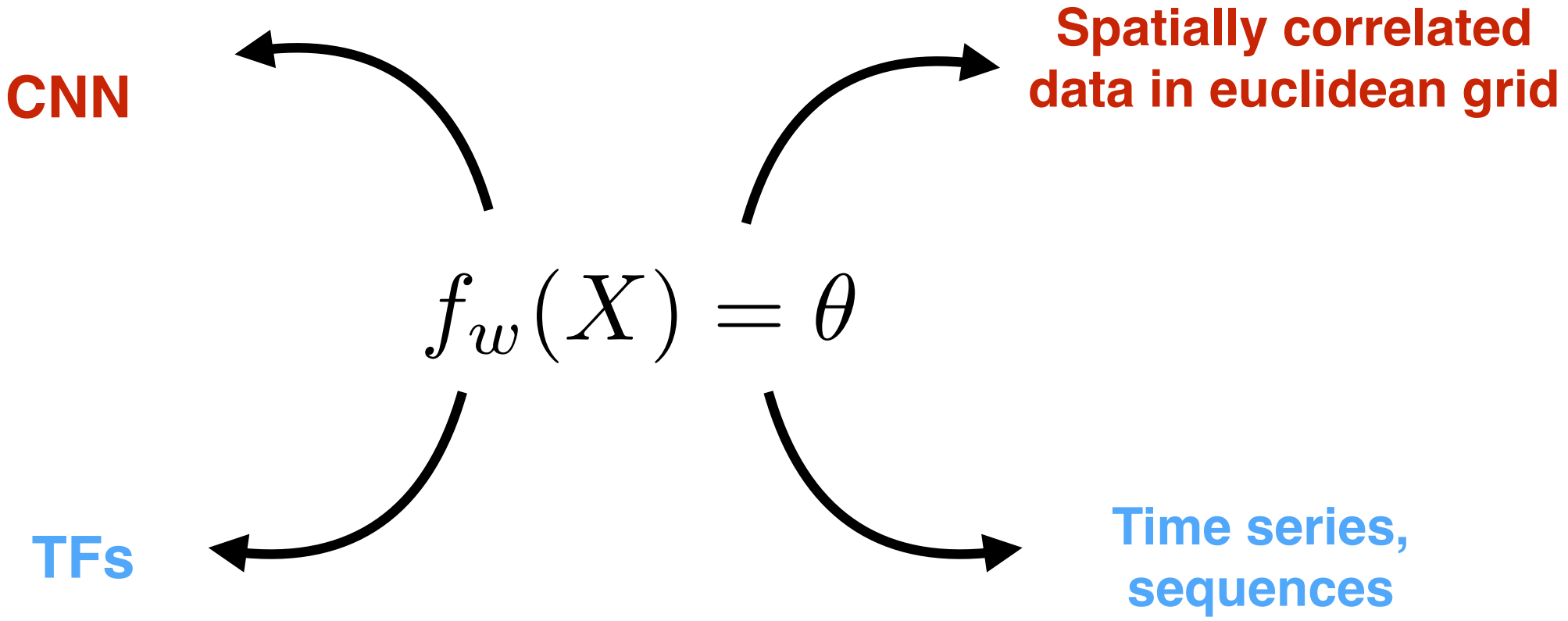
Foundation models and representation learning

Marc Huertas-Company & Hubert Bretonnière

Rodolphe Clédassou School 2025 - Y1

Neural networks as universal feature extractors



Supervised learning pipeline

1. **Gather labeled dataset** (X, θ)
2. **Train** a network to map between labels and data $f_w(x_i) = \theta_i$ for $(x_0, \ldots, x_N) \in X_{tr}$
3. **Validate on a test set** X_t
4. **Apply to new unlabelled dataset** X_u

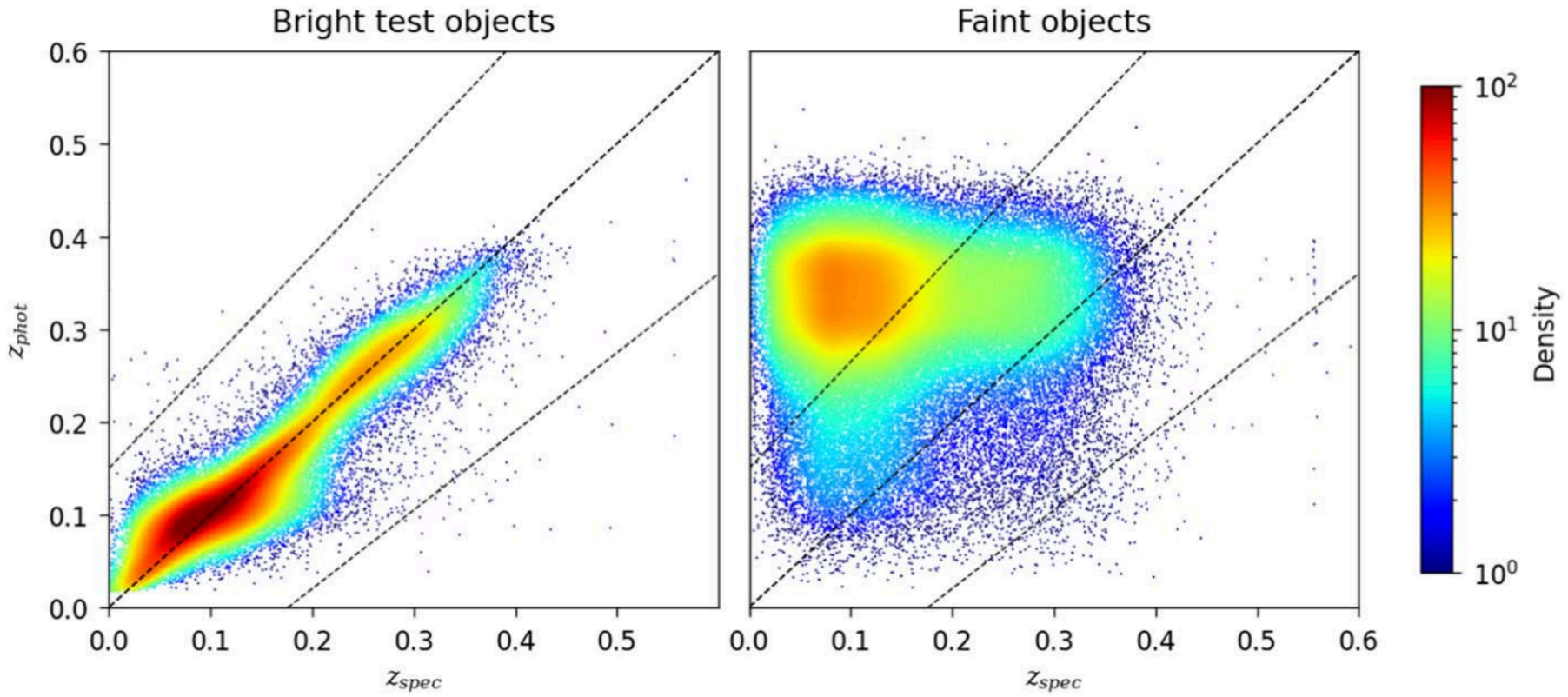
The limits of supervised learning

requires large amount of labeled data
(e.g. galaxy zoo) - the majority of the data is not labelled

limited representation of the information content; network needs to be retrained for new dataset / new labels even for similar tasks

limited discovery potential - rare or novel objects have by definition few labeled examples

photometric redshift estimation is a paradigmatic example



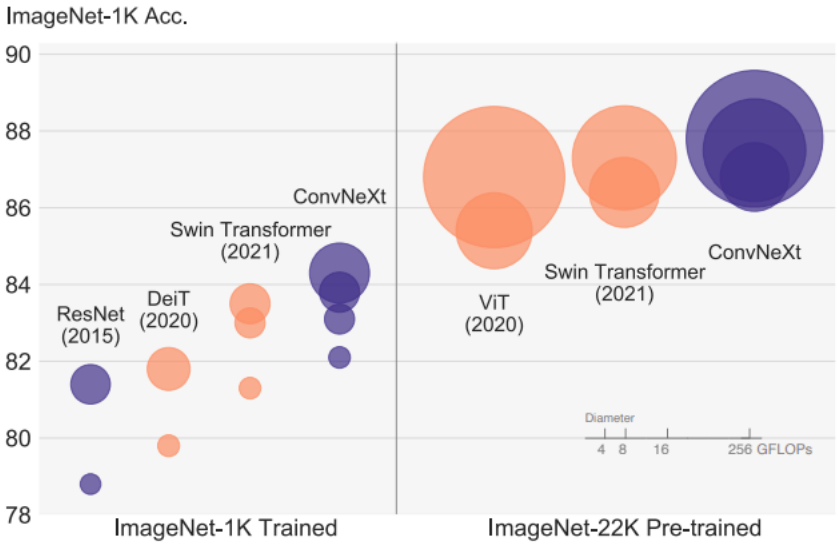
Lee+22

*almost all the extensive literature is focused on mitigating the effects of biased labelling

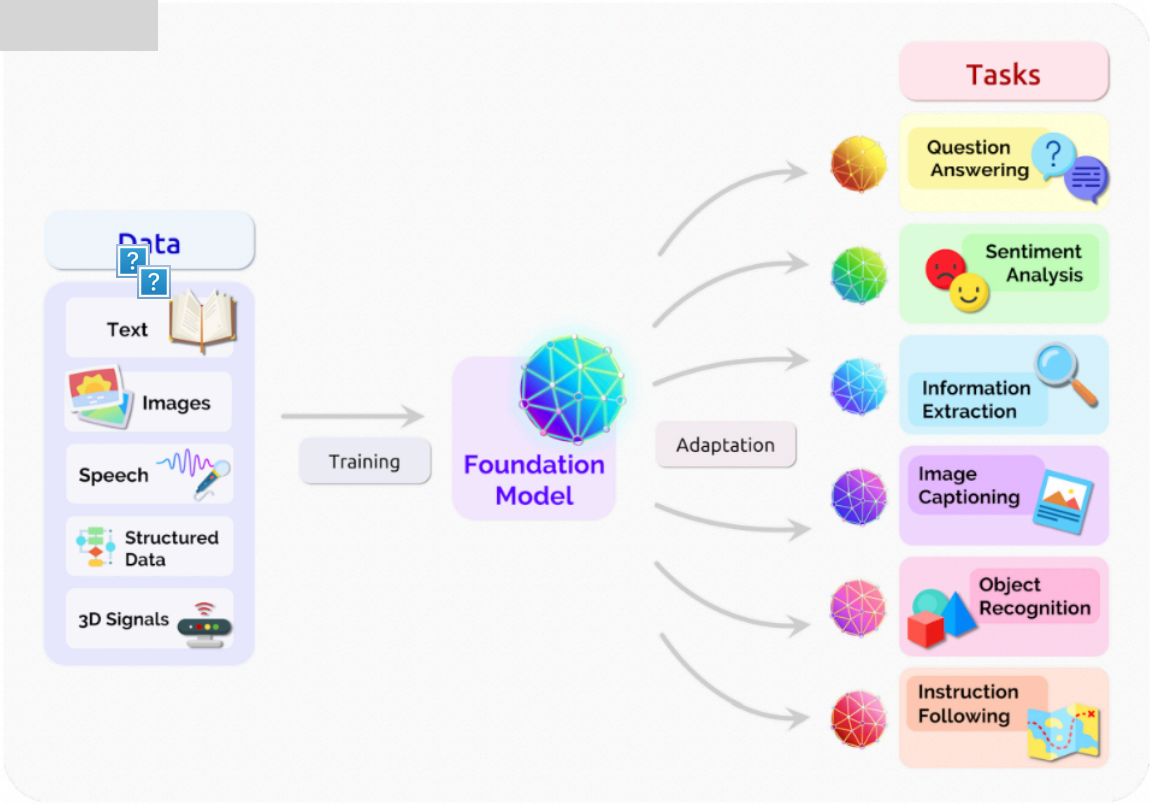
Foundation Models: One model to rule them all

