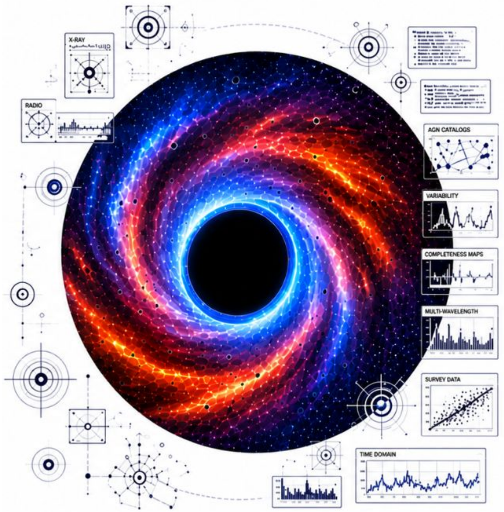
MACHINE LEARNING FOR ASTROPHYSICS
3RD EDITION VALLETTA, 31 AUG - 4 SEP, 2026

Deep Learning the Physics of Galaxy Formation

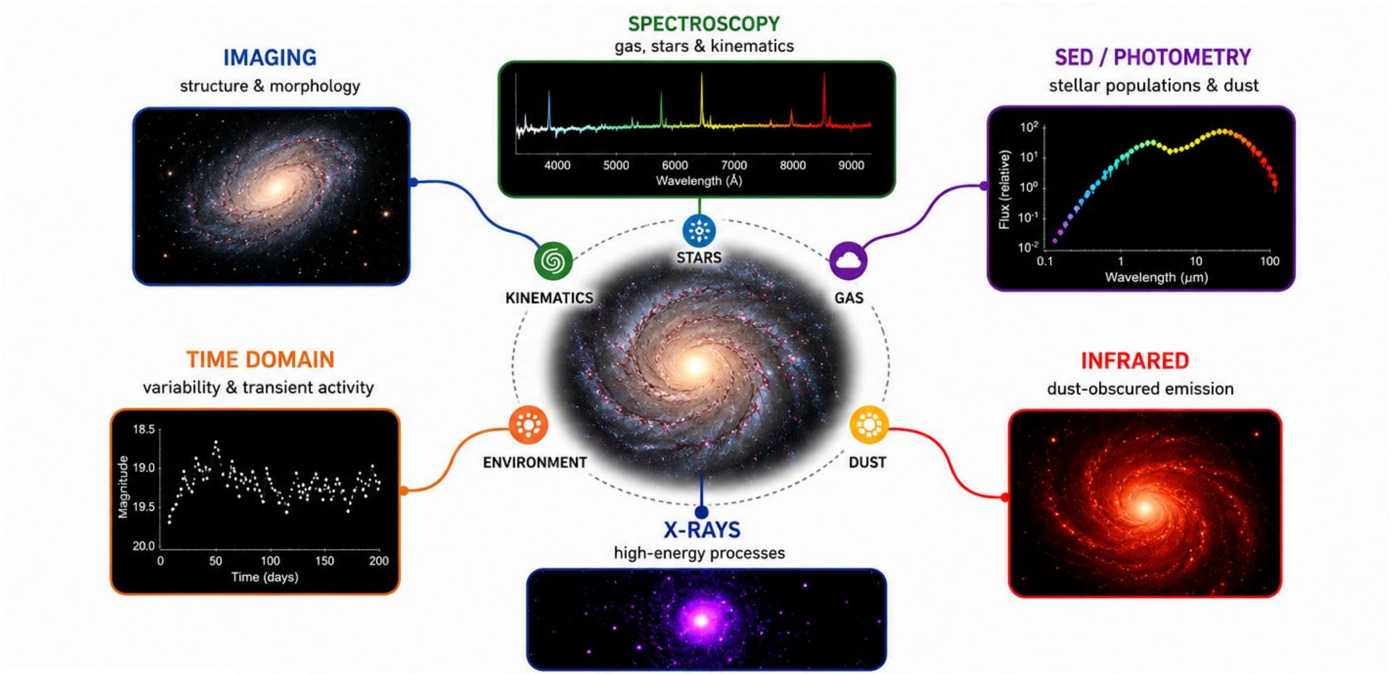
Using Large Observational Surveys

Małgorzata (Gosia) Siudek
Instituto de Astrofísica de Canarias



One galaxy – many views of the same galaxy

Different observables probe different aspects of the same underlying physics.

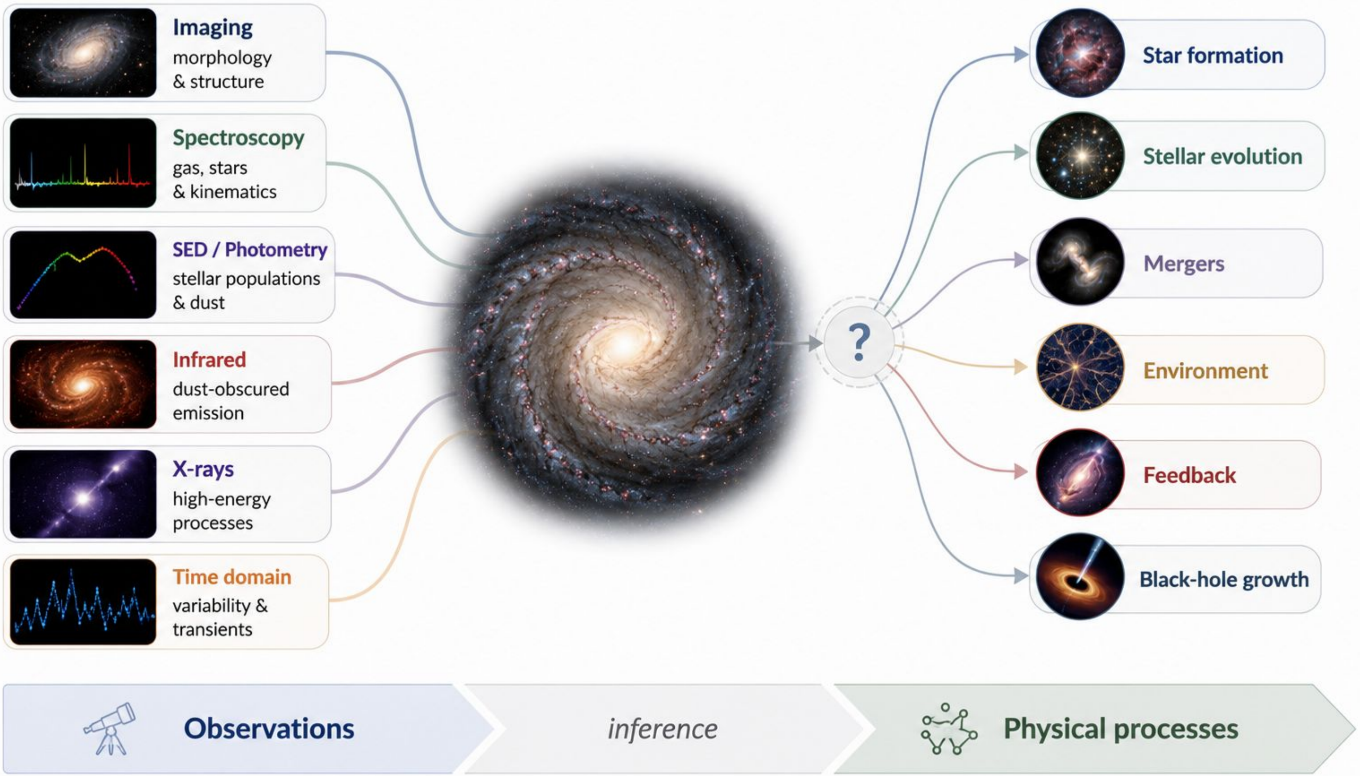


The physics is shared, but the information is distributed

ML4ASTRO 2026 3rd edition: Deep Learning for Galaxy Formation: M. Siudek msiudek@iac.es

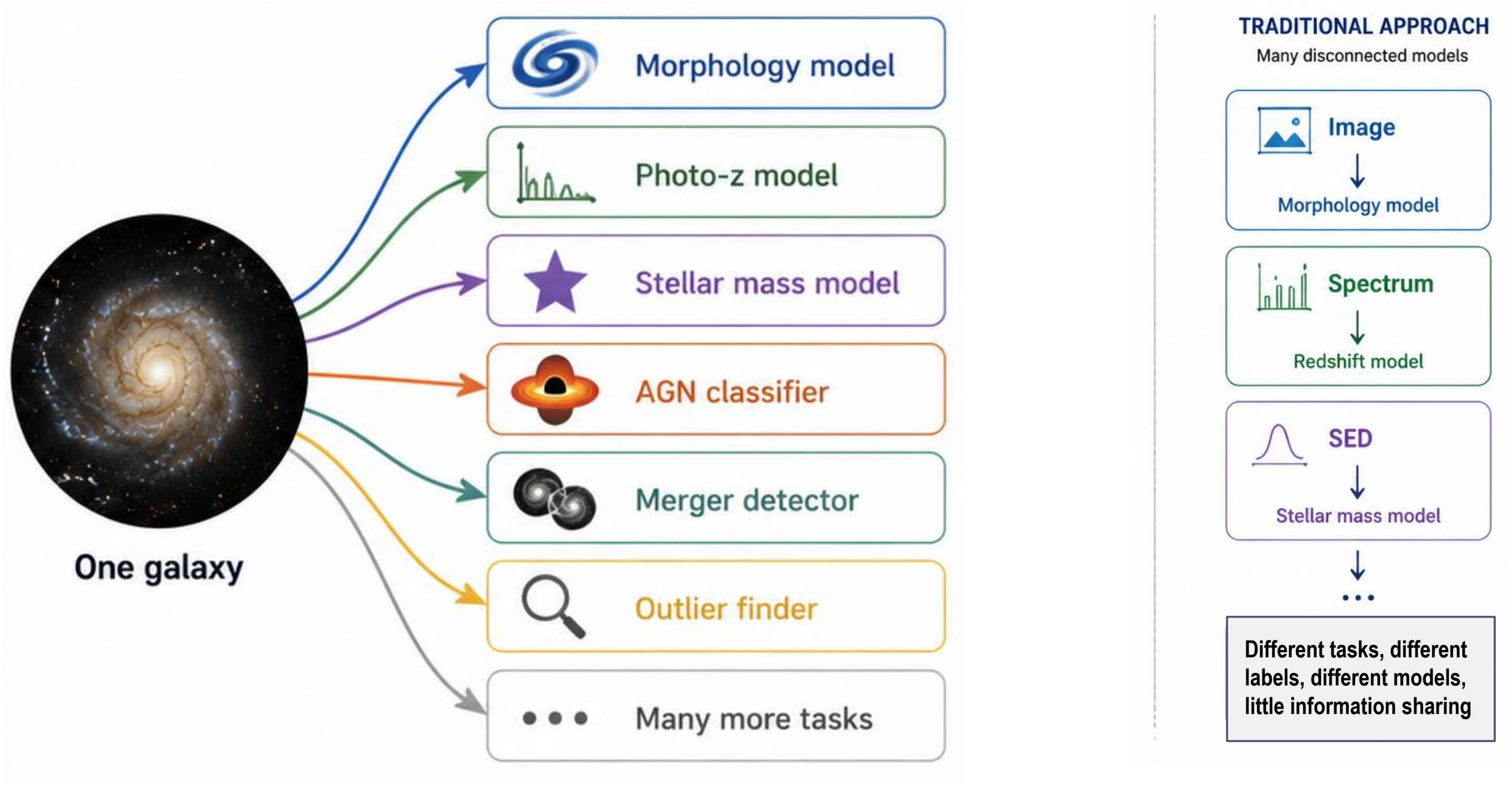
But what we really want is the physics

The challenge is connecting what we observe to the physical processes that shape galaxies.



