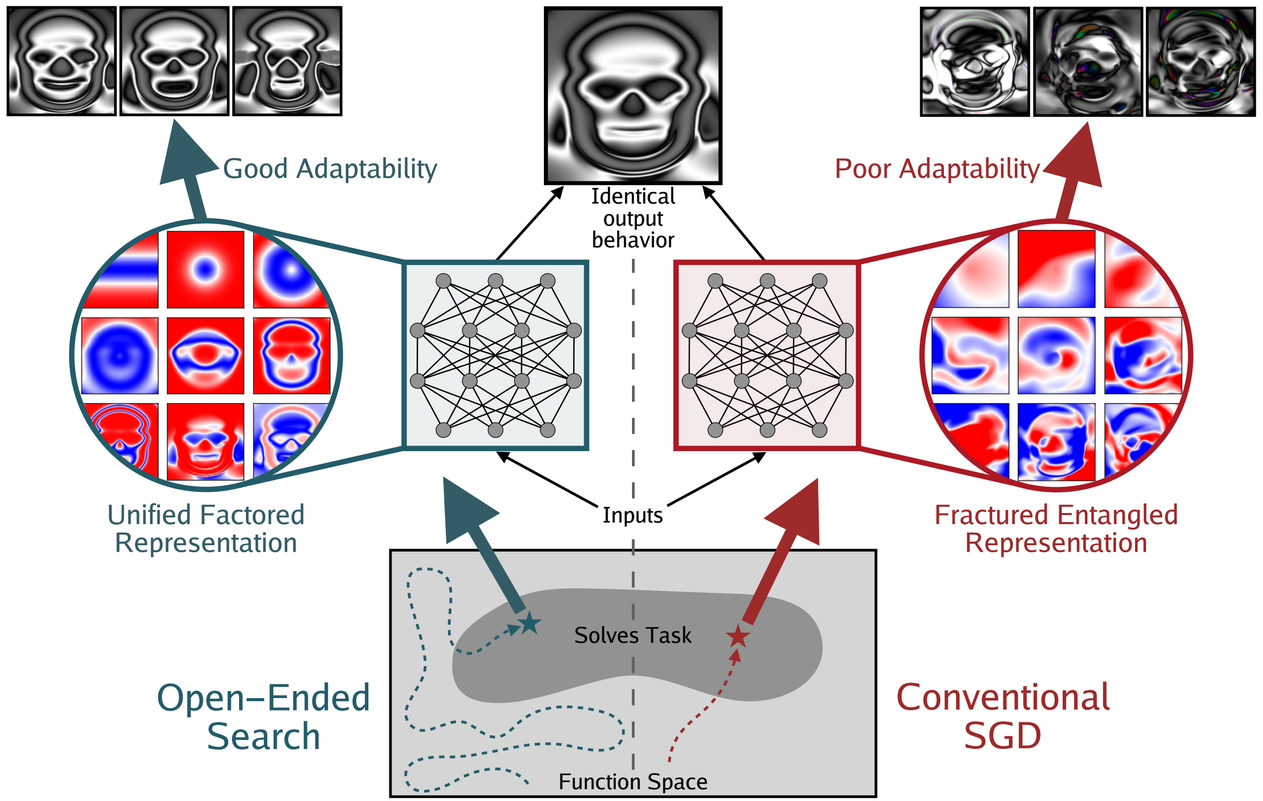
Towards a Platonic Intelligence with Unified Factored Representations

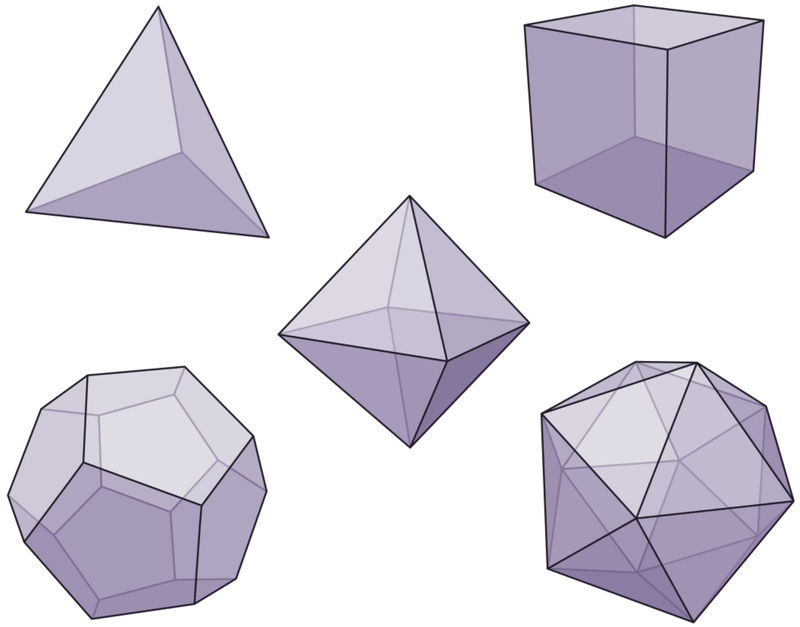


Akarsh Kumar
MIT CSAIL
November 4, 2025

The World is not Random

The World has Structure

Space of Forms



Plato

Real World

Intelligent Agents must capture this Structure

- To solve goals, intelligent agents must understand the world
  - Their *internal representation* must capture the structure of the world

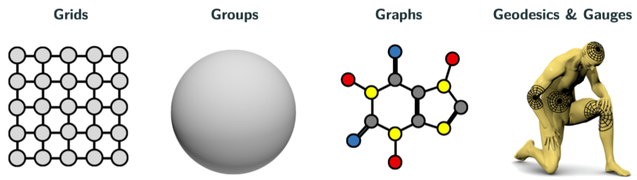
Capturing Structure with AI

- AI systems face the same problem: **how do we capture regularities of the world**?
  - We try to bake in some symmetries through architecture design
  - Inductive bias

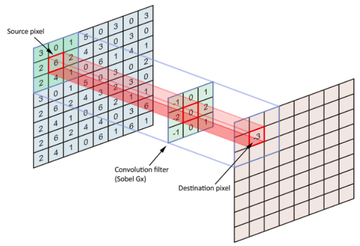
Geometric Deep Learning
Grids, Groups, Graphs,
Geodesics, and Gauges

Michael M. Bronstein[1], Joan Bruna[2], Taco Cohen[3], Petar Veličković[4]

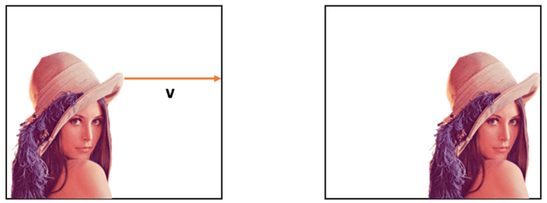
May 4, 2021



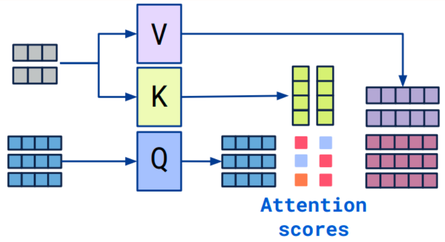
Convolutional Architecture



Translation Equivariance



Attention Architecture



Permutation Equivariance

Capturing Structure with AI

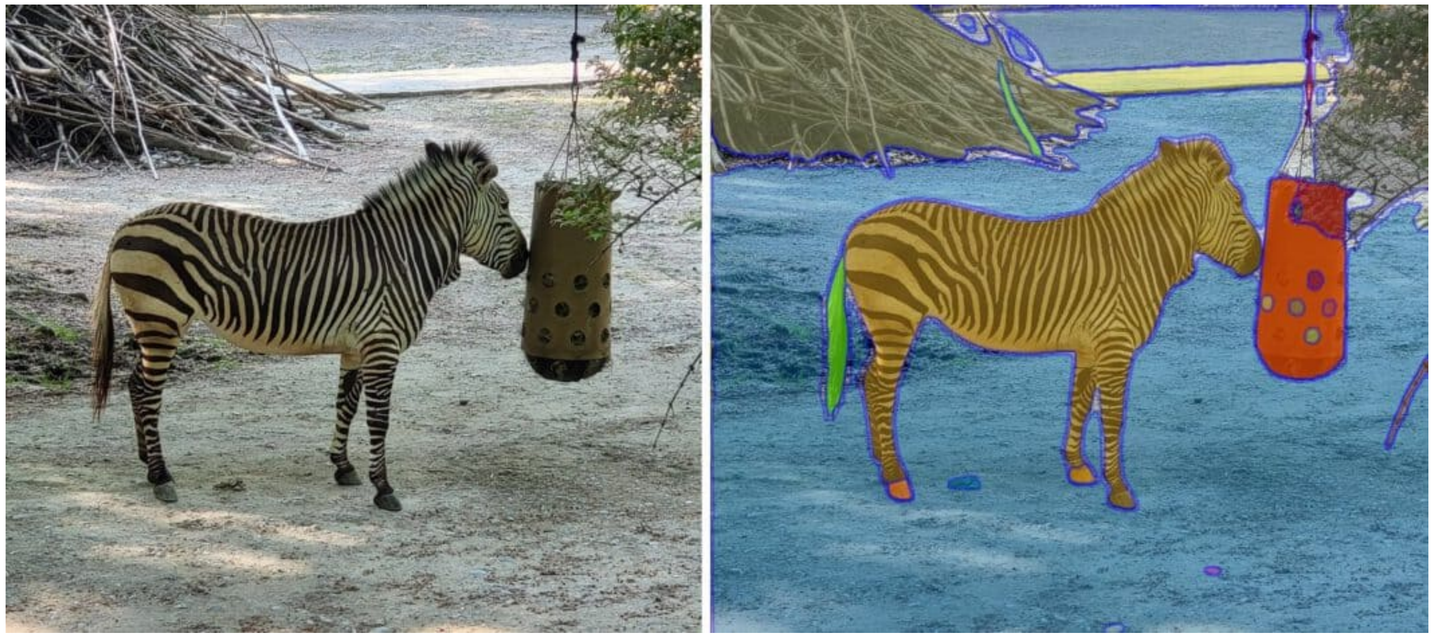
- What about everything else?
  - Example: lighting invariance
  - How do you capture lighting invariance?
  - We don’t know
- **Solution: train on lots of data with SGD**

Architecture

?

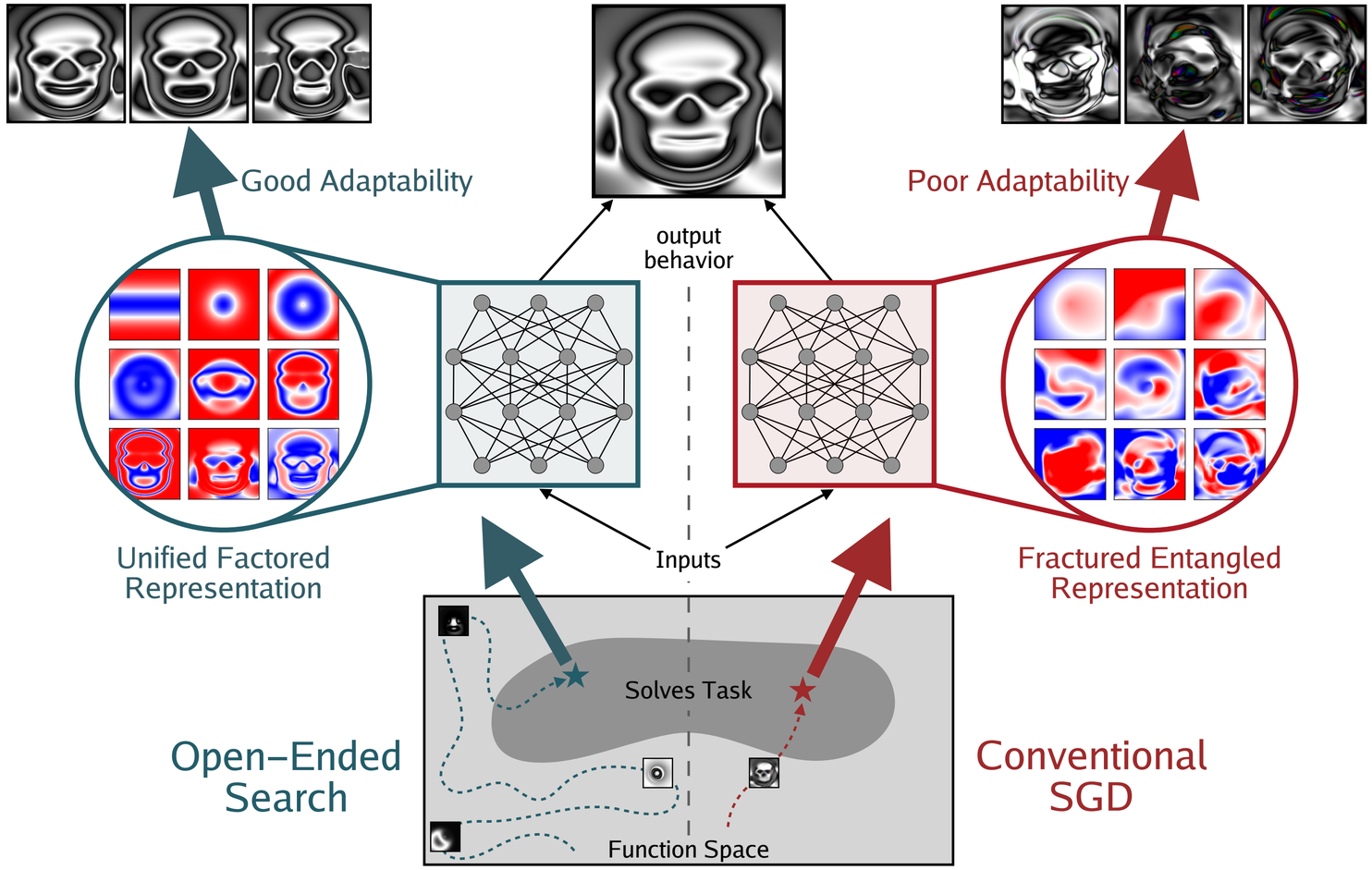
Lighting Invariance

Does this Work?



