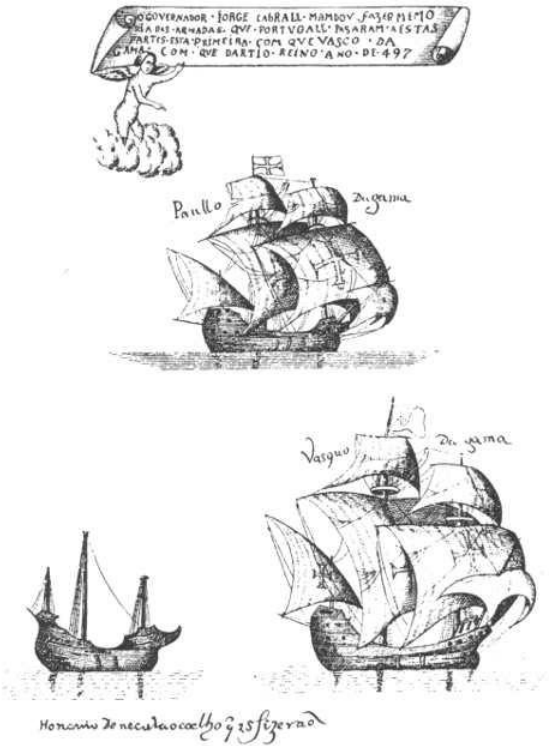
Why Model Design Still Matters in the Age of Scale

Nima Keivan

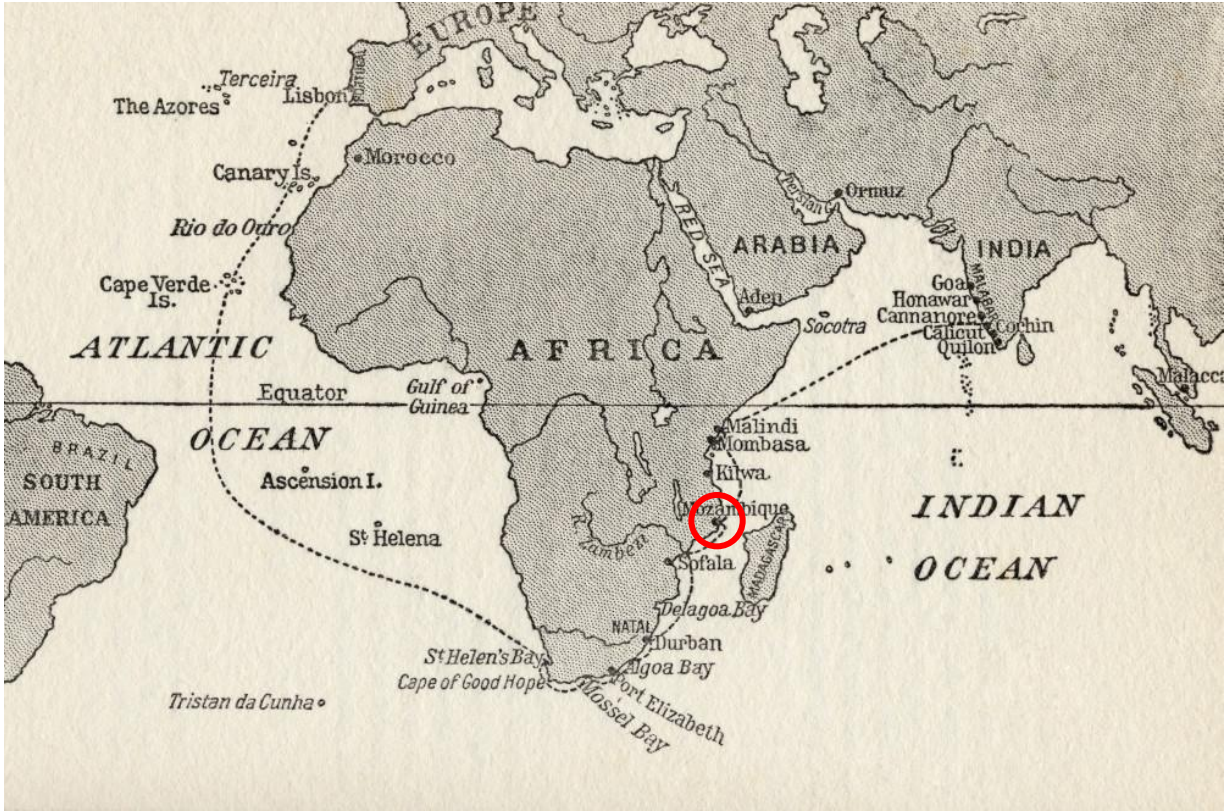
Vasco Da Gama (1460 - 1524)



First Voyage: July 1497

4 ships,170 crew.

January 1498 (6+ months later), Mozambique



“Many of our men fell ill here, their feet and hands swelling, and their gums growing over their teeth, so that they could not eat.”

7–13 April 1498, Malindi

“.. two Almadias approached us. One was laden with fine oranges, better than those of Portugal.”

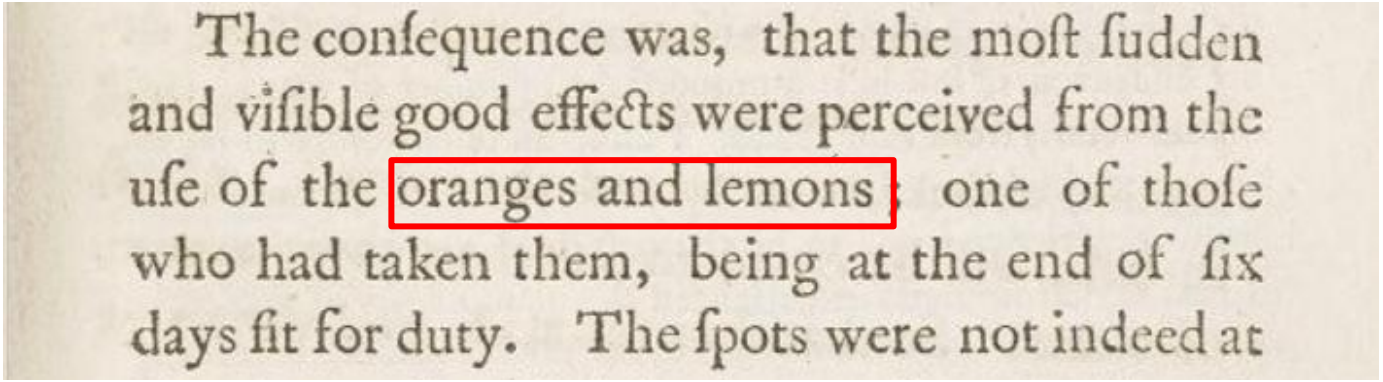
5 days later

“... on arriving at this city all our sick recovered their health, **for the climate (“air”) of this place is very good**…”

Return Voyage: January 1499

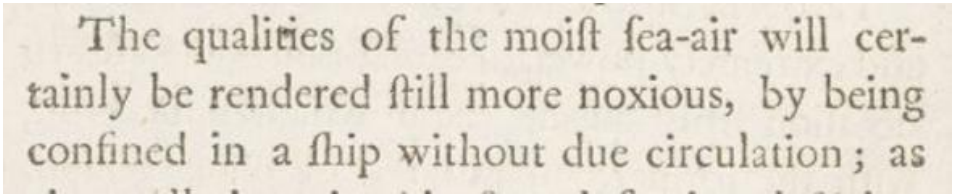
Half the remaining crew die crossing the indian ocean

> “The captain-major sent a man on shore with these messengers with instructions to bring off a supply of oranges, **which were much desired by our sick**.”

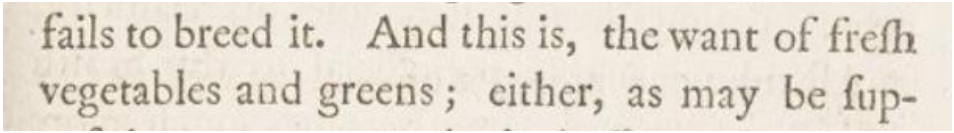

The conſequence was, that the moſt ſudden and viſible good effects were perceived from the uſe of the oranges and lemons; one of thoſe who had taken them, being at the end of ſix days fit for duty. The ſpots were not indeed at

1747 (250 years later): James Lind's randomized trial

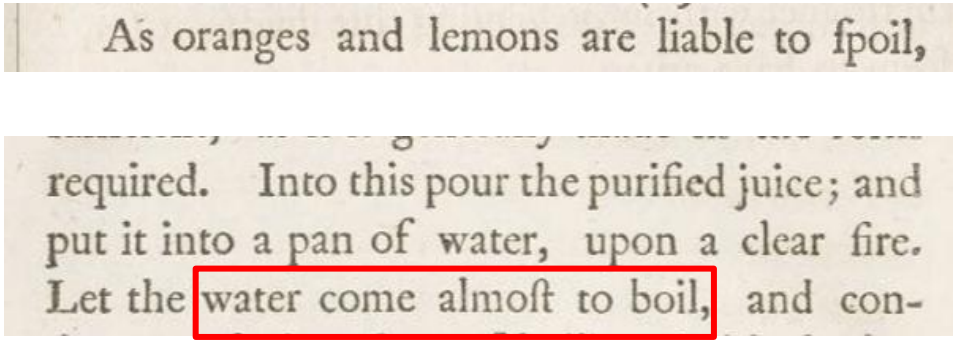
moist sea air


The qualities of the moiſt ſea-air will cer-
tainly be rendered ſtill more noxious, by being
confined in a ſhip without due circulation; as

lack of fresh vegetables / greens


fails to breed it. And this is, the want of freſh
vegetables and greens; either, as may be ſup-

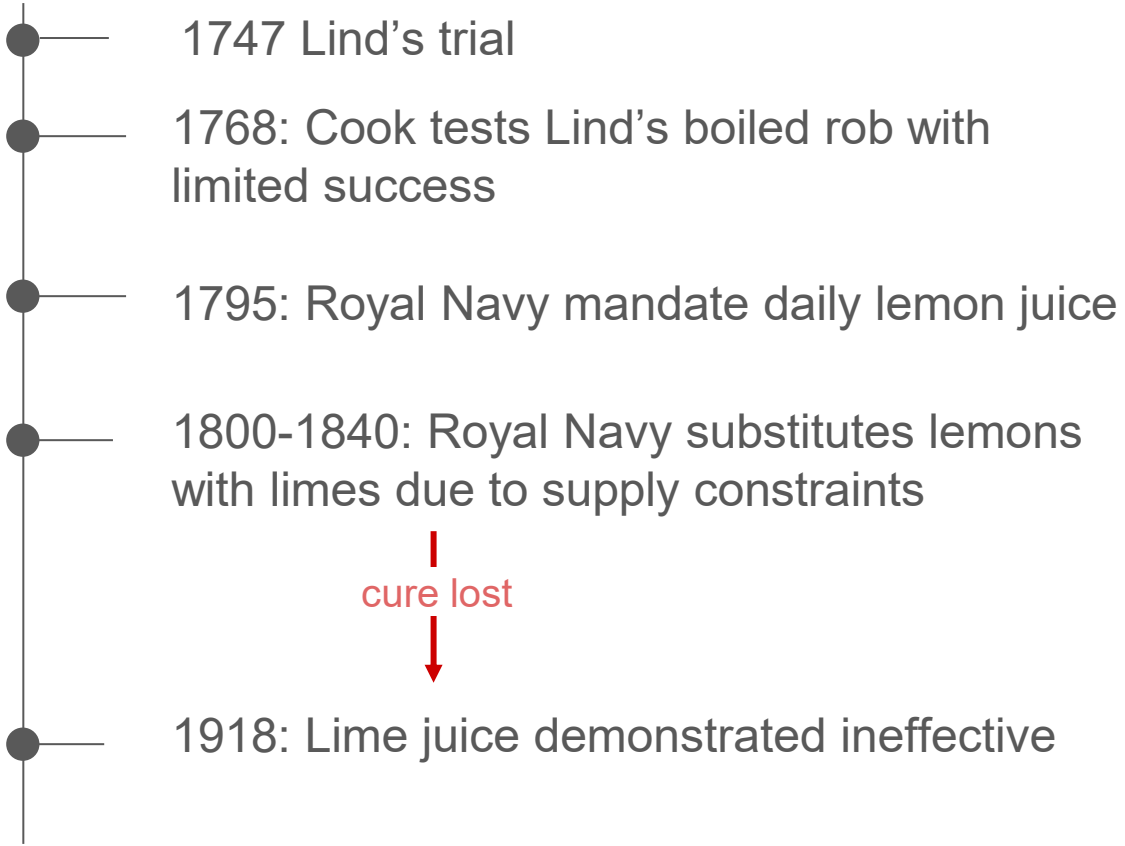
Lind got the causes completely wrong



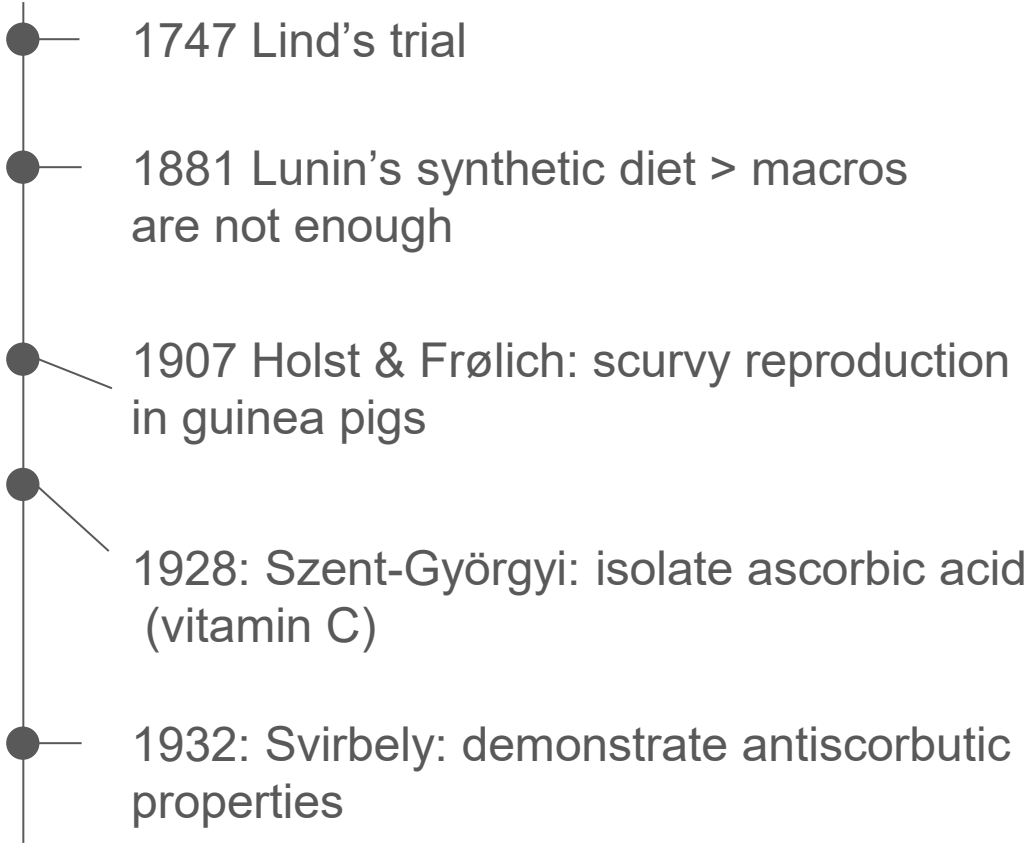
As oranges and lemons are liable to ſpoil,

required. Into this pour the purified juice; and put it into a pan of water, upon a clear fire. Let the water come almoſt to boil, and con-