Foundation Model approach:

1. **Pretrain** models on **pretext tasks**, without labels, on very large scale datasets.
2. **Adapt** pretrained models to d**ownstream tasks.**



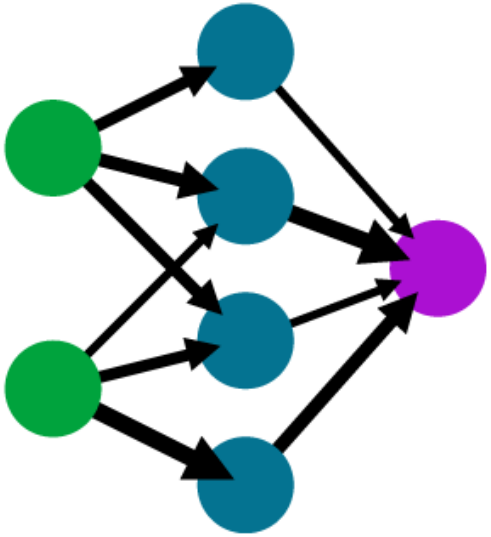
Liu+20



Bommasani+21

representation learning - use neural networks as a universal embedder

X



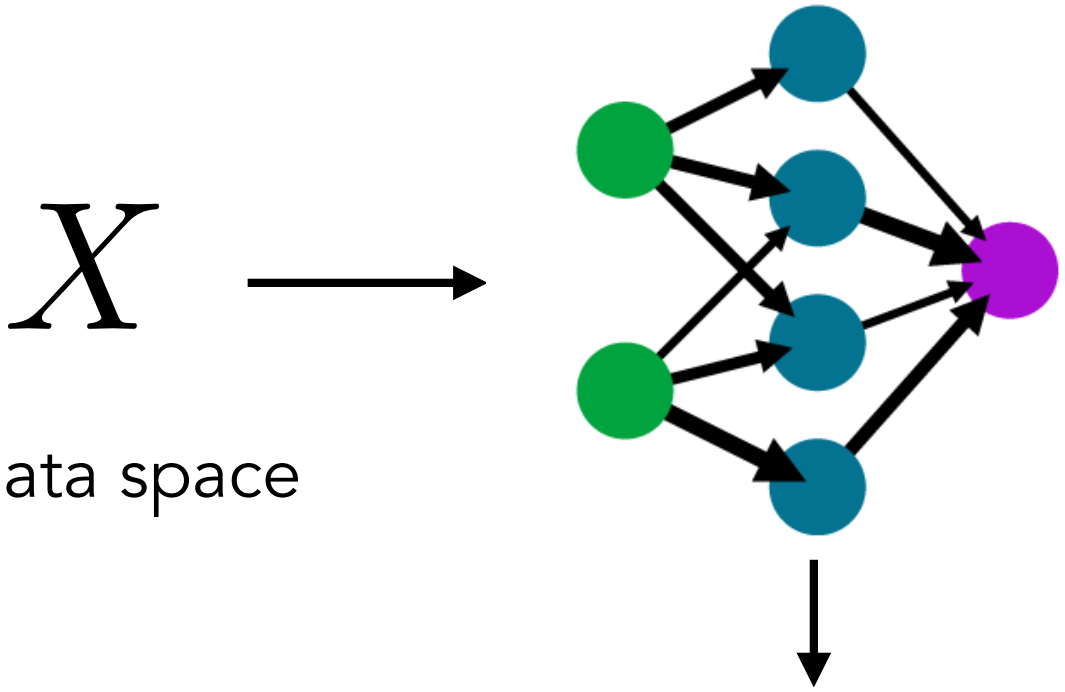
Z

data space

embedding space

general informative representation of X independent of any specific task

representation learning - use neural networks as a universal embedder



X

data space

(No need to be a neural network)

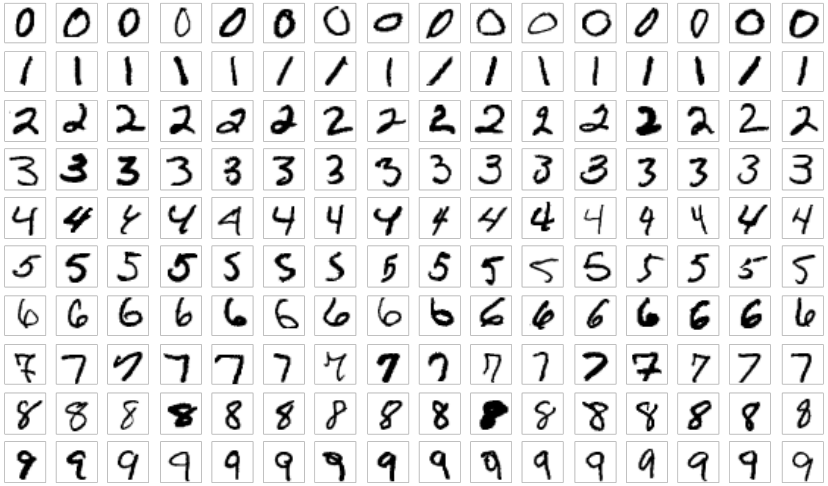
Z

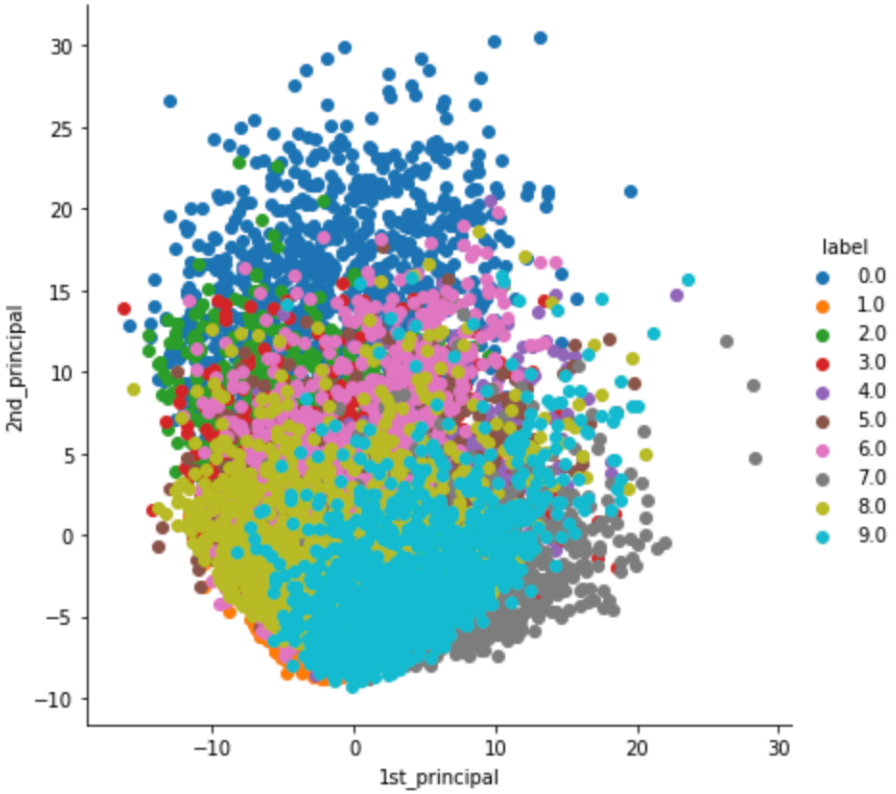
embedding space

general informative representation of X independent of any specific task

limitations of PCA

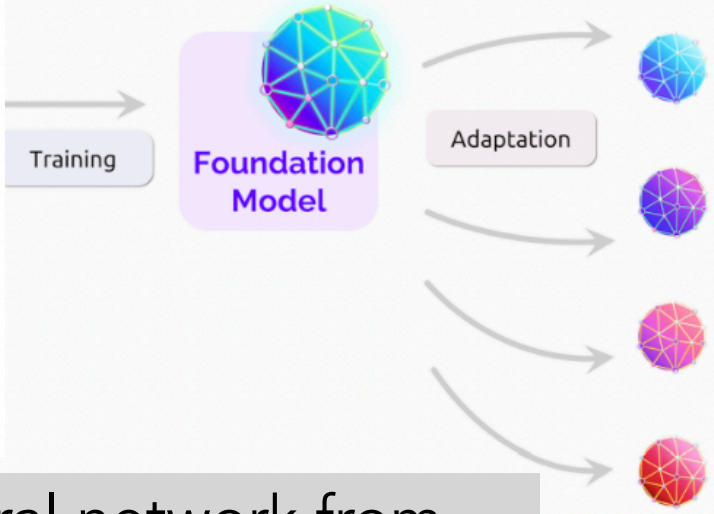
assumes data follows a multivariate gaussian distribution





https://medium.com/analytics-vidhya/principal-component-analysis-pca-with-code-on-mnist-dataset-da7de0d07c22

what does this mean for astro?



