THE PLATONIC UNIVERSE: DO FOUNDATION MODELS SEE THE SAME SKY?

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate when foundation models converge towards shared representations, and how this convergence depends on model capacity, training regime, and model architecture. We take a 'science-for-AI' approach, using astronomy as an experimental instrument to test the Platonic Representation Hypothesis (PRH) and its Aristotelian refinement (ARH) against an external physical reference. The historical success of astrophysics is evidence that a compact, modality-invariant description of galaxy observables exists, and so representation convergence toward reality should be measurable against the physical parameters astronomers already use. Given this framework, we evaluate eleven foundation model families (spanning supervised classification, self-distillation, joint-embedding prediction, masked autoencoding, vision-language pre-training, and astronomy-specific architectures from $\mathcal{O}$(10M)→$\mathcal{O}$(10B) parameters) on crossmatched JWST, HSC, and Legacy Survey imagery, and DESI spectroscopy. All models are evaluated frozen, with no astronomy-specific fine-tuning. We probe redshift, stellar mass, and specific star formation rate via linear probes, and local (MKNN) and global (CKA) embedding geometry within families, between modalities, and across architectures. We find that physics performance scales predictably with capacity; probe directions align consistently with expected astrophysical correlations and selection effects; and local (but not global) embedding alignment tracks physics performance, including between DESI spectra and HSC imagery—modalities that share essentially no low-level statistics. Our results support the ARH over the strict PRH, demonstrate astronomy's value as an experimental framework for neural representation learning, and suggest that astro-foundation models can build on general-purpose pre-trained architectures, capitalizing on the broader open machine learning community's already-spent computational investment.

1 ASTRONOMY AND THE PLATONIC REPRESENTATION HYPOTHESIS

Three historical waves of increasingly automated connectionism have swept the shores of astronomy. The late 1980s brought MLPs tuned for astronomical applications on manually selected inputs (e.g. Adorf & Johnston, 1988; Angel et al., 1990; Odewahn et al., 1992). With the advent of CNNs, RNNs, and deep learning, these MLP models gave way to raw data ingestion (e.g. Dieleman et al., 2015; Charnock et al., 2018; Wu & Peek, 2020; Smith et al., 2021; Khederlarian et al., 2026). And the third wave of unsupervised and self-supervised learning largely removed the need for task-specific human-generated labelling, with connectionist methods inferring astronomical knowledge directly from the raw data (e.g. Sarmiento et al., 2021; Smith et al., 2022). A fourth wave has recently been seeded by the discovery of predictable neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022)—the application of foundation models to astronomical observations, publications, and survey data (Smith & Geach, 2023). This fourth wave has brought with it diverse approaches in the search for a viable path towards a single, canonical, astro-foundation model: from contrastive methods (e.g. Slijepcevic et al., 2024; Parker et al., 2024; Mishra-Sharma et al., 2024; Zhao et al., 2025), to generative architectures (e.g. Leung & Bovy, 2023; Koblischke & Bovy, 2024; Ore et al., 2024), to autoregressive modelling (e.g. Smith et al., 2024; Pan et al., 2024; Euclid Collaboration et al., 2025; Zuo et al., 2025; Heneka et al., 2025; Moriwaki et al., 2025), to finetuning of large language models on astronomical text (e.g. Nguyen et al., 2023; Perkowski et al., 2024; de Haan et al., 2024; Zaman et al., 2025). The sheer diversity of these approaches raises a natural question: does the choice

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of architecture, training regime, and data modality matter, or do sufficiently performant models converge to equivalent representations?

The Platonic Representation Hypothesis (PRH; Huh et al. (2024)) offers one answer: neural networks trained on different modalities are converging toward a shared global statistical model of reality in their representation spaces. The PRH attributes this convergence to three mechanisms—*task generality* (models trained on more diverse tasks require representations that capture more information about underlying reality), *model capacity* (larger models are more likely to find optimal representations), and *simplicity bias* (neural networks naturally favour simpler solutions that generalise better)—and draws inspiration from Plato's Allegory of the Cave (Plato, c. 375 BCE), in which the cave-dwellers mistake shadows on a wall for reality itself. In this analogy, our training data are the shadowy projections of an underlying reality, and our models learn to recover representations (or 'Forms') of the reality that generates them. Under the PRH, larger models trained on more diverse tasks should converge toward a Platonic Ideal: a perfect, lossless Form of underlying reality. A recent refinement, the Aristotelian Representation Hypothesis (ARH; Gröger et al. (2026)), tempers this strong claim. Rather than agreeing on a shared global coordinate system, the ARH proposes that converging models agree on local neighborhood structure rather than global embedding geometry. This distinction has direct empirical consequences in that *local* rank statistics are predicted to track convergence, while *global* similarity measures are not.

We use astronomy as our experimental instrument. Astronomical observations provide a natural testbed as different projections of the same underlying cosmic reality. The historical success of astrophysics is itself evidence that a compact, modality-invariant description of galaxy observables exists, in that distinct instruments observing the same galaxies can be reconciled into a single physical account. These observations are linked through shared physical processes; a galaxy's morphology, kinematics, chemical composition, and star formation history all emerge from the same stellar populations, gas and dust dynamics, and underlying matter distributions (Conroy, 2013). This provides an external physical reference against which we can measure whether foundation models converge towards representations of the underlying physics. Furthermore, the scale and diversity of modern surveys provide the data volume necessary to test convergence across multiple model architectures and training objectives, and recent work (The Multimodal Universe Collaboration, 2024; Caplar et al., 2025) has standardized crossmatching across astronomical modes[1].

Our contributions follow. **We establish the use of astronomy as a instrument for the study of neural representations** by testing both the PRH and the ARH across a basket of eleven foundation model families spanning supervised classification, self-distillation, joint-embedding prediction, masked autoencoding, vision–language pre-training, and astronomy-specific architectures from $\mathcal{O}$(10M)→$\mathcal{O}$(10B) parameters, evaluated on crossmatched JWST, HSC, and Legacy Survey imaging, and DESI spectroscopy. We organise our analysis around answering four 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 find that probe performance scales with model capacity (§3.1 ¶1). **(2) Do different architectures converge to the same physical representation?** We find that physics performance is correlated between modalities, and the relative orientations of probe directions agree across models and with expected astrophysical relationships (§3.1 ¶2-3). **(3) Does geometric alignment scale with performance—locally, globally, or both?** We find that MKNN (local alignment) but not CKA (global alignment) tracks probe performance across model families, between astronomical modalities, and across architectures (§3.2). **(4) What can we learn from our anomalous results?** We find that DINOv3 and PaliGemma do not follow the trend established by other models: larger DINOv3 students align more closely with their fixed teacher but less with an average model, while scaling PaliGemma's language trunk around a fixed vision encoder does not improve physics performance (§3.2; App. C), suggesting that total model capacity alone is insufficient to predict representational convergence.

2 Datasets, models, and metrics

Datasets used. We test across four crossmatched astronomical datasets that capture different projections of galaxy Forms. We use the MMU to crossmatch between data modes at a 1" radius (The

[1]We define an astronomical 'mode' or 'modality' as the information captured by a specific instrument so that (e.g.) JWST and HSC imaging are separate modes.

Multimodal Universe Collaboration, 2024). **Hyper Suprime-Cam (HSC)** (Miyazaki et al., 2018) provides our reference baseline. HSC is an 8.2m ground-based optical survey with ~0.7" seeing and 5-band (*grizy*) coverage. We use HSC as it provides our largest overlap sample across all other modalities, making it a natural anchor for pairwise comparison. **DESI Legacy Imaging Survey** (Dey et al., 2019) is a complementary ground-based optical survey at ~1" seeing, its inclusion allows us to test representational alignment across different ground-based imaging survey strategies. Our crossmatched Legacy-HSC sample contains 102k galaxies. **JWST NIRCam imaging** (Gardner et al., 2023) provides our most extreme imaging comparison: space-based near-infrared and infrared observations that can reveal dust emission and dust-obscured stellar populations invisible to our optical surveys. The JWST-HSC crossmatched sample contains 1.67k galaxies. **DESI spectra** (DESI Collaboration et al., 2024) provide our crossmodal non-imaging test. Spectroscopy probes integrated galaxy properties (e.g. stellar properties, kinematics) that are complementary to the spatial information captured by imaging. The DESI-HSC crossmatched sample contains 18.6k galaxies.

To test the convergence of representations with astronomical parameters, we use HSC and JWST images from **COSMOS-Web** (Casey et al., 2023). The COSMOS field is one of the most observed areas of the night sky; due to imaging from many different telescopes across wide wavelength ranges, galaxies in COSMOS have very accurate measurements of their physical parameters (Shuntov et al., 2025). To ensure that our physical labels are high quality, we select 45 000 galaxies with both HSC i and JWST F150W band flux signal-to-noise ratios greater than 25, and use their redshifts (many-band photometric redshifts), stellar masses, and specific star formation rates (sSFRs) obtained from the template-fitting method LePHARE Arnouts et al. (1999); Ilbert et al. (2006). These parameters are canonical physical descriptors of galaxies; they capture a galaxy's intrinsic luminosity, current evolutionary state, and distance from Earth. To obtain image cutouts, we use the DAWN JWST Archive (DJA)[2] and the image cutout tool from the HSC release website[3].

For our MMU-crossmatched surveys we assemble three-channel RGB composites from the raw flux cubes: g, r, z for HSC and Legacy Survey, and F090W, F277W, F444W for JWST, ensuring maximum wavelength coverage while remaining suitable for our foundation models trained on RGB natural images. Pixel intensities are normalised per band with an arcsinh stretch anchored to precomputed 1st and 99th percentile flux values (estimated per survey over batches of up to 10 000 images): $\bar{x} = \text{arcsinh}(\alpha(x - p_1)/(p_{99} - p_1))/\text{arcsinh}(\alpha)$, with stretch parameter $\alpha = 20$ and the output clipped to $[0, 1]$. Arcsinh stretching preserves faint low-surface-brightness structure while compressing the dynamic range of bright galaxy centres, we choose $\alpha = 20$ following Lupton et al. (2004). In the COSMOS-Web dataset, we use g, r, z for HSC and $F115W$, $F150W$, and $F277W$ for JWST. We crop HSC and Legacy Survey galaxies and bilinearly rescale the image cutouts to match the $3.84'' \times 3.84''$ angular size of the JWST cutouts (96×96 pixels at $0.04''/\text{pixel}$). Finally, before passing data into a model, any model-specific preprocessing is applied according to each model's published processing pipeline.

Model architectures. Our model selection (Tab. 1) is designed to test the PRH along three axes. First, we include models spanning supervised classification (ViT, ConvNeXtv2), self-supervised learning via knowledge distillation (DINOv3), joint-embedding prediction (IJEPA, VJEPA), and masked autoencoding (ViT-MAE), allowing us to test whether convergence depends on training objective. Second, we include vision-language models (CLIP, LLaVA-1.5, PaliGemma) to test whether architectures trained to bridge vision and language converge toward the same representations as unimodal vision models. Third, we include two astronomy-specific models: AstroPTv2, an autoregressive transformer pre-trained exclusively on DESI Legacy Survey imaging, and

Table 1: Foundation models used in this study, spanning supervised, self-supervised, autoregressive, and multimodal training paradigms across vision, language, and spectral modalities.

Model	Training regime	Modality
Supervised		
ViT (Dosovitskiy et al., 2020)	Supervised classification	Images
ConvNeXtv2 (Woo et al., 2023)	Supervised classification	Images
Self-supervised		
DINOv3 (Siméoni et al., 2025)	Self-distillation	Images
IJEPA (Assran et al., 2023)	Joint-embedding prediction	Images
VJEPA (Assran et al., 2025)	Joint-embedding prediction	Video
ViT-MAE (He et al., 2022)	Masked autoencoding	Images
Multimodal / Vision-Language		
LLaVA-1.5 (Liu et al., 2023)	Vision-language	Images + Text
PaliGemma (Beyer et al., 2024)	Vision-language	Images + Text
CLIP (Radford et al., 2021)	Contrastive image-text	Images + Text
Astronomy-specific		
AstroPTv2 (Smith et al., 2024)	Next-token prediction	Astro. images
Specformer (Parker et al., 2024)	Contrastive	Astro. spectra

[2] https://dawn-cph.github.io

[3] https://hsc-release.mtk.nao.ac.jp/doc/index.php/data-access__pdr3/

Specformer, a transformer trained on one-dimensional DESI spectra. AstroPTv2 tests whether domain-specific pre-training yields representations that diverge from or converge with those of general-purpose models. Specformer represents our most out-of-distribution test case: a fundamentally different input modality (spectra vs. images) processed by a model that has never seen natural images. We extract our embeddings from the literature-recommended embedding layer of each model, or from the penultimate layer where there is no recommendation. For each architecture family, we test multiple model sizes to measure whether representational alignment scales with capacity or performance.

Metrics used to measure representational alignment. Following the methodology established in the original PRH work, we measure representational alignment using the mutual k-nearest neighbour (MKNN) metric (Chechik et al., 2010). Given two embeddings $(\mathbf{z}_1, \mathbf{z}_2)$ corresponding to the same object as viewed by two different instruments or models, the MKNN score is computed as the cardinality of intersections for each object's k-nearest neighbours in the embedding space: $\text{MKNN}(\mathbf{z}_1, \mathbf{z}_2) = k^{-1}|N_k(\mathbf{z}_1) \cap N_k(\mathbf{z}_2)|$, averaged over all objects in the dataset, where N_k is the k-nearest neighbours operation under the cosine distance, and $|\cdot|$ denotes set cardinality. We follow Huh et al. (2024) and set $k = 10$ across all experiments. We complement MKNN with the Centered Kernel Alignment (CKA) metric (Kornblith et al., 2019): $\text{CKA} = (\|AB^T\|_F^2)/(\|AA^T\|_F\|BB^T\|_F)$, where A and B are mean centered embedding matrices and $\|\cdot\|_F$ denotes the Frobenius norm. We use both MKNN and CKA as MKNN captures local neighborhood agreement, whereas CKA captures global embedding similarity. To eliminate confounds from network scale we follow Gröger et al. (2026) and use permutation-calibrated versions of the MKNN and CKA metrics in all of our experiments, which subtract an empirical null obtained by shuffling sample correspondences to correct for width-driven baseline inflation.

3 Experiments and Results

We ask first whether foundation models encode galaxy physics (§3.1), and then whether the geometry of their embeddings converges in ways that track this physical content (§3.2).

3.1 Physical interpretation experiments

Larger models encode more physics knowledge. We focus on three physical properties: stellar mass ($M_\star$), specific star formation rate (sSFR), and redshift (z). A model that has internalised the physics of galaxy images should encode all three canonical descriptors. After clipping values to the 1st and 99th percentiles to remove outliers, we train linear probes on the embeddings of each model to predict these properties. For both HSC and JWST, we run 10-fold cross-validation using the 45 000 selected galaxies with matched physical parameter measurements from the COSMOS-Web catalog. Fig. 1 shows the average R^2 score across properties versus the parameter count with full per-property results reported in App. B. A Spearman's rank test gives a significant positive correlation for our JWST ($\rho = 0.508, p = 0.004$) and HSC ($\rho = 0.589, p = 0.0005$) cases, indicating that larger models recover more of the underlying astrophysics from images alone, despite AstroPT being our only tested model pre-trained significantly on galaxies (see App. A.3). This relationship persists after controlling for architecture family in an ANCOVA, $R^2 \sim \log N_{\text{params}} + C(\text{family})$, with a positive capacity coefficient for both HSC ($p = 0.002$) and JWST ($p = 2e{-}5$). The scaling trend therefore cannot be explained by familial differences.

