The Platonic Universe: Do Foundation Models See the Same Sky?

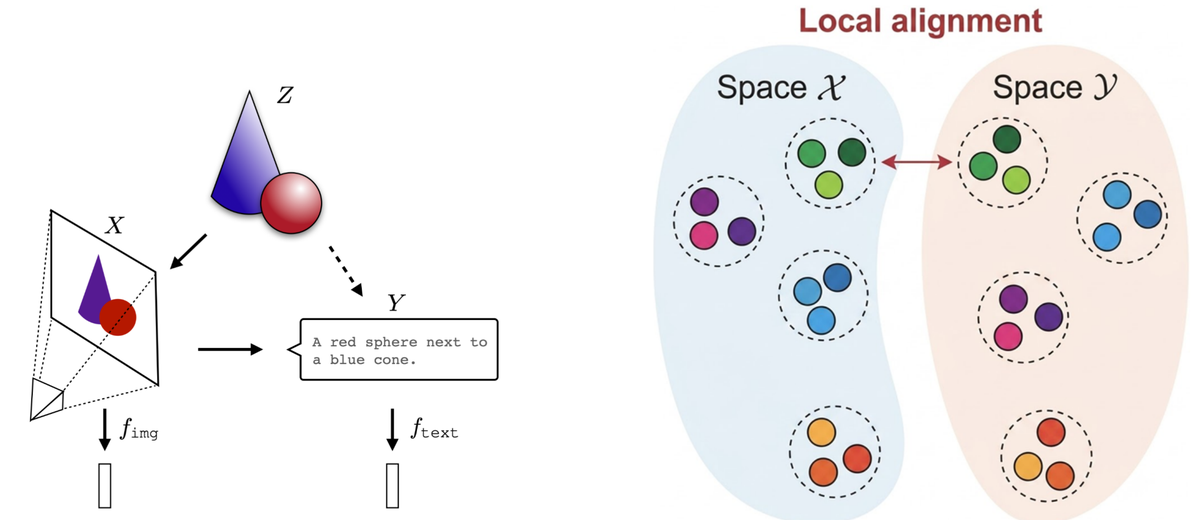
Trinidad Borrell[1,2] Steven Dillmann[3] Kshitij Duraphe[4] Furkan Eris[5] Ashod Khederlarian[6] Aman Kumar[7] Giovanni Marraffini[8,9] Michael J. Smith[10,11] Shashwat Sourav[12,13,14] Rocco Di Tella[10,11] John F. Wu[15,16]

[1]Paris Brain Institute [2]Sorbonne Université [3]Stanford University [4]UniverseTBD [5]Prolific [6]University of Pittsburgh [7]IUCAA [8]INRIA [9]Sigma Nova [10]AstroAI [11]Center for Astrophysics | Harvard & Smithsonian [12]Washington University in St. Louis [13]Oak Ridge National Lab [14]Lawrence Berkeley National Lab [15]Space Telescope Science Institute [16]Johns Hopkins University

ICML International Conference On Machine Learning

Introduction

The **Platonic Representation Hypothesis** says that "all foundation models, regardless of their architecture, training data, or training style", converge to a similar statistical representation of the world as they grow bigger. The **Aristotelian Representation Hypothesis** refutes this, saying that representations are converging to **shared local relationships**.



(a) The Platonic Representation Hypothesis (Huh et al, 2024)

(b) The Aristotelian Refinement (Gröger et al, 2026)

Figure 1. Platonic vs Aristotelian Representation Hypothesis

We investigate this claim for astronomy.

Why astronomy? and how?

Astronomy serves as a natural testbed for representational convergence.

- Observations are different projections of the same underlying physics
- A galaxy's morphology, photometry, and spectra emerge from the same physical processes
- Modern astronomical surveys have massive scale

To test this, we evaluate 11 frozen foundation models families—spanning $\mathcal{O}(10M) \rightarrow \mathcal{O}(10B)$ parameters—on crossmatched JWST, HSC, Legacy Survey, and DESI data using physical linear probes and geometric alignment metrics.