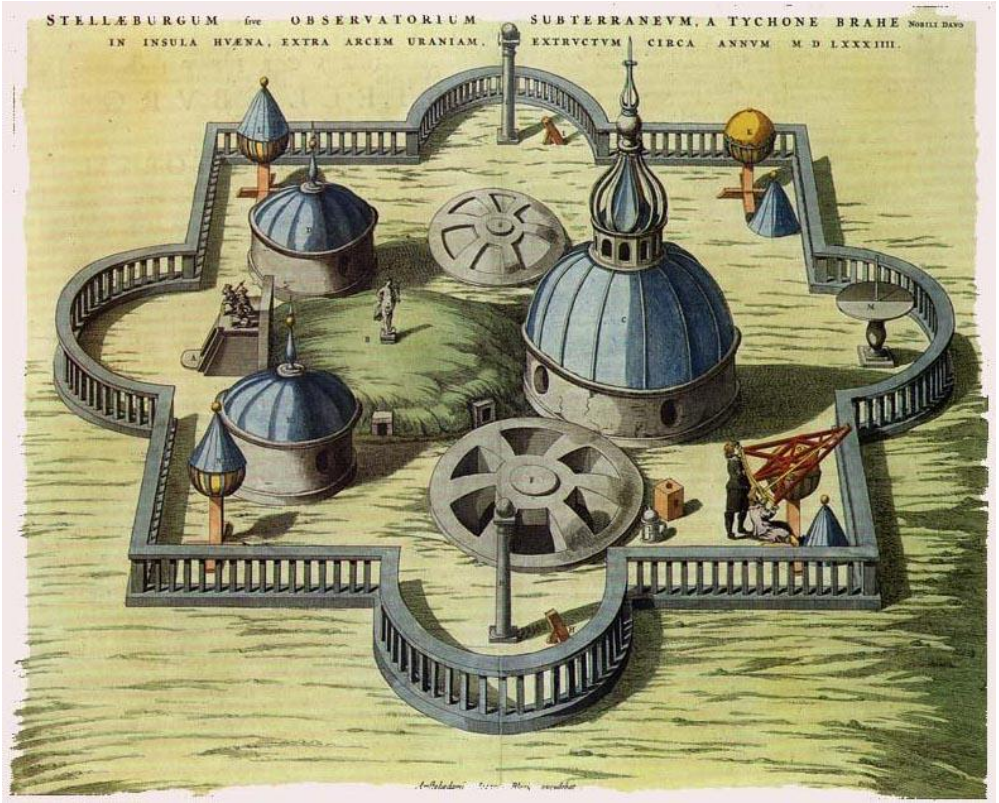
… and recommended boiling citrus juice, which destroys vitamin C



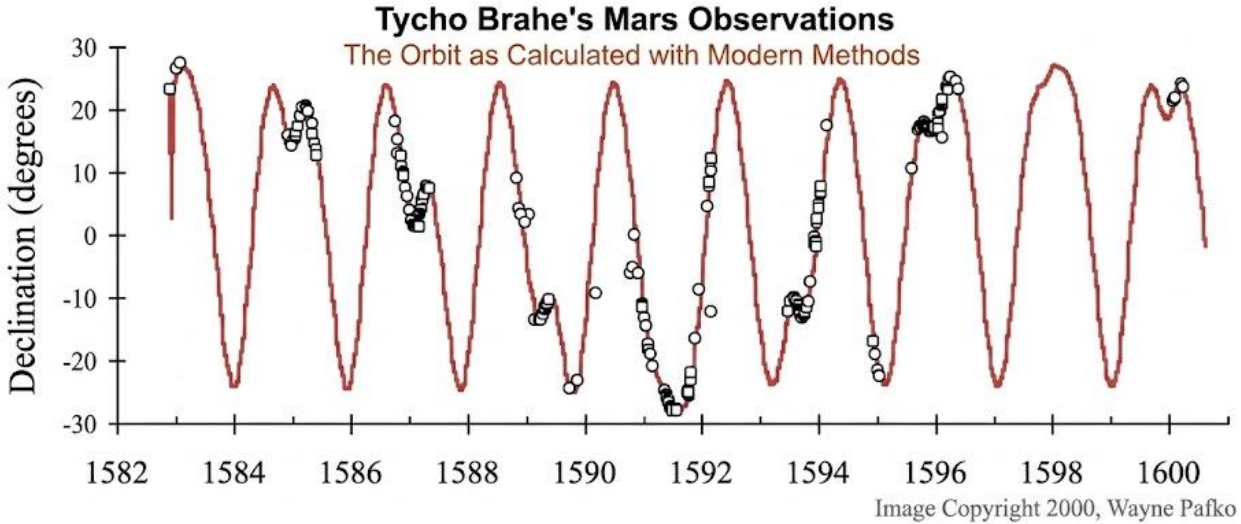
In the absence of understanding, the cure was at times lost



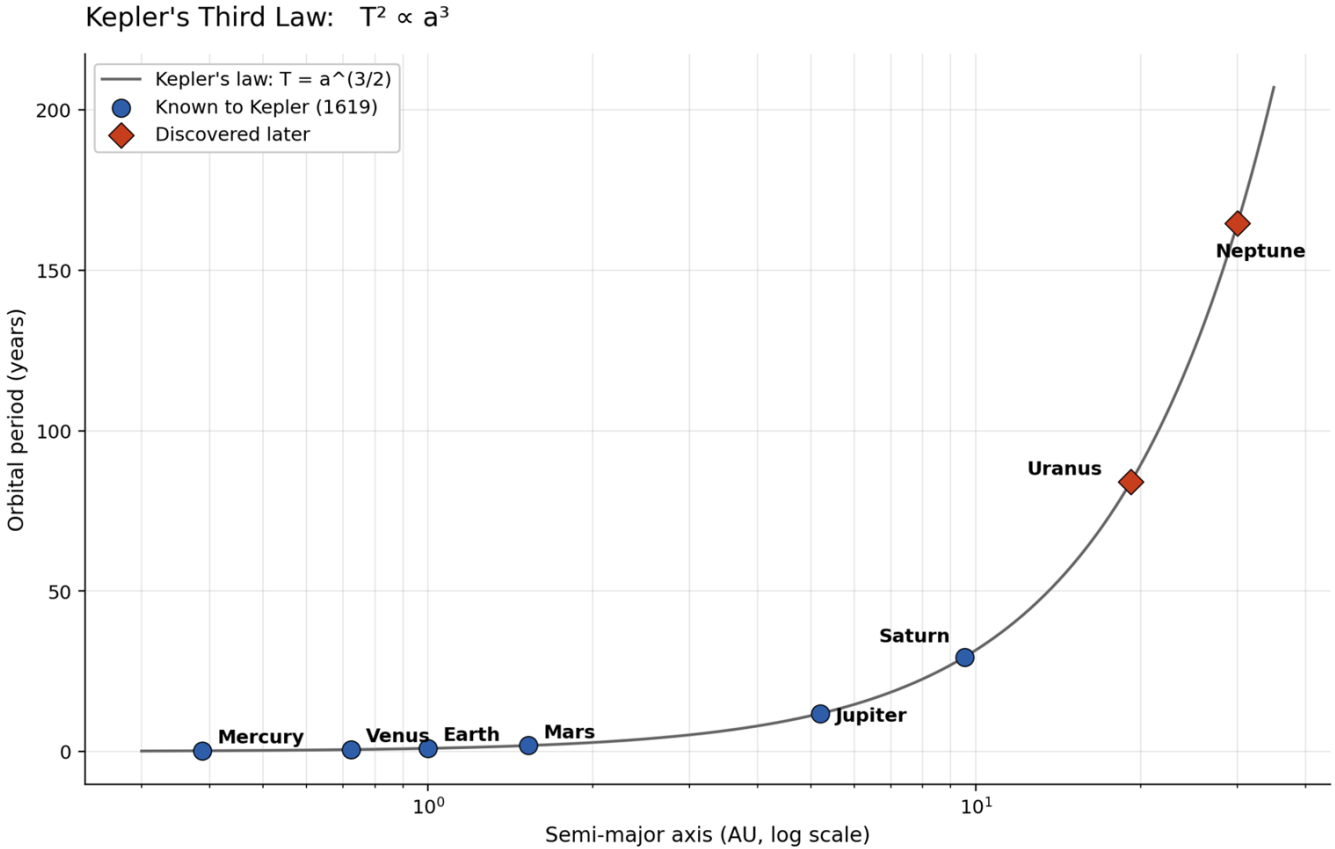
Causal understanding took almost 200 years.



1576: Tycho Brahe builds Uraniborg



20 years, 5-10x accuracy, 7 solar system bodies, 1000s of measurements



1619: Kepler's introduces the *third law*

Kepler's consequential priors:

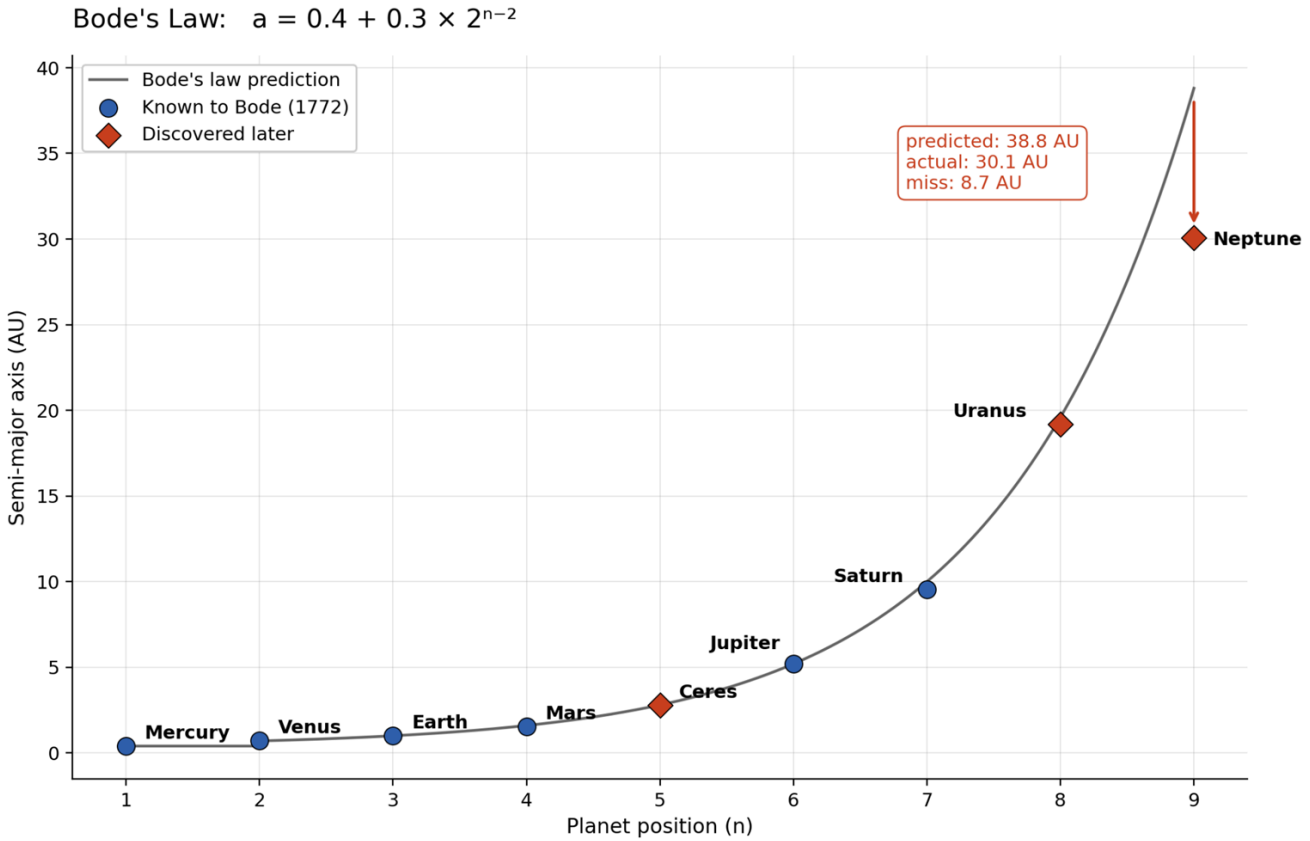
Copernican heliocentricity:
(rejecting Brahe's geocentric model)

Platonic solids (wrong):
prefer simple, mathematical rules

Harmonics in music:
distance and time are related

Proto-gravity (wrong):
driven by causal understanding

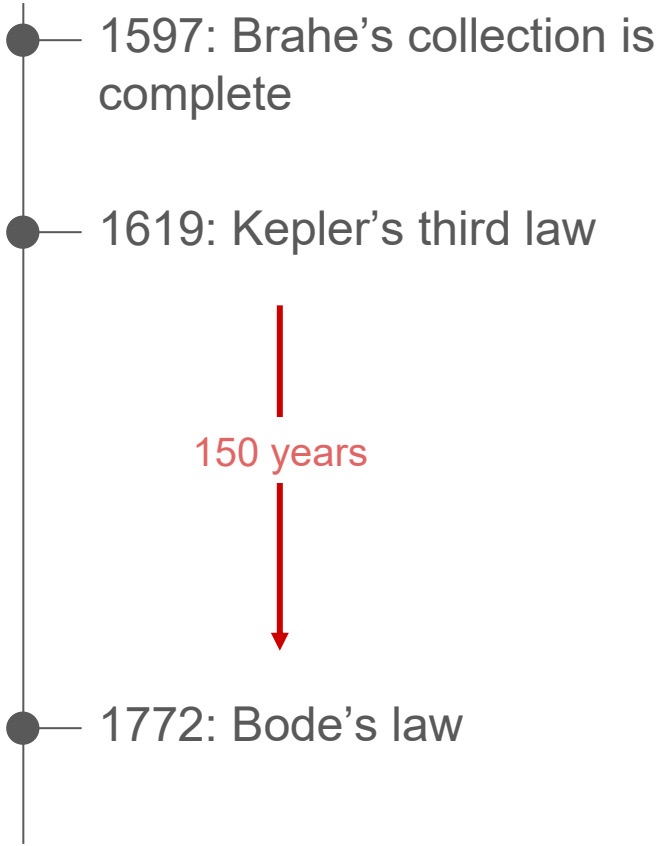
"these eight minutes alone will have led the way to the reformation of all of astronomy" - Kepler in *Astronomia Nova*



1772: (153 years later), Bode introduces his (actually Titius's) law

Bode's law followed Kepler's by **150 years**, **fit the data**, **made correct predictions (Ceres, Uranus)**, and was **simpler**.

