At the Interface of Genetics & Neural Networks

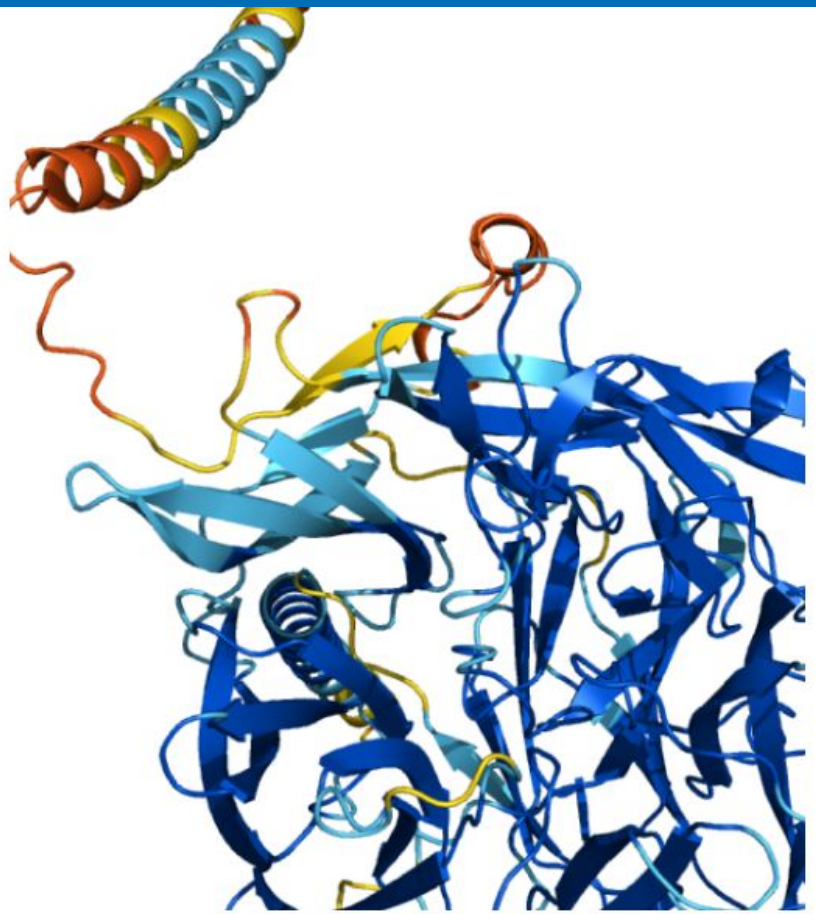
Okonda, Joseph L.

Introduction

AlphaFold

AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The database covers the complete human proteome (including fragments for long proteins) and the proteomes of 47 other key organisms (e.g. mouse), as well as the majority of manually curated UniProt entries (Swiss-Prot). In 2022 we plan to expand the database to cover a large proportion of all catalogued proteins (the over 100 million in UniRef90).

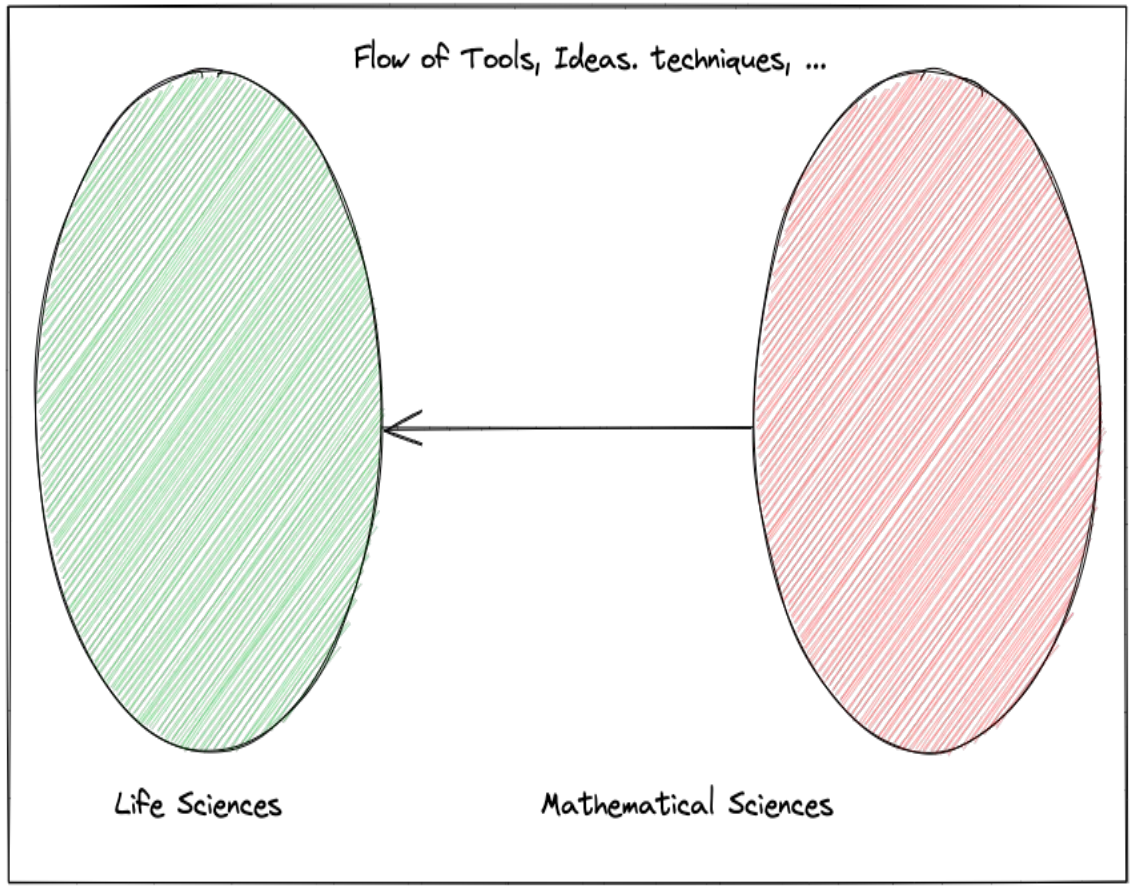


Q8I3H7: May protect the malaria parasite against attack by the immune system.
Mean pLDDT 85.57.

BROAD INSTITUTE

The current paradigm

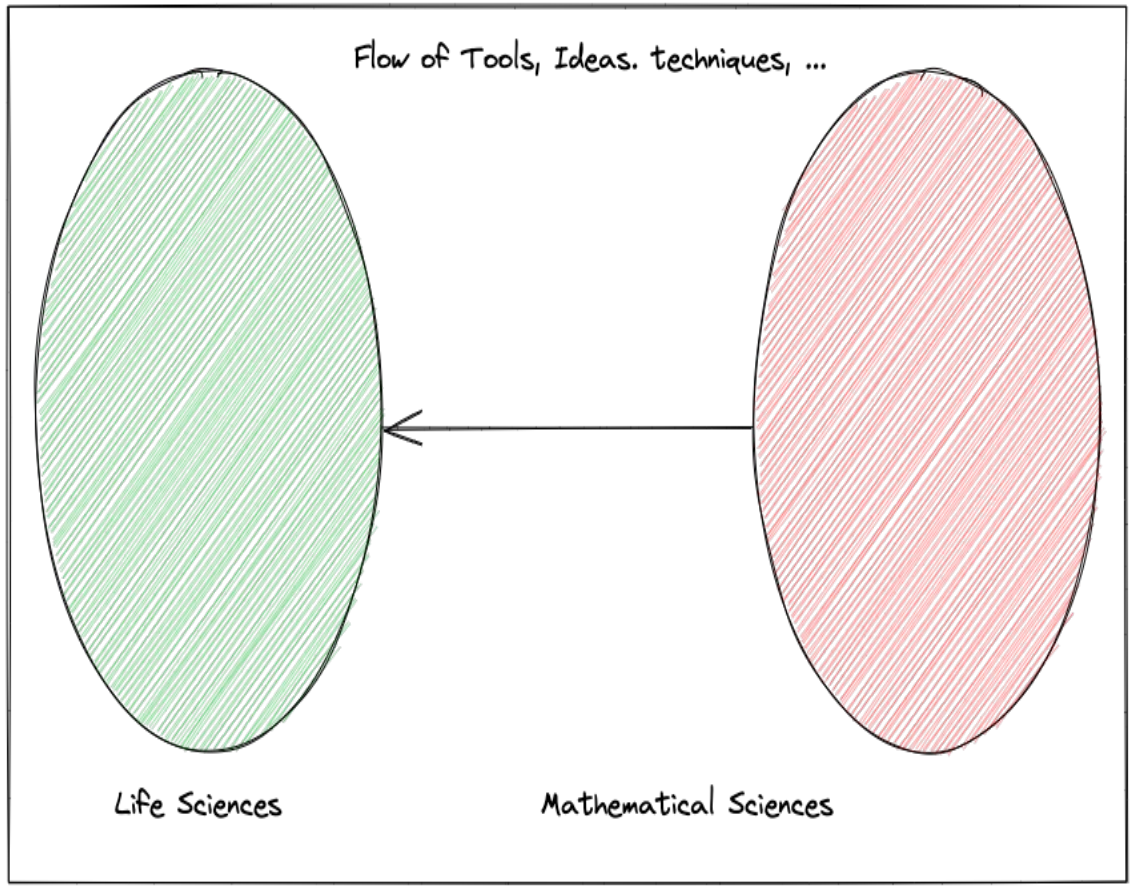
Traditional View



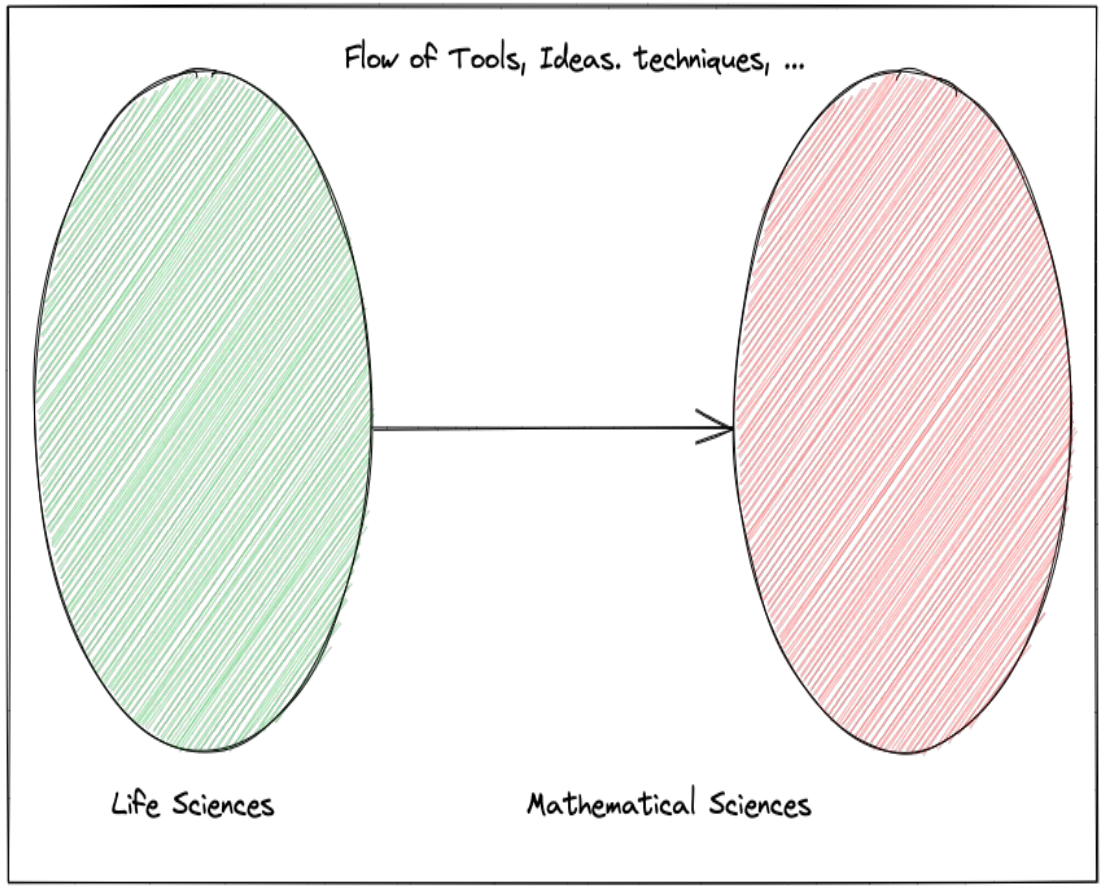
BROAD INSTITUTE

Is the reverse possible? Is it useful?

