Relativity wasn’t in the training set

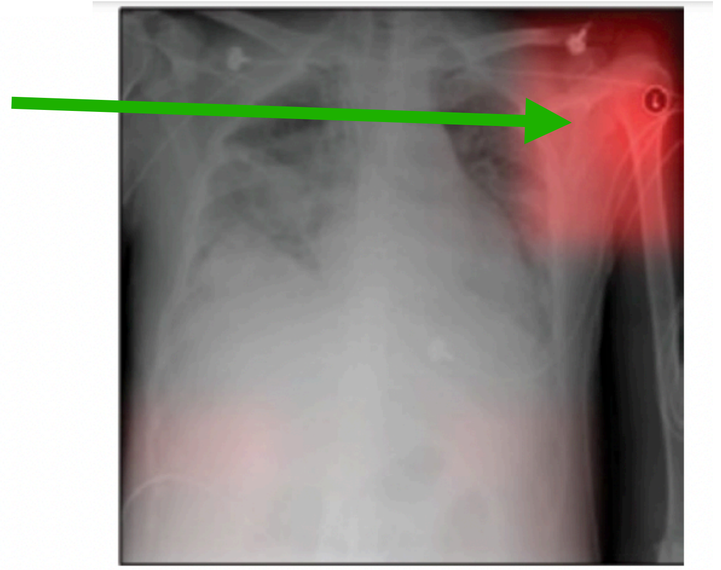
Miles Cranmer
Assistant Prof, DAMTP / IoA

Why is physics so much better at generalising than machine learning?

Problem of unknown unknowns

- Neural networks will pick up on ***<u>everything</u>*** you give them
- Deep neural networks trained for detecting pneumonia
- Later discovered the network was not looking at the lungs at all, but rather looking at **text** in the **background of the image**.

Where the model is most sensitive:



Zech 2018

Task for DNN	Recognize pneumonia
Problem	Fails on scans from new hospitals
Shortcut	Looks at hospital token, not lung

Geirhos et al., 2020

The Illusion of “Simplicity"

Claim 1: **The concept of “complexity" as used in science is wrong**

Claim 2: **Occam's Razor misses the point.**

What we refer to by simplicity is actually ***utility!!***

Core Thesis

“Simplicity" self-deception: We say things are "simple" when we really mean they are generally USEFUL

- e.g., addition: a very abstract and complex concept!
  - But, it's also very useful. “I have 1 less rock today" => “My neighbour stole one!"
  - Even bees can count to 4!
  - Useful for survival

But: addition does not have a built-in register

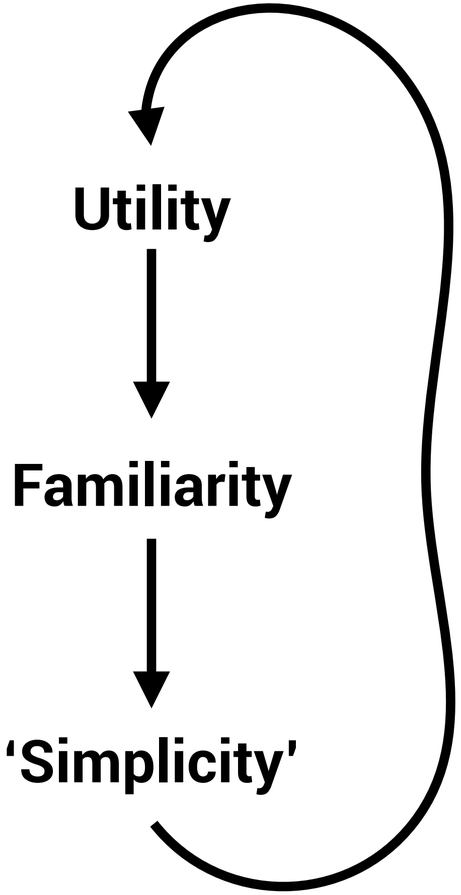
for real numbers in our brain.

Addition is an ***invented concept***!

- The lie:
  - "This model is simple, therefore it's probably true"
- The reality:
  - "This model uses tools that have previously been proven useful for modeling the world"
- We've confused usefulness with aesthetics/simplicity

A note on “simplicity”

- Is “x + y” simple?



The reason “+” is “simple” is ***because*** it is broadly useful.

If something is useful for problems A, B, C; then it seems likely to be also useful for D. (Compositionality)

The Real Process Behind "Occam's Razor"

- Traditional view: "Simple models are more likely to be true"
- **What's actually happening: "Useful models get reused"**
- The models that survive aren't the simplest; they're the ones that keep working
- Simplicity as a concept is just a descriptor based on human psychology and what things we are familiar with. But it's not “real" in this sense.

Why is this bad?

- All “simple" models were once considered “complex"
- We call them simple after becoming familiar with them, which happens after we realise their general utility
- It can be unhealthy to write off complex models purely because of their complexity - because a complex model might become simple in the future!!
  - A completely novel model unlike anything we've seen could be revolutionary if it works across contexts
- The weird observation: models useful in one domain tend to be useful in others
  - Likely due to shared geometry. All problems exist in the same universe with the same physics

How We Actually Build Models

- Most models reuse existing tools
- They're assembled from components that have proven useful:
  - Mathematical operators that helped us build bridges
  - Physical principles that helped us navigate
  - Patterns that helped us survive

What Makes a Model Good?

- **Does it predict things accurately?** That's it!
- Now, we want also want our model to be accurate on unseen, future data.
- The strange thing about our universe is compositionality: Tools useful in one domain often work in others.
- More general concepts are, by definition, more likely to be general.

Symbolic Regression

Symbolic regression is a machine learning task, where the objective is to find ***analytic expressions*** that optimize some objective.

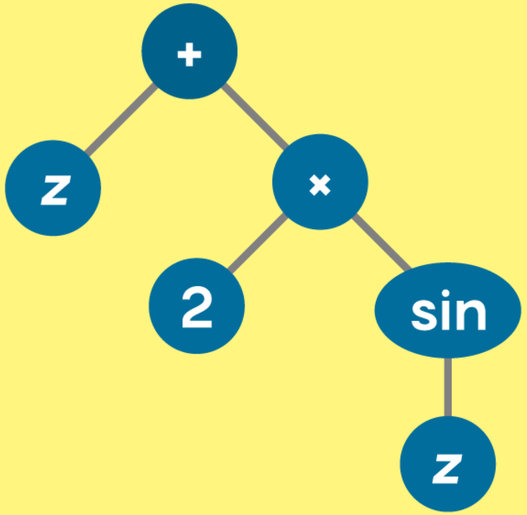
- Popularized by Koza (1990s); and its use in science by Lipson (2000s)

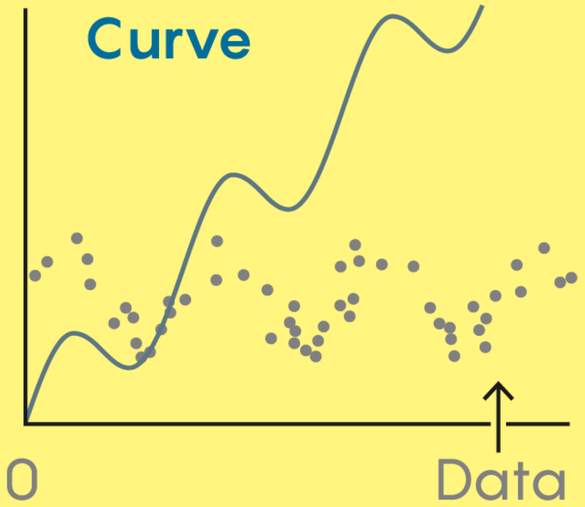
Jointly optimize accuracy and complexity, where complexity is user-defined

EQUATIONS AS TREES

$y = z + 2 \sin z$

can be represented as the following tree and curve.