It was also **wrong.**

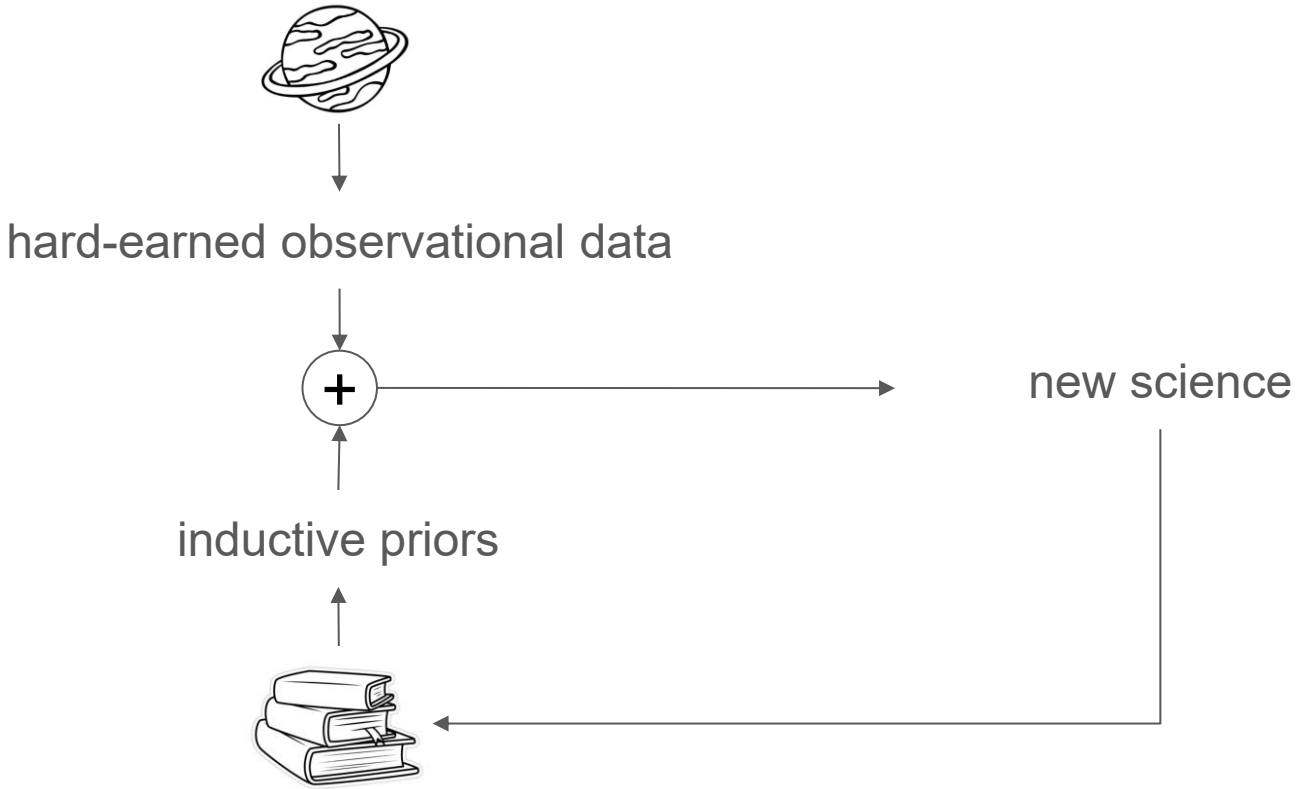


Bode's (Titius's) Theological Priors

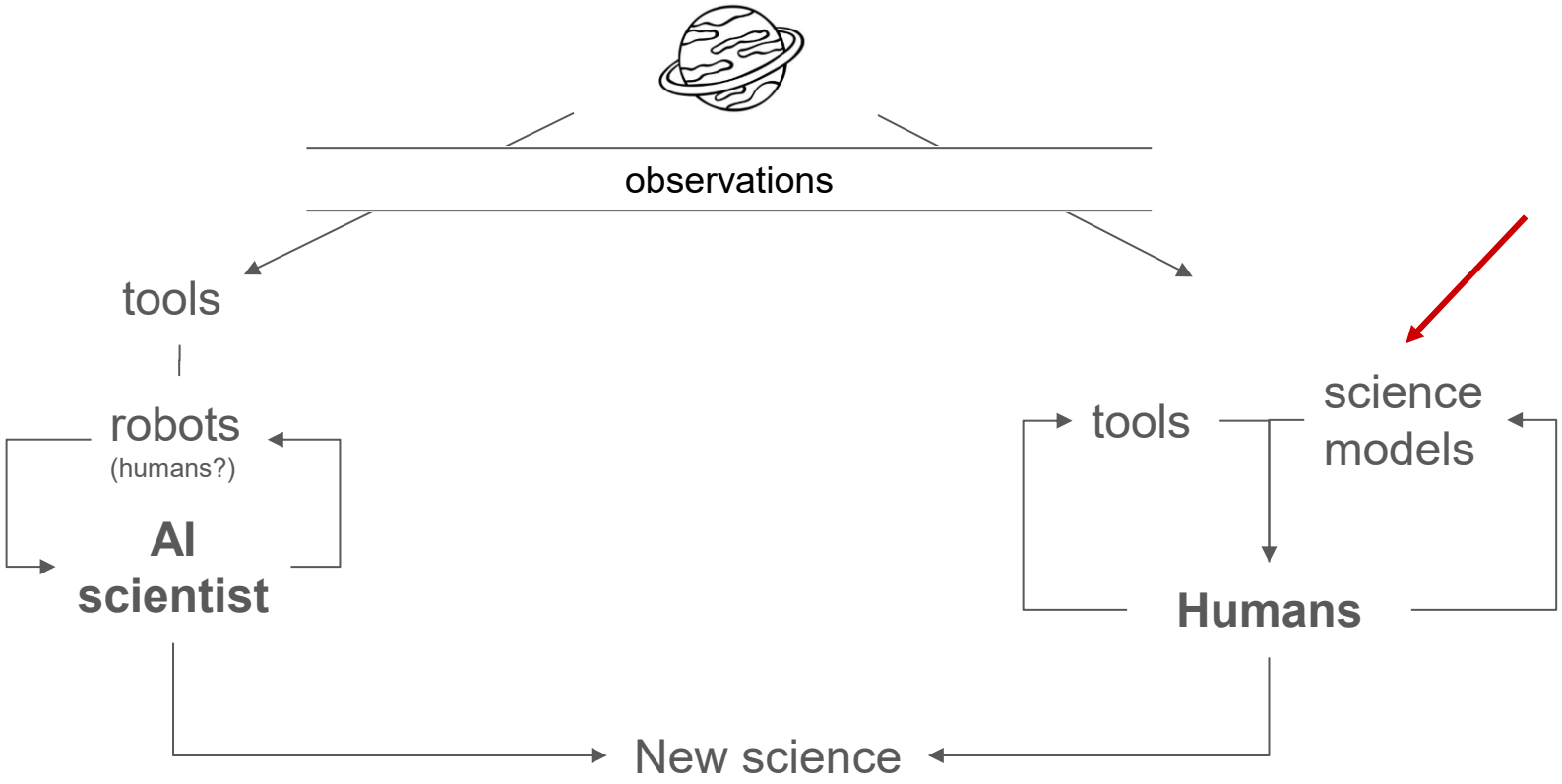
Take notice of the distances of the planets from one another, and recognize that almost all are separated from one another in a proportion which matches their bodily magnitudes. Divide the distance from the Sun to Saturn into 100 parts; then Mercury is separated by 4 such parts from the Sun, Venus by $4+3=7$ such parts, the Earth by $4+6=10$, Mars by $4+12=16$. But notice that from Mars to Jupiter there comes a deviation from this so-exact progression. From Mars there follows a space of $4+24=28$ such parts, but so far no planet or satellite was sighted there. But should the Lord Architect have left that space empty? Not at all.

Stanley L. Jaki, *"The Original Formulation of the Titius-Bode Law,"* Journal for the History of Astronomy 3: 136–138 (1972).

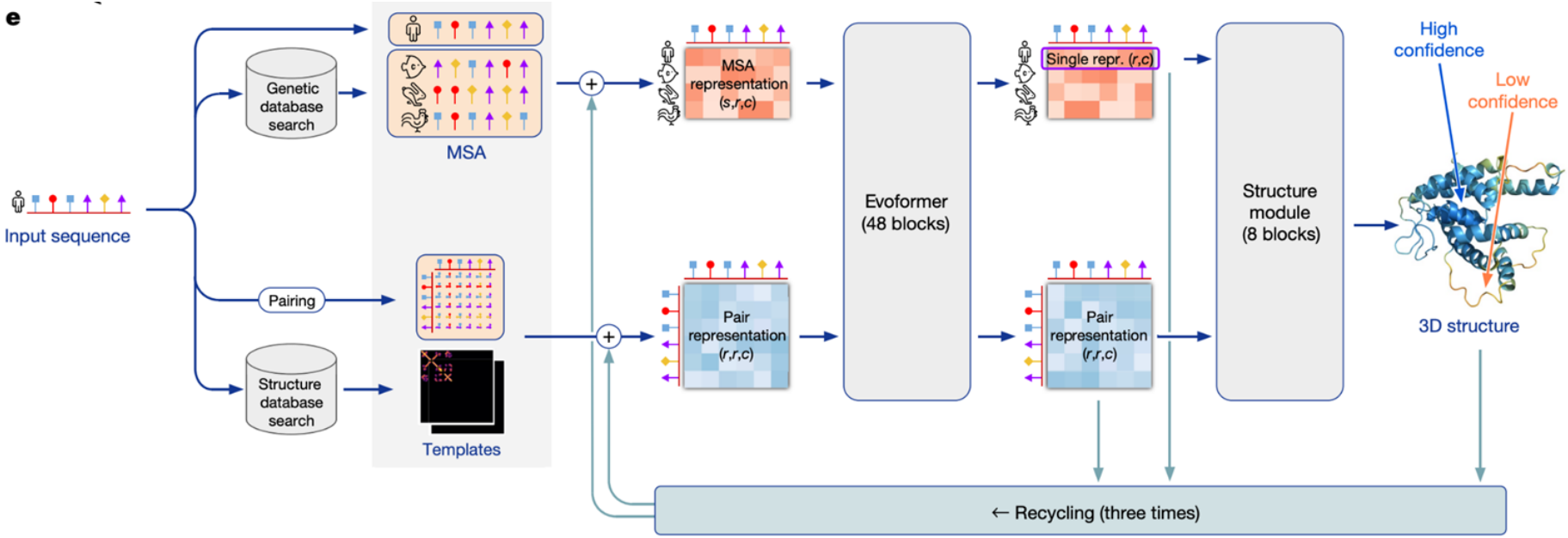
What do these two stories have in common?



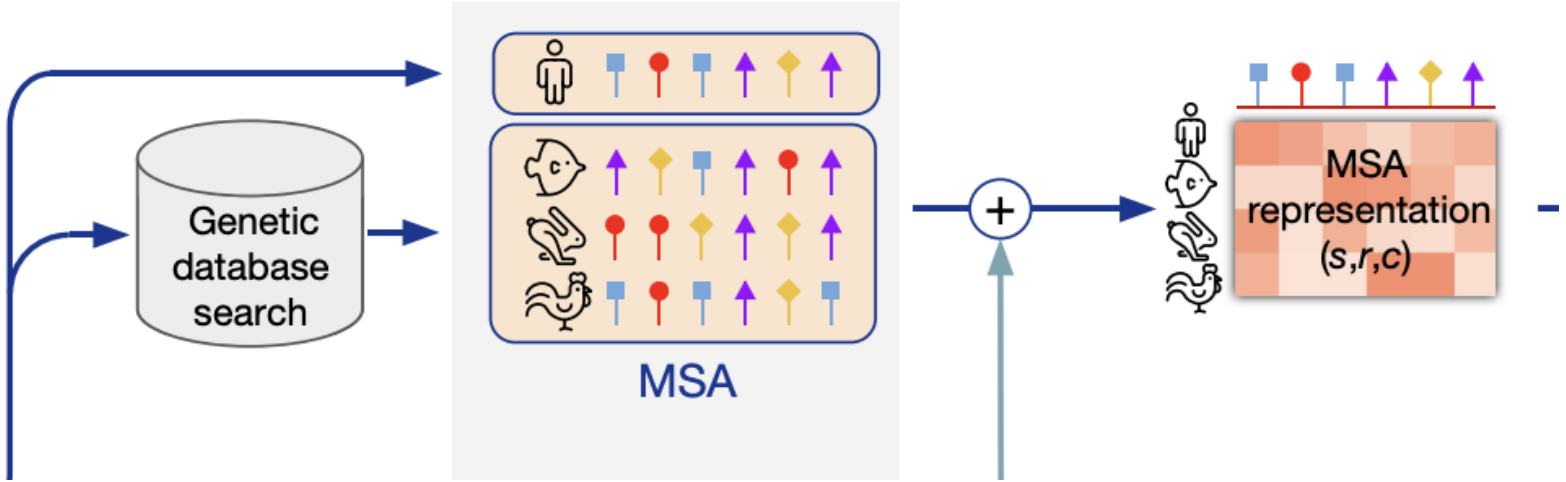
What about AI?