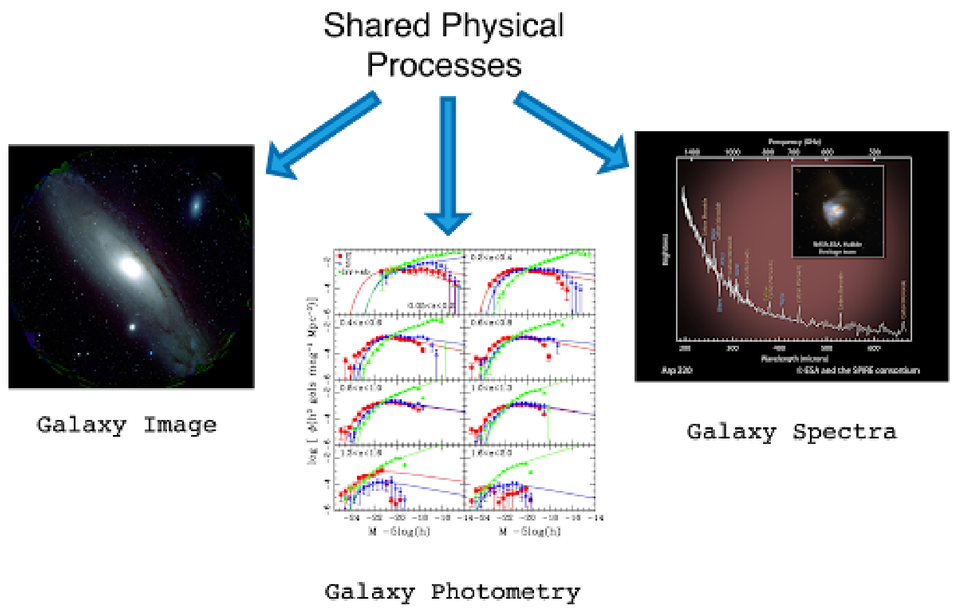


Figure 2. Astronomy as a testbed

Our science questions

We structure our experiments to ask first whether foundation models encode galaxy physics, and then whether the geometry of their embeddings converges in ways that track this physical content. We organize our analysis around three primary questions:

1. **Do larger models encode more galaxy physics?** We train linear probes on frozen embeddings to predict redshift, stellar mass, and specific star formation rate given galaxy images, and ask whether probe performance scales with model capacity.
2. **Do different architectures converge to the same physical representation?** We probe crossmodal performance, measure the distance between probe weight vectors across models, and ask whether they organize physics similarly.
3. **Does geometric alignment scale with performance locally, globally, or both?** We measure local (MKNN) and global (CKA) alignment within model families, between astronomical modalities, and across architectures, regressing alignment against probe performance.

Result 1: Capacity Translates to Physical Knowledge

To answer our first question, we trained linear probes on the frozen embeddings of each model to predict redshift (z), stellar mass (M_*), and specific star formation rate (sSFR).

- **Scaling:** We find a positive correlation between a model's parameter count and its ability to linearly encode physical information.
- **General-Purpose Power:** Larger models recover more astrophysics from images alone, even though almost all tested models were pre-trained on natural imagery or text rather than galaxies.

AstroPT is not always the best!