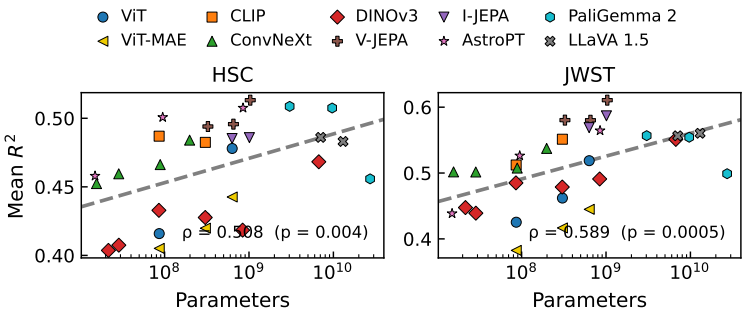


Figure 1: Linear probe R^2 as a function of model size, averaged over mass, sSFR, and redshift. We find a significant positive correlation, even though AstroPT is the only model in our basket pretrained substantially on galaxy imagery.

Physics performance is correlated between modalities. In Fig. 2, we show JWST R^2 scores as a function of HSC R^2 scores in a grid of $M_\star$, sSFR, and z. We find a strong cross-modal correlation (p<<0.01 via a Spearman's rank test) in all cases, indicating that a model's capacity to linearly encode physical information is a consistent property of the model itself rather than that of the observed wavelength range. A principal component analysis across the six probe R^2 scores confirms this with 87% of the variance captured in a single axis with uniformly positive loadings—a 'model quality' axis along which the same models that recover more astrophysics from optical imagery also recover more from near-infrared imagery, despite the two probing largely different physical processes (§2 ¶1). One would expect this if our models are converging on representations of the underlying galaxies rather than on modality-specific image statistics.

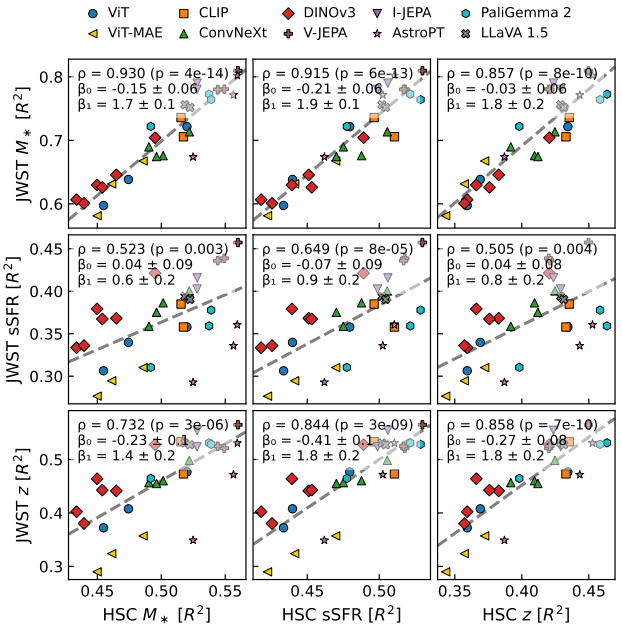


Figure 2: Linear probe R^2 scores of JWST versus HSC modalities on a grid of physical parameters. Model performance is strongly correlated between the instruments, indicating that a model's capacity to linearly encode physical information is a property of its representations rather than a function of the observed wavelength range or modality.

JWST samples near-infrared stellar light, where emission is dominated by relatively low mass (long living) stars. At near-IR wavelengths, the overall spectrum is also less affected by dust attenuation. JWST imagery is therefore a clean tracer of already-built stellar mass but carries little information about ongoing star formation. HSC instead samples UV/optical light, and so is expected to be the more reliable tracer of recent and ongoing star formation (Conroy, 2013). Finally, JWST's longer wavelength baseline pins down redshift more tightly than HSC's (Salvato et al., 2019), so we expect superunity linear regression slopes for $M_\star$ and z on Fig. 2's diagonal and a subunity slope for sSFR—which is what we recover. By definition, sSFR ≡ SFR/$M_\star$, so HSC sSFR carries information about stellar mass as well as star formation. If both modalities encode mass in a shared, modality-invariant way, then the $M_\star$ component of HSC sSFR should align with JWST's mass representation. Because JWST carries little SFR information but traces $M_\star$ cleanly, we therefore expect HSC sSFR to correlate more strongly with JWST $M_\star$ than with JWST sSFR—and indeed we find ρ(HSC sSFR, JWST $M_\star$) = 0.915 vs ρ(HSC sSFR, JWST sSFR) = 0.649. Taken holistically, our crossmodal correlations are consistent with our models having internalised modality invariant astrophysical content rather than instrument-specific image features, and foreshadows our further crossmodal analysis in §3.2.

Relative orientations of probe directions are consistent across models and with known physics. We assess the relative orientations of the redshift, stellar mass, and sSFR directions in embedding space by measuring their pairwise cosine similarities. Fig. 3 shows the cosine similarity matrices for HSC (left) and JWST (right), averaged across all models. Within the star-forming population, galaxies follow the main sequence, $\log \mathrm{SFR} = \alpha \log M_\star + \beta$, with a sub-linear slope $\alpha \lesssim 1$ (Speagle et al., 2014). As sSFR ≡ SFR/$M_\star$ it follows that $\log \mathrm{sSFR} \propto (\alpha - 1) \log M_\star$, and so we would expect the sSFR-$M_\star$ correlation to be mildly negative. This trend is reinforced by the rising quiescent (non-star forming) galaxy fraction with stellar mass, which shifts more massive galaxies to lower sSFR, and both effects are weak relative to the several orders of magnitude that sSFR spans at fixed mass (Wetzel et al., 2012). All of which is consistent with the small negative cosine recovered across all our tested models in Fig. 3. The correlation between $M_\star$ and redshift arises primarily from a Malmquist-style selection effect (Malmquist, 1922): since higher-redshift galaxies appear fainter, only intrinsically luminous—and therefore more massive—galaxies exceed the detection threshold

of telescopes, biasing high-redshift samples toward larger stellar masses. Finally, higher redshift galaxies are observed to have higher star-forming rates (Madau & Dickinson, 2014), which explains the sSFR and redshift correlation. In short, the correlations between these three parameters result from intrinsic galaxy physics and selection effects, yet they are captured by all models despite all-but-one of our models having been pre-trained on natural imagery or text (App. A.3).

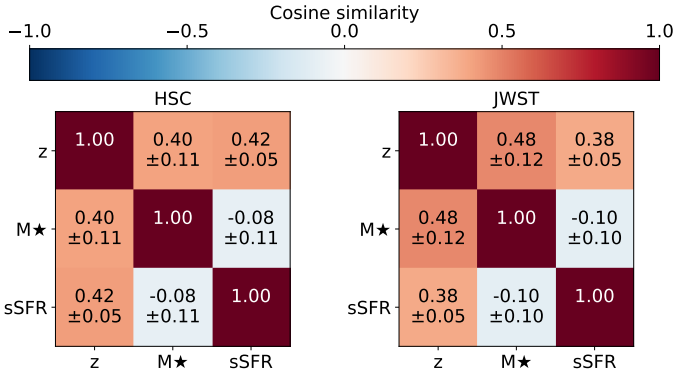


Figure 3: Cosine similarity matrices between the redshift (z), stellar mass ($M_\star$), and sSFR probe weight vectors for HSC (left) and JWST (right), averaged over all models. The relationships between the three probe weight vectors are consistent across all models, suggesting that the embedding spaces encode physics via similar internal organisation.

To summarize, we find that model capacity translates into physical knowledge, despite only one model in our basket being substantially pre-trained on galaxy imagery. We also see that probe performance is tightly correlated between HSC and JWST across all three physical parameters, indicating that a model's ability to linearly encode galaxy physics transfers across observed wavelength ranges and observables. And the relative orientations of the physical probe directions agree across models, and agreement reflects genuine astrophysical correlations, which suggests the formation of a shared physics subspace within the models' embeddings.

3.2 Global and local embedding alignment experiments

We test whether embedding alignment scales with model capacity and physics performance within model families, across modalities, and across architectures, using permutation-calibrated MKNN and CKA (§2 ¶3; Gröger et al., 2026).

DINOv3 and PaliGemma reveal limits to convergence under scale. We first ask whether MKNN and CKA scores increase in the direction the PRH predicts. We construct ordered pairs in two ways. The first method takes consecutively sized pairs of pre-trained models, ordered by model size within a family (Tab. 5); the PRH predicts that embedding similarity of pairs should increase with scale, since better-performing models sit closer to the Platonic Ideal. The second method pairs a single model's embeddings of two distinct astronomical modalities (e.g. JWST and HSC imagery, Tab. 6); here we expect better-performing models to produce more aligned cross-modal embeddings. We treat each ordered pair as an independent Bernoulli trial with success if the similarity score increases along the predicted direction (larger model pair in the intra-architectural case, the larger model in the crossmodal case), and failure otherwise. Under the null that scale carries no information about alignment, the success rate is 50%, and we test the basket rate against this null via a binomial test. Across our full basket, including DINOv3, the MKNN trend is in the predicted direction (60/96 = 62%, $p = 0.02$), confirming the PRH's basic prediction. However, this effect is concentrated outside DINOv3, and excluding it sharpens the signal (51/69 = 74%, $p = 9e{-}5$) while DINOv3 alone trends in the opposite direction (9/27 = 33%). A Fisher exact test confirms that DINOv3 sits in a statistically distinct scaling regime from the rest of the basket ($p = 4e{-}4$ for MKNN, $p = 0.03$ for CKA). CKA shows no significant trend in either case.

We investigate DINOv3's behavior by comparing its distilled students with their fixed ViT-7B/16 teacher and with the non-DINOv3 model basket. On paired HSC and JWST imaging we find that teacher alignment as measured by MKNN generally increases with student size, while mean basket alignment decreases despite maintained or improving mean probe R^2 (App. C.1). This differs from the overall positive relationship between physical performance and basket alignment observed in our other models (Fig. 6). PaliGemma provides a complementary case, with crossmodal MKNN also trending negatively with scale. Its SigLIP vision encoder remains fixed in size, with additional capacity confined to the language trunk. Layerwise probing shows that physical recovery declines through this trunk, and the largest model recovers less physics than its smaller counterparts

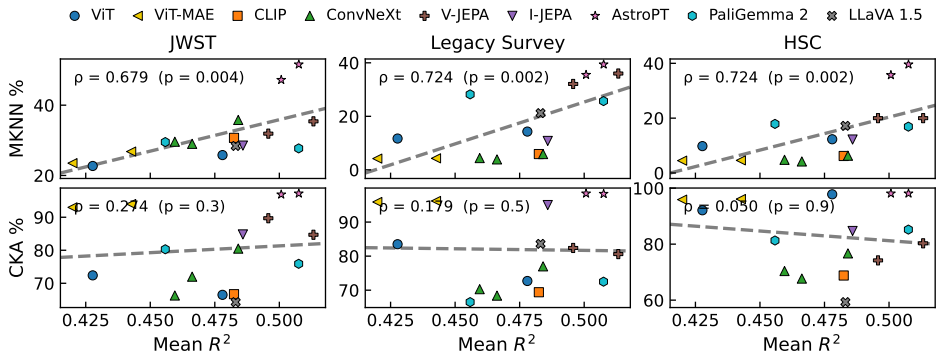


Figure 4: Intra-architectural scaling results. We plot our MKNN and CKA similarity values within each architecture family against the mean R^2 probe performance across the physical tasks described in §3.1. In each pane we state the ρ and p values as measured via a Spearman's rank test.

(App. C.2). Taken together, these results suggest that convergence under scale depends on the training target and where capacity is added when scaling. Given this statistically distinct scaling regime, we exclude DINOv3 from the remaining tests in this section. We retain PaliGemma, as the same Fisher exact test does not reach significance ($p = 0.74$ for MKNN).

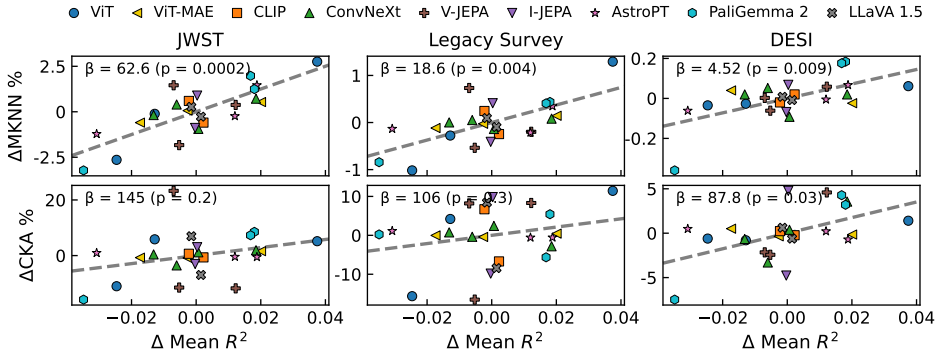


Figure 5: Crossmodal MKNN and CKA metrics vs R^2 (§3.1). R^2 and MKNN/CKA are demeaned per architecture family to show the trend without architecture family confounders. We state β and p from the ANCOVA tests MKNN ~ $R^2 + C(\text{family})$, and CKA ~ $R^2 + C(\text{family})$.

Within-family local embedding alignment correlates with physics performance. We further test whether embedding similarities of models within the same architecture family tracks how well those models recover galaxy physics. Concretely, for each model family we pair adjacent size rungs (e.g. ConvNeXtv2 Nano vs. Tiny, Tiny vs. Base, Base vs. Large), compute their MKNN and CKA scores on each of our three imaging surveys (JWST, Legacy, HSC; §2 ¶1), and compare these pairwise alignment scores against the physical-probe R^2 of the larger of the two models in the pair (§3.1). The PRH predicts that better-performing pairs should also be more internally aligned, since both models in a high-performing pair should sit closer to the Platonic Ideal. Fig. 4 shows this is borne out for MKNN: across all three surveys we find a strong, significant Spearman correlation ($\rho = 0.679$–0.724, $p < 0.005$). CKA shows no significant trend in any survey ($p \geq 0.3$). Per-property breakdowns (mass, sSFR, redshift) and the full table of pairwise scores are reported in App. D.