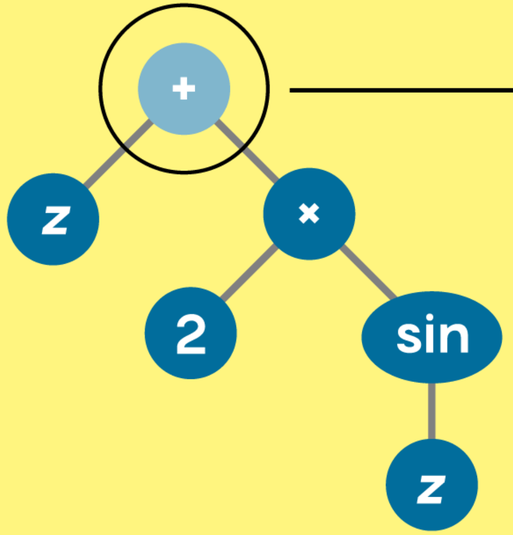


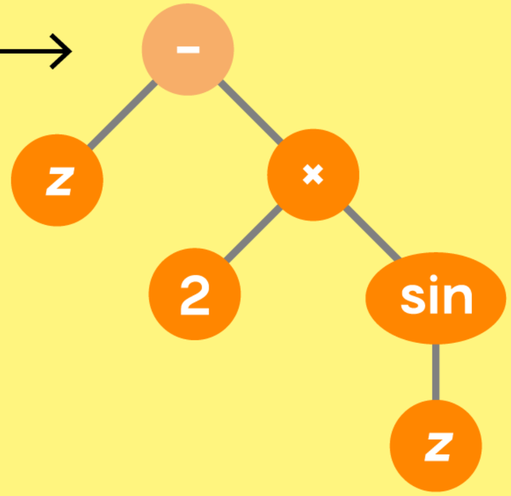


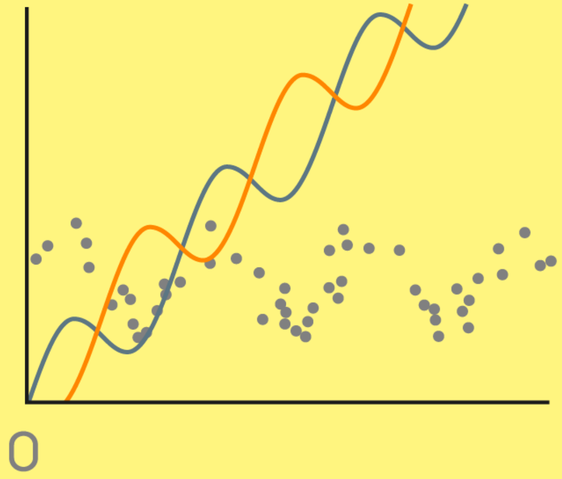
Kouzou Sakai
for Quanta Magazine

MUTATION

The algorithm might mutate one node of the tree.





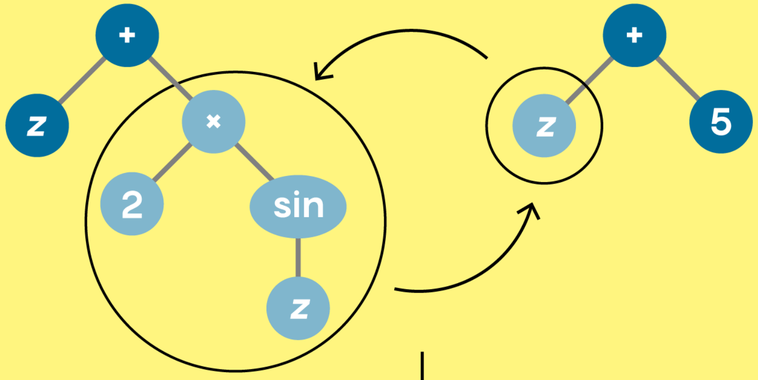


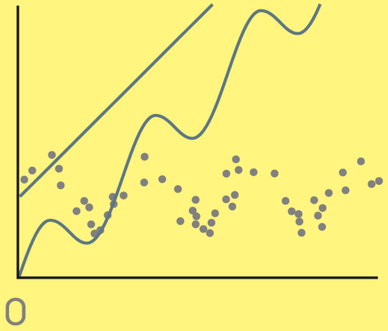
$y = z + 2 \sin z$

$y = z - 2 \sin z$

CROSSBREEDING

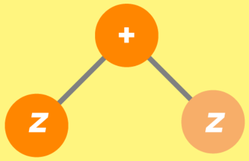
It may also breed new equations by swapping the branches of existing ones.

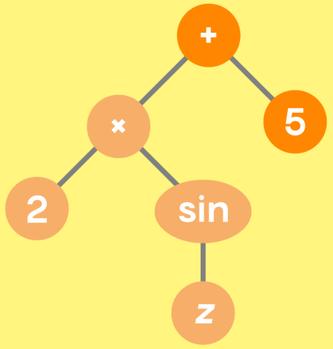


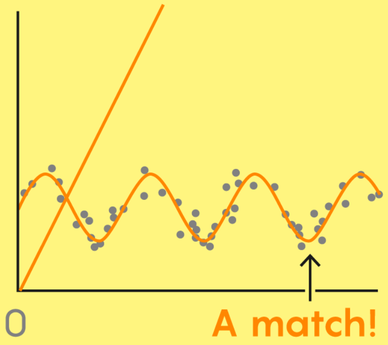


$y = z + 2 \sin z$

$y = z + 5$







$y = z + z$

$y = 2 \sin z + 5$

Tools:

PySR &
SymbolicRegression.jl

High-Performance Symbolic Regression in Python and Julia

ai.damtp.cam.ac.uk/pysr

Apache-2.0 license

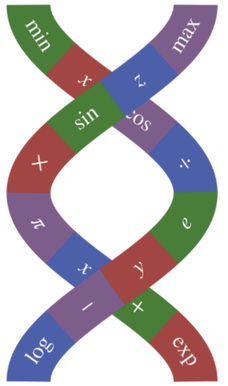
2.8k stars **247** forks **31** watching **60** Branches **184** Tags

SymbolicRegression.jl Search ⌘K Home Examples API Losses Types Customization dev

SymbolicRegression.jl
Discover Mathematical Laws from Data

A flexible, user-friendly framework that automatically finds interpretable equations from your data

Get Started API Reference View on GitHub



Insight from SR

- SR works best, and generalises best, when ***you use operators common for the particular subfield*** (dilogarithm makes sense for particle physics, but not for biology!)
  - Kind of a quantitative way to frame this argument
  - Not because those operators are "simpler"
- Expressions also generalise better when the expressions are smaller (overfitting)
- Perhaps because they align with patterns that actually exist in that domain

Another perspective on this

(2021-) Shift in industry machine learning to favour general pre-trained “foundation models"