classifications

inference

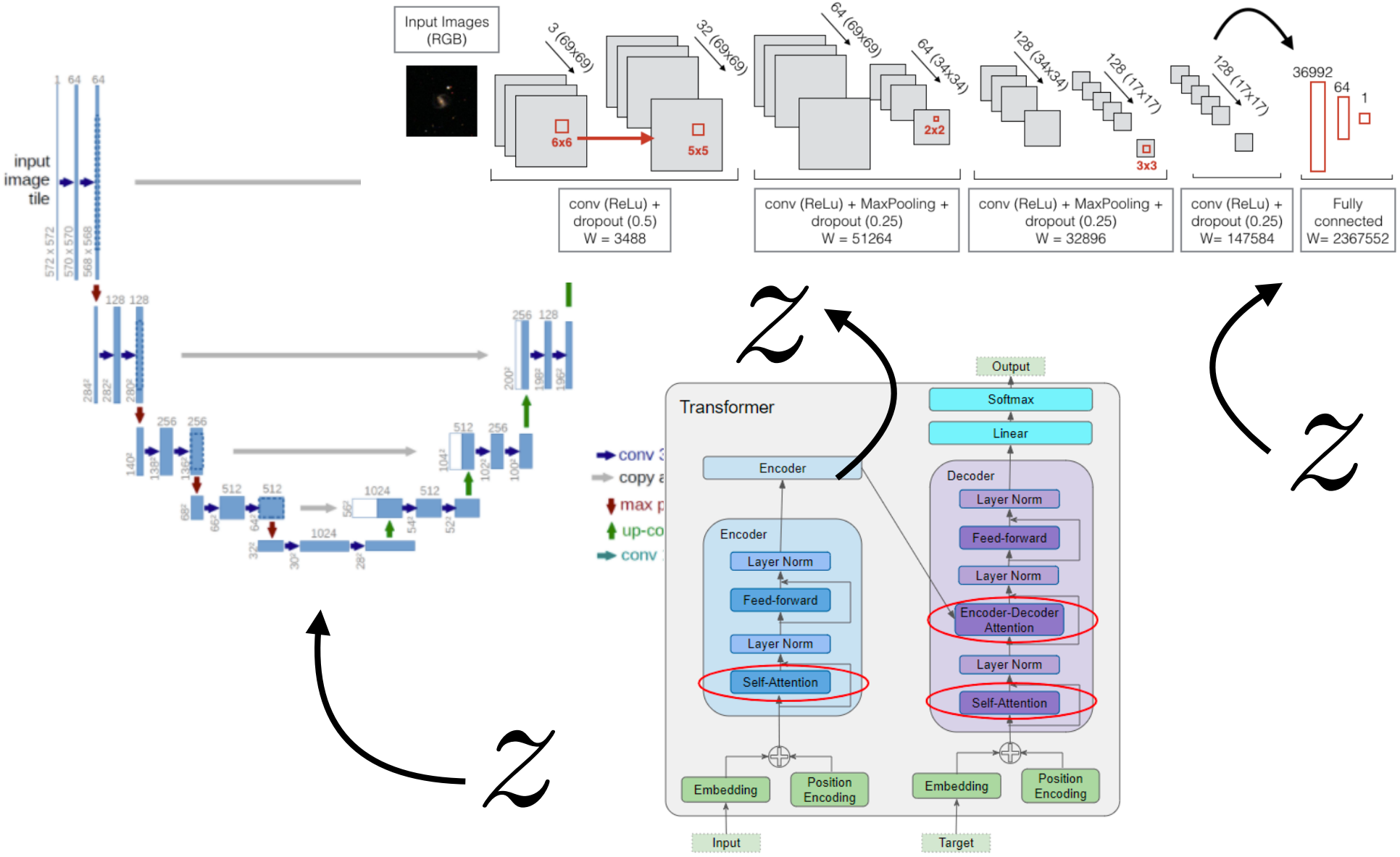
segmentation

anomaly

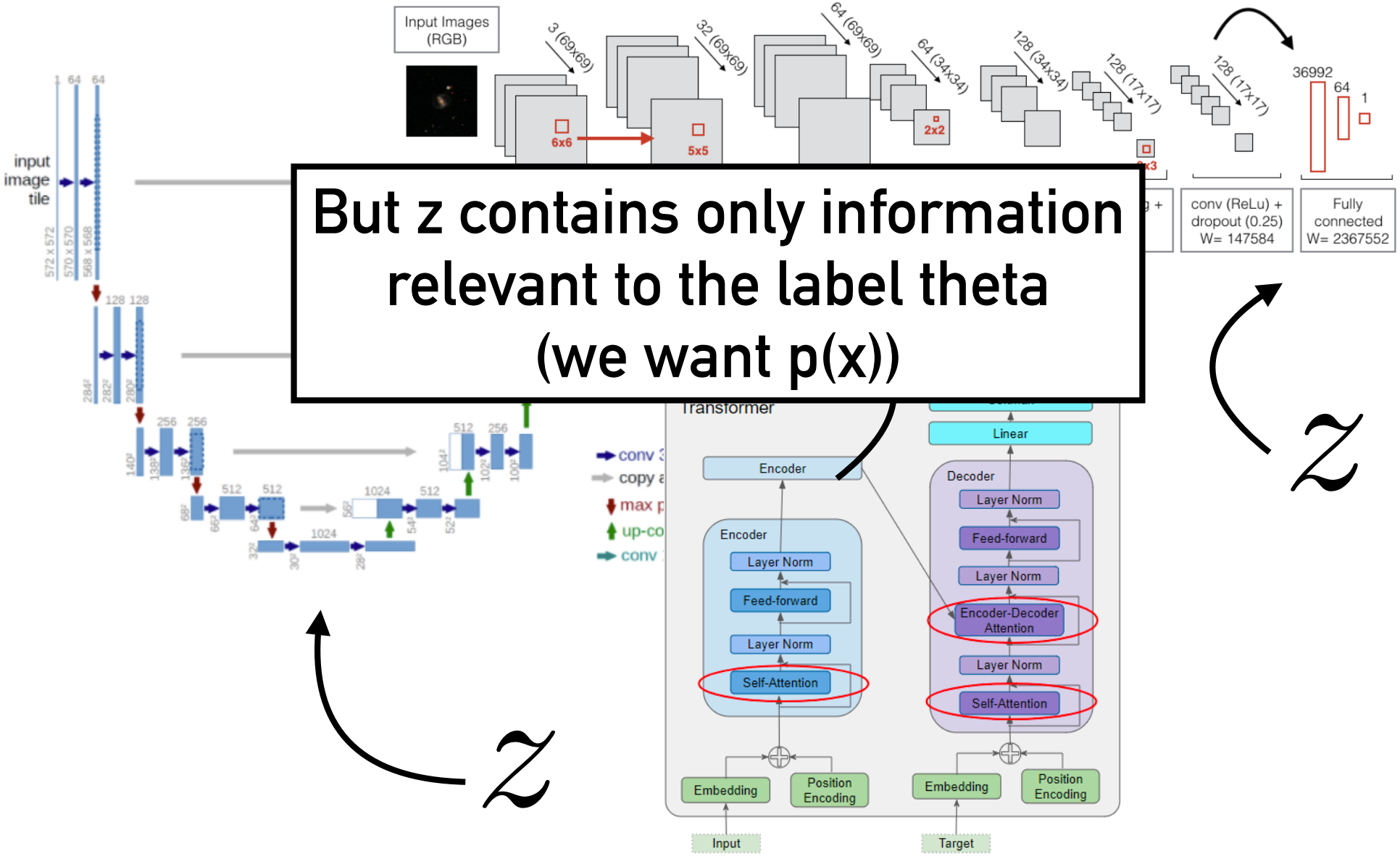
sim / obs

1. never again train a neural network from scratch
2. no labels required (or very few)
3. integrates all data beyond images
4. very simple analysis tools for downstream tasks

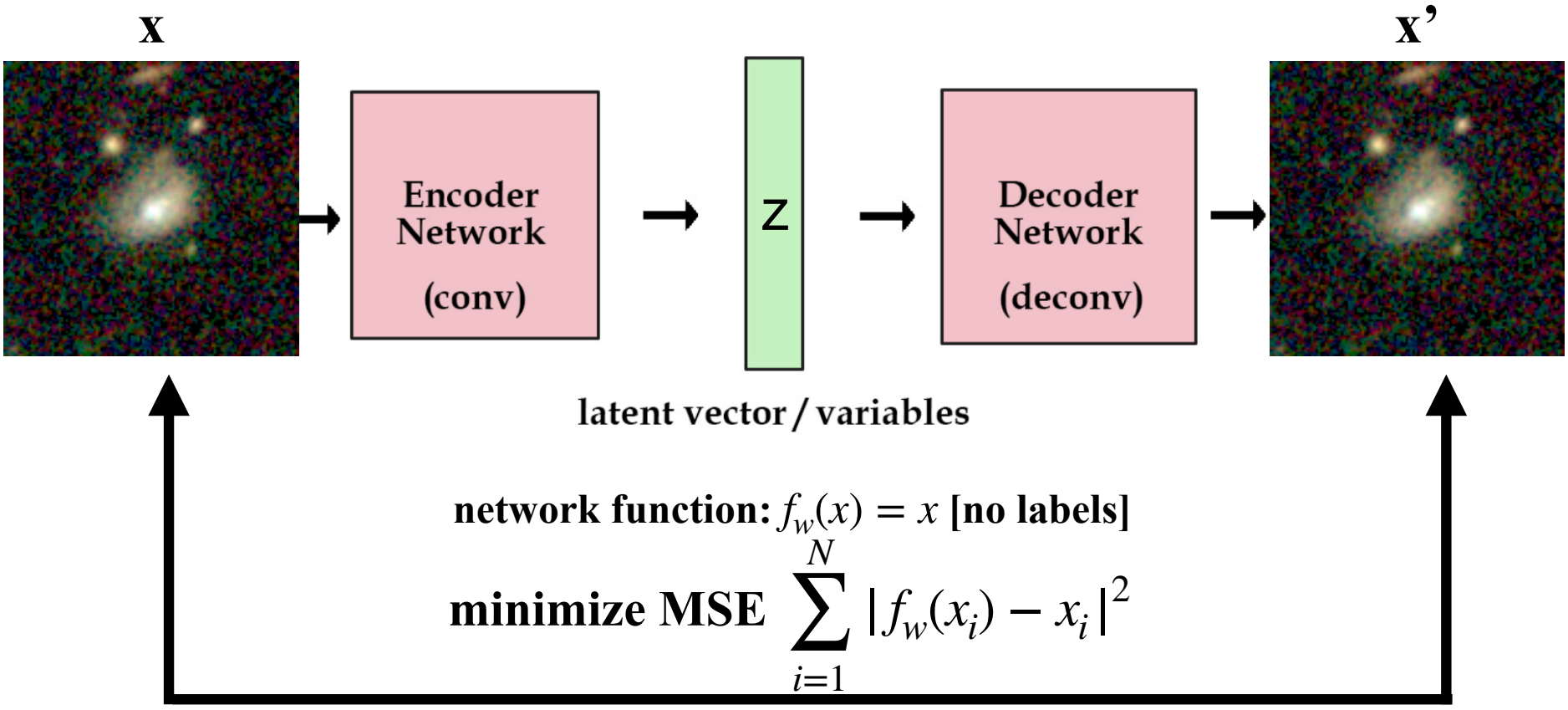
any supervised network obtains a non-linear representation of the data



