The Galaxy's Guide to the Tokenizer: A Benchmark for Scientific Foundation Models

Sogol Sanjaripour
University of California, Riverside
Astro-AI, Harvard University

2nd Annual Cosmic Horizons Conference · Charlottesville, Virginia · July 2026

The Galaxy's Guide to the Tokenizer: A Benchmark for Scientific Foundation Models

Sogol Sanjaripour[1,*] **Michael J. Smith**[2,3,4] **Manuel Pérez-Carrasco**[2,3]
Juan Rafael Martínez-Galarza[2,3] **Bahram Mobasher**[1] **Gabriela Canalizo**[1]

[1]University of California, Riverside [2]AstroAI
[3]Center for Astrophysics | Harvard & Smithsonian [4]UniverseTBD

*sogol.sanjaripour@email.ucr.edu

The big data era in astronomy

- **Exponential increase in the astronomical data:**
  - Billions of galaxy images, millions of spectra.
  - Billions of star positions and motions.
  - Tens of millions of brightness measurements over time.
- **Different surveys mapping astronomical objects across the electromagnetic spectrum:**

M31 (the Andromeda Galaxy) across the electromagnetic spectrum · image: R. Simpson (orbitingfrog), via K. Masters

Radio | Microwave | Infrared | Visible | Ultraviolet | X-ray

WHY ASTRONOMY

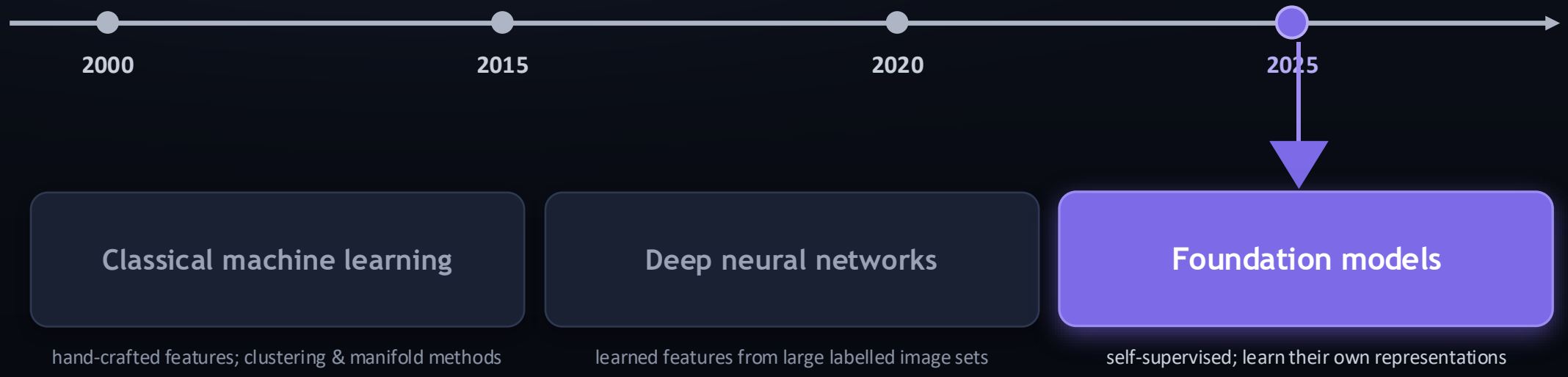
Astronomy: A Natural Match for AI

All of that data is exactly what modern representation learning was built for.

1. **Beyond human scale** billions of objects — far more than anyone could ever label or inspect manually.
2. **Multi-modal** images, spectra, light curves, and catalogs all describe the same physical objects.
3. **Physics as ground truth** redshift, stellar mass, and star-formation rate give objective labels for free.