Language

Vision

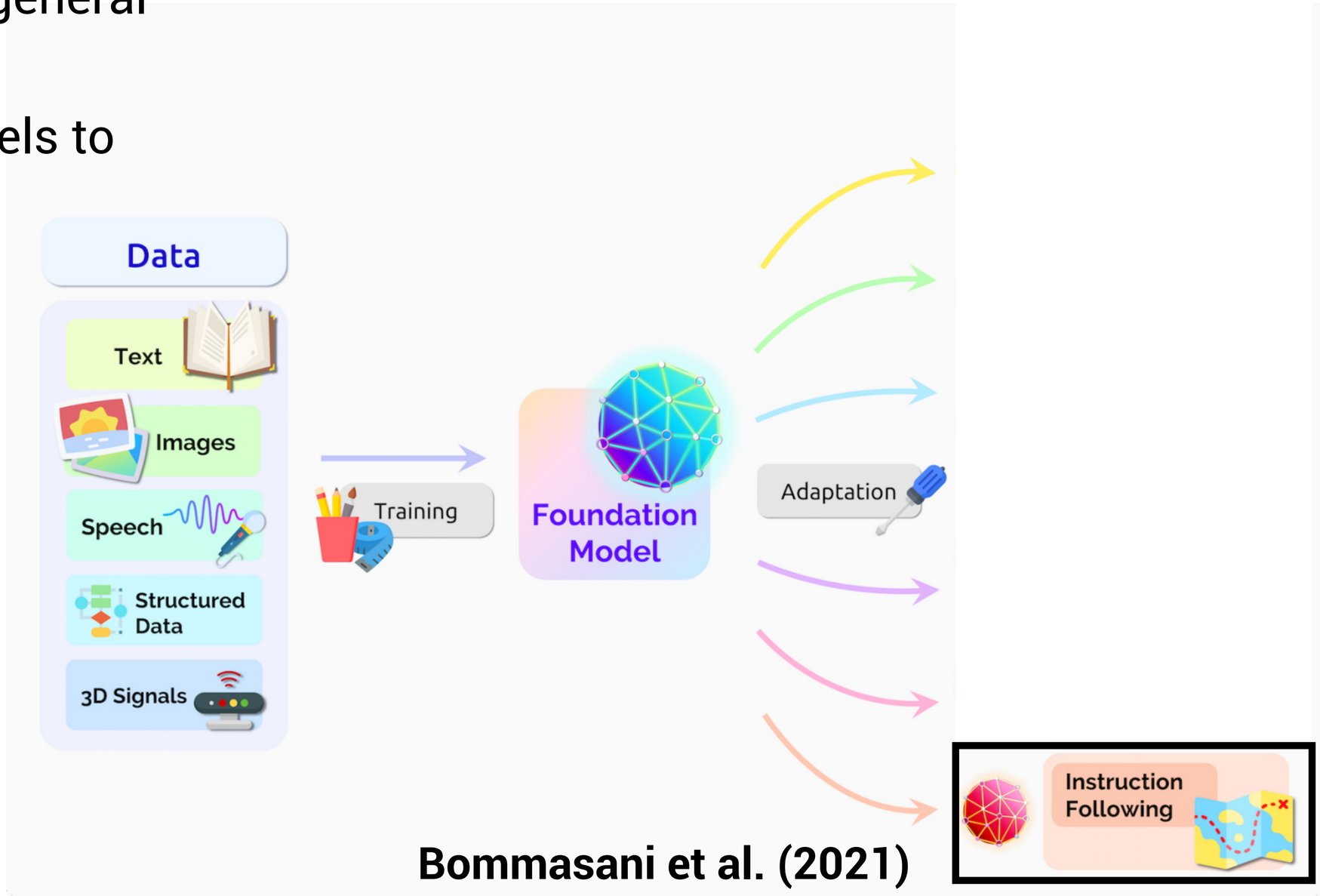
MISTRAL AI_

ChatGPT

LangChain

GitHub Copilot

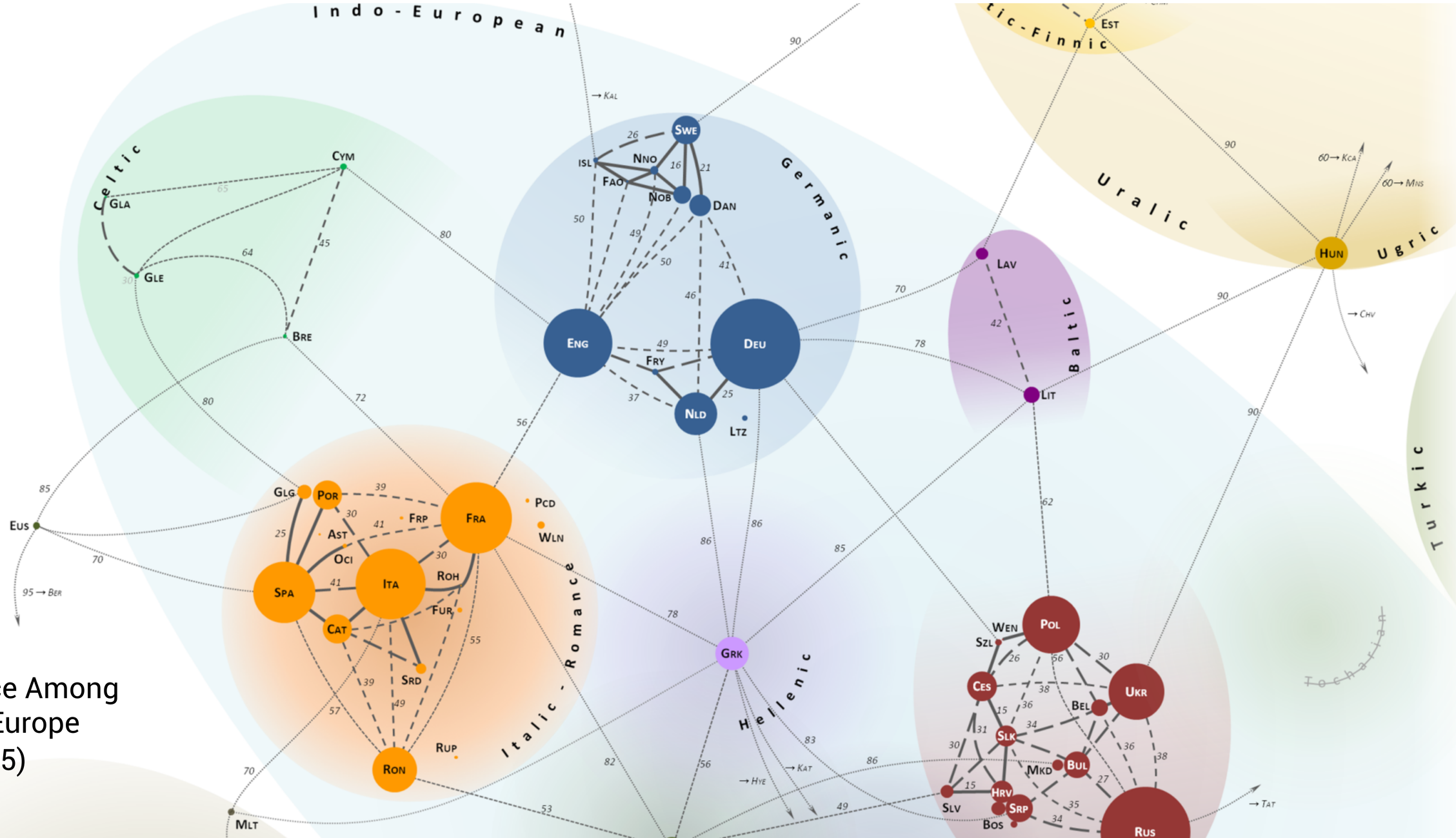
- Foundation Model Approach:
  - **Pretrain** models on tasks without supervision on large, general datasets
  - **Adapt** pretrained models to downstream tasks



Bommasani et al. (2021)