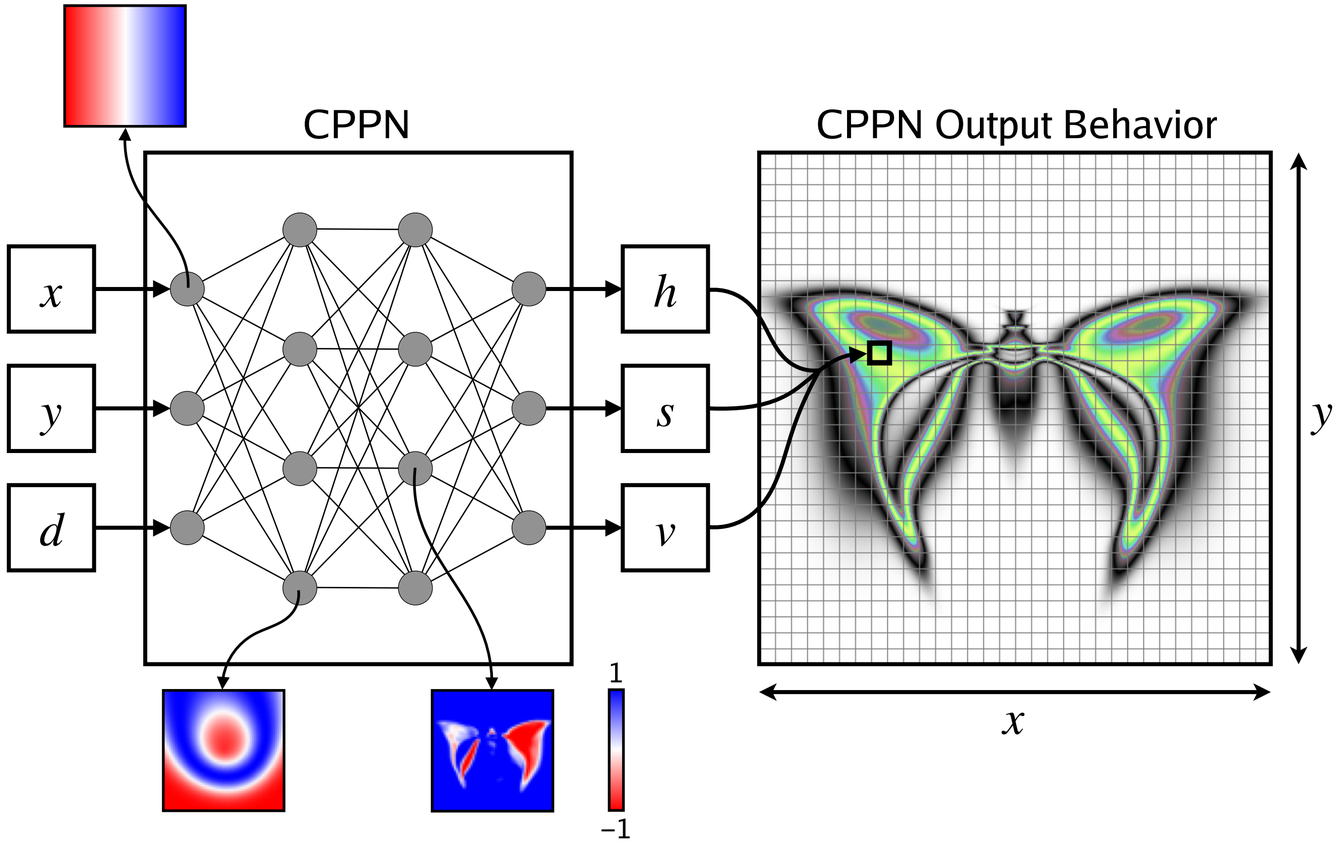
Hypothesis: Fractured Entangled Representations

- **Conventional SGD** training finds neural representations which are **fractured** and **entangled**
  - Doesn't capture the underlying regularities of the world
- Position: **Open–Ended Search** may be the solution to learn **unified** and **factored** neural representations
- Internal representation affects **generalization**, **creativity**, and **continual learning**



Compositional Pattern Producing Network (CPPN)

- Toy domain to study neural representations: implicitly represent an image
  - Inspired by biological developmental process
- Easy to visualize *how output behavior* is *internally represented*, neuron by neuron

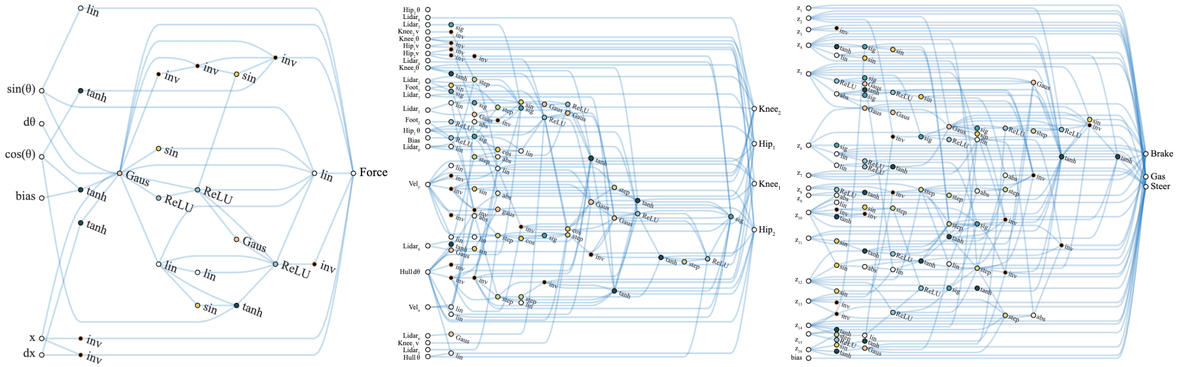


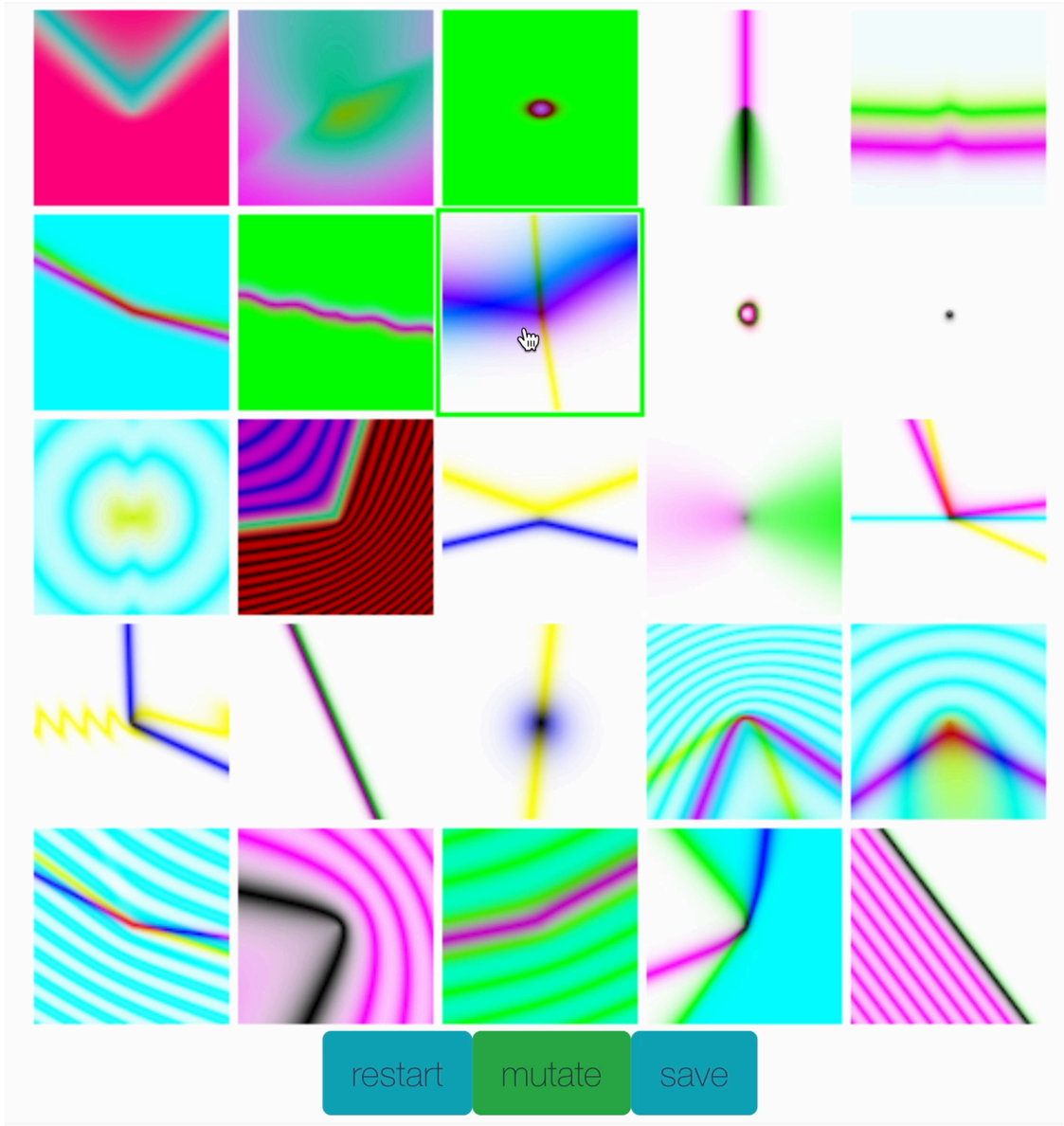
CPPNs are an Analogy

- **CPPNs are a useful visual analogy to think about LLMs**
- CPPN's output image $\iff$ LLM's output behavior
- CPPN's internal visualization $\iff$ LLM's internal representation
  - Visualize how behaviors are actually constructed holistically
- Two CPPNs may have same output, but inner encodings could be qualitatively different
  - Two LLMs may have same behavior, but their inner representation might be qualitatively different

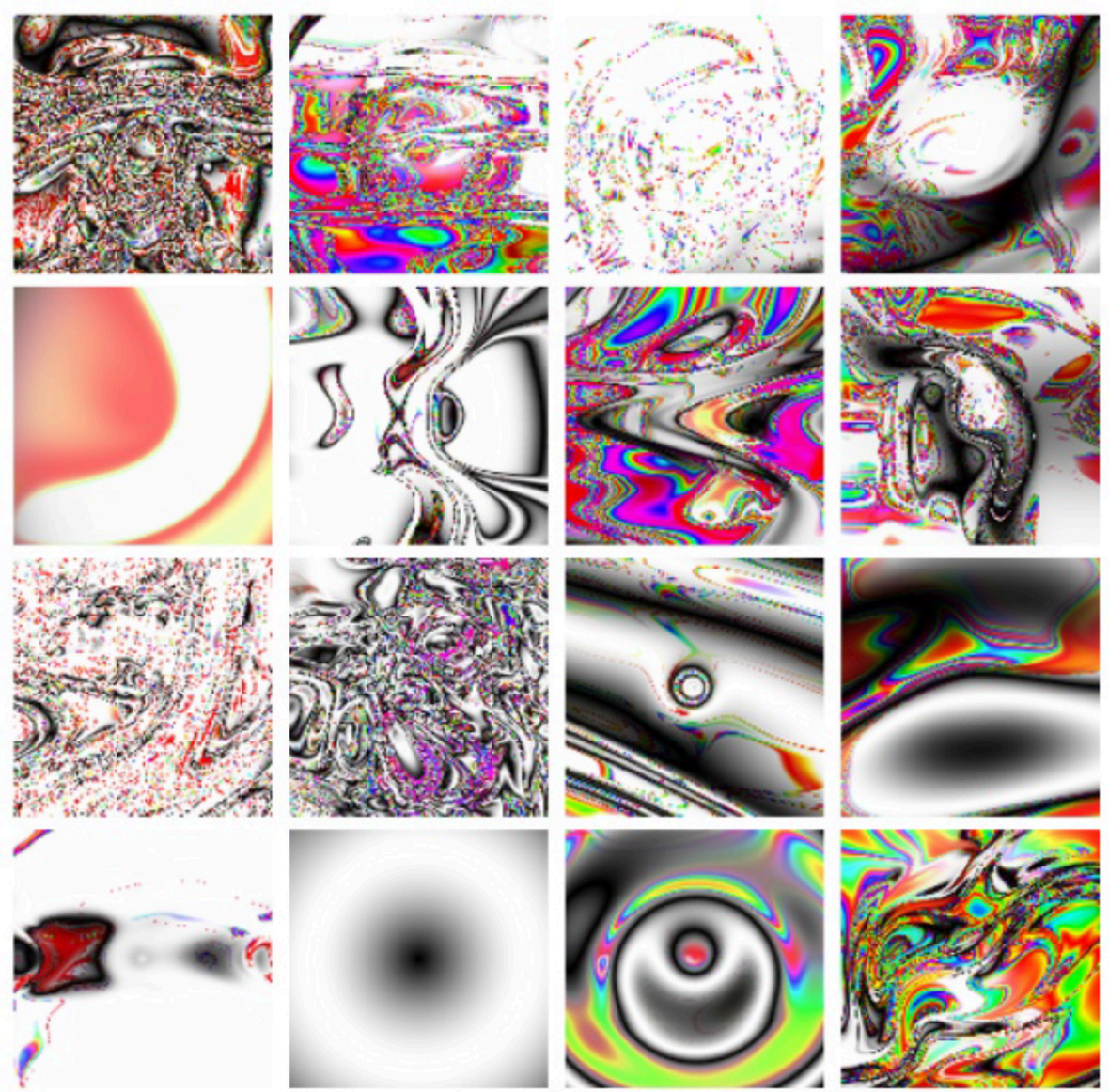
Picbreeder!

- Online website for humans to breed images to their desire
  - Evolve the underlying CPPNs
- **No end goal, do whatever you want!**
- *NEAT based CPPNs

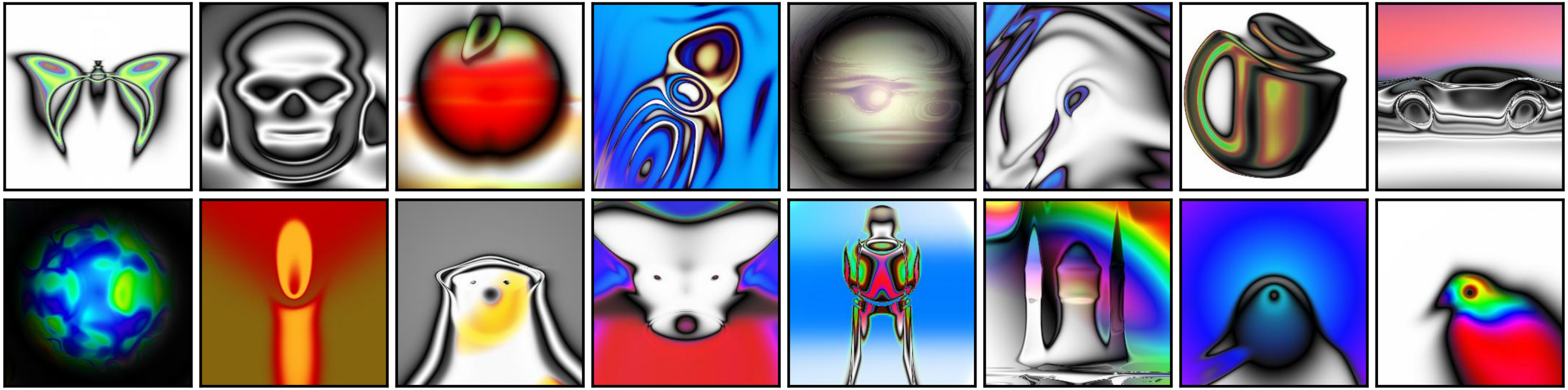




What Do You Expect to Find?