The Evolution of Machine Learning in Astronomy

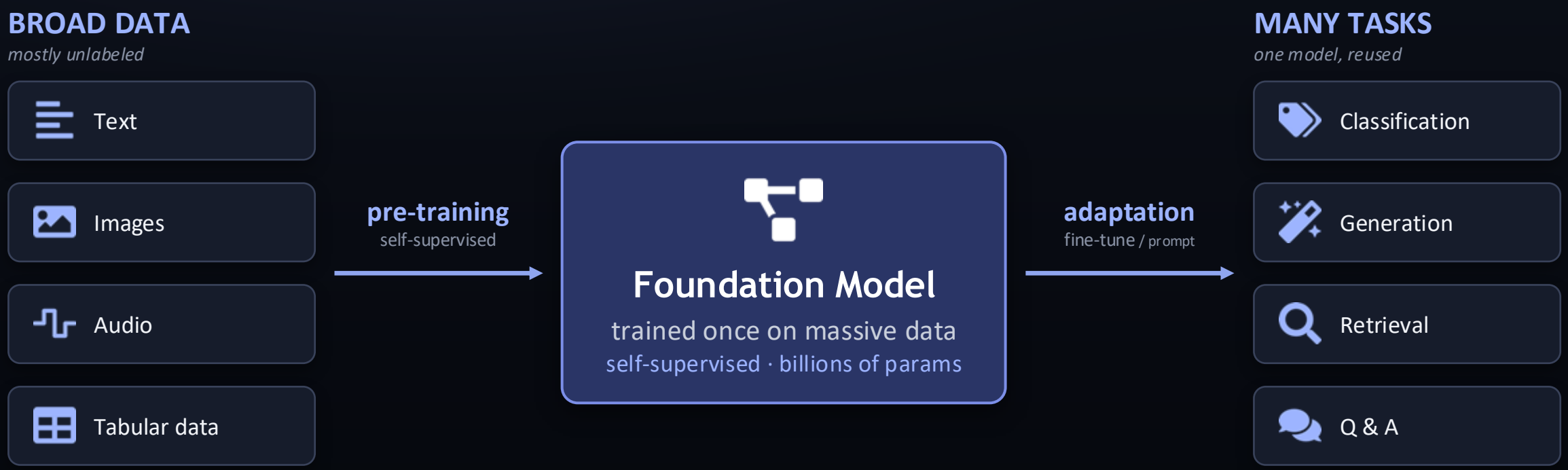
PART 2 · FOUNDATION MODELS



*Now we move to the modern era — **transformer foundation models** and the **AstroPT tokenizer benchmark**.*

What is a foundation model?

A single large model **pretrained on broad, mostly unlabeled data** — then **adapted to many downstream tasks**.
They learn by compressing data into a meaningful representation



Term coined by Stanford's Center for Research on Foundation Models — *Bommasani et al. (2021).*

A Familiar Starting Point: ChatGPT

The architecture powering today's foundation models is the **transformer** — and **ChatGPT** is its most familiar example.

01

Sequential structure

Learns how each element in a sequence relates to all the others.

02

Context awareness

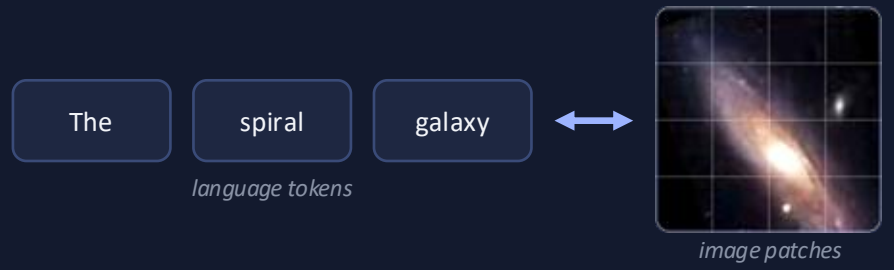
Reads every token from the tokens around it — never in isolation.

03

Generative prediction

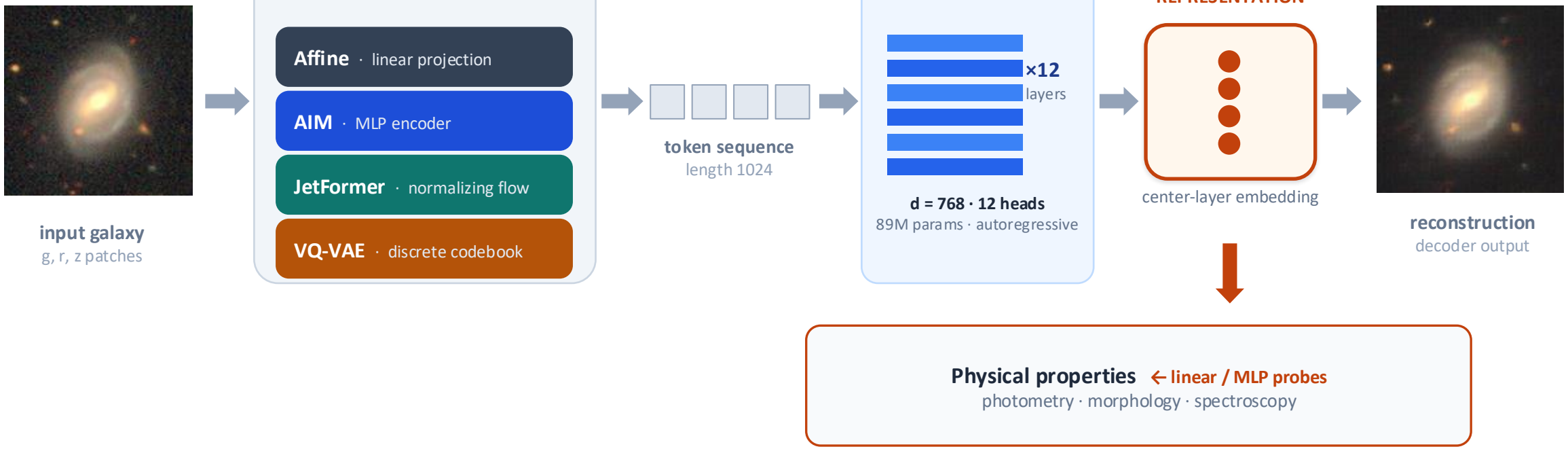
Builds its output one token at a time, each conditioned on the last.