Local embedding alignment across modalities correlates with physics performance. We now test whether cross-modal embeddings of the same galaxies converge as model performance grows. For each model we compute the MKNN and CKA alignment between the embeddings of HSC imagery and the embeddings of a second modality (JWST imagery, Legacy Survey imagery, or DESI spectra as processed by Specformer), and regress these alignment scores against probe R^2 (§3.1). In this test, if two instruments observing the same galaxies result in embeddings that align,

the alignment cannot be inherited from raw pixel similarity and must instead reflect shared latent content recovered by the model. Fig. 5 summarizes our results. Controlling for familial effects via an ANCOVA, we find a significant correlation ($p < 0.01$) in all tested MKNN cases (JWST: $\beta = 62.6, p = 2e-4$; Legacy: $\beta = 18.6, p = 4e{-}3$; DESI: $\beta = 4.52, p = 9e{-}3$), and no significant trend in the JWST and Legacy Survey CKA tests. We find a borderline correlation in our CKA DESI test case ($p = 0.03$). When additionally controlling for model capacity via $\mathrm{MKNN} \sim R^2 + \log_{10} N_{\mathrm{params}} + C(\mathrm{family})$ we find that the positive association between local alignment and physics performance persists for all three pairings (JWST: $\beta = 69.1$, $p = 4e{-}4$; Legacy: $\beta = 21.2$, $p = 5e{-}3$; DESI: $\beta = 6.79$, $p = 1e{-}4$). This indicates that parameter count does not explain the relationship, leaving the primary driver as probe performance. Notably, we find a significant relation for DESI spectra vs HSC imagery: spectra and images share essentially no low-level statistics—a 1D sequence of fluxes vs a 2D RGB cutout, processed by a spectra model that has never seen an image, and image models that have never seen spectra—so alignment between them must be driven by shared physical content. The finding that better-performing models produce more aligned crossmodal embeddings is direct evidence that convergence is mediated by underlying astrophysics rather than by input correlations or shared inductive biases inherited from natural-image pre-training. Per-property breakdowns are reported in App. E.

Cross-architectural alignment correlates with physics performance. We ask whether our basket of models is converging toward a single 'ideal' embedding space across architectures. For each model in our basket we compute the MKNN and CKA scores against every other model and take the mean across comparisons, yielding a single per-model alignment score that quantifies how close a model sits to the rest of the basket in embedding space. Fig. 6 plots the mean cross-architectural MKNN and CKA against the average linear probe R^2 as described in §3.1. We restrict this analysis to the JWST-HSC crossmatched MMU dataset for computational tractability, since pairwise comparison across the full basket scales quadratically with model count. We find a significant positive Spearman correlation between MKNN alignment and physics performance on both probe-spaces ($\rho = 0.455, p = 0.02$ for JWST, $\rho = 0.432, p = 0.03$ for HSC), indicating that models with better physics performance sit closer to the basket's average embedding geometry. This trend holds across all tested physical properties (App. F). As in our earlier tests, we do not see a significant trend for CKA vs physics performance ($p > 0.1$ in both probe-spaces).

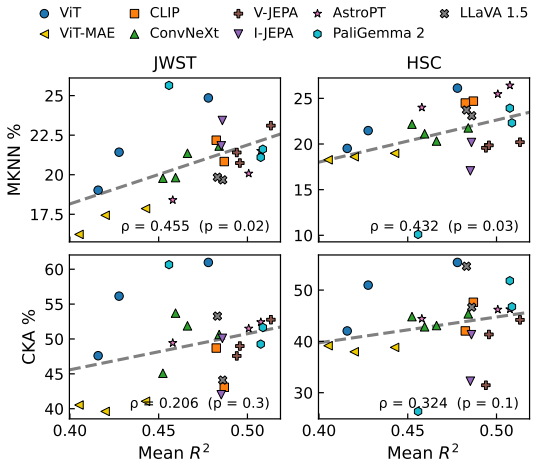


Figure 6: Cross-architecture metric distance (MKNN, CKA), plotted against average R^2 over the three physical properties for HSC and JWST images (§3.1). We find that probe performance is significantly correlated with MKNN, but not with CKA.

Probe-defined physical directions contribute to embedding alignment. To complement our above analysis we remove the joint subspace spanned by our physical probe directions in 21 ViT–ConvNeXtv2 pairs evaluated on HSC imagery (App. F.1). We find that MKNN and CKA decrease in all cases, while not decreasing when removing a random subspace. This supports the conclusion that the results reported here are caused by a learnt physical subspace in the models' embeddings.

Across all tests in this subsection we find that local embedding similarity (MKNN) tracks physics performance, while global similarity (CKA) does not. The intra-architectural test shows that within a family, better-performing size-pairs are more locally aligned. The crossmodal test extends this to pairs of instruments observing the same galaxies, and crucially holds for DESI spectra vs. HSC imagery—a comparison which isolates shared physical content as the driver of alignment. The cross-architectural test shows that models that encode more physics sit closer to the basket's mean embedding geometry, suggesting convergence toward a shared representation rather than family-specific solutions. The CKA null result across all four tests indicates that convergence is a property of local neighbourhood structure rather than global embedding geometry, consistent with the Aristotelian rather than the strict Platonic reading of the convergence hypothesis (Gröger et al., 2026).

4 CONCLUSIONS: THINK LOCAL, NOT GLOBAL

We find strong support for the Aristotelian Representation Hypothesis. Across our tests in §3.2 we find consistently that local neighborhood structure (MKNN) tracks model performance, while global embedding similarity (CKA) does not. We must remind ourselves that, aside from AstroPT and Specformer, our tested models are not significantly pre-trained on astronomical data (App. A.3; that these models identify any correspondence between fundamentally different astronomical observations is remarkable, and suggests that sufficiently scaled up neural networks learn universal structural patterns transcending their training domains. Our local-not-global embedding convergence pattern supports the Aristotelian reading of representational convergence (Gröger et al., 2026): foundation models trained on different objectives, modalities, and data appear to agree on which galaxies are like other galaxies, without agreeing on a shared global coordinate system in which to place them. We can also see that our natural image-trained models achieve embedding alignment that increases with model performance with Specformer's DESI spectral embeddings, therefore showing correspondence between fundamentally different modalities and data types they have never encountered that do not share low-level statistics.

AstroPT does not dominate. AstroPTv2 is the only model in our basket pre-trained substantially on galaxy imagery, and yet it sits comfortably within the trends defined by general-purpose models rather than above them, in physics probe performance and in representational similarity. We read this as evidence that sufficiently scaled general-purpose models already recover representations close to those delivered by domain-specific pre-training—precisely as Sutton's Bitter Lesson would predict (Sutton, 2019). This follows directly from the representational convergence documented throughout our remaining results. Domain-specific architectures may still be useful where the input modality is genuinely outside the natural-image distribution (Specformer on 1D spectra is a clear example), but for galaxy imaging, scale and data diversity appear to substitute for domain specificity.

DINOv3's and PaliGemma's anomalous behaviors are informative. Unlike our other tested models, DINOv3's distilled student models' embedding spaces converge towards their teacher model as their capacity increases rather than the remainder of the model basket. PaliGemma's larger language trunks do not improve physical information recovery due to its fixed-capacity vision tower. These results show that total parameter count is not sufficient to predict representational convergence while also providing guidance for pre-training and architectural choices for foundation models.

Limitations. While our results provide compelling evidence for the ARH, we note several limitations that suggest avenues for future exploration. Our cross-modal samples are uneven in size (1.67k galaxies for JWST–HSC, 102k for Legacy–HSC, 18.6k for DESI–HSC). We note however that our dataset sizes are in all cases larger than the datasets of size $n = 1024$ used in the establishing works of Huh et al. (2024, Apps. C1, C2) and Gröger et al. (2026, §6.3). We probe only three physical properties; convergence on morphology, kinematics, or environment may behave differently. Our cross-architectural test is restricted to our 1.67k JWST/HSC dataset for tractability. The DINOv3 checkpoints confound distillation with other training differences, and isolating this effect requires a computationally-expensive controlled pre-training study. We leave these questions for future studies.

Implications for AI. This work demonstrates the usefulness of astronomy as an experimental instrument for understanding machine learning models. Open, crossmatched observations provide heterogeneous views of the same systems, while independently derived physical parameters provide an external reference. We release the code, embeddings, and data so that the Platonic Universe can be reused to study neural representations beyond the hypotheses tested here.

Implications for astronomical foundation modelling. We observe general improvement in representational alignment at larger model scales, suggesting that each architecture is converging towards a shared representation. Taken to its conclusion, this convergence implies that future efforts in astronomical foundation modelling should focus less on astronomy-specific architectures and more on scale and data diversity. It also follows that the astronomy community should embrace pre-trained foundation models rather than training from scratch: if all architectures converge toward the same representations, then starting from models pre-trained on natural images or text—with their billions of parameters and massive computational investment already spent—offers both superior performance and dramatic reductions in environmental impact. The broader open source machine learning community has already invested the GPU-centuries needed to learn general-purpose representations, we need now only gently guide these models toward astronomical use-cases.

AI USE STATEMENT

We used generative AI tools to aid data processing and coding efforts, checking any code changes via a manual PR review before merging into our repository. In addition to this, we used generative AI tools to aid in editing and reviewing this manuscript after an initial human-written draft was in place. We take responsibility for the final content of this manuscript, including text, claims, and any artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

During the review process we make our code available anonymously at `https://anonymous.4open.science/r/platonic-universe-7330`, and our embeddings and data available at `https://huggingface.co/HCVYM5w6Gn`. Upon deanonymization we will link to all code and data publicly.

REFERENCES

H. M. Adorf and M. D. Johnston. Artificial neural nets in astronomy. In *Arbeitspapier der Gesellschaft für Mathematik and Datenverarbeitung*, volume 329 of *Arbeitspapier der Gesellschaft für Mathematik and Datenverarbeitung*, 1988.

J. R.P. Angel, P. Wizinowich, M. Lloyd-Hart, and D. Sandler. Adaptive optics for array telescopes using neural-network techniques. *Nature*, 348(6298):221–224, 1990. ISSN 0028-0836. doi: 10.1038/348221a0.

S. Arnouts, S. Cristiani, L. Moscardini, S. Matarrese, F. Lucchin, A. Fontana, and E. Giallongo. Measuring and modelling the redshift evolution of clustering: the hubble deep field north. *Monthly Notices of the Royal Astronomical Society*, 310:540–556, 1999. doi: 10.1046/j.1365-8711.1999.02978.x.

M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *ArXiv e-prints*, 2023. doi: 10.48550/arXiv.2301.08243.

M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, Mojtaba, Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2506.09985.

L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai. PaliGemma: A versatile 3B VLM for transfer. *ArXiv e-prints*, 2024. doi: 10.48550/arXiv.2407.07726.

N. Caplar, W. Beebe, D. Branton, S. Campos, A. Connolly, M. DeLucchi, D. Jones, M. Juric, J. Kubica, K. Malanchev, R. Mandelbaum, and S. McGuire. Using LSDB to enable large-scale catalog distribution, cross-matching, and analytics. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2501.02103.

C. M. Casey, Jeyhan S. Kartaltepe, Nicole E. Drakos, Maximilien Franco, Santosh Harish, Louise Paquereau, Olivier Ilbert, Caitlin Rose, Isabella G. Cox, James W. Nightingale, Brant E. Robertson, John D. Silverman, Anton M. Koekemoer, Richard Massey, Henry Joy McCracken, Jason Rhodes, Hollis B. Akins, Natalie Allen, Aristeidis Amvrosiadis, Rafael C. Arango-Toro, Micaela B. Bagley, Angela Bongiorno, Peter L. Capak, Jaclyn B. Champagne, Nima Chartab, Óscar A. Chávez Ortiz, Katherine Chworowsky, Kevin C. Cooke, Olivia R. Cooper, Behnam

Darvish, Xuheng Ding, Andreas L. Faisst, Steven L. Finkelstein, Seiji Fujimoto, Fabrizio Gentile, Steven Gillman, Katriona M. L. Gould, Ghassem Gozaliasl, Christopher C. Hayward, Qiuhan He, Shoubaneh Hemmati, Michaela Hirschmann, Knud Jahnke, Shuowen Jin, Ali Ahmad Khostovan, Vasily Kokorev, Erini Lambrides, Clotilde Laigle, Rebecca L. Larson, Gene C. K. Leung, Daizhong Liu, Tobias Liaudat, Arianna S. Long, Georgios Magdis, Guillaume Mahler, Vincenzo Mainieri, Sinclaire M. Manning, Claudia Maraston, Crystal L. Martin, Jacqueline E. McCleary, Jed McKinney, Conor J. R. McPartland, Bahram Mobasher, Rohan Pattnaik, Alvio Renzini, R. Michael Rich, David B. Sanders, Zahra Sattari, Diana Scognamiglio, Nick Scoville, Kartik Sheth, Marko Shuntov, Martin Sparre, Tomoko L. Suzuki, Margherita Talia, Sune Toft, Benny Trakhtenbrot, C. Megan Urry, Francesco Valentino, Brittany N. Vanderhoof, Eleni Vardoulaki, John R. Weaver, Katherine E. Whitaker, Stephen M. Wilkins, Lilan Yang, and Jorge A. Zavala. COSMOS-Web: An Overview of the JWST Cosmic Origins Survey. ApJ, 954(1):31, September 2023. doi: 10.3847/1538-4357/acc2bc.

T. Charnock, G. Lavaux, and B. D. Wandelt. Automatic physical inference with information maximizing neural networks. *Physical Review D*, 97(8):083004, 2018. ISSN 2470-0029. doi: 10.1103/PhysRevD.97.083004.

G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(36):1109–1135, 2010. URL `http://jmlr.org/papers/v11/chechik10a.html`.

C. Conroy. Modeling the Panchromatic Spectral Energy Distributions of Galaxies. *Annual Review of Astronomy and Astrophysics*, 51:393–455, 2013. doi: 10.1146/annurev-astro-082812-141017.

T. de Haan, Y.-S. Ting, T. Ghosal, T. D. Nguyen, A. Accomazzi, A. Wells, N. Ramachandra, R. Pan, and Z. Sun. AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model. *ArXiv e-prints*, 2024. doi: 10.1038/s41598-025-97131-y.