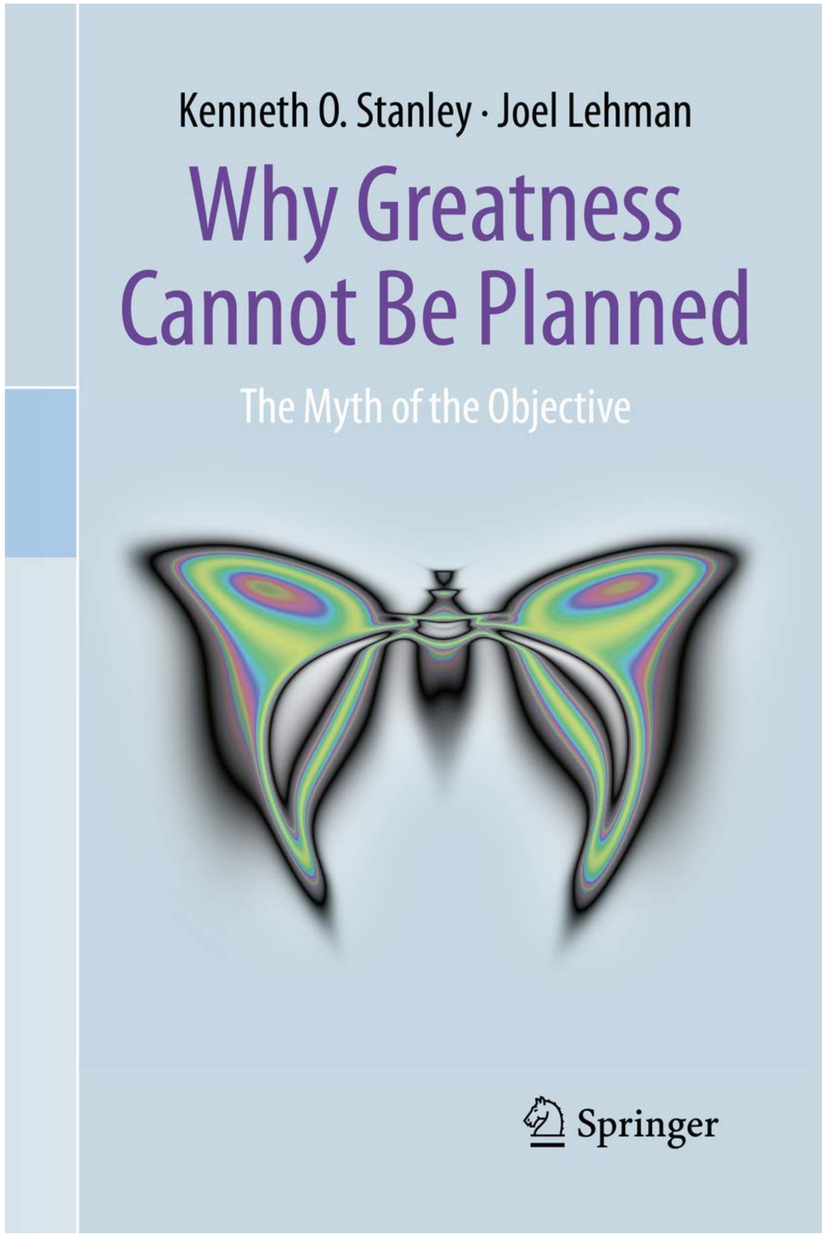


What People Actually Found!



Why Greatness Cannot be Planned

- Many insights on the nature of search
  - Deception
  - Serendipity
  - Open–Endedness
- Case studies:
  - Natural Evolution
  - Scientific Innovation

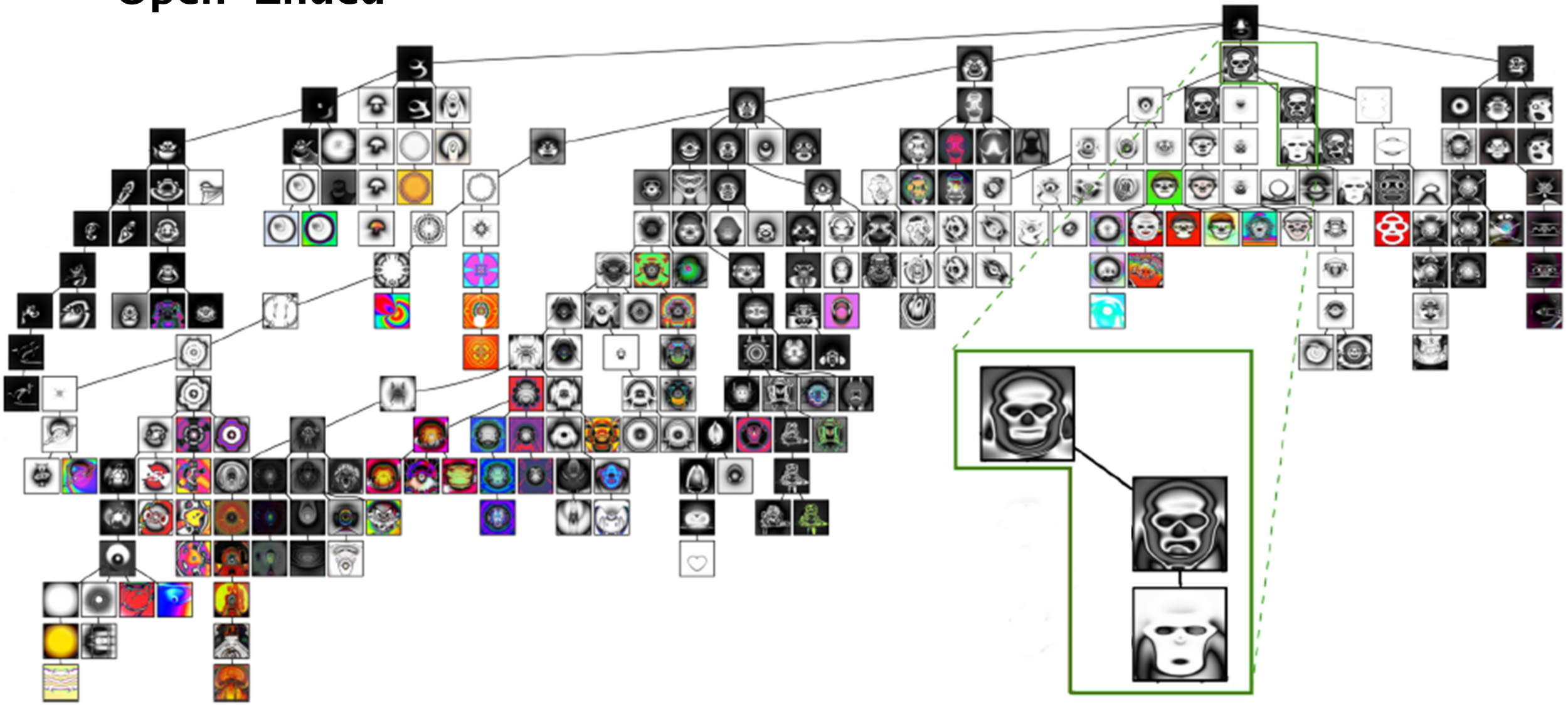


Picbreeder has Intriguing Properties

Open-Ended

Picbreeder has Intriguing Properties

Open-Ended



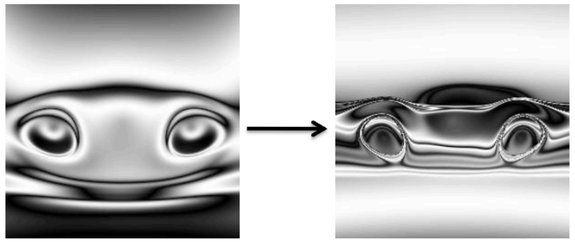
Picbreeder has Intriguing Properties

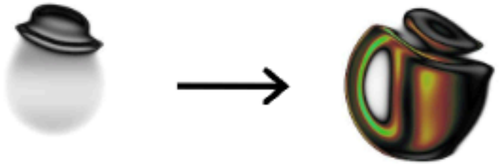
Serendipitous Exaptation

- When a trait evolved for one function but gets **repurposed** for another function

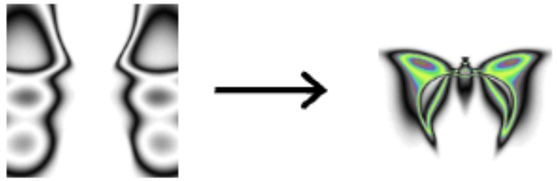
Picbreeder has Intriguing Properties

Serendipitous Exaptation



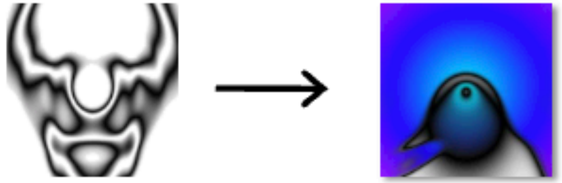


Stepping stone to the Teapot



Stepping stone to the Butterfly

Stepping stone to the Skull



Stepping stone to the Penguin

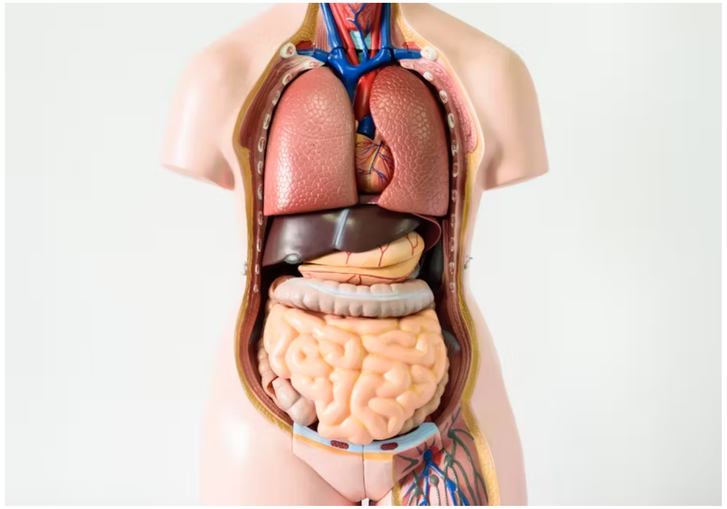
Stepping stone to Jupiter

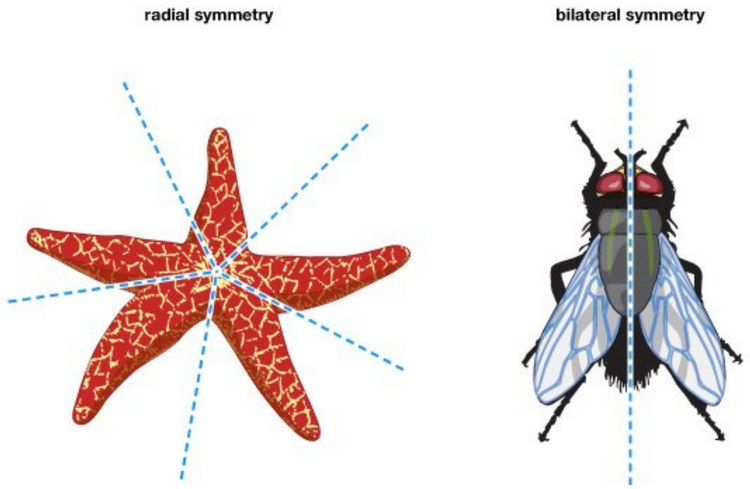
Stepping stone to the Lamp

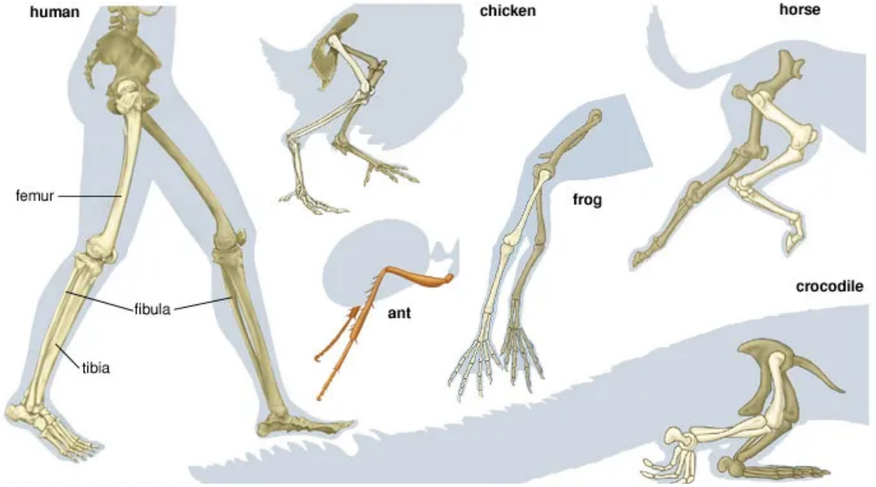
Picbreeder has Intriguing Properties

Emergence of Evolvability

- Natural evolution has developed *adaptable* genotypes:
  - Canalization
  - Regularity
  - Modularity
  - Symmetry
- Certain axes of variation become more likely while others become impossible