any supervised network obtains a non-linear representation of the data



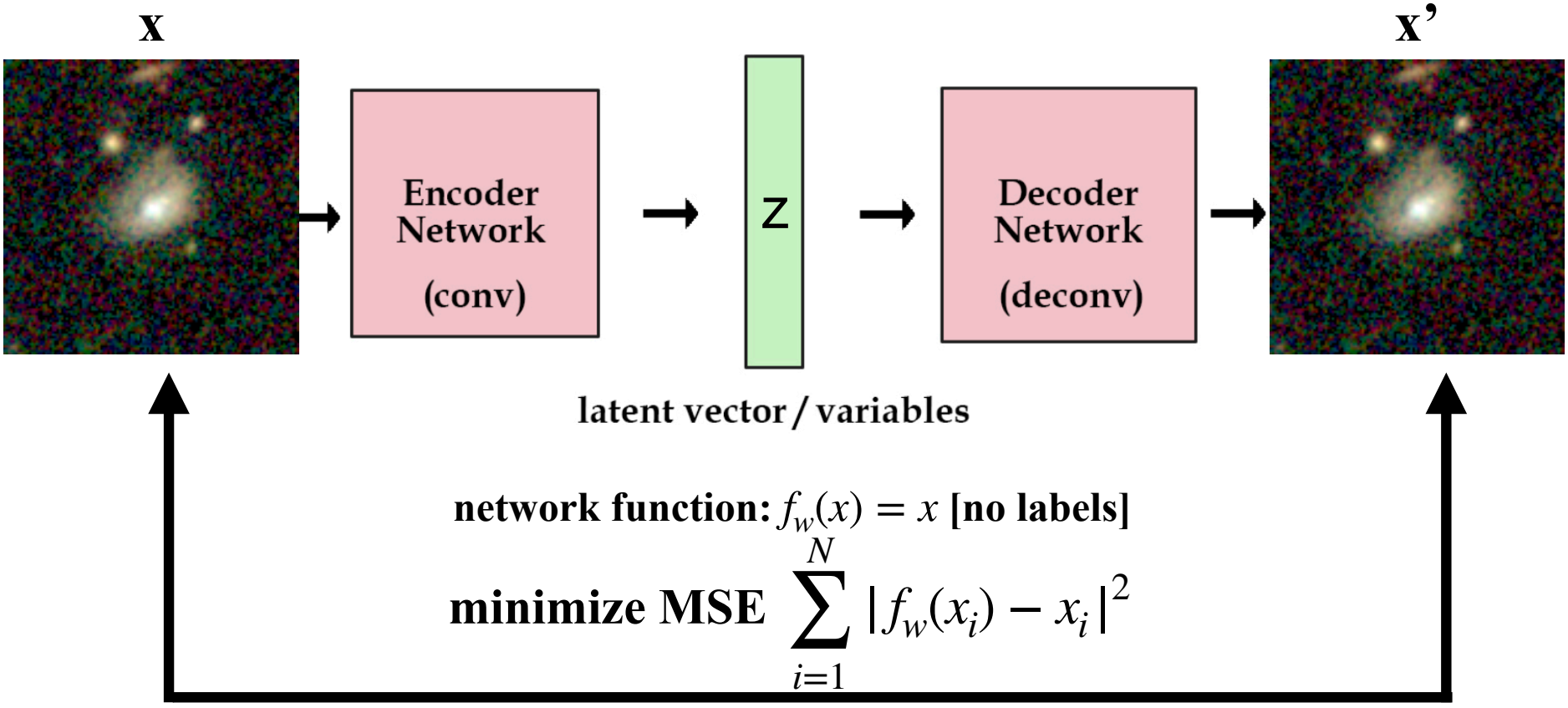
auto-encoder



network function: $f_w(x) = x$ **[no labels]**

minimize MSE $\sum_{i=1}^{N} |f_w(x_i) - x_i|^2$

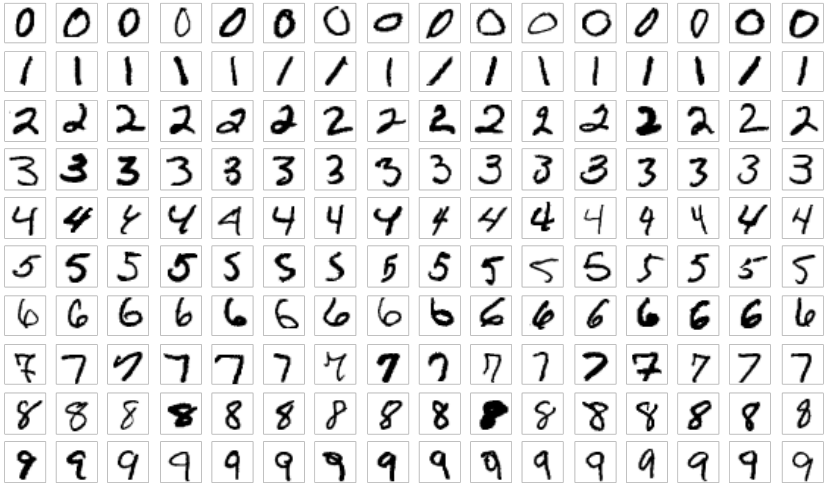
auto-encoder

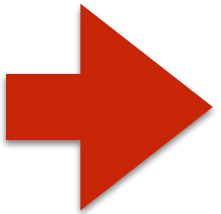


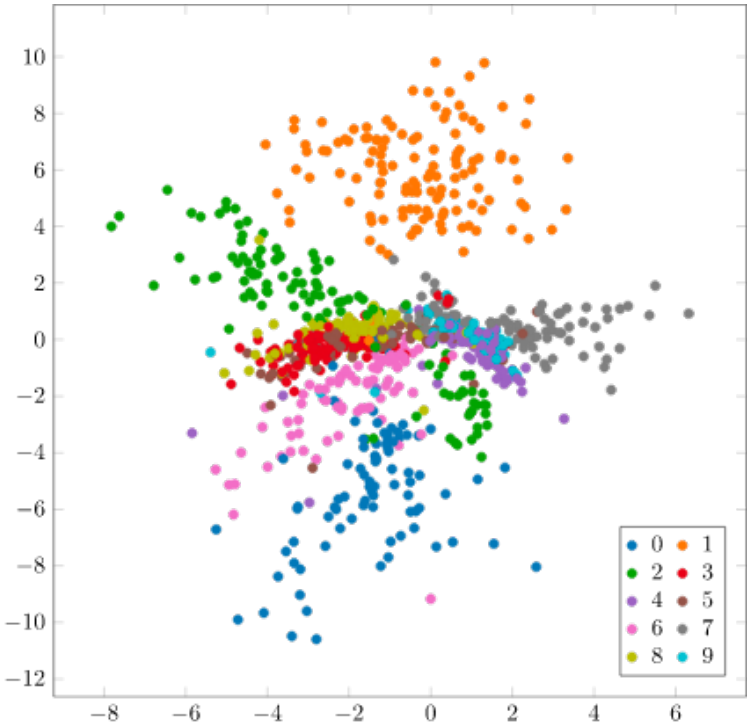
By creating a bottleneck, we encourage z to encode a lower dimensional informative representation of the data

* **question**: what would happen if we train an auto encoder with no activation functions - only linear operations?

AUTOENCODER REPRESENTATION OF MNIST

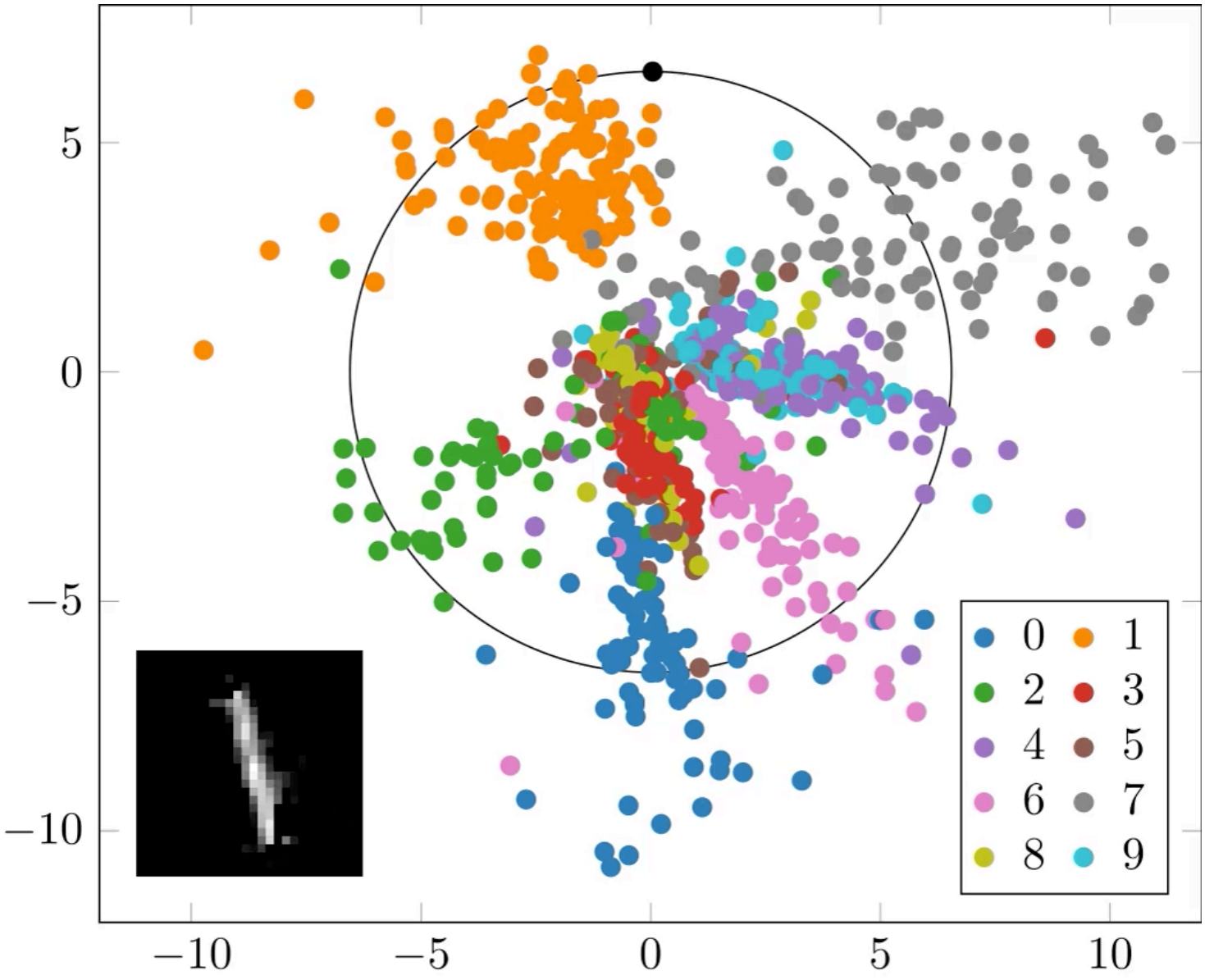






https://gertjanvandenburg.com/blog/autoencoder/

AUTOENCODER REPRESENTATION OF MNIST



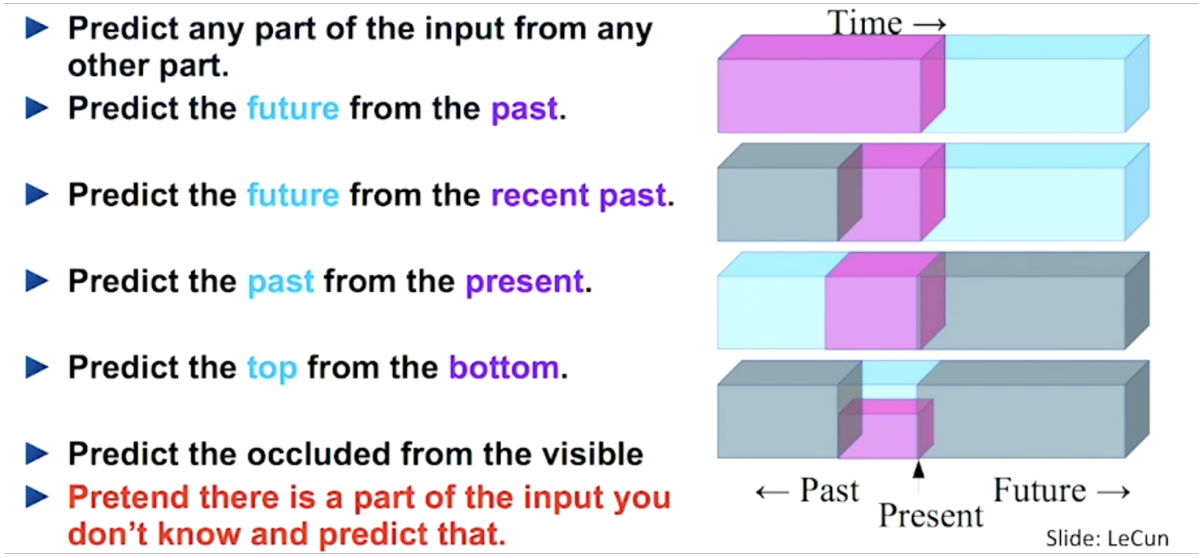
https://gertjanvandenburg.com/blog/autoencoder/