*Swap **words** for **image patches**, and the same machinery learns the structure of **galaxies**.*



Our Foundation-Model Architecture

Four interchangeable tokenizers feed one shared AstroPT transformer backbone.



Only the tokenizer changes — the 12-layer backbone is held fixed across all four methods, isolating tokenization's effect.

The Galaxy's Guide to the Tokenizer

A Benchmark for Scientific Foundation Models

The question

Does tokenization — the way images become sequences — shape what a foundation model learns, and how that knowledge is organized?

The approach

- Four tokenizers compared inside one shared AstroPT backbone.
- Evaluated on physical-property prediction and image reconstruction.
- Grounded in independently measured physics — redshift, stellar mass, star-formation rate.

Four Tokenization Strategies

Affine

Linear projection of image patches to and from the embedding space — the simplest possible baseline.

linear · continuous · deterministic

AIM

An MLP encodes and decodes each patch; trained end-to-end alongside the backbone, no pre-training.

non-linear · continuous · deterministic

JetFormer

An invertible normalizing flow maps pixels ↔ latents, preserving all input information.

flow-based · continuous · probabilistic

VQ-VAE

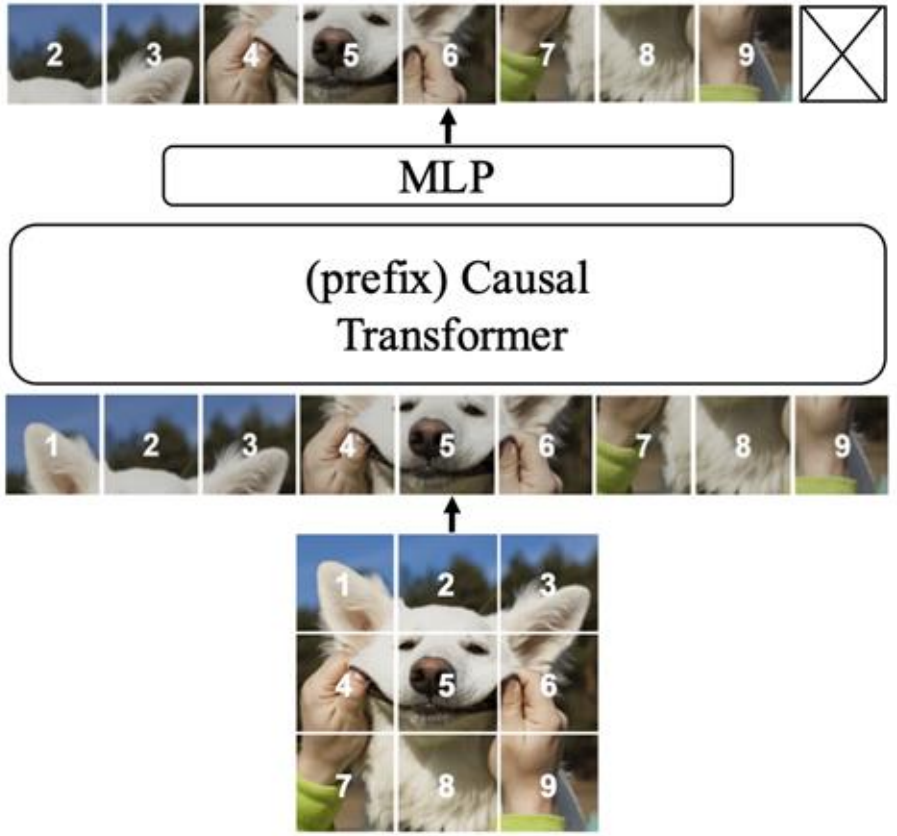
Patches are quantized to a learned codebook of visual tokens; pre-trained, then frozen during training.

non-linear · discrete · bottleneck

Spanning the design space: linear ↔ non-linear · continuous ↔ discrete · deterministic ↔ probabilistic

AIM:

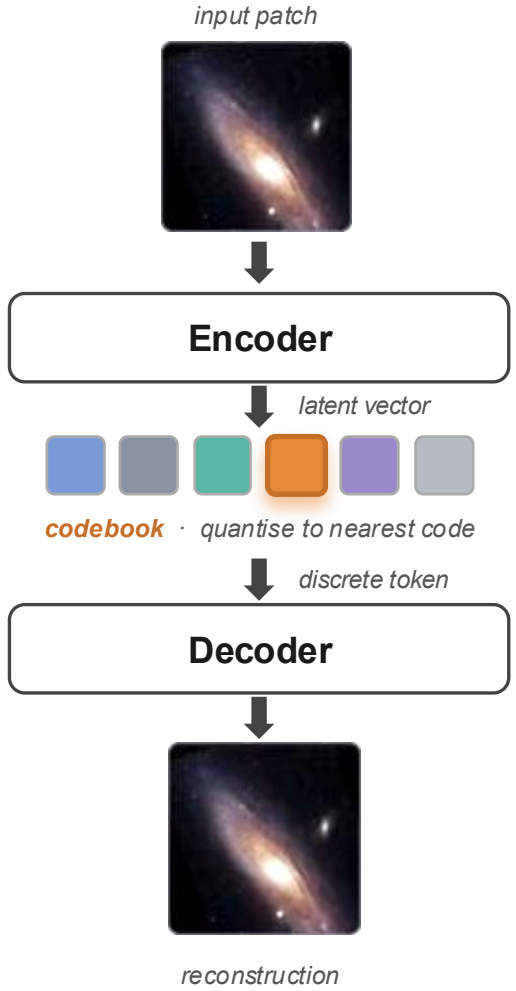
- This is the tokeniser that is already included in AstroPT.
- Consists of two MLPs for each modality; one encoding from data to latent space, and one decoding from latent to pixel space.
- It is flexible; does not need to be pre-trained and can be trained end-to-end alongside the foundation model.



El-Nouby et al 2024

VQ-VAE:

- Learns a discrete latent space — each patch maps to the nearest entry in a learned codebook, a fixed 'visual vocabulary'.
- An encoder compresses each patch to a latent vector; that vector is quantised to its closest codebook entry, and a decoder rebuilds the patch.
- Unlike AIM, the codebook is usually pre-trained first — the foundation model then learns over the resulting discrete tokens.

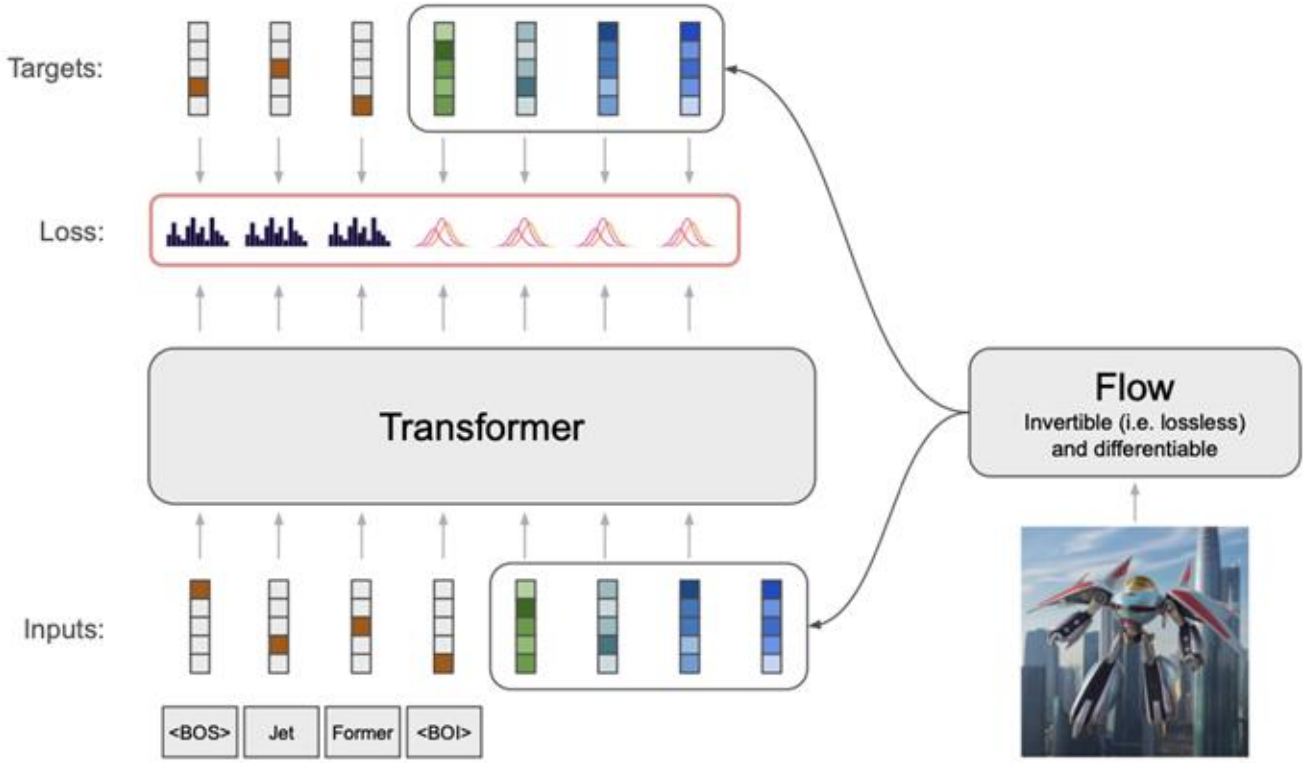


van den Oord et al. 2017

Jetformer:

- Introduced in deep learning; here applied to astronomical imaging for the first time.
- An invertible `**jet-flow**' (normalizing flow) maps back and forth between pixel space and latent space, losing no information.

- Trained alongside the foundation model with no **quantization bottleneck** — and in our benchmark it produces the **sharpest reconstructions**.



Tschannen et al 2025

Experimental Setup

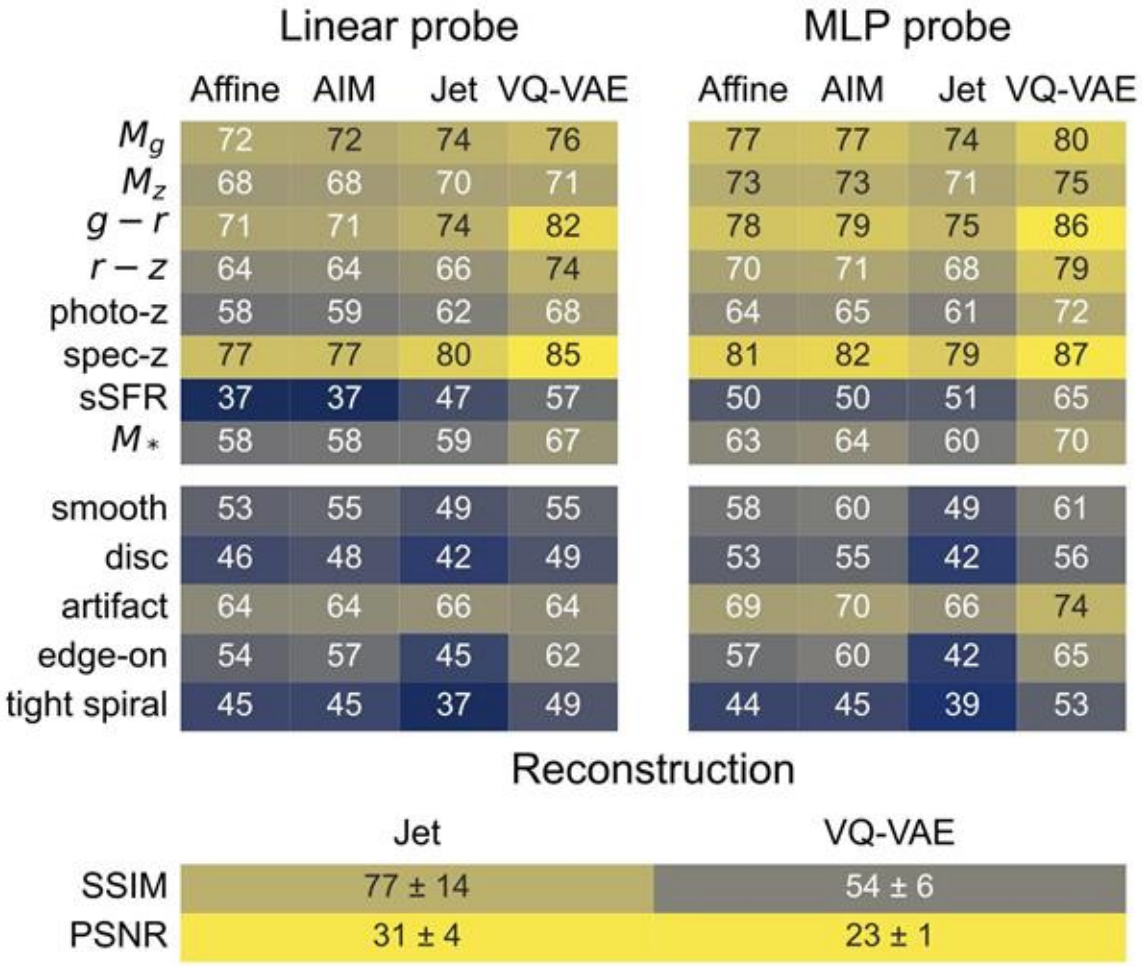
- **Data** 640,000 galaxies from the DESI Legacy Survey; 256×256pixel g, r, z postage stamps at 0.262”/pixel.
- **Shared backbone** one 89M-parameter AstroPT (12 layers, dim 768, 1024 tokens) — only the tokenizer changes between runs.
- **Physical probes** linear & MLP probes predict 13 measured properties on 167,000 held-out galaxies (k = 10-fold cross-validation).
- **Reconstruction** SSIM & PSNR measured on 5,000 held-out galaxies; independently measured physics gives objective ground truth.