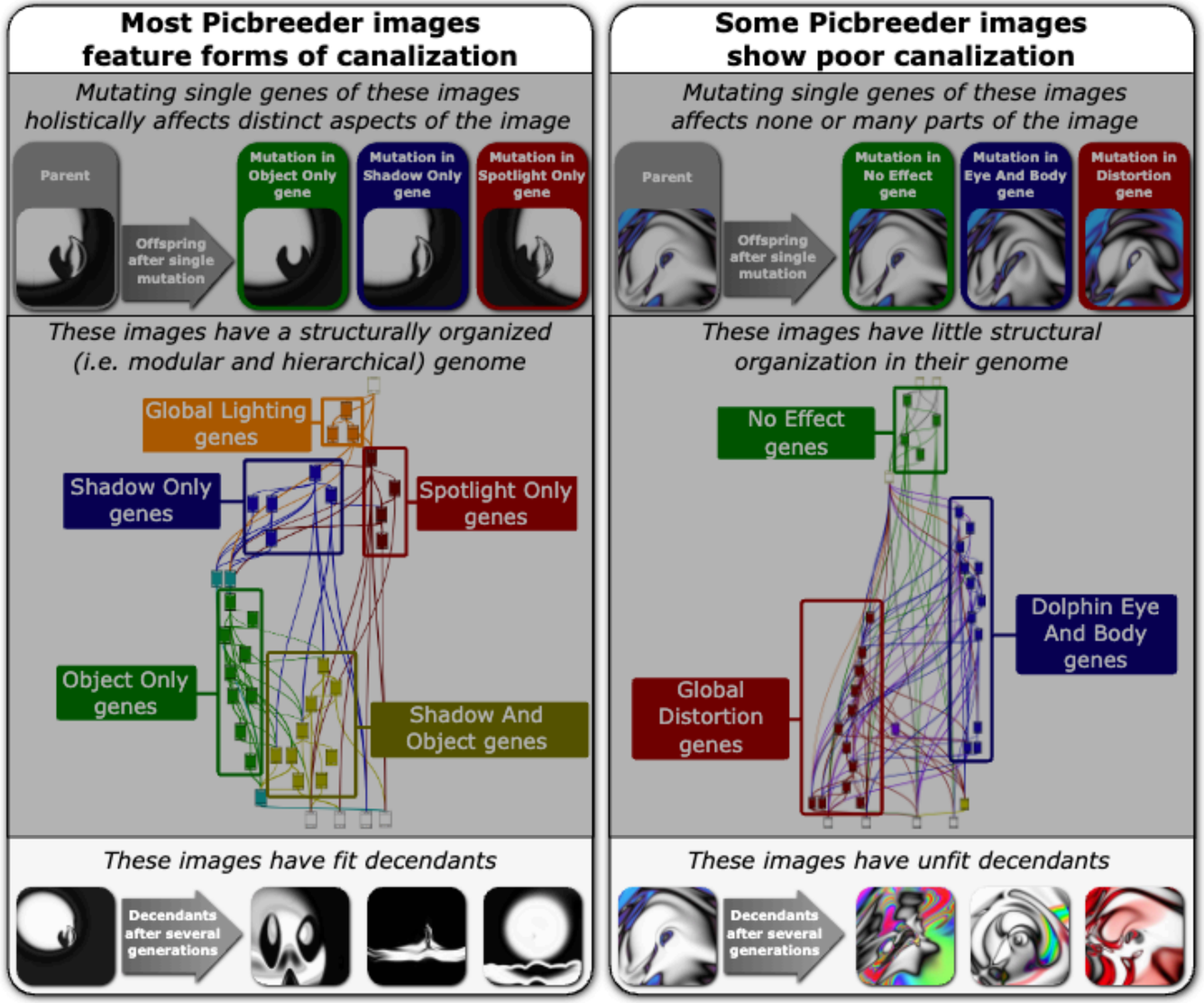




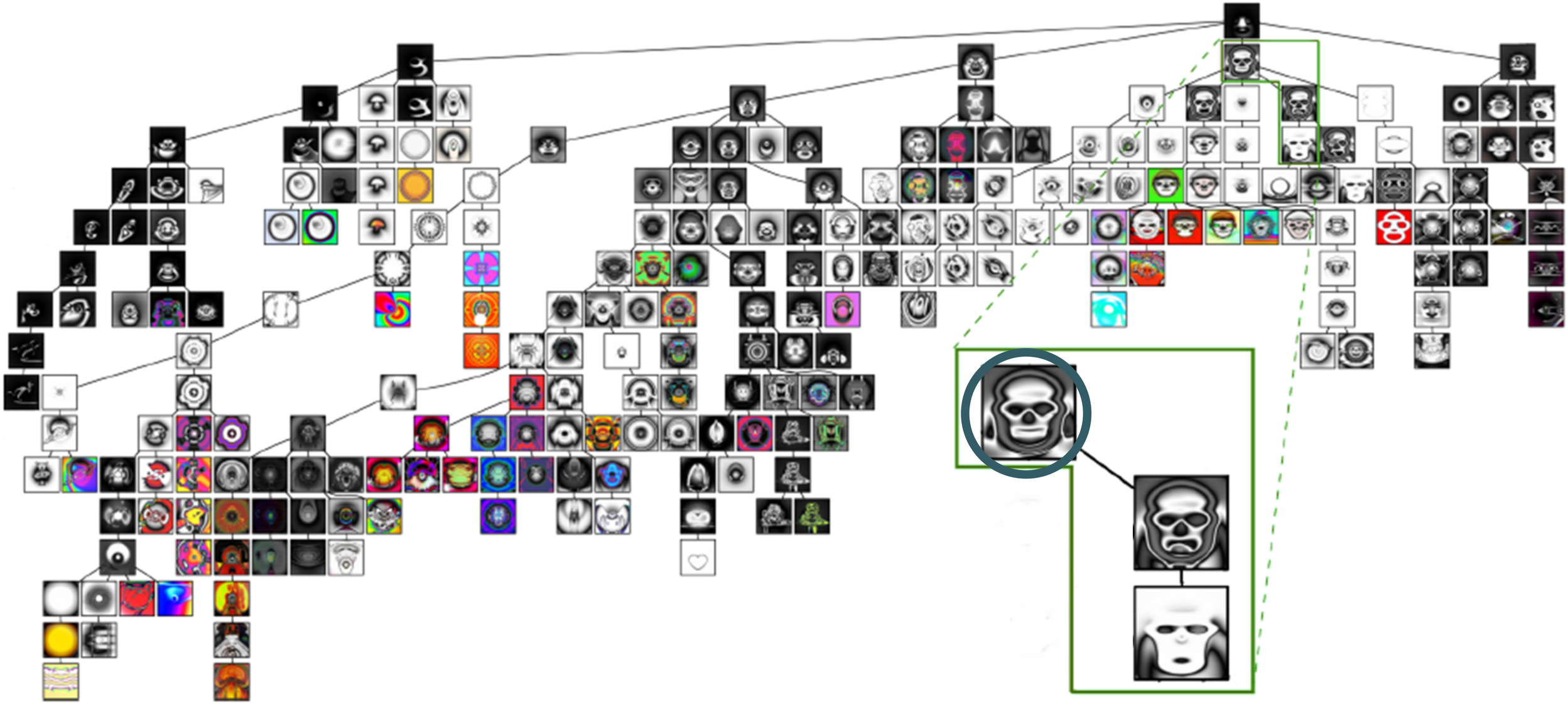


Picbreeder has Intriguing Properties

Emergence of Evolvability



Great, Let's Look Into an Image



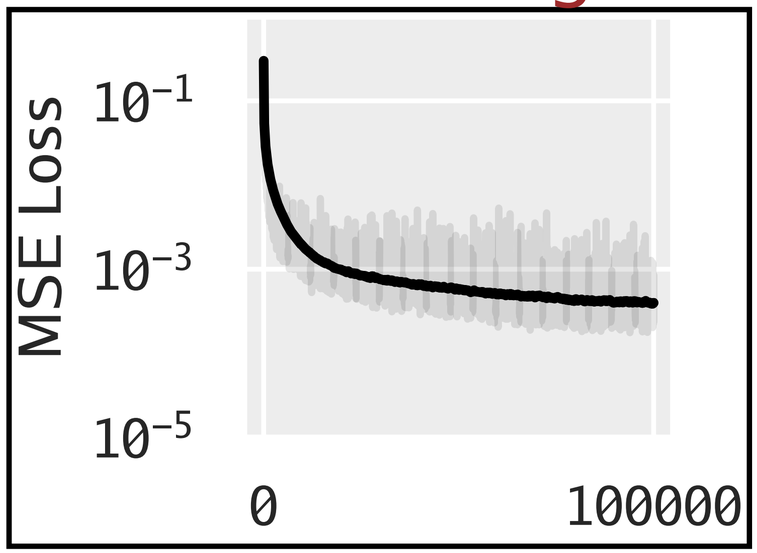
Learning the Picbreeder Skull with SGD

- Let's train a conventional network to recreate the skull
  - Perfect reconstruction!

Picbreeder Skull

$$\begin{bmatrix}(x,y,d) \rightarrow (h,s,v)\\(x,y,d) \rightarrow (h,s,v)\end{bmatrix}$$
$$\dots$$
$$\begin{bmatrix}(x,y,d) \rightarrow (h,s,v)\\(x,y,d) \rightarrow (h,s,v)\end{bmatrix}$$

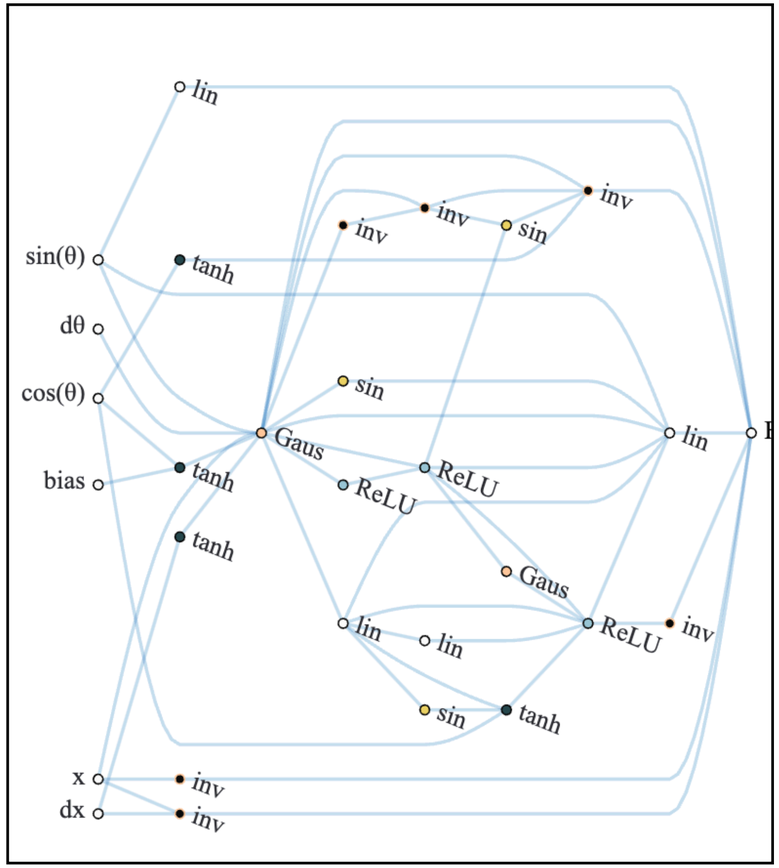
SGD Training

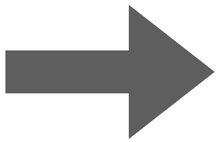


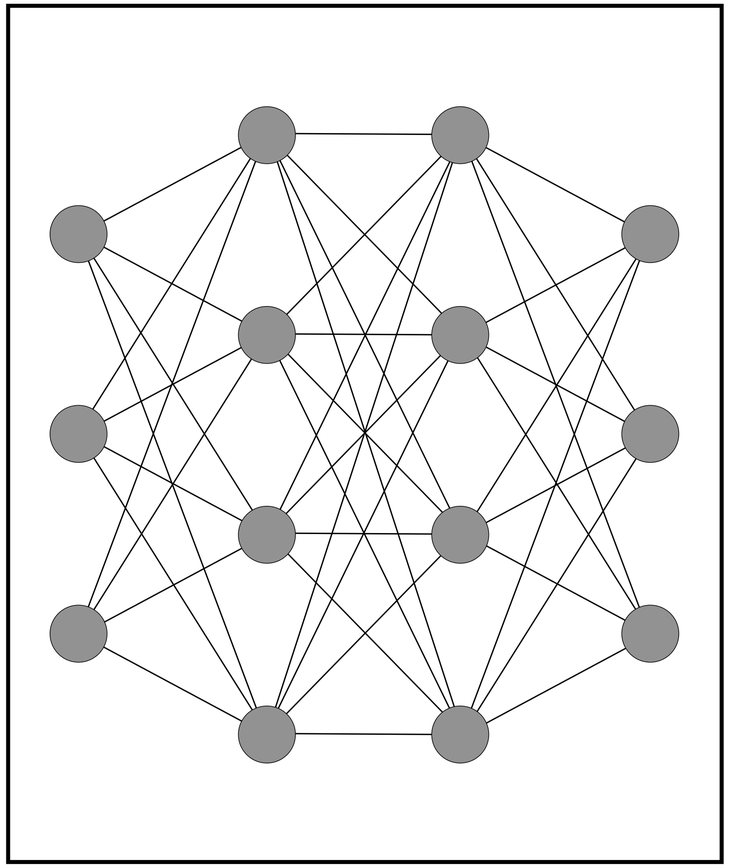
SGD Skull

Layerization

- Convert everything to a universal architecture space: MLP
  - Existence proof of Picbreeder solution MLP weight space

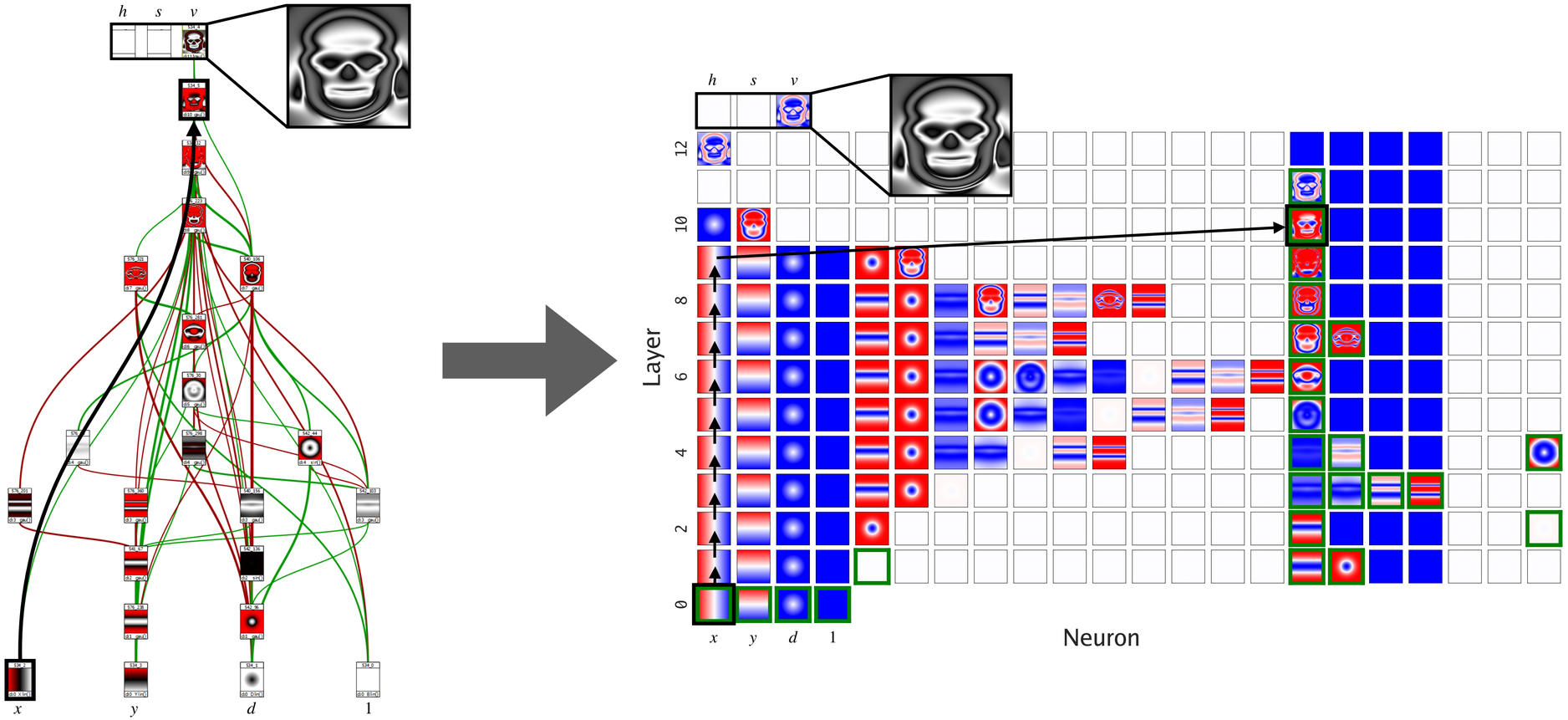






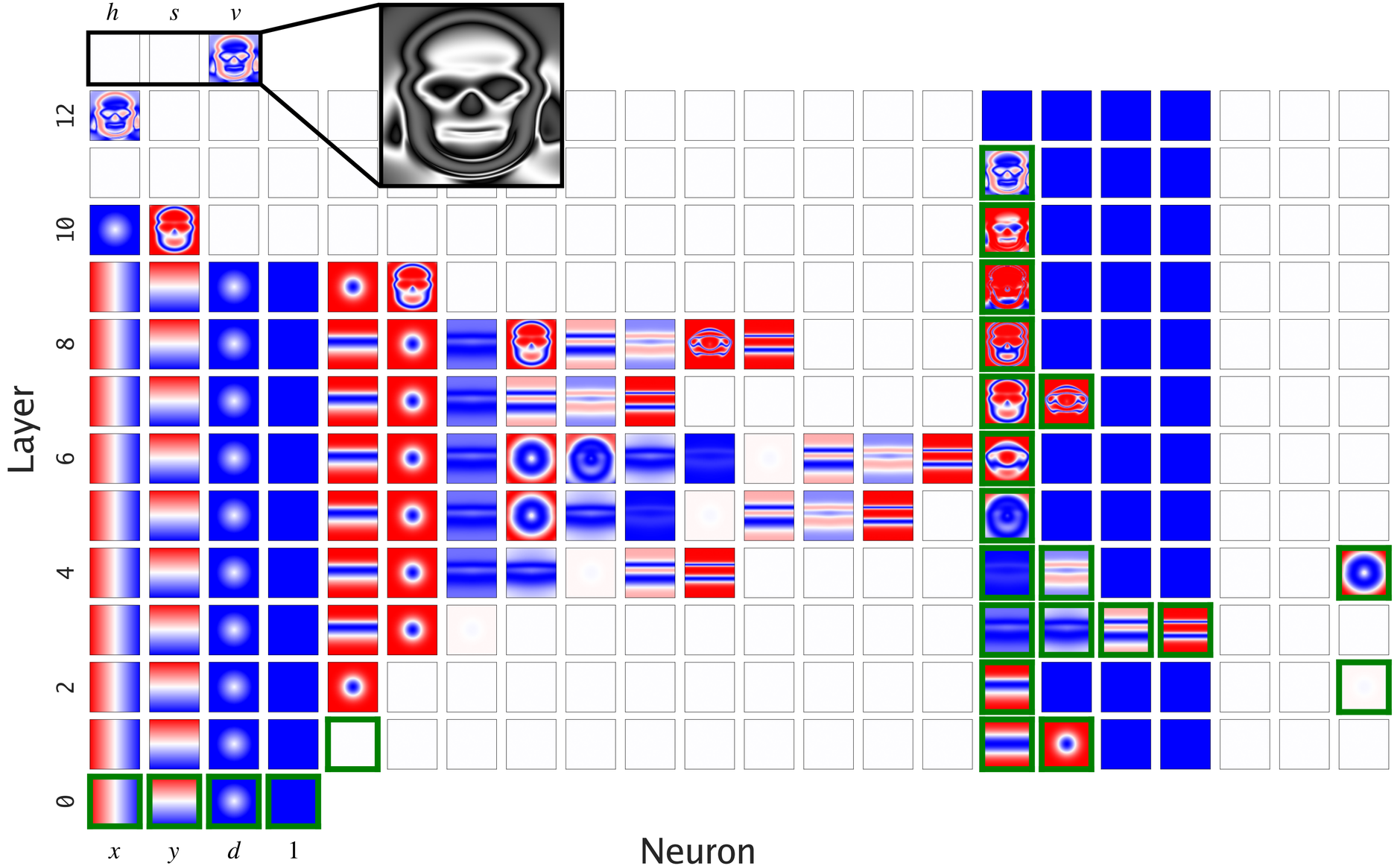
Layerization

- Convert everything to a universal architecture space: MLP
  - Existence proof of Picbreeder solution MLP weight space