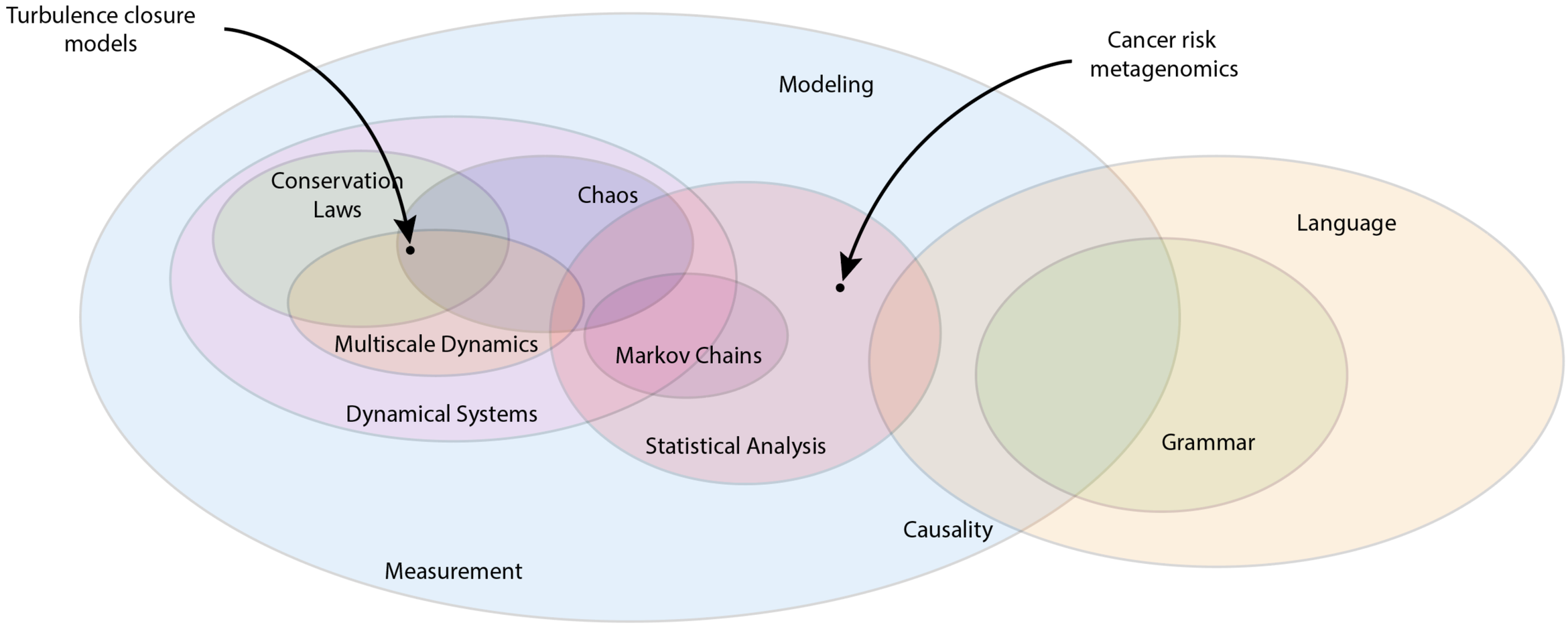
Teaching the model "general" knowledge before specialising it!

Why does ChatGPT work?



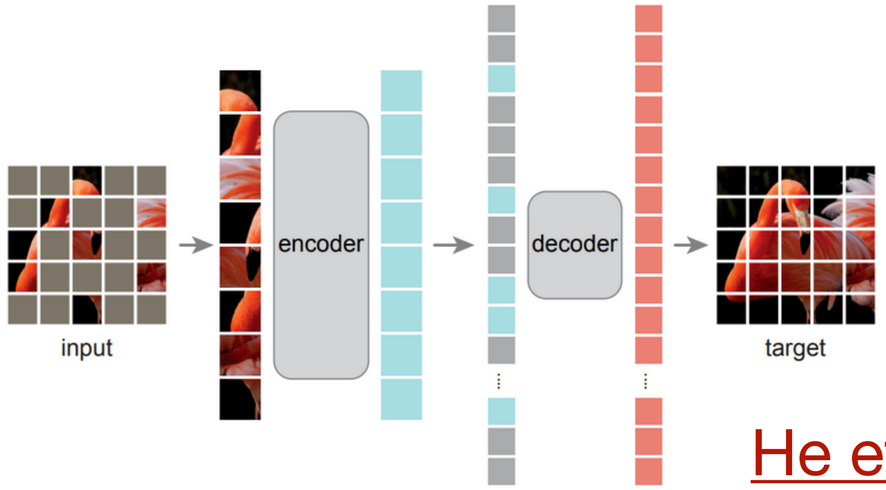
Lexical Distance Among Languages of Europe
Steinbach (2015)

“Universality classes”

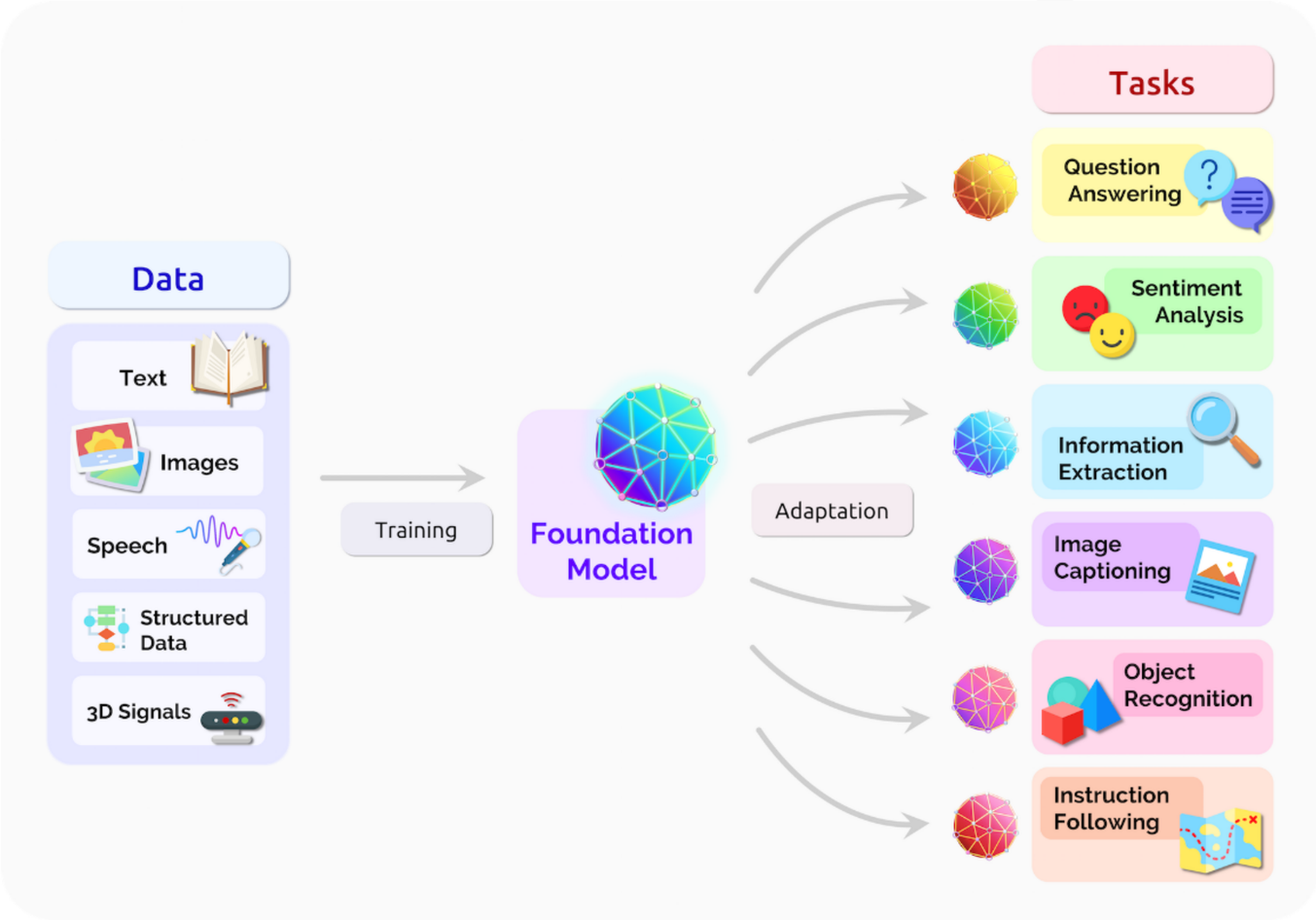


The Rise of the Foundation Model Paradigm

- **Foundation Model approach**
  - **Pretrain** models on objectives that do not require manual labeling to get access to very large datasets.
  - **Adapt** pretrained models to downstream tasks.
  - **Combine** pretrained models in more complex systems