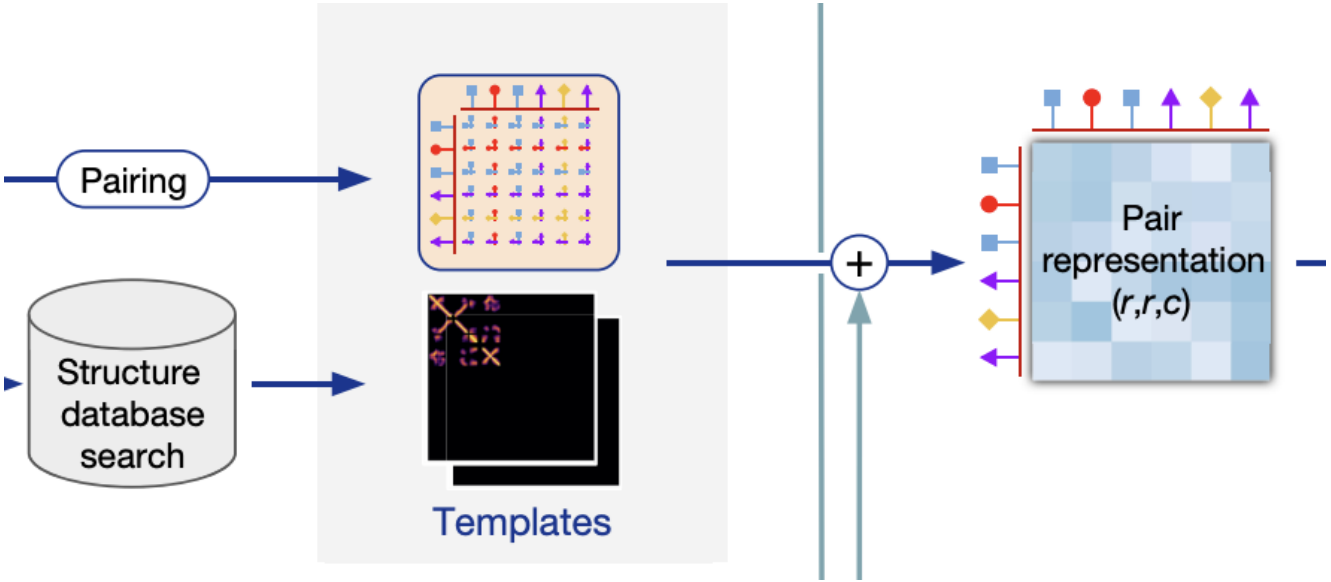
Two timelines



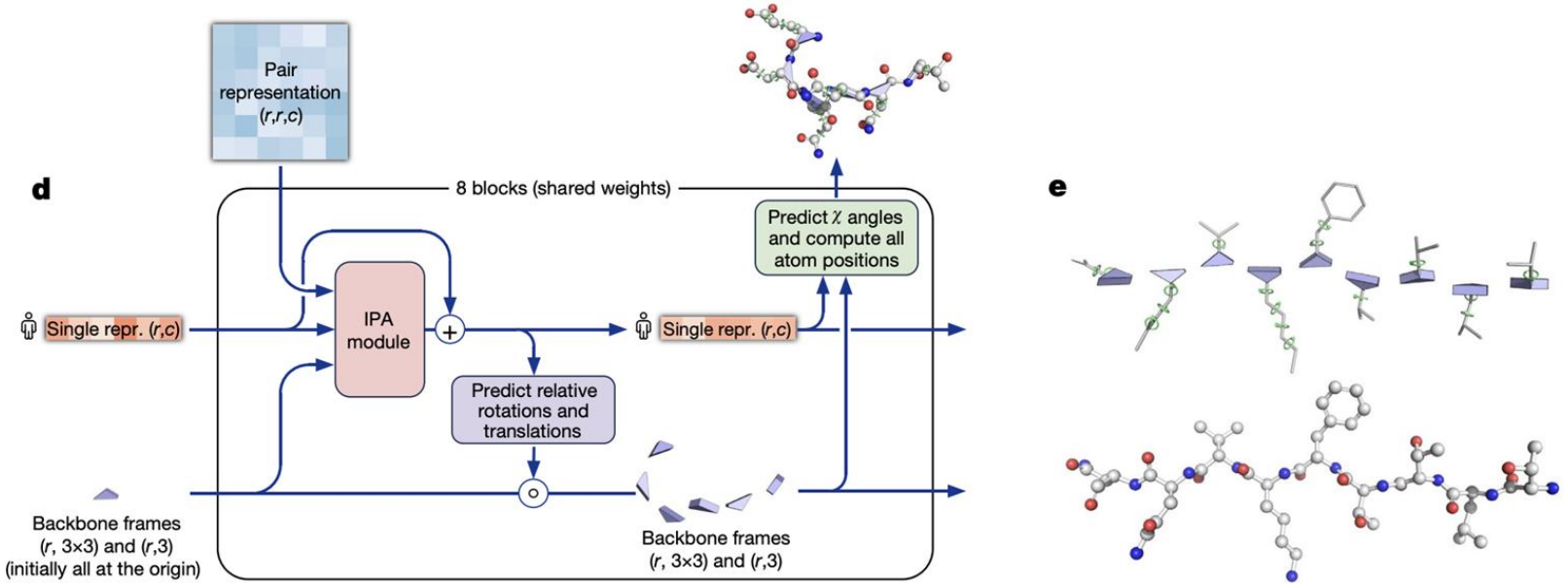
AlphaFold 2 (2021) uses strong architectural priors



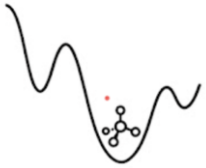
Input prior: Evolution re-uses structures



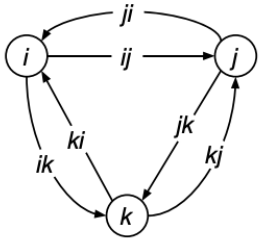
Input Prior: the *relationships* between the residuals are important



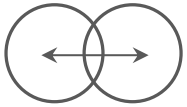
Output Prior: the *relative* positions/rotations matter



Post-output energy minimization



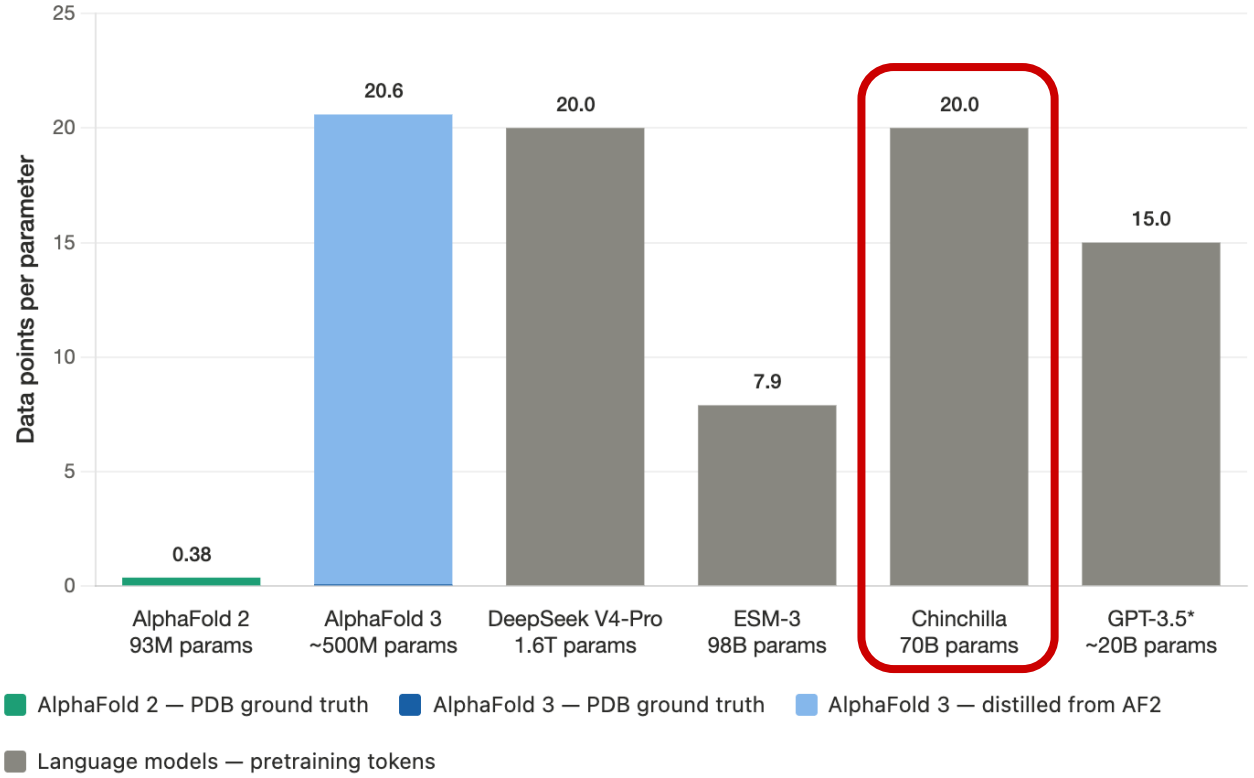
Enforce triangle inequality in pairwise updates



