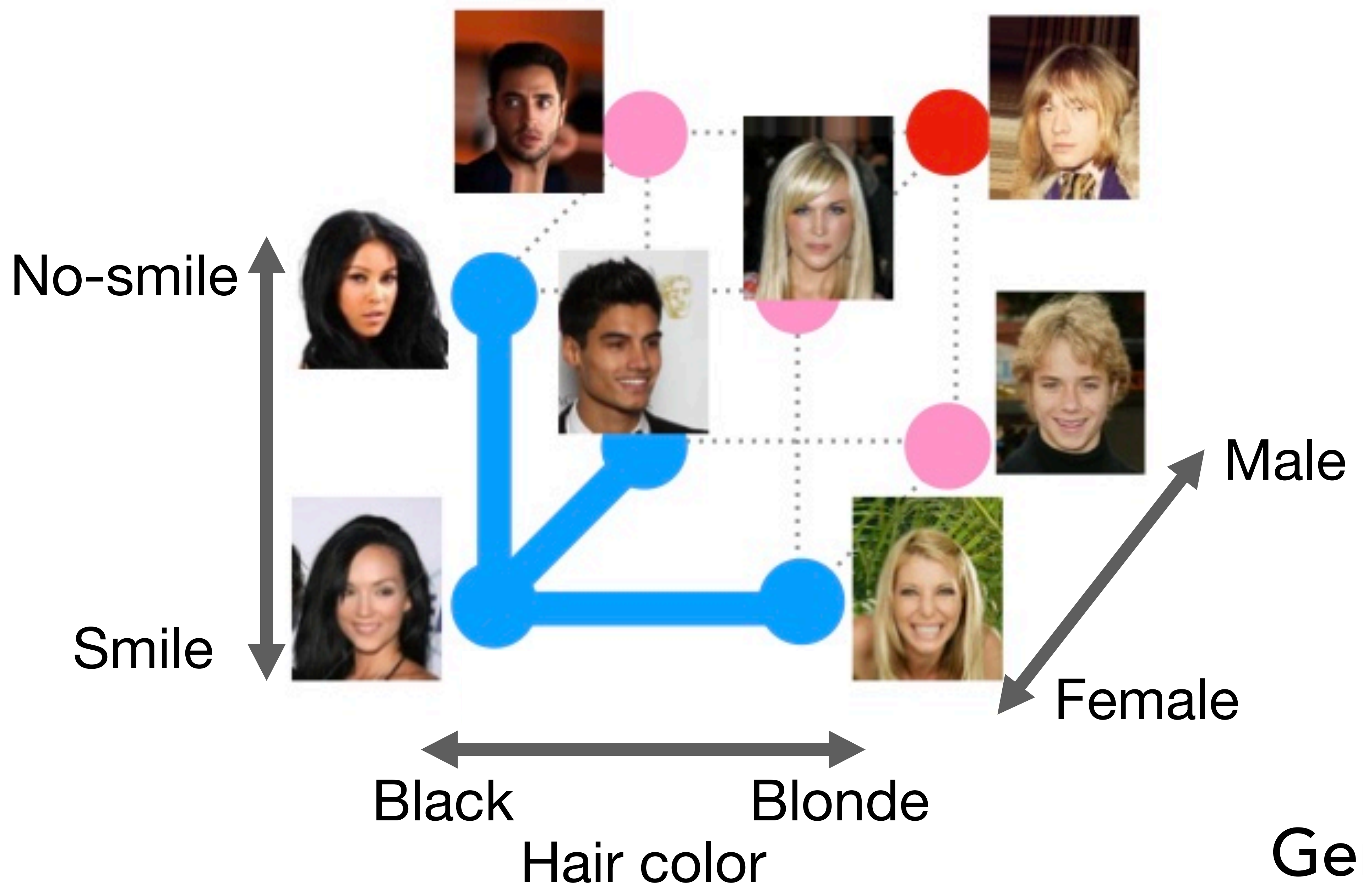
The learning curve becomes sharper as the task becomes more compositional!

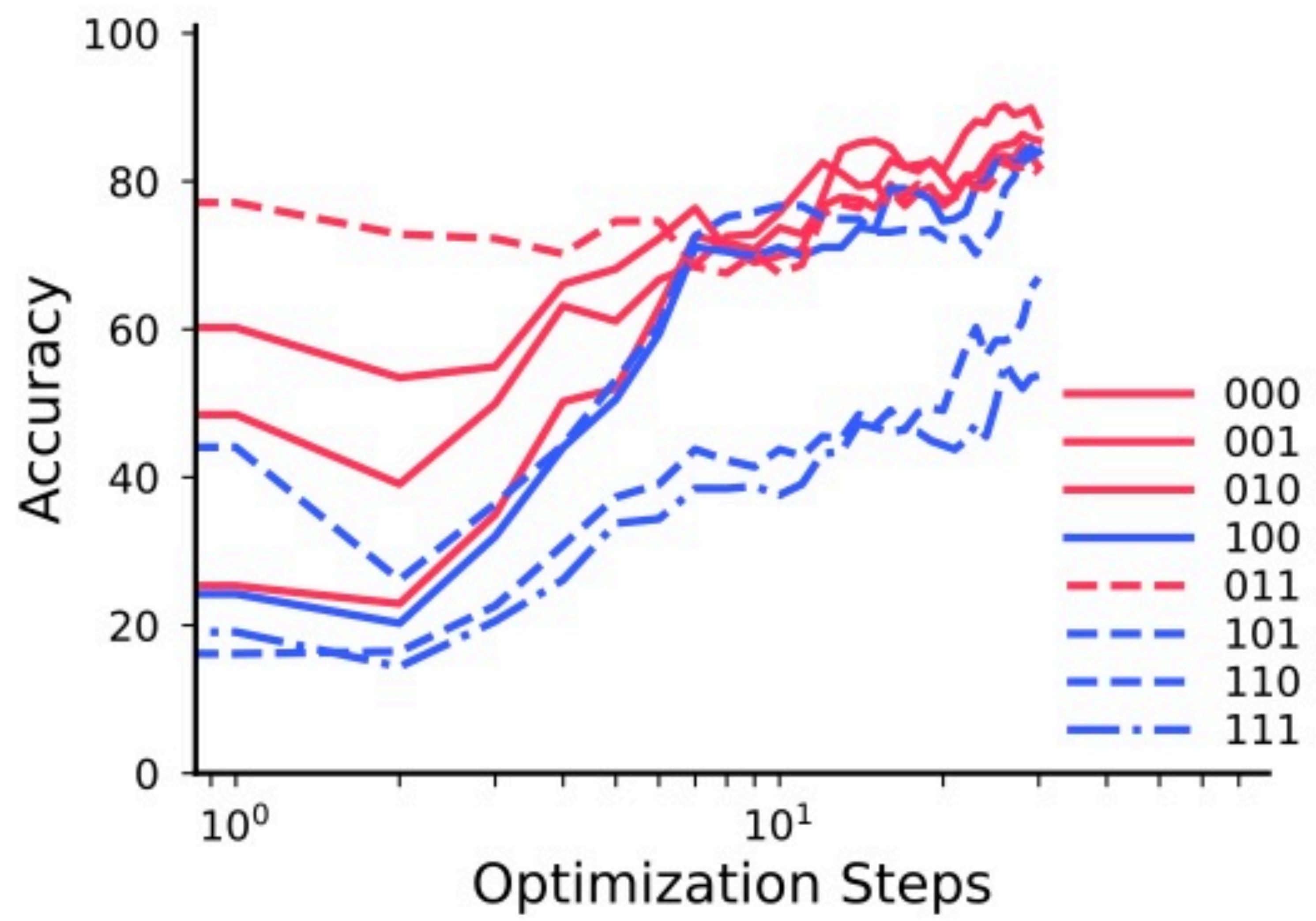
Practical Insights for Trustworthy AI:
Compositional generalization to minority class requires extensive training





Towards Fair AI: Verifying our hypotheses in realistic settings





Generalization to minority gender happens at the end!

What about large language models? Signatures of compositionally

Composition of arithmetic operations

What is the sum of the digits of the square of the cube of 8. Give the answer step-by-step.

Cube of 8:

8^3 = 8 * 8 * 8 = 512

Now, find the square of 512:

512^2 = 512 * 512 = 262,144

To find the sum of the digits of this answer, add up the individual digits:

2 + 6 + 2 + 1 + 4 + 4 = 19

So, the sum of the digits of the square of the cube of 8 is 19.

✓

Composition of content and style in writing

Tell my friend that I am going to the mall. Write it in legalese.

Dear [Friend's Name],

I hereby notify you, in accordance with applicable legal standards, that I shall be departing for the shopping center forthwith.

Sincerely,

[Your Name]

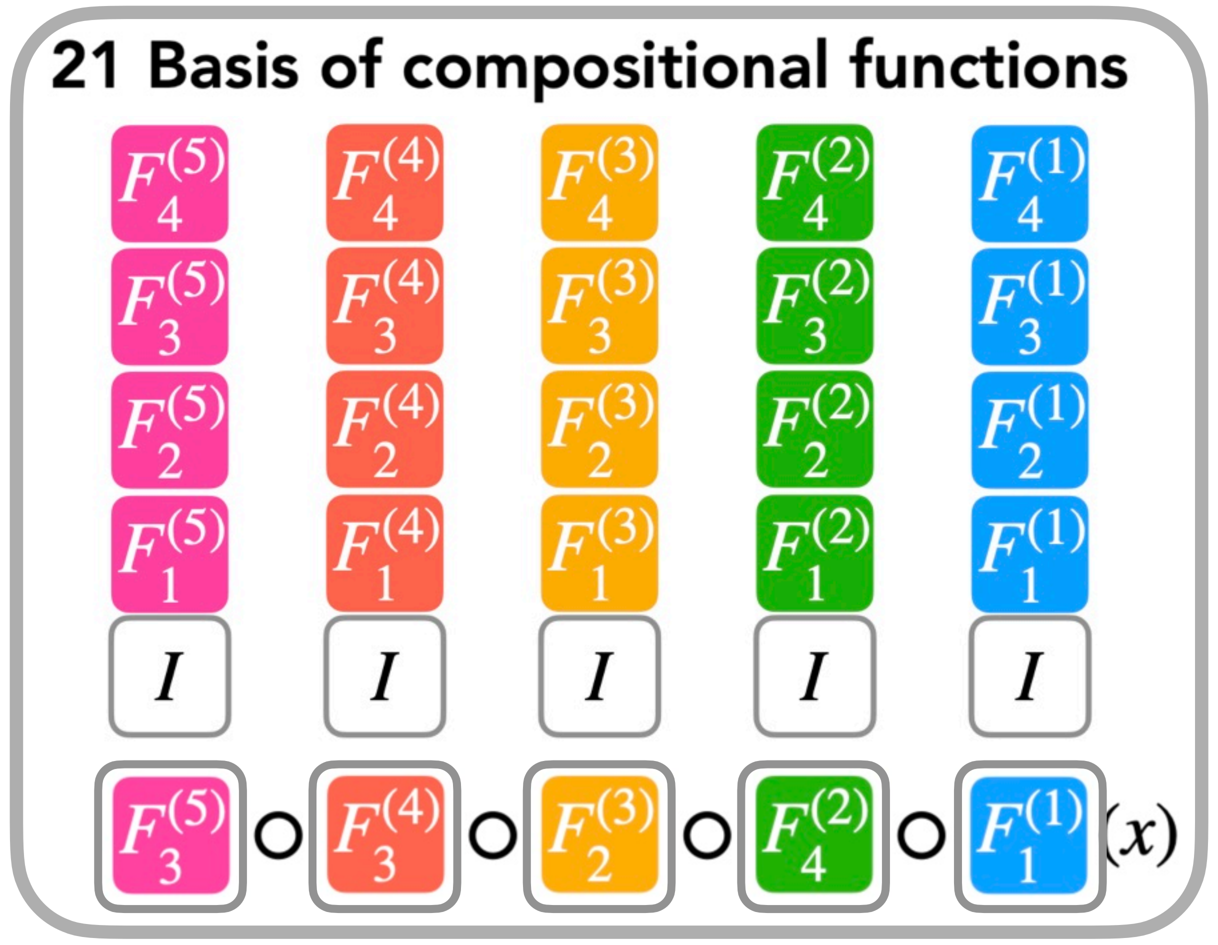
✓

"Compositional Capabilities of Autoregressive Transformers: A Study on Synthetic, Interpretable Tasks"

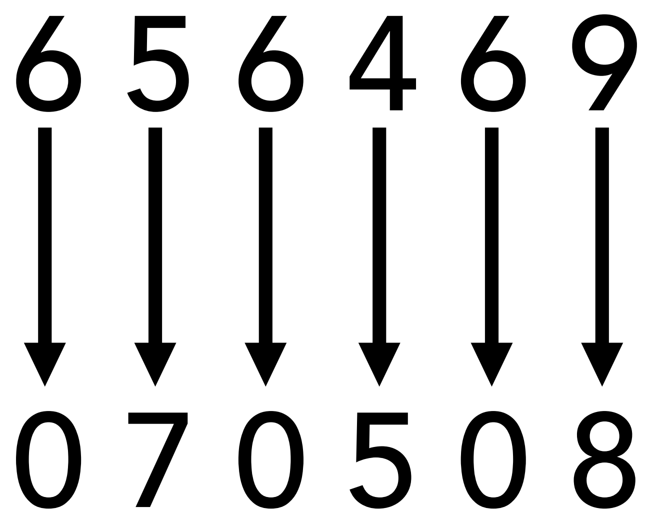
R. Ramesh, E.S. Lubana, M. Khona, R.P. Dick, and <u>H. Tanaka</u>

Compositional Task on Sequential Data

Task: Function Composition



e.g.) bijection: $F_1^{(1)}$



Prompt Structure:

S $F_1^{(1)} F_4^{(2)} F_2^{(3)} F_3^{(4)} F_3^{(5)}$ 656469 070508 … 121416 979490

Task tokens — Input — Output

Vectorization:

Character "6" ⟶ One-hot vector $x \in \mathrm{R}^{\#\text{ of characters}}$

Sequence prediction task:

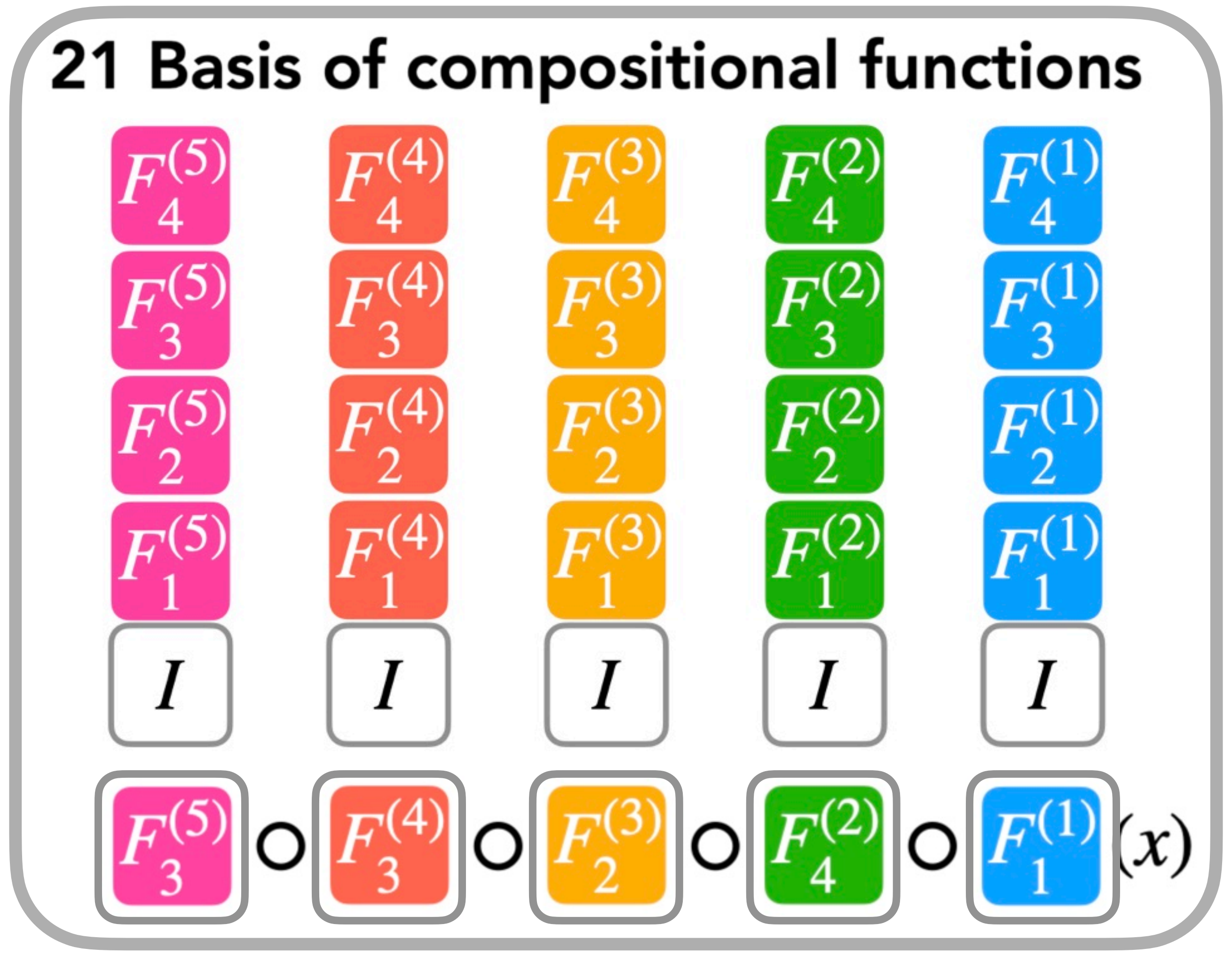
$$\hat{x}_{t+1} = f(x_1, x_2, \ldots, x_t; \Theta)$$

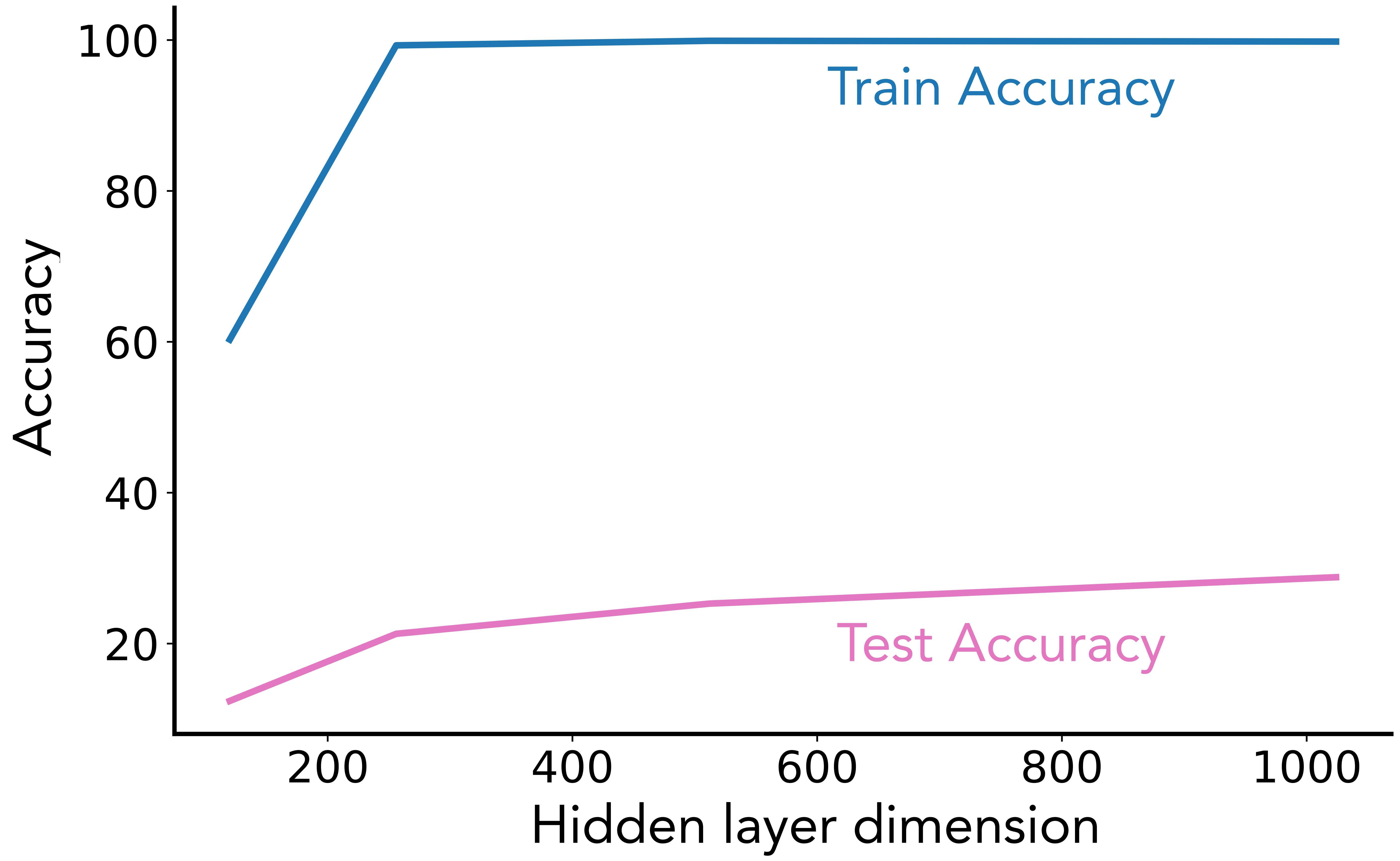
$$\operatorname{argmin}_{\Theta} \left[\sum_t - x_{t+1} \cdot \log \hat{x}_{t+1} \right]$$

Correct answer (x_{t+1}) — Prediction ($\hat{x}_{t+1}$)

LSTM (RNN) fails to compositionally generalize

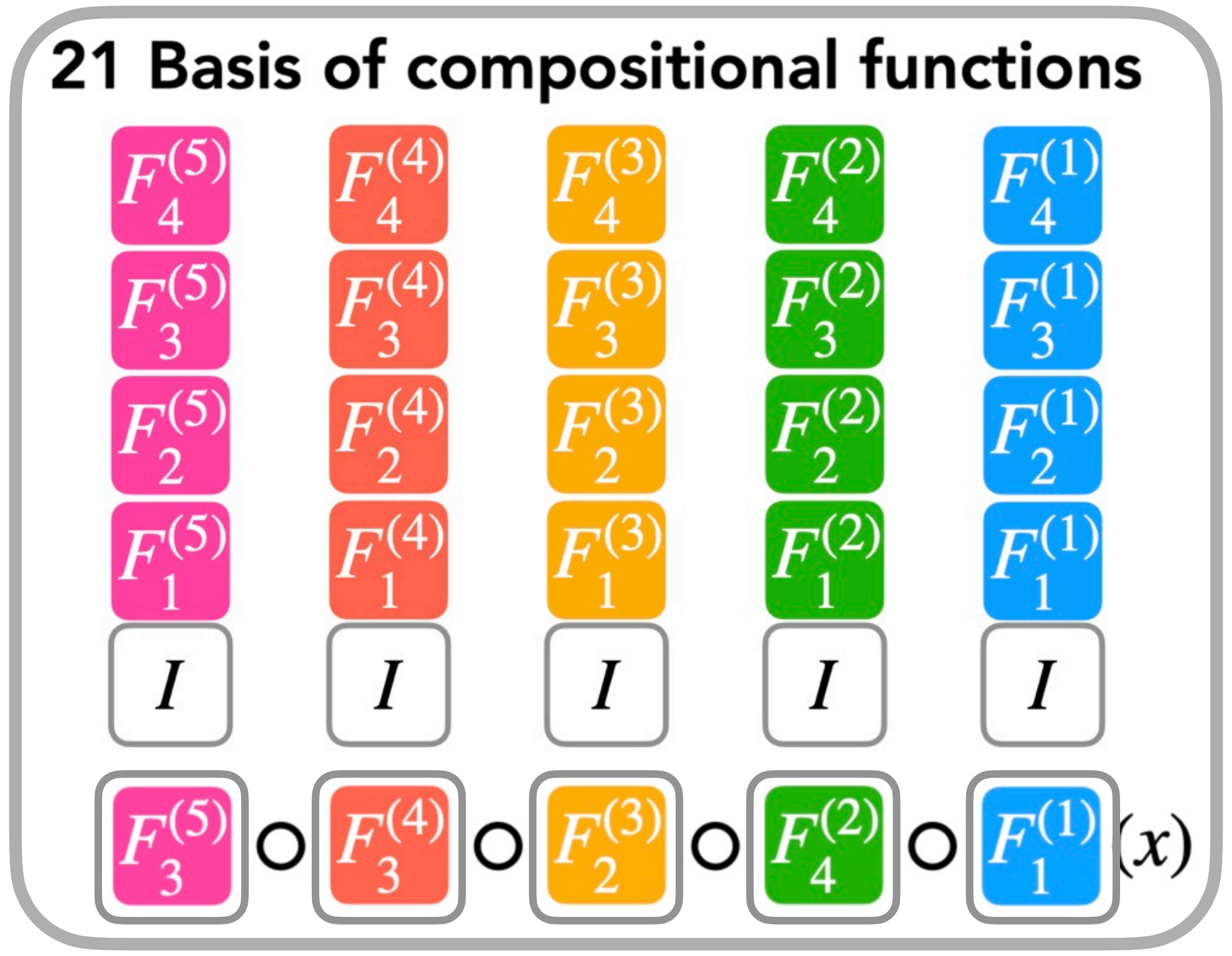
Train a model on 50 random compositions of 5 functions.
Test it on all (5^5 = 3125) compositions.