640,000
training galaxies

89M
parameter backbone

13
measured properties

Physical Knowledge in the Latent Space



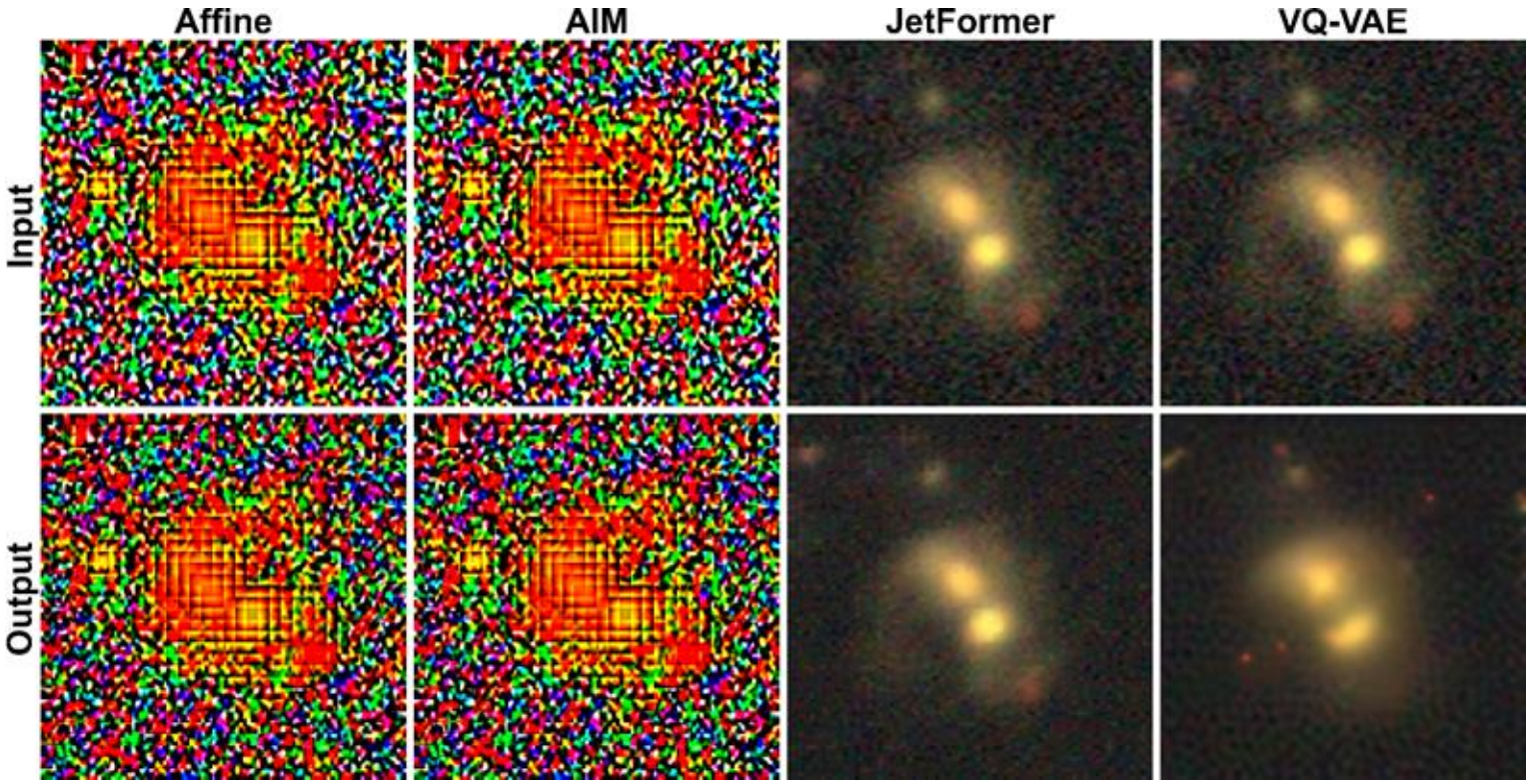
Linear & MLP probe R^2 (×100 in figure). Brighter = higher R^2.

What the probes reveal

- **VQ-VAE wins.** Strongest probe R^2 for most properties; its discrete bottleneck organizes physical information more linearly.
- **JetFormer** leads on colour and spectral quantities — magnitudes, colours, redshifts, star-formation rate, stellar mass.
- **AIM ≈ Affine.** Both edge ahead on morphology; the MLP head adds little over a simple linear projection.