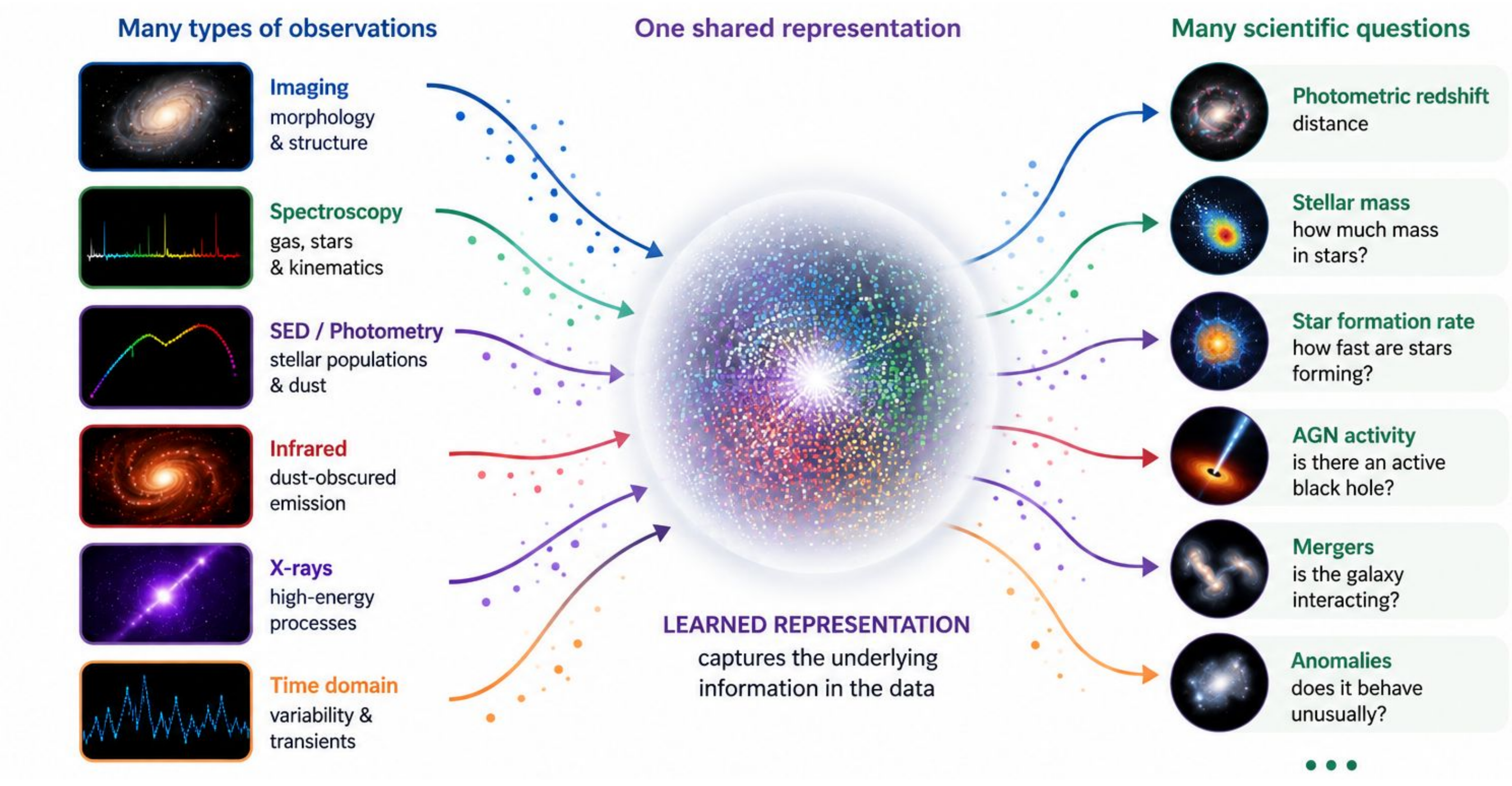
The challenge is not just obtaining data – it is connecting the observables back to the underlying physics

Traditional approach: information is fragmented



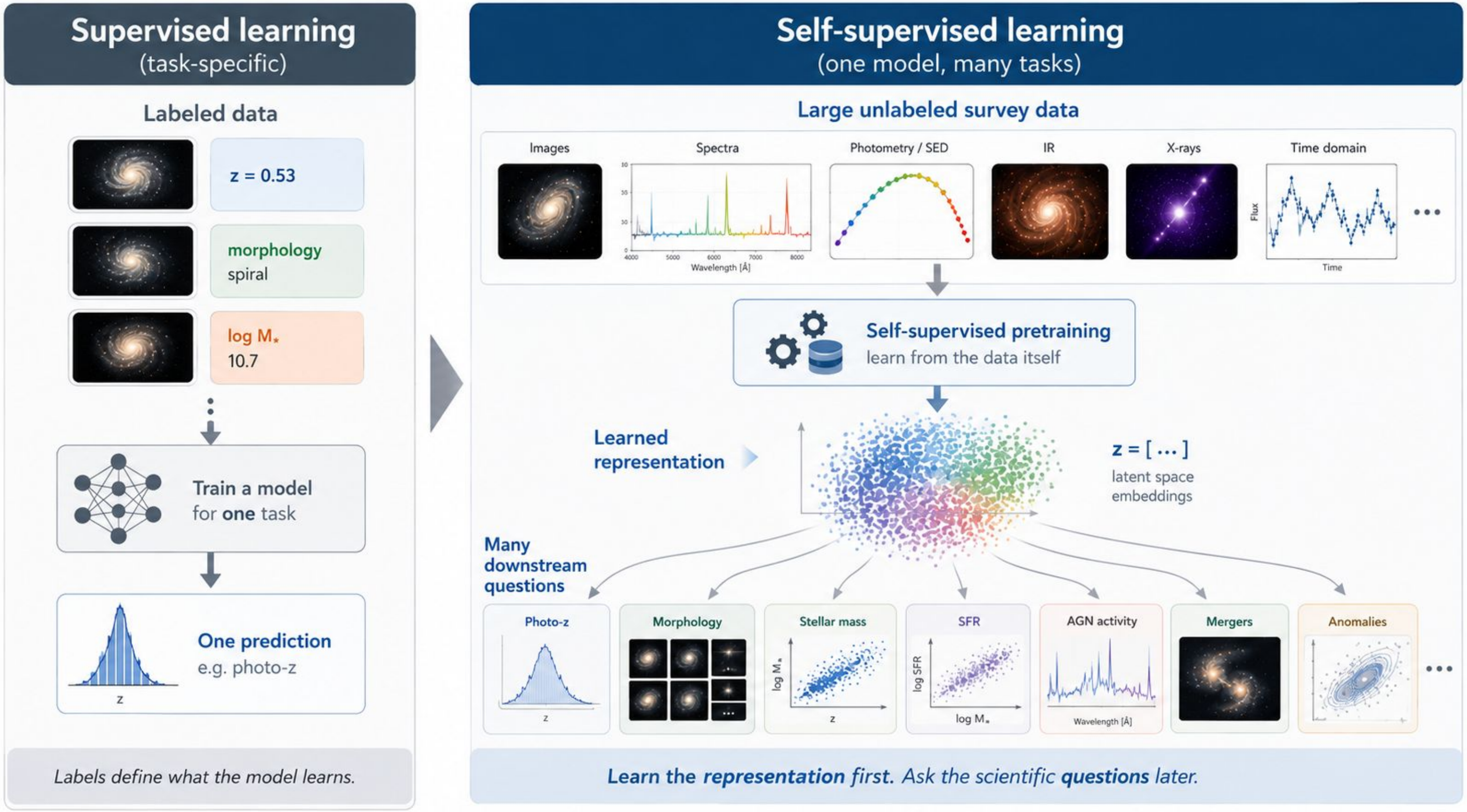
One galaxy. Many models. Limited information sharing

Can we learn **one** representation for **many** scientific questions?



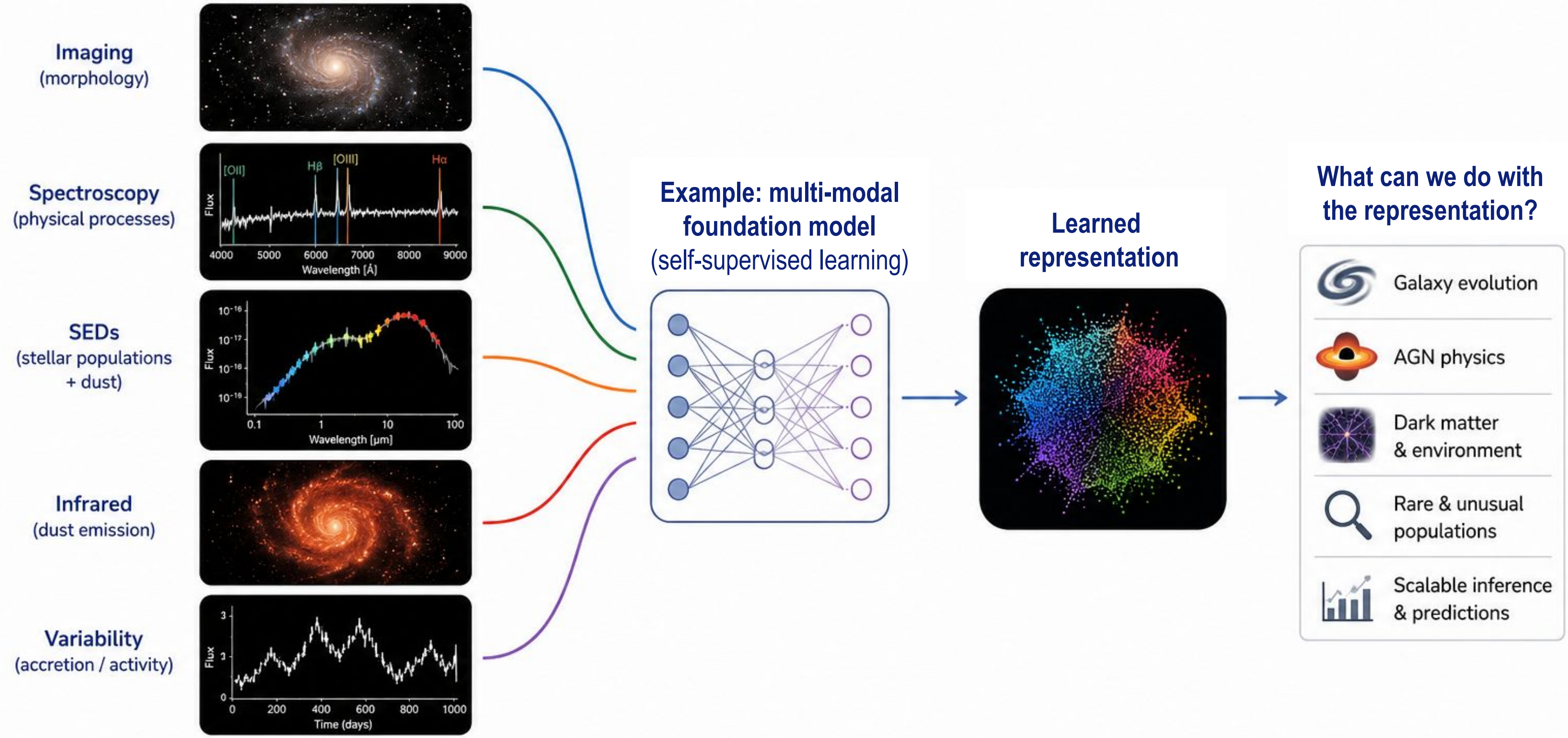
Learn once – reuse across scientific questions

The key shift: learn the representation before the labels



From learned representations to foundation models

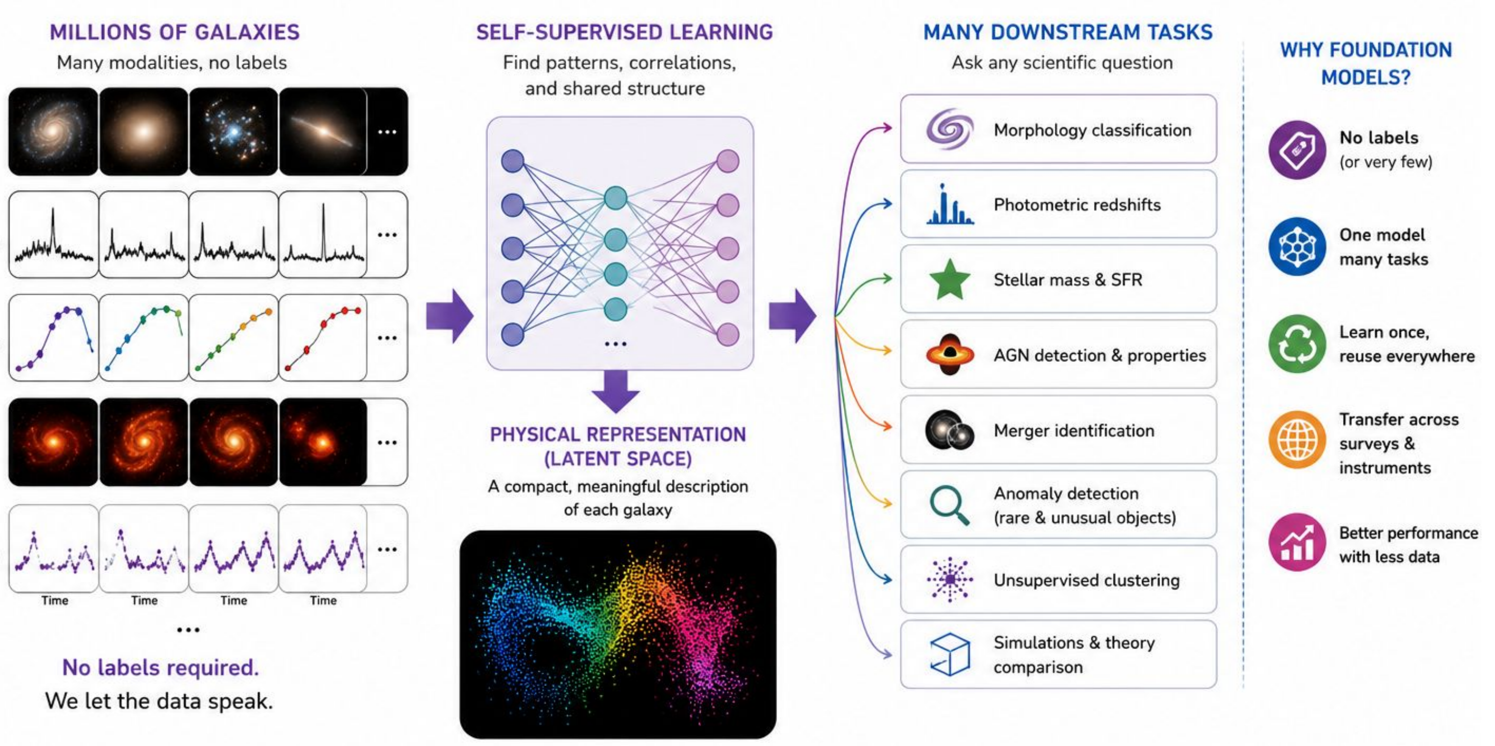
Pretrain once on large, diverse datasets – then adapt to many scientific tasks.