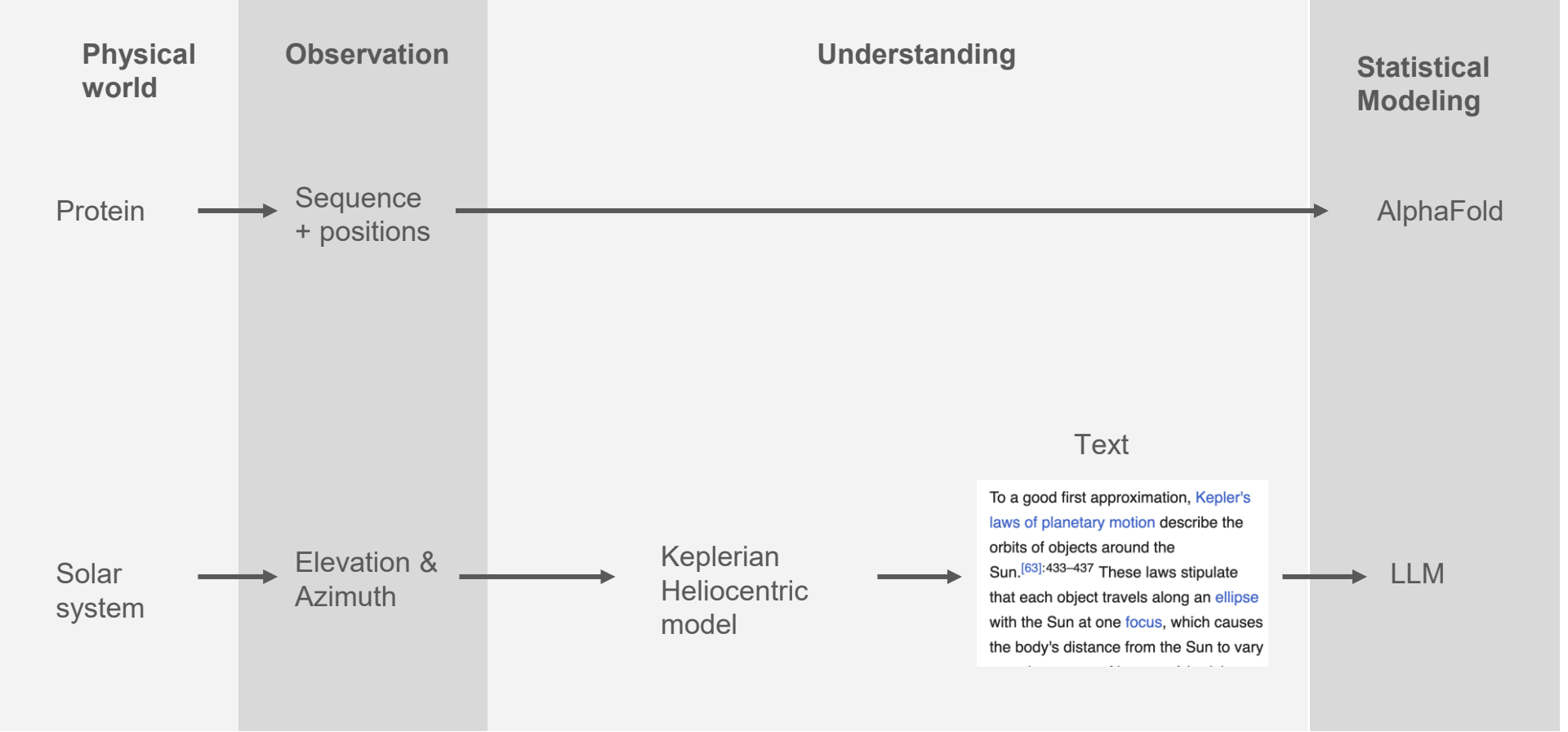
Atom clash loss

Output Priors: physics should not be violated

data points (tokens, residues) per model parameter



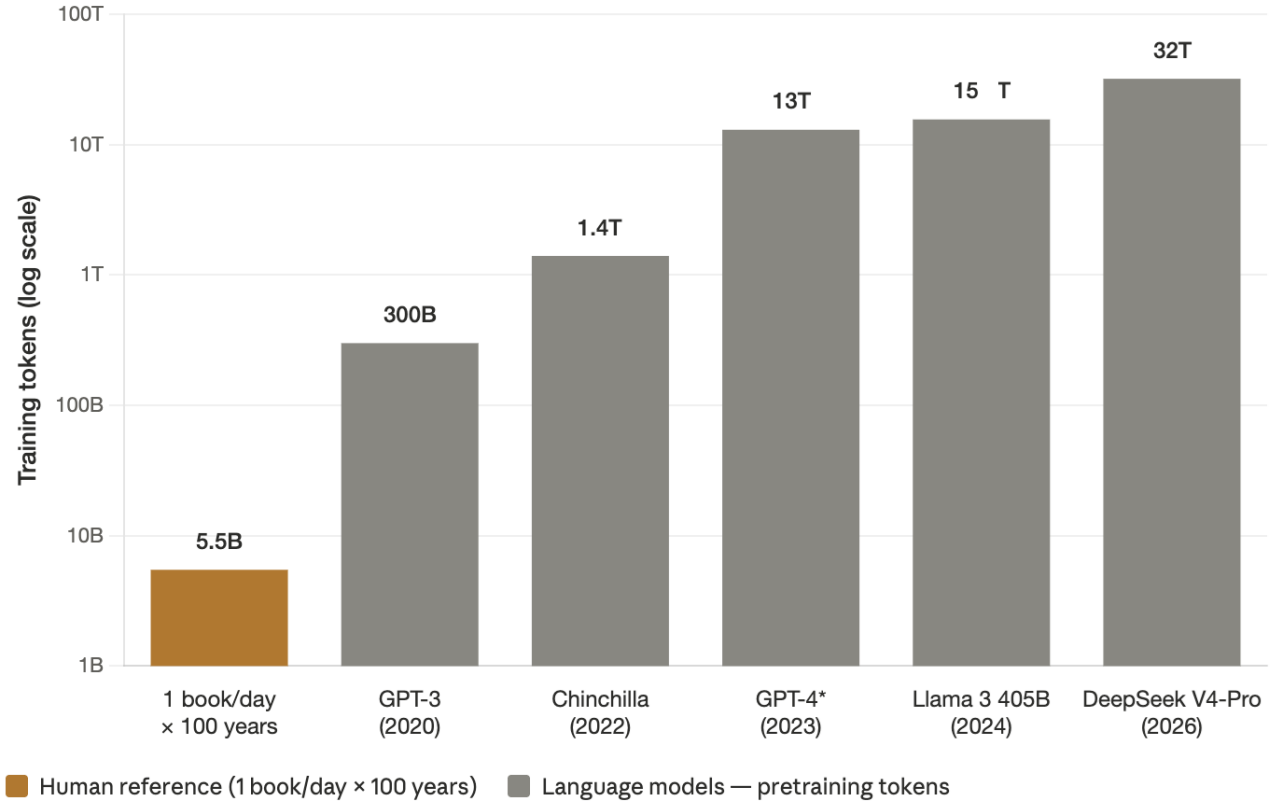
Why AlphaFold 2 needs specialized inductive priors



Text, unlike raw observations, encodes the output of human causal understanding

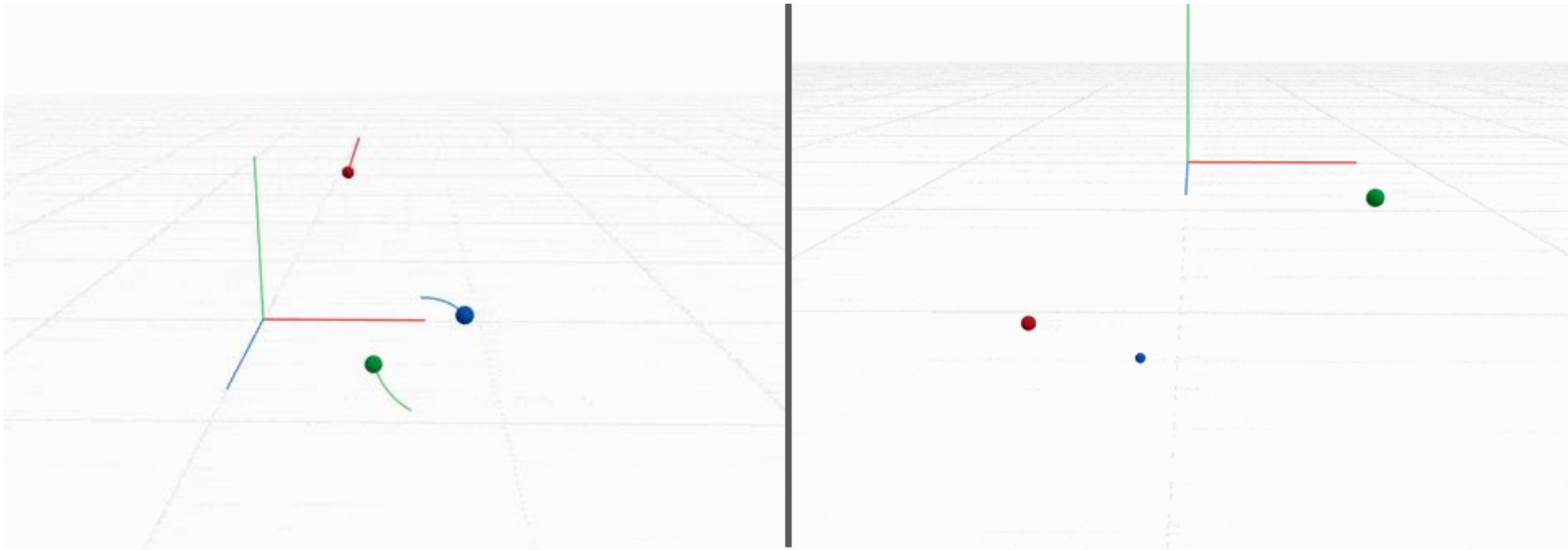
Problem: Small datasets necessitate specialized inductive priors

Solution: build **bigger** datasets?



The Age of Scale

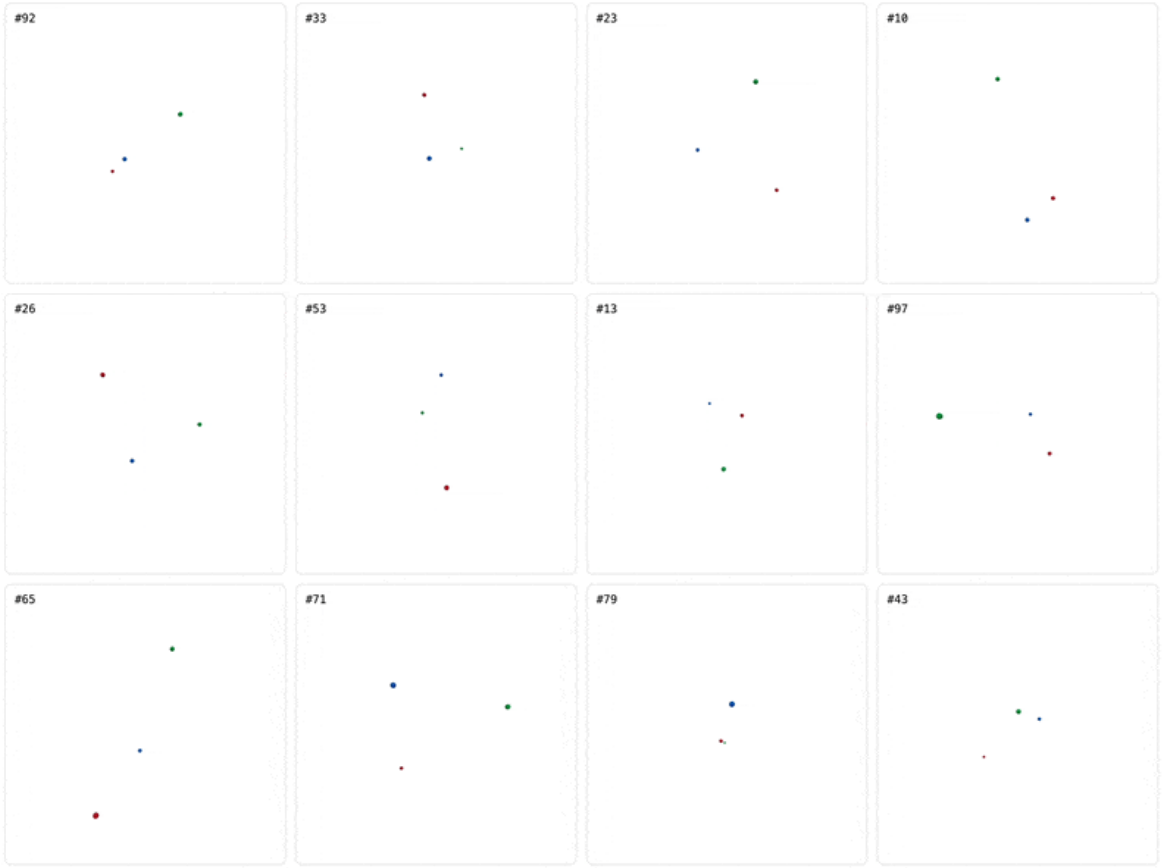
Will general model architectures (e.g. transformers) + enormous **observational** datasets generate **new science**?



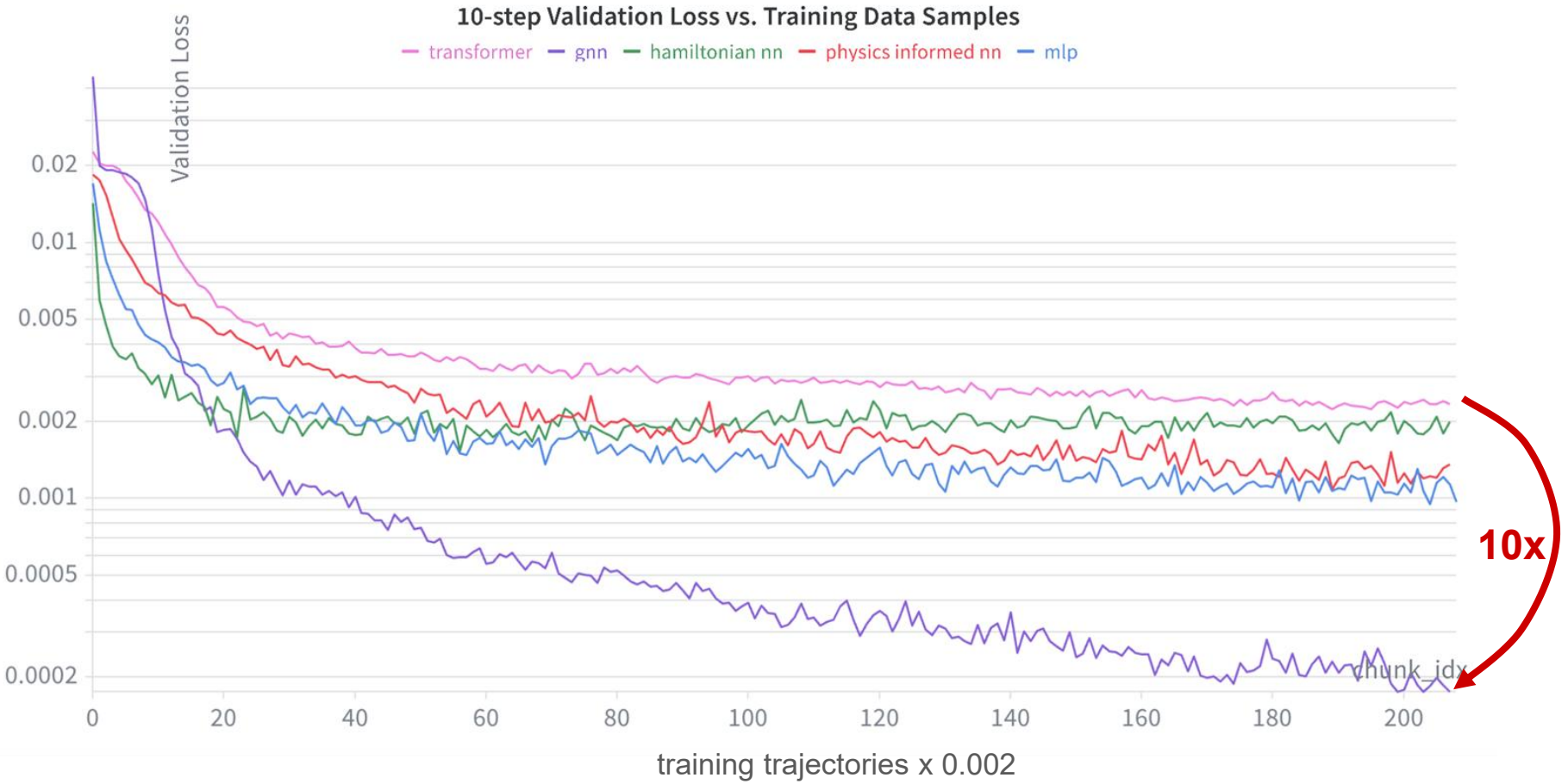
3-body dynamics: simple physics → complex observations

Model	# params	configuration	Inductive Priors
transformer	36,522	`d=36, n_heads=4 n_layers=2, ctx_len=8`	
mlp	37,001	`d=128, n_layers=3`	
physics loss mlp	37,001	same as mlp	energy conservation loss
hamiltonian mlp	35,969	same as mlp	model computes hamiltonian, derivatives give velocities
gnn	38,774	`n_nodes=3,d_node=40, d_msg=32 num_layers=2, n_messages=3`	pairwise message-passing between bodies

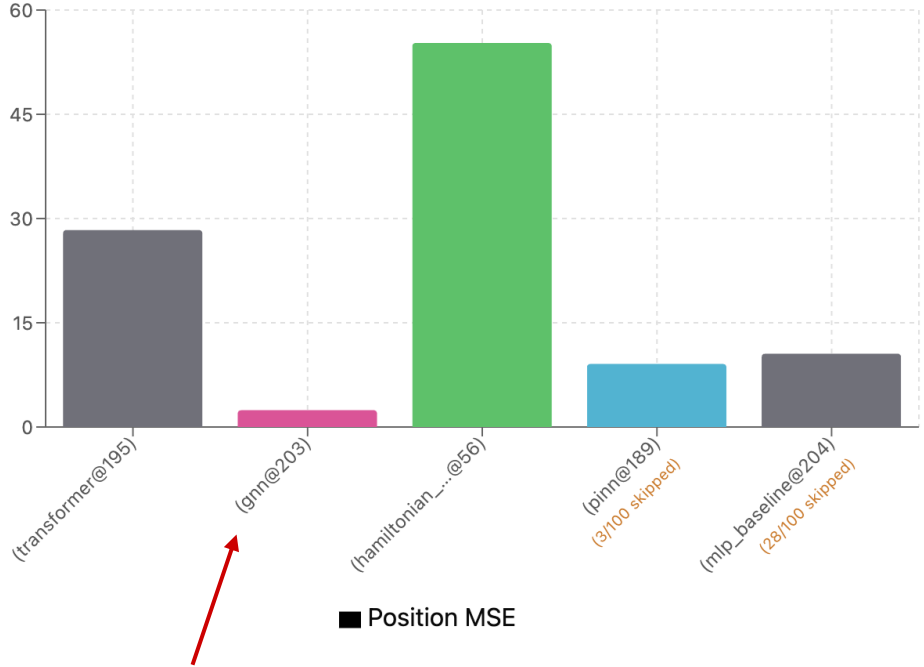
Baseline: 5 models, ~equal param #, ~560 D/N