Reconstruction quality does not predict any of this.

Image Reconstruction



Top row: input galaxies. Bottom row: reconstructions. AIM & Affine are shown in their patch-normalized representation.

Reconstruction fidelity

- **JetFormer:** sharpest reconstructions — best PSNR (31.1 dB) and SSIM (0.76); recovers spiral arms, cores, and faint emission.
- **VQ-VAE:** lower fidelity (PSNR 23.6 dB, SSIM 0.54); plausible but with cloudy artifacts around bright cores.
- **AIM / Affine:** patch-wise z-score normalization gives a block-like appearance but stabilizes training.

Best reconstruction ≠ most informative embedding.

Reconstruction ≠ Representation

The best image reconstruction and the most useful representation come from ***different*** tokenizers.

BEST RECONSTRUCTION

JetFormer

PSNR 31.1 dB · SSIM 0.76

Sharp images, faithful pixels.
Information is encoded but hard to access

≠

MOST INFORMATIVE EMBEDDING

VQ-VAE

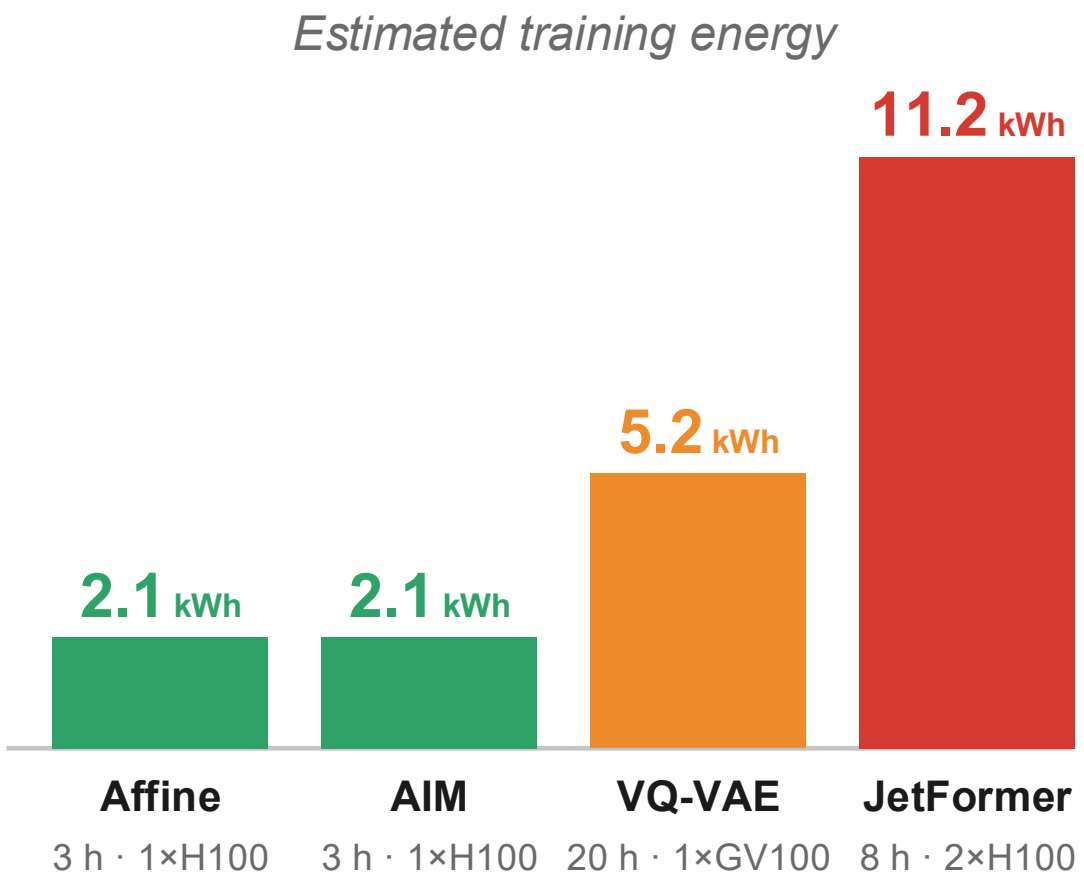
Highest probe R^2 for physical properties

Compressed, semantically organized.
Information is not fully encoded

Choose your tokenizer by the downstream task — reconstruction quality alone does not predict representation quality.

Compute & Carbon:

- The two simplest tokenisers — Affine and AIM — are also the greenest, using about 5× less energy than JetFormer.
- Wall-clock time ≠ energy: VQ-VAE trained longest (20 h) yet drew under half of JetFormer's power on its older, lower-TDP GPU.



Conclusions

1. **Tokenization is not neutral.** It shapes both what a foundation model learns and how that information is organized in the representation.