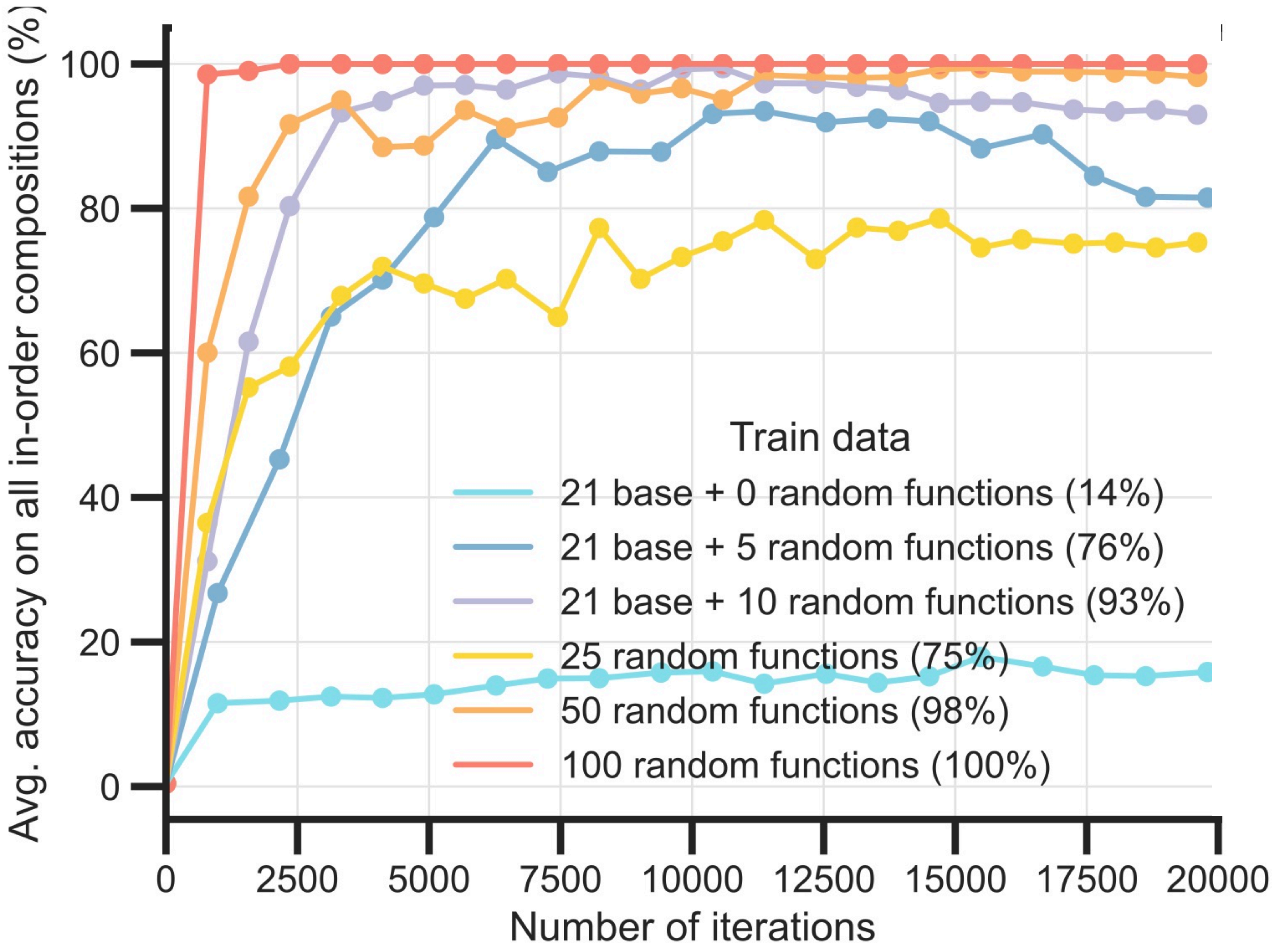




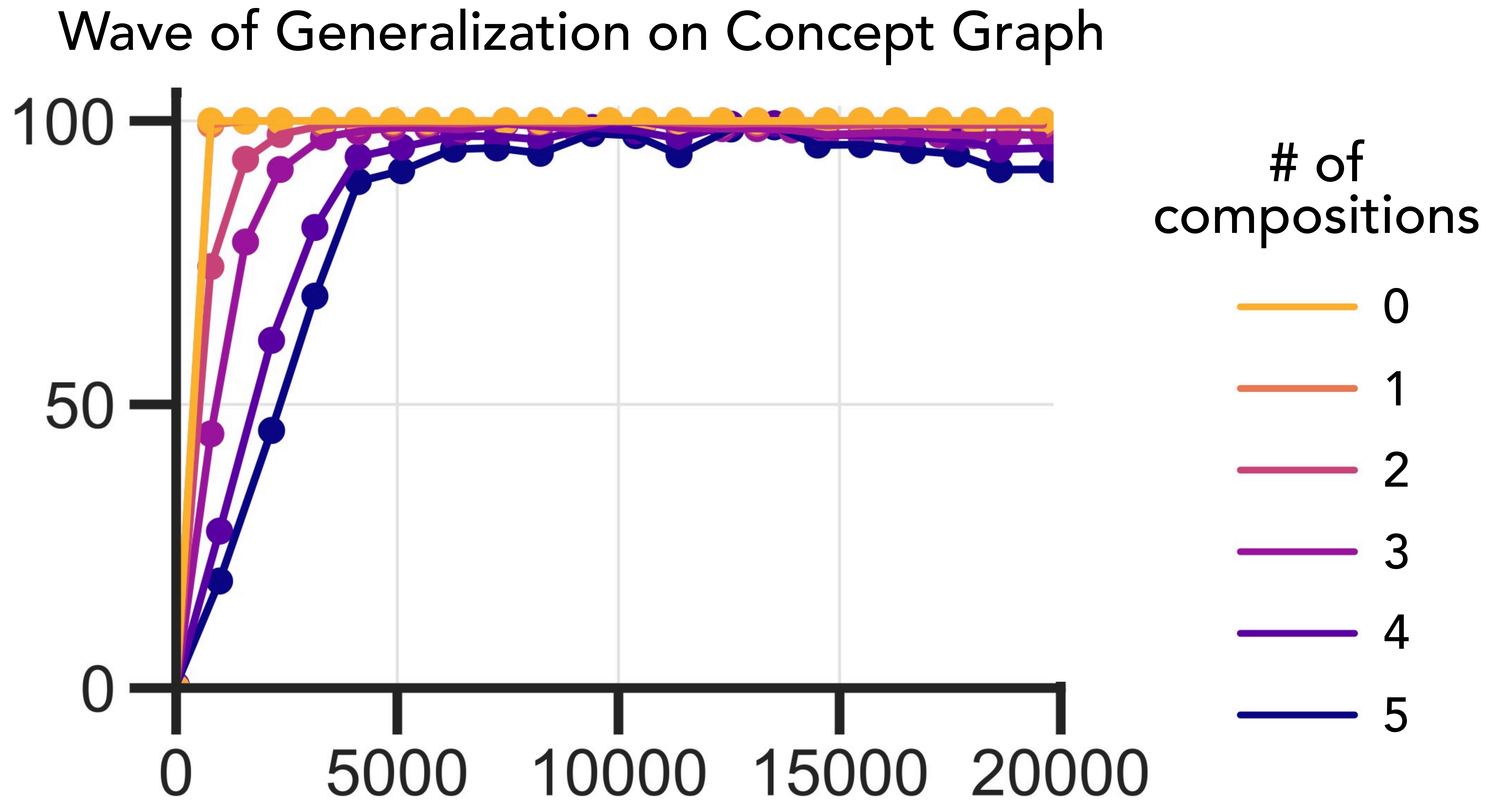
Transformers compositionally generalize successfully!

Generalizing to 3125 (=5^5) compositional functions by just seeing 50~100 examples!

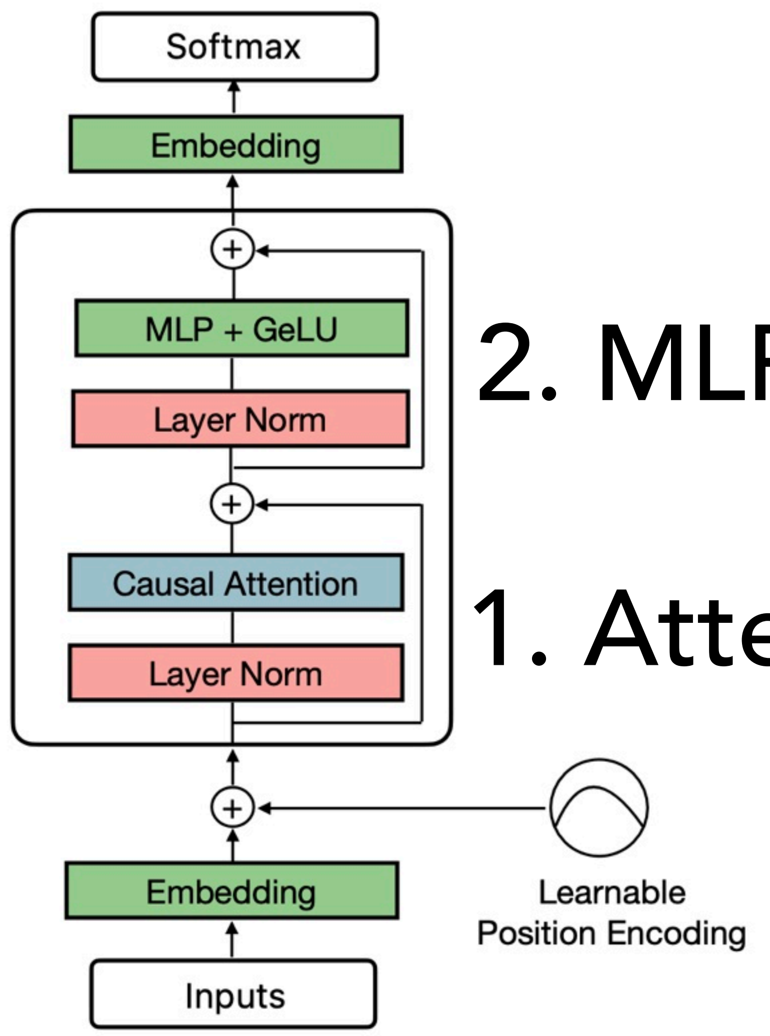




Compositional structure in the "task" induces "universal" learning dynamics!



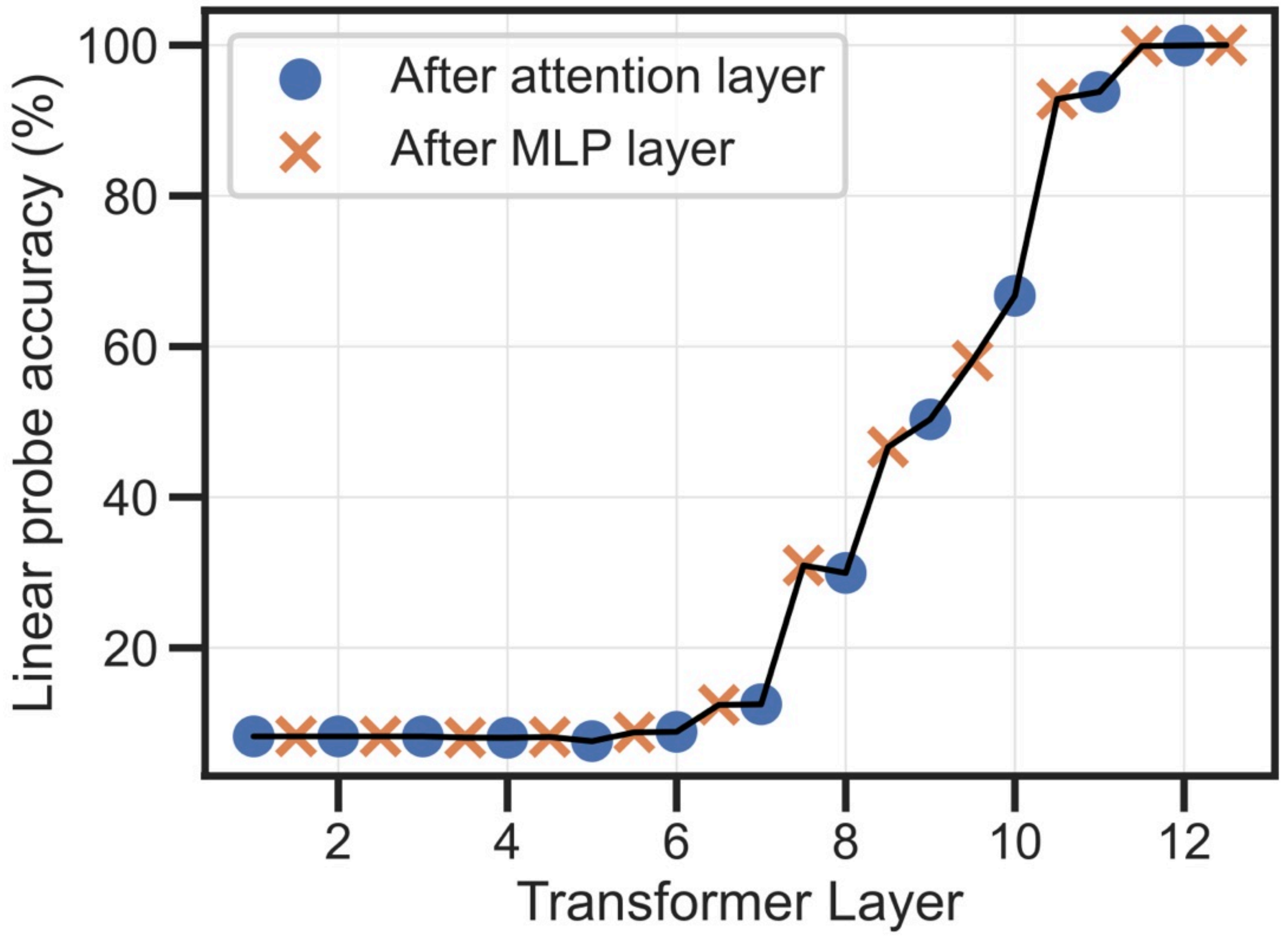
Attention Mechanism Enables Compositional Generalization



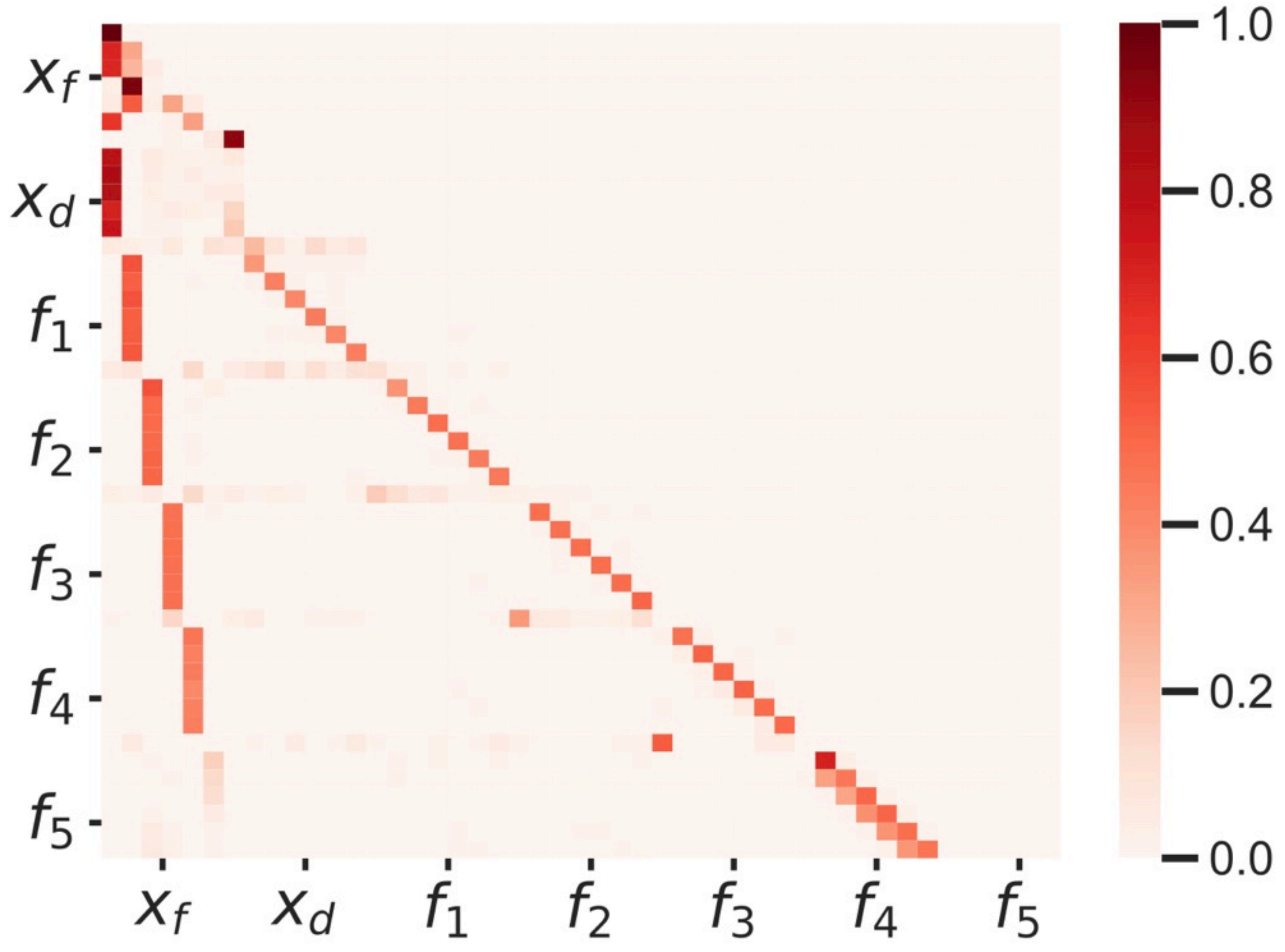
2. MLP: Applies the Selected Function

1. Attention: Picks a Function to Apply

Linear probe accuracy jumps after MLP layers



Attention focuses on task and current tokens



A neural computational mechanism for compositional generalization!

Future Direction:
Towards Neural Principles for Concept Learning and Generalization

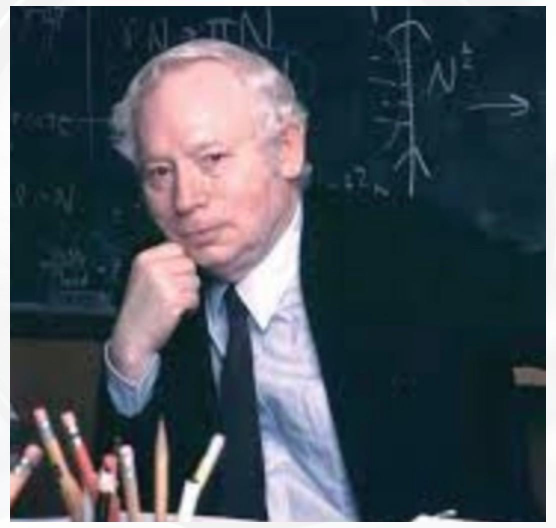
Objects' Geometry | Numbers | Mechanics | Optics

What are good mathematical models of the data & task?