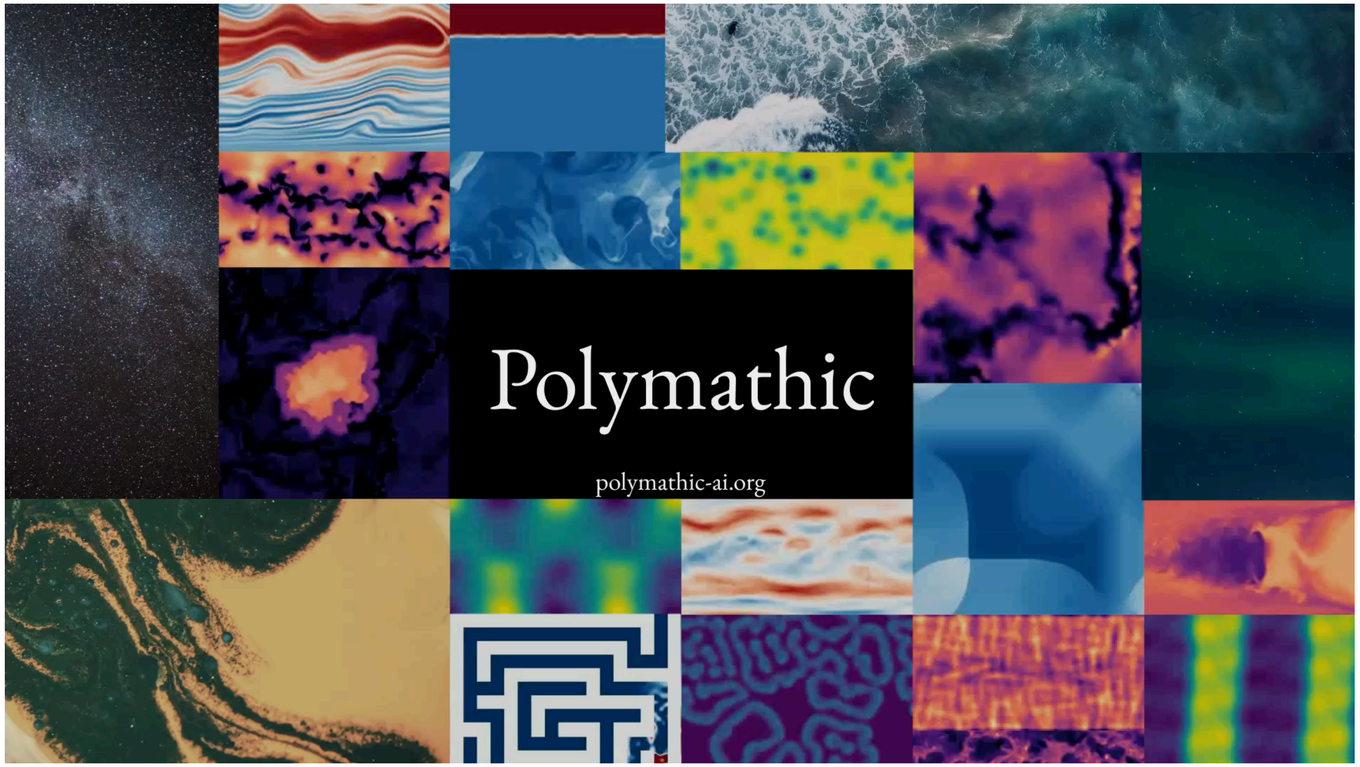
He et al. 2021



Bommansani et al. 2021

Introducing: Polymathic AI

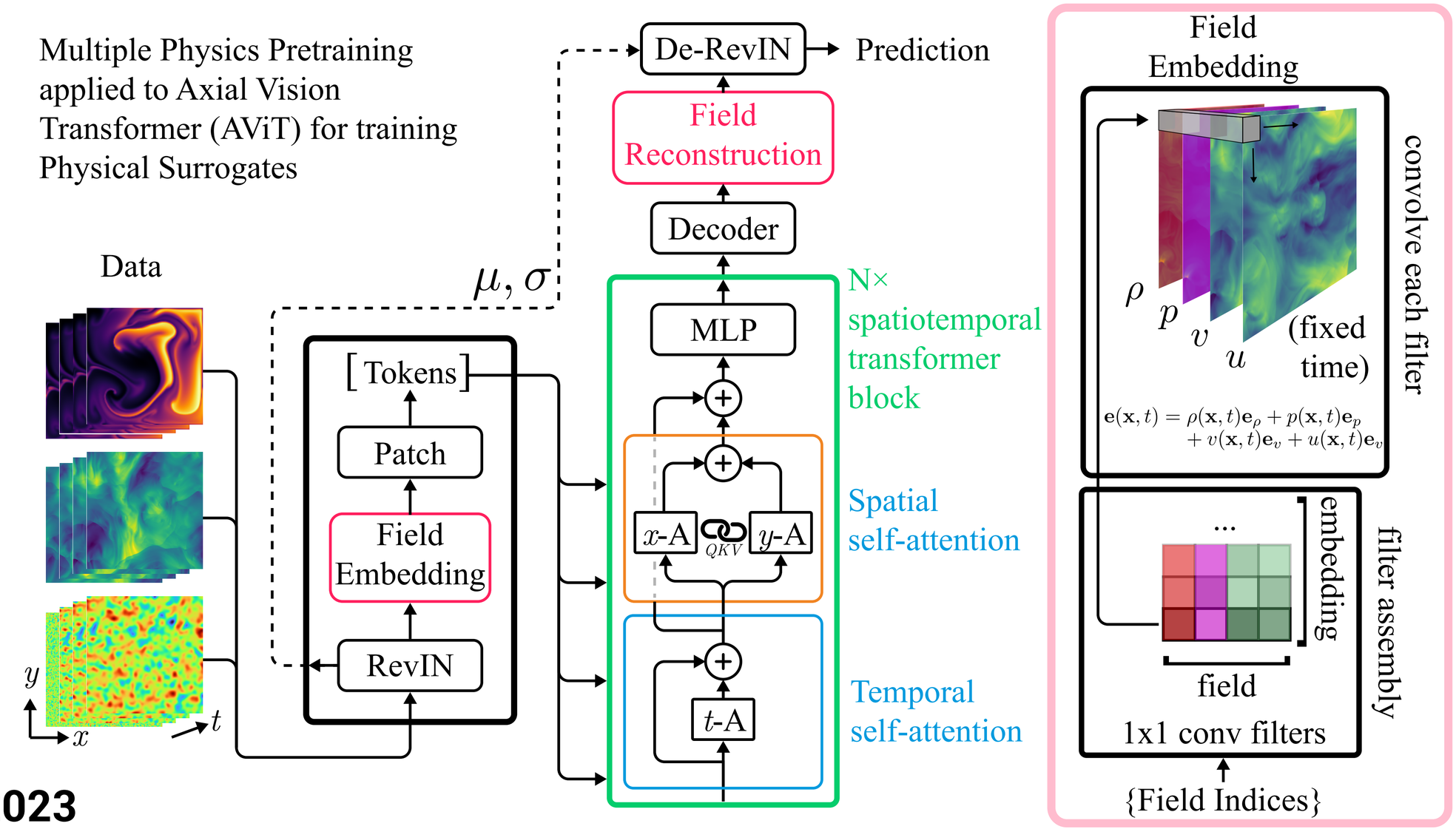
- An international **research collaboration** to build foundation models for science
  - Mostly split between the University of Cambridge and Flatiron Institute
  - Shared compute for training models at **industry scale**



Multiple Physics Pretraining
(a general emulator for continuum physics)

Many different physics!