Foundation models themselves can become scientific objects

Foundation models learn galaxies, not labels

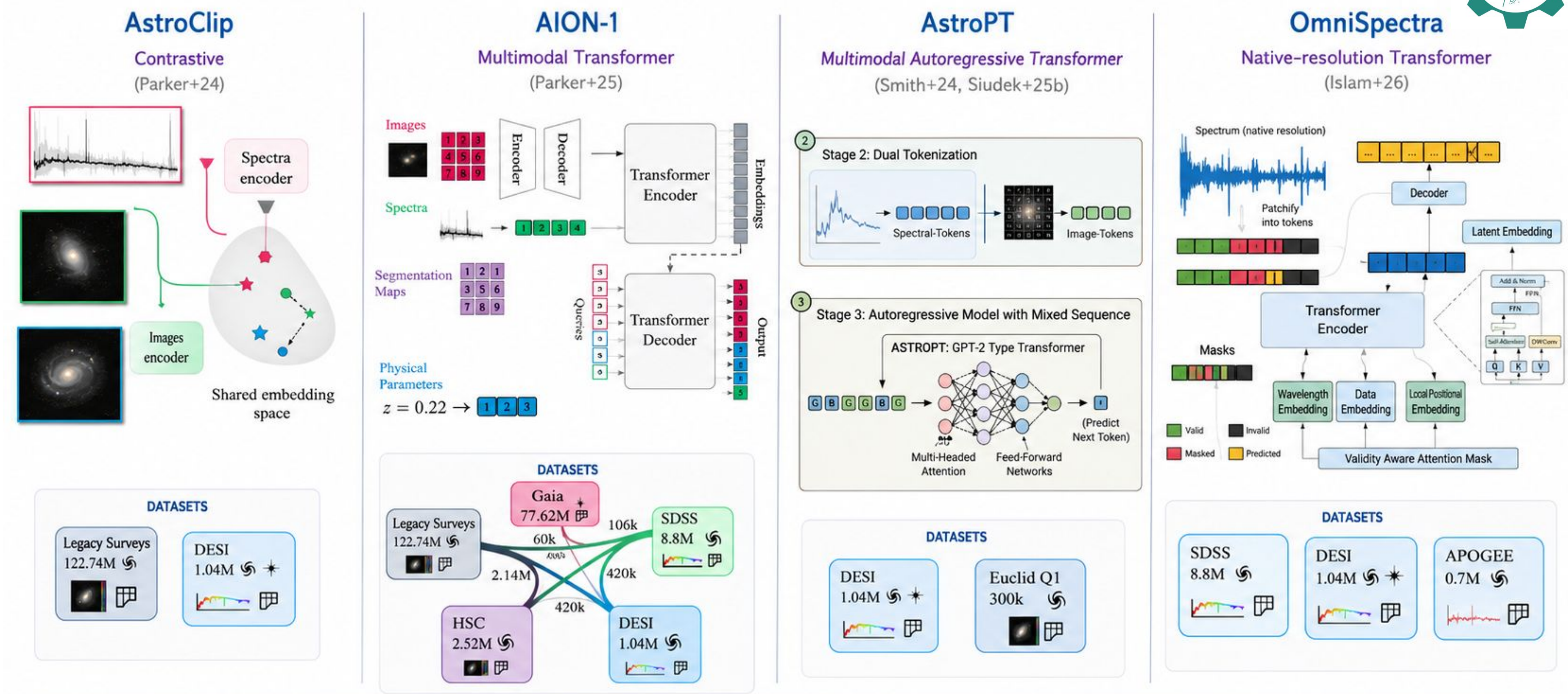
We train on data, not labels, to learn general representation of galaxies that can be reused for many scientific problems.



Learn once – reuse everywhere.

A foundation for the next generation of astronomical discovery

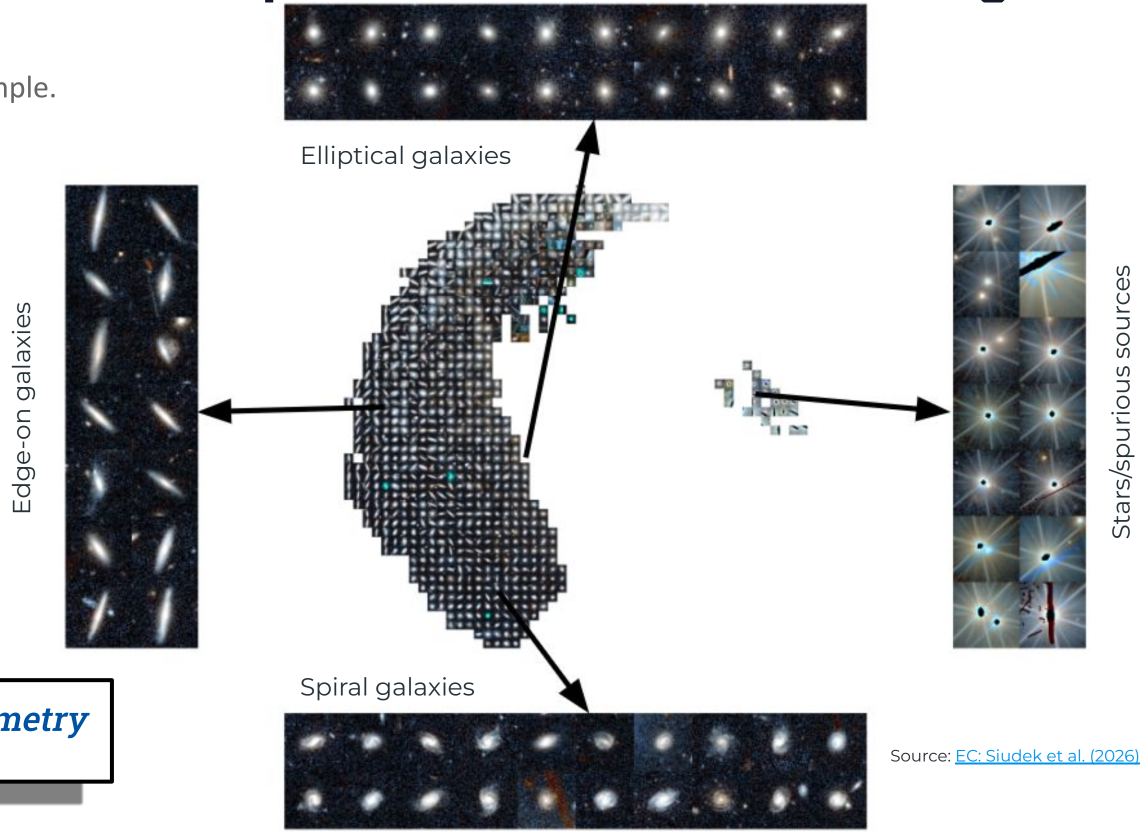
A new generation of astronomical foundation models



Astronomy is moving from task-specific networks to reusable foundation models

What does the representation know about galaxies?

AstroPT as an example.



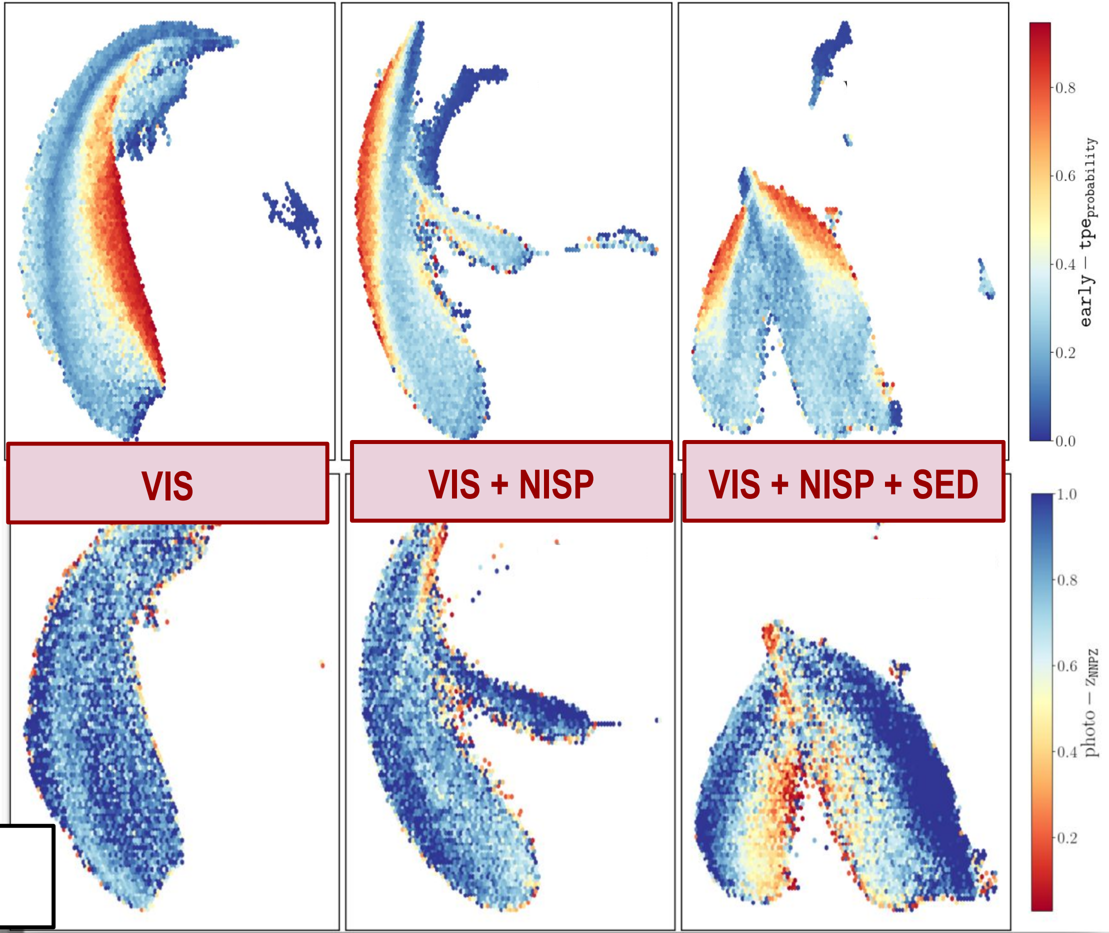
Trained on photometry (VIS)

Source: EC: Siudek et al. (2026)

A representation learned without physical labels can still organize galaxies by meaningful properties

What changes when we add another view of galaxy?

AstroPT as an example.

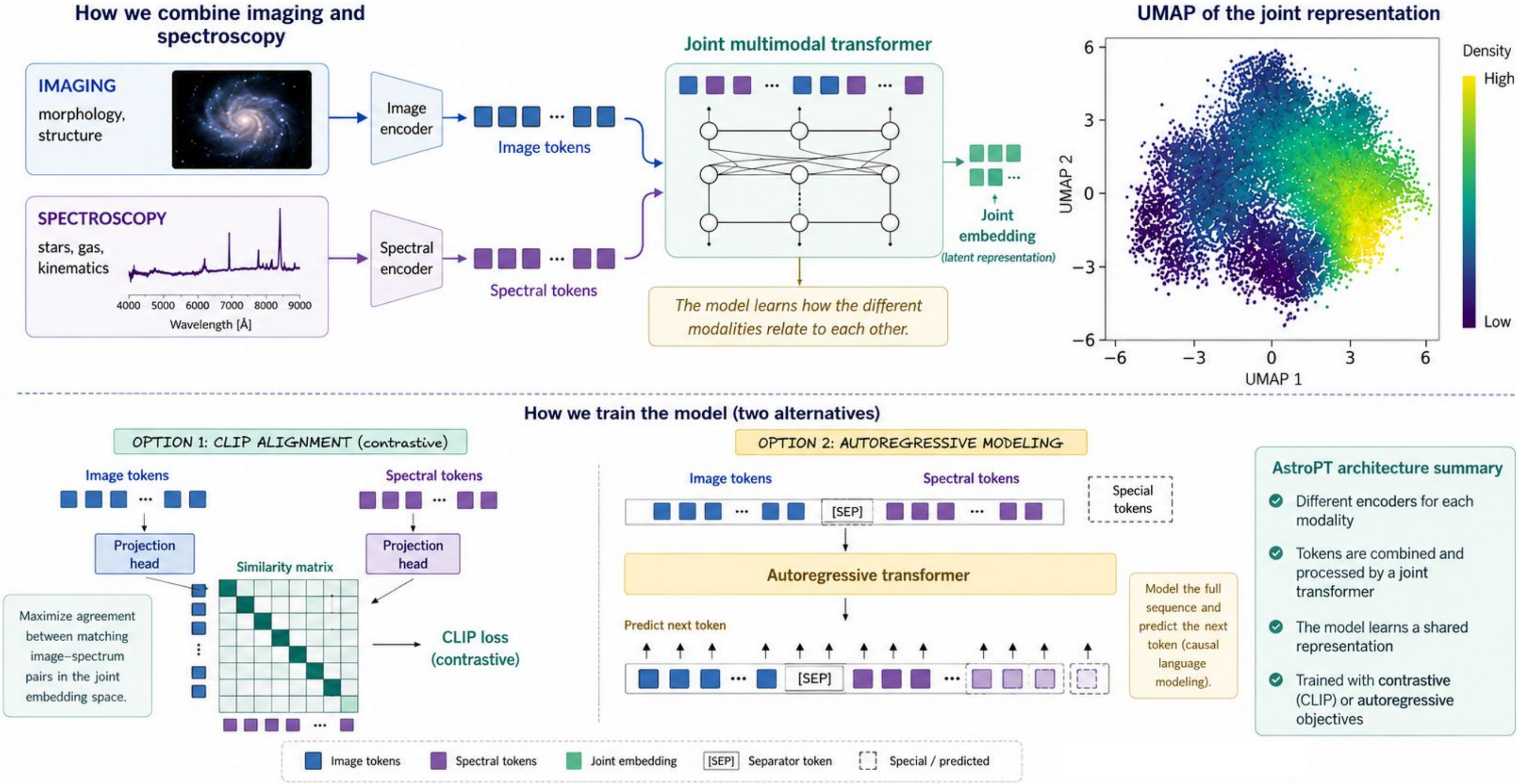


Trained on photometry

ML4ASTRO 2026 3rd edition: Deep Learning for Galaxy Formation: M. Siudek msiudek@iac.es