Traditional View



This Talk



BROAD INSTITUTE

Can *Biology* help us to *Understand* Neural Networks?

We can *repurpose experimental methods, and analysis tools* developed in **Genetics**, to understand cells and organisms, to *mechanistically explain the function of Neural Networks.*

BROAD INSTITUTE

Outline

1. What are NNs? How are they made?
2. What do we mean by "*understanding neural networks*?"
3. How can Biology help?
4. A rudimentary experiment.
5. What remains to be done?

BROAD INSTITUTE

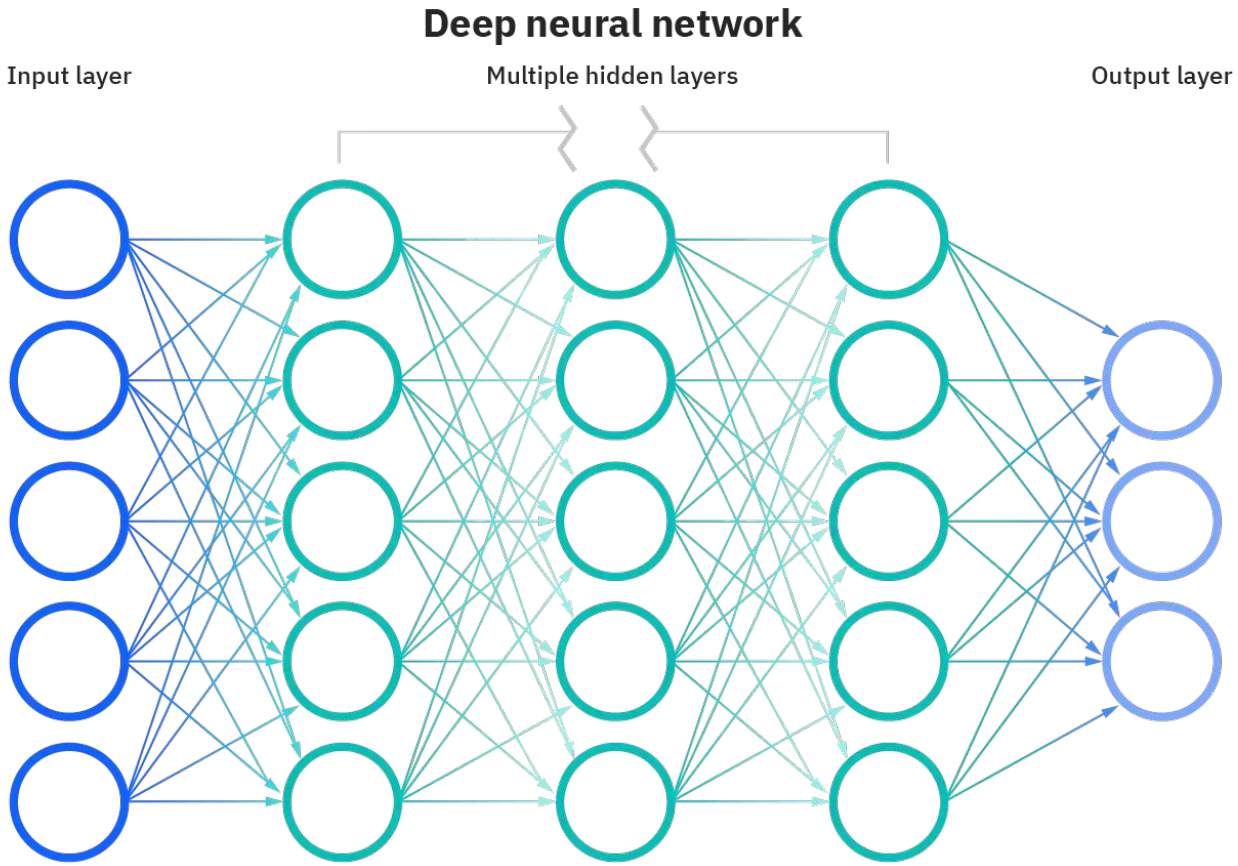
What are Neural Networks?

Computational Black Boxes

BROAD INSTITUTE

Neural Networks: The Network

- Basically a *collection of matrices* with numbers in them.



BROAD INSTITUTE

How are Neural Networks Made?

Computational Black Boxes

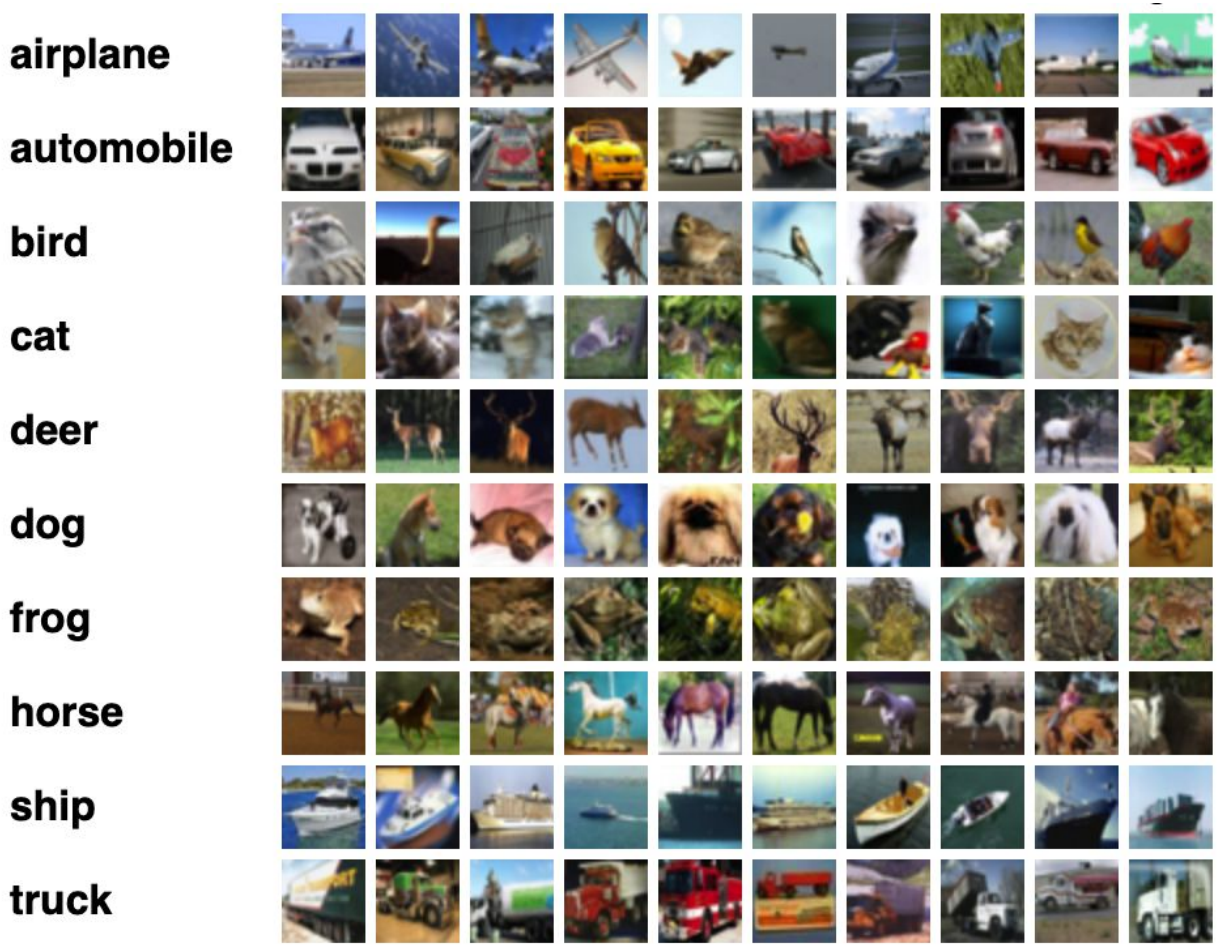
BROAD INSTITUTE

Neural Networks

- Training Data.
- Training/Optimization Process.
- Final Result: What do we end up with?

Neural Networks: Training Data

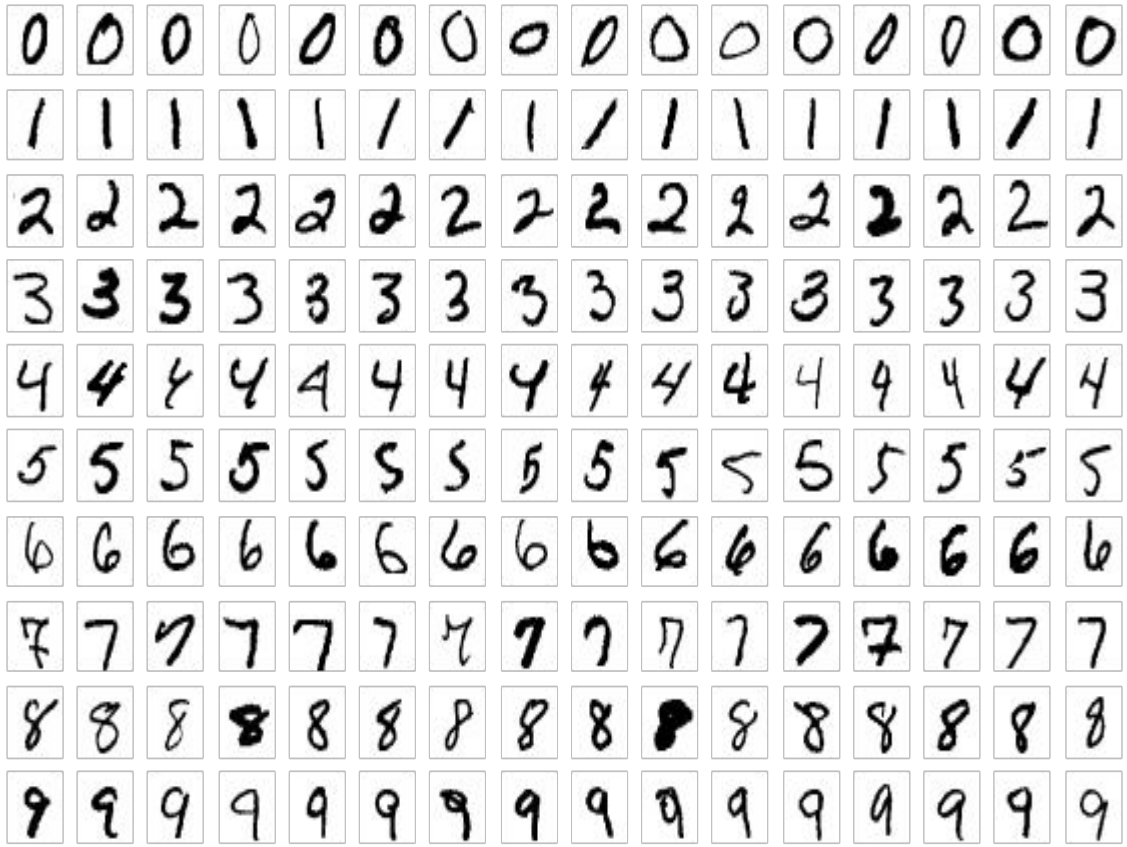
- Examples that we'll *use to tune the weights* of the network.
- *Weights* are the numbers in the matrices.