2. **Reconstruction ≠ representation.** JetFormer reconstructs best, encodes all the information, yet VQ-VAE yields the most informative embeddings — the two are decoupled.

3. **Choose by the downstream task.** VQ-VAE for physical inference; JetFormer for generation and reconstruction; Affine when compute is limited.

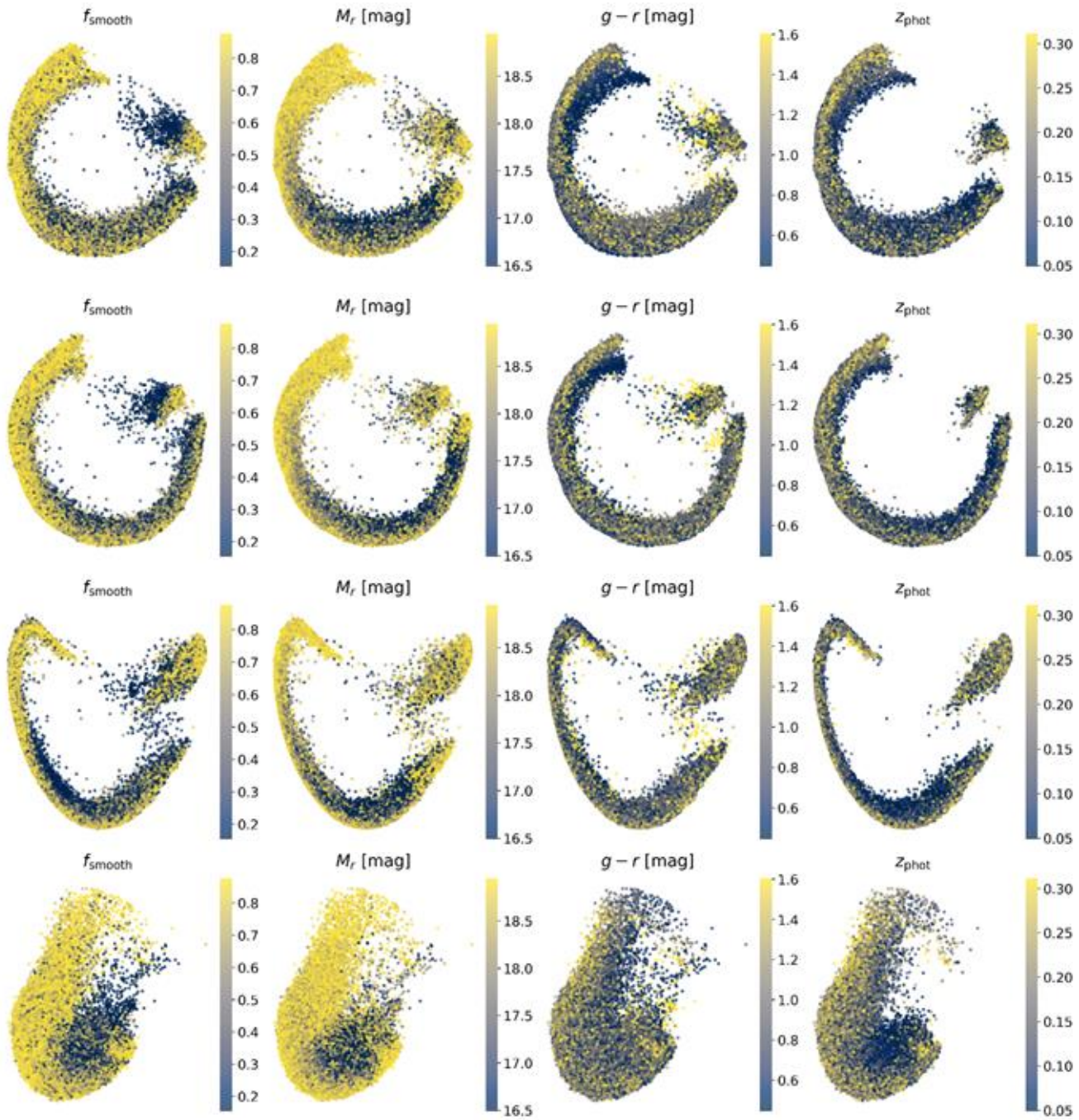
4. **Astronomy as a benchmark.** Independently measured physics gives objective ground truth — a path toward a physics-grounded “tokenization bench.”

Thank you

Questions and discussion welcome.

Sogol Sanjaripour · University of California, Riverside
. AstroAI, Harvard University

Selected work The Galaxy's Guide to the Tokenizer (2026)



- Each row shows one tokenizer's learned latent embeddings, projected to two dimensions with PCA.
- Within a row, the four columns color that same embedding by a different galaxy property — left to right: smoothness, absolute magnitude, color, and photometric redshift.
- Tokenizers are labelled at right.