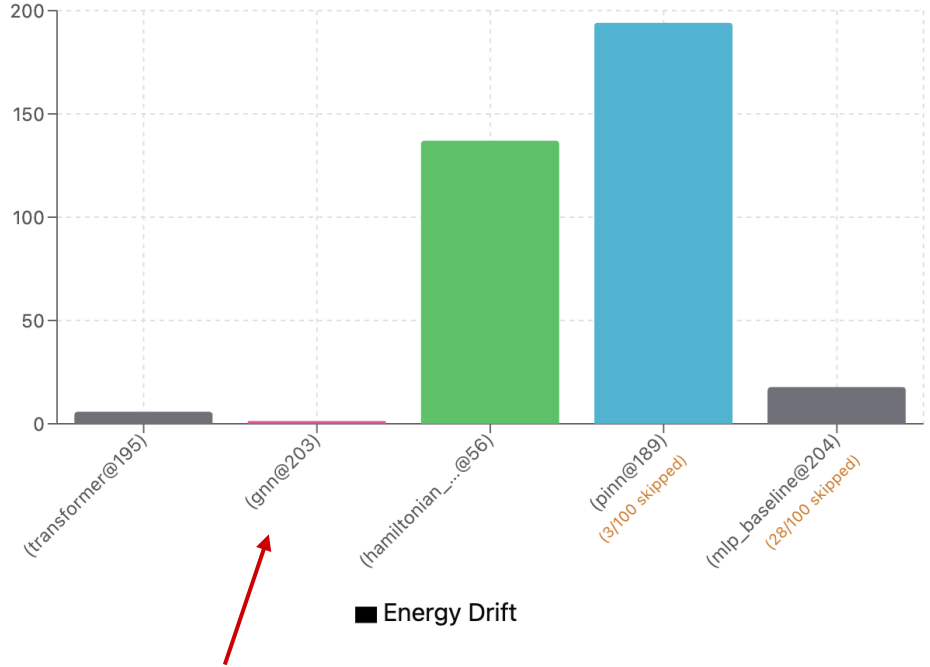


Trained on 100k trajectories (2M observations)
filtered for linearity within dt=0.05s

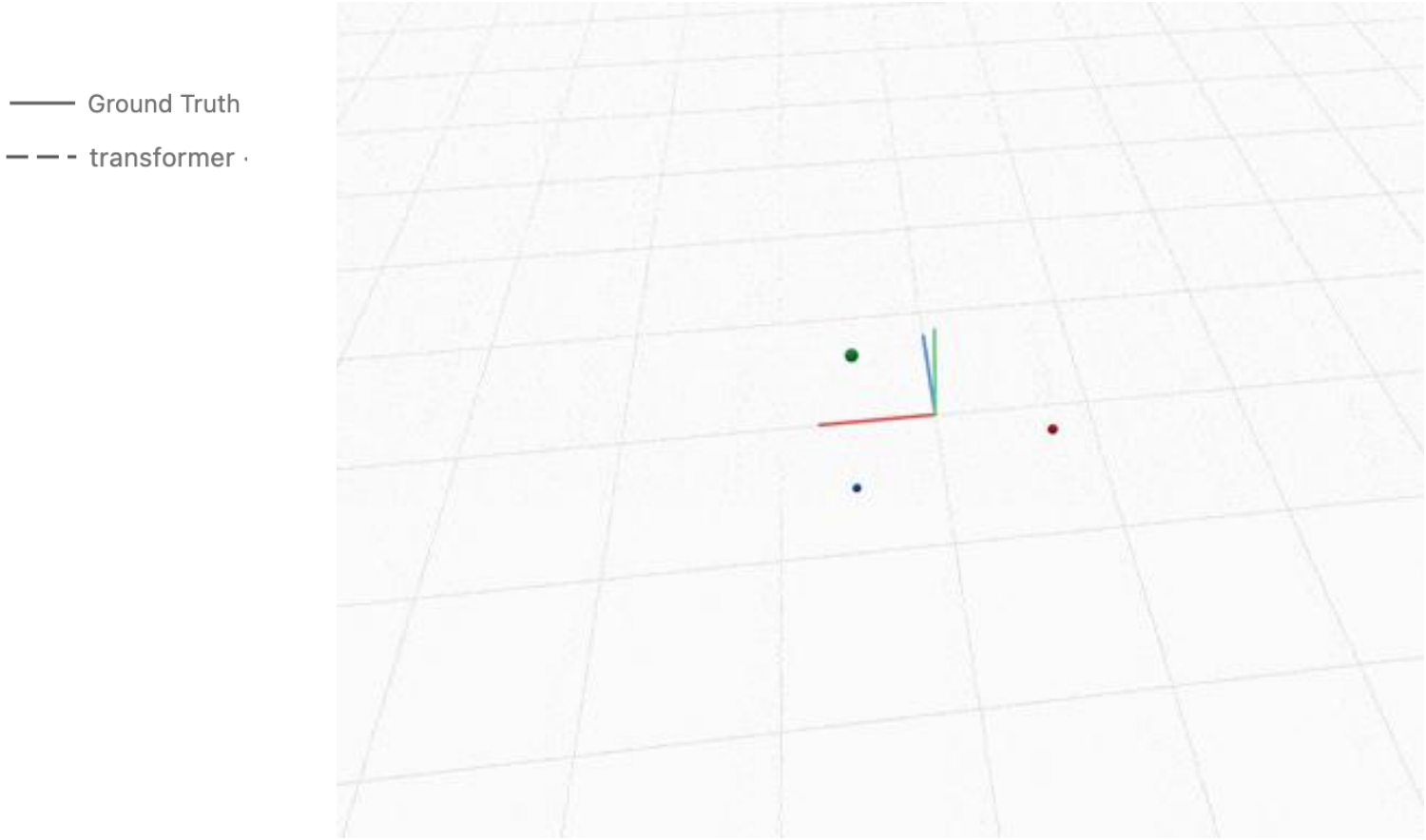


GNN prior results in >10x performance over the same data (~560 D/N)





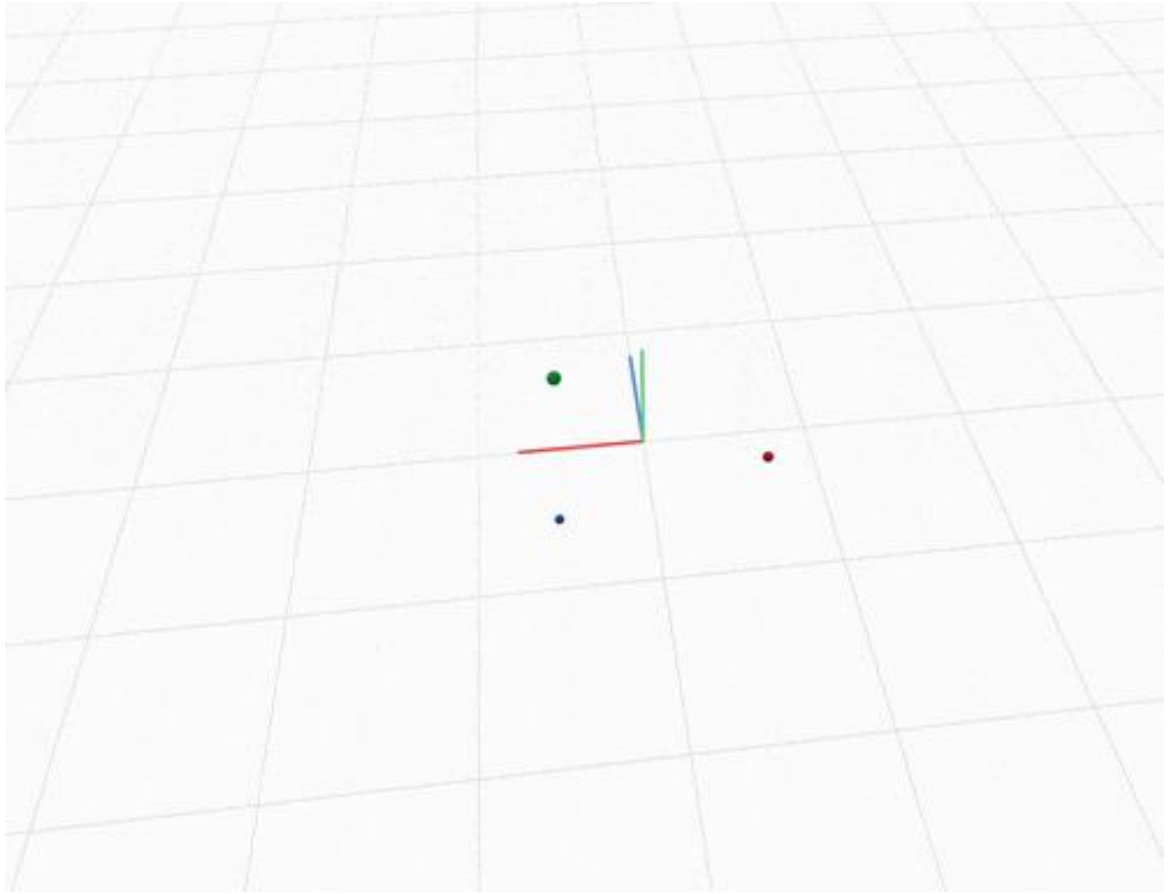
GNN wins in 100-trajectory validation dataset



Transformer sample prediction

Ground Truth

gnn



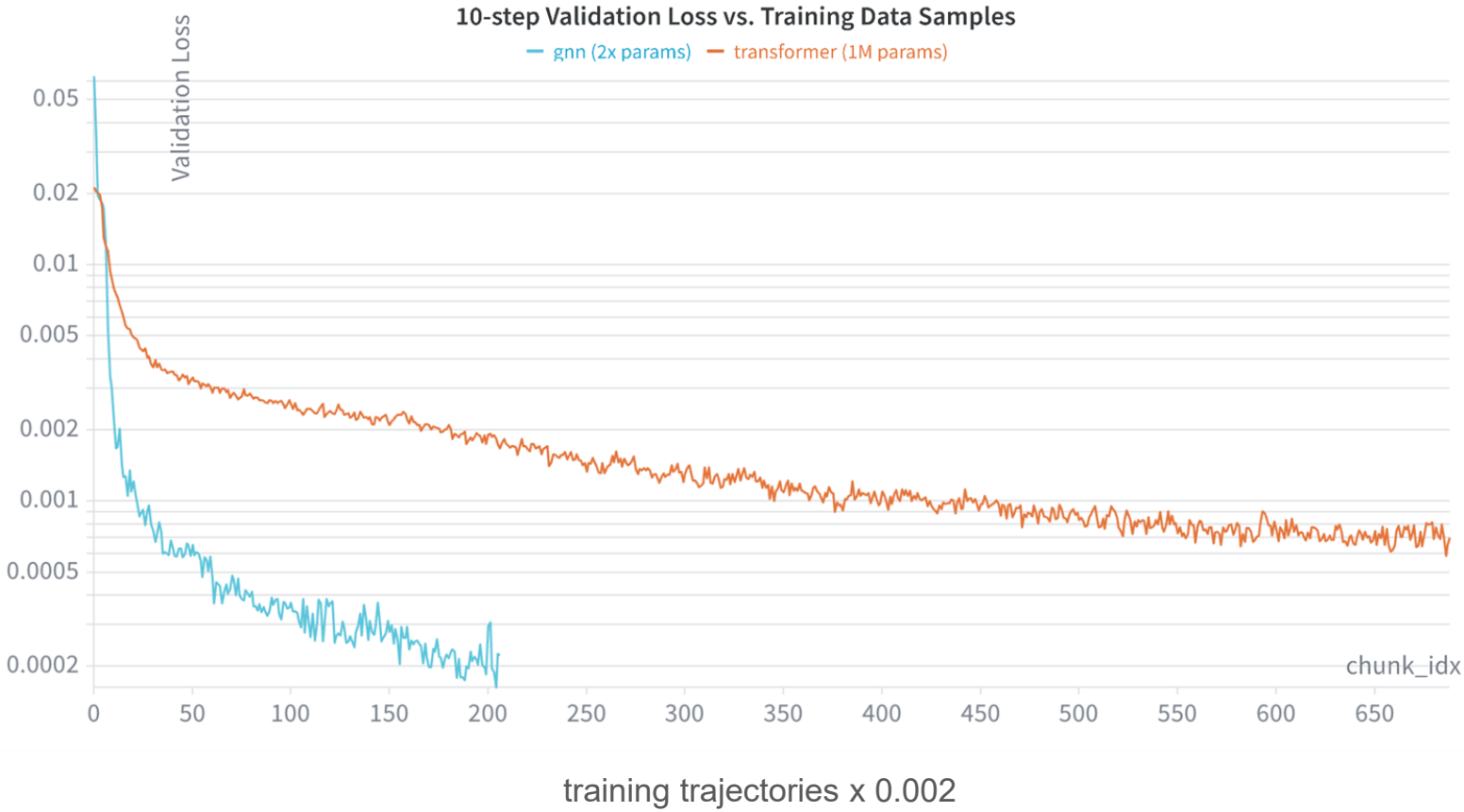
GNN prediction sample

Models with inductive priors are more data-efficient

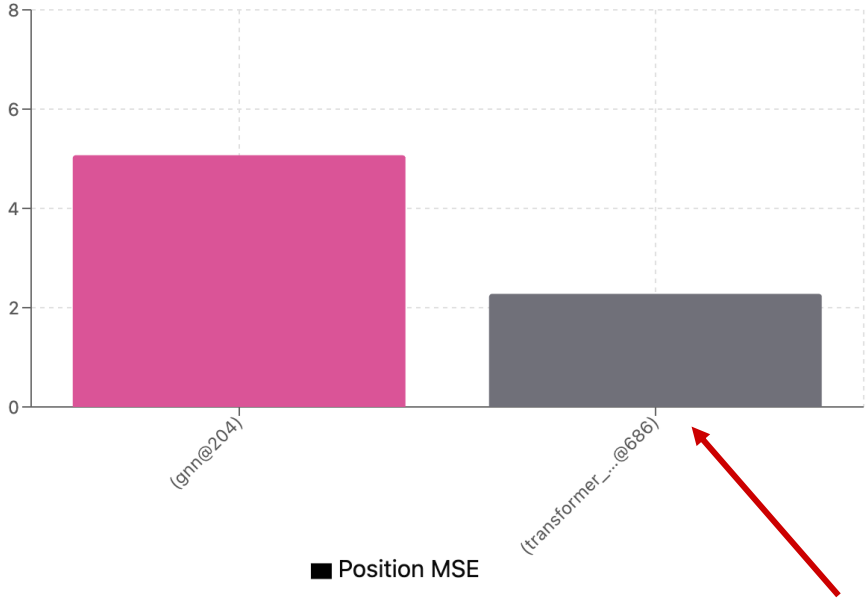
What if data weren’t the limitation?