BROAD INSTITUTE

Neural Networks: Training Data

- Examples that we'll *use to tune the weights* of the network.
- *Weights* are the numbers in the matrices.



BROAD INSTITUTE

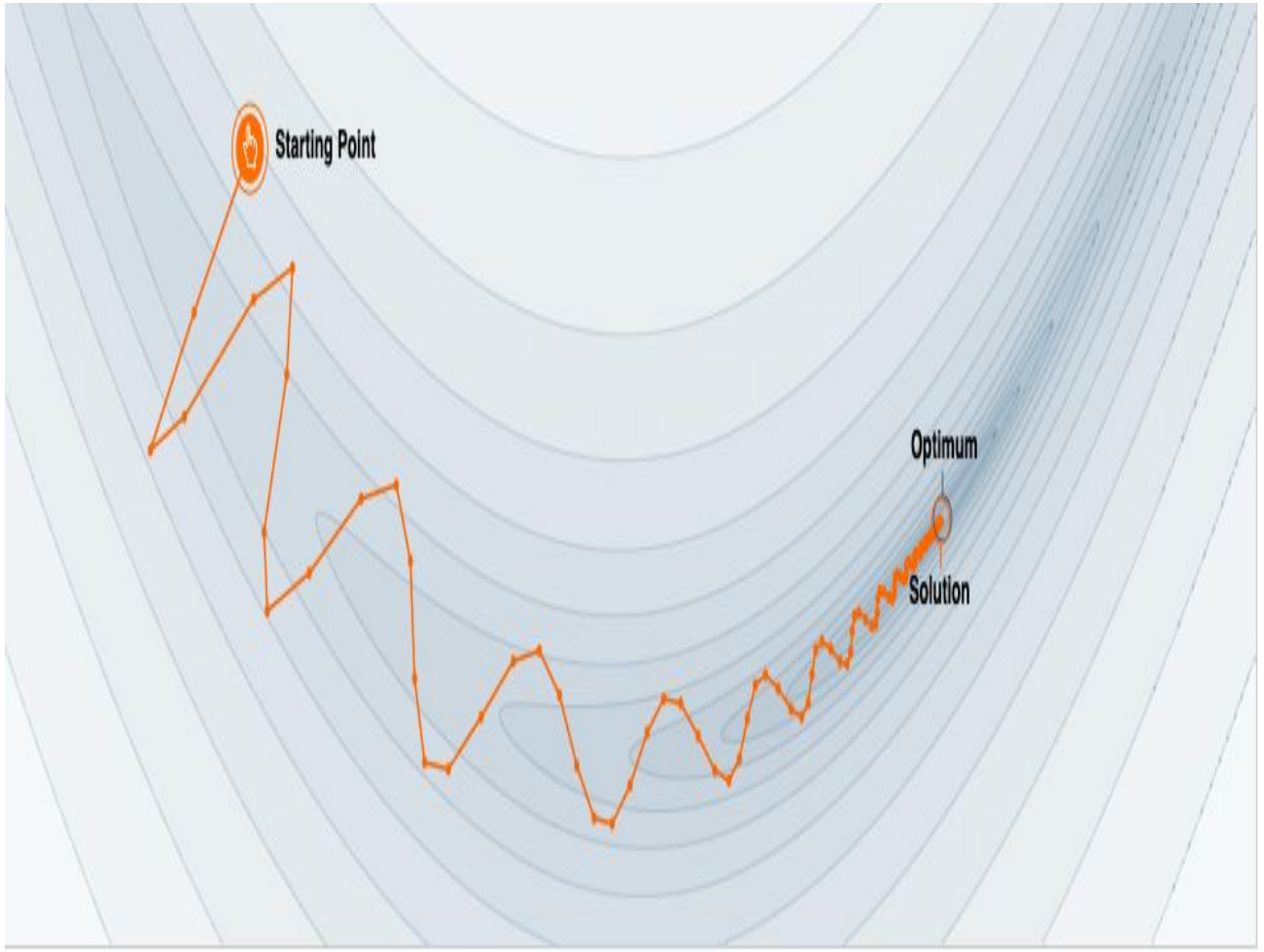
Neural Networks

- Training Data.
- Training/Optimization Process.
- Final Result: What do we end up with?

BROAD INSTITUTE

Neural Networks: Training Process

- We *iteratively* change the numbers in the matrices
- Stop when network is able to perform well on the task.



BROAD INSTITUTE

Neural Networks

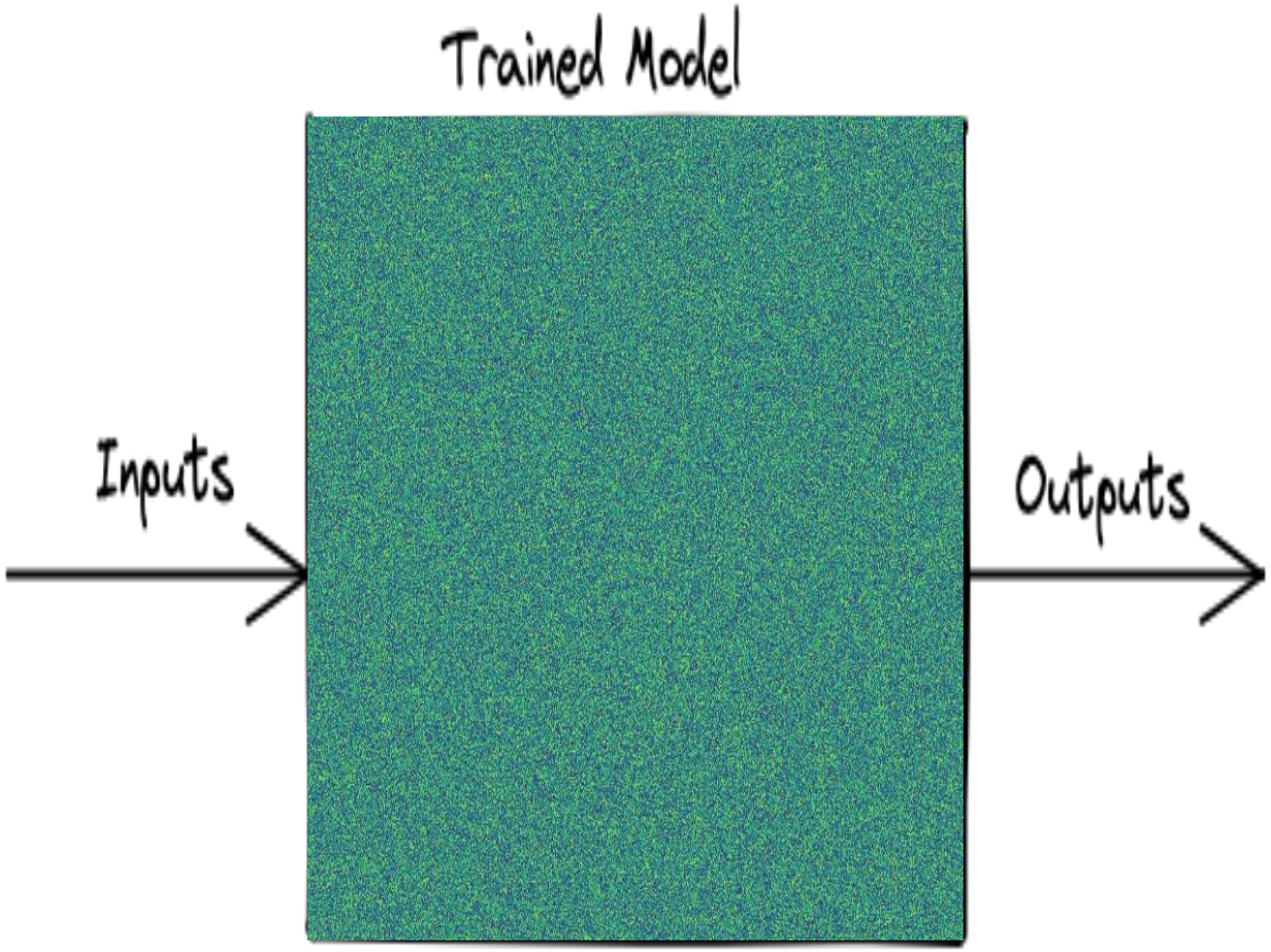
- Training Data.
- Training/Optimization Process.
- Final Result: What do we end up with?

BROAD INSTITUTE

Neural Networks: Final Result

- A Trained Network.
- A *computational black box*

  exhibiting some useful traits



BROAD INSTITUTE

Outline

1. What are NNs? How are they made?
2. What do we mean by "*understanding neural networks*?"
3. How can Biology help?
4. A rudimentary Experiment.
5. What remains to be done?