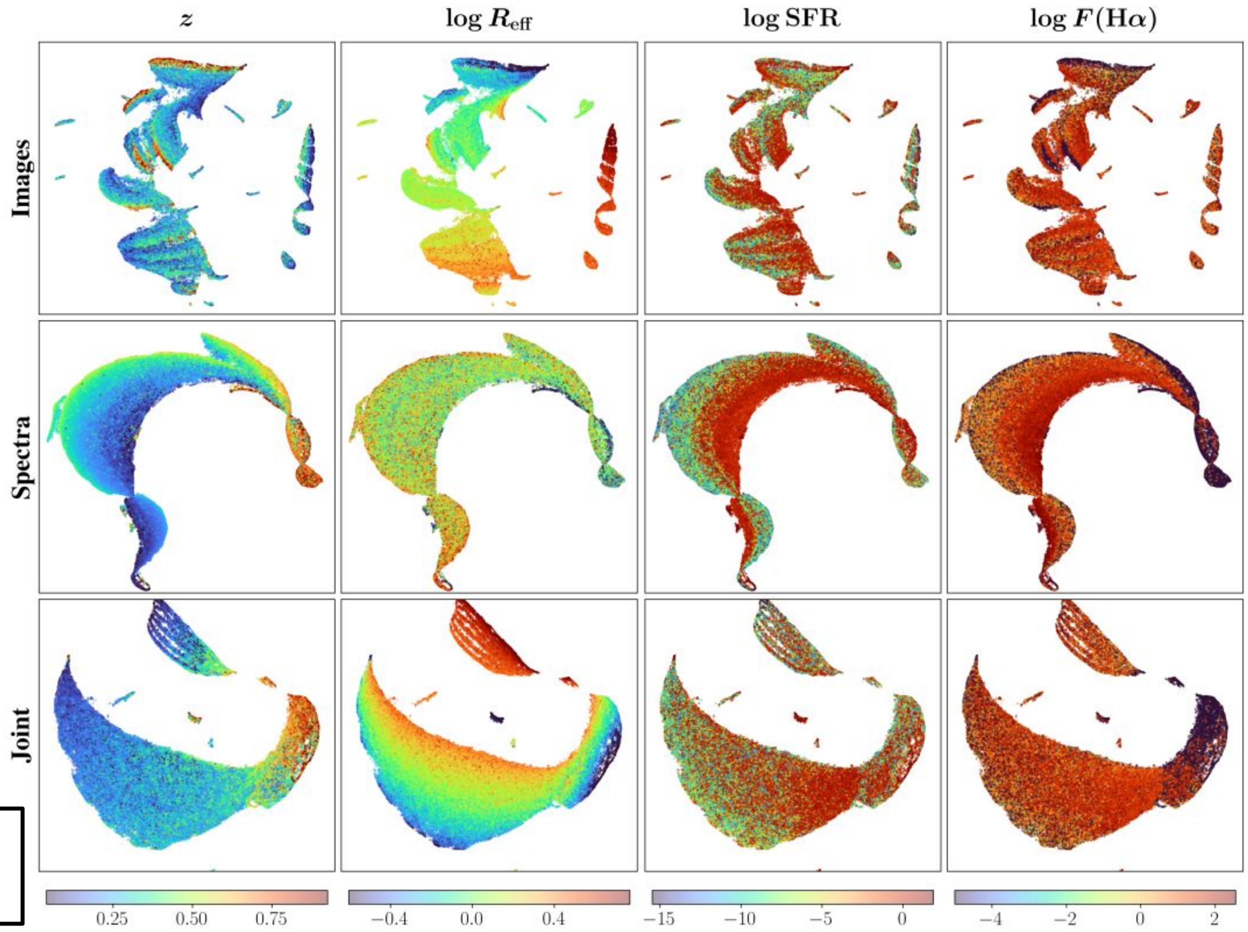
Multimodal foundation models – what if we add spectroscopy?

AstroPT trained on photometry and spectroscopy as an example.



Multimodal foundation models – what does each modality contribute?

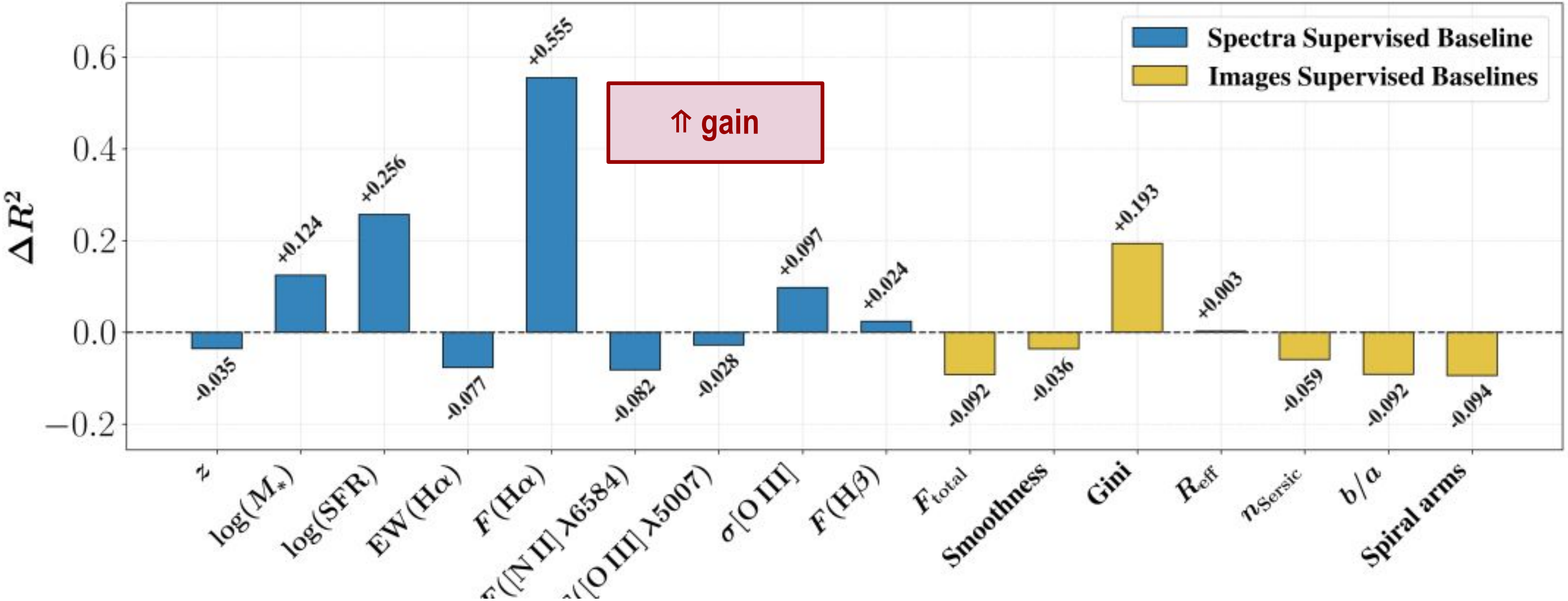
AstroPT as an example.



Different modalities encode different physical information

Multimodal foundation models – what we can gain?

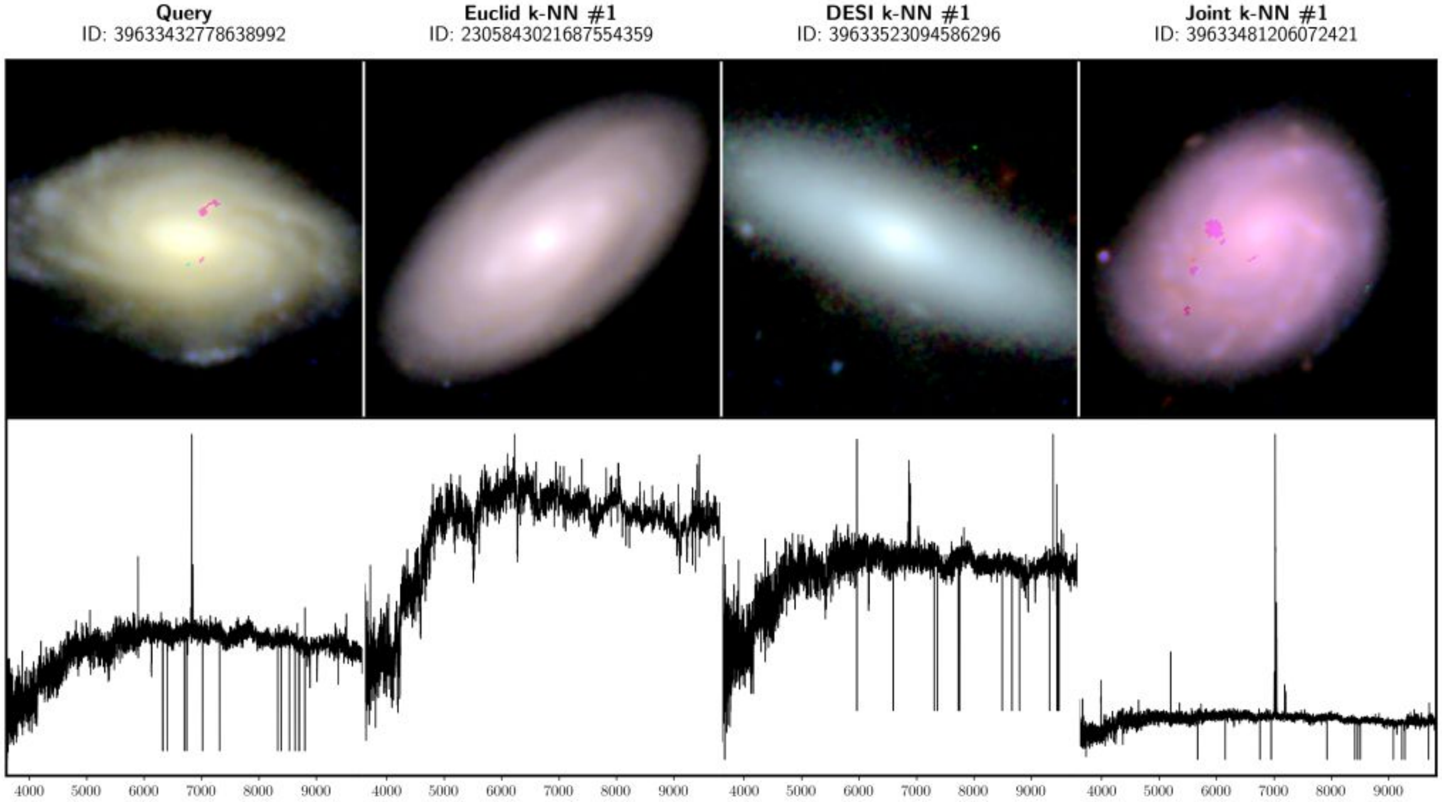
AstroPT as an example.



Trained on photometry + spectroscopy

Can we use this representation to search?

AstroPT as an example.



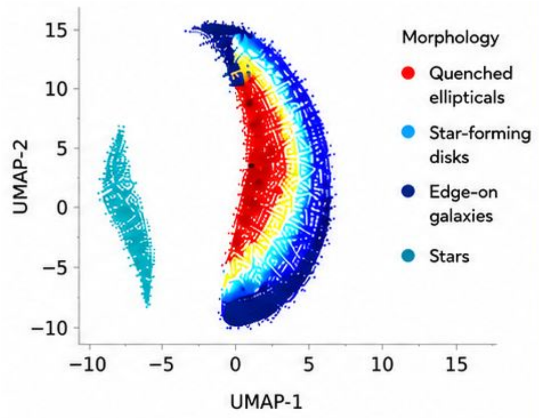
Joint similarity captures both imaging and spectral properties

ML4ASTRO 2026 3rd edition: Deep Learning for Galaxy Formation: M. Siudek msiudek@iac.es

Does multimodality can reveal something new?

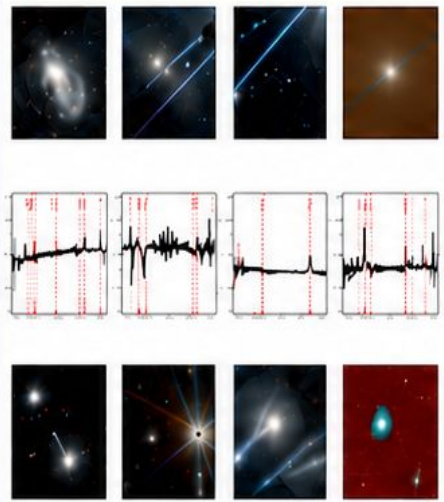
AstroPT as an example.