DESI Collaboration, A. G. Adame, J. Aguilar, S. Ahlen, S. Alam, G. Aldering, D. M. Alexander, R. Alfarsy, C. A. Prieto, M. Alvarez, O. Alves, A. Anand, F. Andrade-Oliveira, E. Armengaud, J. Asorey, S. Avila, A. Aviles, S. Bailey, A. Balaguera-Antolínez, O. Ballester, C. Baltay, A. Bault, J. Bautista, J. Behera, S. F. Beltran, S. BenZvi, L. B. e. Silva, J. R. Bermejo-Climent, A. Berti, R. Besuner, F. Beutler, D. Bianchi, C. Blake, R. Blum, A. S. Bolton, S. Brieden, A. Brodzeller, D. Brooks, Z. Brown, E. Buckley-Geer, E. Burtin, L. Cabayol-Garcia, Z. Cai, R. Canning, L. Cardiel-Sas, A. C. Rosell, F. J. Castander, J. L. Cervantes-Cota, S. Chabanier, E. Chaussidon, J. Chaves-Montero, S. Chen, X. Chen, C. Chuang, T. Claybaugh, S. Cole, A. P. Cooper, A. Cuceu, T. M. Davis, K. Dawson, R. de Belsunce, R. de la Cruz, A. de la Macorra, J. D. Costa, A. de Mattia, R. Demina, U. Demirbozan, J. DeRose, A. Dey, B. Dey, G. Dhungana, J. Ding, Z. Ding, P. Doel, R. Doshi, K. Douglass, A. Edge, S. Eftekharzadeh, D. J. Eisenstein, A. Elliott, J. Ereza, S. Escoffier, P. Fagrelius, X. Fan, K. Fanning, V. A. Fawcett, S. Ferraro, B. Flaugher, A. Font-Ribera, J. E. Forero-Romero, D. Forero-Sánchez, C. S. Frenk, B. T. Gänsicke, L. Á. García, J. García-Bellido, C. Garcia-Quintero, L. H. Garrison, H. Gil-Marín, J. Golden-Marx, S. G. A. Gontcho, A. X. Gonzalez-Morales, V. Gonzalez-Perez, C. Gordon, O. Graur, D. Green, D. Gruen, J. Guy, B. Hadzhiyska, C. Hahn, J. J. Han, M. M. S. Hanif, H. K. Herrera-Alcantar, K. Honscheid, J. Hou, C. Howlett, D. Huterer, V. Iršič, M. Ishak, A. Jacques, A. Jana, L. Jiang, J. Jimenez, Y. P. Jing, S. Joudaki, R. Joyce, E. Jullo, S. Juneau, N. G. Karaçaylı, T. Karim, R. Kehoe, S. Kent, A. Khederlarian, S. Kim, D. Kirkby, T. Kisner, F. Kitaura, N. Kizhuprakkat, J. Kneib, S. E. Koposov, A. Kovács, A. Kremin, A. Krolewski, B. L'Huillier, O. Lahav, A. Lambert, C. Lamman, T.-W. Lan, M. Landriau, D. Lang, J. U. Lange, J. Lasker, A. Leauthaud, L. Le Guillou, M. E. Levi, T. S. Li, E. Linder, A. Lyons, C. Magneville, M. Manera, C. J. Manser, D. Margala, P. Martini, P. McDonald, G. E. Medina, L. Medina-Varela, A. Meisner, J. Mena-Fernández, J. Meneses-Rizo, M. Mezcua, R. Miquel, P. Montero-Camacho, J. Moon, S. Moore, J. Moustakas, E. Mueller, J. Mundet, A. Muñoz-Gutiérrez, A. D. Myers, S. Nadathur, L. Napolitano, R. Neveux, J. A. Newman, J. Nie, R. Nikutta, G. Niz, P. Norberg, H. E. Noriega, E. Paillas, N. Palanque-Delabrouille, A. Palmese, Z. Pan, D. Parkinson, S. Penmetsa, W. J. Percival, A. Pérez-Fernández, I. Pérez-Ràfols, M. Pieri, C. Poppett, A. Porredon, S. Pothier, F. Prada, R. Pucha, A. Raichoor, C. Ramírez-Pérez, S. Ramirez-Solano, M. Rashkovetskyi, C. Ravoux, A. Rocher, C. Rockosi, A. J. Ross, G. Rossi, R. Ruggeri, V. Ruhlmann-Kleider, C. G. Sabiu, K. Said, A. Saintonge,

L. Samushia, E. Sanchez, C. Saulder, E. Schaan, E. F. Schlafly, D. Schlegel, D. Scholte, M. Schubnell, H. Seo, A. Shafieloo, R. Sharples, W. Sheu, J. Silber, F. Sinigaglia, M. Siudek, Z. Slepian, A. Smith, M. T. Soumagnac, D. Sprayberry, L. Stephey, J. Suárez-Pérez, Z. Sun, T. Tan, G. Tarlé, R. Tojeiro, L. A. Ureña-López, R. Vaisakh, D. Valcin, F. Valdes, M. Valluri, M. Vargas-Magaña, A. Variu, L. Verde, M. Walther, B. Wang, M. S. Wang, B. A. Weaver, N. Weaverdyck, R. H. Wechsler, M. White, Y. Xie, J. Yang, C. Yèche, J. Yu, S. Yuan, H. Zhang, Z. Zhang, C. Zhao, Z. Zheng, R. Zhou, Z. Zhou, H. Zou, S. Zou, and Y. Zu. The Early Data Release of the Dark Energy Spectroscopic Instrument. *Astronomical Journal*, 168(2):58, 2024. ISSN 1538-3881. doi: 10.3847/1538-3881/ad3217.

A. Dey, D. J. Schlegel, D. Lang, R. Blum, K. Burleigh, X. Fan, J. R. Findlay, D. Finkbeiner, D. Herrera, S. Juneau, et al. Overview of the DESI Legacy Imaging Surveys. *Astronomical Journal*, 157 (5):168, 2019. ISSN 1538-3881. doi: 10.3847/1538-3881/ab089d.

S. Dieleman, K. W. Willett, and J. Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015. doi: 10.1093/mnras/stv632.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv e-prints*, 2020. doi: 10.48550/arXiv.2010.11929.

Euclid Collaboration, M. Siudek, M. Huertas-Company, M. Smith, G. Martinez-Solaeche, F. Lanusse, S. Ho, E. Angeloudi, P. A. C. Cunha, H. D. Sánchez, M. Dunn, Y. Fu, P. Iglesias-Navarro, J. Junais, J. H. Knapen, B. Laloux, M. Mezcua, W. Roster, G. Stevens, J. Vega-Ferrero, N. Aghanim, B. Altieri, A. Amara, S. Andreon, N. Auricchio, H. Aussel, C. Baccigalupi, M. Baldi, S. Bardelli, P. Battaglia, A. Biviano, A. Bonchi, E. Branchini, M. Brescia, J. Brinchmann, S. Camera, G. Cañas-Herrera, V. Capobianco, C. Carbone, J. Carretero, S. Casas, F. J. Castander, M. Castellano, G. Castignani, S. Cavuoti, K. C. Chambers, A. Cimatti, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, F. Courbin, H. M. Courtois, M. Cropper, A. Da Silva, H. Degaudenzi, G. De Lucia, A. M. Di Giorgio, J. Dinis, C. Dolding, H. Dole, F. Dubath, C. A. J. Duncan, X. Dupac, S. Dusini, S. Escoffier, M. Farina, R. Farinelli, F. Faustini, S. Ferriol, F. Finelli, S. Fotopoulou, M. Frailis, E. Franceschi, S. Galeotta, K. George, B. Gillis, C. Giocoli, J. Gracia-Carpio, B. R. Granett, A. Grazian, F. Grupp, S. Gwyn, S. V. H. Haugan, W. Holmes, I. M. Hook, F. Hormuth, A. Hornstrup, K. Jahnke, M. Jhabvala, E. Keihänen, S. Kermiche, A. Kiessling, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, Q. L. Boulc'h, A. M. C. L. Brun, D. L. Mignant, S. Ligori, P. B. Lilje, V. Lindholm, I. Lloro, G. Mainetti, D. Maino, E. Maiorano, O. Mansutti, S. Marcin, O. Marggraf, M. Martinelli, N. Martinet, F. Marulli, R. Massey, S. Maurogordato, H. J. McCracken, E. Medinaceli, S. Mei, M. Melchior, Y. Mellier, M. Meneghetti, E. Merlin, G. Meylan, A. Mora, M. Moresco, L. Moscardini, R. Nakajima, C. Neissner, S.-M. Niemi, J. W. Nightingale, C. Padilla, S. Paltani, F. Pasian, K. Pedersen, W. J. Percival, V. Pettorino, S. Pires, G. Polenta, M. Poncet, L. A. Popa, L. Pozzetti, F. Raison, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, R. Saglia, Z. Sakr, A. G. Sánchez, D. Sapone, B. Sartoris, J. A. Schewtschenko, P. Schneider, T. Schrabback, M. Scodeggio, A. Secroun, G. Seidel, M. Seiffert, S. Serrano, P. Simon, C. Sirignano, G. Sirri, L. Stanco, J. Steinwagner, P. Tallada-Crespí, A. N. Taylor, I. Tereno, S. Toft, R. Toledo-Moreo, F. Torradeflot, I. Tutusaus, L. Valenziano, J. Valiviita, T. Vassallo, G. V. Kleijn, A. Veropalumbo, Y. Wang, J. Weller, A. Zacchei, G. Zamorani, F. M. Zerbi, I. A. Zinchenko, E. Zucca, V. Allevato, M. Ballardini, M. Bolzonella, E. Bozzo, C. Burigana, R. Cabanac, A. Cappi, D. Di Ferdinando, J. A. E. Vigo, L. Gabarra, J. Martín-Fleitas, S. Matthew, N. Mauri, R. B. Metcalf, A. Pezzotta, M. Pöntinen, C. Porciani, I. Risso, V. Scottez, M. Sereno, M. Tenti, M. Viel, M. Wiesmann, Y. Akrami, I. T. Andika, S. Anselmi, M. Archidiacono, F. Atrio-Barandela, C. Benoist, K. Benson, D. Bertacca, M. Bethermin, L. Bisigello, A. Blanchard, L. Blot, M. L. Brown, S. Bruton, A. Calabro, B. C. Quevedo, F. Caro, C. S. Carvalho, T. Castro, Y. Charles, F. Cogato, A. R. Cooray, O. Cucciati, S. Davini, F. De Paolis, G. Desprez, A. Díaz-Sánchez, J. J. Diaz, S. Di Domizio, J. M. Diego, P.-A. Duc, A. Enia, Y. Fang, A. G. Ferrari, P. G. Ferreira, A. Finoguenov, A. Fontana, A. Franco, K. Ganga, J. García-Bellido, T. Gasparetto, V. Gautard, E. Gaztanaga, F. Giacomini, F. Gianotti, G. Gozaliasl, M. Guidi, C. M. Gutierrez, A. Hall, W. G. Hartley, S. Hemmati, C. Hernández-Monteagudo, H. Hildebrandt, J. Hjorth, J. J. E. Kajava, Y. Kang, V. Kansal, D. Karagiannis, K. Kiiveri, C. C.

Kirkpatrick, S. Kruk, J. L. Graet, L. Legrand, M. Lembo, F. Lepori, G. Leroy, G. F. Lesci, J. Lesgourgues, L. Leuzzi, T. I. Liaudat, A. Loureiro, J. Macias-Perez, G. Maggio, M. Magliocchetti, E. A. Magnier, F. Mannucci, R. Maoli, C. J. A. P. Martins, L. Maurin, M. Miluzio, P. Monaco, C. Moretti, G. Morgante, C. Murray, K. Naidoo, A. Navarro-Alsina, S. Nesseris, F. Passalacqua, K. Paterson, L. Patrizii, A. Pisani, D. Potter, S. Quai, M. Radovich, S. Sacquegna, M. Sahlén, D. B. Sanders, E. Sarpa, A. Schneider, D. Sciotti, D. Scognamiglio, E. Sellentin, L. C. Smith, K. Tanidis, G. Testera, R. Teyssier, S. Tosi, A. Troja, M. Tucci, C. Valieri, A. Venhola, D. Vergani, G. Verza, P. Vielzeuf, N. A. Walton, and J. G. Sorce. Euclid Quick Data Release (Q1) Exploring galaxy properties with a multi-modal foundation model. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2503.15312.

J. P. Gardner, John C. Mather, Randy Abbott, James S. Abell, Mark Abernathy, Faith E. Abney, John G. Abraham, Roberto Abraham, Yasin M. Abul-Huda, Scott Acton, Cynthia K. Adams, Evan Adams, David S. Adler, Maarten Adriaensen, Jonathan Albert Aguilar, Mansoor Ahmed, Nasif S. Ahmed, Tanjira Ahmed, Rüdeger Albat, Loïc Albert, Stacey Alberts, David Aldridge, Mary Marsha Allen, Shaune S. Allen, Martin Altenburg, Serhat Altunc, Jose Lorenzo Alvarez, Javier Álvarez-Márquez, Catarina Alves de Oliveira, Leslie L. Ambrose, Satya M. Anandakrishnan, Gregory C. Andersen, Harry James Anderson, Jay Anderson, Kristen Anderson, Sara M. Anderson, Julio Aprea, Benita J. Archer, Jonathan W. Arenberg, Ioannis Argyriou, Santiago Arribas, Étienne Artigau, Amanda Rose Arvai, Paul Atcheson, Charles B. Atkinson, Jesse Averbukh, Cagatay Aymergen, John J. Bacinski, Wayne E. Baggett, Giorgio Bagnasco, Lynn L. Baker, Vicki Ann Balzano, Kimberly A. Banks, David A. Baran, Elizabeth A. Barker, Larry K. Barrett, Bruce O. Barringer, Allison Barto, William Bast, Pierre Baudoz, Stefi Baum, Thomas G. Beatty, Mathilde Beaulieu, Kathryn Bechtold, Tracy Beck, Megan M. Beddard, Charles Beichman, Larry Bellagama, Pierre Bely, Timothy W. Berger, Louis E. Bergeron, Antoine-Darveau Bernier, Maria D. Bertch, Charlotte Beskow, Laura E. Betz, Carl P. Biagetti, Stephan Birkmann, Kurt F. Bjorklund, James D. Blackwood, Ronald Paul Blazek, Stephen Blossfeld, Marcel Bluth, Anthony Boccaletti, Martin E. Boegner, Jr., Ralph C. Bohlin, John Joseph Boia, Torsten Böker, N. Bonaventura, Nicholas A. Bond, Kari Ann Bosley, Rene A. Boucarut, Patrice Bouchet, Jeroen Bouwman, Gary Bower, Ariel S. Bowers, Charles W. Bowers, Leslye A. Boyce, Christine T. Boyer, Martha L. Boyer, Michael Boyer, Robert Boyer, Larry D. Bradley, Gregory R. Brady, Bernhard R. Brandl, Judith L. Brannen, David Breda, Harold G. Bremmer, David Brennan, Pamela A. Bresnahan, Stacey N. Bright, Brian J. Broiles, Asa Bromenschenkel, Brian H. Brooks, Keira J. Brooks, Bob Brown, Bruce Brown, Thomas M. Brown, Barry W. Bruce, Jonathan G. Bryson, Edwin D. Bujanda, Blake M. Bullock, A. J. Bunker, Rafael Bureo, Irving J. Burt, James Aaron Bush, Howard A. Bushouse, Marie C. Bussman, Olivier Cabaud, Steven Cale, Charles D. Calhoon, Humberto Calvani, Alicia M. Canipe, Francis M. Caputo, Mihai Cara, Larkin Carey, Michael Eli Case, Thaddeus Cesari, Lee D. Cetorelli, Don R. Chance, Lynn Chandler, Dave Chaney, George N. Chapman, S. Charlot, Pierre Chayer, Jeffrey I. Cheezum, Bin Chen, Christine H. Chen, Brian Cherinka, Sarah C. Chichester, Zachary S. Chilton, Dharini Chittiraibalan, Mark Clampin, Charles R. Clark, Kerry W. Clark, Stephanie M. Clark, Edward E. Claybrooks, Keith A. Cleveland, Andrew L. Cohen, Lester M. Cohen, Knicole D. Colón, Benee L. Coleman, Luis Colina, Brian J. Comber, Thomas M. Comeau, Thomas Comer, Alain Conde Reis, Dennis C. Connolly, Kyle E. Conroy, Adam R. Contos, James Contreras, Neil J. Cook, James L. Cooper, Rachel Aviva Cooper, Michael F. Correia, Matteo Correnti, Christophe Cossou, Brian F. Costanza, Alain Coulais, Colin R. Cox, Ray T. Coyle, Misty M. Cracraft, Keith A. Crew, Gary J. Curtis, Bianca Cusveller, Cleyciane Da Costa Maciel, Christopher T. Dailey, Frédéric Daugeron, Greg S. Davidson, James E. Davies, Katherine Anne Davis, Michael S. Davis, Ratna Day, Daniel de Chambure, Pauline de Jong, Guido De Marchi, Bruce H. Dean, John E. Decker, Amy S. Delisa, Lawrence C. Dell, and Gail Dellagatta. The James Webb Space Telescope Mission. *Publications of the Astronomical Society of the Pacific*, 135(1048):068001, 2023. doi: 10.1088/1538-3873/acd1b5.