Processed with the SAME transformer



McCabe et al., 2023

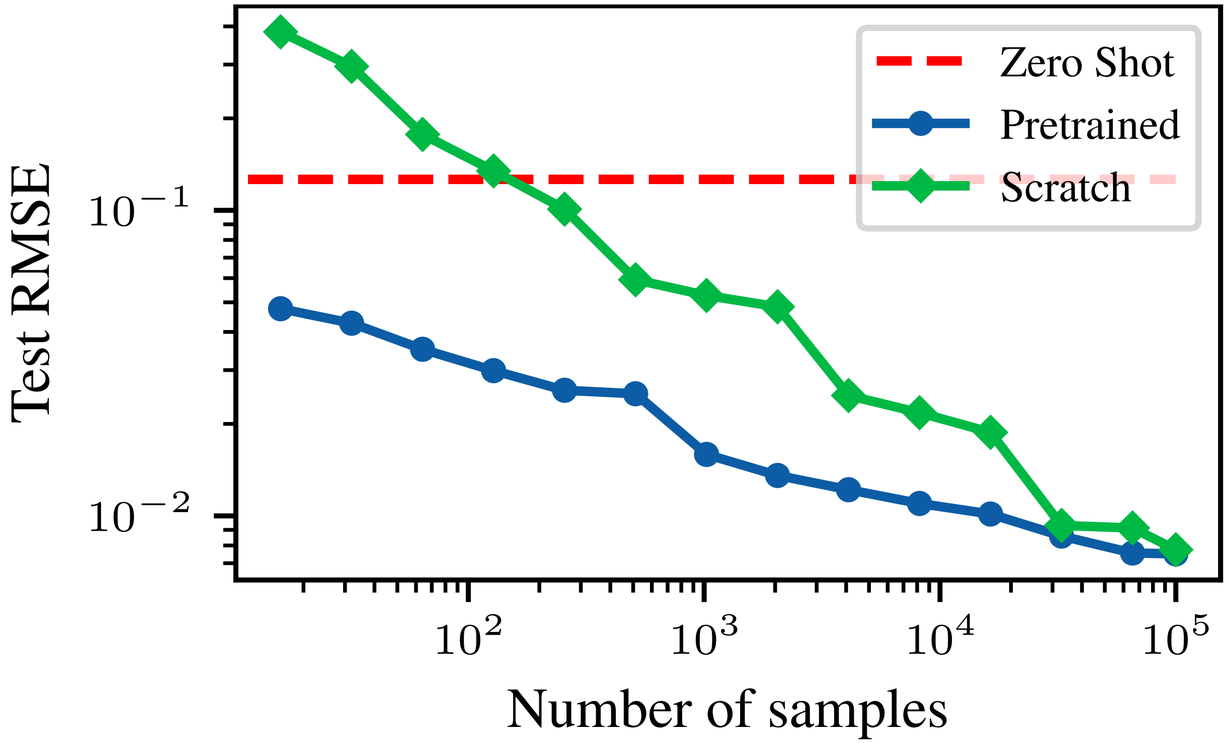
Generalisation in Physics Foundation Models

Scalar Transport - Compositional Systems

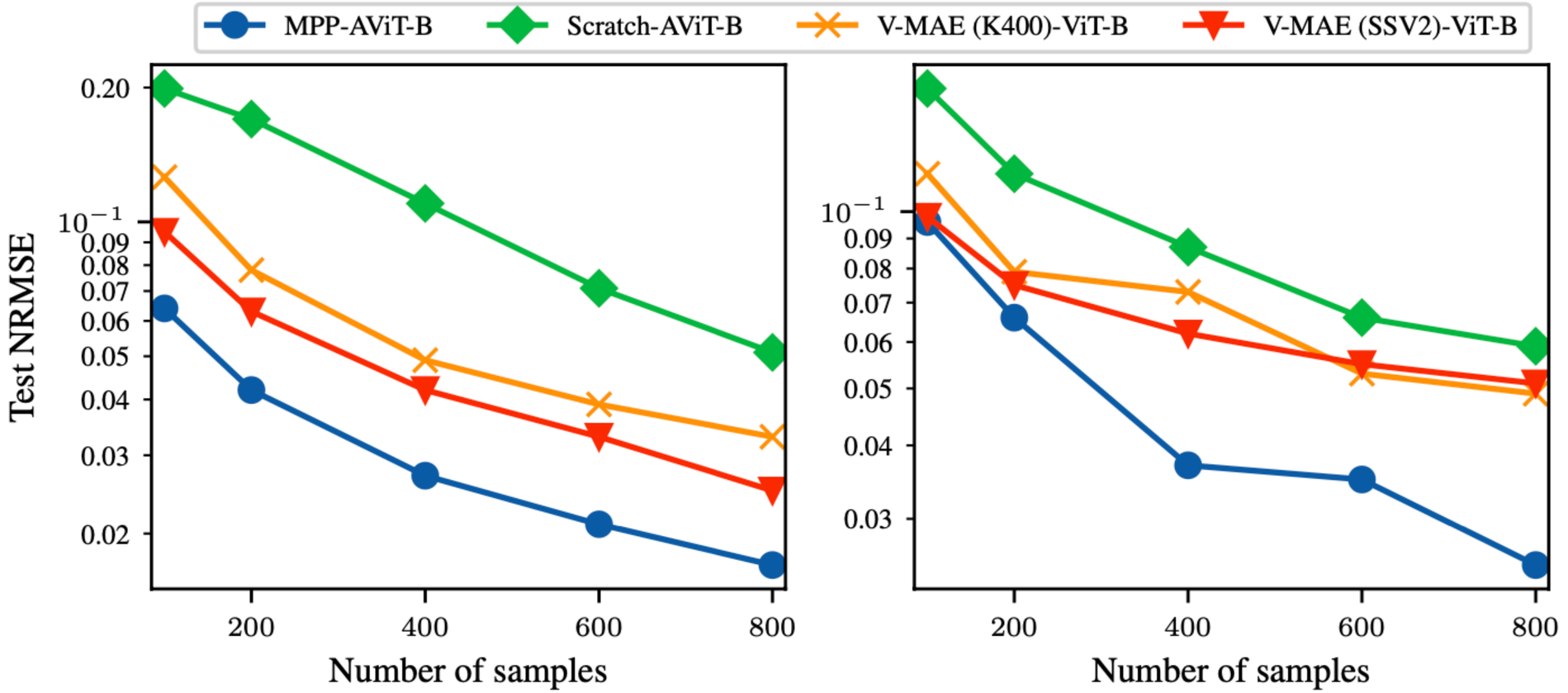
Advection: $\frac{\partial \psi}{\partial t} + \nabla \cdot (v\psi) = 0$

Diffusion: $\frac{\partial \psi}{\partial t} + \nabla \cdot (-\delta \nabla \psi) = 0$

Advection-Diffusion: $\frac{\partial \psi}{\partial t} + \nabla \cdot (v\psi - \delta \nabla \psi) = 0.$



McCabe et al., 2023



Even pre-training on cat videos helps…

(Due to shared concepts of spatiotemporal continuity?)