How do the laws of the physical world shape the dynamics of neural learning and computation?

What neural network principles enable compositional generalization?

"Particle physics really was a mess in the 1960s.
Go for the messes — that's where the action is."
by Steven Weinberg



Computing and Learning as Physical Processes

1. Can generative AI (diffusion models) imagine? If so, how?

"Compositional Abilities Emerge Multiplicatively: Exploring Diffusion Models on a Synthetic Task"
NeurIPS 2023
M. Okawa*, E.S. Lubana*, R.P. Dick, ***H. Tanaka****

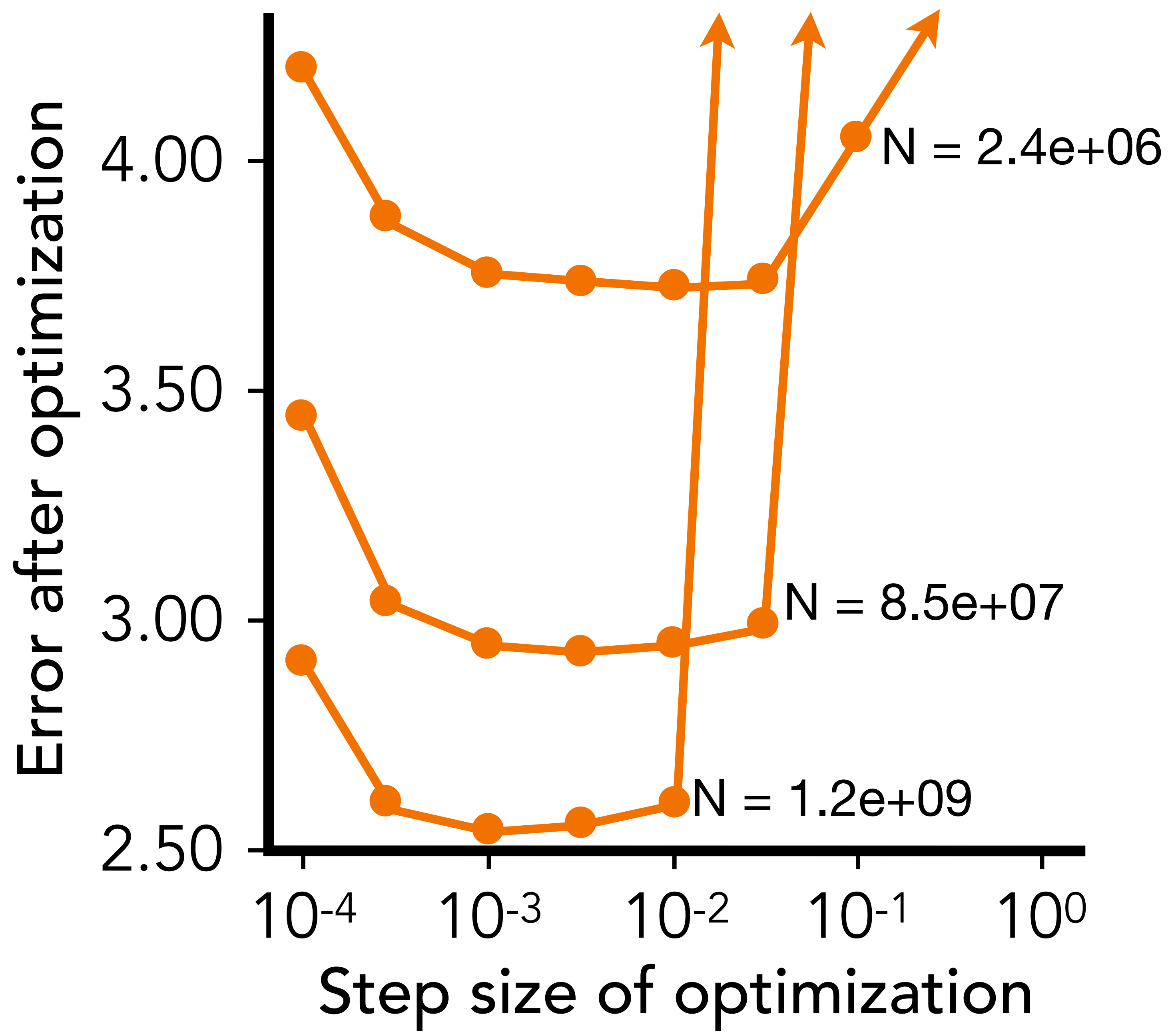
2. Learning as physical dynamics:

"Noether's Learning Dynamics: Role of Symmetry Breaking in Neural Networks" *NeurIPS 2021*
H. Tanaka, D. Kunin

"Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics" *ICLR 2021*
D. Kunin*, J. Sagastuy, S. Ganguli, D.L.K Yamins, H. Tanaka*



Large neural networks are extremely fragile to choices we make at initialization



M. Wortsman et al. Google 2023

A single failure can cost $~millions!

e.g., LLaMA/ChatGPT-3: ~100billion (10^9) parameters
trained on ~1trillion (10^{12}) words
Each training run of modern AI costs $2~3 million!

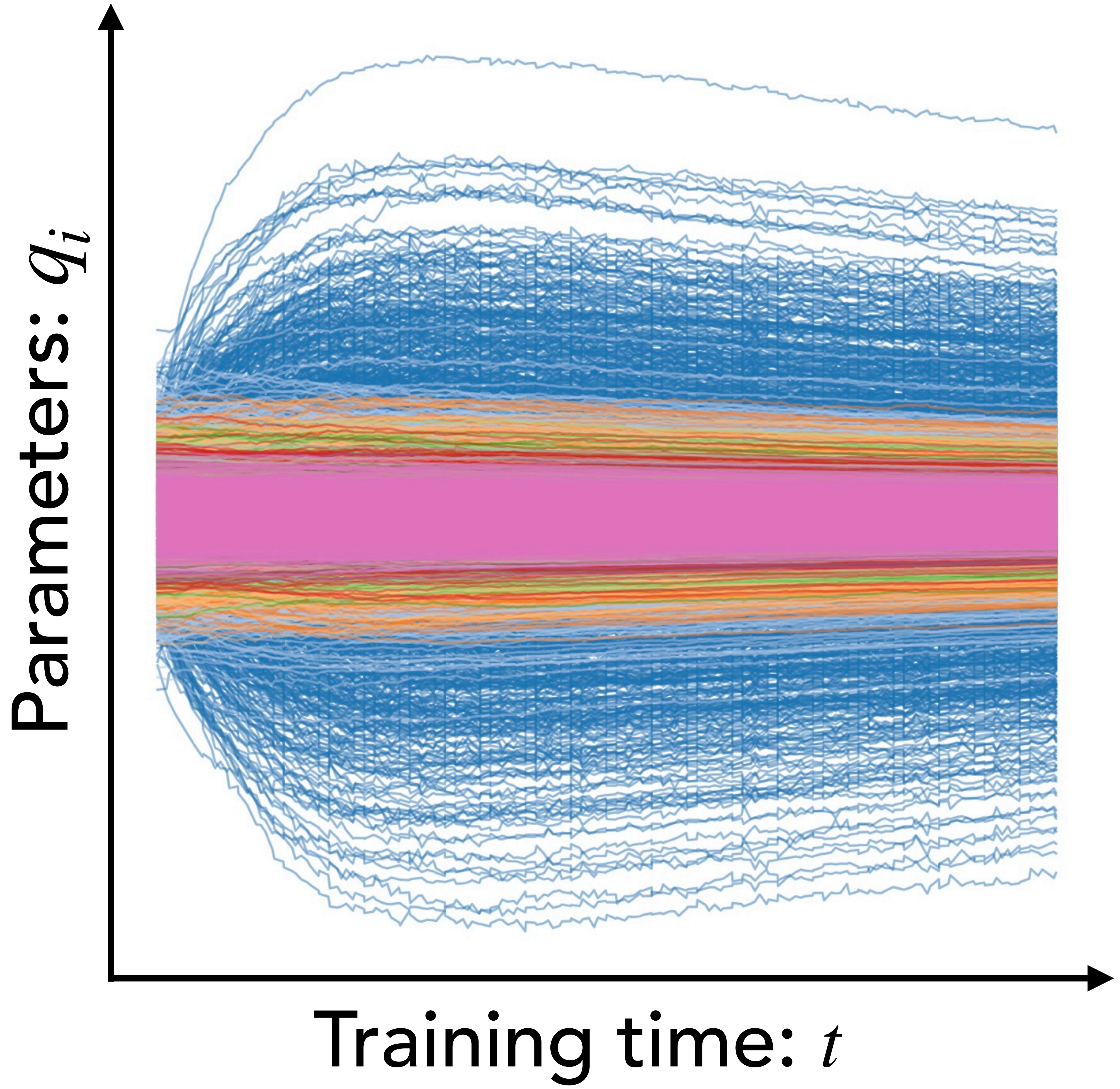
Q. What are the laws that govern complex deep learning dynamics?

Symmetry

"Deep": Architectures

ReLU
BatchNorm
Layer Norm
GroupNorm
WeightNorm
SoftMax
Convolution
Transformer
Residual connection
...

VGG16 trained on Tiny ImageNet



Lagrangian

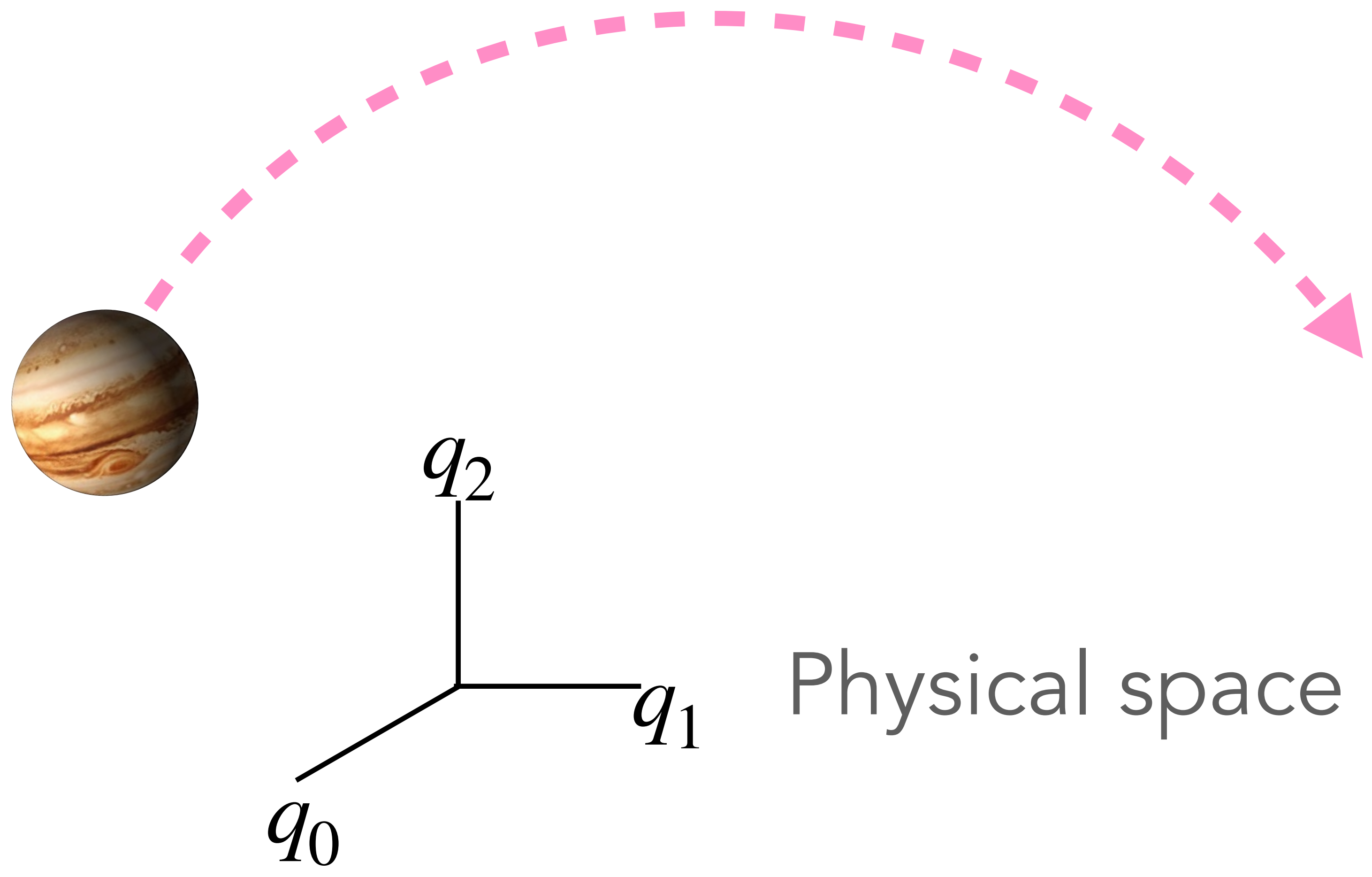
"Learning": Optimizers

Stochastic Gradient Descent
AdaGrad
Adam
RMSProp
AdamW
Heavy-ball momentum
Nesterov momentum
Natural gradients
...