1 PHYSICAL CORRELATIONS



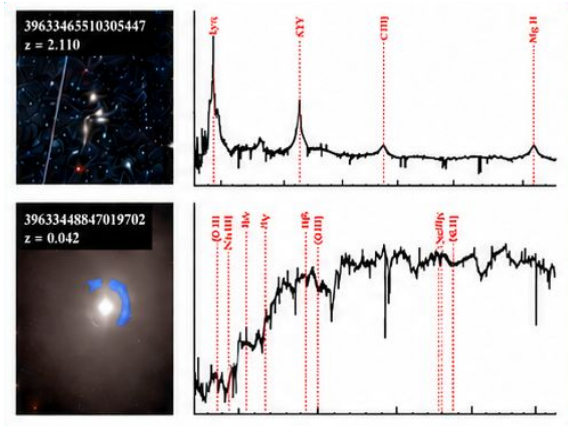
Learns correlations between galaxies

2 SINGLE-MODALITY ANOMALIES



Bad cutouts, cosmic rays, weird spectra

3 MULTI-MODALITY ANOMALIES



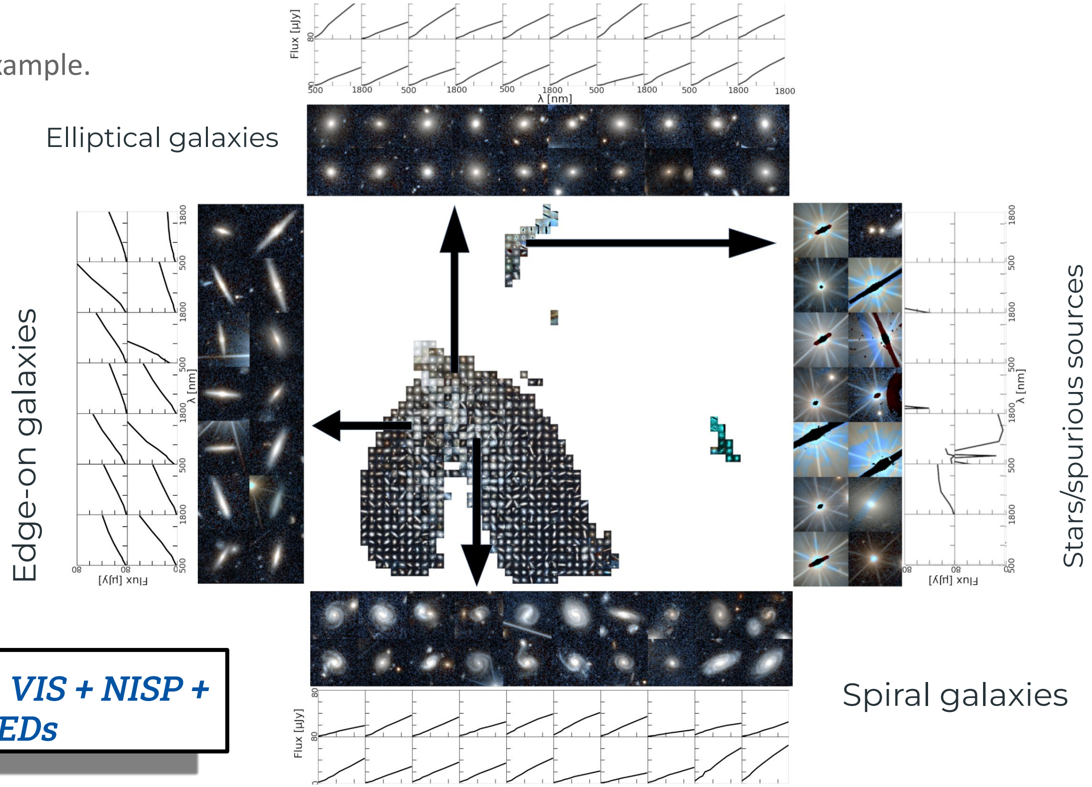
Real rare sources - QSOs, strong gravitational lens

- images + SEDs/spectra,
- ~300k sources (Siudek+25a, Ronceray+26),
- improves photometric redshift, physical property estimates (1% of labels),
- identify rare sources.

Multimodal disagreement/similarity can expose rare physical sources

But what exactly does the model learn?

AstroPT as an example.



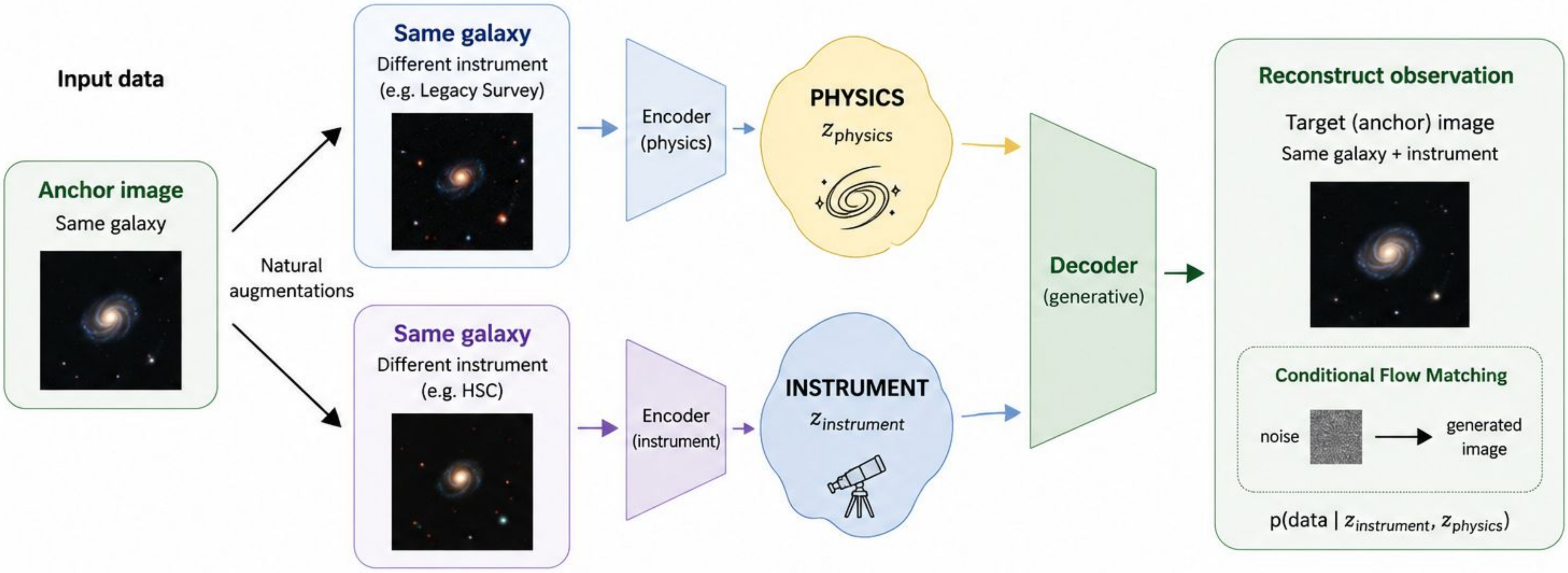
Trained on VIS + NISP + SEDs

Source: EC: Siudek et al. (2026)

The representation captures both astrophysical structure and patterns introduced by the observations

Can we separate astrophysics from the instrument?

Learn what belongs to the galaxy and what belongs to the observation.

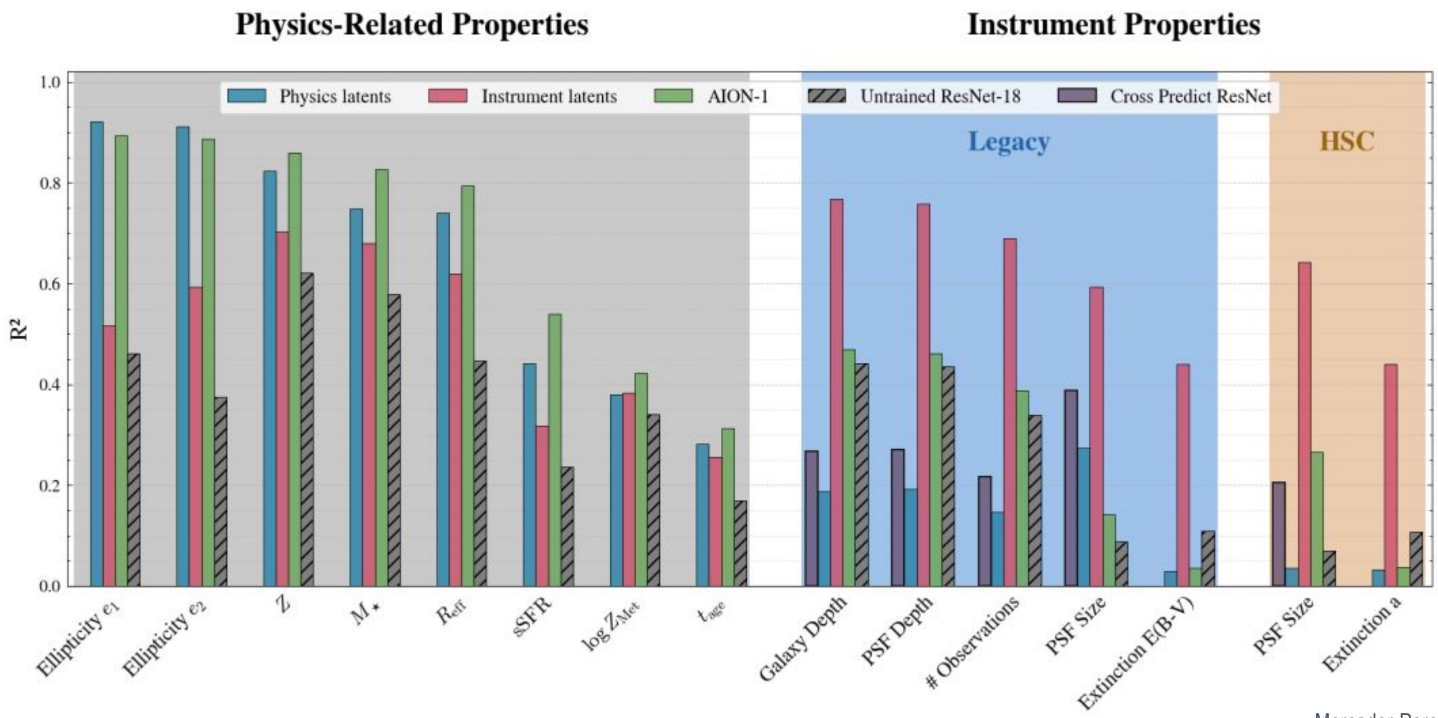


Mercader-Perez+26 (adapted)

We learn to separate intrinsic properties of galaxies from the effects of the instrument

Does the separation actually work?

The physics latent retains galaxy properties while suppressing instrument-specific information.

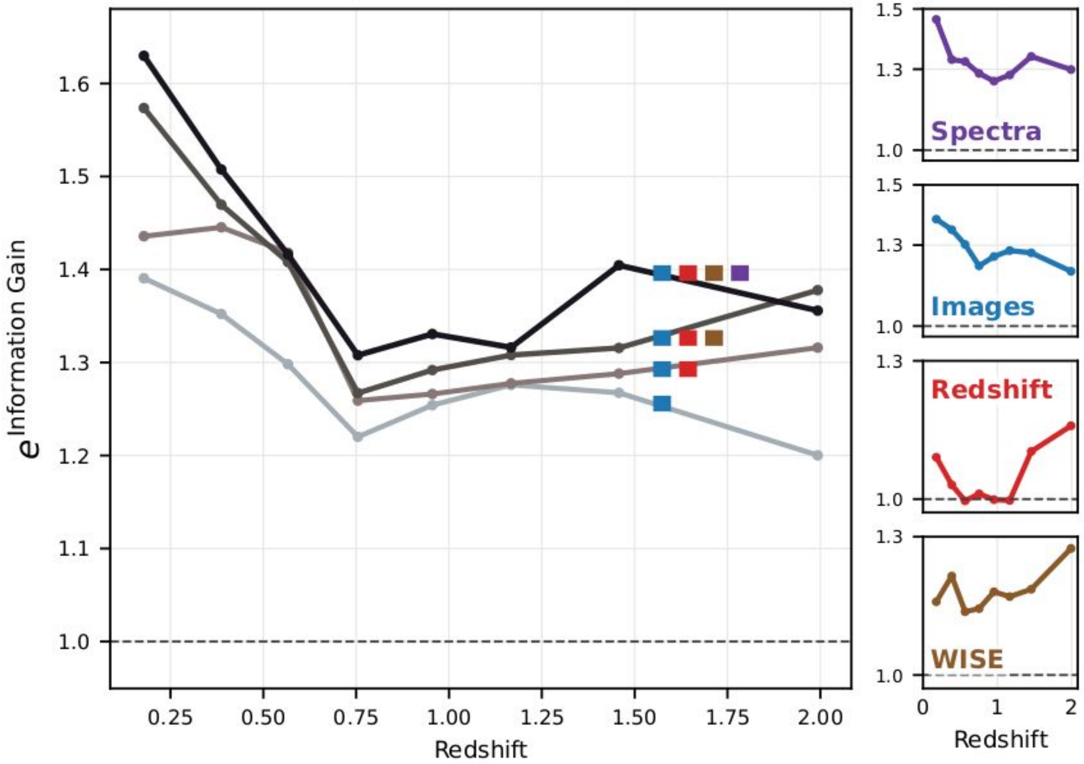


Mercader-Perez+26

The two latent spaces encode different kinds of information

Can we connect optical observations to X-ray emission?