self-supervised learning

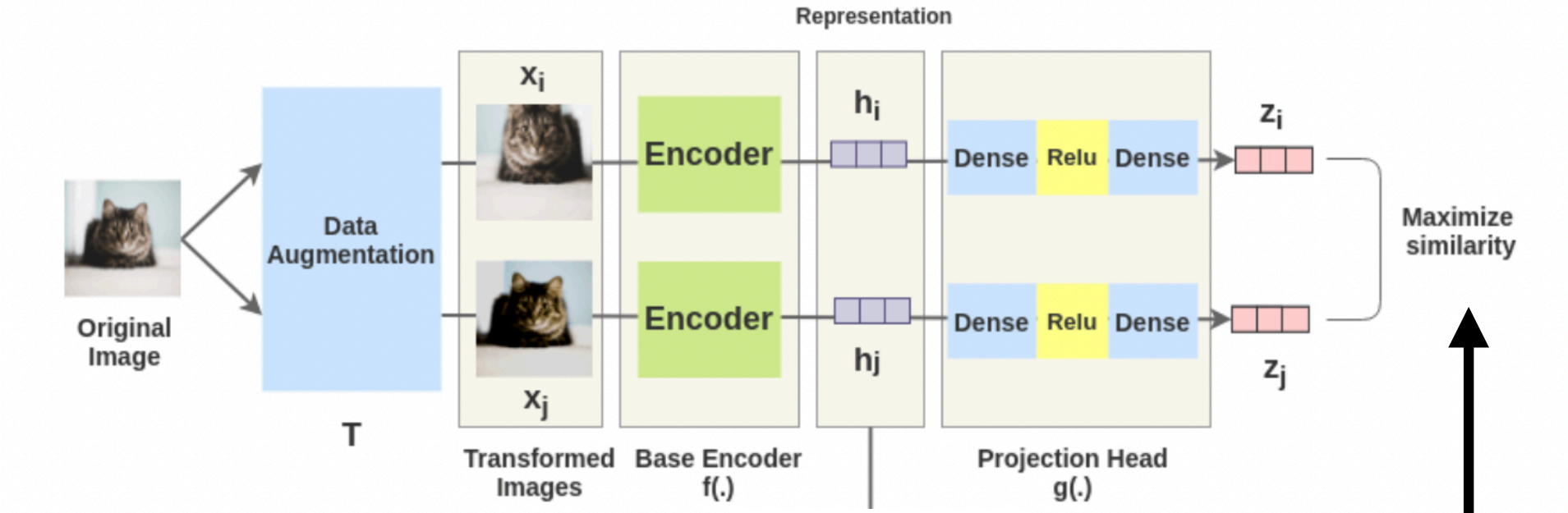
Another alternative is to invent a target to obtain z

$$f_w(X) = \hat{\theta}$$

(Still using a discriminative model, with a general target, and use the latent z as a representation of the data)



contrastive self-supervised learning



z

Downstream tasks

The key is the loss function

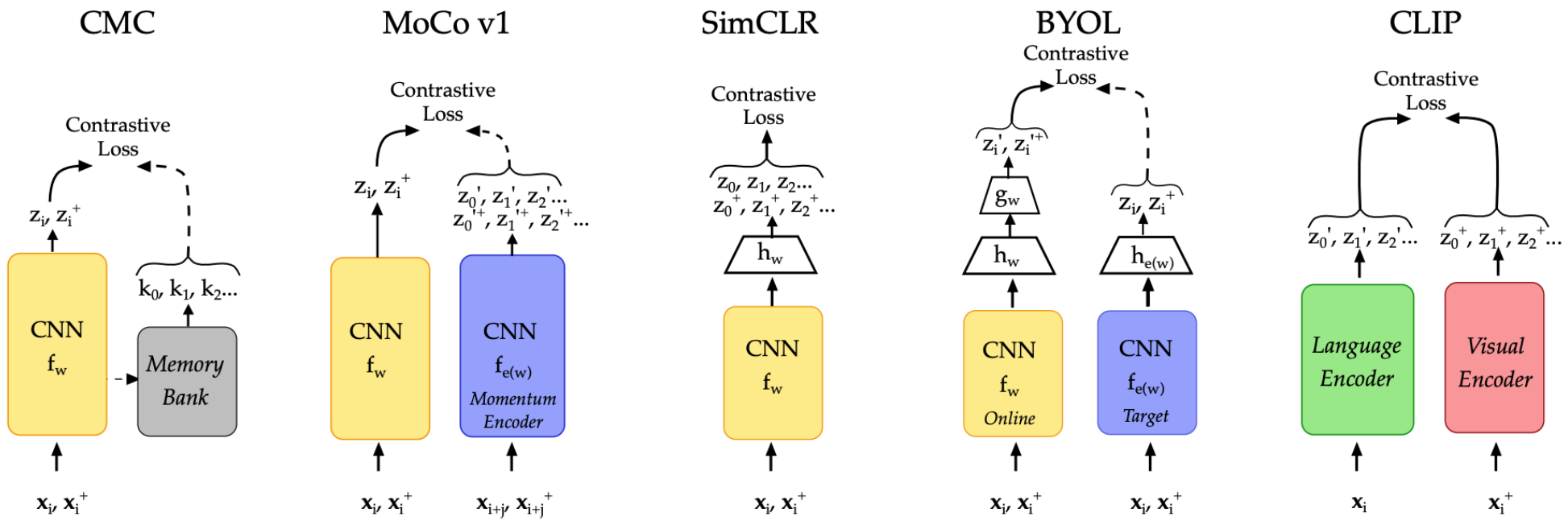
The contrastive loss:

Similarly between
two representations of positive pairs

$$l_{i,j} = -\log \frac{\exp(\langle z_i, z_j \rangle / h)}{\sum_{k=1,k\neq i}^{2N} \exp(\langle z_i, z_k \rangle / h)},$$

Sum of all similarities between
negative pairs

contrastive learning



- - ► No gradient update

z_i target representation

z_i^+ target's pair representation

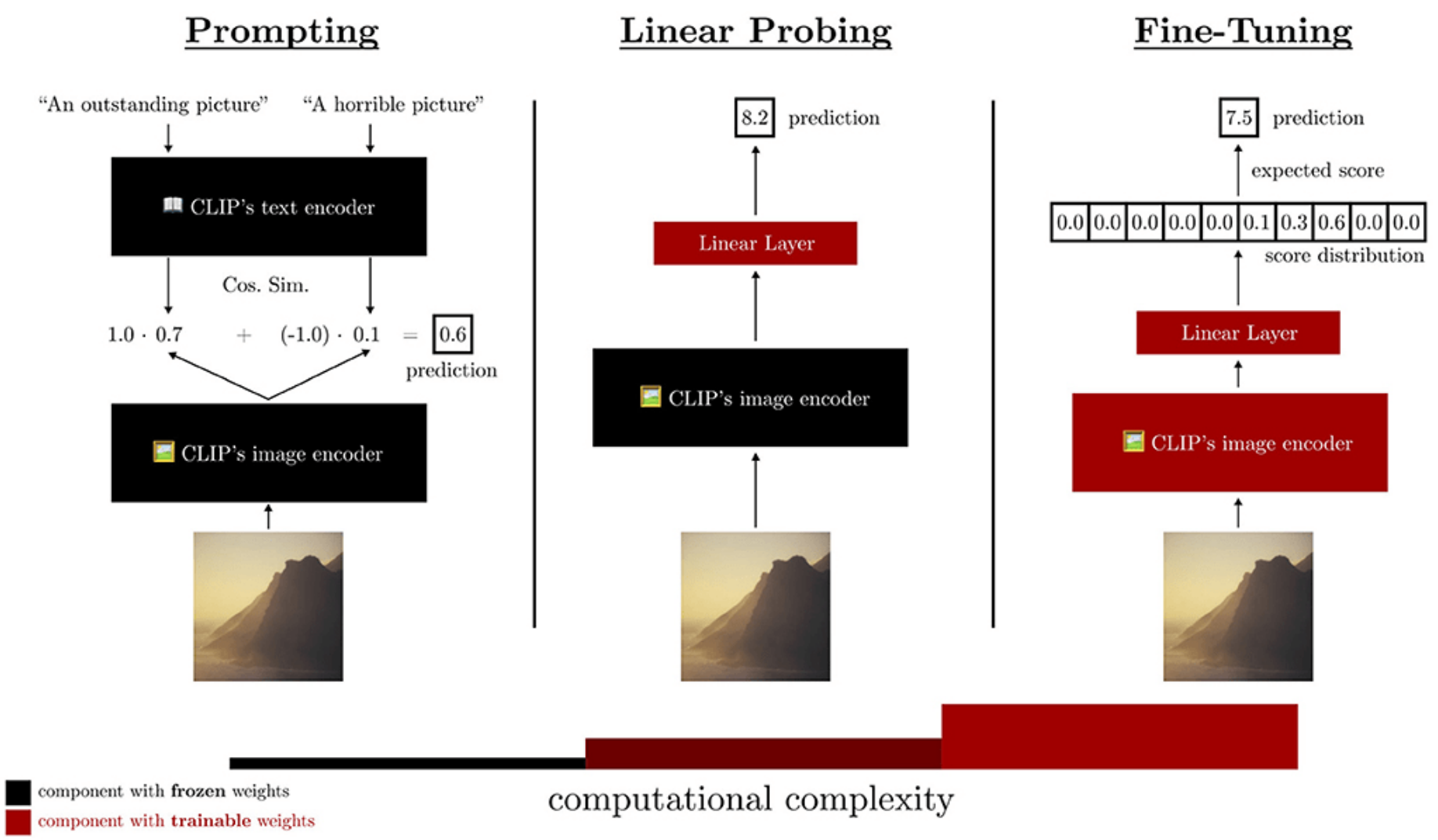
z_i' parallel network's representation

k_i saved key representation

x, x^+ input element and positive pair

f, g, h Trainable functions

MHC+23


Prompting
Linear Probing
Fine-Tuning
"An outstanding picture"
"A horrible picture"
CLIP's text encoder
Cos. Sim.
1.0 · 0.7 + (-1.0) · 0.1 = 0.6
prediction
CLIP's image encoder
8.2 prediction
Linear Layer
CLIP's image encoder
7.5 prediction
expected score
0.0 0.0 0.0 0.0 0.0 0.1 0.3 0.6 0.0 0.0
score distribution
Linear Layer
CLIP's image encoder
component with frozen weights
component with trainable weights
computational complexity