BROAD INSTITUTE

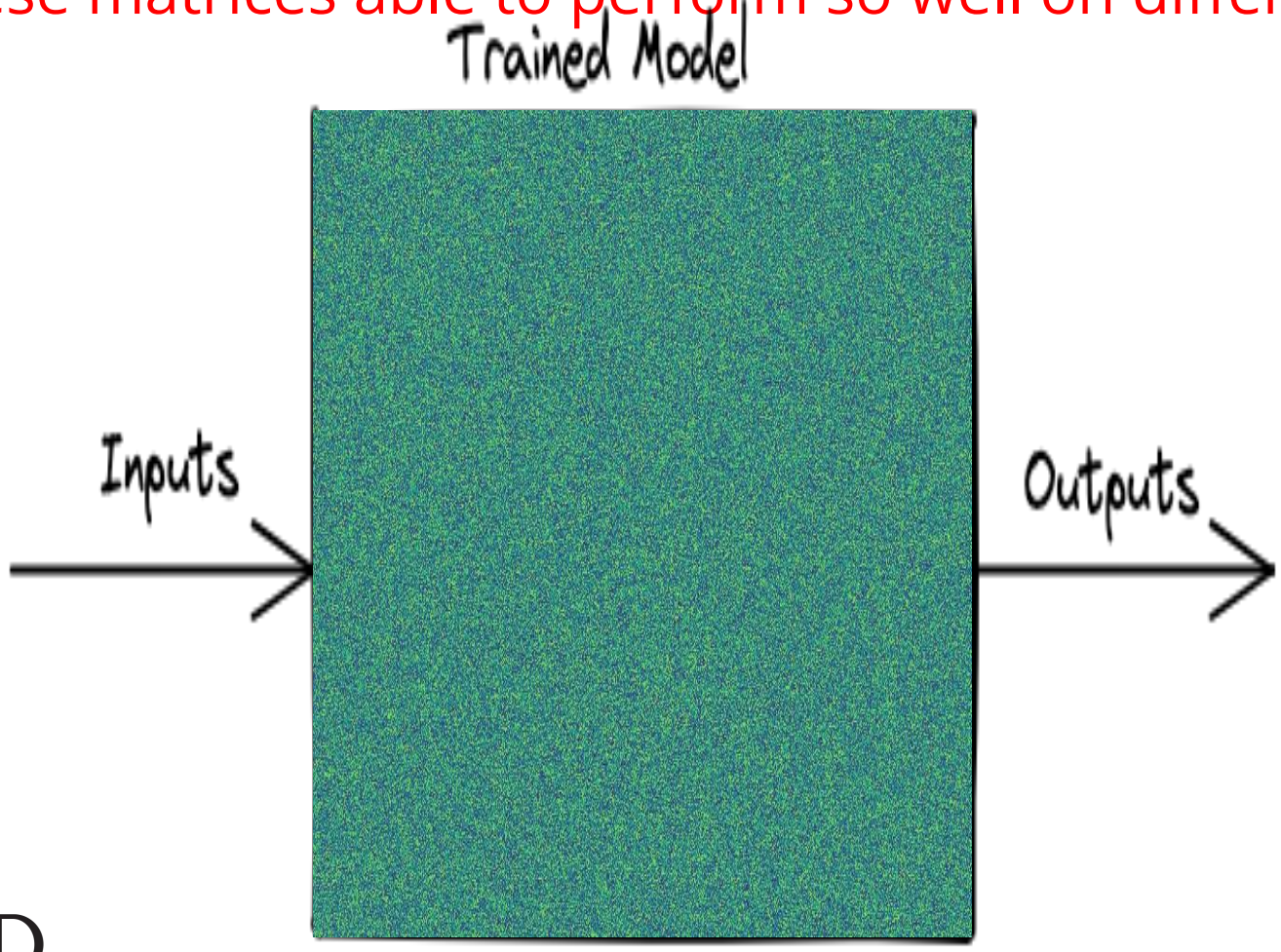
The Neural Network Interpretation Task

How are these matrices able to perform so well on different tasks?

BROAD INSTITUTE

The Neural Network Interpretation Task

How are these matrices able to perform so well on different tasks?

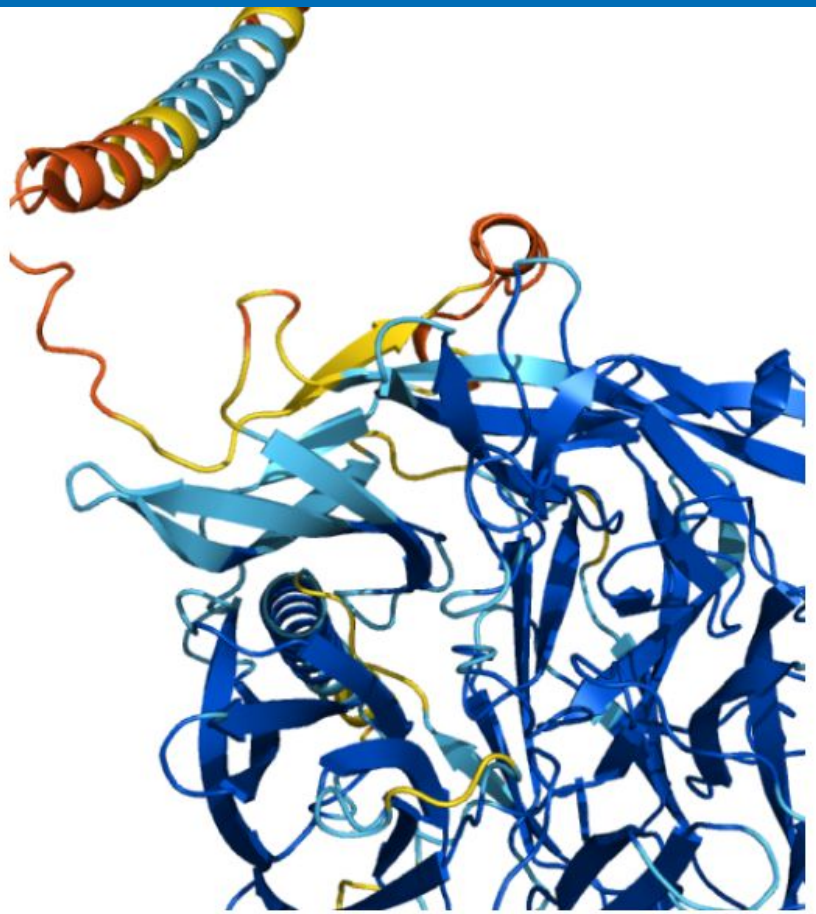


BROAD INSTITUTE

How does AlphaFold predict proteins' 3D structure?

AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The database covers the complete human proteome (including fragments for long proteins) and the proteomes of 47 other key organisms (e.g. mouse), as well as the majority of manually curated UniProt entries (Swiss-Prot). In 2022 we plan to expand the database to cover a large proportion of all catalogued proteins (the over 100 million in UniRef90).



Q8I3H7: May protect the malaria parasite against attack by the immune system.
Mean pLDDT 85.57.

BROAD INSTITUTE

Current Interpretation Techniques

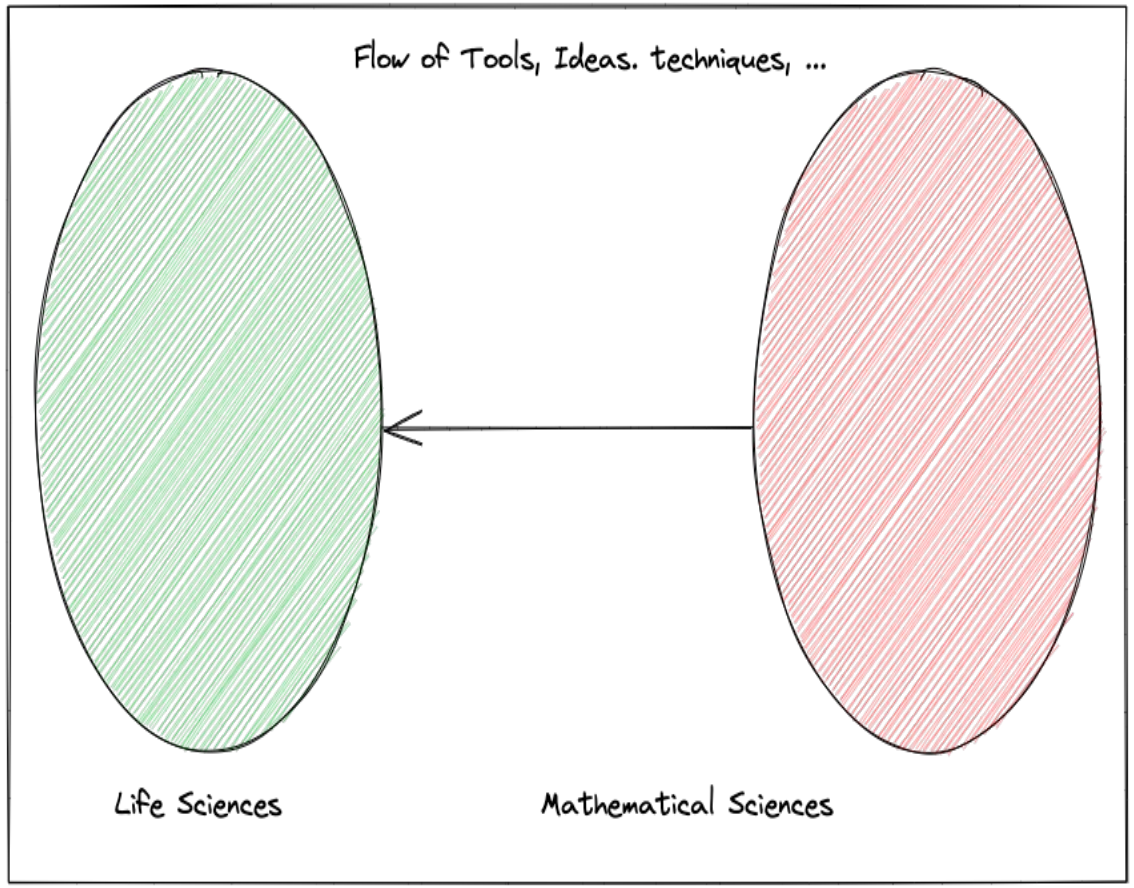
- Feature Attribution
- Feature Visualization
- *Probing the Networks Learned Representations*

1. https://distill.pub/2018/building-blocks/
2. https://distill.pub/2020/circuits/zoom-in/
3. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html

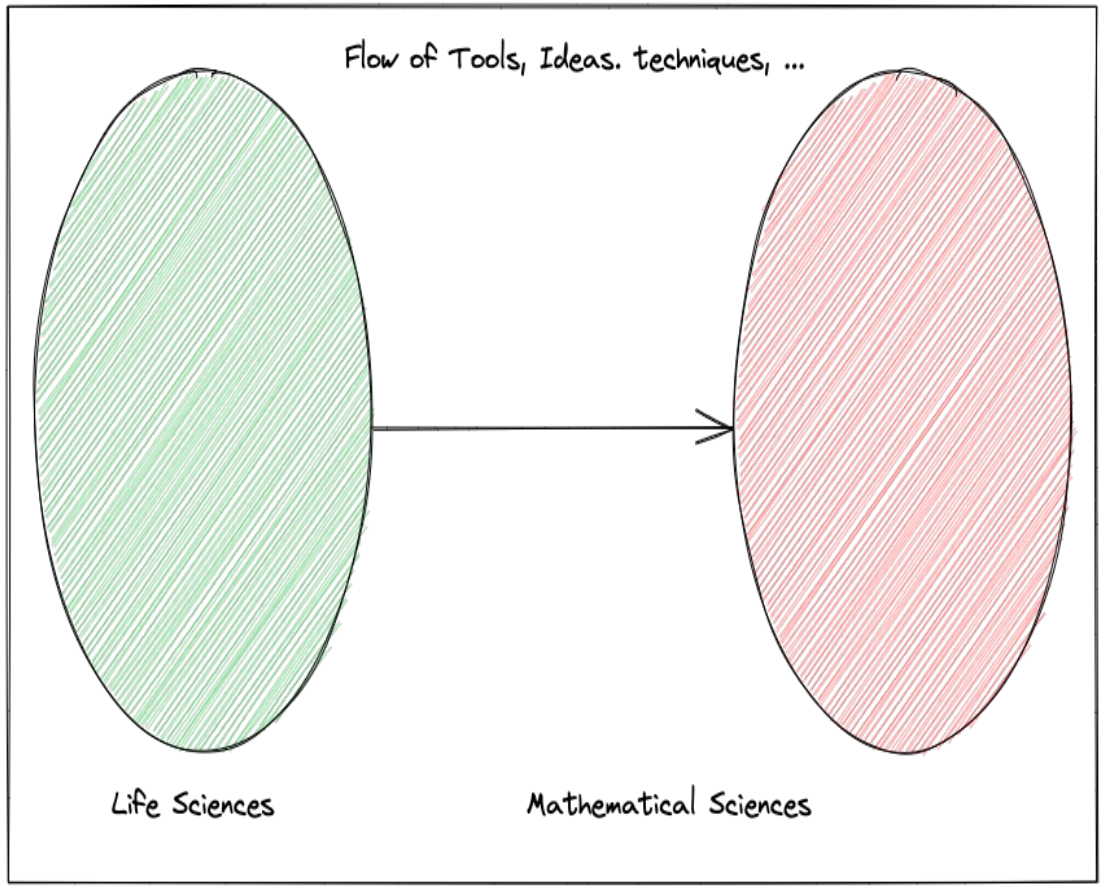
BROAD INSTITUTE

Can Biology help us Interpret Neural Networks?