Different wavelengths probe complementary aspects of galaxy and black-hole activity.

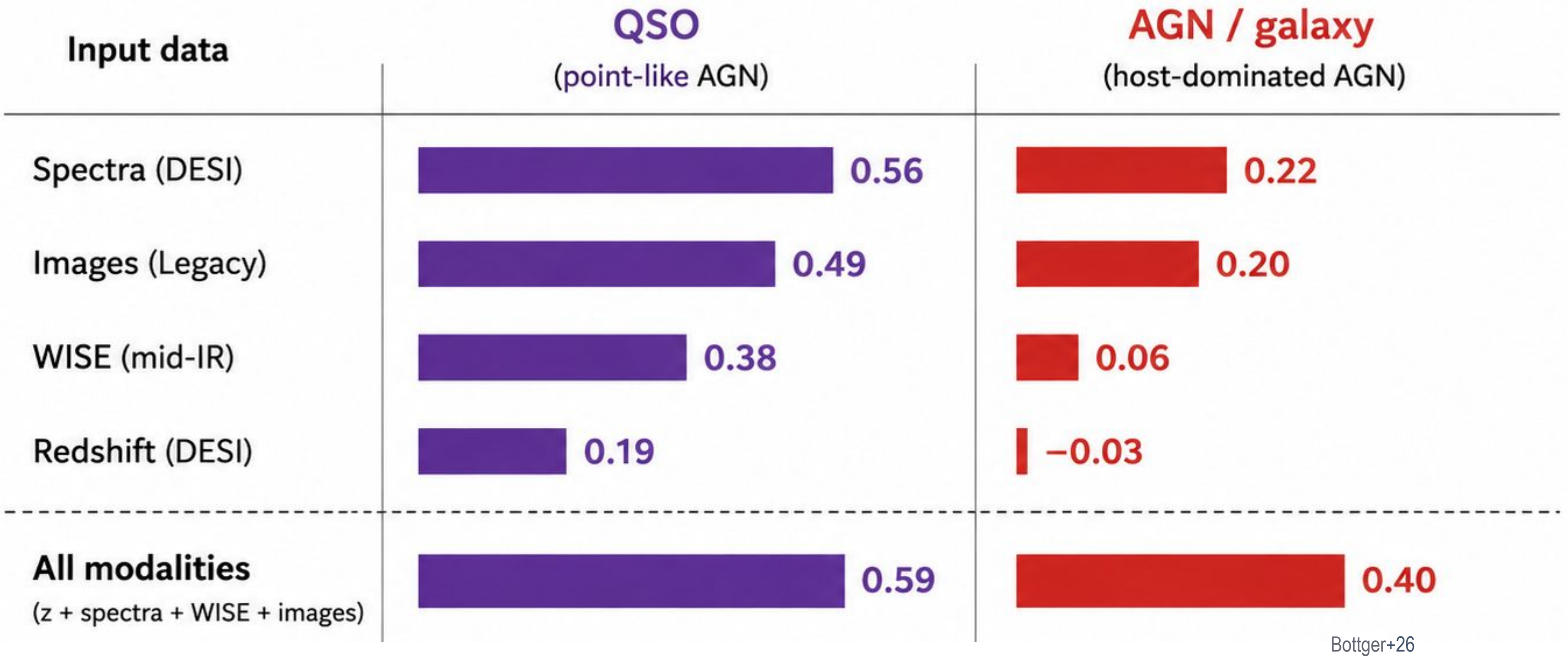


Bottger+26

The information is distributed across wavelengths — combining them gives the strongest constraints

How much of the X-ray emission can we recover?

Joint model nearly doubles the explained variance for AGN/galaxies — but most X-ray variance remains unexplained.



Bottger+26

The multimodal gain is largest for AGN/galaxies

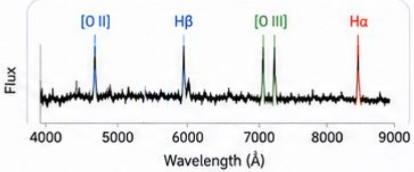
FROM LEARNING REPRESENTATIONS TO DISCOVERING NEW PHYSICS

We now turn to science: using these representations to understand **active black holes (AGN)** and **galaxy evolution.**

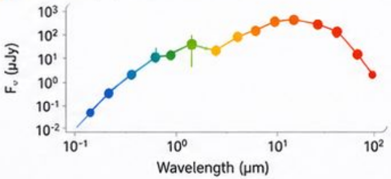
Each observation reveals a **different piece** of the same system.

IMAGING
structure & morphology

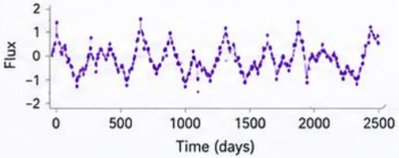
SPECTROSCOPY
kinematics & physical conditions



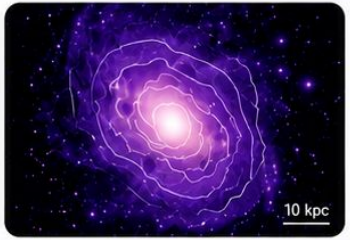
SEDs / PHOTOMETRY
stellar populations + dust



VARIABILITY
accretion & transient activity



X-RAYS
high-energy processes



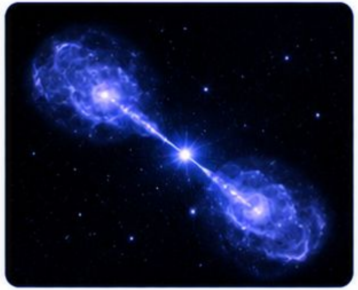
INFRARED
dust & obscured emission

MULTIPLE MESSENGERS
complementary views

Why conventional AGN selection is incomplete?

Different methods see different AGN.

X-RAY SELECTION



Sees energetic regions close to the black hole

+

OPTICAL SELECTION

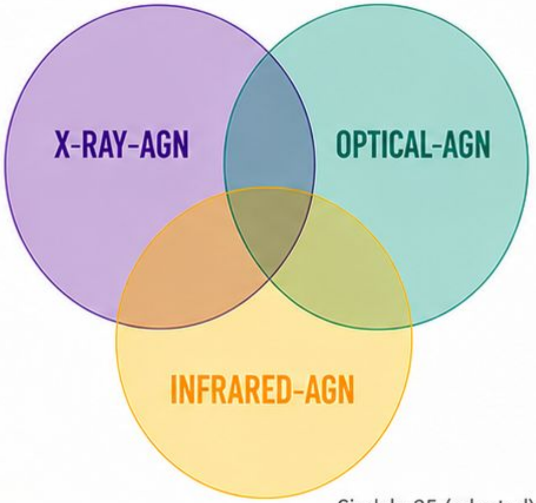
Sees stars and gas in the galaxy

+

INFRARED SELECTION

Sees dust-obscured active black holes

=



Siudek+25 (adapted)

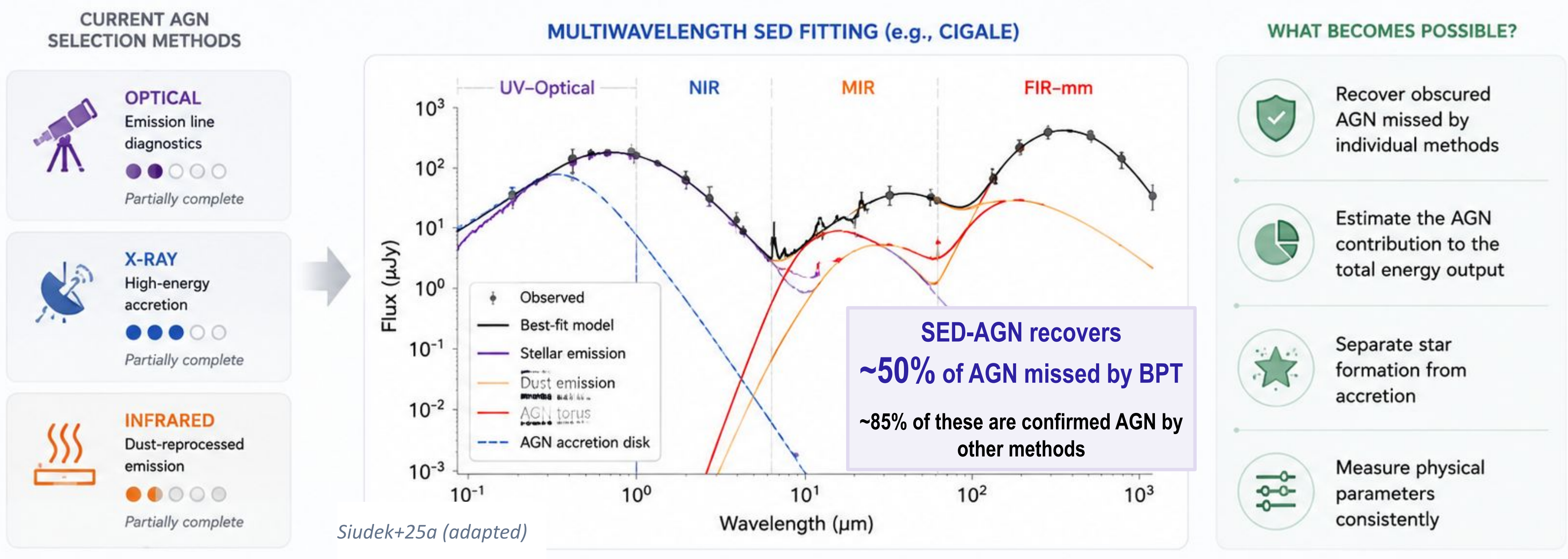
ACTIVE BLACK HOLE CENSUS:
Only **1 in 10** active black holes are identified by multiple methods.
(Siudek+25)

DWARF GALAXIES:
Theory predicts many active black holes (20%), observations find very few (<1%).
(Reines+13, Pacucci+21)

The AGN population depends on how we look at it

ML4ASTRO 2026 3rd edition: Deep Learning for Galaxy Formation: M. Siudek msiudek@iac.es

Multimodal information recovers hidden AGN

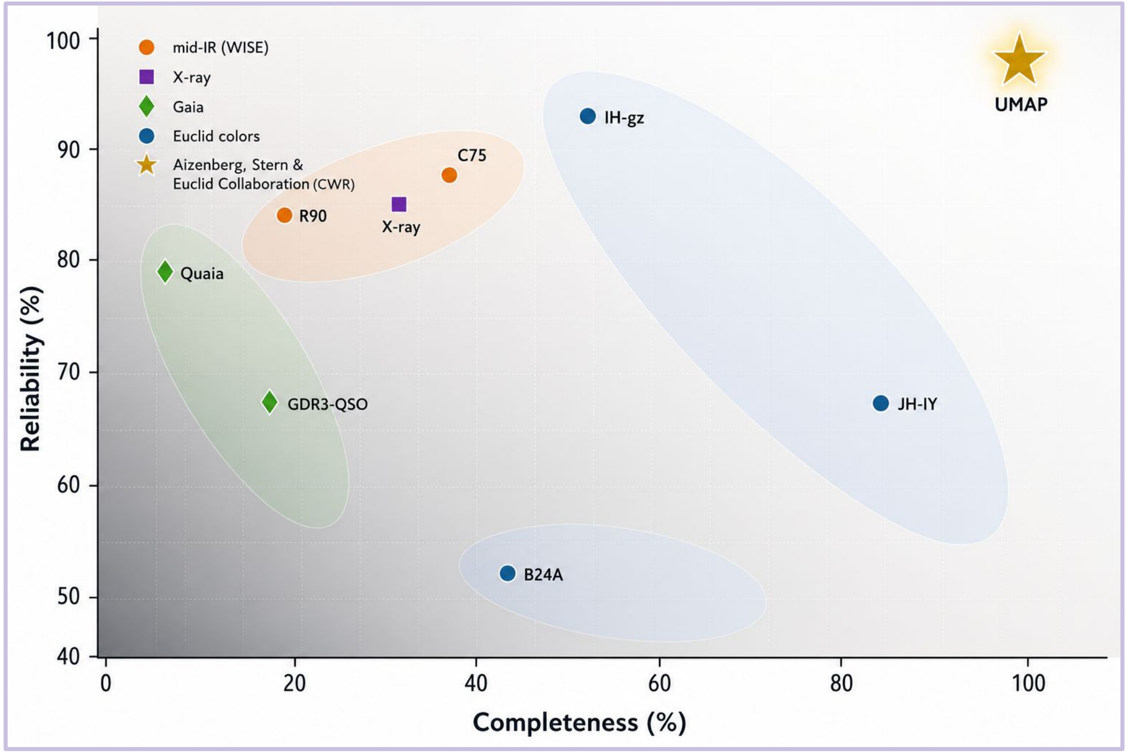


Siudek+25a (adapted)

Combining complementary observables reveals AGN missed by individual diagnostics

Scaling multimodal learning to millions of galaxies

Combining imaging and morphology enables population-scale AGN discovery.