F. Gröger, S. Wen, and M. Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian View. *ArXiv e-prints*, 2026. doi: 10.48550/arXiv.2602.14486.

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked Autoencoders Are Scalable Vision Learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18–24. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553.

C. Heneka, F. Nieser, A. Ore, T. Plehn, and D. Schiller. Large Language Models – the Future of Fundamental Physics? *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2506.14757.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training Compute-Optimal Large Language Models. *ArXiv e-prints*, 2022. doi: 10.48550/arXiv.2203.15556.

M. Huh, B. Cheung, T. Wang, and P. Isola. Position: The Platonic Representation Hypothesis. In *International Conference on Machine Learning*, pp. 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

O. Ilbert, S. Arnouts, H. J. McCracken, M. Bolzonella, E. Bertin, O. Le Fèvre, Y. Mellier, G. Zamorani, et al. Accurate photometric redshifts for the cfht legacy survey calibrated using the vimos vlt deep survey. *Astronomy & Astrophysics*, 457:841–856, 2006. doi: 10.1051/0004-6361:20065138.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling Laws for Neural Language Models. *arXiv*, 2020. doi: 10.48550/arXiv.2001.08361.

A. Khederlarian, Brett H. Andrews, Jeffrey A. Newman, Tianqing Zhang, and Biprateep Dey. Optimizing Deep Learning Photometric Redshifts for the Roman Space Telescope with HST/CANDELS. *arXiv e-prints*, art. arXiv:2602.10207, February 2026. doi: 10.48550/arXiv.2602.10207.

N. Koblischke and J. Bovy. SpectraFM: Tuning into Stellar Foundation Models. *ArXiv e-prints*, 2024. doi: 10.48550/arXiv.2411.04750.

S. Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

H. W. Leung and J. Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, 2023. ISSN 0035-8711. doi: 10.1093/mnras/stad3015.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved Baselines with Visual Instruction Tuning. *ArXiv e-prints*, 2023. doi: 10.48550/arXiv.2310.03744.

R. Lupton, M. R. Blanton, G. Fekete, D. W. Hogg, W. O'Mullane, A. Szalay, and N. Wherry. Preparing Red-Green-Blue Images from CCD Data. *Publications of the Astronomical Society of the Pacific*, 116(816):133, 2004. ISSN 1538-3873. doi: 10.1086/382245.

Piero Madau and Mark Dickinson. Cosmic Star-Formation History. ARA&A, 52:415–486, August 2014. doi: 10.1146/annurev-astro-081811-125615.

K. G. Malmquist. On some relations in stellar statistics. *Meddelanden fran Lunds Astronomiska Observatorium Serie I*, 100:1–52, 1922. URL https://ui.adsabs.harvard.edu/abs/1922MeLuF.100....1M/exportcitation.

S. Mishra-Sharma, Y. Song, and J. Thaler. PAPERCLIP: Associating Astronomical Observations and Natural Language with Multi-Modal Models. *ArXiv e-prints*, 2024. doi: 10.48550/arXiv.2403.08851.

S. Miyazaki, Y. Komiyama, S. Kawanomoto, Y. Doi, H. Furusawa, T. Hamana, Y. Hayashi, H. Ikeda, Y. Kamata, H. Karoji, M. Koike, T. Kurakami, S. Miyama, T. Morokuma, F. Nakata, K. Namikawa, H. Nakaya, K. Nariai, Y. Obuchi, Y. Oishi, N. Okada, Y. Okura, P. Tait, T. Takata, Y. Tanaka, M. Tanaka, T. Terai, D. Tomono, F. Uraguchi, T. Usuda, Y. Utsumi, Y. Yamada, H. Yamanoi, H. Aihara, H. Fujimori, S. Mineo, H. Miyatake, M. Oguri, T. Uchida, M. M. Tanaka, N. Yasuda, M. Takada, H. Murayama, A. J. Nishizawa, N. Sugiyama, M. Chiba, T. Futamase, S.-Y. Wang, H.-Y. Chen, P. T. P. Ho, E. J. Y. Liaw, C.-F. Chiu, C.-L. Ho, T.-C. Lai, Y.-C. Lee, D.-Z. Jeng, S. Iwamura, R. Armstrong, S. Bickerton, J. Bosch, J. E. Gunn, R. H. Lupton, C. Loomis,

P. Price, S. Smith, M. A. Strauss, E. L. Turner, H. Suzuki, Y. Miyazaki, M. Muramatsu, K. Yamamoto, M. Endo, Y. Ezaki, N. Ito, N. Kawaguchi, S. Sofuku, T. Taniike, K. Akutsu, N. Dojo, K. Kasumi, T. Matsuda, K. Imoto, Y. Miwa, M. Suzuki, K. Takeshi, and H. Yokota. Hyper Suprime-Cam: System design and verification of image quality. *Publications of the Astronomical Society of Japan*, 70(SP1):S1, 2018. ISSN 0004-6264. doi: 10.1093/pasj/psx063.

K. Moriwaki, R. L. Jun, K. Osato, and N. Yoshida. CosmoGLINT: Cosmological Generative Model for Line Intensity Mapping with Transformer. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2506.16843.

T. D. Nguyen, Y.-S. Ting, I. Ciucă, C. O'Neill, Z.-C. Sun, M. Jabłońska, S. Kruk, E. Perkowski, J. Miller, J. Li, et al. AstroLLaMA: Towards Specialized Foundation Models in Astronomy. *ArXiv e-prints*, 2023. doi: 10.48550/arXiv.2309.06126.

S. C. Odewahn, E. B. Stockwell, R. L. Pennington, R. M. Humphreys, and W. A. Zumach. Automated Star/Galaxy Discrimination With Neural Networks. *The Astronomical Journal*, 103:318, 1992. doi: 10.1086/116063.

A. Ore, C. Heneka, and T. Plehn. SKATR: A Self-Supervised Summary Transformer for SKA. *ArXiv e-prints*, 2024. doi: 10.21468/SciPostPhys.18.5.155.

J.-S. Pan, Y.-S. Ting, Y. Huang, J. Yu, and J.-F. Liu. The Scaling Law in Stellar Light Curves. *ArXiv e-prints*, 2024. doi: 10.48550/arXiv.2405.17156.

L. Parker, F. Lanusse, S. Golkar, L. Sarra, M. Cranmer, A. Bietti, M. Eickenberg, G. Krawezik, M. McCabe, R. Morel, R. Ohana, M. Pettee, B. Régaldo-Saint Blancard, K. Cho, S. Ho, and The Polymathic AI Collaboration. AstroCLIP: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531(4):4990–5011, 2024. ISSN 0035-8711. doi: 10.1093/mnras/stae1450.

E. Perkowski, R. Pan, T. D. Nguyen, Y.-S. Ting, S. Kruk, T. Zhang, C. O'Neill, M. Jablonska, Z. Sun, M. J. Smith, et al. AstroLLaMA-Chat: Scaling AstroLLaMA with Conversational and Diverse Datasets. *Research Notes of the AAS*, 8(1):7, 2024. ISSN 2515-5172. doi: 10.3847/2515-5172/ad1abe.

Plato. *The Republic*. c. 375 BCE.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. *ArXiv e-prints*, 2021. doi: 10.48550/arXiv.2103.00020.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *ArXiv e-prints*, 2014. doi: 10.48550/arXiv.1409.0575.

M. Salvato, O. Ilbert, and B. Hoyle. The many flavours of photometric redshifts. *Nature Astronomy*, 3:212–222, 2019. ISSN 2397-3366. doi: 10.1038/s41550-018-0478-0.

R. Sarmiento, M. Huertas-Company, J. H. Knapen, S. F. Sánchez, H. D. Sánchez, N. Drory, and J. Falcón-Barroso. Capturing the Physics of MaNGA Galaxies with Self-supervised Machine Learning. *The Astrophysical Journal*, 921(2):177, 2021. ISSN 0004-637X. doi: 10.3847/1538-4357/ac1dac.

M. Shuntov, Hollis B. Akins, Louise Paquereau, Caitlin M. Casey, Olivier Ilbert, Rafael C. Arango-Toro, Henry Joy McCracken, Maximilien Franco, Santosh Harish, Jeyhan S. Kartaltepe, Anton M. Koekemoer, Lilan Yang, Marc Huertas-Company, Edward M. Berman, Jacqueline E. McCleary, Sune Toft, Raphaël Gavazzi, Mark J. Achenbach, Emmanuel Bertin, Malte Brinch, Jackie Champagne, Nima Chartab, Nicole E. Drakos, Eiichi Egami, Ryan Endsley, Andreas L. Faisst, Xiaohui Fan, Carter Flayhart, William G. Hartley, Hossein Hatamnia, Ghassem Gozaliasl, Fabrizio Gentile, Iris Jermann, Shuowen Jin, Koki Kakiichi, Ali Ahmad Khostovan, Martin Kümmel, Clotilde Laigle, Ronaldo Laishram, Erini Lambrides, Daizhong Liu, Jianwei Lyu, Georgios Magdis, Bahram Mobasher, Thibaud Moutard, Alvio Renzini, R. Michael Rich, David B. Sanders, Zahra Sattari, Brant E. Robertson, Marc Schefer, Diana Scognamiglio, Nick Scoville,

John D. Silverman, Sina Taamoli, Benny Trakhtenbrot, Francesco Valentino, Feige Wang, John R. Weaver, and Jinyi Yang. COSMOS2025: The COSMOS-Web galaxy catalog of photometry, morphology, redshifts, and physical parameters from JWST, HST, and ground-based imaging. A&A, 704:A339, December 2025. doi: 10.1051/0004-6361/202555799.

O. Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv e-prints*, art. arXiv:2508.10104, August 2025. doi: 10.48550/arXiv.2508.10104.

I. V. Slijepcevic, A. M. M. Scaife, M. Walmsley, M. Bowles, O. I. Wong, S. S. Shabala, and S. V. White. Radio galaxy zoo: towards building the first multipurpose foundation model for radio astronomy with self-supervised learning. *RAS Techniques and Instruments*, 3(1):19–32, 2024. ISSN 2752-8200. doi: 10.1093/rasti/rzad055.

M. J. Smith and J. E. Geach. Astronomia ex machina: a history, primer and outlook on neural networks in astronomy. *R. Soc. Open Sci.*, 10(5):221454, 2023. ISSN 2054-5703. doi: 10.1098/rsos.221454.

M. J. Smith, N. Arora, C. Stone, S. Courteau, and J. E. Geach. Pix2Prof: fast extraction of sequential information from galaxy imagery via a deep natural language 'captioning' model. *Monthly Notices of the Royal Astronomical Society*, 503(1):96–105, 5 2021. ISSN 0035-8711. doi: 10.1093/mnras/stab424.

M. J. Smith, J. E. Geach, Ryan A. Jackson, Nikhil Arora, Connor Stone, and Stéphane Courteau. Realistic galaxy image simulation via score-based generative models. *Monthly Notices of the Royal Astronomical Society*, 511(2):1808–1818, 2022. doi: 10.1093/mnras/stac130.

M. J. Smith, R. J. Roberts, E. Angeloudi, and M. Huertas-Company. AstroPT: Scaling Large Observation Models for Astronomy. *ArXiv e-prints*, 2024. doi: 10.48550/arXiv.2405.14930.

J. S. Speagle, C. L. Steinhardt, P. L. Capak, and J. D. Silverman. A Highly Consistent Framework for the Evolution of the Star-Forming "Main Sequence" from z ~0-6. ApJS, 214(2):15, October 2014. doi: 10.1088/0067-0049/214/2/15.

R. Sutton. The Bitter Lesson. `https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf`, 2019. URL `http://incompleteideas.net/IncIdeas/BitterLesson.html`.

The Multimodal Universe Collaboration. The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data. *Advances in Neural Information Processing Systems*, 37:57841–57913, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/hash/6a57493d35fefea59d06396c7cb69228-Abstract-Datasets_and_Benchmarks_Track.html`.

A. R. Wetzel, J. L. Tinker, and C. Conroy. Galaxy evolution in groups and clusters: star formation rates, red sequence fractions and the persistent bimodality. *Monthly Notices of the Royal Astronomical Society*, 424(1):232–243, 2012. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2012.21188.x.

S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. *ArXiv e-prints*, 2023. doi: 10.48550/arXiv.2301.00808.

J. F. Wu and J. E. G. Peek. Predicting galaxy spectra from images with hybrid convolutional neural networks. *ArXiv e-prints*, 2020. doi: 10.48550/arXiv.2009.12318.

S. Zaman, M. J. Smith, P. Khetarpal, R. Chakrabarty, M. Ginolfi, M. Huertas-Company, M. Jabłońska, S. Kruk, M. L. Lain, S. J. R. Méndez, and D. Tanoglidis. AstroLLaVA: towards the unification of astronomical data and natural language. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2504.08583.

X. Zhao, Y. Huang, G. Xue, X. Kong, J. Liu, X. Tang, T. C. Beers, Y.-S. Ting, and A.-L. Luo. SpecCLIP: Aligning and Translating Spectroscopic Measurements for Stars. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2507.01939.

X. Zuo, Y. Tao, Y. Huang, Z. Kang, H. Chen, C. Cui, J. Pan, X. Kong, X. Tang, H. Han, H. Mu, Y. Xu, D. Fan, G. Xue, A. Luo, and J. Liu. FALCO: a Foundation model of Astronomical Light Curves for time dOmain astronomy. *ArXiv e-prints*, 2025. doi: 10.48550/arXiv.2504.20290.

CONTENTS

1 Astronomy and the Platonic Representation Hypothesis **1**

2 Datasets, models, and metrics **2**

3 Experiments and results **4**

3.1 Physical interpretation experiments 4

3.2 Global and local embedding alignment experiments 6

4 Conclusions: think local, not global **9**

A Code, data, and compute **19**

A.1 Code and data availability 19

A.2 Compute resources and experimental cost 19

A.3 Dataset provenance 19

B Extended physics results **20**

B.1 Embedding width and probe capacity 20

C Anomaly investigation **24**

C.1 DINOv3 24

C.2 PaliGemma 24

D Extended intra-architectural results **25**

E Extended crossmodal results **27**

F Extended cross-architectural results **29**

F.1 Intervening on the shared physics subspace 29

A CODE, DATA, AND COMPUTE

A.1 CODE AND DATA AVAILABILITY

During the review process we make our code available anonymously at `https://anonymous.4open.science/r/platonic-universe-7330`, and our embeddings and data available at `https://huggingface.co/HCVYM5w6Gn`. Upon deanonymization we will link to all code and data publicly.

A.2 COMPUTE RESOURCES AND EXPERIMENTAL COST

Extracting embeddings across the JWST, HSC, Legacy Survey, and DESI datasets cost the most GPU-hours, particularly for the larger vision-language models. In total, the experiments required approximately 3000 GPU-hours across various hardware tiers. Various downstream tasks (MKNN/CKA computation, linear probing, and statistical analyses) were computationally cheaper compared to extraction, requiring about one CPU-hour per probe once embeddings were cached.