Traditional View



This Talk



BROAD INSTITUTE

Outline

1. What are NNs? How are they made?
2. What do we mean by "*understanding neural networks*?"
3. How can Genetics help?
4. A rudimentary Experiment.
5. What remains to be done?

BROAD INSTITUTE

What is Genetics, really?

A general framework for mechanistically interpreting and understanding

biological black boxes

BROAD INSTITUTE

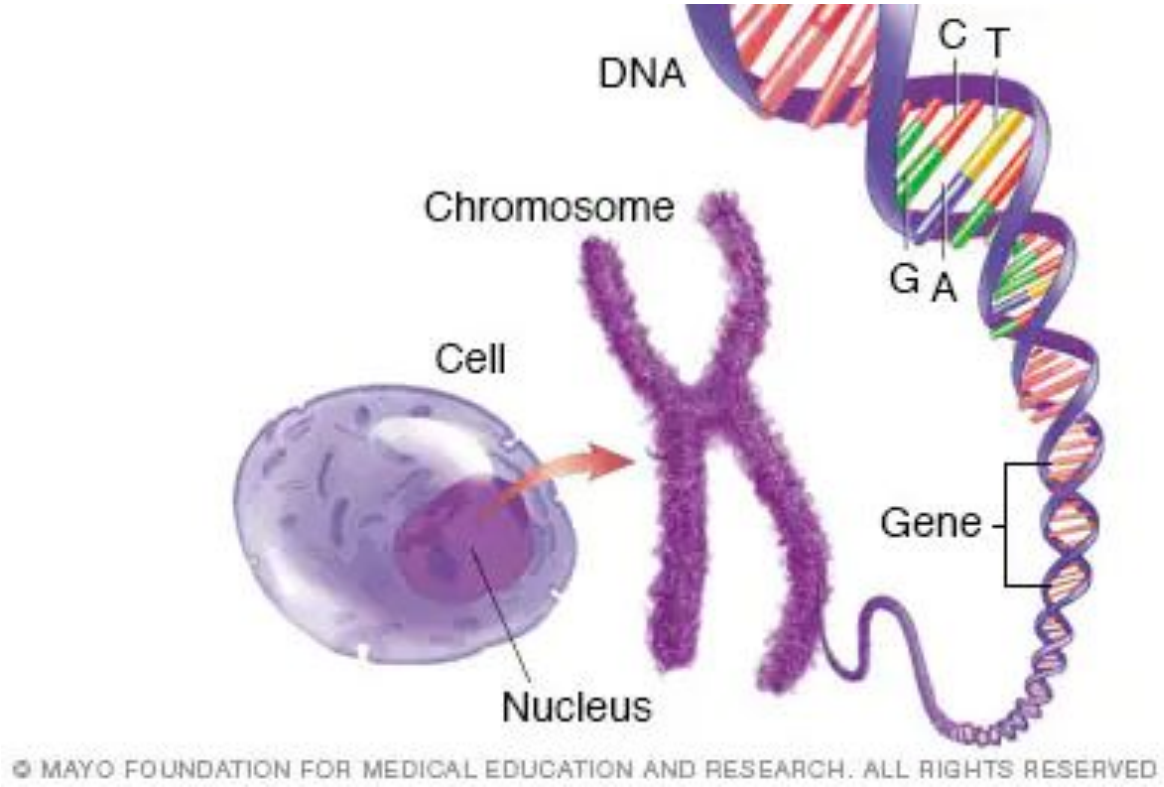
Genetics: Key Components in the Framework

- Phenotypic *Variation*

BROAD INSTITUTE

Genetics: Key Components in the Framework

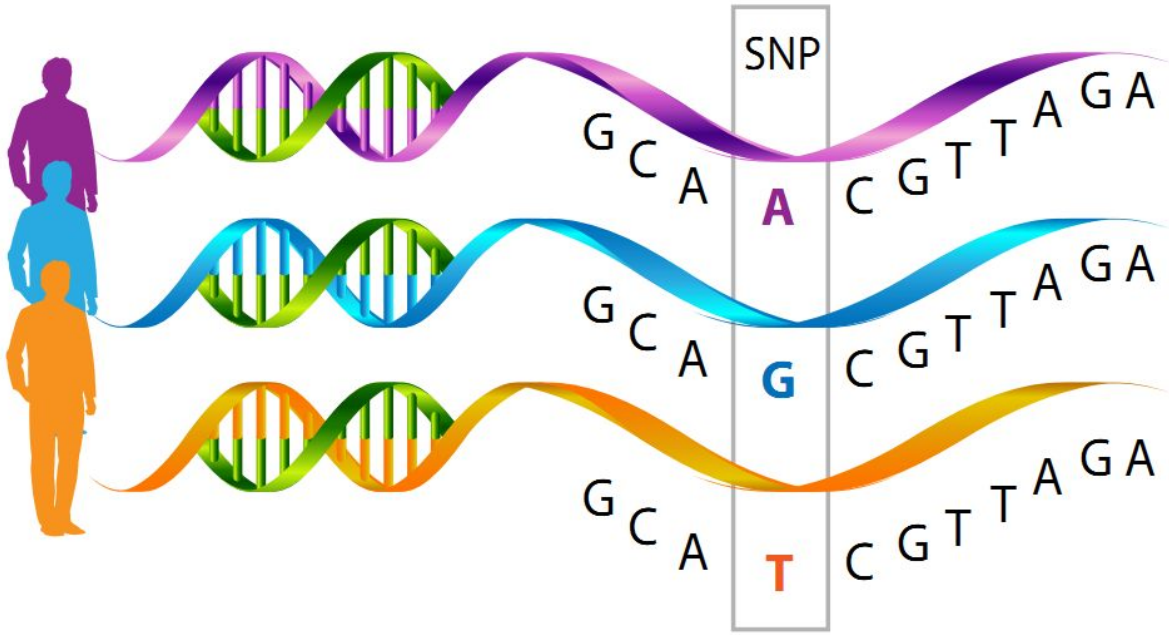
- Phenotypic *Variation*
- Genes: *functional* modules



BROAD INSTITUTE

Genetics: Key Components in the Framework

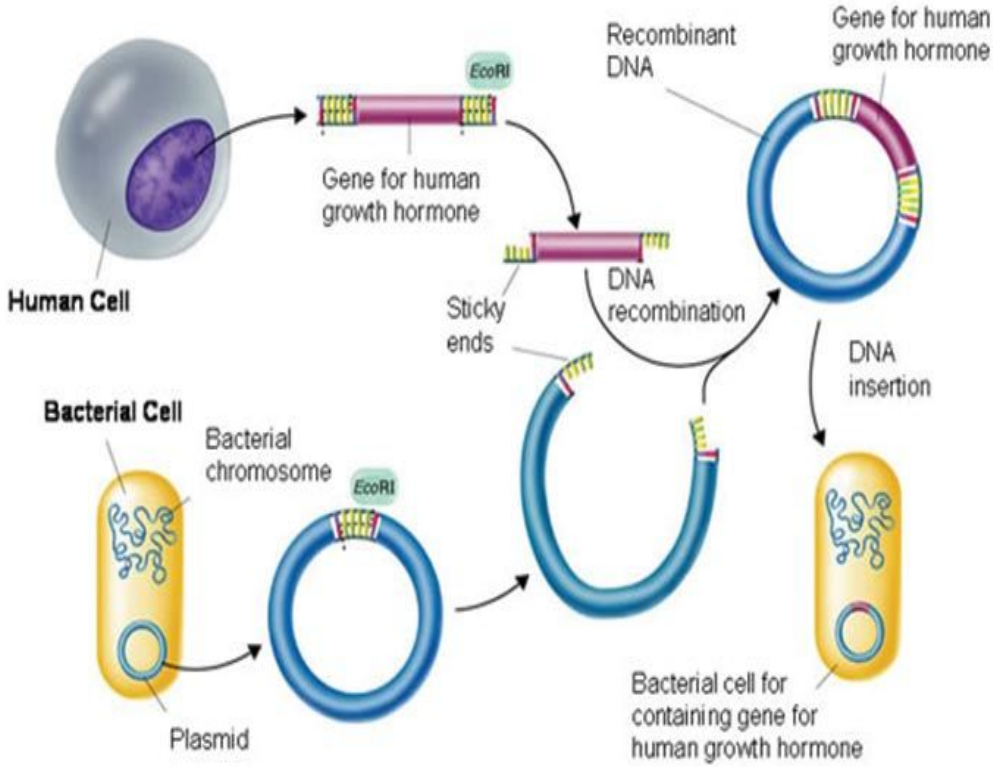
- Phenotypic *Variation*
- Genes: *functional* modules
- Genetic *Variation*



BROAD INSTITUTE

Genetics: Key Components in the Framework

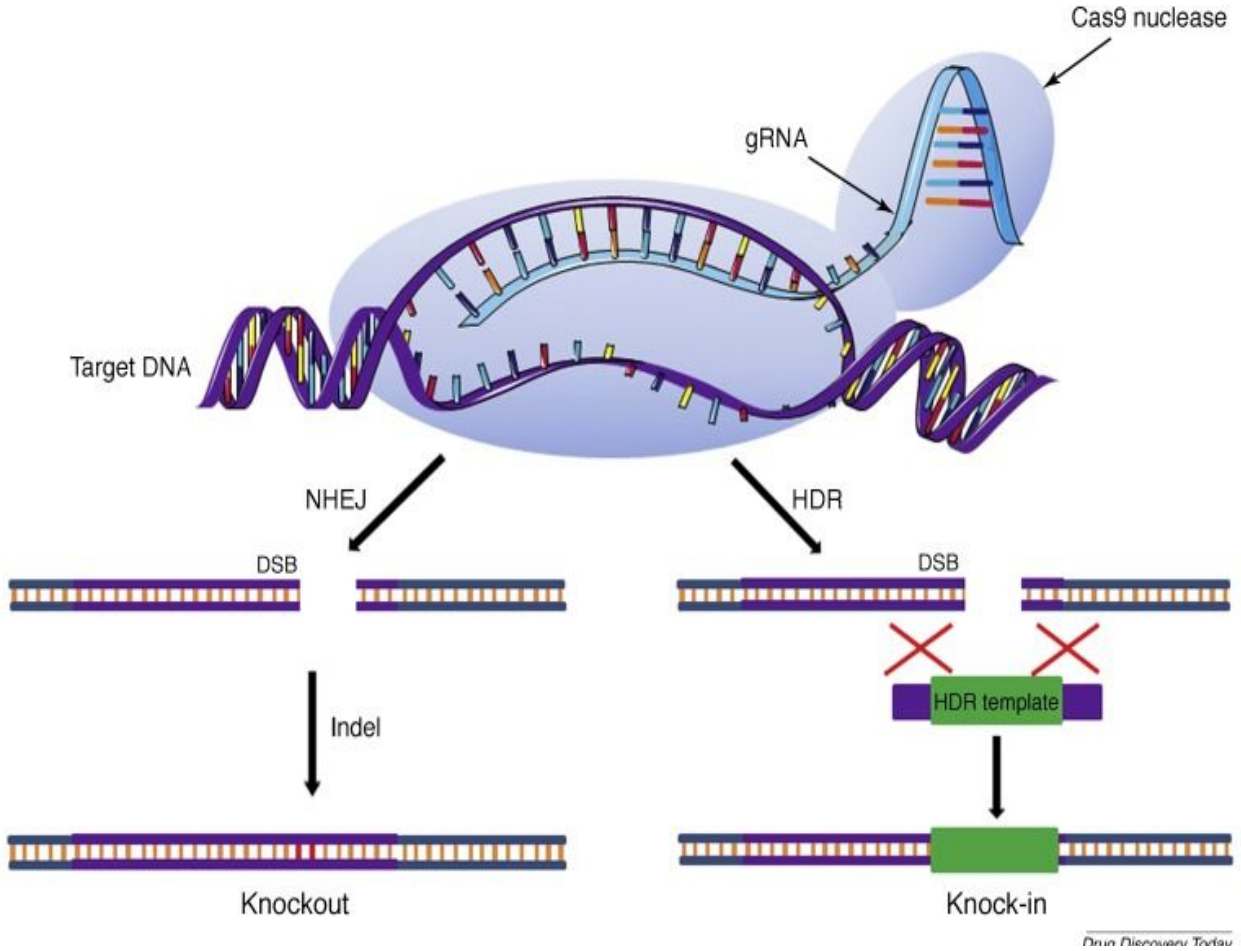
- Phenotypic *Variation*
- Genes: *functional* modules
- Genetic *Variation*
- Gene Cloning: *Isolating* functional modules.