EUCLID Q1: 8 AGN-selection methods

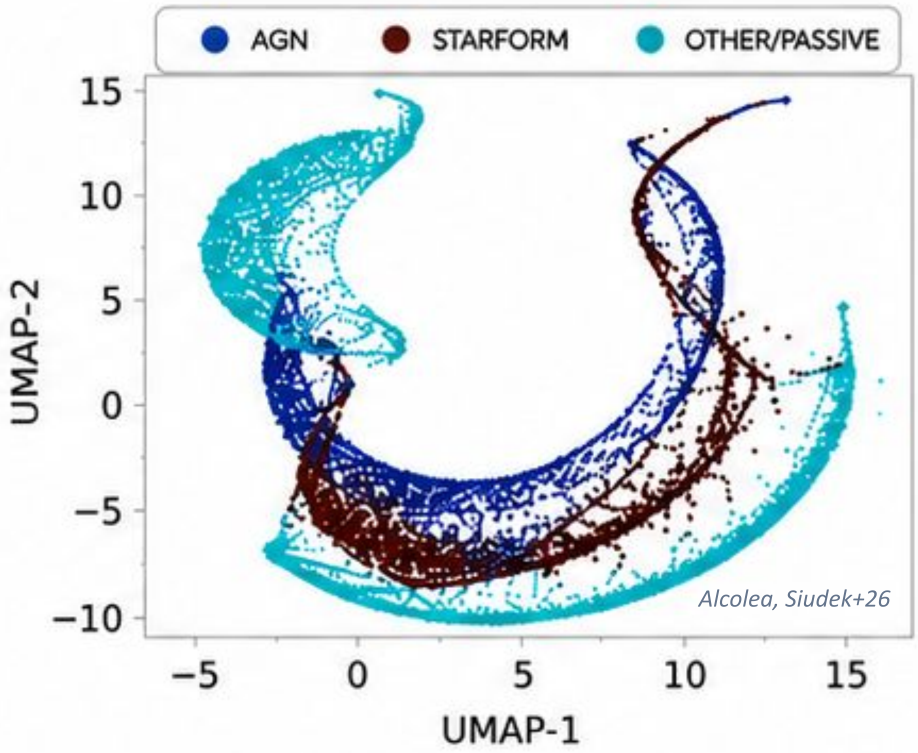
- traditional selections ⟶ reliable but incomplete,
- Euclid colors ⟶ better trade-off,
- UMAP: A 25-D colour + morphology representation improves the completeness–reliability trade-off

Aizenberg, Stern + Euclid Collaboration (under Collaboration review)
See also: Siudek+18a,b, Hviding+24, Dubois+24, Gouldman (in prep), Siudek+25a

A learned representation can scale AGN selection to millions of galaxies

Recovering hidden AGN in latent space

Hidden AGN recovered from low SNR spectra.



Spectrum → learned representation → latent space → AGN candidates

DESI EDR: Recovering ~10% of AGN from noise

- ~10,000 AGN spectra in the sample,
- ~1,000 (10%) missed by optical diagnostic (BPT) due to low SNR,
- 94% (+/-2%) of these are confirmed AGN by other methods & stacked spectrum.

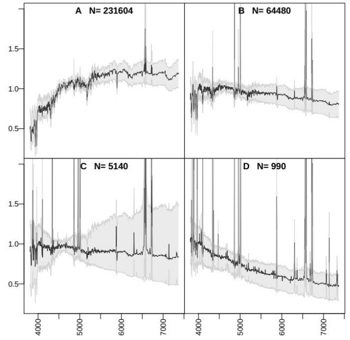
AGN diagnostic methods used

AGN Diagnostic Method	Number of Sources (%)
Total unique active black holes detected	1048 (93.5%)
Detected by ≥ 2 methods	887 (78.9%)
Detected by ≥ 3 methods	523 (46.6%)
Optical AGN (non-BPT)	974 (86.9%)
BPT: Seyfert (robust)	176 (15.7%)
Mass-excitation AGN	822 (73.3%)
WHAN: Strong AGN	792 (70.6%)
WISE colour diagnostic	317 (28.3%)
Kinematics-excitation AGN	294 (26.2%)
BPT (Seyfert/LINER)	239 (21.3%)
Broad emission lines	216 (15.3%)
QuasarNet reclassifies as QSO	167 (14.9%)
BLUE diagram: AGN	97 (8.6%)
Ne V AGN	85 (7.6%)
Redrock QSO template	81 (7.2%)
MgII broad line (afterburner)	46 (4.1%)
WHAN: Weak AGN	34 (3.0%)
He II BPT AGN	22 (2.0%)

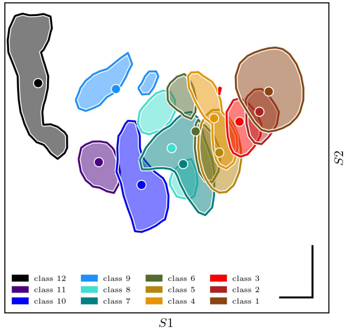
Latent space can reveal structure beyond classification

Similar objects cluster — and continuous paths can trace galaxy evolution.

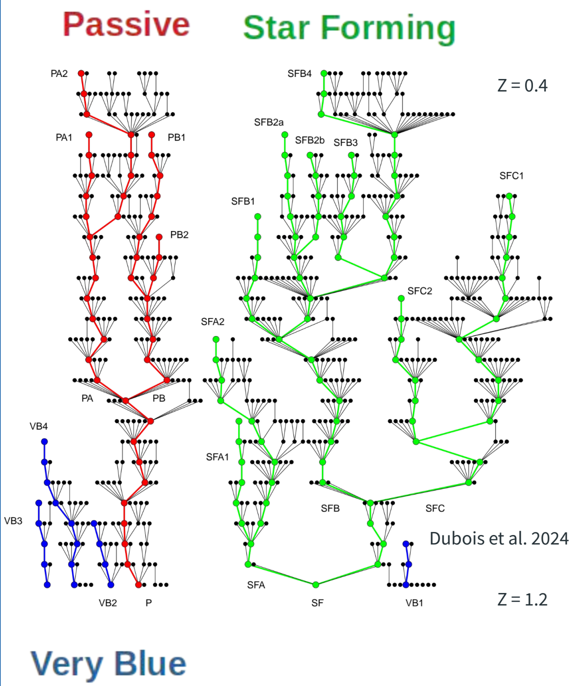
1 SDSS (z~0.2)



2 VIPERS (z~0.7)



3 SDSS -VIPERS (z~0.2 - 0.7)



LATENT SPACE ACROSS TIME

- SDSS (spectra):
  → spectra clusters in classes,
- VIPERS (UV-FIR SEDs):
  → superpass WISE AGN selection,
  → found overmassive AGN in dwarf galaxies,
- SDSS + VIPERS (spectra):
  → evolutionary tree,
  → AGN branch.
- SDSS + DESI (spectra):
  → pushing to high-z.

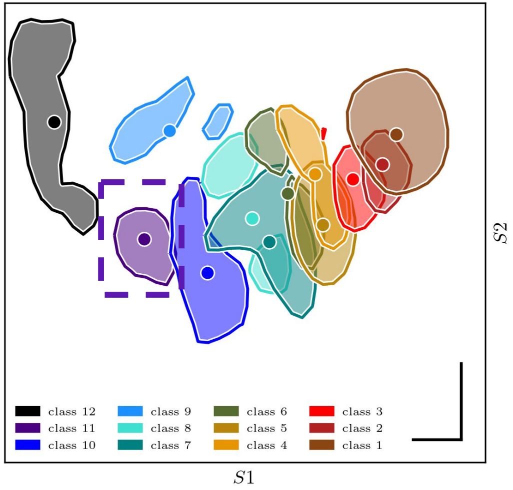
Fraix-Brunet+21, Siudek+18, Dubois+24, Chambon+25
See also: Siudek+18a,b, 22, Pan+20…

ML4ASTRO 2026 3rd edition: Deep Learning for Galaxy Formation: M. Siudek msiudek@iac.es

Overmassive black holes in dwarf galaxies

Finding unexpected AGN - from VIPERS to JWST.

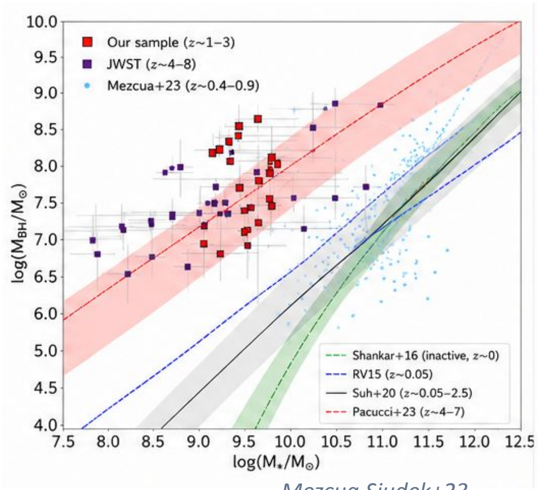
1 Where we found them



Siudek+18a,b, 22, 23

Hidden population of AGN

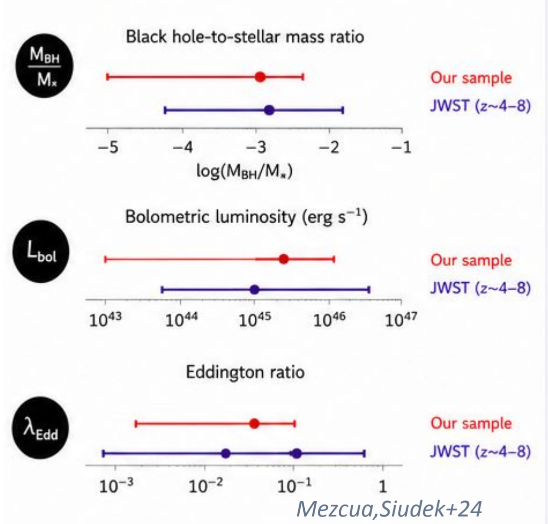
2 Two orders of magnitude overmassive



Mezcua,Siudek+23

12 overmassive BHs in dwarfs at z~0.7

3 Similar to the early Universe



Mezcua,Siudek+24

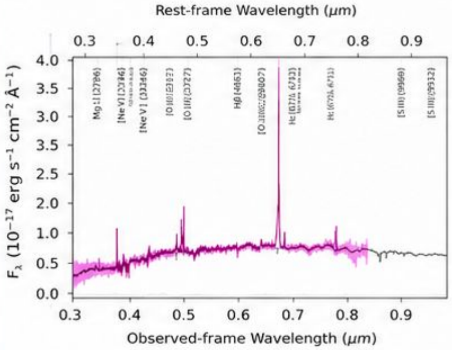
Dwarfs at z ~2 share properties with JWST AGN

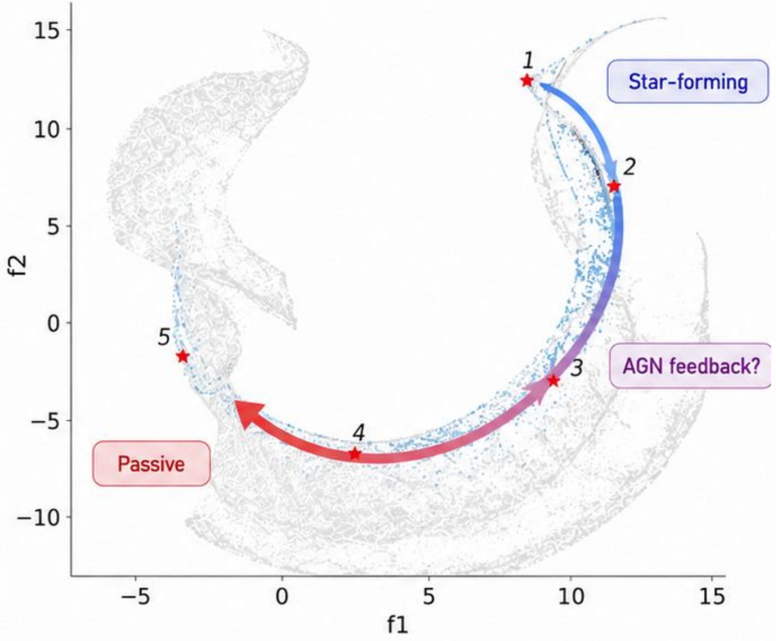
AGN feedback written in the data

The latent space reveals a sequence from star formation to AGN activity to quenching.