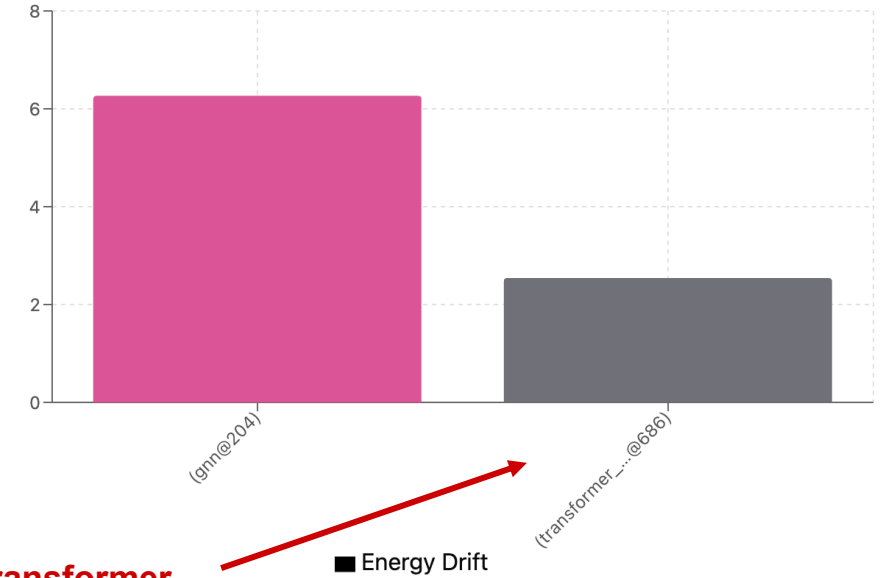
train **bigger** models on more data?

2x params (~60k params), same data (~256 D/N) → minimal improvement



1M param transformer (67 D/N) vs. 60k param GNN (256 D/N)

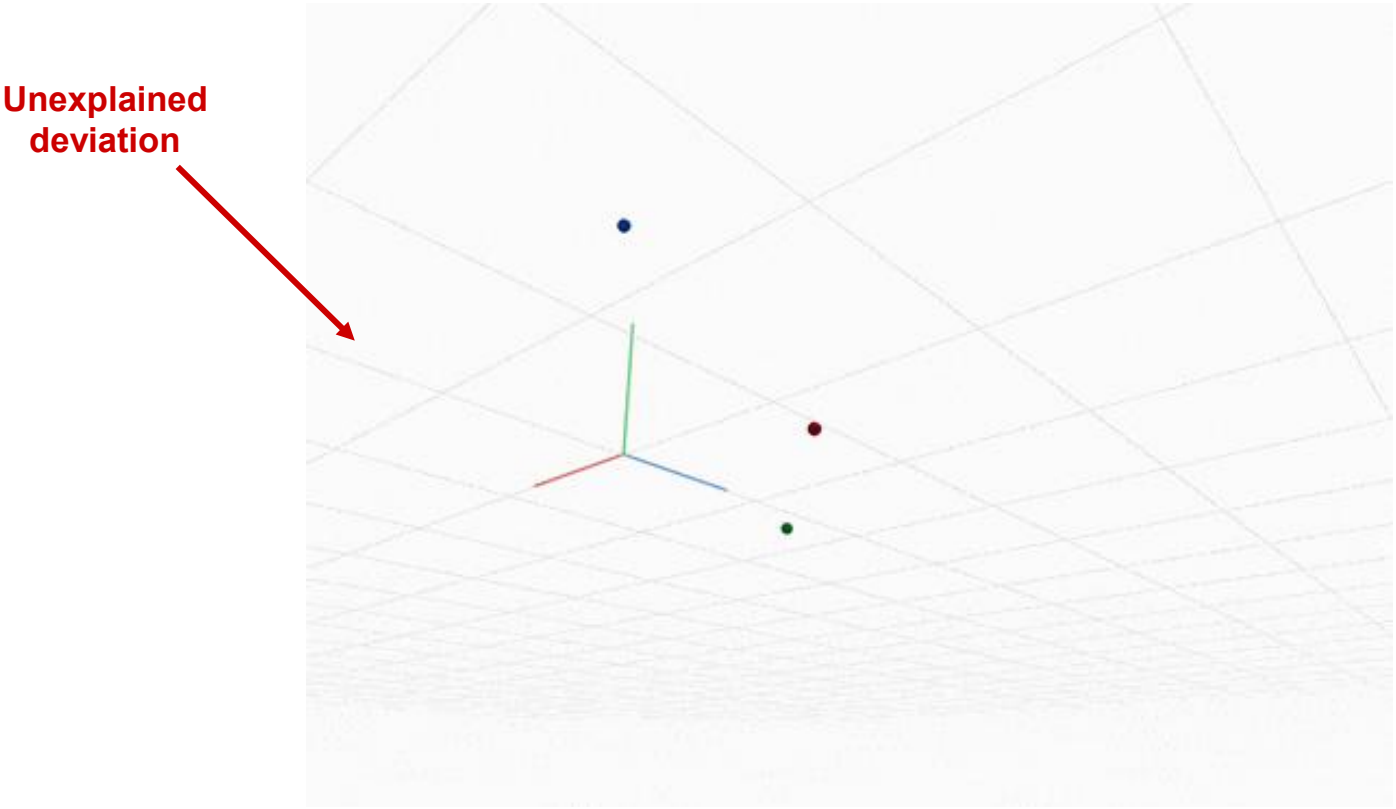




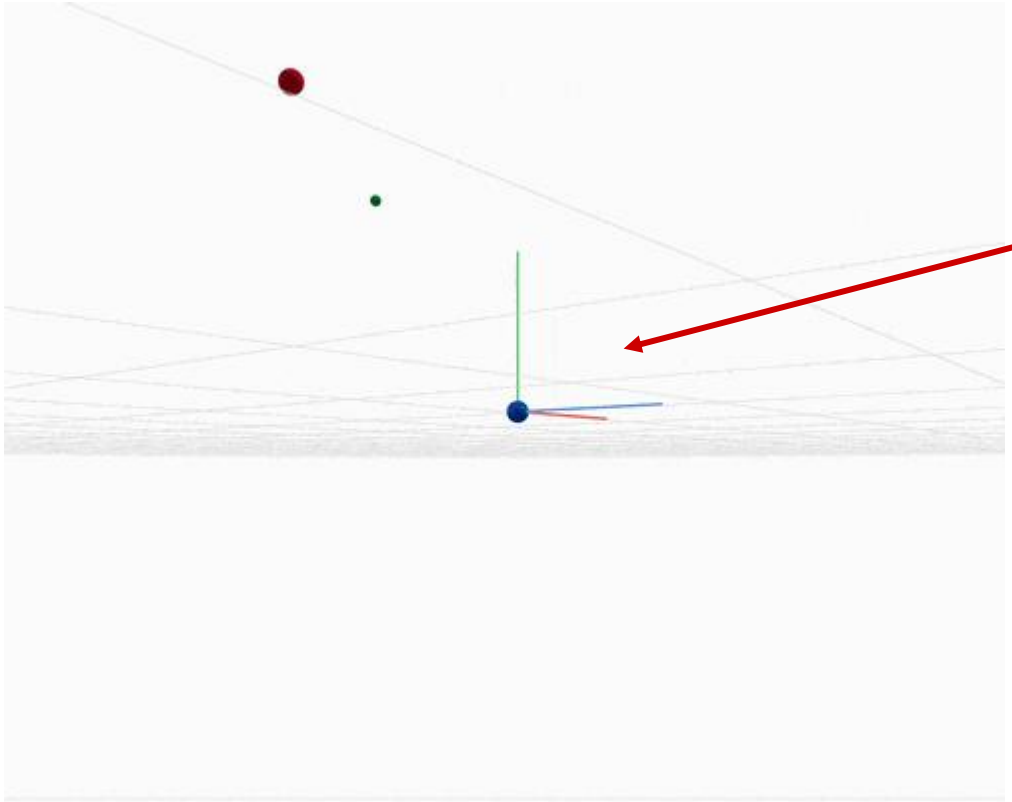
Transformer wins by >2x

Transformer (1M params) wins vs GNN (60k params) on In-distribution 100-trajectory validation

Will massive data and a large enough model (>10x kolmogorov complexity) learn the **underlying dynamics**?



Transformer (1M params): In-distribution initial conditions



Transformer (1M params):
out of distribution (planar motion) initial conditions

In the age of scale, model design still matters because...

Nutriment	Effect on pigeons	Effect on guinea-pigs	Number of examined animals	Molar teeth: Examined in number of animals	Molar teeth: Found loose in number of animals	Number of animals the osseous system of which was examined microscopically
Oats	0	Death after 22–28 days	4	4	4	4
Rye	Not examined	24–28 ,,	4	4	4	4
Wheat	0	25–29 ,,	4	4	4	4
Barley	0	26–46 ,,	7	3	3	3
Oaten groats	0	28–33 ,,	9	9	9	9
Barley groats	Death after on an average 5 weeks	21–41 ,,	13	2	2	2
Rye-bread baked with yeast	In some cases some loss of weight; but otherwise no effect after 4 months	15–46 ,,	10	6	6	9
Do. baked with baking-powder	Do.	32–37 ,,	4	0	0	3
Rye-bread baked with yeast, and oats	Not examined	27–29 ,,	2	2	2	2
Wheat-bread baked with yeast	Death after 3–3½ months	23–36 ,,	6	4	4	4
Do. baked with baking-powder	In most cases death after 30–51 days, in some cases after about 3 months	33 ,,	2	0	0	0
		Total	65	36	36	44

Holst, A., & Frølich, T. (1907). Experimental studies relating to ship-beri-beri and scurvy

1 - Frontier observational data is scarce.
Data-efficient training relies on strong inductive priors.

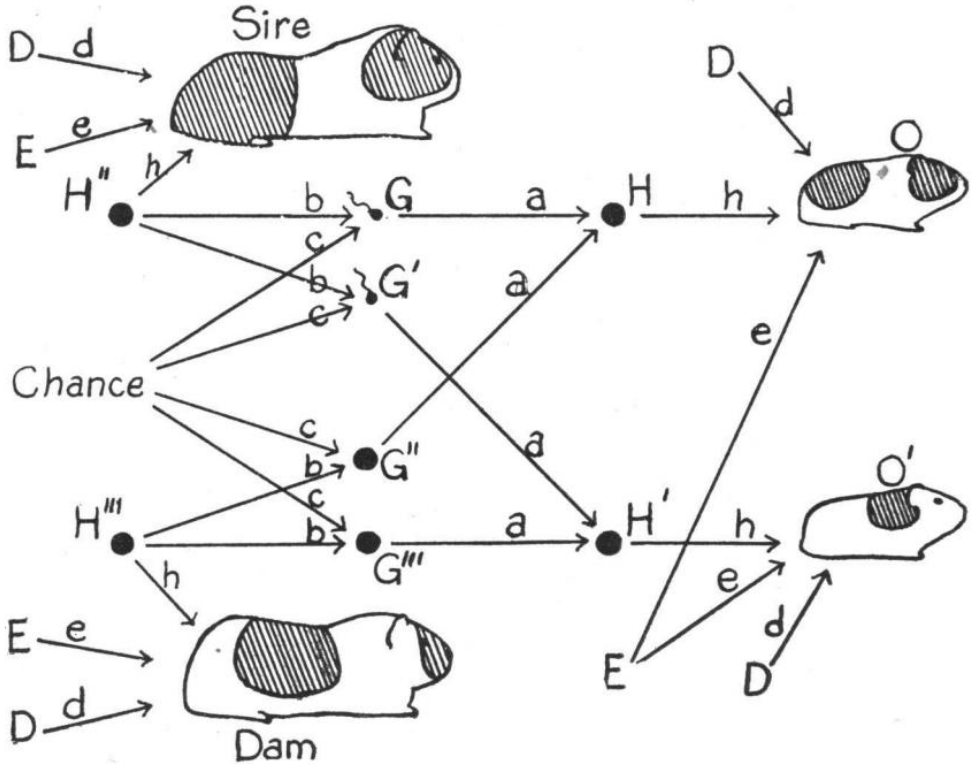
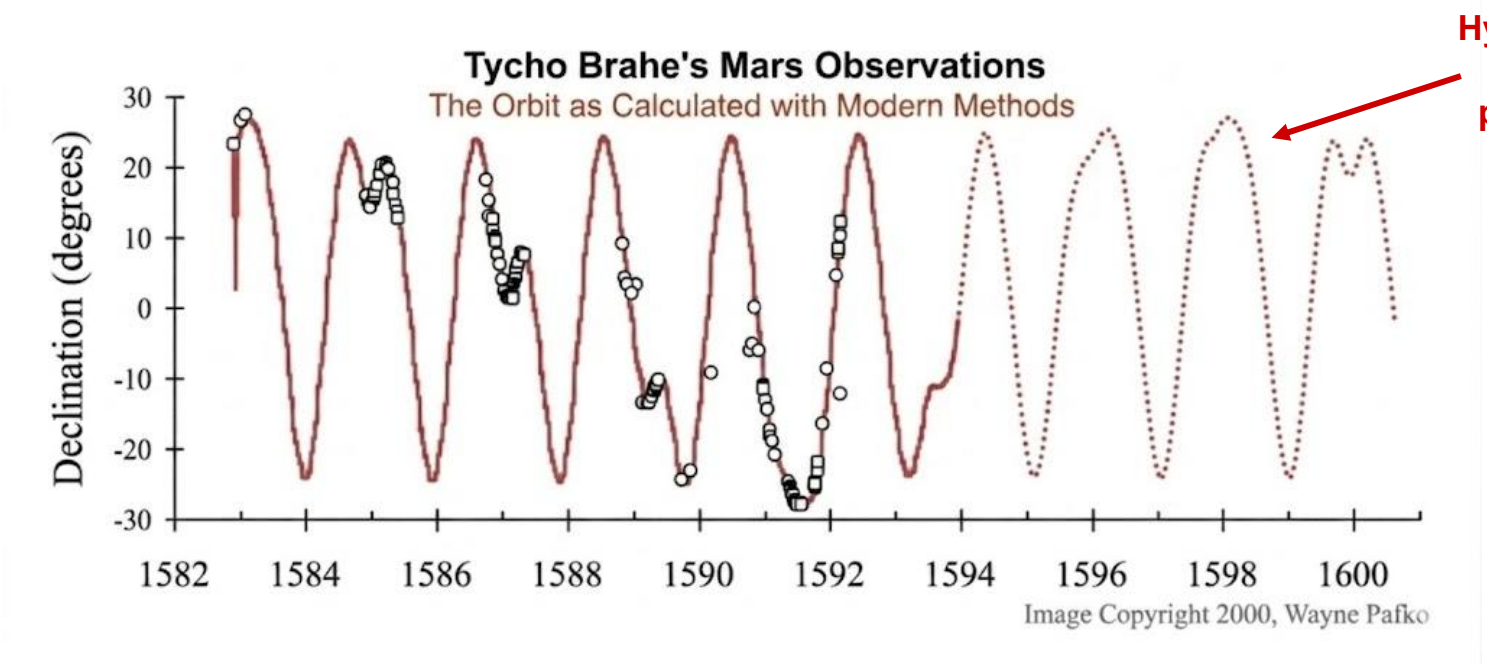