BROAD INSTITUTE

Genetics: Key Components in the Framework

- Phenotypic *Variation*
- Genes: *functional* modules
- Genetic *Variation*
- Gene Cloning: *Isolating* functional modules.
- Loss & Gain of Function Screening



BROAD INSTITUTE

Genetics: Key Components in the Framework

- Phenotypic *Variation*
- Genes: *functional* modules
- Genetic *Variation*
- Gene Cloning: *Isolating* functional modules.
- Loss & Gain of Function Screening

BROAD INSTITUTE

Neural Networks + Genetics

A general framework for mechanistically interpreting and understanding **computational black boxes?**

BROAD INSTITUTE

Neural Networks + Genetics: Linking Questions

1. Do NN weights contain functional modules?
2. Are these modules associated with NN traits?
3. Can we isolate said modules?

Neural Networks + Genetics

Can we draw inspiration from Genetics to design experiments to answer these questions?

BROAD INSTITUTE

At the Interface of Genetics & Neural Networks

At the Interface

- Experimental Setup
- Rudimentary Analysis
- Remaining Work

BROAD INSTITUTE

At the Interface

- Experimental Setup
- Rudimentary Analysis
- Remaining Work

BROAD INSTITUTE

At the Interface: Experimental Setup

- **In Biology:** *Phenotypic Variation* helps us uncover the existence and function of functional modules (genes).
- **Starting Point:** Generate *a population of neural networks* with *Variation* in their traits.

BROAD INSTITUTE

At the Interface: Experimental Setup

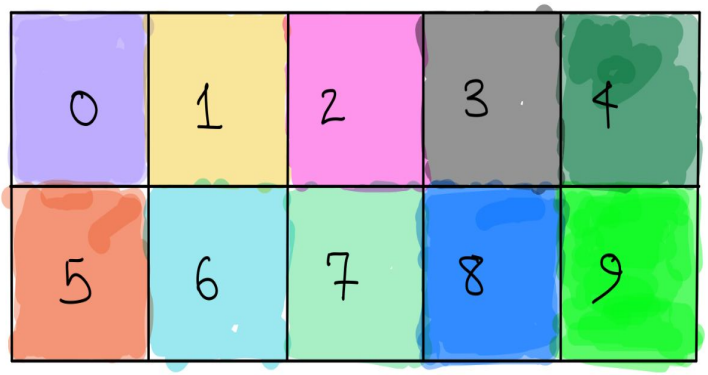
- Train on MNIST.
- *Generate Variation* by hiding some

  of the classes.

BROAD INSTITUTE

At the Interface: Experimental Setup

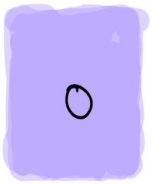
- We'll call a network **Wild** if it was trained on all the training data.
- A **wild** model has seen all classes

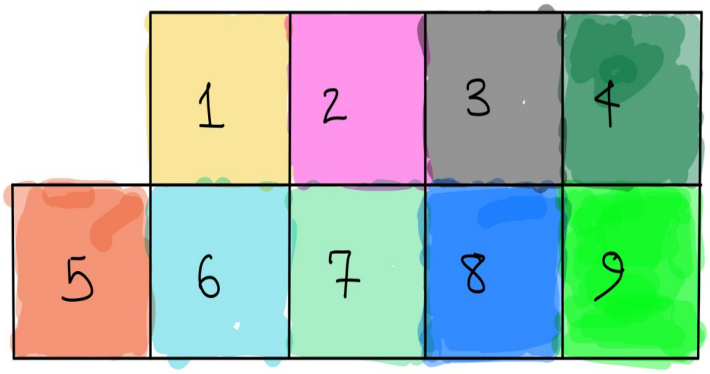


BROAD INSTITUTE

At the Interface: Experimental Setup

- We'll call a network **Mutant^{X}** if it was trained on all the training data except data for the class **X**.
- A **Mutant^{X}** model has seen all classes except **X**

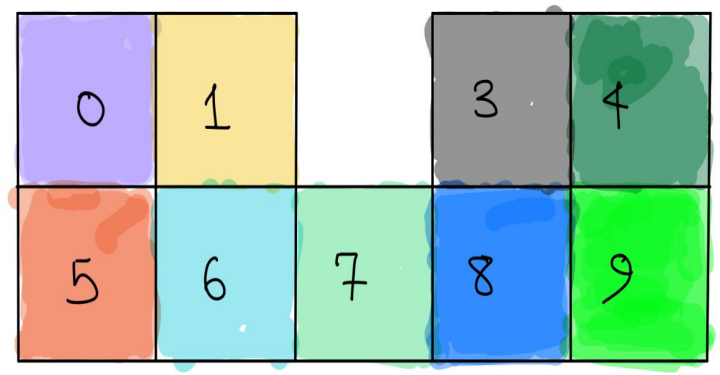




BROAD INSTITUTE

At the Interface: Experimental Setup

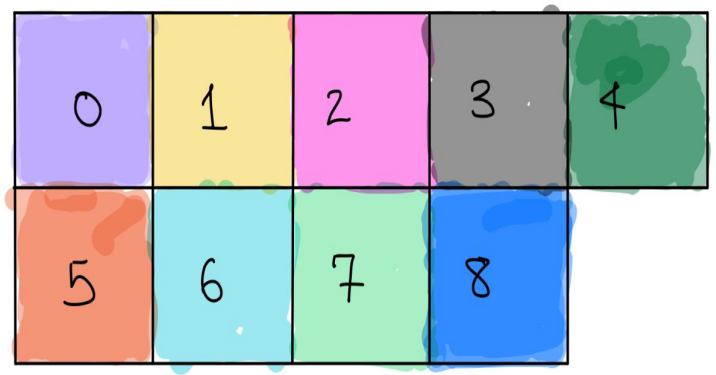
- We'll call a network **Mutant^{X}** if it was trained on all the training data except data for the class **X**.
- A **Mutant^{X}** model has seen all classes except **X**



BROAD INSTITUTE

At the Interface: Experimental Setup

- We'll call a network **Mutant**^{X} if it was trained on all the training data except data for the class **X**.
- A **Mutant**^{X} model has seen all classes except **X**



BROAD INSTITUTE

At the Interface: The Difference Model

- The **Diff^{X}** Model is the model that results if we subtract a **Mutant^{X}** from a **Wild** model.
- ***Diff^{X} = Wild - Mutant^{X}***

BROAD INSTITUTE

At the Interface: Experimental Summary

- For each of the 10 MNIST Classes, we'll train a **Mutant** model.
- We'll also train 1 **Wild** model.
- From this, we'll generate 10 **Diff** models.
- We can repeat this procedure from different seeds (starting points) to generate a population of models.

BROAD INSTITUTE

At the Interface: Evaluating the Models

- We used 5 seeds to generate a total of **50 Mutant** Models, **5 Wild** Models and **50 Diff** Models.
- We evaluate each model the dataset with only a single class, (**X-Only**) for all 10 classes.

BROAD INSTITUTE

At the Interface

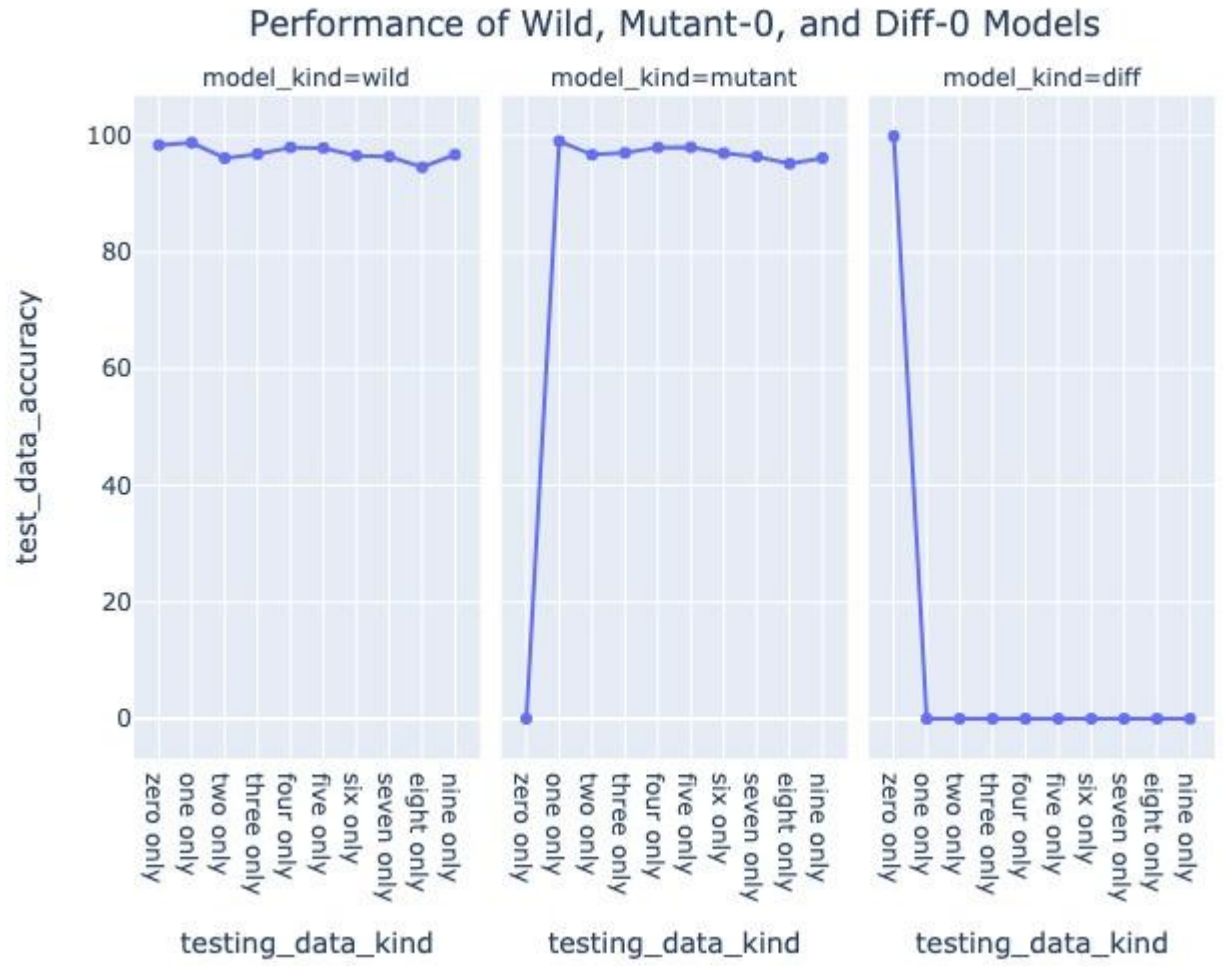
- Experimental Setup
- Rudimentary Analysis
- Remaining Work

BROAD INSTITUTE

How do the test accuracies of the different models compare?