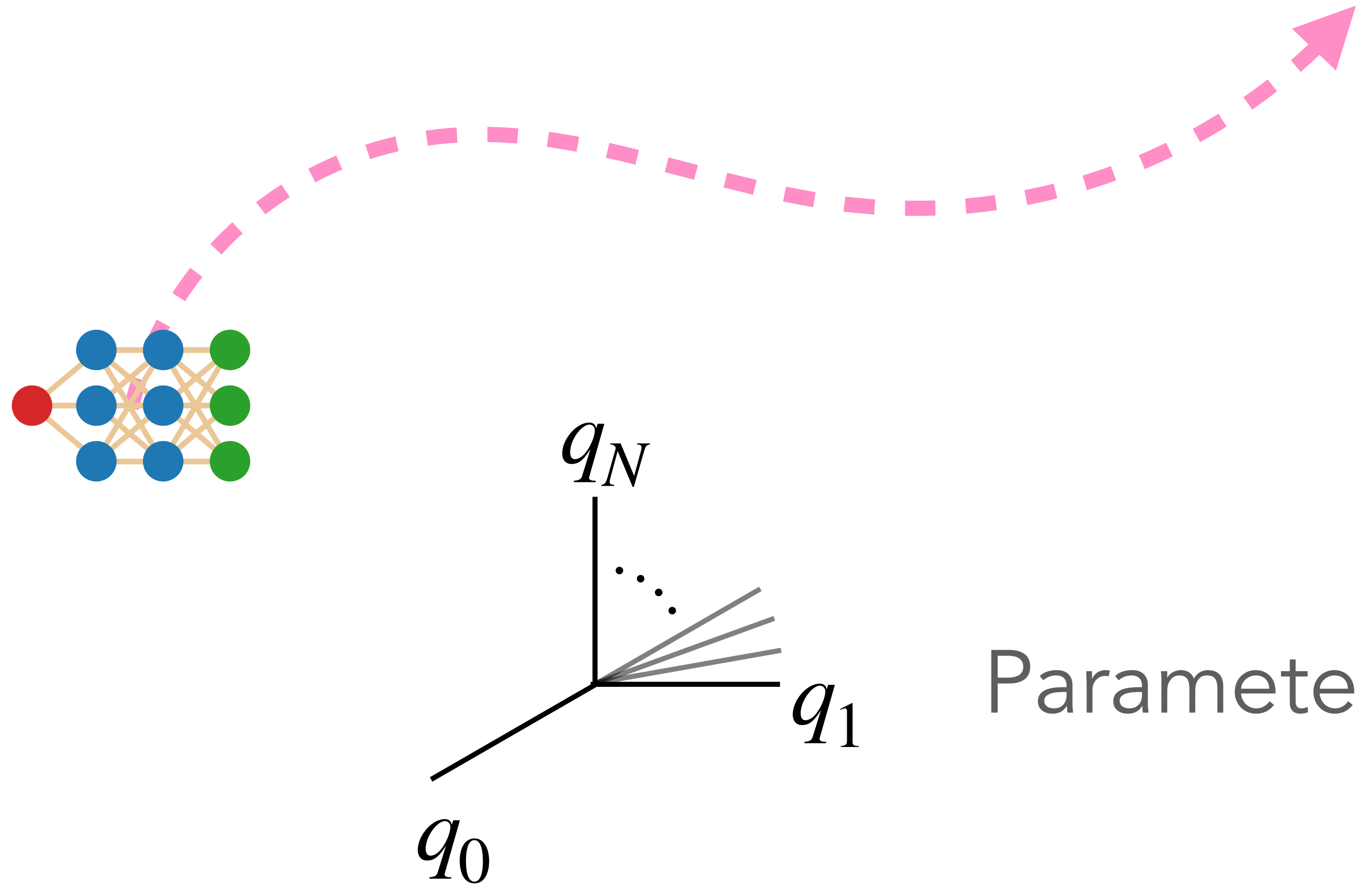
We construct a Lagrangian framework to understand the dynamics of learning!

Neural learning as physical dynamics

Classical Mechanics v.s. Neural Mechanics



Physical space



Parameter space

Forces:
Gravity, Electric/Magnetic, Friction etc…

Equation of motion:
$F(x) = m\partial_t^2 x$

Symmetries in Lagrangian:
Translation in time/space, Rotation

Conservation laws:
Energy, momentum, angular momentum

Forces:
Gradients driven by real world dataset

Equation of learning:
Gradient Descent: $q(t+\eta) = q(t) - \eta \nabla f(q)$

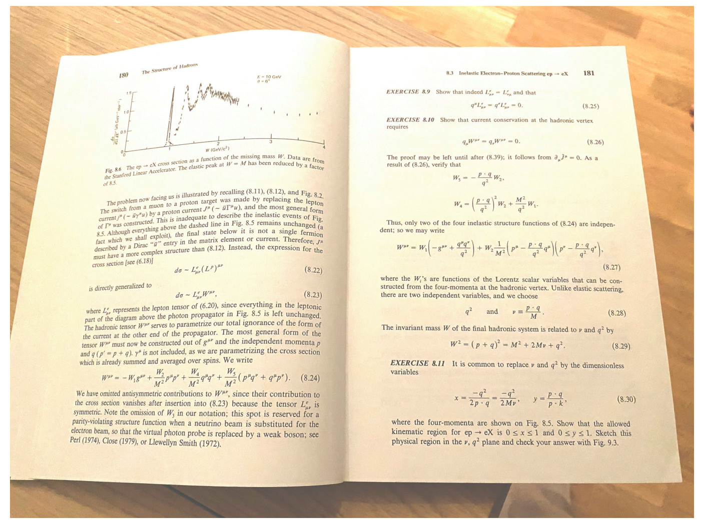
Symmetries in the Loss function:
Translation, Scale, Rescale

(Broken) conservation laws:
Dynamics of parameter combinations

Scale invariance $f(sq) = f(q)$ is one of the most ubiquitous symmetries in neural networks

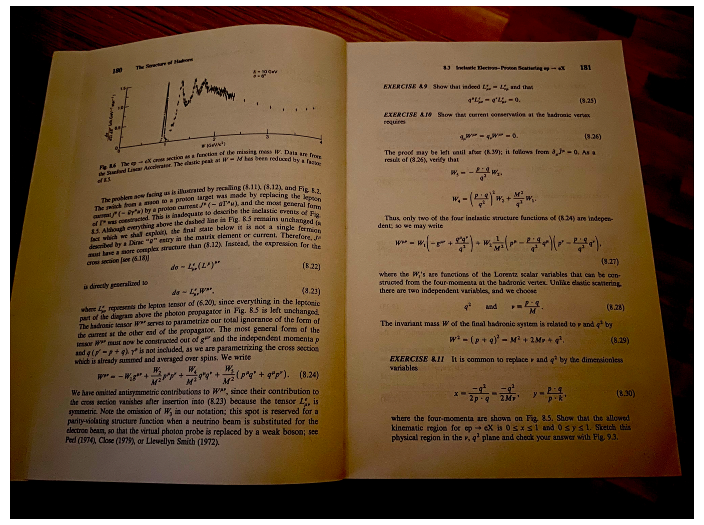
Visual signal

In a bright room



q

In a dark room



$\underset{\text{Intensity}}{\underline{s}} \times q$

Mechanism: Normalizing signal at each step of neural computation

$$\mathrm{Norm}(q) = \frac{q - \mathrm{E}[q]}{\sqrt{\mathrm{Var}[q]}}$$

$$\mathrm{Norm}(sq) = \frac{\cancel{s}q - \mathrm{E}[\cancel{s}q]}{\sqrt{\mathrm{Var}[\cancel{s}q]}} = \mathrm{Norm}(q)$$

How does scale symmetry $f(sq) = f(q)$ affect the "dynamics" of learning?
Let's generalize Noether's theorem for scale symmetry!

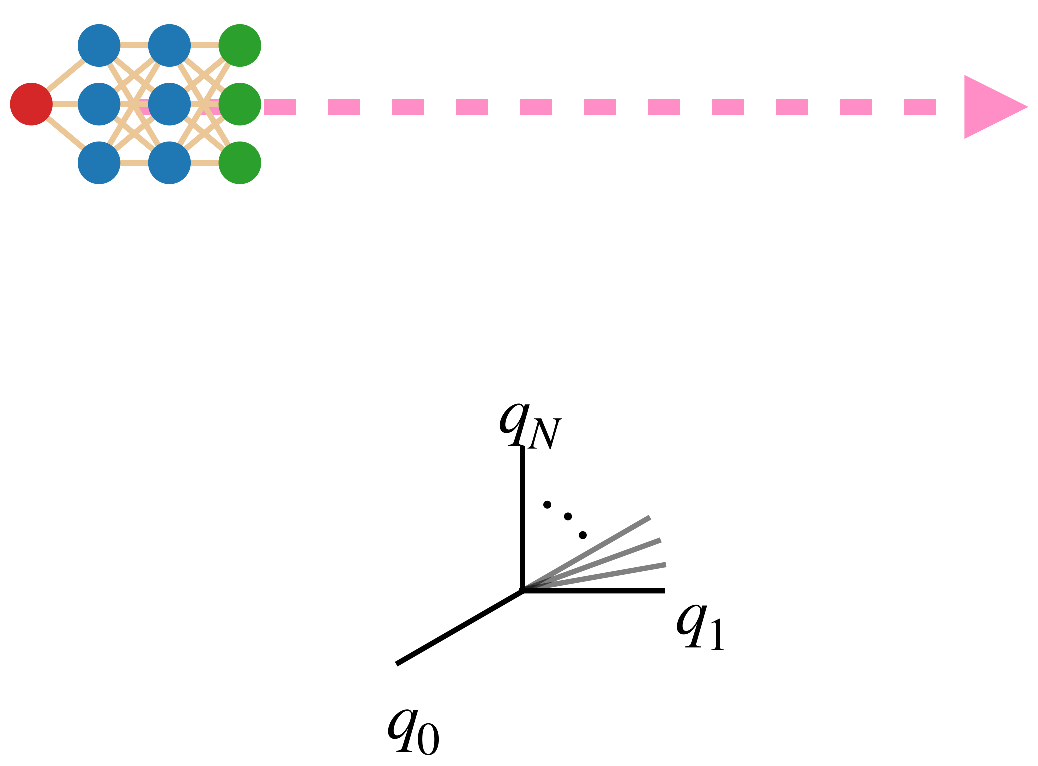
Symmetry unifies neural architectures

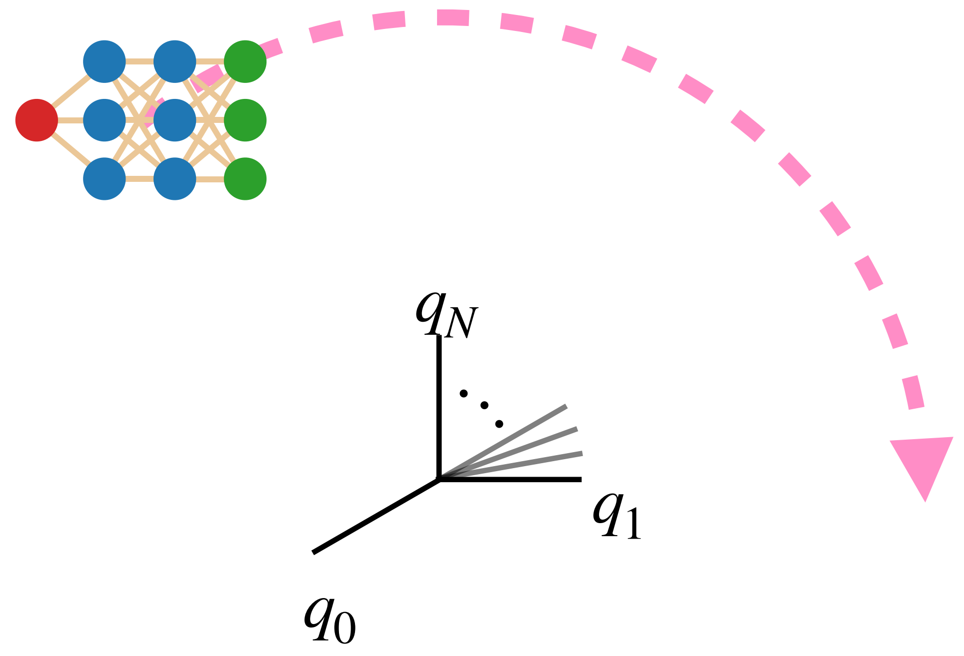
$\mathscr{L} = T - V$

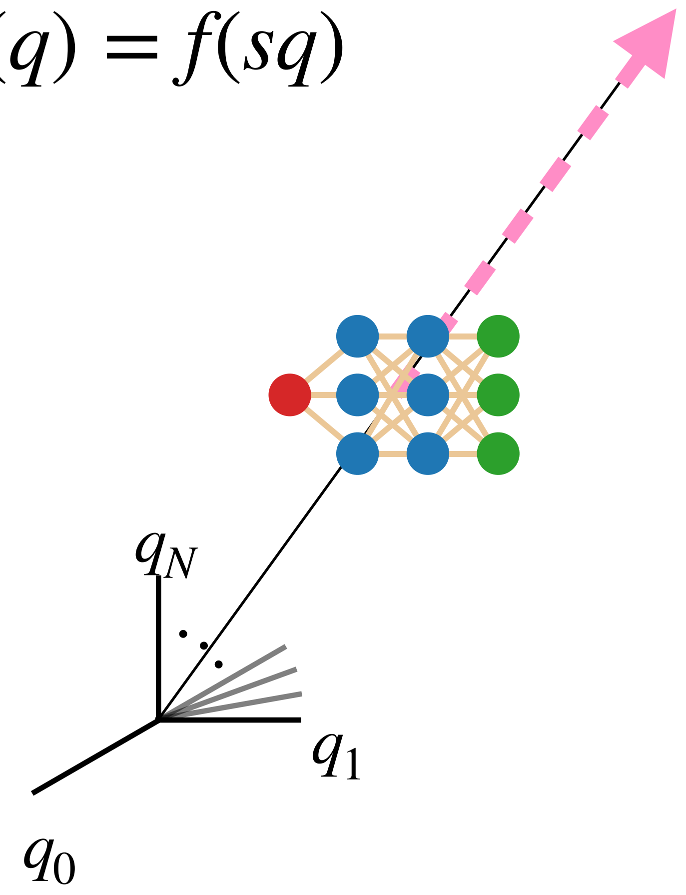
Symmetry: A function $f(q)$ posses a symmetry if it is invariant under the transformation $q \mapsto Q(q, s)$, i.e. if $f(Q(q,s)) = f(q)$ for any (q, s).