Zhai+22

autoregressive

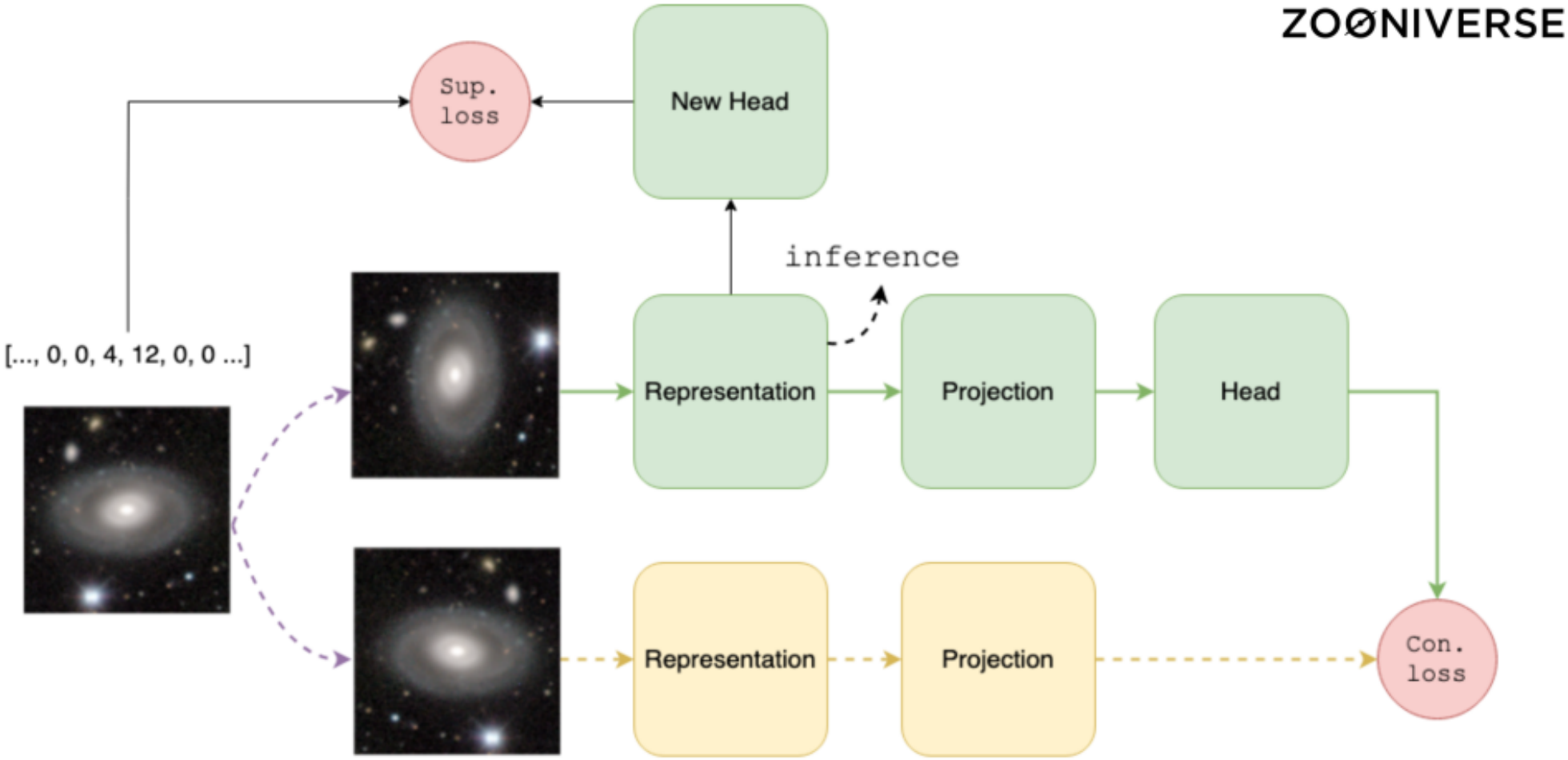
$$P_\theta(x) = \prod_{t=2}^{T} P_\theta(X_t \mid \{x_{<t}\})$$

$$= P_{\theta_2}(x_2 \mid \{x_1\}) P_{\theta_3}(x_3 \mid \{x_1, x_2\}) \cdots P_{\theta_T}(x_T \mid \{x_1, \cdots, x_{T-1}\})$$

$o \leftarrow \{\bullet\}$ $o \leftarrow \{\bullet\ \bullet\}$ $o \leftarrow \{\bullet\ \cdots\ \bullet\}$

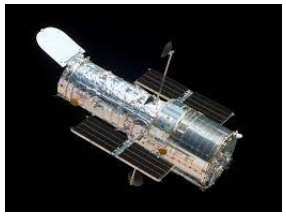
$x_2\ x_1$ $x_3\ x_1\ x_2$ $x_T\ x_1\ x_{T-1}$

ZOOBOT: an hybrid foundation model

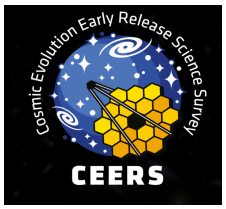


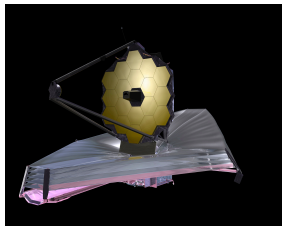
Adding a **new supervised head** to guide the contrastive representation

walmsley+

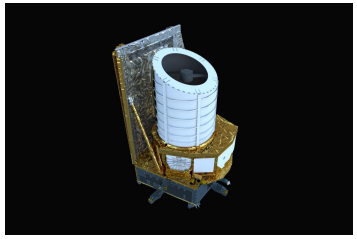


4 years





3 months



0 months

Credit: ESA/Euclid/Euclid Consortium/NASA, image processing by M. Walmsley, M. Huertas-Company, J.-C. Cuillandre

EC: M. Walmsley. MHC+25

Everything else

Smooth or featured?

Answer

Smooth

Vote Fraction

0.00 1.00

0.00 1.00

Disk edge on?

Answer

Yes

Vote Fraction

0.00 1.00

0.00 1.00

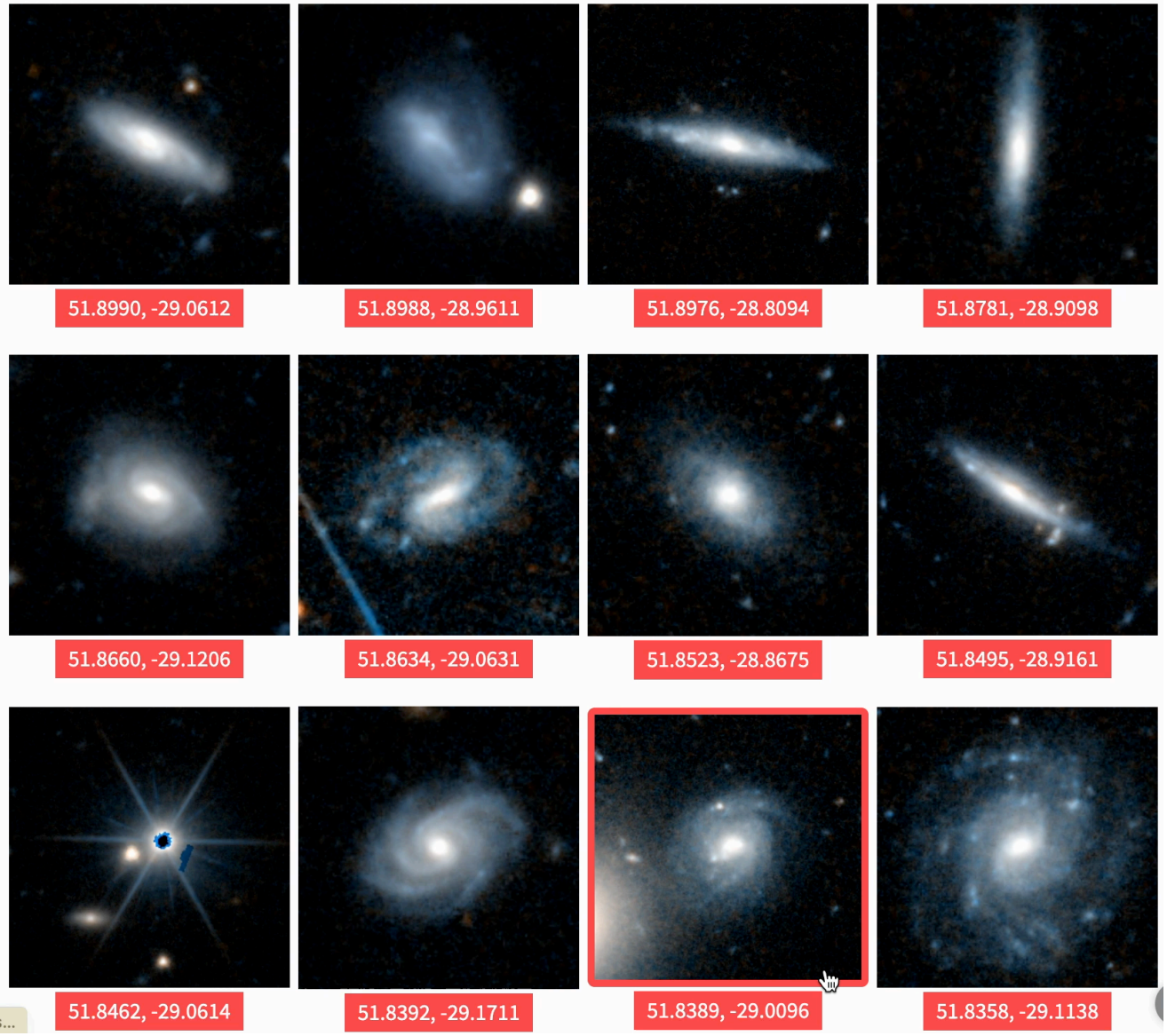
Bar?