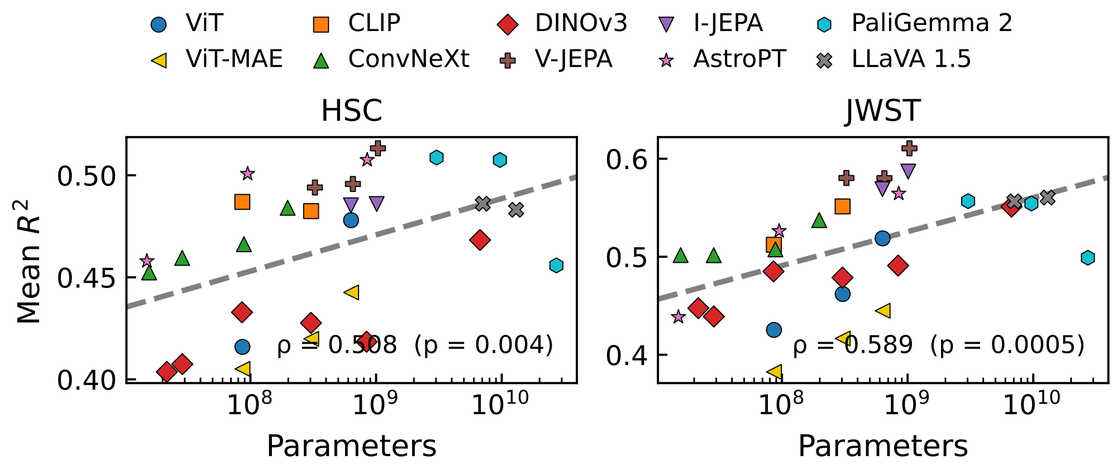


Figure 3. Linear probe R^2 as a function of model size, averaged over mass, sSFR, and redshift for HSC and JWST imagery.

Result 2: Stronger support for the Aristotelian Hypothesis

We test whether embedding geometries converge locally (via MKNN) or globally (via CKA).

- **MKNN > CKA** Across our tests, local neighborhood structure (MKNN) consistently tracks model physics performance, while global embedding similarity (CKA) does not.
- **Cross-Modal Alignment:** Local alignment holds even between DESI spectra and HSC imagery—modalities that share essentially no low-level statistics. Alignment must be driven by shared physical content.
- **Conclusion:** Foundation models agree on which galaxies are similar to each other (local structure), but do not agree on a shared global coordinate system. This strongly supports the Aristotelian Representation Hypothesis (ARH) over the strict Platonic hypothesis

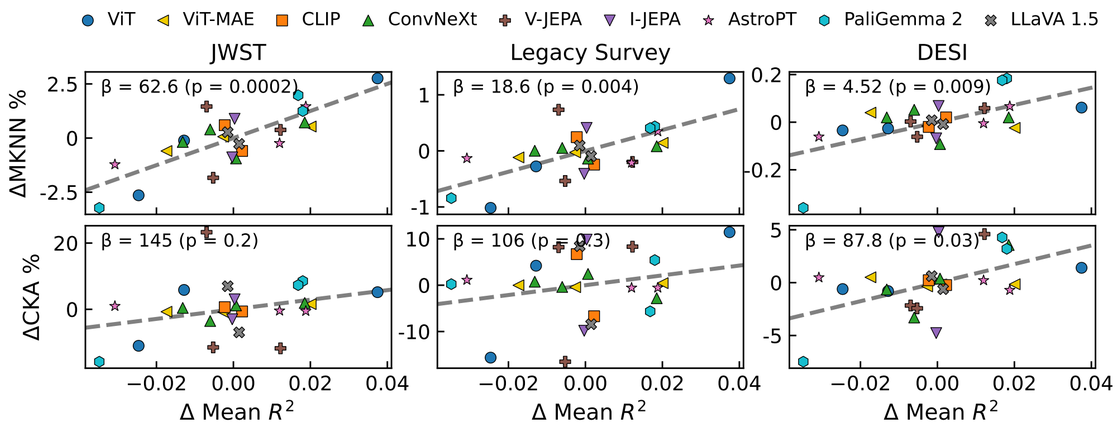


Figure 4. Local embedding similarity (MKNN) significantly correlates with physics probe performance across modalities, but global (CKA) doesn't correlate as much

Conclusions & Implications

- **Do foundation models see the same sky?: Locally**, but not globally.
- **Out-of-distribution performance:** Domain-specific architectures (like AstroPT) do not dominate. Sufficiently scaled general-purpose models recover representations just as well, indicating that scale and data diversity substitute for domain specificity.
- **Universal Structural Patterns:** The successful alignment between natural image-trained models and 1D spectral models demonstrates that scaled networks learn universal features that transcend their original training domains. This convergence is driven purely by the shared underlying astrophysics rather than shared low-level input statistics.

References

[1] F. Gröger, S. Wen, and M. Brbić.
Revisiting the Platonic Representation Hypothesis: An Aristotelian View.

[2] M. Huh, B. Cheung, T. Wang, and P. Isola.
Position: The Platonic Representation Hypothesis.

[3] The Multimodal Universe Collaboration.
The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data.