Euclidean learning rules: $T \propto |\frac{dq}{dt}|^2$ Loss function: $V \propto f(q)$

	Translation	**Rotation**	**Scale**	
Symmetry	$f(q) = f(q + s\hat{n})$	$f(q) = f\left(R(s)q\right)$	$f(q) = f(sq)$	







Parameter space

Example	Softmax	Self-Supervised Learning	BatchNorm	ReLU* (*Re-scale)
	$\sigma(qx)_i = \frac{e^{q_i x}}{\sum_j e^{q_j x}}$	$L(f^T(x)f(x'))$	$\mathrm{BN}(qx) = \frac{qx - \mathrm{E}[qx]}{\sqrt{\mathrm{Var}(qx)}}$	$q_2\mathrm{ReLU}(q_1 x) = q_2\max(0, q_1 x)$

Lagrangian unifies learning rules

Modeling discrete learning dynamics in continuous time

Gradient descent:

$$q_{n+1} = q_n - \eta \nabla f(q)$$

Forward Euler discretization:

$$\frac{1}{\eta}\left(q(t+\eta) - q(t)\right) = \frac{1}{\eta}\left(q(t) + \eta\frac{dq}{dt} + \frac{\eta^2}{2}\frac{d^2q}{dt^2} - q(t)\right) = -\nabla f(q)$$

Newton's equation of motion:

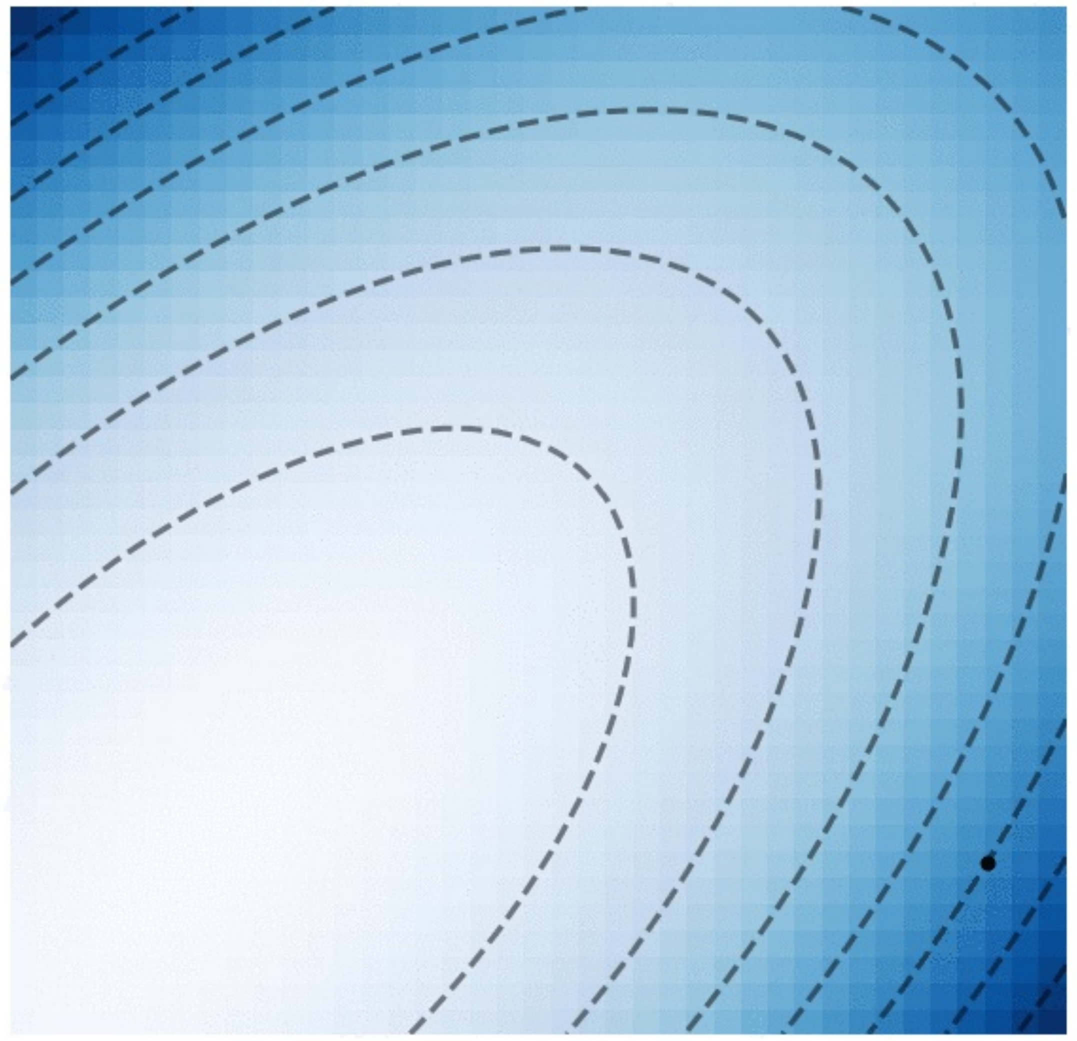
$$m\frac{dq^2}{dt^2} = -\nabla f(q)$$

Gradient flow:

$$\frac{dq}{dt} = -\nabla f(q)$$

Modified gradient flow:

$$\frac{\eta}{2}\frac{dq^2}{dt^2} + \frac{dq}{dt} = -\nabla f(q)$$



Blue curve: gradient flow
Red curve: modified trajectory
Black dots: discrete SGD steps

Lagrangian of modern practical optimizer with finite step size η,

Damping

$$\mathcal{L}(q, \dot{q}, t) = e^{\frac{2}{\eta}t}\left[\frac{\eta}{4}|\dot{q}|^2 - f(q)\right]$$

Kinetic energy (*T*) ⇔ Learning rules

Potential energy (*V*) ⇔ Loss function

(S)GD becomes Lagrangian dynamics in practical settings with a finite learning rate

Lagrangian unifies learning rules

Modeling discrete learning dynamics in continuous time

Gradient descent:

$$q_{n+1} = q_n - \eta \nabla f(q)$$

Forward Euler discretization:

$$\frac{1}{\eta}\left(q(t+\eta) - q(t)\right) = -\nabla f(q)$$

Newton's equation of motion:

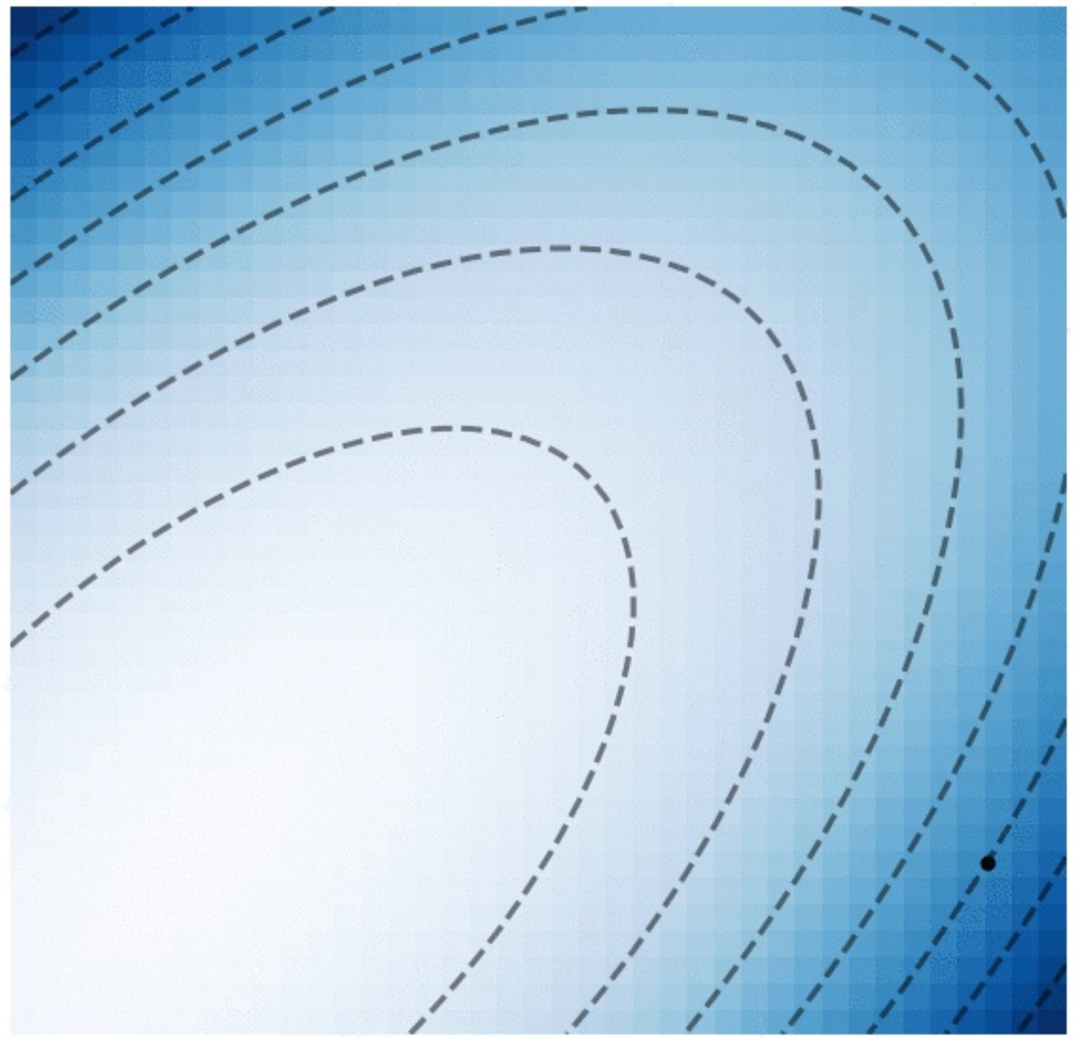
$$m\frac{dq^2}{dt^2} = -\nabla f(q)$$

Gradient flow:

$$\frac{dq}{dt} = -\nabla f(q)$$

Modified gradient flow:

$$\frac{\eta}{2}\frac{dq^2}{dt^2} + \frac{dq}{dt} = -\nabla f(q)$$



Blue curve: gradient flow
Red curve: modified trajectory
Black dots: discrete SGD steps

Lagrangian of modern practical optimizer with finite step size η, momentum β, and weight decay k.

Damping

$$\mathcal{L}(q, \dot{q}, t) = e^{\frac{2(1-\beta)}{\eta(1+\beta)}t}\left[\frac{\eta(1+\beta)}{4}|\dot{q}|^2 - \left(f(q) + \frac{k}{2}|q|^2\right)\right]$$

Kinetic energy (T) ⇔ Learning rules

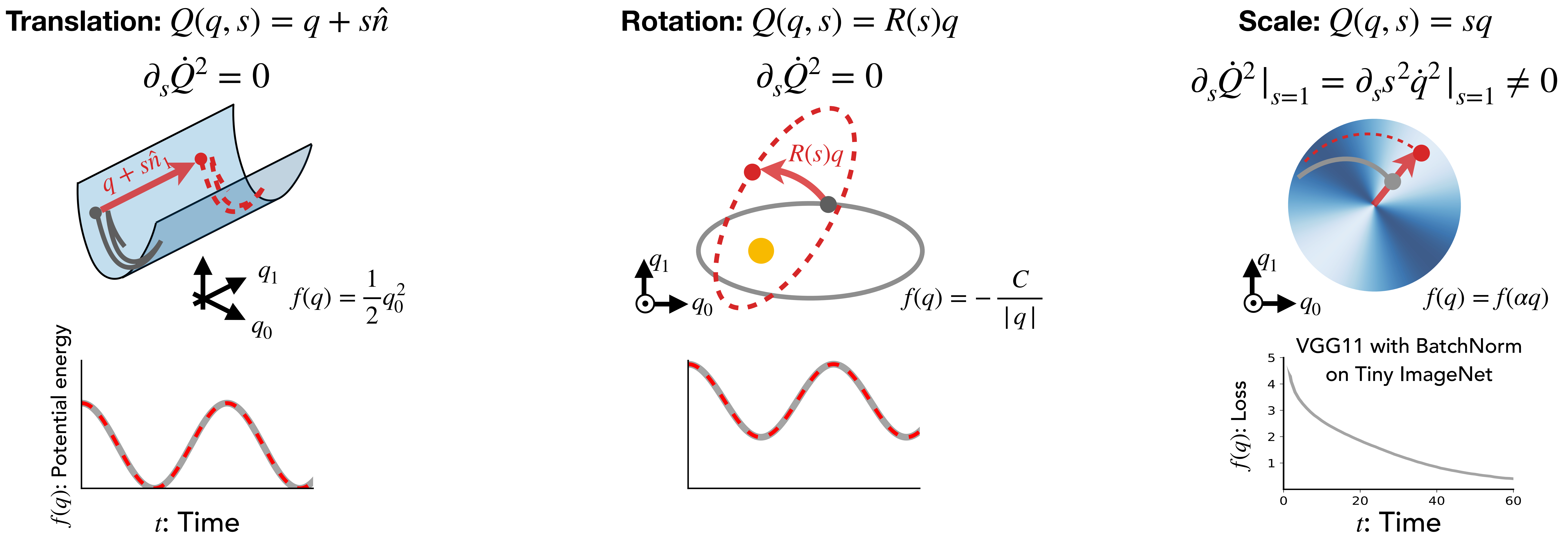
Potential energy (V) ⇔ Loss function

SGD becomes Lagrangian dynamics in practical settings with a finite learning rate

Kinetic energy of learning T breaks the symmetry in deep learning

$$\mathscr{L} = T - V$$

Euclidean learning rules: $T \propto |\frac{dq}{dt}|^2$ Loss function: $V \propto f(q)$



Kinetic asymmetry: The kinetic energy does not observe the same symmetry as the potential function unique to learning systems.

Noether's learning dynamics

Noether's learning dynamics:

Noether charge — damping — kinetic asymmetry — non-Euclidean metric

$$\frac{d}{dt}\overbrace{\langle \Delta_h, \partial_s Q\rangle}^{\text{Noether charge}} + \overbrace{\eta_t\langle \Delta_h, \partial_s Q\rangle}^{\text{damping}} = \overbrace{\langle \Delta_h, \partial_s \dot{Q}\rangle}^{\text{kinetic asymmetry}} + \overbrace{e^{\alpha_t}\langle \Delta_h - e^{-\alpha_t}\nabla^2 h(q)\dot{q}, \partial_s Q\rangle}^{\text{non-Euclidean metric}}$$

$$\Delta_h(q, \dot{q}, \alpha_t) \equiv \nabla h(q + e^{-\alpha_t}\dot{q}) - \nabla h(q)\,.$$

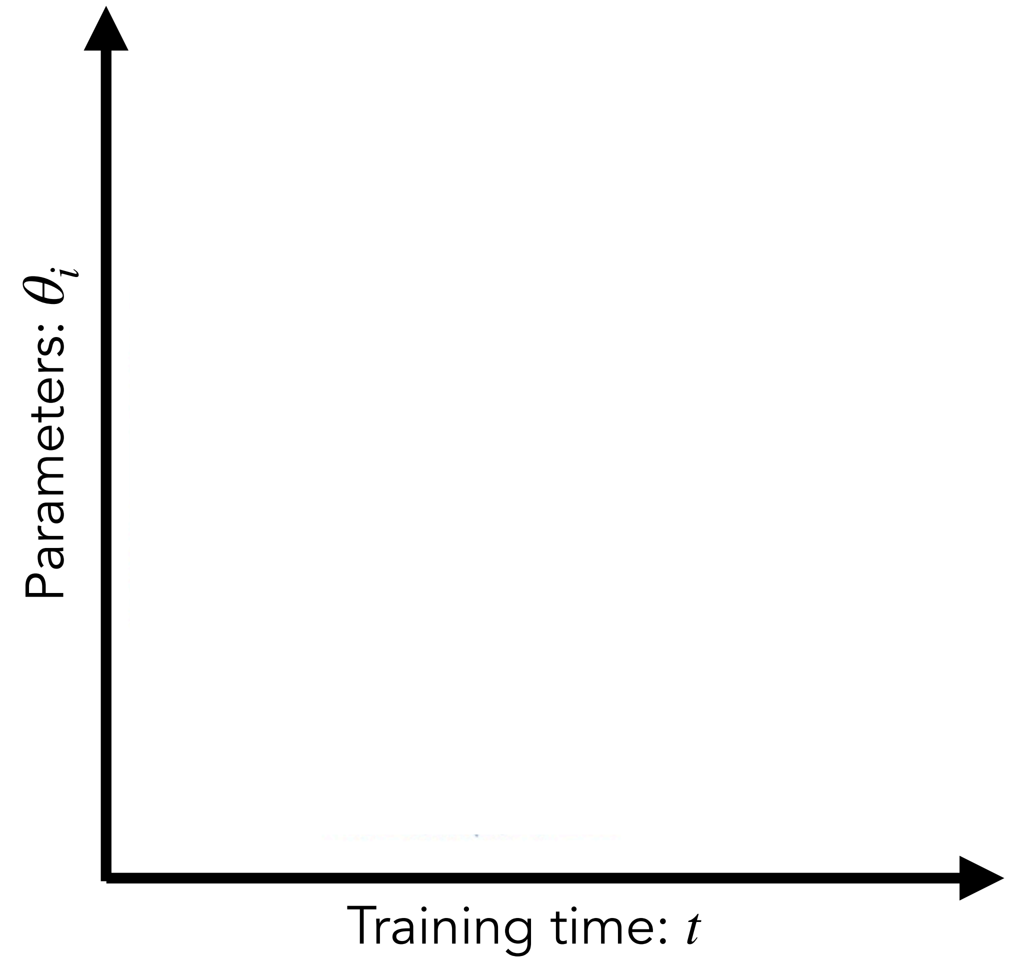
Noether charge for scale symmetry:

$$\langle \Delta_h, \partial_s Q\rangle \propto = \frac{1}{2}\frac{d}{dt}|q|^2$$

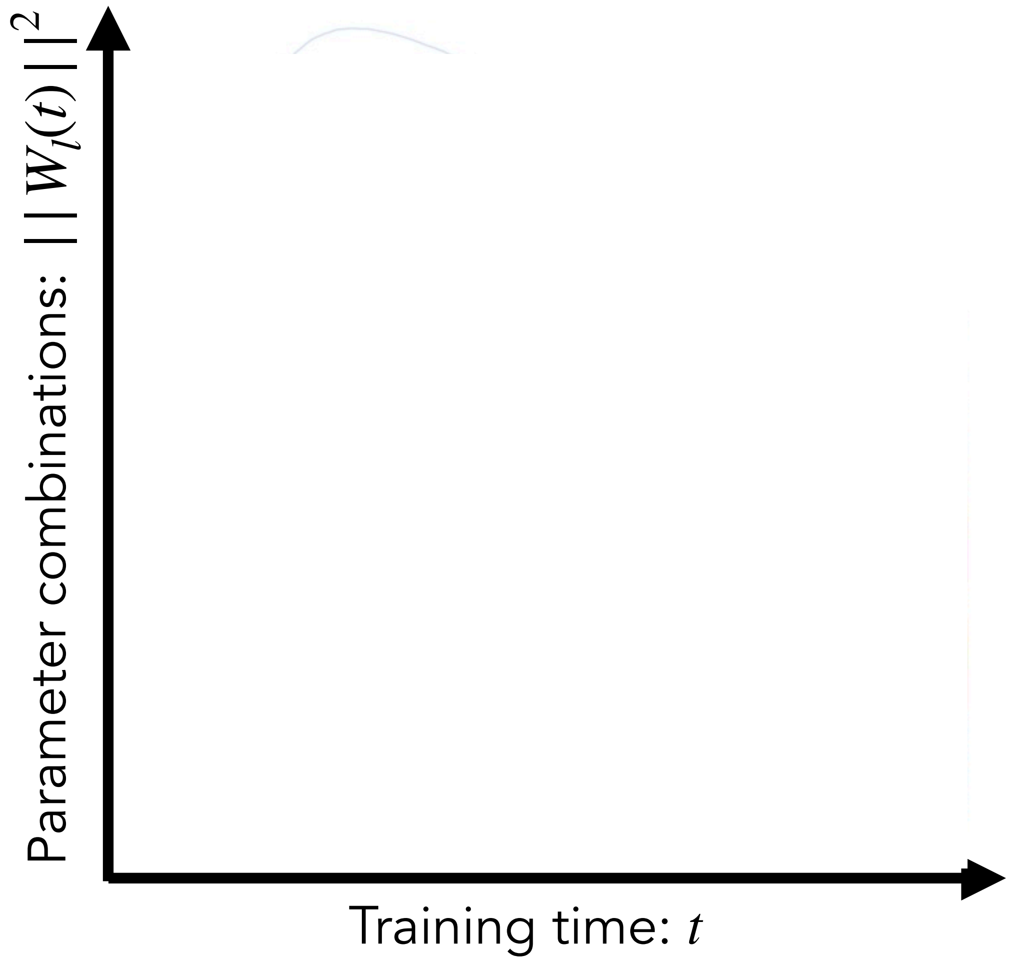
VGG16

- conv. 1
- conv. 2
- conv. 3
- conv. 4
- conv. 5
- conv. 6
- conv. 7
- conv. 8
- conv. 9
- conv. 10
- conv. 11
- conv. 12

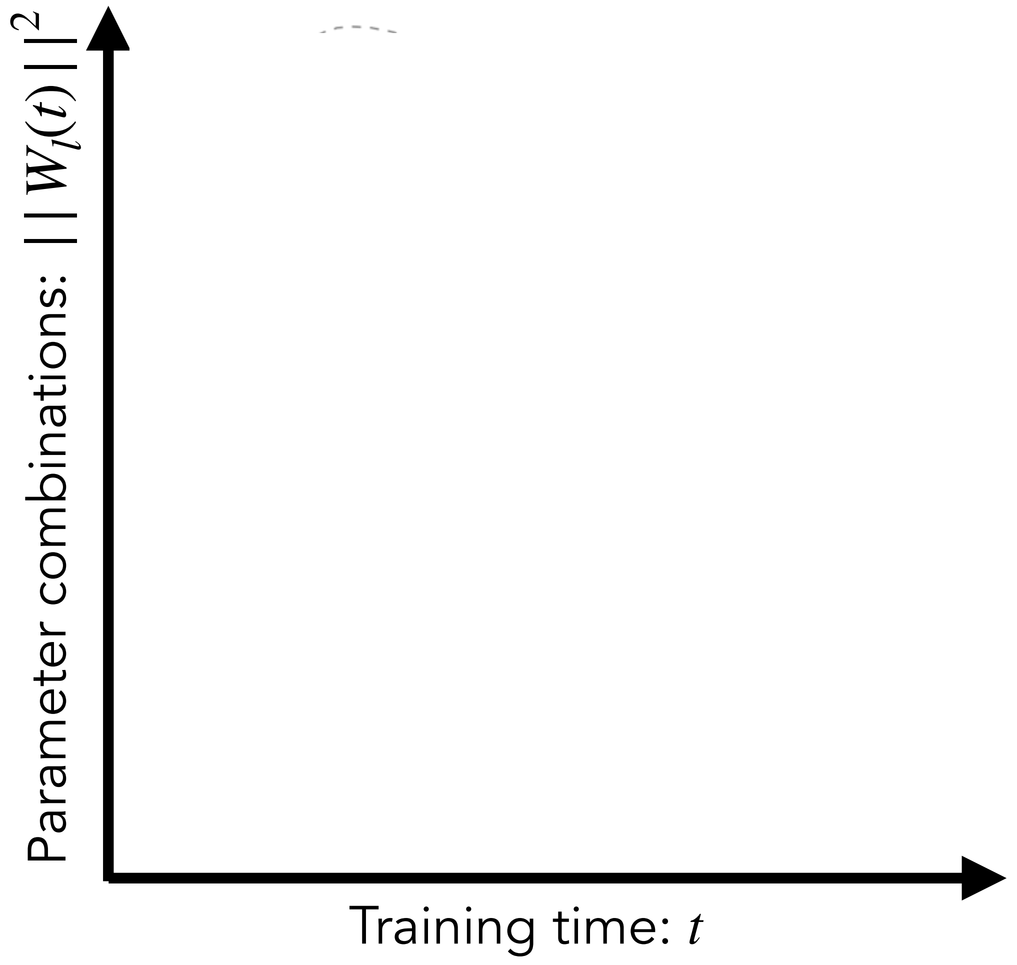
parameter dynamics



combination dynamics



theory



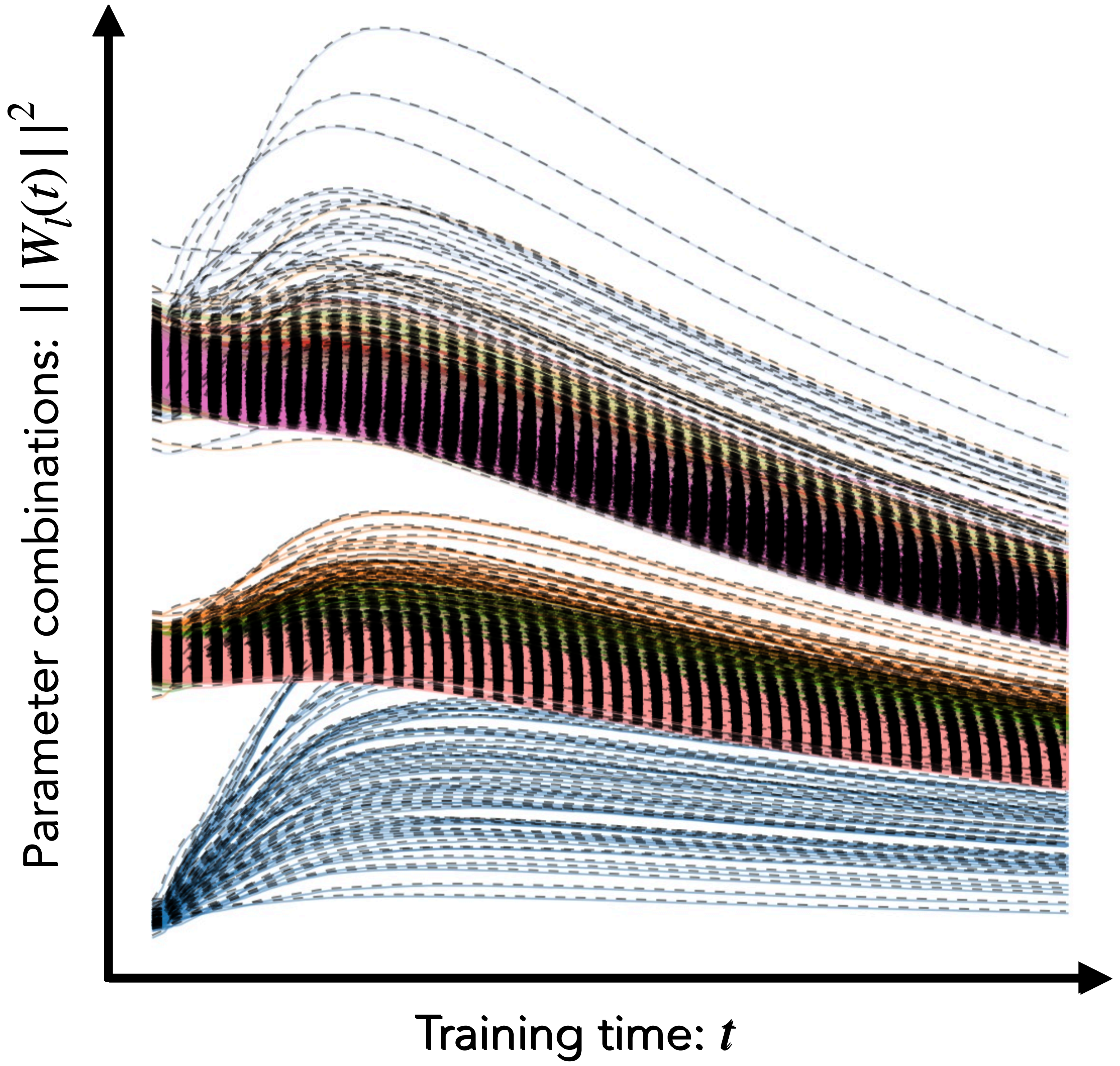
VGG 16 trained on Tiny-ImageNet

Validating the Noether's learning dynamics (Scale symmetry)

Noether's learning dynamics:

$$\frac{d}{dt}\overbrace{\langle \Delta_h, \partial_s Q\rangle}^{\text{Noether charge}} + \overbrace{\dot{\gamma}_t\langle \Delta_h, \partial_s Q\rangle}^{\text{dissipation}} = \overbrace{\langle \Delta_h, \partial_s \dot{Q}\rangle}^{\text{dynamic asymmetry}} + \overbrace{e^{\alpha_t}\langle \Delta_h - e^{-\alpha_t}\nabla^2 h(q)\dot{q}, \partial_s Q\rangle}^{\text{non-Euclidean metric}}$$

$$\Delta_h(q, \dot{q}, \alpha_t) \equiv \nabla h(q + e^{-\alpha_t}\dot{q}) - \nabla h(q)\,.$$



Our theory matches experiment exactly!

Noether's Learning Dynamics offers practical insights and algorithms!

1. Demystifying the role of normalization layers in deep learning

"Machine learning has become alchemy! Batch Normalization works amazingly well. But we know almost nothing about it." by Ali Rahimi 2017

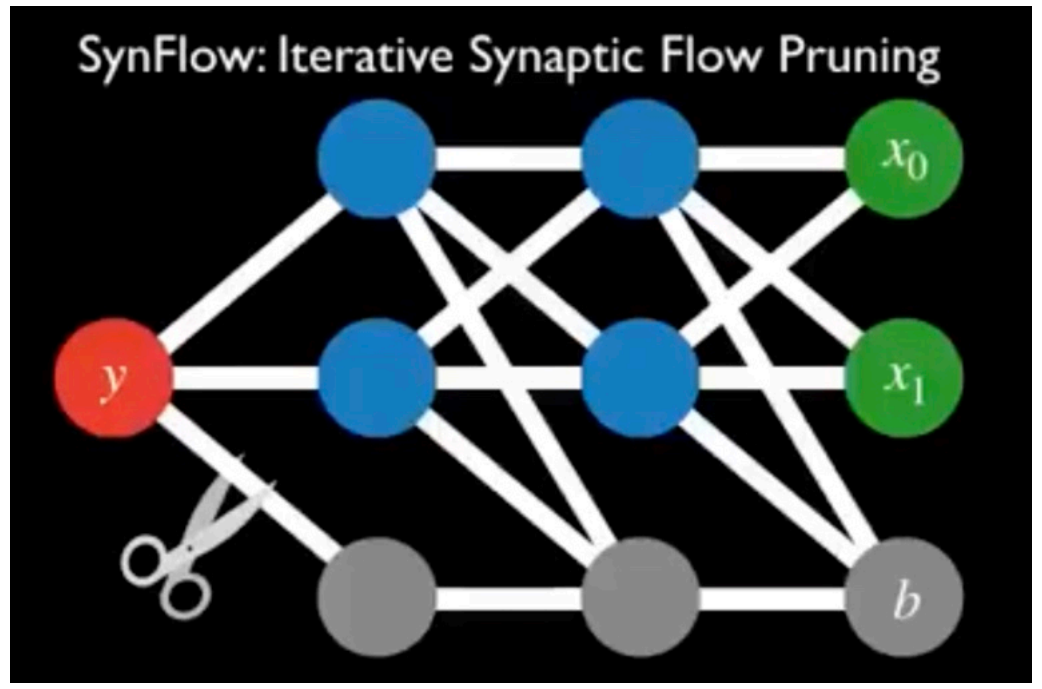
"Normalization ~ Adaptive Optimization"