Table 2: Hardware usage summary by chip type.

Hardware	Approximate Usage
NVIDIA A40 / A80	300 GPU-hours
NVIDIA A100 / A800	600 GPU-hours
NVIDIA GH200 / H100	2000 GPU-hours
NVIDIA RTX 5070 and RTX 3090	50 GPU-hours
Standard CPU cores	$\approx$ 1 CPU-hour per probe
Total	**3,000 GPU-hours**

A.3 DATASET PROVENANCE

Here we discuss the datasets used to pretrain our examined models in order to rule out significant exposure to galaxy imagery. ViT, ViT-MAE, ConvNeXtv2, and I-JEPA are pretrained on ImageNet-1k or ImageNet-22k (Russakovsky et al., 2014), neither of which includes a galaxy category. V-JEPA 2 uses ImageNet and video data. These datasets provide no dedicated galaxy-image training, unlike AstroPTv2, which is explicitly pretrained on astronomical imagery. Incidental galaxy exposure remains possible in internet-derived corpora, including those used by CLIP, DINOv3, LLaVA, and PaliGemma 2. We note that I-JEPA and V-JEPA 2 sit above the overall physics-scaling trend (Fig. 1), including above AstroPTv2. Dedicated galaxy-image pretraining is therefore not necessary to recover the physical information measured by our probes. We also note that we observe MKNN scaling between HSC imagery and DESI spectra (Fig. 5), with Specformer pretrained solely on spectra, further weakening an explanation based on shared image-level training examples.

B EXTENDED PHYSICS RESULTS

Here we show the full per-property results for the physics linear-probe experiments. Tab. 3 lists mean and standard deviation R^2 values across 10-fold cross-validation for each model on HSC and JWST imagery, broken down by redshift, stellar mass, and sSFR. Fig. 7 shows probe R^2 against parameter count separately for each of the three physical properties. Figs. 8 and 9 show the per-model cosine similarity matrices between the redshift, stellar mass, and sSFR probe weight vectors for HSC and JWST imagery respectively.

Table 3: Downstream regression performance (R^2) when predicting redshift (z), stellar mass ($\log M_\star$), and specific star formation rate (sSFR) from frozen embeddings of each model, evaluated on 45 000 HSC and JWST galaxies. Values are mean ± standard deviation over 10 k-folds.

	HSC (R^2)			JWST (R^2)		
Model	z	$\log M_\star$	sSFR	z	$\log M_\star$	sSFR
AstroPTv2 Small	$0.387_{\pm 0.013}$	$0.525_{\pm 0.009}$	$0.462_{\pm 0.002}$	$0.349_{\pm 0.010}$	$0.674_{\pm 0.010}$	$0.293_{\pm 0.008}$
AstroPTv2 Base	$0.443_{\pm 0.017}$	$0.556_{\pm 0.012}$	$0.502_{\pm 0.006}$	$0.472_{\pm 0.015}$	$0.771_{\pm 0.008}$	$0.336_{\pm 0.005}$
AstroPTv2 Large	$0.453_{\pm 0.007}$	$0.559_{\pm 0.013}$	$0.510_{\pm 0.007}$	$0.531_{\pm 0.012}$	$0.802_{\pm 0.006}$	$0.360_{\pm 0.005}$
ViT Base	$0.359_{\pm 0.006}$	$0.455_{\pm 0.013}$	$0.434_{\pm 0.006}$	$0.372_{\pm 0.012}$	$0.597_{\pm 0.004}$	$0.306_{\pm 0.012}$
ViT Large	$0.369_{\pm 0.007}$	$0.474_{\pm 0.007}$	$0.440_{\pm 0.005}$	$0.408_{\pm 0.010}$	$0.638_{\pm 0.006}$	$0.340_{\pm 0.009}$
ViT Huge	$0.434_{\pm 0.011}$	$0.520_{\pm 0.008}$	$0.480_{\pm 0.008}$	$0.477_{\pm 0.008}$	$0.722_{\pm 0.004}$	$0.358_{\pm 0.009}$
ViT-MAE Base	$0.343_{\pm 0.010}$	$0.450_{\pm 0.009}$	$0.422_{\pm 0.008}$	$0.290_{\pm 0.011}$	$0.581_{\pm 0.009}$	$0.277_{\pm 0.015}$
ViT-MAE Large	$0.357_{\pm 0.012}$	$0.461_{\pm 0.009}$	$0.441_{\pm 0.007}$	$0.324_{\pm 0.012}$	$0.631_{\pm 0.010}$	$0.295_{\pm 0.011}$
ViT-MAE Huge	$0.372_{\pm 0.009}$	$0.486_{\pm 0.008}$	$0.470_{\pm 0.005}$	$0.357_{\pm 0.011}$	$0.668_{\pm 0.010}$	$0.310_{\pm 0.010}$
DINOv3 vits16	$0.359_{\pm 0.005}$	$0.434_{\pm 0.008}$	$0.418_{\pm 0.008}$	$0.402_{\pm 0.009}$	$0.607_{\pm 0.006}$	$0.334_{\pm 0.008}$
DINOv3 vits16plus	$0.357_{\pm 0.008}$	$0.440_{\pm 0.008}$	$0.426_{\pm 0.008}$	$0.380_{\pm 0.013}$	$0.601_{\pm 0.008}$	$0.336_{\pm 0.011}$
DINOv3 vitb16	$0.383_{\pm 0.005}$	$0.465_{\pm 0.008}$	$0.451_{\pm 0.011}$	$0.441_{\pm 0.008}$	$0.646_{\pm 0.007}$	$0.368_{\pm 0.008}$
DINOv3 vitl16	$0.376_{\pm 0.004}$	$0.454_{\pm 0.006}$	$0.453_{\pm 0.008}$	$0.443_{\pm 0.008}$	$0.626_{\pm 0.006}$	$0.367_{\pm 0.012}$
DINOv3 vith16plus	$0.366_{\pm 0.010}$	$0.450_{\pm 0.006}$	$0.439_{\pm 0.008}$	$0.464_{\pm 0.012}$	$0.630_{\pm 0.007}$	$0.379_{\pm 0.011}$
DINOv3 vit7b16	$0.421_{\pm 0.007}$	$0.495_{\pm 0.008}$	$0.489_{\pm 0.013}$	$0.528_{\pm 0.004}$	$0.704_{\pm 0.003}$	$0.421_{\pm 0.011}$
ConvNeXtv2 Nano	$0.392_{\pm 0.005}$	$0.490_{\pm 0.009}$	$0.475_{\pm 0.007}$	$0.457_{\pm 0.010}$	$0.690_{\pm 0.007}$	$0.359_{\pm 0.011}$
ConvNeXtv2 Tiny	$0.412_{\pm 0.009}$	$0.496_{\pm 0.009}$	$0.470_{\pm 0.006}$	$0.455_{\pm 0.016}$	$0.675_{\pm 0.007}$	$0.375_{\pm 0.008}$
ConvNeXtv2 Base	$0.409_{\pm 0.004}$	$0.501_{\pm 0.008}$	$0.488_{\pm 0.011}$	$0.460_{\pm 0.009}$	$0.676_{\pm 0.003}$	$0.386_{\pm 0.006}$
ConvNeXtv2 Large	$0.425_{\pm 0.008}$	$0.522_{\pm 0.010}$	$0.506_{\pm 0.008}$	$0.498_{\pm 0.012}$	$0.713_{\pm 0.004}$	$0.400_{\pm 0.010}$
I-JEPA Huge	$0.429_{\pm 0.008}$	$0.528_{\pm 0.007}$	$0.499_{\pm 0.006}$	$0.524_{\pm 0.011}$	$0.780_{\pm 0.006}$	$0.403_{\pm 0.009}$
I-JEPA Giant	$0.424_{\pm 0.004}$	$0.528_{\pm 0.007}$	$0.506_{\pm 0.012}$	$0.555_{\pm 0.013}$	$0.790_{\pm 0.006}$	$0.416_{\pm 0.008}$
V-JEPA2 Large	$0.421_{\pm 0.008}$	$0.544_{\pm 0.010}$	$0.517_{\pm 0.006}$	$0.525_{\pm 0.010}$	$0.780_{\pm 0.006}$	$0.436_{\pm 0.007}$
V-JEPA2 Huge	$0.420_{\pm 0.006}$	$0.549_{\pm 0.012}$	$0.518_{\pm 0.006}$	$0.521_{\pm 0.011}$	$0.781_{\pm 0.005}$	$0.439_{\pm 0.012}$
V-JEPA2 Giant	$0.450_{\pm 0.006}$	$0.560_{\pm 0.011}$	$0.530_{\pm 0.007}$	$0.565_{\pm 0.011}$	$0.810_{\pm 0.003}$	$0.457_{\pm 0.009}$
CLIP Base	$0.433_{\pm 0.009}$	$0.517_{\pm 0.008}$	$0.511_{\pm 0.007}$	$0.473_{\pm 0.009}$	$0.706_{\pm 0.003}$	$0.358_{\pm 0.011}$
CLIP Large	$0.435_{\pm 0.008}$	$0.515_{\pm 0.010}$	$0.496_{\pm 0.007}$	$0.533_{\pm 0.009}$	$0.736_{\pm 0.004}$	$0.385_{\pm 0.010}$
PaliGemma 2 3B	$0.458_{\pm 0.007}$	$0.539_{\pm 0.008}$	$0.529_{\pm 0.011}$	$0.528_{\pm 0.011}$	$0.764_{\pm 0.005}$	$0.378_{\pm 0.007}$
PaliGemma 2 10B	$0.464_{\pm 0.009}$	$0.537_{\pm 0.011}$	$0.522_{\pm 0.011}$	$0.531_{\pm 0.015}$	$0.773_{\pm 0.004}$	$0.359_{\pm 0.008}$
PaliGemma 2 28B	$0.398_{\pm 0.009}$	$0.492_{\pm 0.009}$	$0.478_{\pm 0.010}$	$0.465_{\pm 0.014}$	$0.722_{\pm 0.006}$	$0.310_{\pm 0.014}$
LLaVA 1.5 7B	$0.431_{\pm 0.009}$	$0.522_{\pm 0.013}$	$0.504_{\pm 0.006}$	$0.528_{\pm 0.014}$	$0.752_{\pm 0.005}$	$0.390_{\pm 0.010}$
LLaVA 1.5 13B	$0.429_{\pm 0.008}$	$0.518_{\pm 0.009}$	$0.502_{\pm 0.004}$	$0.531_{\pm 0.011}$	$0.756_{\pm 0.003}$	$0.394_{\pm 0.012}$

B.1 EMBEDDING WIDTH AND PROBE CAPACITY

Wider embeddings give a linear probe more trainable coefficients, raising the possibility that differences in physics-probe performance reflect probe capacity. To assess this possibility, we group 19 models by embedding width (384, 768, 1024, and 1280) and perform a one-way ANOVA comparing physics-probe performance across these groups, separately for HSC and JWST. We find no statistically significant differences between width groups: $F(3, 15) = 0.92$, $p = 0.46$ for HSC and $F(3, 15) = 1.29$, $p = 0.31$ for JWST.

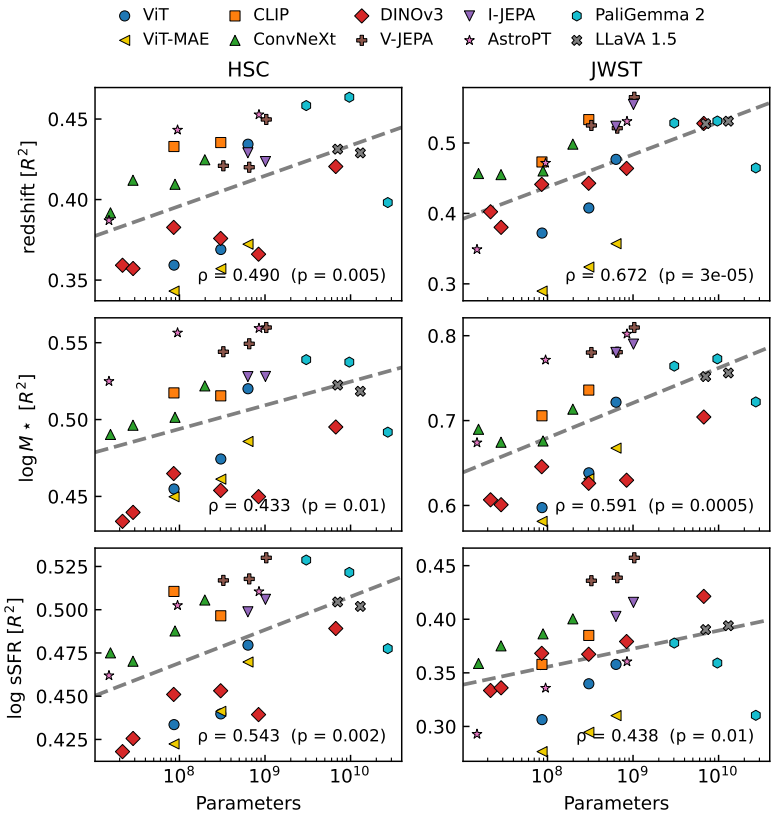


Figure 7: Parameter count vs probe performance for our tested physical properties. Spearman's ρ and p values are stated for each panel. In all tested cases we find a significant correlation between parameter count and physics probe performance.

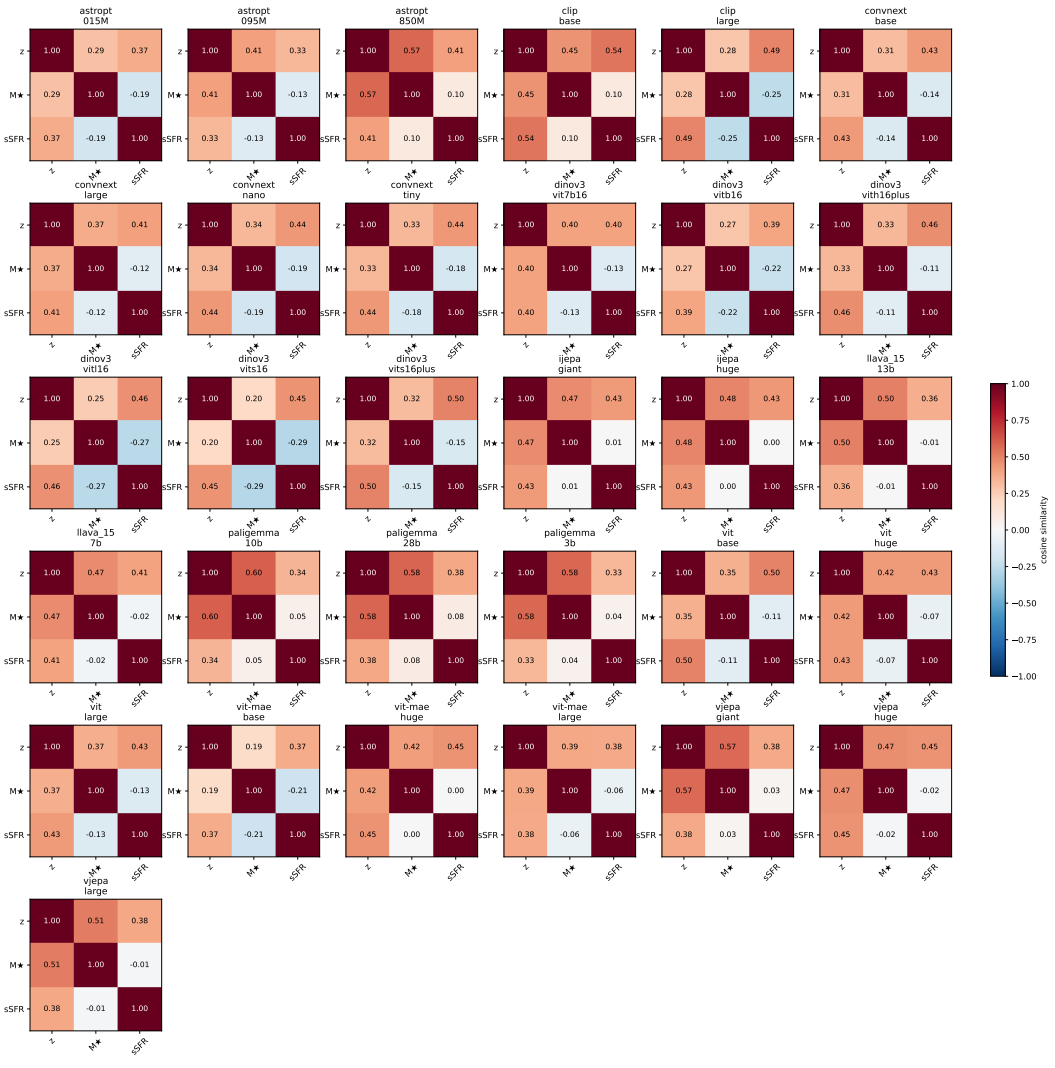


Figure 8: Cosine similarity matrices between the redshift (z), stellar mas ($M_\star$), and sSFR probe weight vectors using HSC images.