Answer

Strong

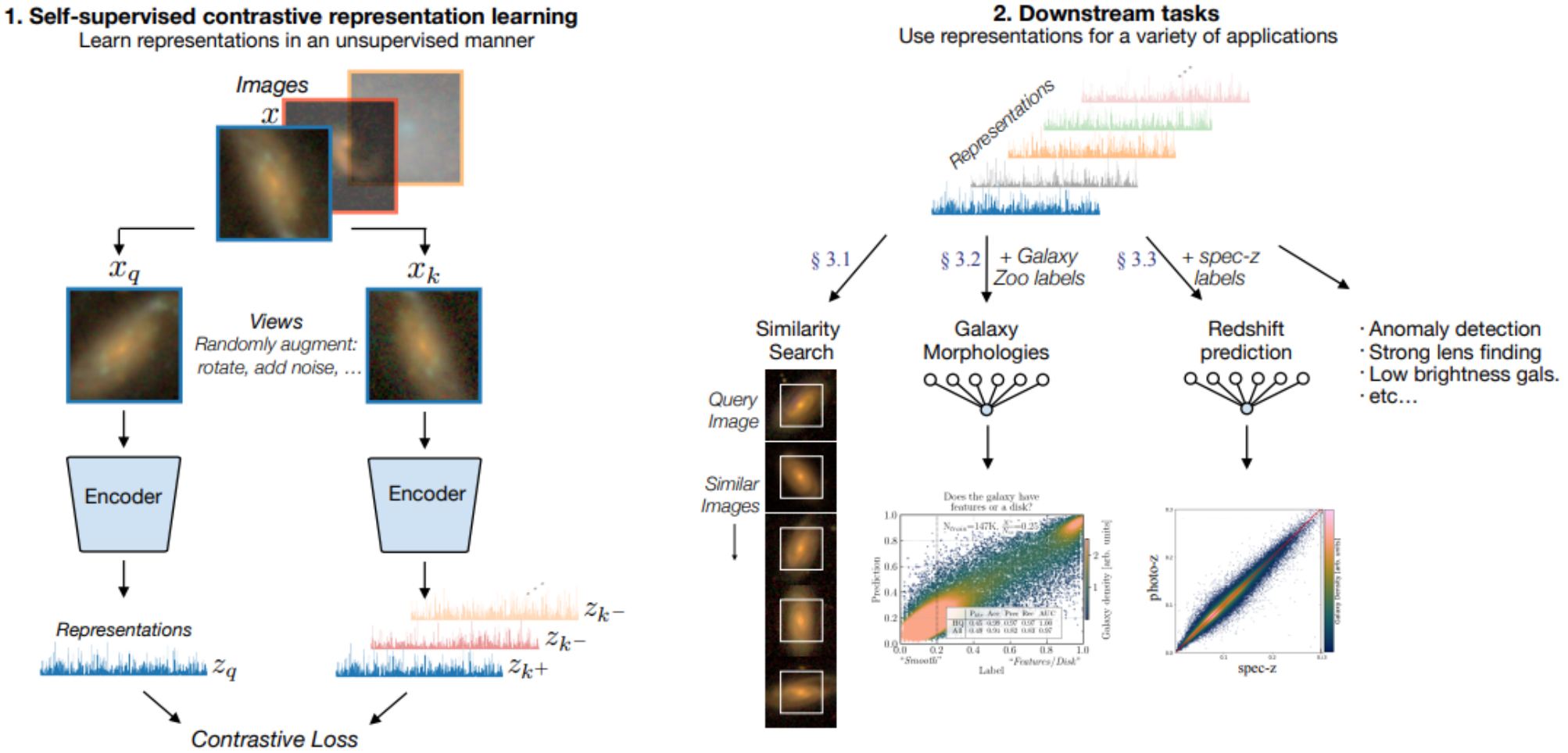
Vote Fraction

0.00 1.00



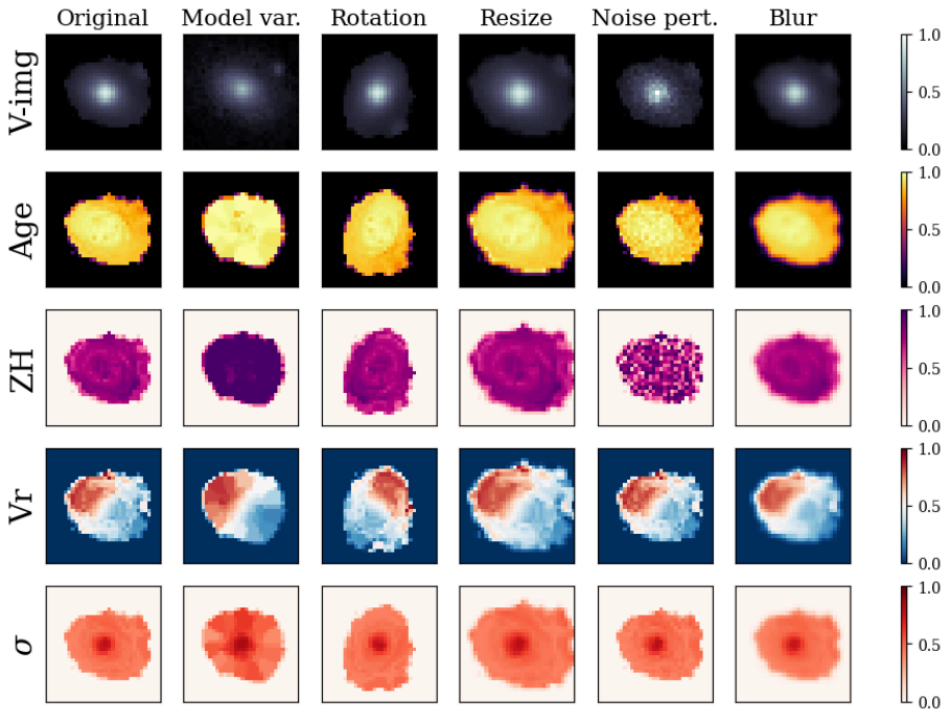
...googleapis.com/.../102042913_NEG518388643290096129_gz_arcsinh_vis...