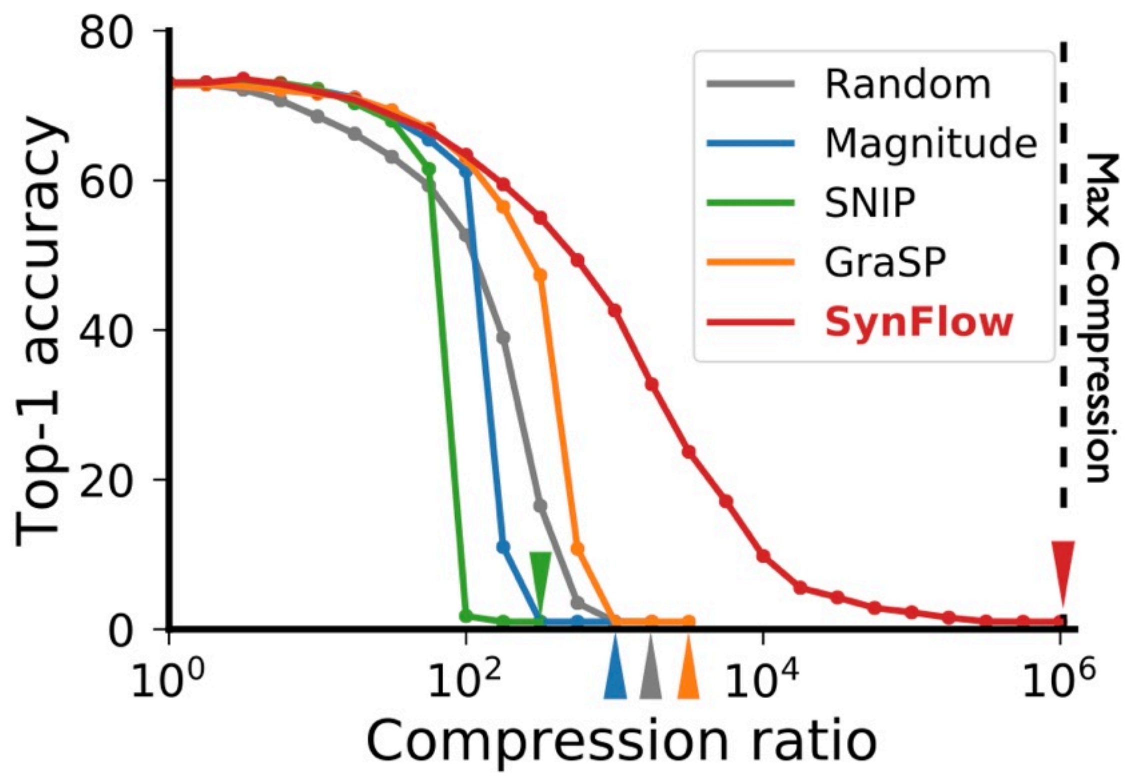
(Architecture) (Learning rule)

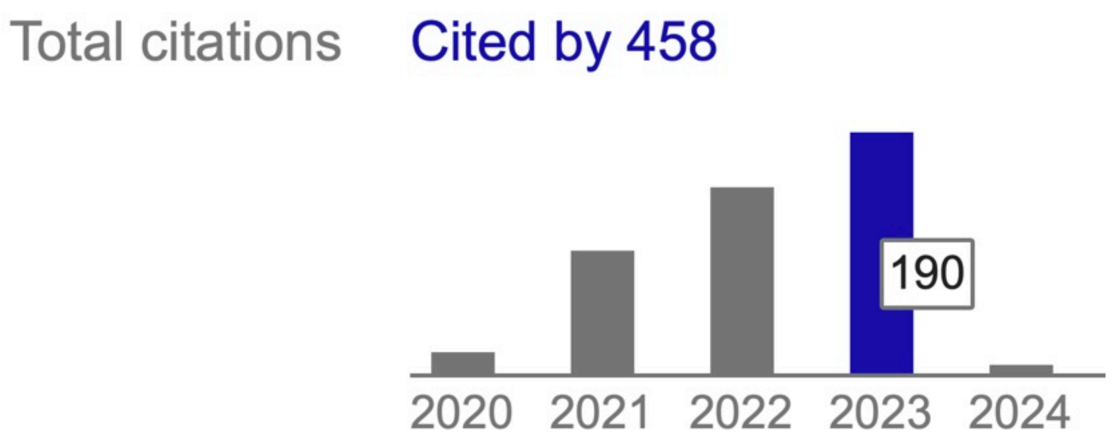
$$|q(t)|^2 = \sqrt{\frac{2\eta(1+\beta)}{(1-\beta)^3}\int_0^t e^{-\frac{4k}{1-\beta}(t-\tau)}|\hat{g}(\tau)|^2 d\tau + e^{-\frac{4k}{1-\beta}t}|q(0)|^4}$$

$$\sqrt{G(t)} = \sqrt{\frac{1-\rho}{\eta}\int_0^t e^{-\frac{1-\rho}{\eta}(t-\tau)}|g(\tau)|^2 d\tau + e^{-\frac{1-\rho}{\eta}t}G(0)}$$

2. Using scale symmetry to compress networks by ~100x for energy efficiency







"Physics of AI" inspired algorithm with practical impact!

Pruning neural networks without any data by iteratively conserving synaptic flow
H. Tanaka*, D. Kunin*, D. Yamins, S. Ganguli (**NeurIPS 2020**)

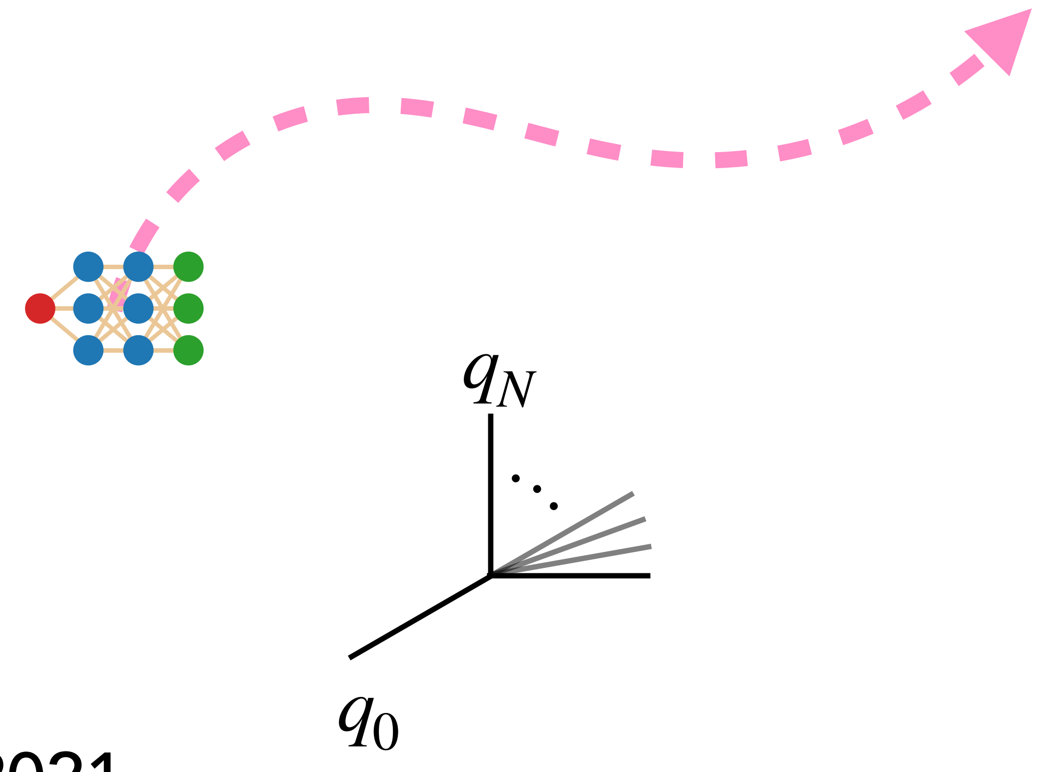
Summary

- **Task Induced Universality:** Symmetry of the data and task shapes symmetries in artificial and biological neural networks.

- **Lagrangian Formulation of Learning:** Symmetry unifies the architectures and kinetic energy unifies learning rules.

- **Generalizing Physics:** Noether's learning dynamics accounts for kinetic asymmetry, dissipation, and non-Euclidean geometry inherent in learning.

- **Practical Insights:** Demystifying the alchemy of normalization layer. "Normalization ~ Adaptive Optimization"



"Noether's Learning Dynamics: Role of Symmetry Breaking in Neural Networks" NeurIPS 2021
H. Tanaka, D. Kunin

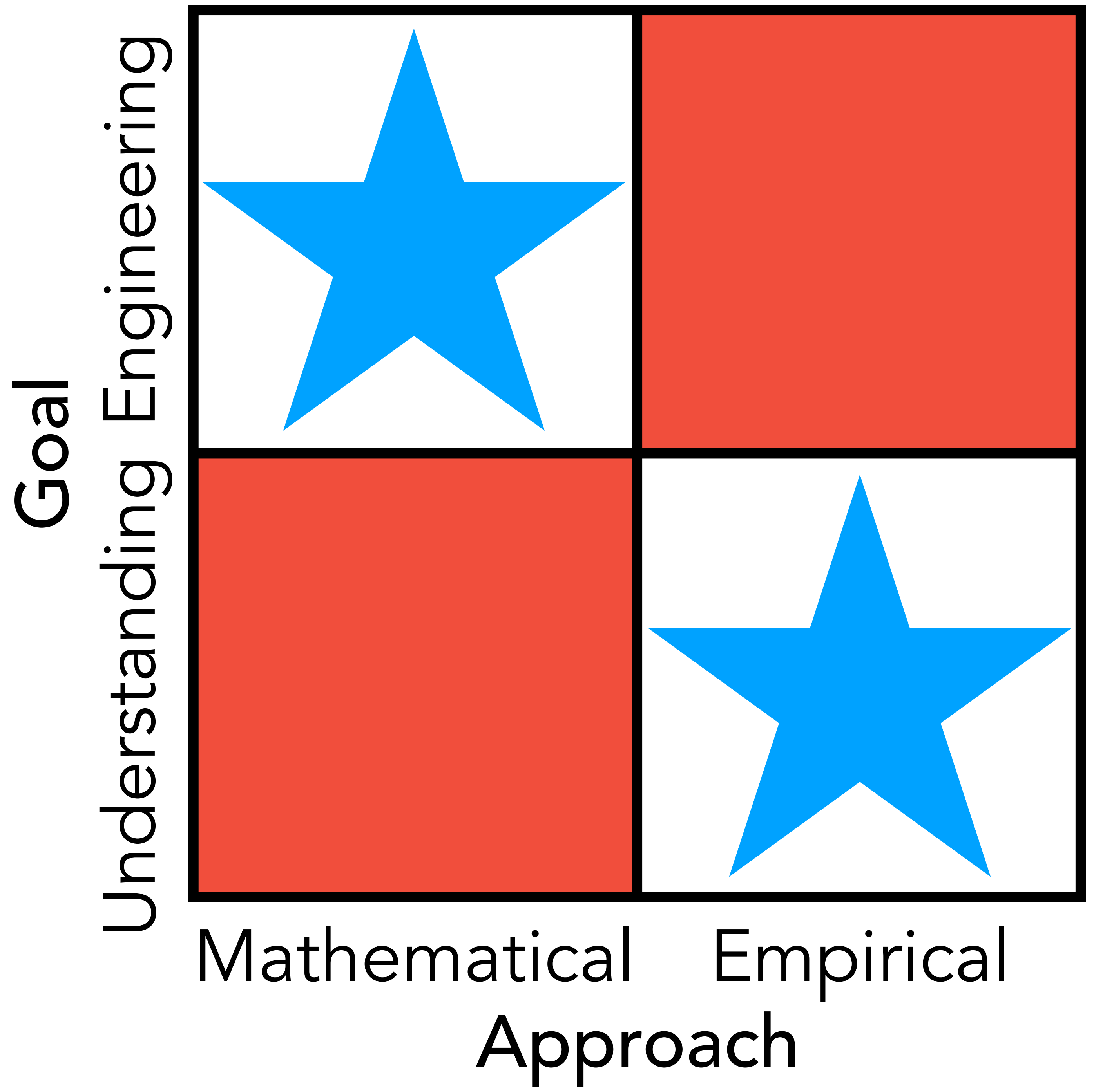
"Neural Mechanics: Symmetry and Broken Conservation Laws in Deep Learning Dynamics" ICLR 2021
D. Kunin*, J. Sagastuy, S. Ganguli, D.L.K Yamins, H. Tanaka*

Conclusion: Shared evolutionary pressures (task) and experiences (data) within the physical world → Shared neural mechanisms for learning and computation!

Compositional structure of concepts/functions
→ Universal wave of generalization in concept graph

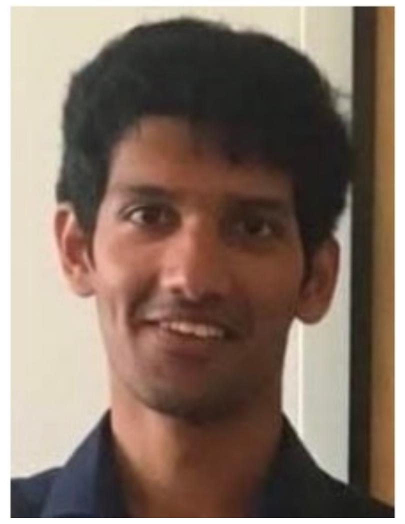
Scale symmetry in the task
→ Generalized Noether's theorem for learning dynamics

Blue oceans in AI Research



Thank you!

Research Scientists



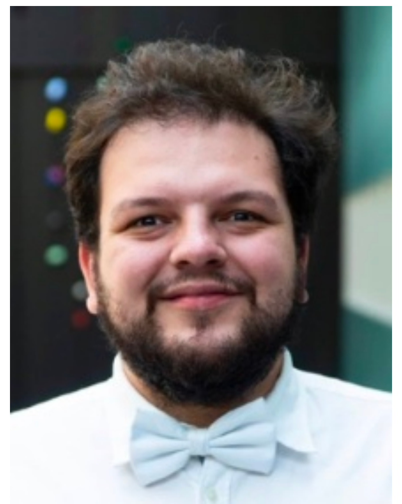
Gautam Reddy
Research Scientist
Assist. Prof. Princeton (Physics)

Logan Wright
Research Scientist
Assist. Prof. Yale (Applied Physics)

Maya Okawa
Research Scientist

Long-term Ph.D. Students

Ekdeep Singh Lubana
U. Mich (EECS)



Fatih Dinc
Stanford (Applied Physics)

Summer Ph.D. Student Interns

Mikail Khona
MIT (Physics)

Rahul Ramesh
U.Penn (CS)

William Tong
Harvard (Applied Math)

Kento Nishi
Harvard (CS)

Daniel Kunin
Stanford (Applied Math)

Ziyin Liu
UofTokyo (Physics)

Max Aalto
MIT (EECS)

Physics of Intelligent Systems Group at Harvard