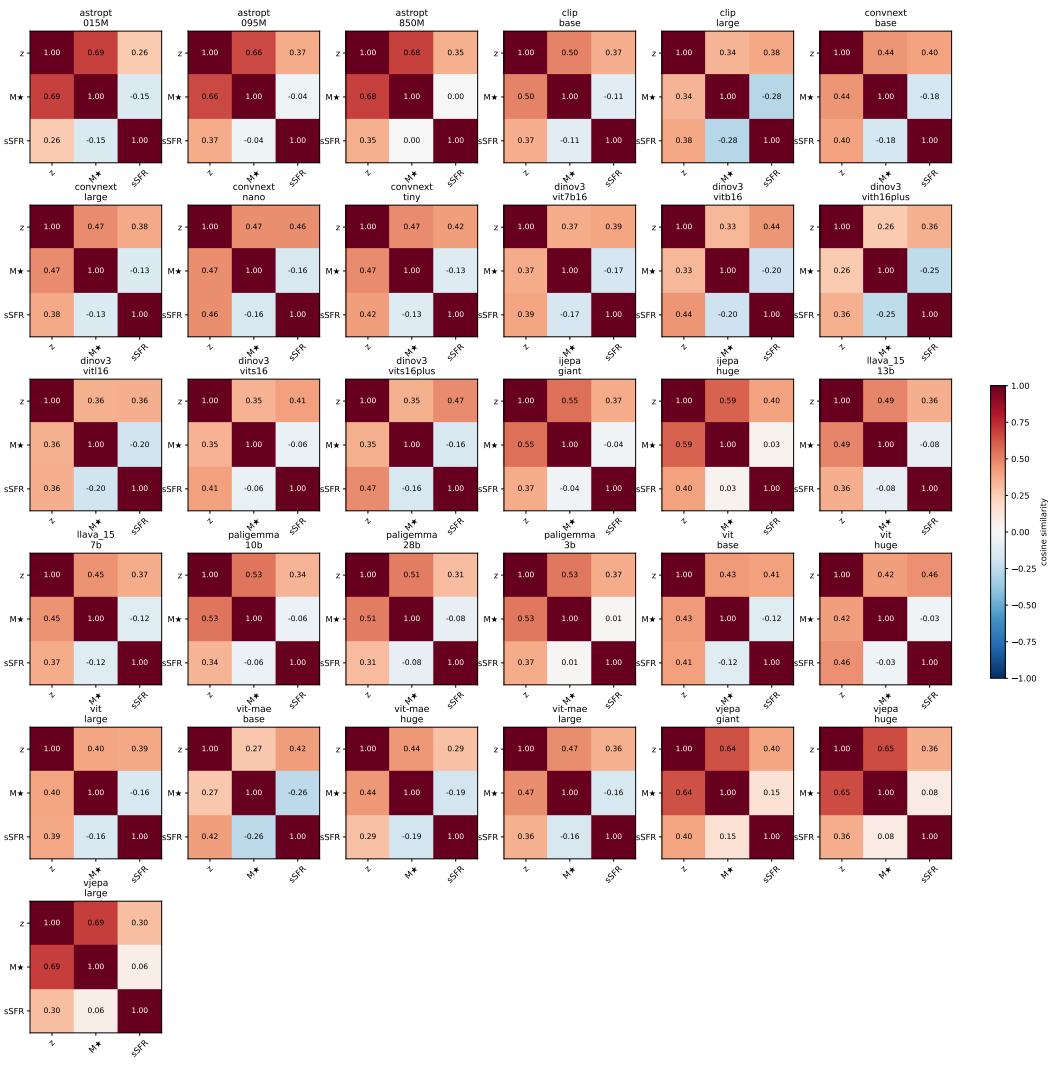


Figure 9: Cosine similarity matrices between the redshift (z), stellar mas ($M_\star$), and sSFR probe weight vectors using JWST images.

C ANOMALY INVESTIGATION

Here we provide extended analysis on DINOv3's and PaliGemma's anomalous behavior.

C.1 DINOv3

DINOv3's VIT-S/16 through ViT-H+/16 models are distilled from a fixed ViT-7B/16 teacher trained from scratch. On the 1.67k galaxies with paired HSC and JWST imaging we compute the student models' MKNN alignment with the teacher and its mean pairwise alignment of the non-DINOv3 model basket. Tab. 4 reports the results. We see that teacher alignment generally increases with student model size while basket alignment decreases. This is unlike our other tested models where basket alignment tracks model performance (§3.2 ¶4). This pattern supports the interpretation that increasing student capacity improves alignment to a teacher model without necessarily improving alignment with independently trained models.

Table 4: DINOv3 MKNN alignment with the ViT-7B/16 teacher and the non-DINOv3 model basket, evaluated on 1.67k galaxies with paired HSC and JWST imaging. Basket alignment is averaged over non-DINOv3 models. Scores are percentages, with uncertainties estimated from 200 draws containing 80% of the sample each.

		HSC MKNN (%)		JWST MKNN (%)	
Model	Params (M)	Teacher	Basket	Teacher	Basket
ViT-S/16	21	25.0 ± 0.3	24.3 ± 0.2	27.8 ± 0.3	23.0 ± 0.2
ViT-S+/16	29	25.7 ± 0.3	23.2 ± 0.2	27.8 ± 0.3	21.5 ± 0.2
ViT-B/16	86	26.0 ± 0.3	23.1 ± 0.2	31.3 ± 0.3	22.4 ± 0.2
ViT-L/16	300	29.4 ± 0.3	18.7 ± 0.2	37.4 ± 0.4	18.0 ± 0.2
ViT-H+/16	840	29.3 ± 0.3	15.0 ± 0.2	40.5 ± 0.3	16.0 ± 0.2
ViT-7B/16 (teacher)	6716	—	17.6 ± 0.2	—	19.4 ± 0.2

C.2 PALIGEMMA

Here we investigate PaliGemma's physics performance in more depth. We run the linear probing tests described in §3.1 over each layer in the PaliGemma model of families. Fig. 10 shows our results. We can see that the fixed-size SigLIP image encoder network eliminates the correlation between network size and physics recovery that we observe in other networks (Fig. 1); we even find that the scaling of the Gemma LLM backbone network is detrimental to physics recovery! This is likely due to nuisance noise being introduced by the language trunk to the astronomical signal passed by SigLIP, with more noise added the further the signal propagates through the Gemma trunk.

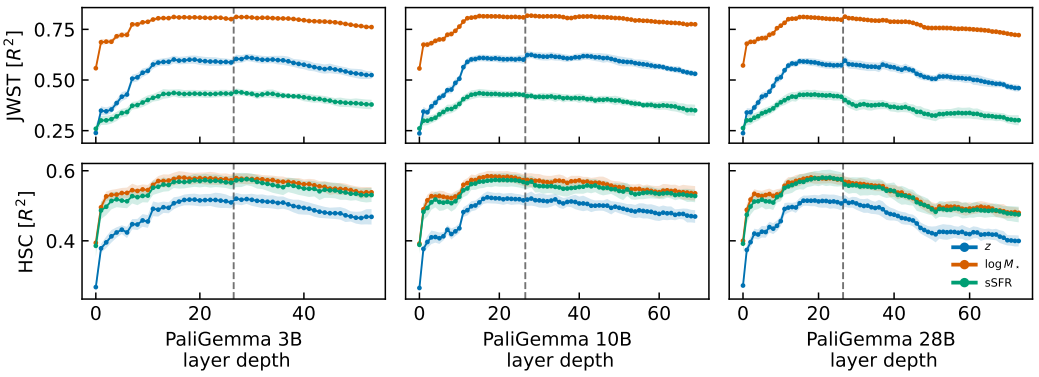


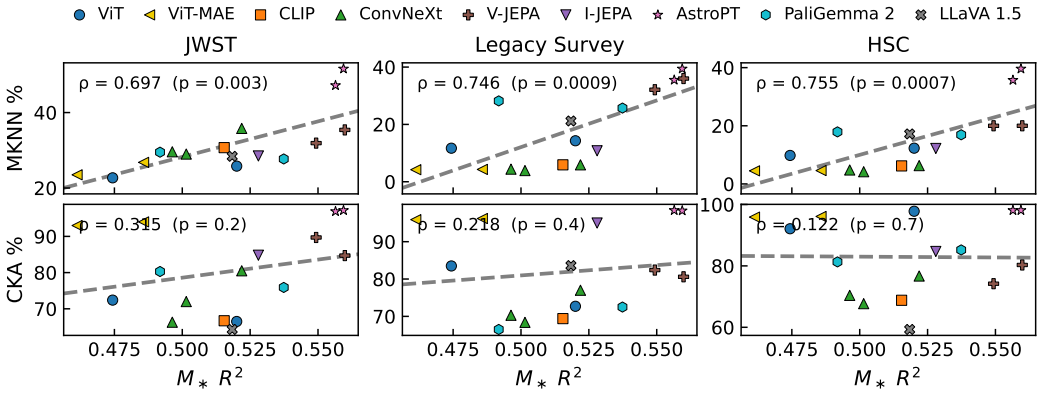
Figure 10: Physics linear R^2 probing results for the PaliGemma 3B, 10B, and 28B networks on a layer-wise basis. The $k = 10$-fold standard deviation is shaded. We demarcate the boundary between the fixed-size SigLIP image encoder and Gemma decoder networks with a gray dashed line.

D EXTENDED INTRA-ARCHITECTURAL RESULTS

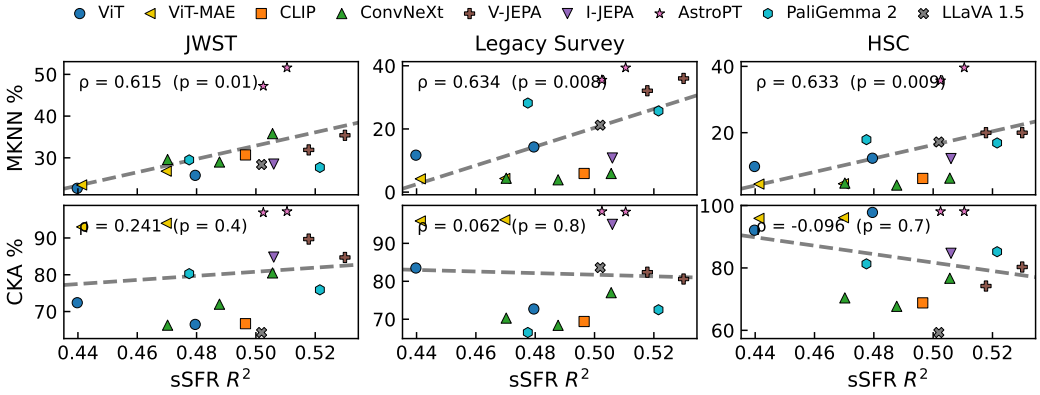
We list the full results across pairs of models within the same family in Tab. 5. Each entry is the MKNN or CKA score as a percentage. Fig. 11 plots intra-architectural MKNN and CKA against linear-probe R^2 for each of redshift, stellar mass, and sSFR individually.

Table 5: Intra-architectural embedding alignment within a model family, measured by MKNN and CKA scores. The PRH predicts that both MKNN and CKA scores will increase as we compare the embeddings of larger model pairs within a model family, since larger models will generate embeddings closer to the Platonic ideal representation.

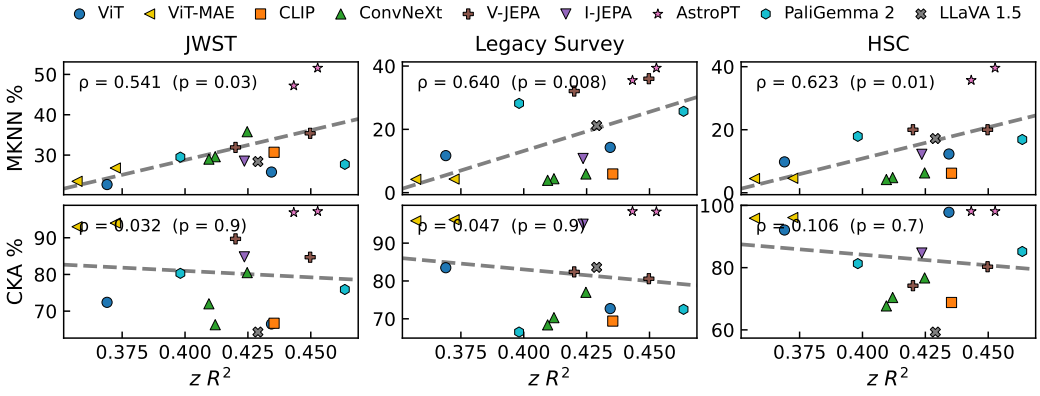
	MKNN (%)			CKA (%)		
Model Pairs	JWST	Legacy	HSC	JWST	Legacy	HSC
AstroPTv2 Small vs Base	47.2	35.5	35.7	96.9	98.4	98.1
AstroPTv2 Base vs Large	51.6	39.4	39.6	97.2	98.3	98.1
CLIP Base vs Large	30.7	5.9	6.2	66.7	69.4	68.8
ConvNeXtv2 Nano vs Tiny	29.6	4.4	4.8	66.3	70.3	70.4
ConvNeXtv2 Tiny vs Base	29.0	3.9	4.2	72.0	68.4	67.7
ConvNeXtv2 Base vs Large	35.8	5.9	6.3	80.5	77.0	76.7
DINOv3 vits16 vs vits16plus	43.6	10.8	16.5	88.5	89.7	86.7
DINOv3 vits16plus vs vitb16	38.2	9.6	14.3	86.5	88.5	86.3
DINOv3 vitb16 vs vitl16	32.7	7.1	9.4	73.6	73.1	68.0
DINOv3 vitl16 vs vith16plus	34.0	6.3	8.8	78.5	68.7	66.6
DINOv3 vith16plus vs vit7b16	40.5	7.8	11.7	81.8	62.1	64.2
IJEPA Huge vs Giant	28.5	10.8	12.2	84.8	95.0	84.7
LLaVA-1.5 7B vs 13B	28.4	21.2	17.2	64.3	83.6	59.3
ViT-MAE Base vs Large	23.5	4.2	4.5	93.0	95.9	95.9
ViT-MAE Large vs Huge	26.8	4.3	4.6	94.0	96.2	96.1
ViT Base vs Large	22.7	11.7	9.8	72.4	83.5	92.1
ViT Large vs Huge	25.8	14.3	12.3	66.5	72.7	97.8
PaliGemma 2 3B vs 10B	27.7	25.7	16.9	75.9	72.5	85.2
PaliGemma 2 10B vs 28B	29.5	28.2	17.9	80.3	66.5	81.3
V-JEPA2 ViT-L vs H	31.9	32.1	20.0	89.7	82.4	74.2
V-JEPA2 ViT-H vs G	35.4	36.0	20.0	84.7	80.6	80.3



(a) Mass ($M_\star$).