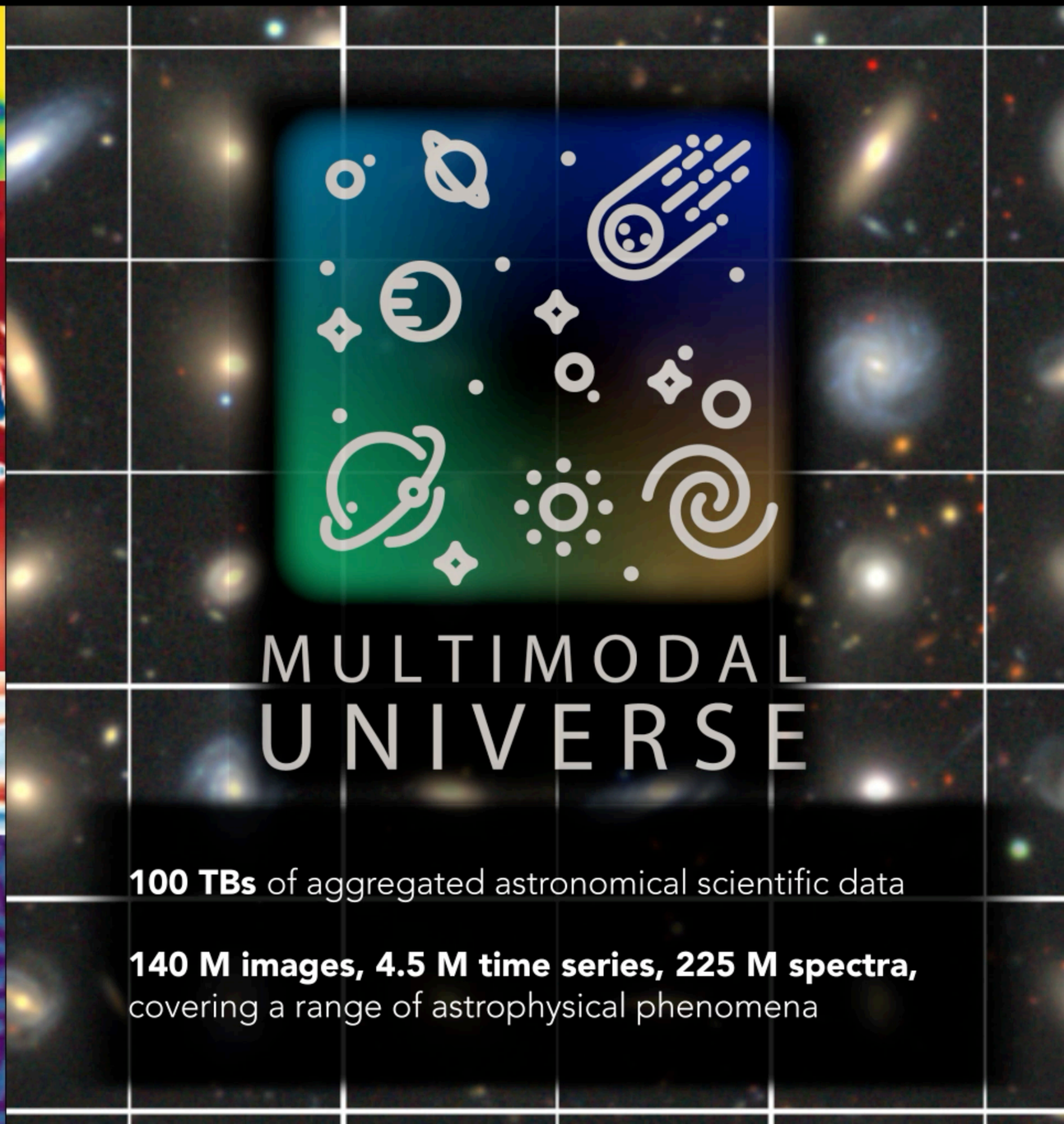
github.com/PolymathicAI/the_well

The Well

55 B tokens from **3 M** frames

16 physical settings with ranges of physical parameters spanning problems in astro, bio, aerospace, chemistry, fundamental turbulence, atmospheric science and more.

github.com/MultimodalUniverse/MultimodalUniverse



MULTIMODAL UNIVERSE

100 TBs of aggregated astronomical scientific data

140 M images, 4.5 M time series, 225 M spectra, covering a range of astrophysical phenomena

The Well: a Large-Scale Collection of Diverse Physics Simulations for Machine Learning

Ruben Ohana [1,2,*], **Michael McCabe** [1,*], **Lucas Meyer** [1], **Rudy Morel** [1,2], **Fruzsina J. Agocs** [2,3,†], **Miguel Beneitez** [4,†], **Marsha Berger** [2,5,†], **Blakesley Burkhart** [2,6,†], **Stuart B. Dalziel** [4,†], **Drummond B. Fielding** [2,7,†], **Daniel Fortunato** [2,†], **Jared A. Goldberg** [2,†], **Keiya Hirashima** [1,2,8,†], **Yan-Fei Jiang** [2,†], **Rich R. Kerswell** [4,†], **Suryanarayana Maddu** [2,†], **Jonah Miller**[9,†], **Payel Mukhopadhyay** [10,†], **Stefan S. Nixon** [4,†], **Jeff Shen** [11,†], **Romain Watteaux** [12,†], **Bruno Régaldo-Saint Blancard** [1,2], **François Rozet**[1,13], **Liam H. Parker**[1,2,10], **Miles Cranmer** [1,4], **Shirley Ho** [1,2,5,11]

The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data

The Multimodal Universe Collaboration
Eirini Angeloudi[1,2], **Jeroen Audenaert**[3], **Micah Bowles**[4,5], **Benjamin M. Boyd**[6], **David Chemaly**[6], **Brian Cherinka**[7], **Ioana Ciucă**[9,10,8], **Miles Cranmer**[6,5] **Aaron Do**[6], **Matthew Grayling**[6], **Erin E. Hayes**[6], **Tom Hehir**[6,5] **Shirley Ho**[11,13,14,5], **Marc Huertas-Company**[1,2,10], **Kartheik G. Iyer**[15,11,10], **Maja Jablonska**[8,10] **Francois Lanusse**[11,5,16], **Henry W. Leung**[17], **Kaisey Mandel**[6], **Juan Rafael Martínez-Galarza**[18,19] **Peter Melchior**[14], **Lucas Meyer**[11,5], **Liam H. Parker**[11,5,12], **Helen Qu**[20] **Jeff Shen**[14], **Michael J. Smith**[21,10], **Connor Stone**[22,23,24], **Mike Walmsley**[17], **John F. Wu**[7,25]

Hot off the press!

Polymathic

AION-1: Omnimodal Foundation Model for Astronomical Sciences

Liam Parker[*,1,2,3,4], **Francois Lanusse**[*,5,2], **Jeff Shen**[*,6], **Ollie Liu**[7], **Tom Hehir**[8], **Leopoldo Sarra**[3], **Lucas Meyer**[3], **Micah Bowles**[9], **Sebastian Wagner-Carena**[2,3], **Helen Qu**[2], **Siavash Golkar**[2,3], **Alberto Bietti**[2], **Hatim Bourfoune**[10], **Nathan Cassereau**[10], **Pierre Cornette**[10], **Keiya Hirashima**[2,11], **Geraud Krawezik**[2], **Ruben Ohana**[2], **Nicholas Lourie**[3], **Michael McCabe**[2,3], **Rudy Morel**[2], **Payel Mukhopadhyay**[1,8], **Mariel Pettee**[12], **Bruno Regaldo-Saint Blancard**[2], **Kyunghyun Cho**[3], **Miles Cranmer**[8], **Shirley Ho**[2,3,6]

[1]University of California, Berkeley, [2]Flatiron Institute, [3]New York University, [4]Lawrence Berkeley National Laboratory, [5]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, [6]Princeton University, [7]University of Southern California, [8]University of Cambridge, [9]University of Oxford, [10]IDRIS, CNRS, [11]RIKEN Center for iTHEMS, [12]University of Wisconsin–Madison

[*]Equal Contribution.