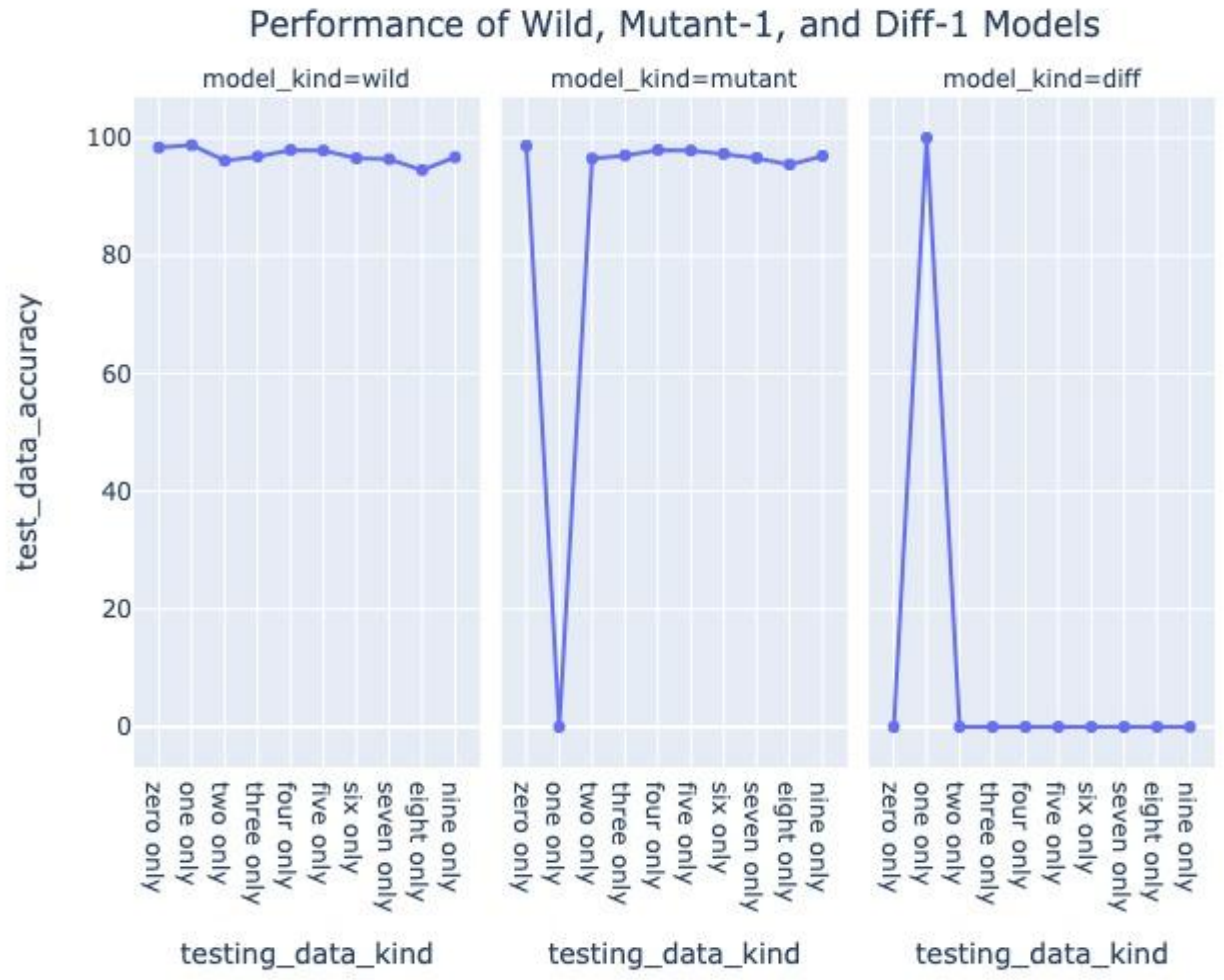
BROAD INSTITUTE

Rudimentary Analysis: Model Accuracy



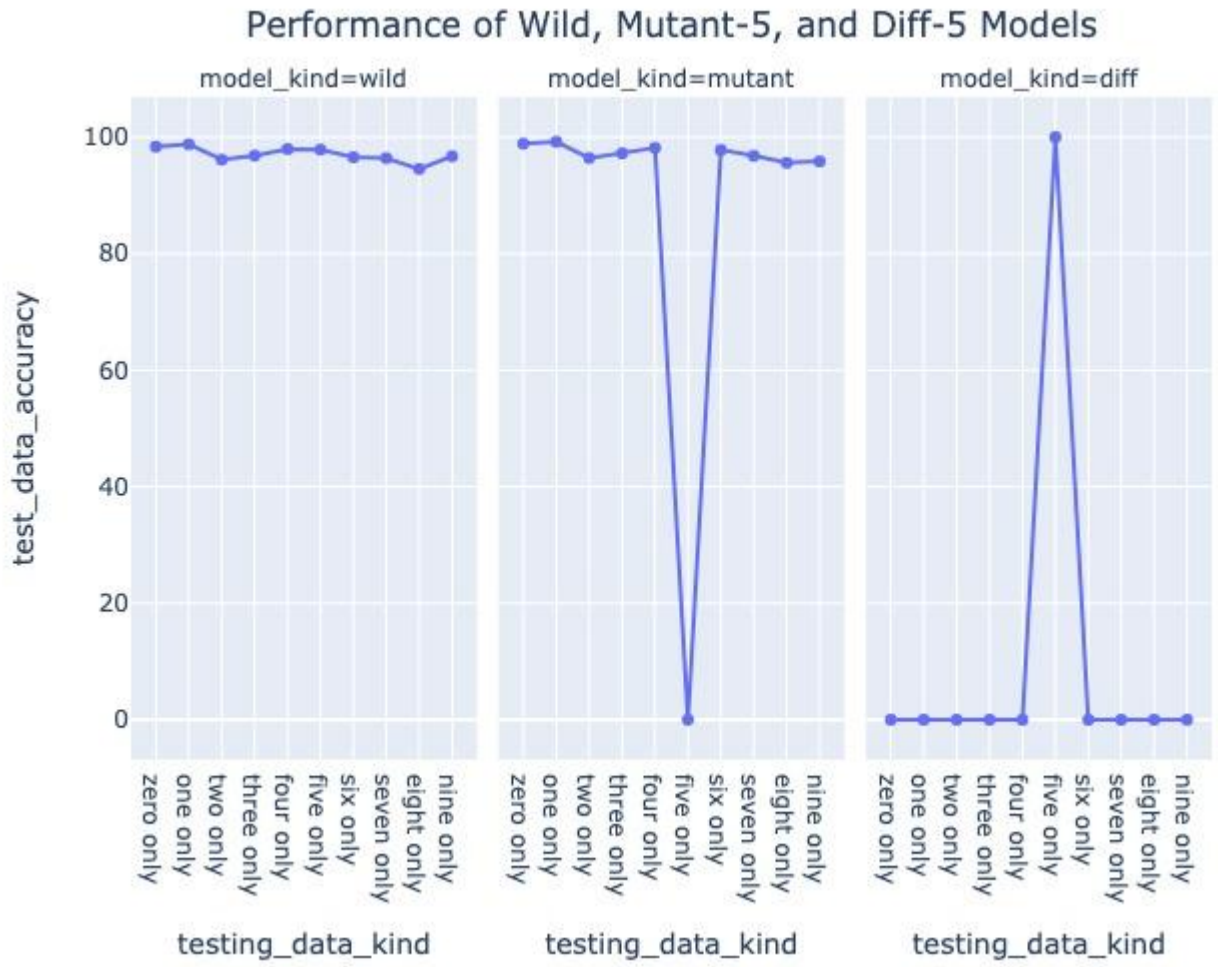
BROAD INSTITUTE

Rudimentary Analysis: Model Accuracy



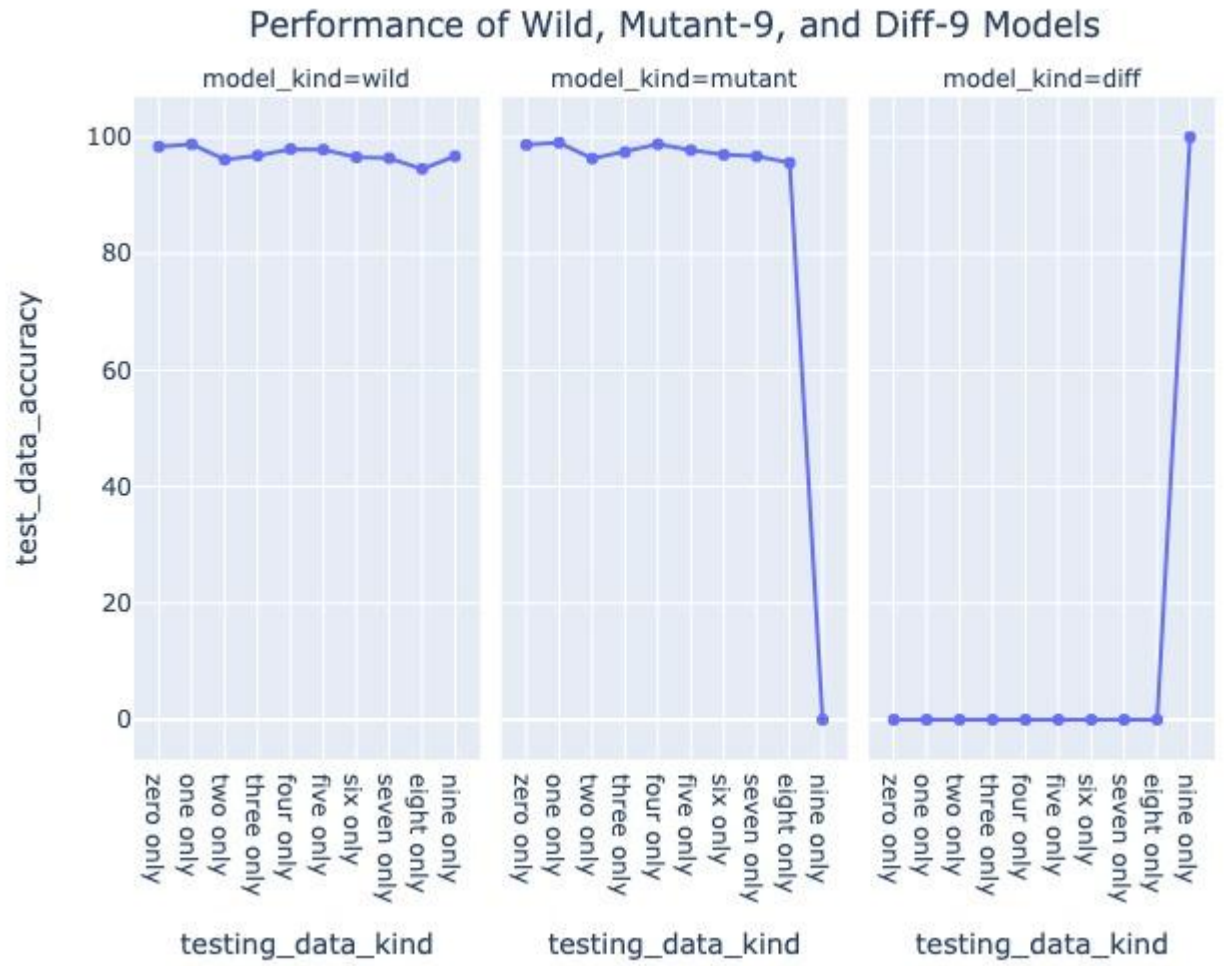
BROAD INSTITUTE

Rudimentary Analysis: Model Accuracy



BROAD INSTITUTE

Rudimentary Analysis: Model Accuracy



BROAD INSTITUTE

How does the test accuracy of the different models compare?