(b) Specific Star Formation Rate (sSFR).



(c) Redshift (z).

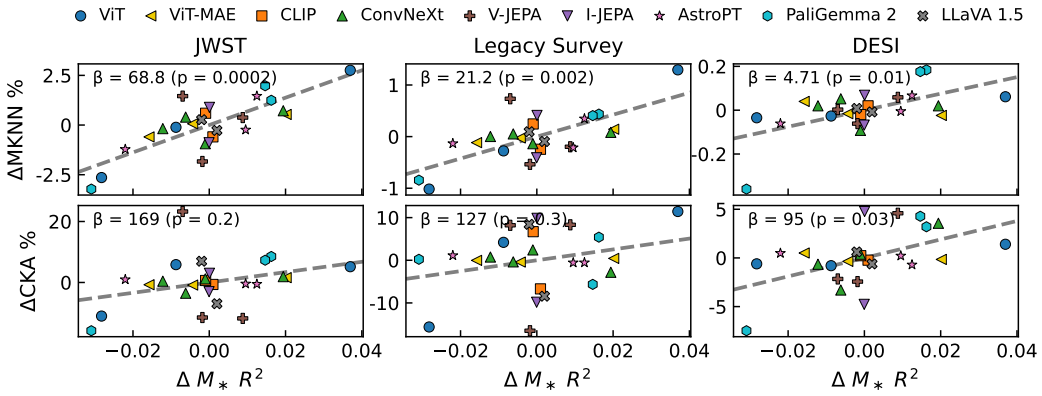
Figure 11: Intra-architectural embedding alignment (MKNN and CKA) vs probe performance for our tested physical properties. Spearman's ρ and p values are stated for each panel. In all tested cases we find a significant correlation between MKNN and physics probe performance. We do not find a significant correlation between CKA and physics probe performance in any tested case.

E EXTENDED CROSSMODAL RESULTS

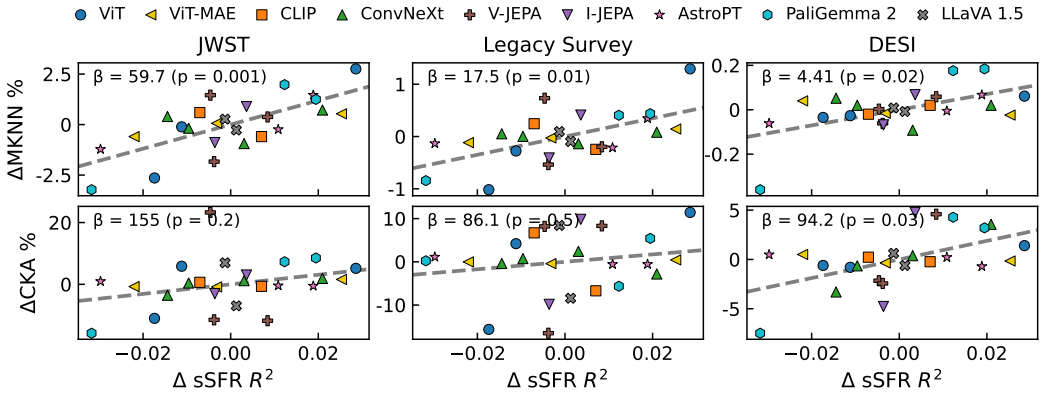
We list the full results across pairs of data modalities for each model variant in Tab. 6. Each entry is the MKNN or CKA score as a percentage. Fig. 12 plots cross modal MKNN and CKA against linear-probe R^2 for each of redshift, stellar mass, and sSFR individually.

Table 6: Crossmodal embedding alignment between a model's embeddings of different astronomical modalities and its embeddings of HSC imaging, measured by MKNN and CKA scores. The DESI spectra column uses Specformer embeddings for the DESI side compared against each listed vision model's HSC image embeddings. The PRH predicts that both MKNN and CKA scores will increase as models within a family grow in size, since larger models should produce embeddings closer to the Platonic ideal representation.

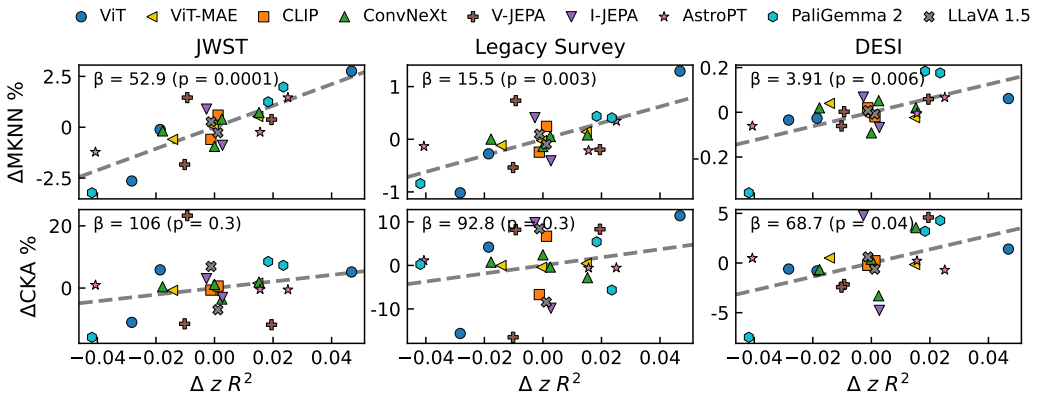
	MKNN (%)			CKA (%)		
Model	JWST	Legacy	DESI	JWST	Legacy	DESI
AstroPTv2 Small	11.62	3.36	1.25	43.4	86.6	45.66
AstroPTv2 Base	12.60	3.28	1.32	42.0	84.9	45.39
AstroPTv2 Large	14.30	3.84	1.41	41.9	84.9	44.47
CLIP Base	12.88	1.79	1.29	30.59	59.32	33.39
CLIP Large	14.07	2.28	1.24	31.89	72.73	33.85
ConvNeXtv2 Nano	11.33	1.13	1.01	32.52	54.06	32.29
ConvNeXtv2 Tiny	11.91	1.18	1.05	28.57	52.93	29.64
ConvNeXtv2 Base	10.57	0.99	0.87	33.28	55.73	33.32
ConvNeXtv2 Large	12.23	1.21	1.01	34.01	50.49	36.50
DINOv3 vits16	14.55	2.42	0.98	52.18	70.23	34.80
DINOv3 vits16plus	13.09	2.09	0.97	48.53	68.06	35.39
DINOv3 vitb16	14.03	2.48	0.92	49.44	70.29	33.25
DINOv3 vitl16	11.80	1.70	0.79	45.31	66.07	30.14
DINOv3 vith16plus	10.35	1.16	0.67	31.58	43.12	22.21
DINOv3 vit7b16	12.62	1.89	0.82	40.14	41.65	30.13
IJEPA Huge	9.85	1.27	0.56	15.22	31.64	15.86
IJEPA Giant	11.63	2.09	0.73	21.25	51.31	25.40
LLaVA-1.5 7B	11.16	1.67	1.04	31.10	56.45	36.88
LLaVA-1.5 13B	11.70	1.86	1.06	45.03	73.30	38.11
PaliGemma 2 3B	11.50	1.84	1.18	43.20	58.58	32.89
PaliGemma 2 10B	12.23	1.81	1.17	41.98	47.51	33.97
PaliGemma 2 28B	7.03	0.56	0.50	31.2	53.39	22.21
ViT-MAE Base	10.09	1.89	1.01	59.34	94.89	29.25
ViT-MAE Large	10.75	1.98	0.94	59.23	94.49	28.39
ViT-MAE Huge	11.22	2.15	0.93	61.66	95.35	28.59
ViT Base	10.48	1.20	1.02	30.13	47.35	35.17
ViT Large	13.01	1.94	1.03	47.00	67.19	34.98
ViT Huge	15.88	3.51	1.14	46.32	74.42	37.18
V-JEPA2 ViT-L	14.27	3.04	0.82	59.63	77.48	23.98
V-JEPA2 ViT-H	10.98	1.77	0.74	24.86	52.76	23.71
V-JEPA2 ViT-G	13.19	2.11	0.89	24.57	77.59	30.73



(a) Mass ($M_\star$).



(b) Specific Star Formation Rate (sSFR).



(c) Redshift (z).

Figure 12: Crossmodal embedding alignment between the stated modality and HSC plotted against probe performance for our tested physical properties. ANCOVA β and p values are stated for each panel. In all tested cases we find a significant correlation between MKNN and physics probe performance. We only find a significant correlation between CKA and physics probe performance for DESI Spectra vs HSC.

F EXTENDED CROSS-ARCHITECTURAL RESULTS

This appendix expands the cross-architectural analysis. Figure 13 plots each model's mean MKNN and CKA against the rest of the basket against its linear-probe R^2 for redshift, stellar mass, and sSFR individually, on both JWST and HSC imagery.

F.1 INTERVENING ON THE SHARED PHYSICS SUBSPACE

The analyses in §3.2 show that local embedding alignment tracks physics-probe performance across model families and astronomical modalities. These relationships are nevertheless observational: a third property of the representations could in principle improve both physical recovery and embedding similarity. We therefore perform a direct intervention on the embedding spaces to ask whether the physical content identified by our probes itself contributes to representational alignment.

We test 21 pairs drawn from the ViT and ConvNeXtv2 families on HSC imagery. We choose these two families as both are supervised ImageNet-pretrained vision models, reducing variation in the pretraining regime. For each model we fit the three linear probes described in §3.1, corresponding to redshift (z), stellar mass ($M_\star$), and sSFR. Because these quantities are not independent (Fig. 3), we treat their probe directions jointly and remove the subspace spanned by all three from each model's embeddings. We then recompute MKNN and CKA between each pair after the intervention.

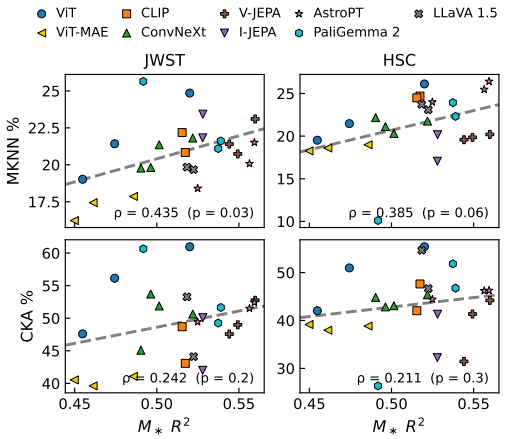
Our relevant control is not about removing information from the embedding: any low-rank projection may perturb its geometry. We therefore repeat the experiment after removing a random subspace of the same rank. If the alignment observed in §3.2 is specifically supported by shared physical content, removing the probe-defined physics subspace should reduce similarity more than removing an equivalently sized random subspace.

Table 7: Effect of removing the probe-defined physics subspace from model embeddings. Values give the mean [min, max] percentage decrease in alignment over 21 tested model pairs. Negative values indicate that alignment increased after the intervention. The random control removes a subspace of equal rank.

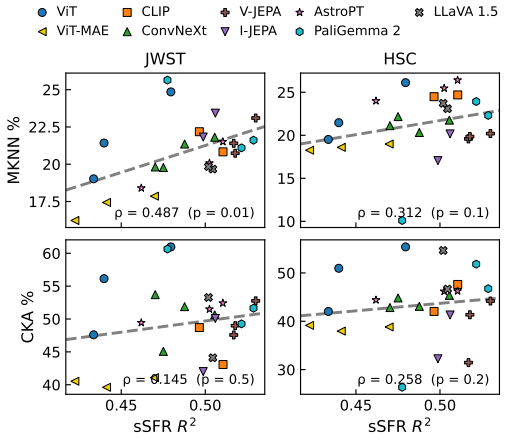
	Physics subspace	Random subspace
CKA (global)	21% $[9, 28]$	0% $[-0, 0]$
MKNN (local)	7% $[2, 12]$	-7% $[-12, -2]$

Removing the physics subspace reduces MKNN for every tested pair, with a mean decrease of 7% and a range of 2–12%. Removing a random subspace of the same rank has the opposite effect: MKNN increases by 7% on average. We find an even larger change in absolute CKA, which decreases by 21% after removing the physics subspace while remaining effectively unchanged under the random control.

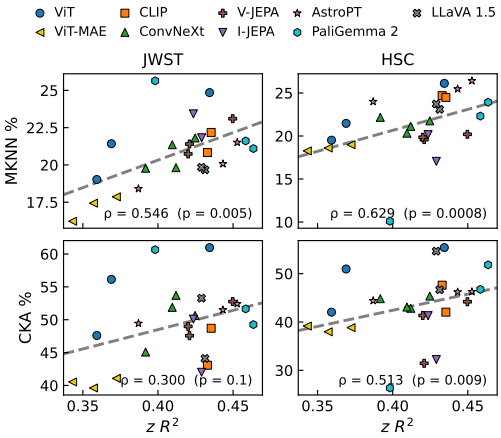
This intervention answers a different question from the scaling experiments in §3.2. The latter ask whether representational similarity *tracks* physics performance as models improve, for which we consistently find a relationship for MKNN but not CKA. Here we instead ask whether the probe-defined physical content makes a direct contribution to the similarity between two fixed embedding spaces. The decrease following targeted removal, together with the equal-rank random control, shows that the shared physics subspace is not merely correlated with the observed alignment: intervening on that subspace directly changes it. This provides causal evidence that the astrophysical content recovered by our probes contributes to the representational convergence measured in the main experiments.



(a) Mass ($M_\star$).



(b) Specific Star Formation Rate (sSFR).



(c) Redshift (z).

Figure 13: Cross-architectural embedding alignment between the stated modality and HSC plotted against probe performance for our tested physical properties. Spearman's ρ and p values are stated for each panel. In all tested JWST cases we find a significant correlation between MKNN and physics probe performance. For HSC we find significant or borderline correlations between MKNN and physics probe performance. We only find a significant correlation between CKA and physics probe performance for HSC redshift.