FIG. 5.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea-pigs

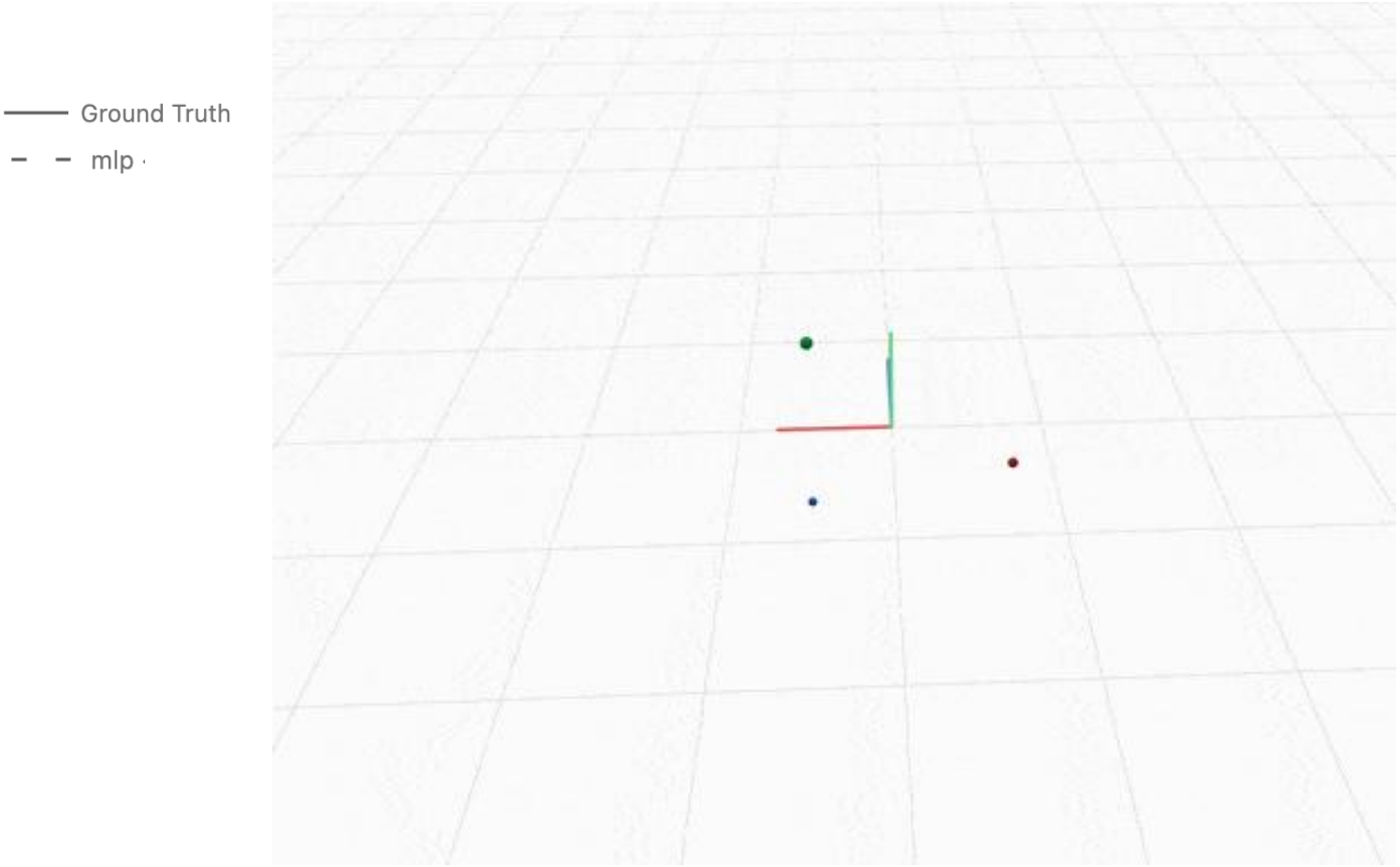
2 - Models trained on observations, **unlike text,** need inductive priors to uncover causal mechanisms, even with large datasets



3 - Even perfect prediction is a precursor to understanding. Inductive priors *hint* at how models make predictions.

Appendix

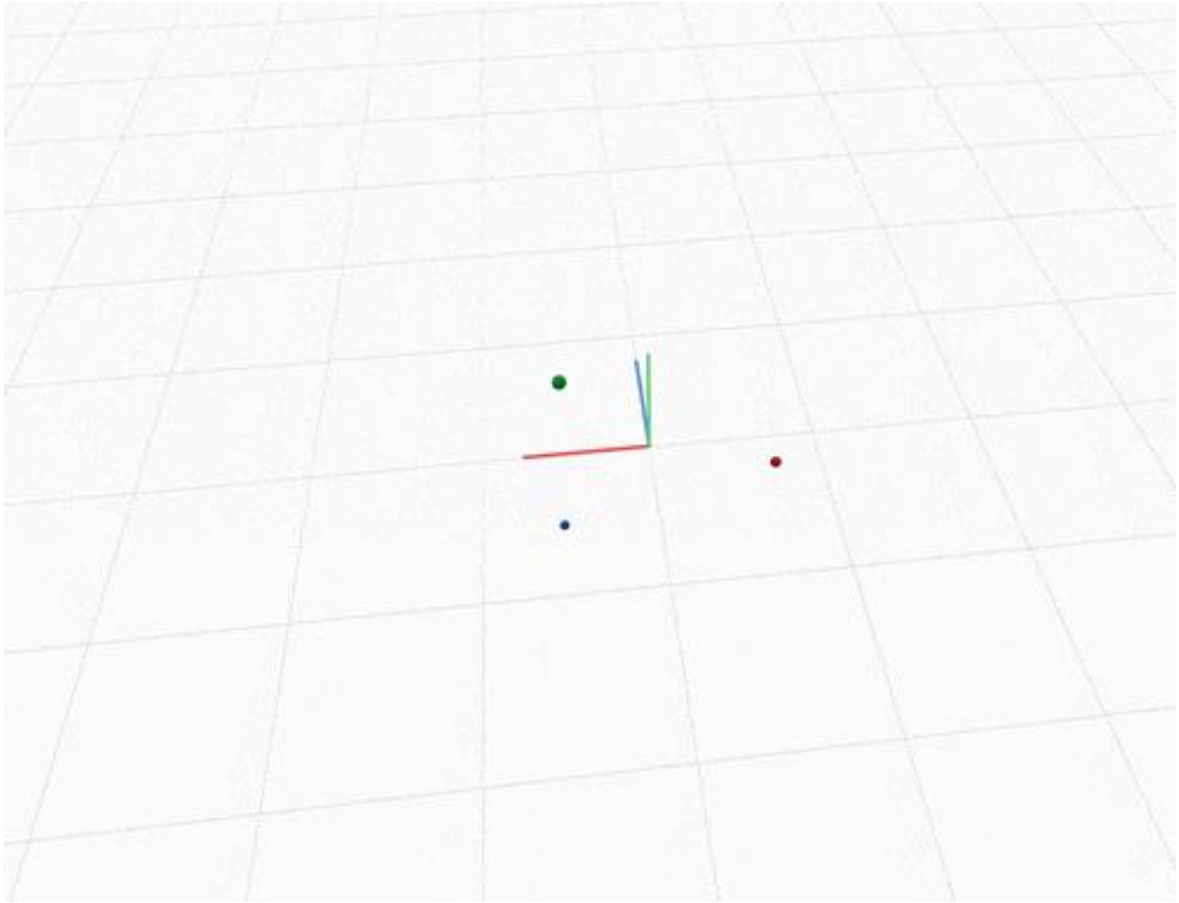
Nima Keivan



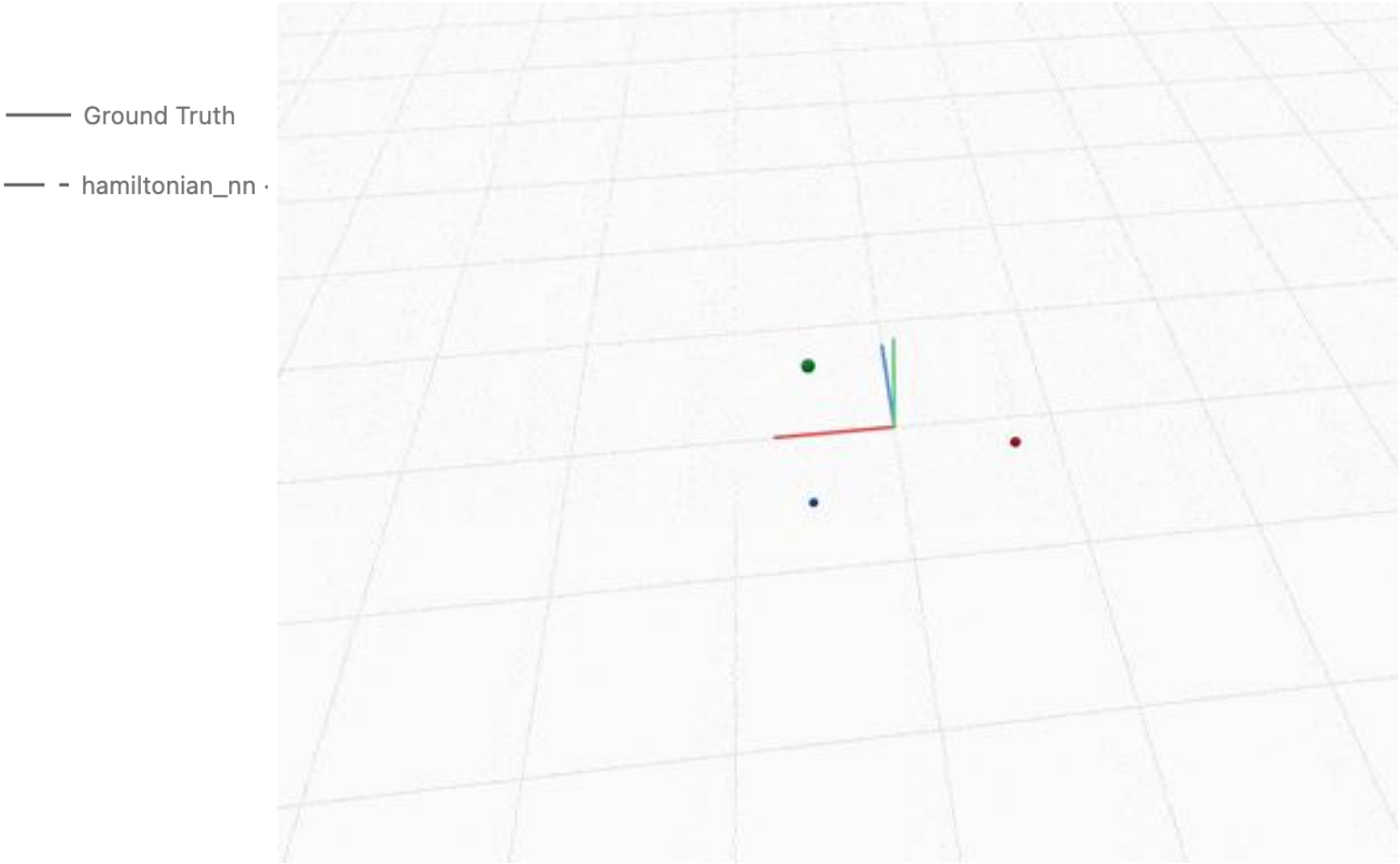
MLP sample prediction

Ground Truth

pinn



Physics-loss MLP sample prediction



Hamiltonian MLP sample prediction

Planet	T (years)	a (AU)
Mercury	0.2408	0.3871
Venus	0.6152	0.7233
Earth	1.0000	1.0000
Mars	1.8808	1.5237
Jupiter	11.863	5.2029
Saturn	29.448	9.5367

Kepler’s data