- The *Difference Model* is, apparently, only able to identify a single class. **Why**?

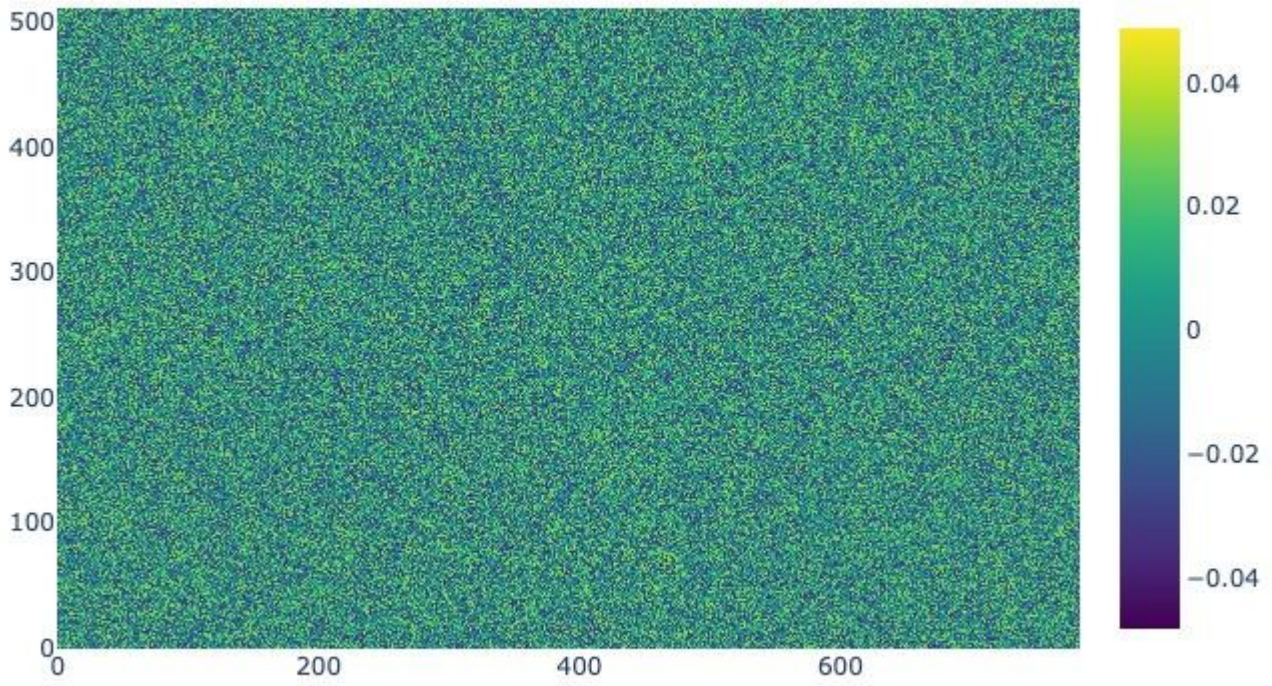
- Does subtracting the weights of the mutant model from those of wild model *isolate the functional modules responsible for a given trait*?

BROAD INSTITUTE

Do the weights of the models have any structure?

Do the weights of the models have any structure?

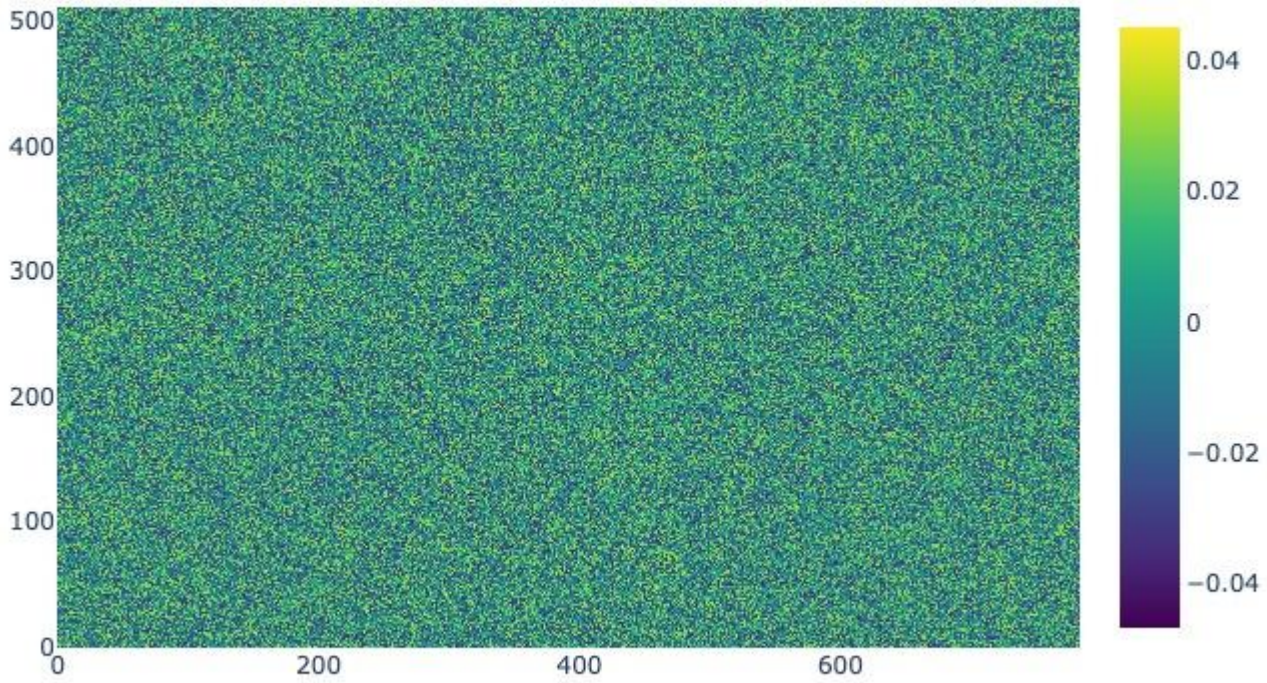
Heatmap of the First Layer of Wild Model



BROAD INSTITUTE

Do the weights of the models have any structure?

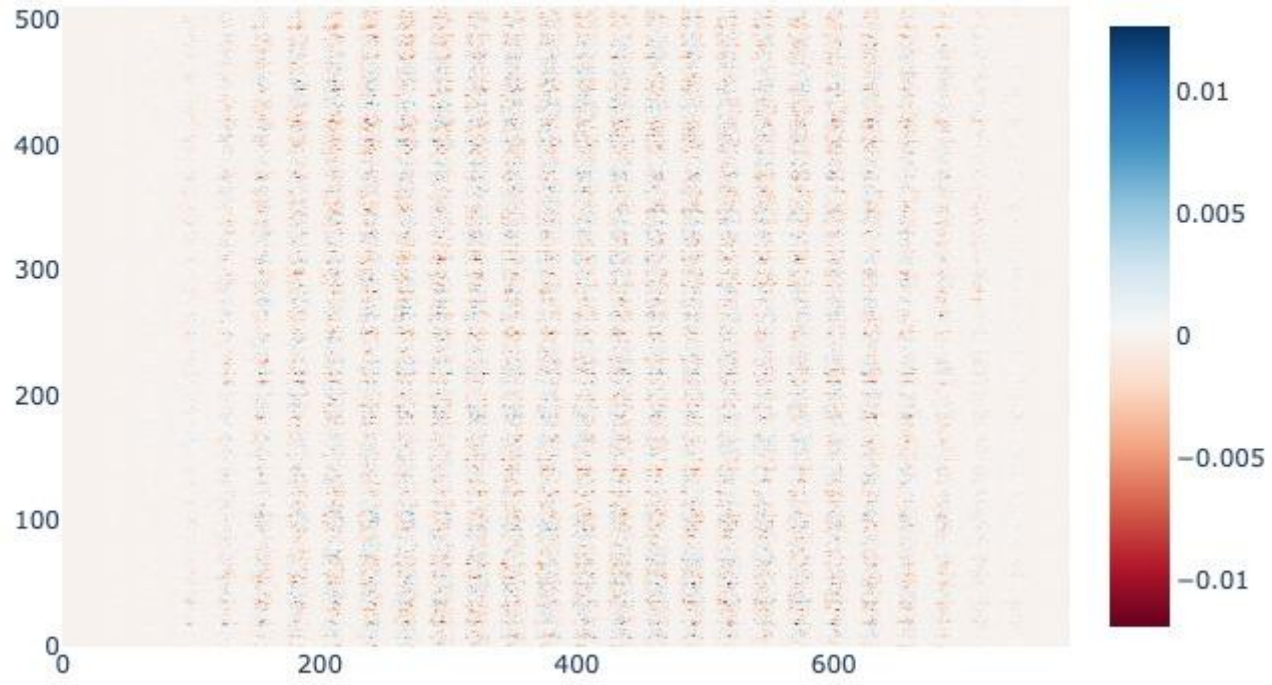
Heatmap of the First Layer of Mutant Model



BROAD INSTITUTE

Do the weights of the models have any structure?

Heatmap of the First Layer of Diff Model



BROAD INSTITUTE

Do the weights of the models have any structure?

- At first glance, the weights in the first layer of the Difference Model is *sparse* and exhibits some *interesting structure*.

- Why is this?

BROAD INSTITUTE

At the Interface

- Experimental Setup
- Rudimentary Analysis
- Remaining Work

BROAD INSTITUTE

At the Interface: Remaining Work

- Isolate/Pinpoint the location of the modules.
- Verify their function through Loss/Gain of function screens.

BROAD INSTITUTE

Questions & Feedback

BROAD INSTITUTE

Neural Networks + Genetics

We can draw inspiration from Genetics to design experiments to mechanistically explain how Neural Network work!