4–5 Passive

Absorption-dominated, flat continuum, no emission

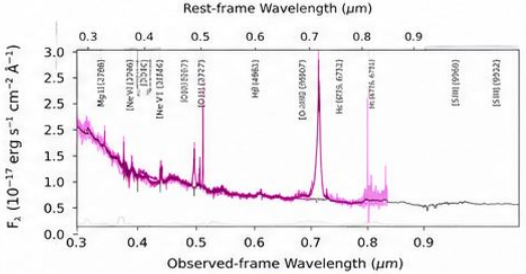




Alcolea, Siudek, submitted

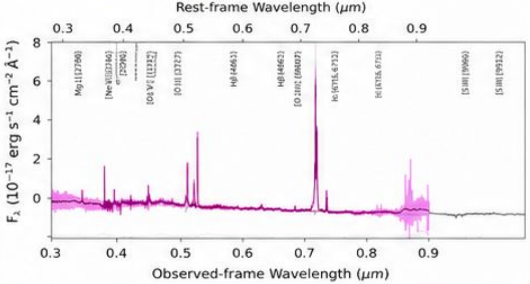
1–2 Star-forming

Strong emission lines, blue continuum



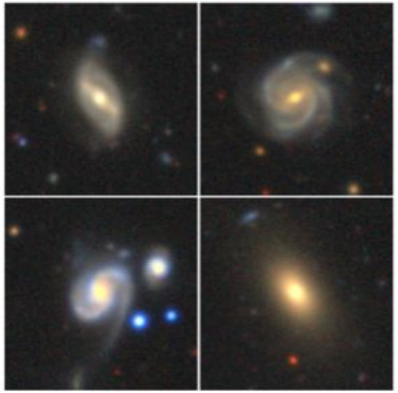
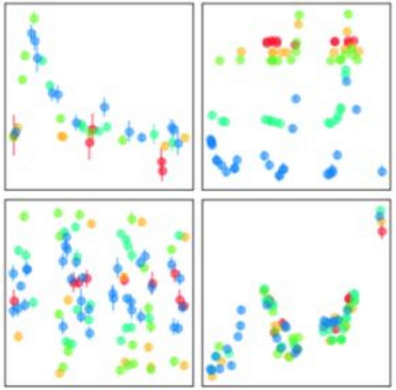
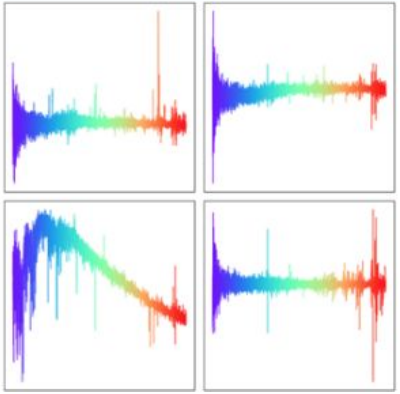
3 AGN signatures

Strong narrow lines, elevated continuum



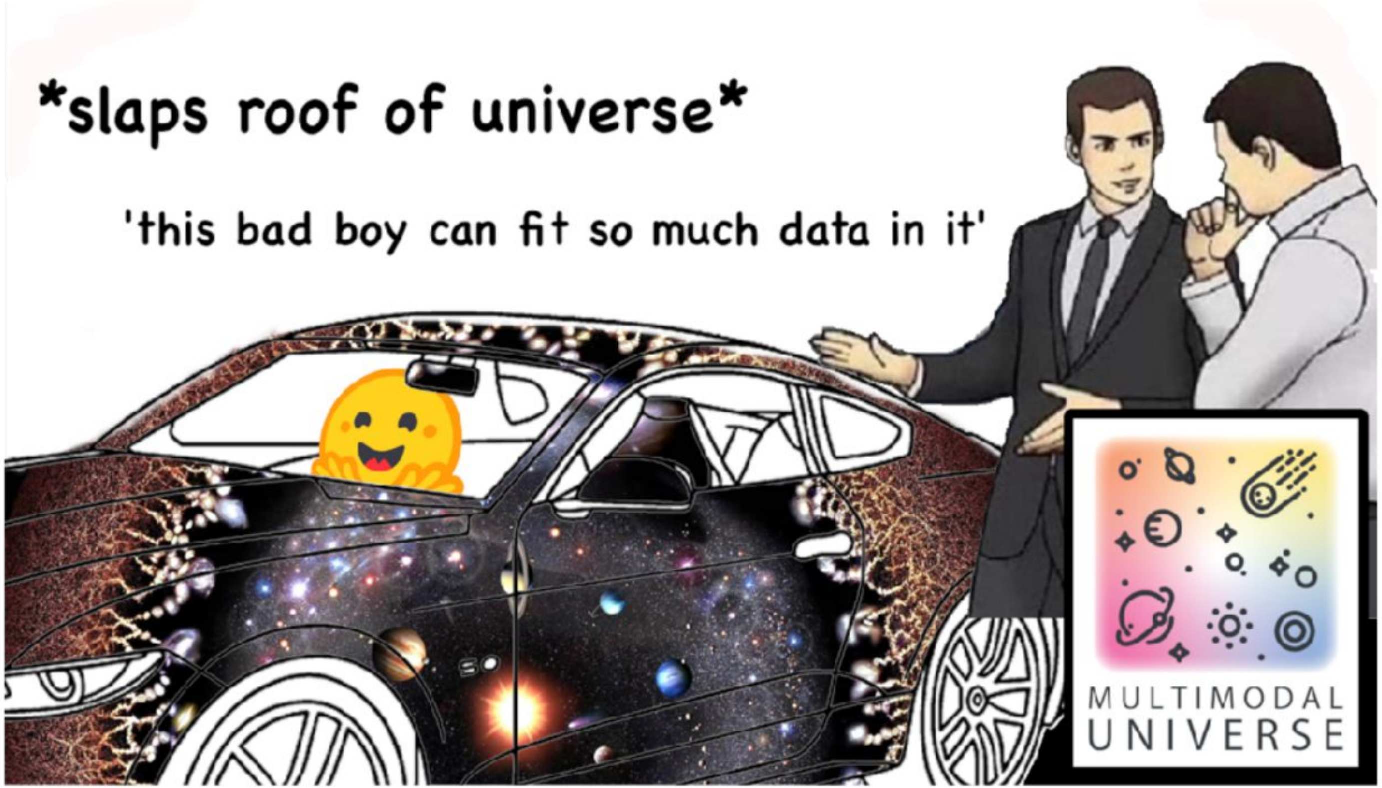
Representation learns real astrophysics

The next step: the multimodal astronomical universe

	Images	Time-Series	Spectra
# examples	140M	4.5M	225M
Description	images in a variety of wavelength ranges, including optical and infrared	multivariate time-series of flux + uncertainty in different wavelength ranges	flux as a function of wavelength
Tasks	galaxy classification, physical property estimation	time-series classification, redshift estimation	physical property estimation
Examples			

MMU Collaboration+24

Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data



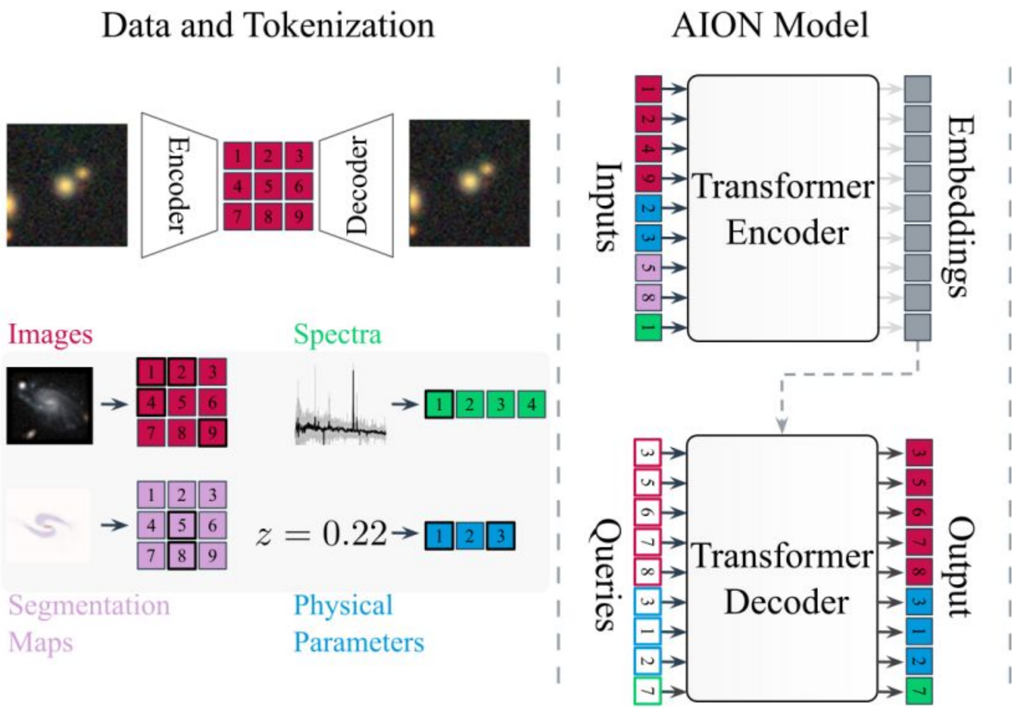
MMUv1.5 now includes Euclid Q1, JWST COSMOS-Web, JADES, ZTF

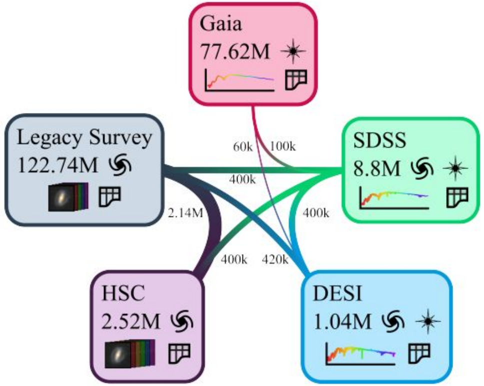
MMUxHATSxHuggingFace

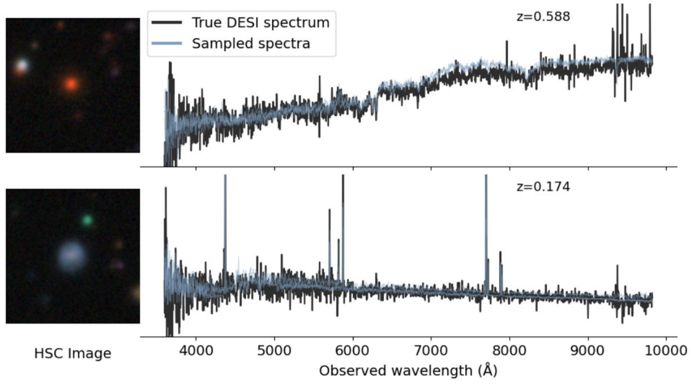
Polymathic

From models to astronomical infrastructure

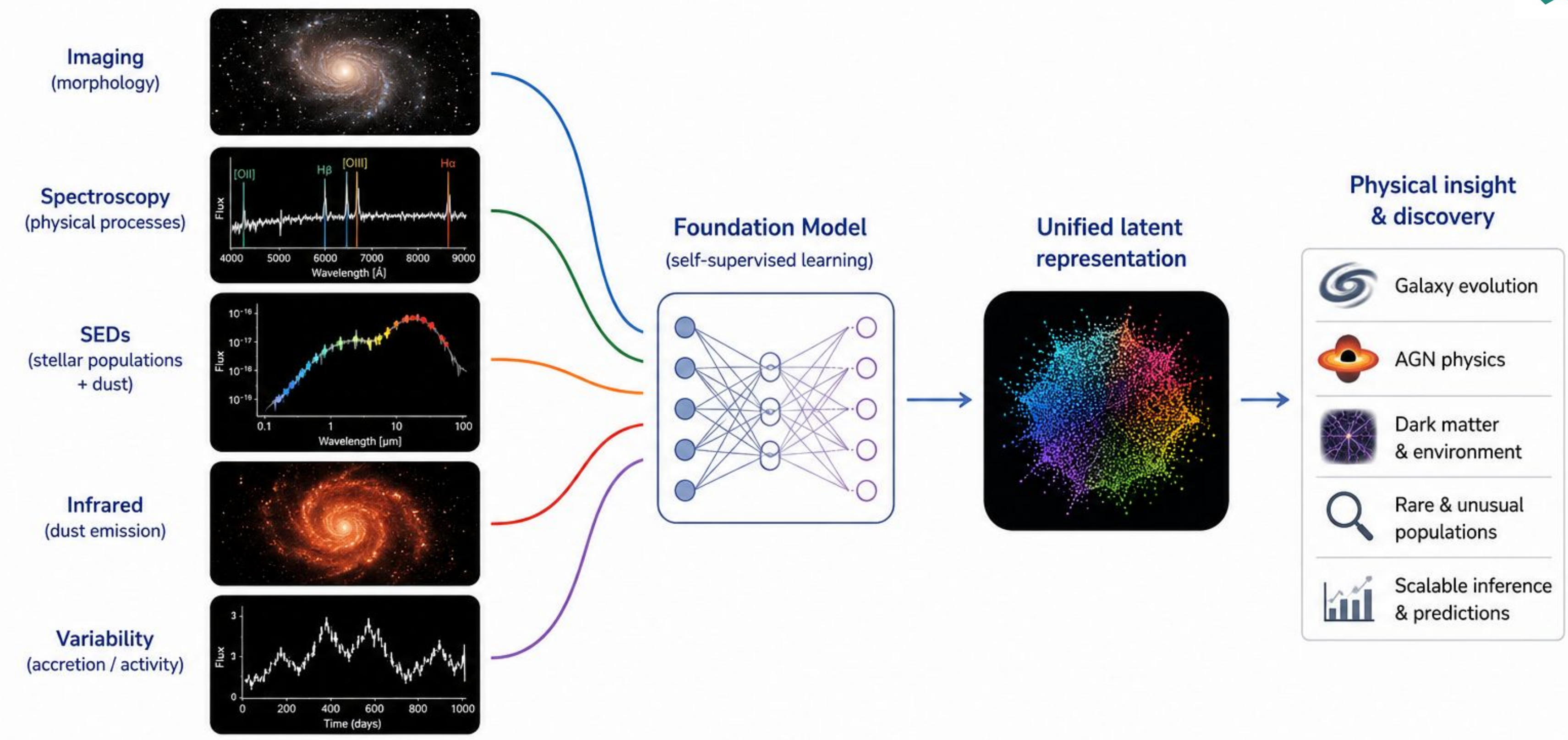
AION-1: Astronomical Omnimodal Network.







Astronomy has entered the era of multi-modal surveys



Machine learning is transforming astronomy by moving us from measuring individual properties to learning the physical representations hidden across complementary observations