While foundation models have shown promise across a variety of fields, astronomy still lacks a unified framework for joint modeling across its highly diverse data modalities. In this paper, we present AION-1, a family of large-scale multimodal foundation models for astronomy. AION-1 integrates heterogeneous imaging, spectroscopic, and scalar data using a two-stage architecture: modality-specific tokenization followed by transformer-based masked modeling of cross-modal token sequences. The model is pretrained on five large-scale surveys: Legacy Survey, Hyper Suprime-Cam (HSC), Sloan Digital Sky Survey (SDSS), Dark Energy Spectroscopic Instrument (DESI), and Gaia. These span more than 200 million observations of stars, galaxies, and quasars. With a single frozen encoder, AION-1 achieves strong results on a broad suite of downstream tasks, including galaxy and stellar property estimation, galaxy morphology classification, similarity-based retrieval, galaxy image segmentation, and spectral super-resolution. We release AION-1 model variants ranging from 300 M to 3.1 B parameters. Beyond astronomy, AION-1 provides a scalable blueprint for multimodal scientific foundation models that can seamlessly integrate noisy, instrument-specific observations. All code, tokenizers, pretrained weights, and a lightweight evaluation suite are released under an open-source license.

Date: October 22, 2025

Correspondence: Liam Parker: lhparker@berkeley.edu; Francois Lanusse: francois.lanusse@cnrs.fr

Code: https://github.com/PolymathicAI/AION/

arXiv: 2510.17960

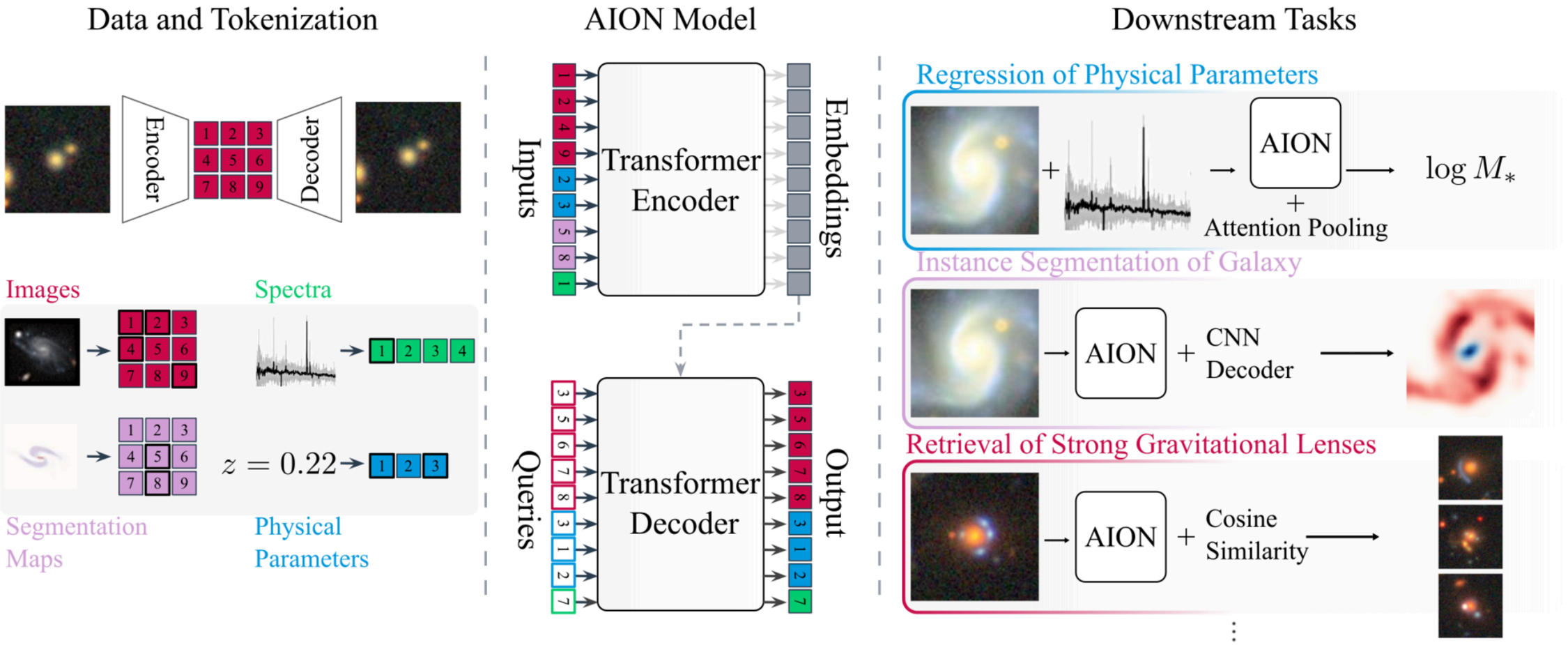
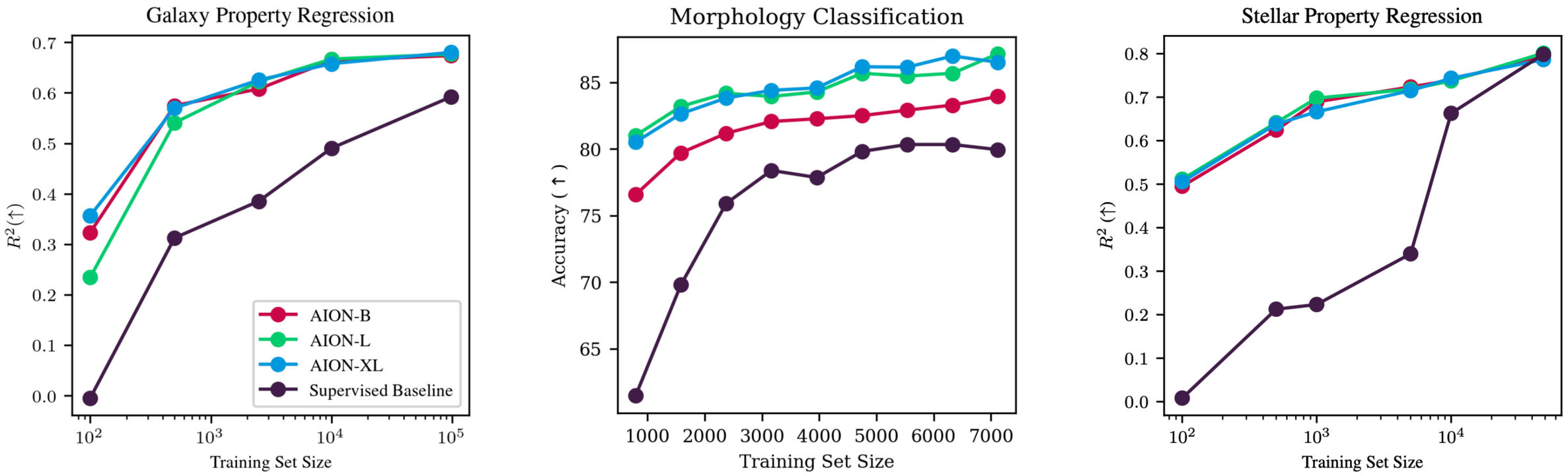


Figure 1: AION-1 integrates 39 different data modalities — multiband images, optical spectra, and various properties and measurements — into a single model usable for a wide range of downstream applications. It implements a two-step process: first, bespoke tokenization strategies that homogenize the diverse scientific data, followed by multimodal masked modeling that learns how different observations relate, inducing a deep understanding of the underlying physical objects. Astronomers can then leverage AION-1's rich astrophysical understanding for a variety of downstream tasks.



(a): Galaxy Properties **(b):** Morphology Classification **(c):** Stellar Properties

Figure 10: Model performance vs downstream task training set size. We regress **(a)** galaxy physical properties from images, **(b)** classify galaxy morphology, and **(c)** regress stellar properties from spectra on the same held-out test sample equating to 20% of the available data. However, we artificially reduce the training set size, and train a lightweight head on top of the frozen AION-1 encoder and a supervised model on the raw input data for each training set size. For galaxy and stellar properties, we report the R^2 averaged over all the properties, while for morphology classification we report the average accuracy.

Summary

- Symbolic regression and foundation models have a connection in the sense they RE-USE core pieces
- (If time left, can show exciting new results on interpretability of these foundation models)