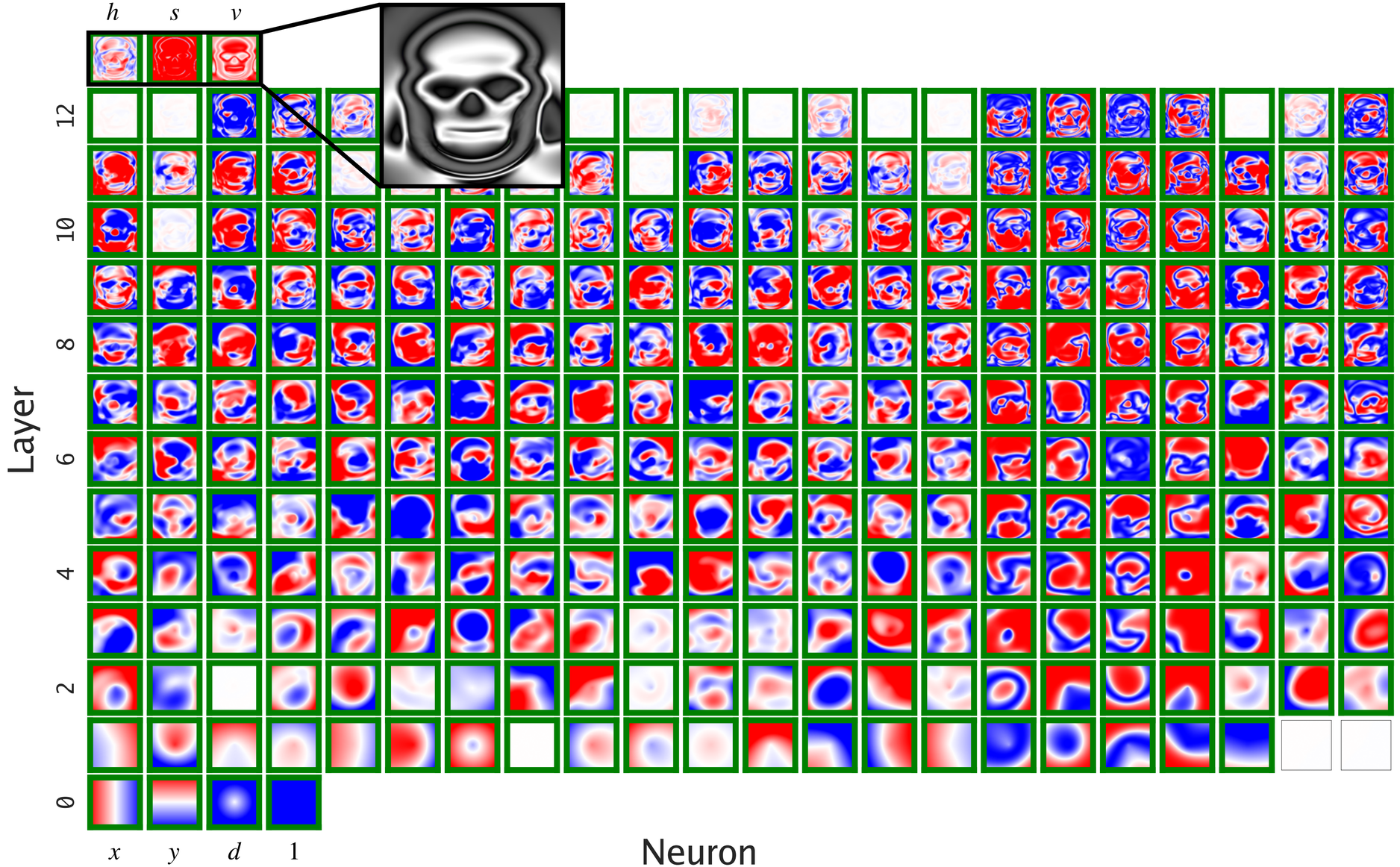
Picbreeder Skull

Unified Factored Representation



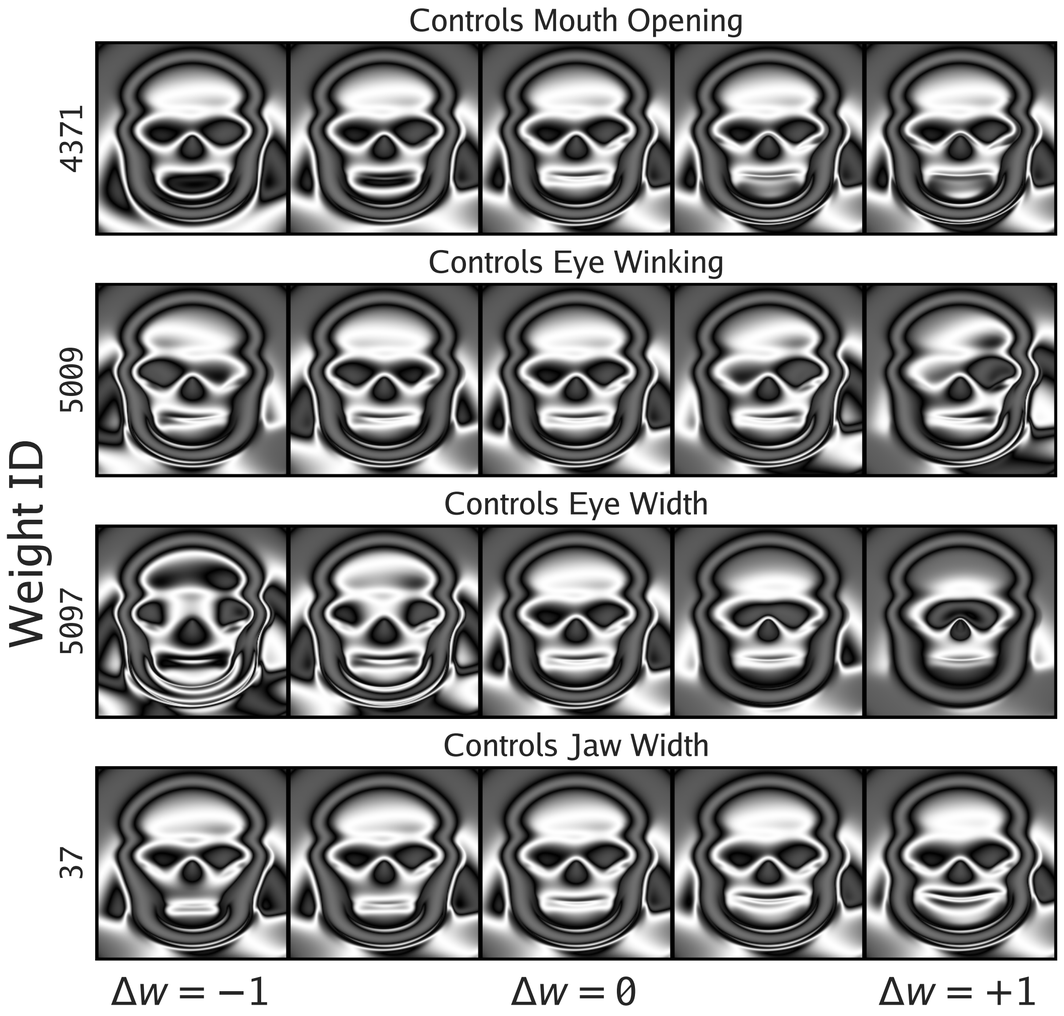
SGD Skull

Fractured Entangled Representation



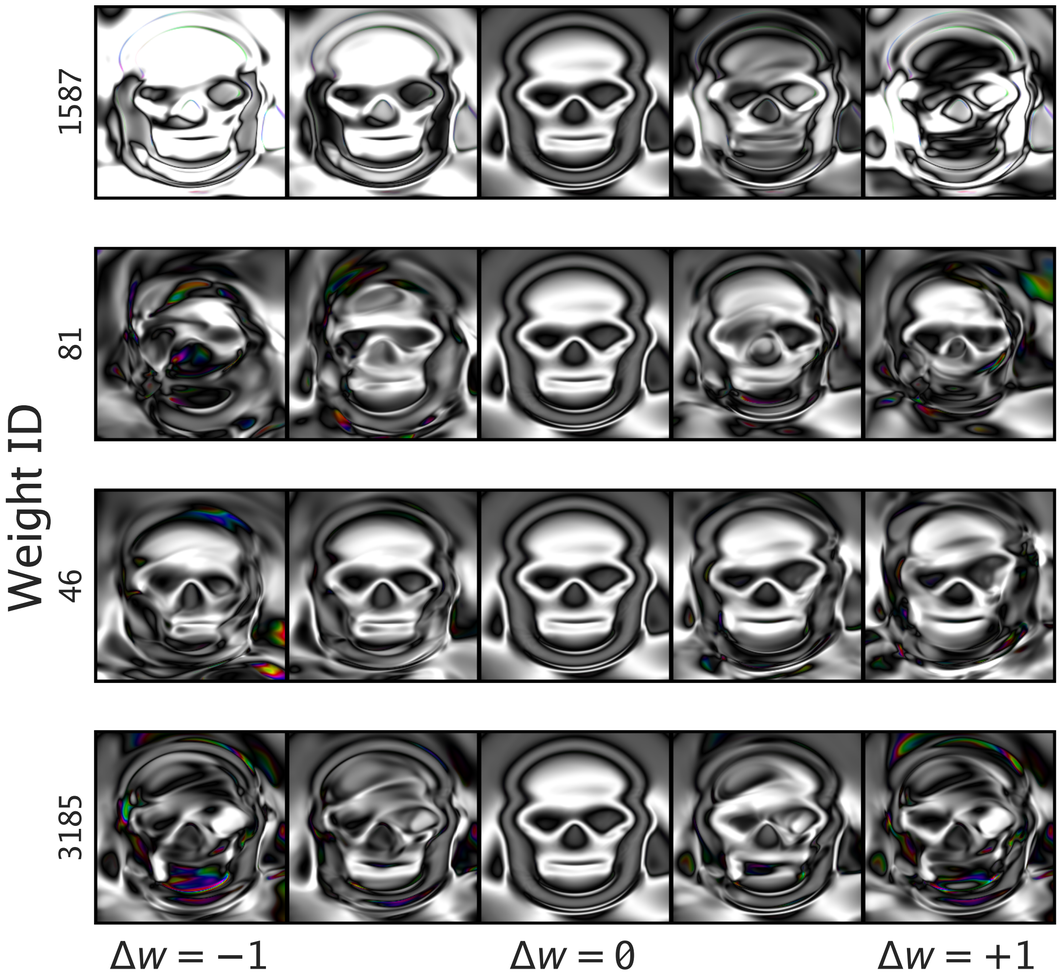
Picbreeder Skull

Unified Factored Representation



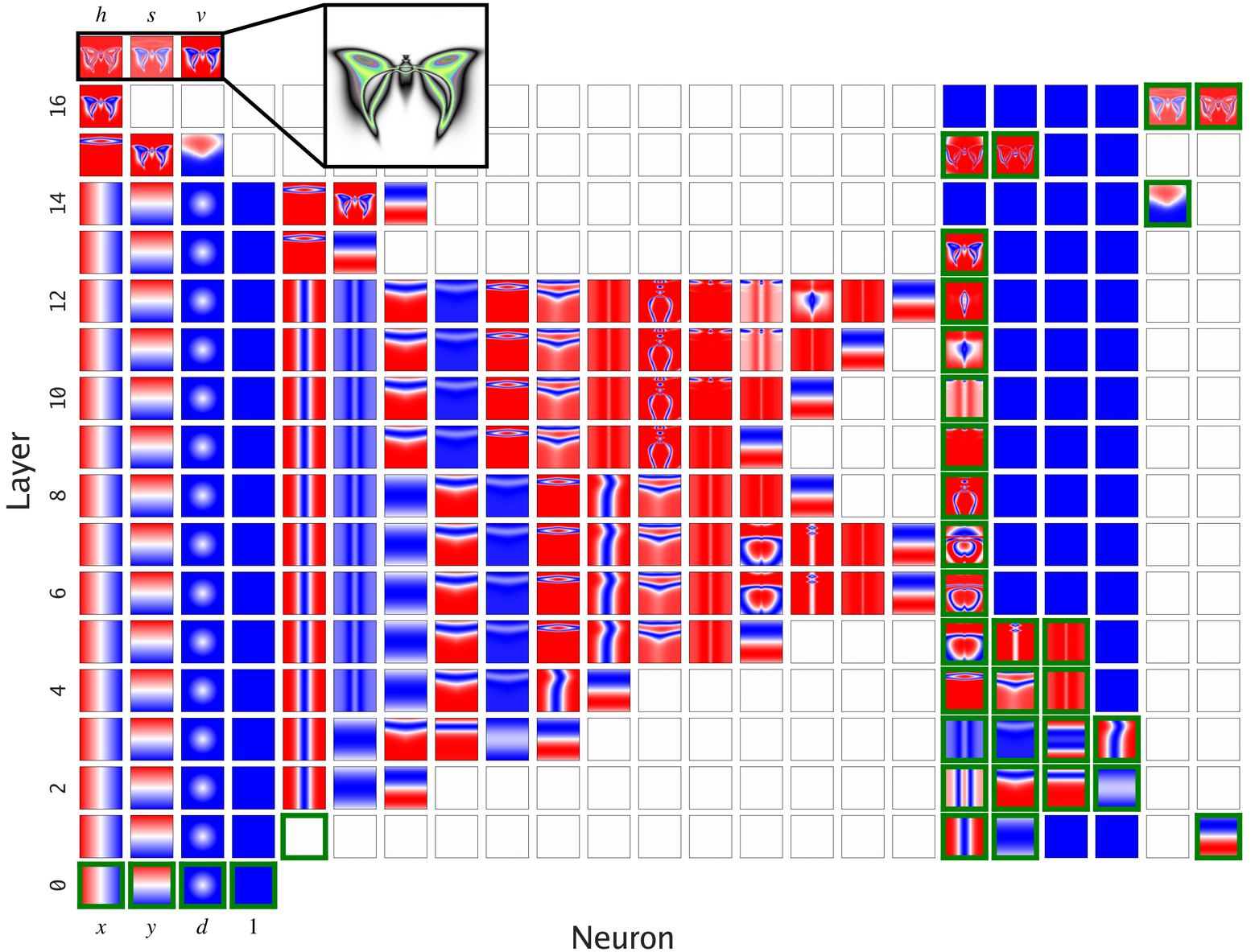
SGD Skull

Fractured Entangled Representation



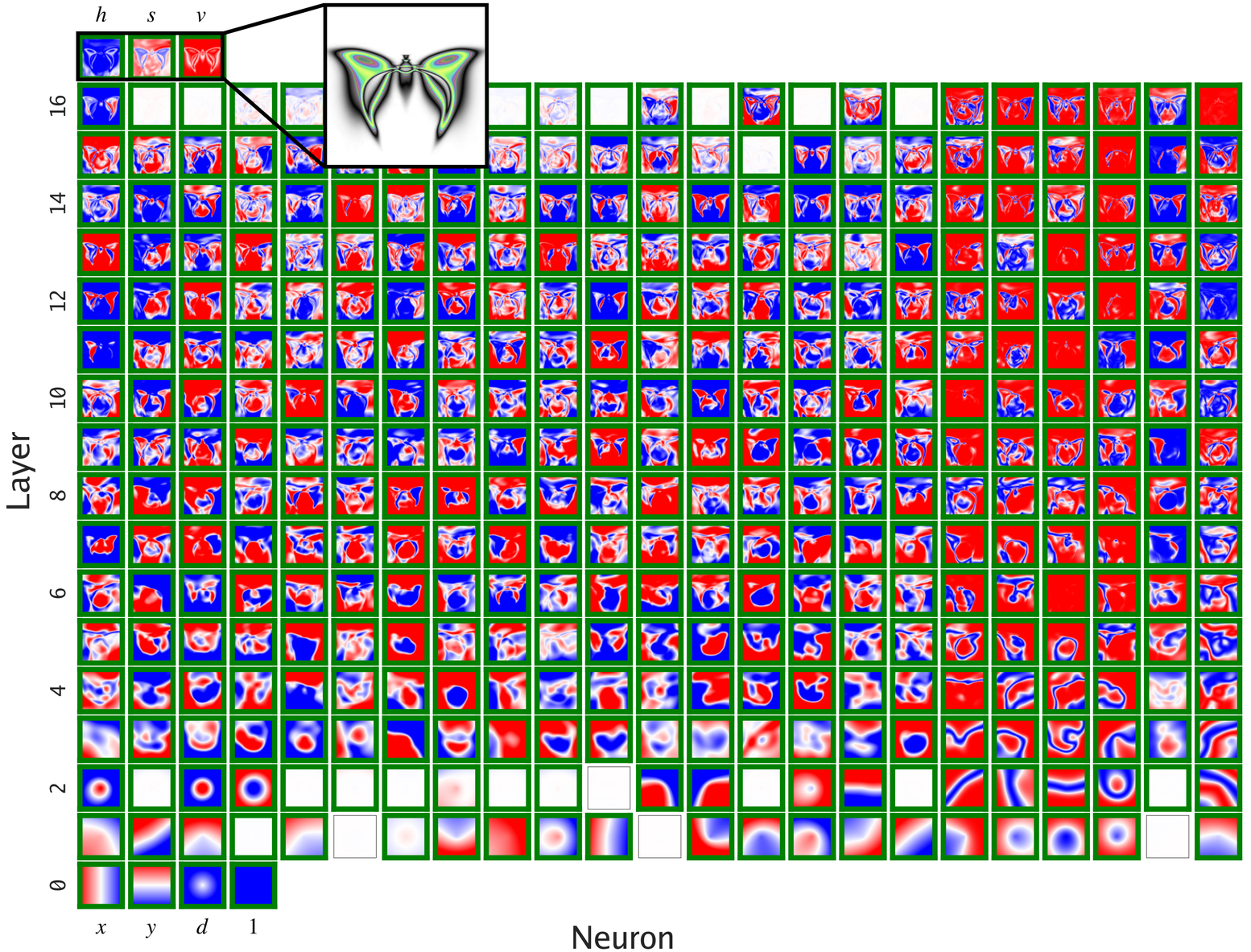
Picbreeder Butterfly

Unified Factored Representation



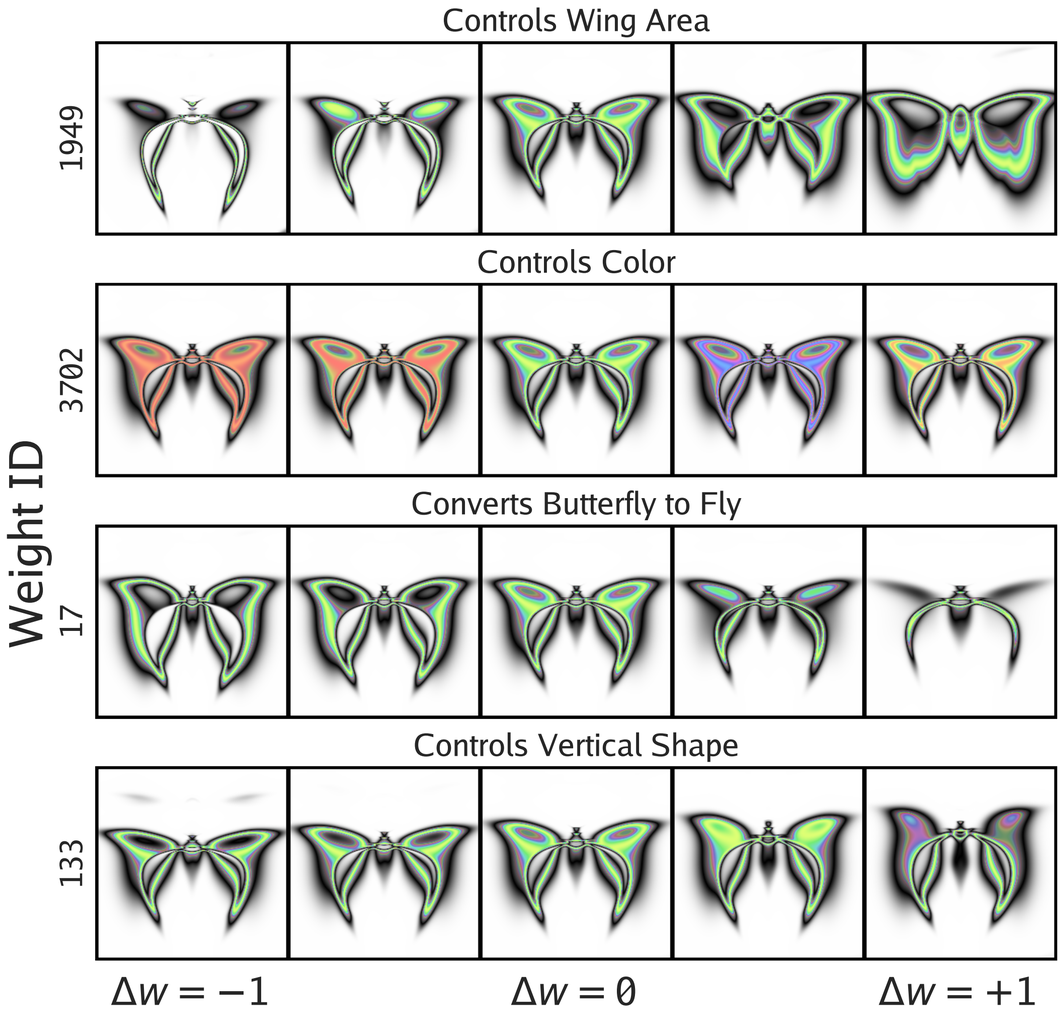
SGD Butterfly

Fractured Entangled Representation



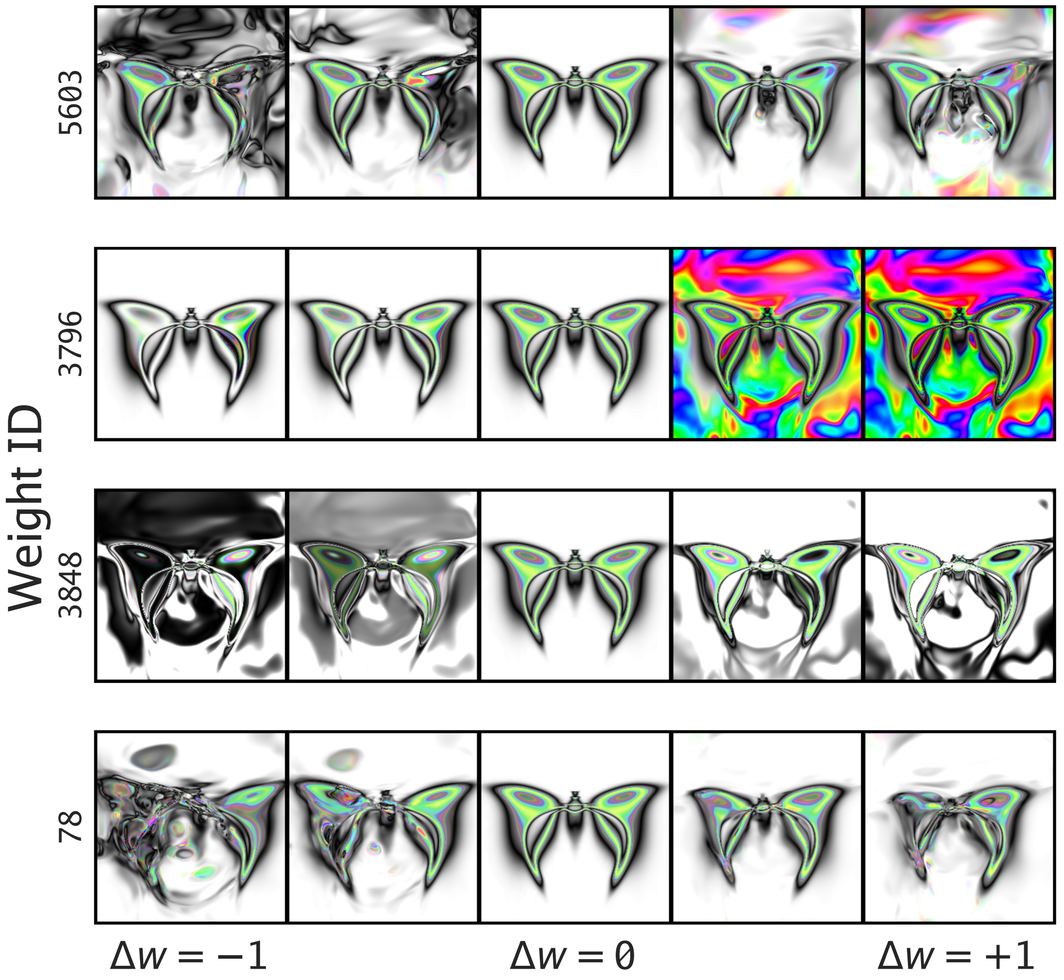
Picbreeder Butterfly

Unified Factored Representation



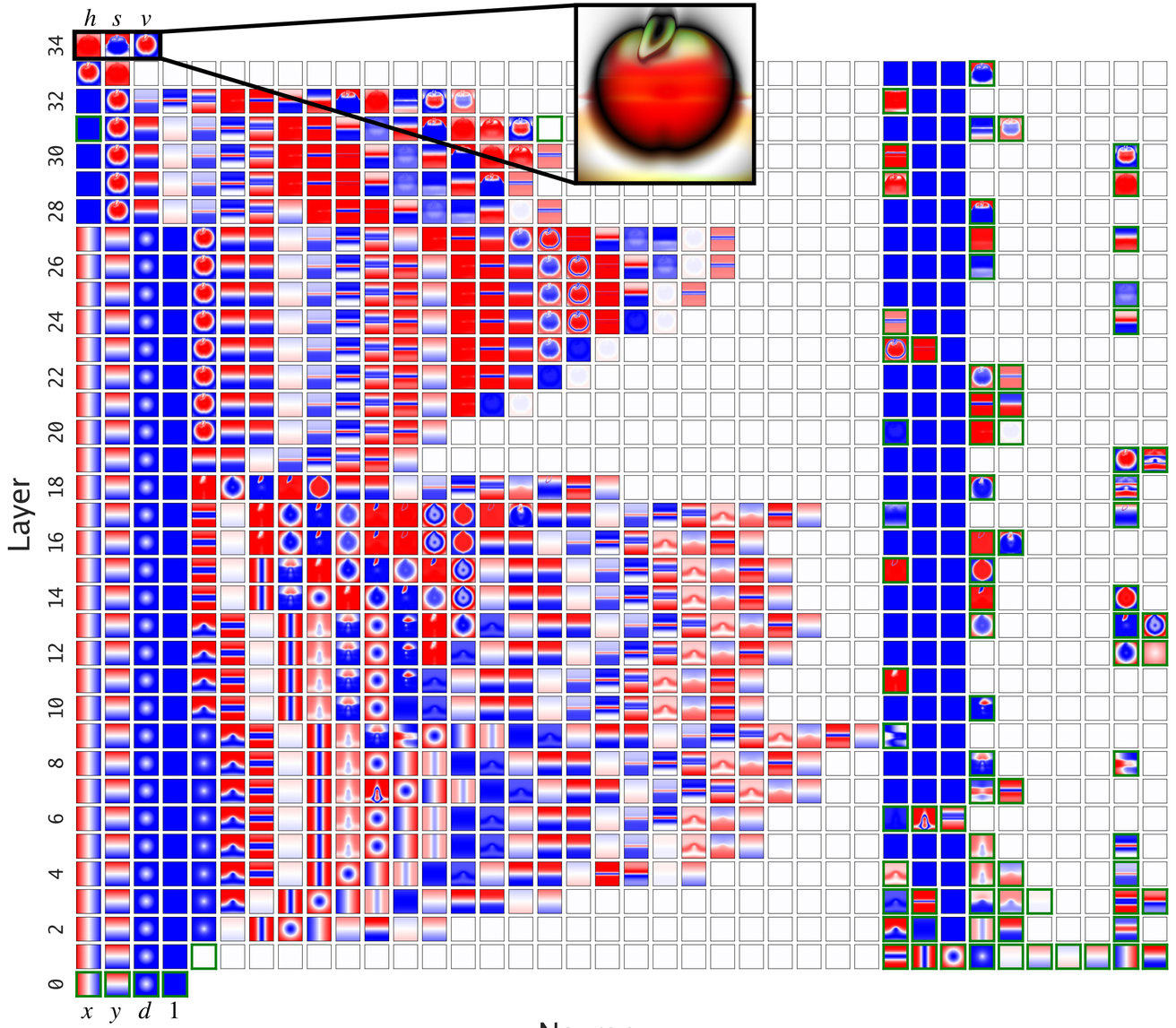
SGD Butterfly

Fractured Entangled Representation



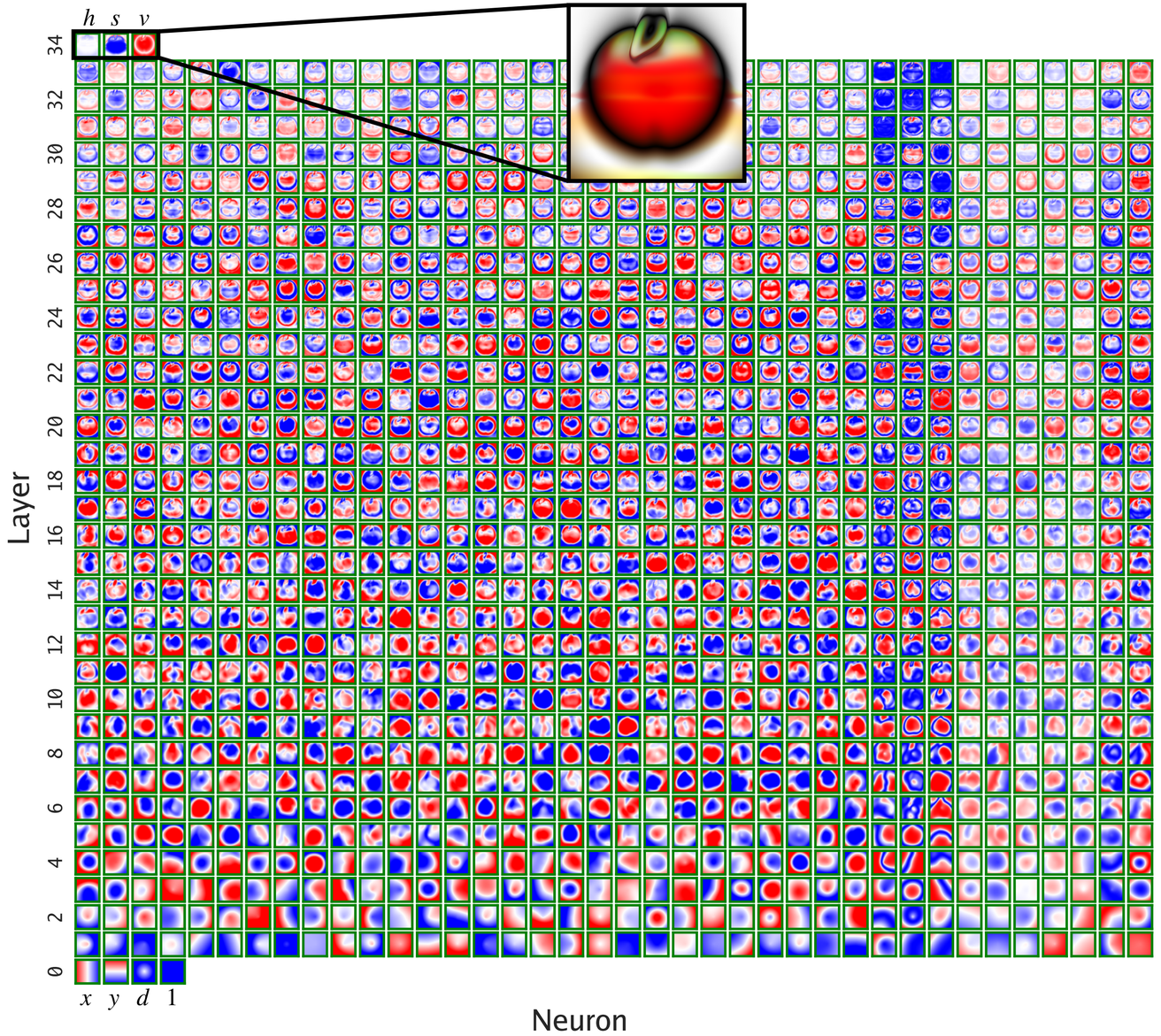
Picbreeder Apple

Unified Factored Representation



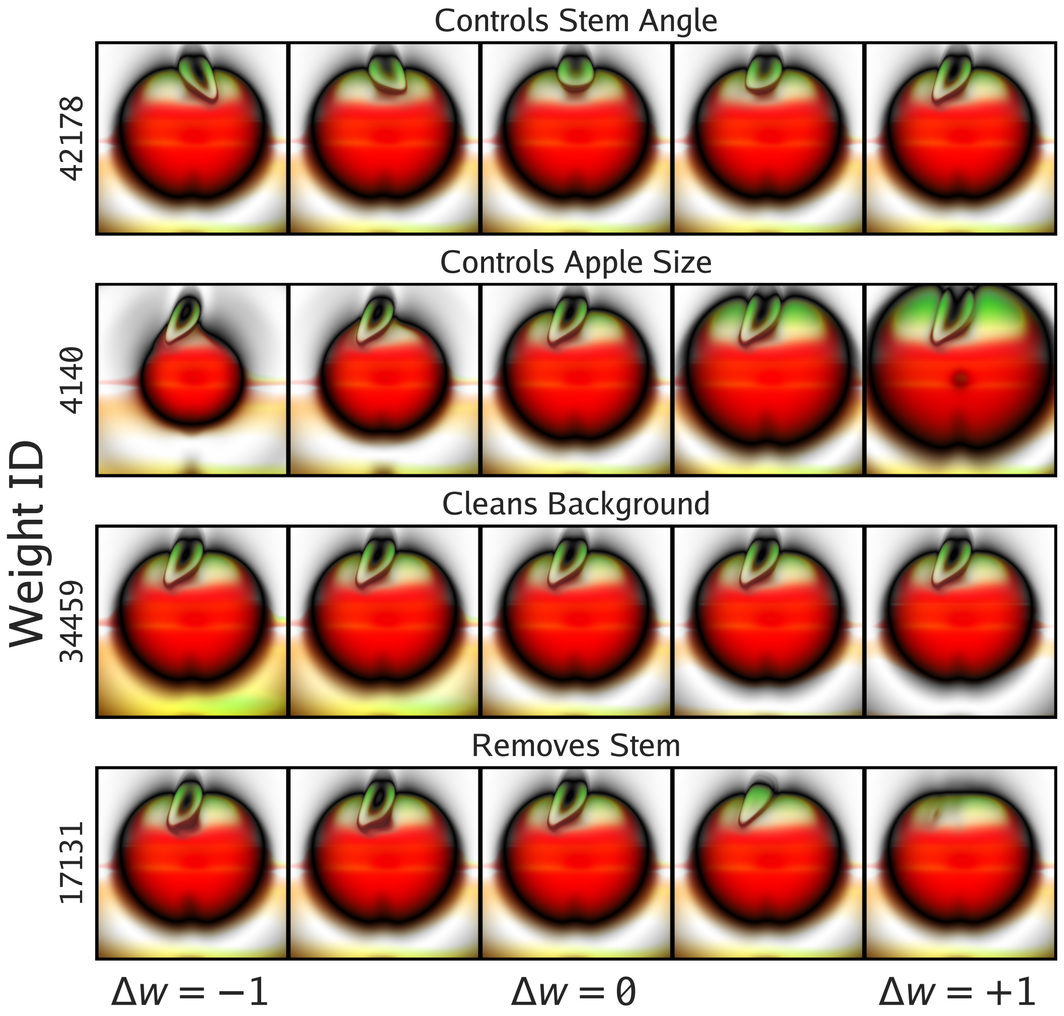
SGD Apple

Fractured Entangled Representation



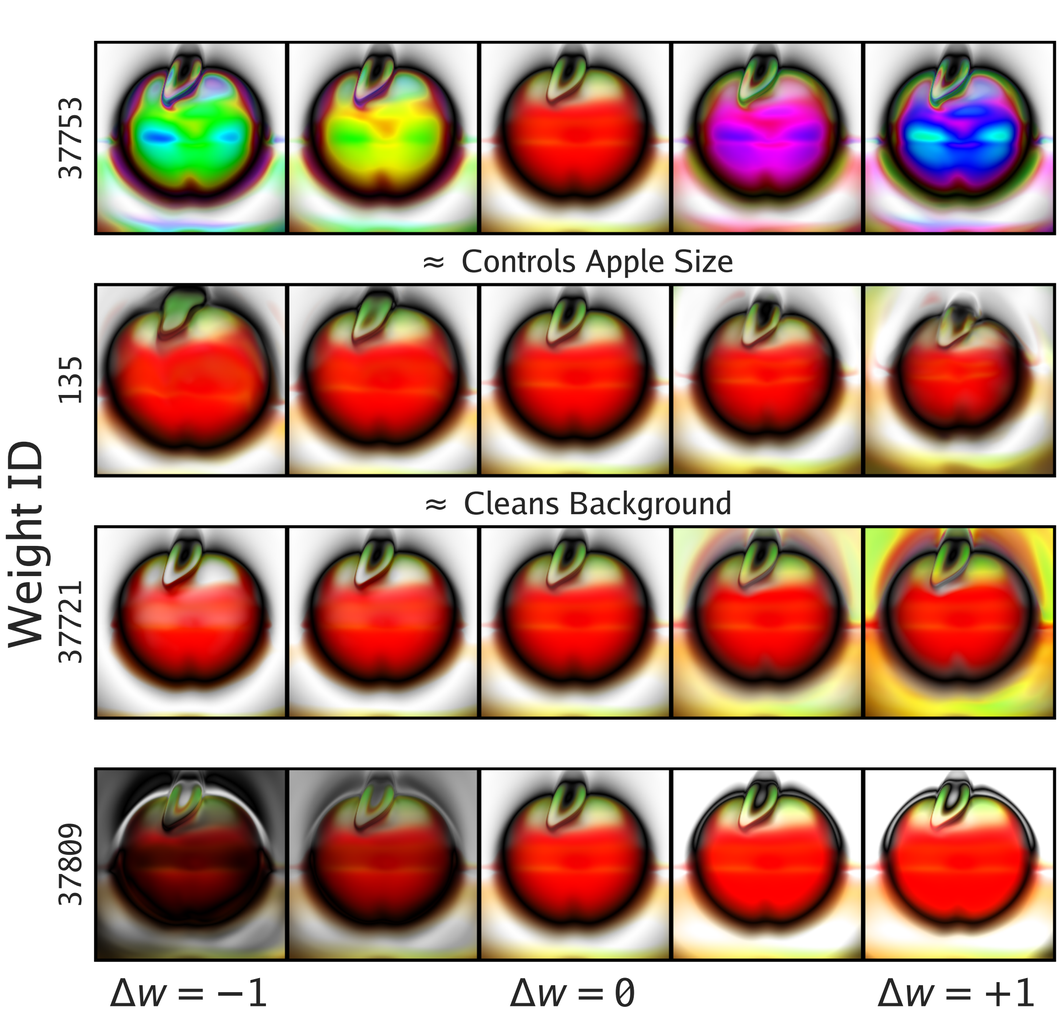
Picbreeder Apple

Unified Factored Representation



SGD Apple

Fractured Entangled Representation



How does this Apply to LLMs?

FER In LLMs

Evidence in GPT-3

Example 1:

Me: I have 3 pencils, 2 pens, and 4 erasers. How many things do I have?

GPT-3: You have 9 things. [always correct]

Example 2:

Me: I have 3 chickens, 2 ducks, and 4 geese. How many things do I have?

GPT-3: You have 10 animals total. [always incorrect]

GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models

Iman Mirzadeh[†] Keivan Alizadeh Hooman Shahrokhi[*]
Oncel Tuzel Samy Bengio Mehrdad Farajtabar[†]

Apple

Abstract

Recent advancements in Large Language Models (LLMs) have sparked interest in their formal reasoning capabilities, particularly in mathematics. The GSM8K benchmark is widely used to assess the mathematical reasoning of models on grade-school-level questions. While the performance of LLMs on GSM8K has significantly improved in recent years, it remains unclear whether their mathematical reasoning capabilities have genuinely advanced, raising questions about the reliability of the reported metrics. To address these concerns, we conduct a large-scale study on several state-of-the-art open and closed models. To overcome the limitations of existing evaluations, we introduce GSM-Symbolic, an improved benchmark created from symbolic templates that allow for the generation of a diverse set of questions. GSM-Symbolic enables more controllable evaluations, providing key insights and more reliable metrics for measuring the reasoning capabilities of models.Our findings reveal that LLMs exhibit noticeable variance when responding to different instantiations of the same question. Specifically, the performance of all models declines when only the numerical values in the question are altered in the GSM-Symbolic benchmark. Furthermore, we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data. When we add a single clause that appears relevant to the question, we observe significant performance drops (up to 65%) across all state-of-the-art models, even though the added clause does not contribute to the reasoning chain needed to reach the final answer. Overall, our work provides a more nuanced understanding of LLMs' capabilities and limitations in mathematical reasoning.

Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks

Zhaofeng Wu **Linlu Qiu** **Alexis Ross** **Ekin Akyürek** **Boyuan Chen**
Bailin Wang **Najoung Kim** **Jacob Andreas** **Yoon Kim**
MIT Boston University
zfw@csail.mit.edu

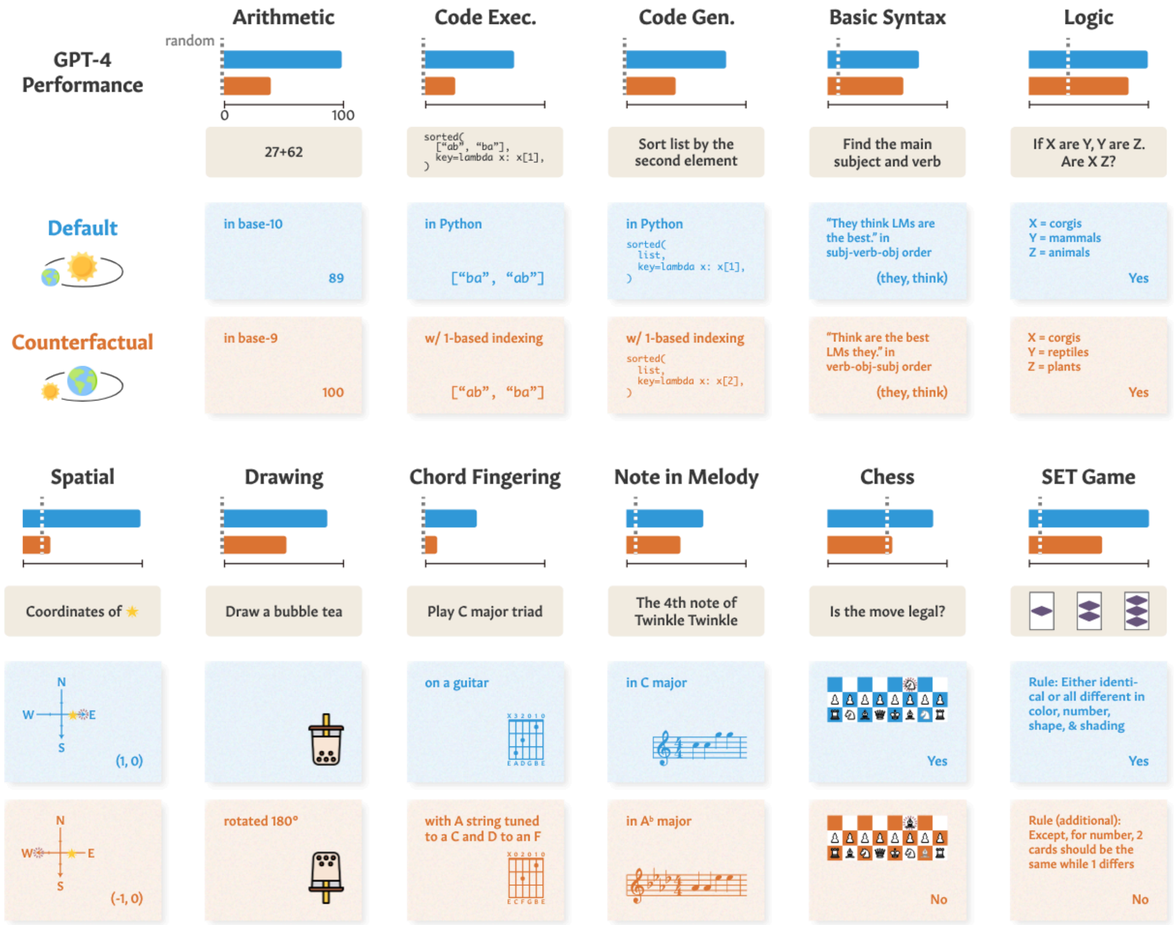
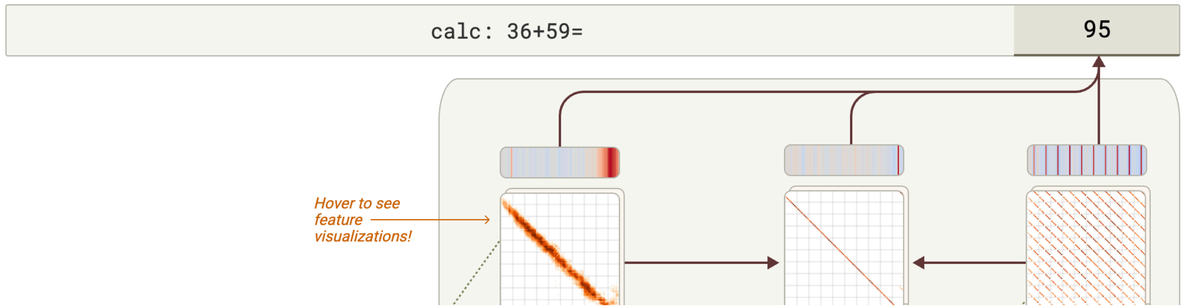


Figure 1: GPT-4's performance on the default version of various tasks (blue) and counterfactual counterparts (orange). The shown results use 0-shot chain-of-thought prompting (§4; Kojima et al., 2023). GPT-4 consistently and substantially underperforms on counterfactual variants compared to default task instantiations.

On the Biology of a Large Language Model

We investigate the internal mechanisms used by Claude 3.5 Haiku — Anthropic's lightweight production model — in a variety of contexts, using our circuit tracing methodology.



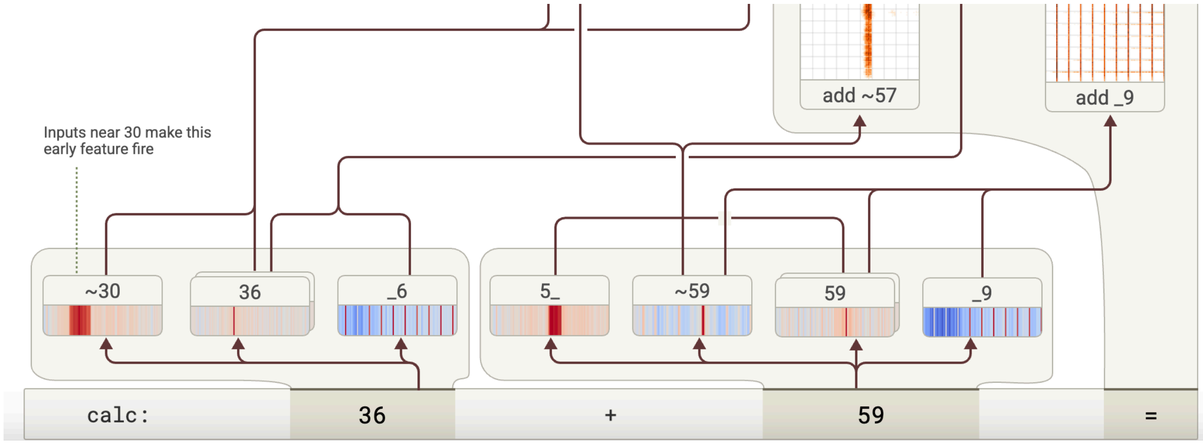
Sum Features

The model has finally computed information about the sum: its value mod 10, mod 100, and its approximate magnitude.

ARITHMETIC WITHOUT ALGORITHMS: LANGUAGE MODELS SOLVE MATH WITH A BAG OF HEURISTICS

Yaniv Nikankin[1]* **Anja Reusch**[1] **Aaron Mueller**[1,2] **Yonatan Belinkov**[1]

[1]Technion – Israel Institute of Technology [2]Northeastern University



approximate magnitude. Operand plots show vertical or horizontal stripes.

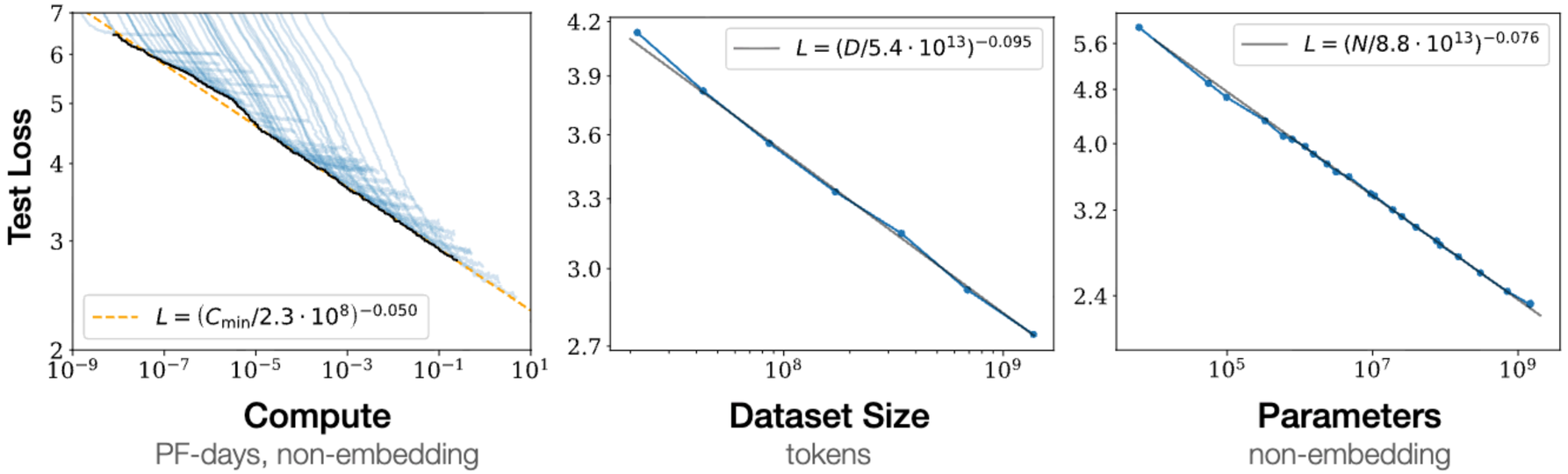
Input Features

The model has features specific to the ones digit and to the approximate magnitude, at various scales.

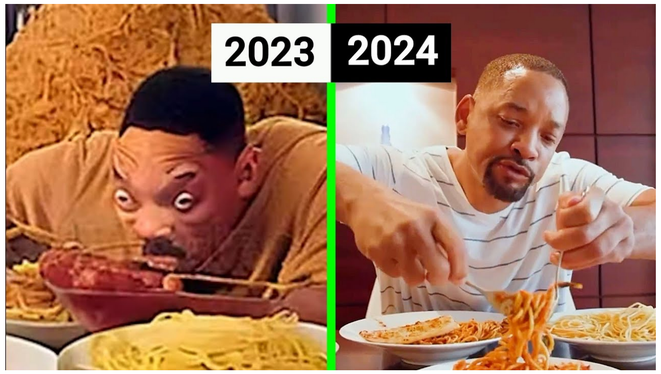
Most computation takes place on the "=" token

Scaling helps… but in what way?

Scaling Laws for Neural Language Models



Kaplan et al. (2020)

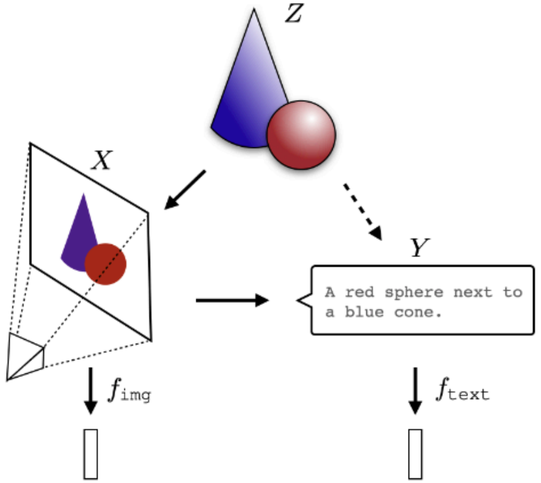


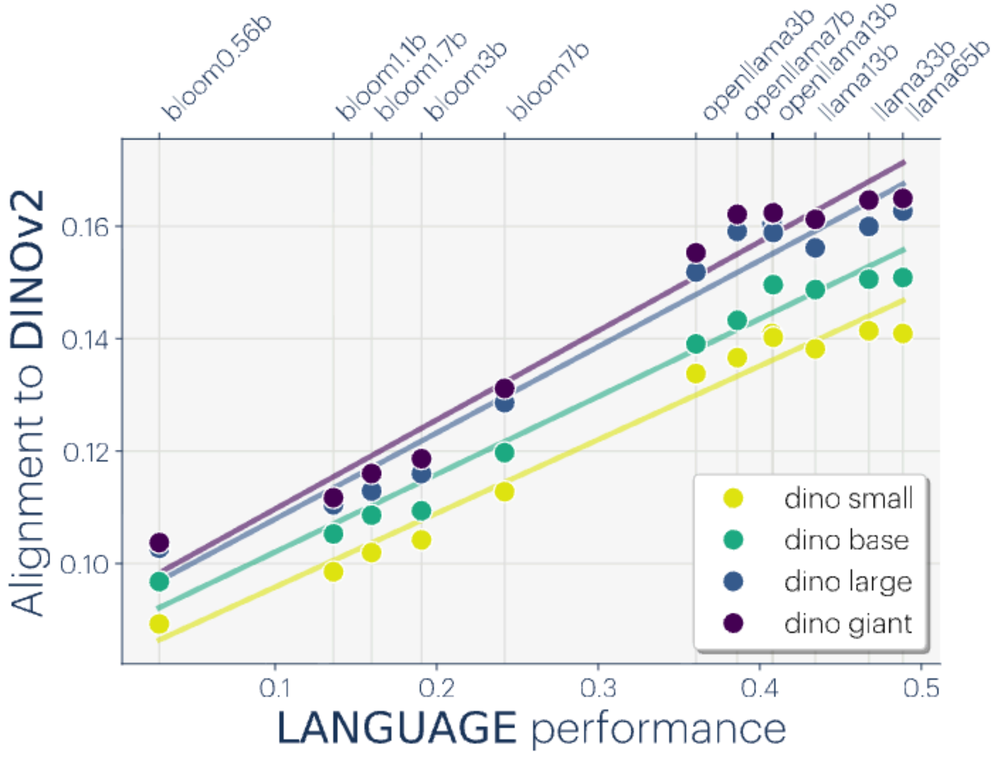
Scaling helps… but in what way?

Platonic Representation Hypothesis

The Platonic Representation Hypothesis

Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.





Hypothesis space

Huh et al. (2020)

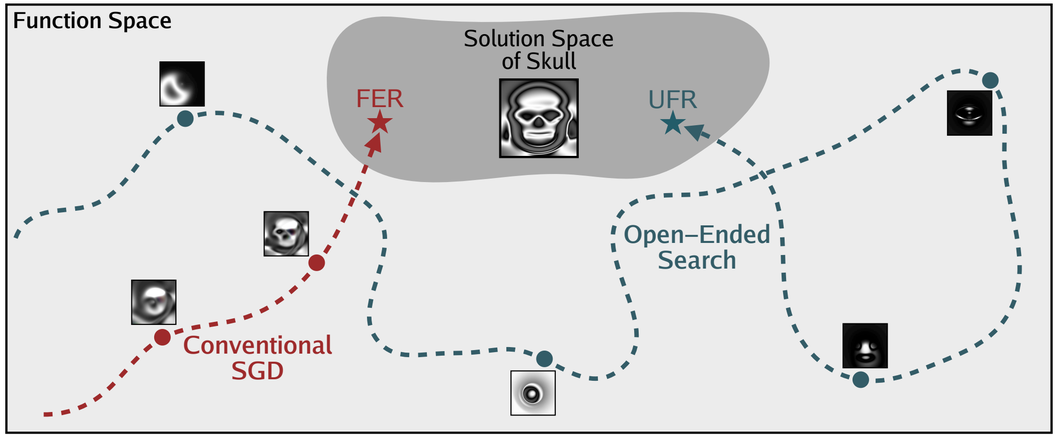
Scaling helps… but in what way?

- **Statistical:** Scaling laws are **statistical** observations
  - How does it relate to **regularities**?
- **Efficiency:** we don’t have infinite data
  - Is 10T tokens enough?
- **Practicality:** Why do our LLMs still have **“jagged intelligence”?**
  - Can get IMO gold, but can’t reliably book a hotel!

- Deep learning is a **data–driven** learning paradigm
- Does there exist a more efficient **regularity–driven** learning paradigm?

What could be better?

- **Complexification** (ex: morphogenesis, etc.)
  - Builds regularities on top of other regularities (bottom up)
  - Emergence
- **Adaptability**
  - Pressures the learned regularities to be robust to environmental changes
  - Representation must capture axes of variation which "carve nature at its joints”
- **Serendipity** (order matters for learning!)
  - Much higher chance of finding a useful learning curriculum
- What learning paradigm captures all of these? **Open–Endedness!**



Is this a Platonic Intelligence?

Space of Forms

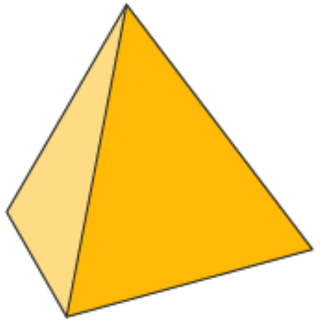
Real World

Intelligent Agents

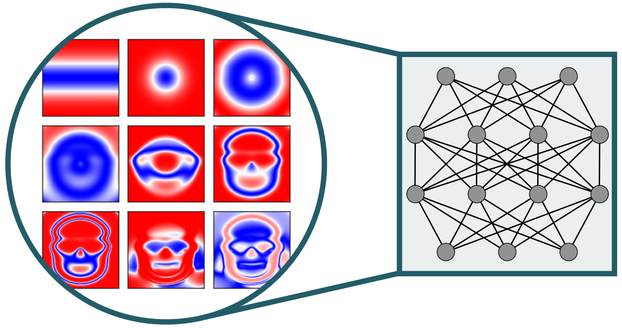
Unified Factored Representation

Is this a Platonic Intelligence?

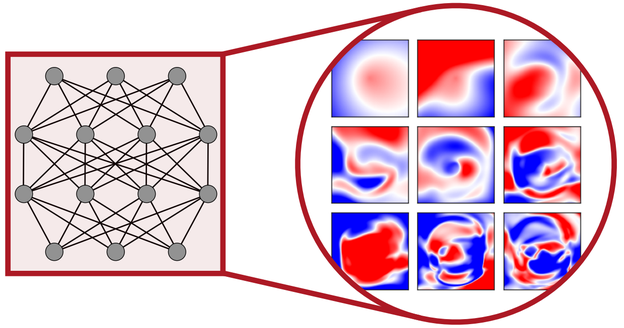
Aspirational Ideal



Instantiation



Unified Factored Representation



Fractured Entangled Representation

Collaborators

Jeff Clune
UBC
Vector Institute

Joel Lehman
University of Oxford

Kenneth Stanley
Lila Sciences


Thank You!