first applications show good generalisation to downstream tasks



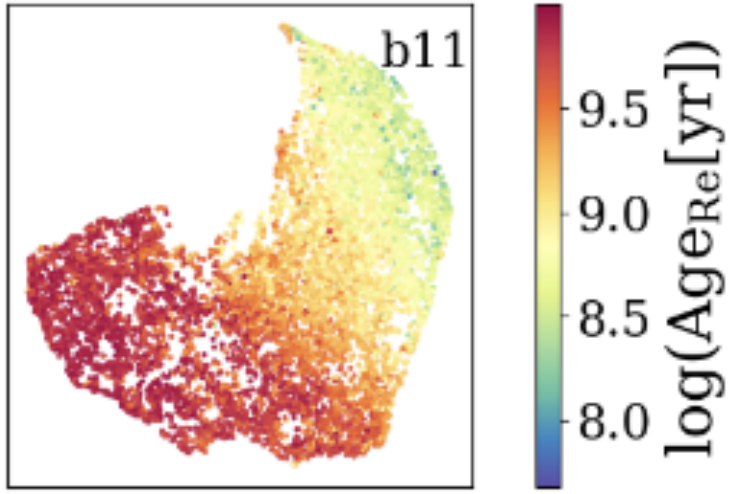
Hayat+21

first applications show informative embeddings

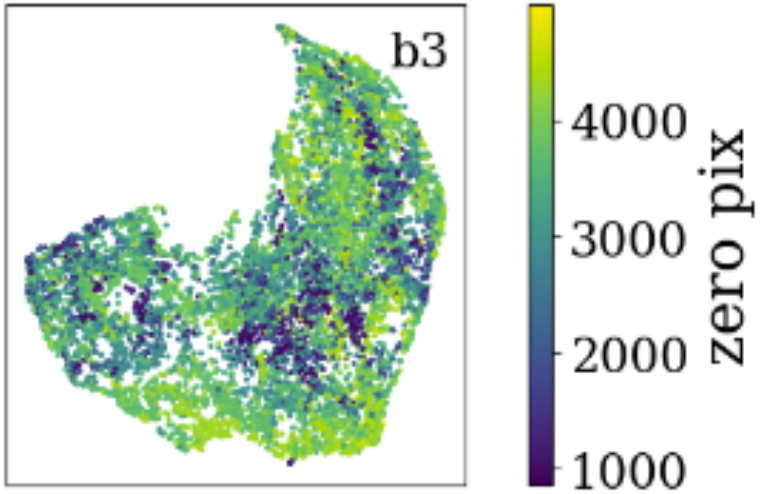


Sarmiento,MHC+21

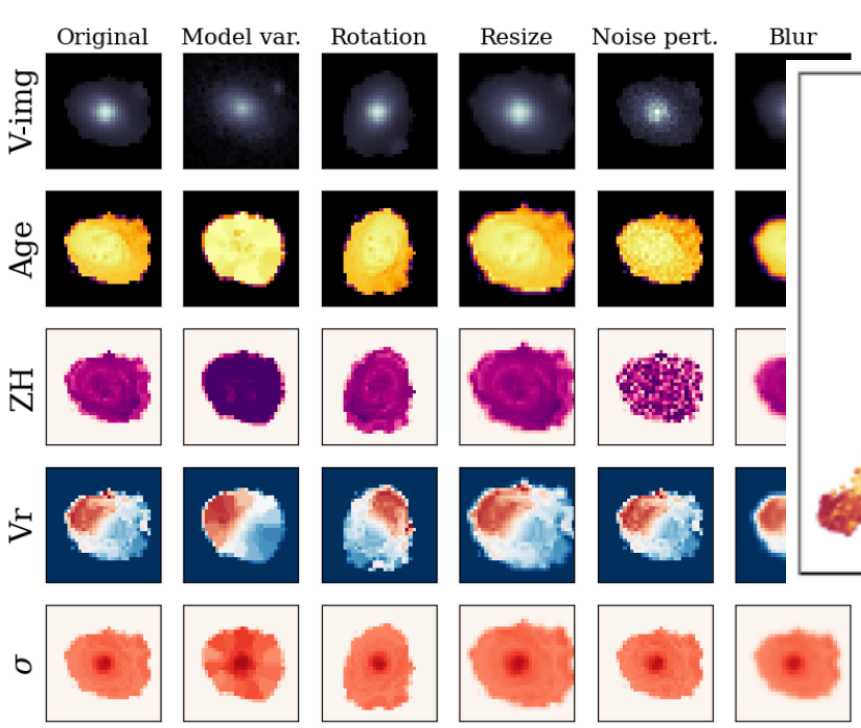
Physical



Instrumental

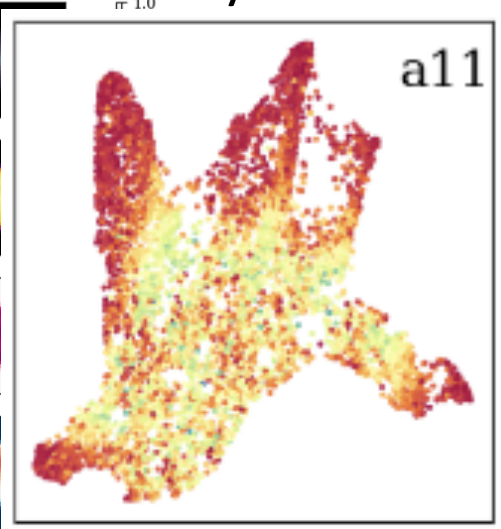


first applications show informative embeddings



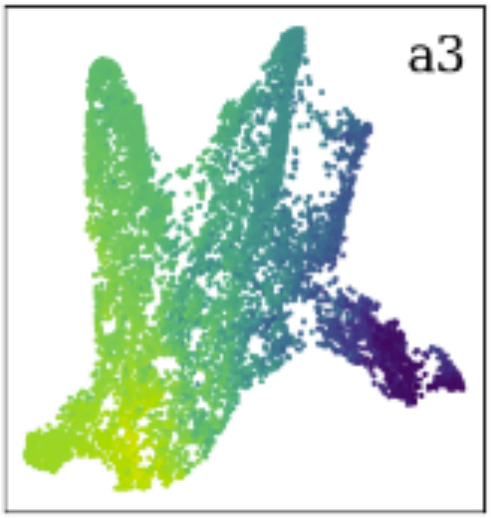
Sarmiento,MHC+21

Physical

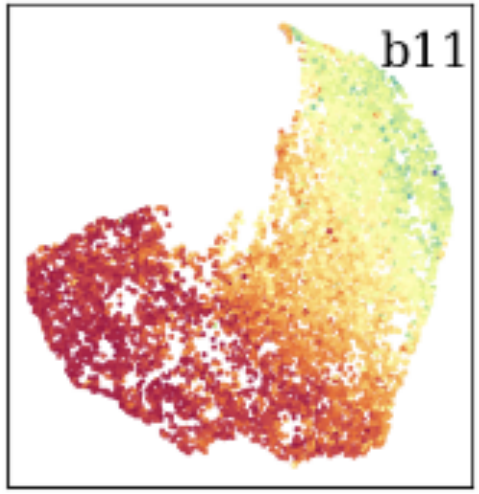


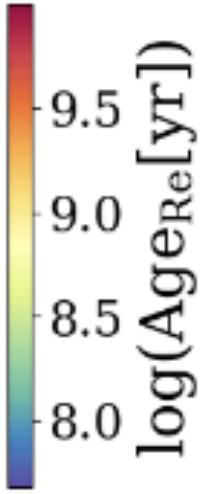
PCA

Instrumental

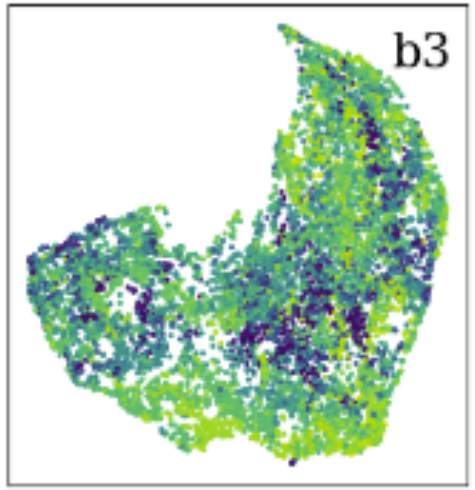


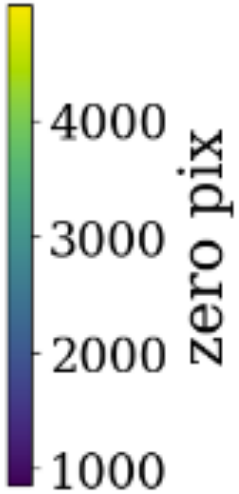
Physical



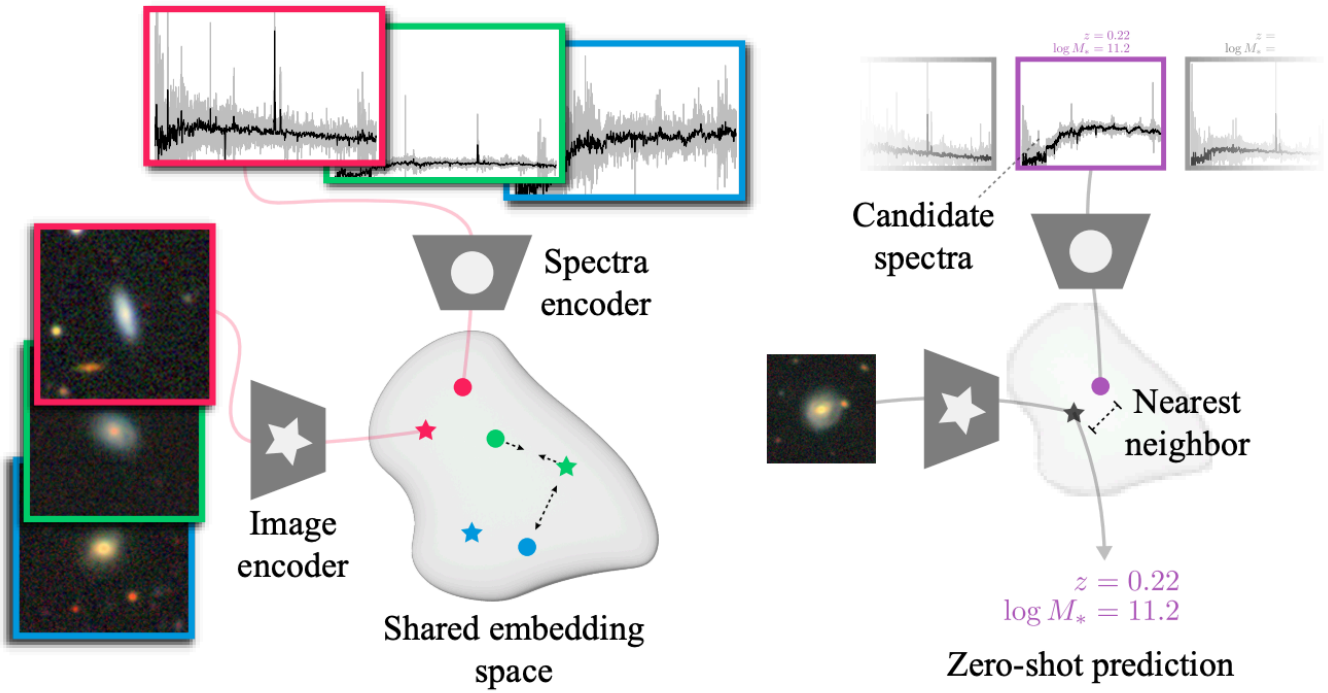


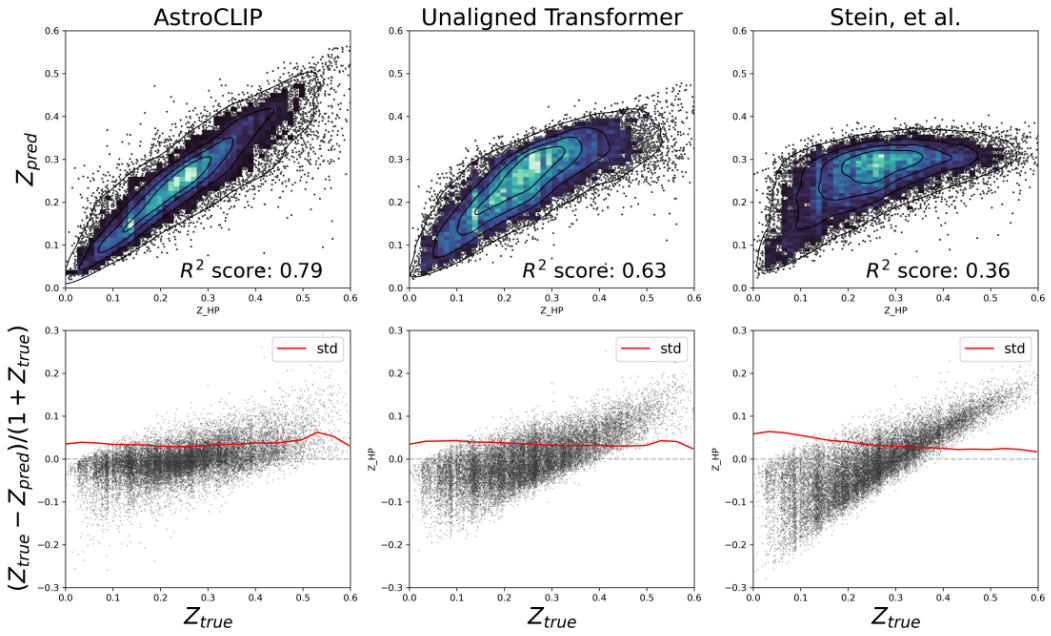
Instrumental





straightforward generalization to two modalities

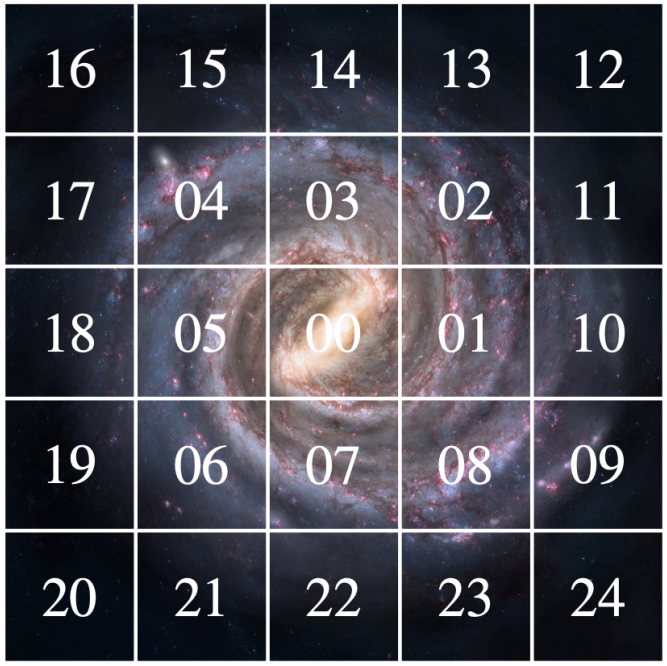


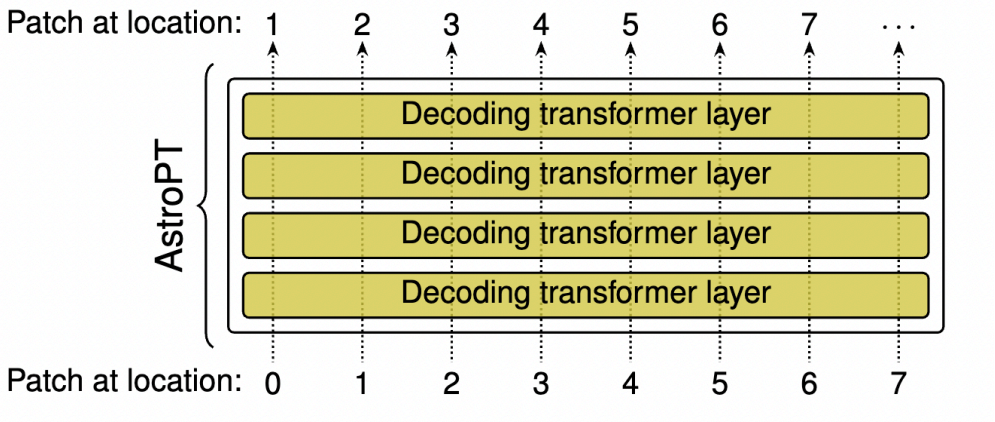


(a) Self-Supervised Zero-Shot Performance.

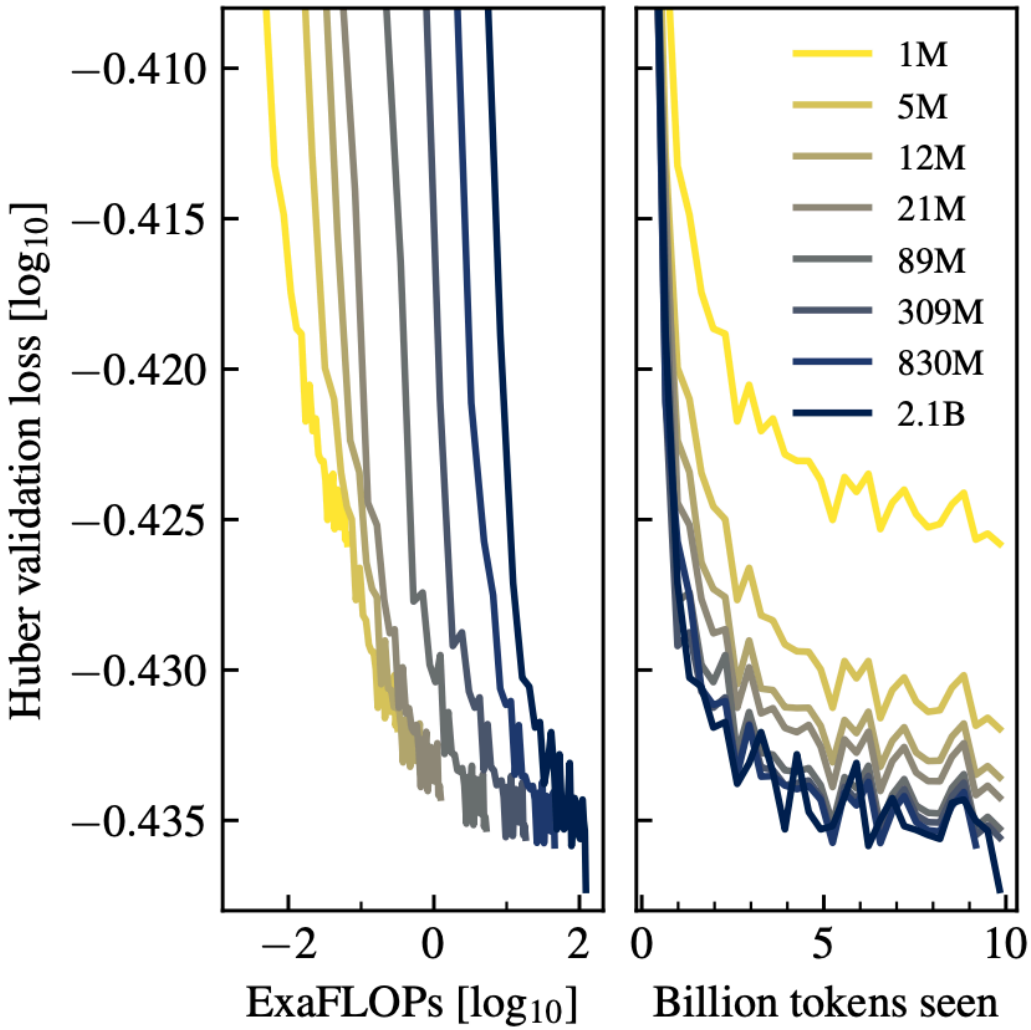
Parker+23

astro-PT: auto-regressive encoding

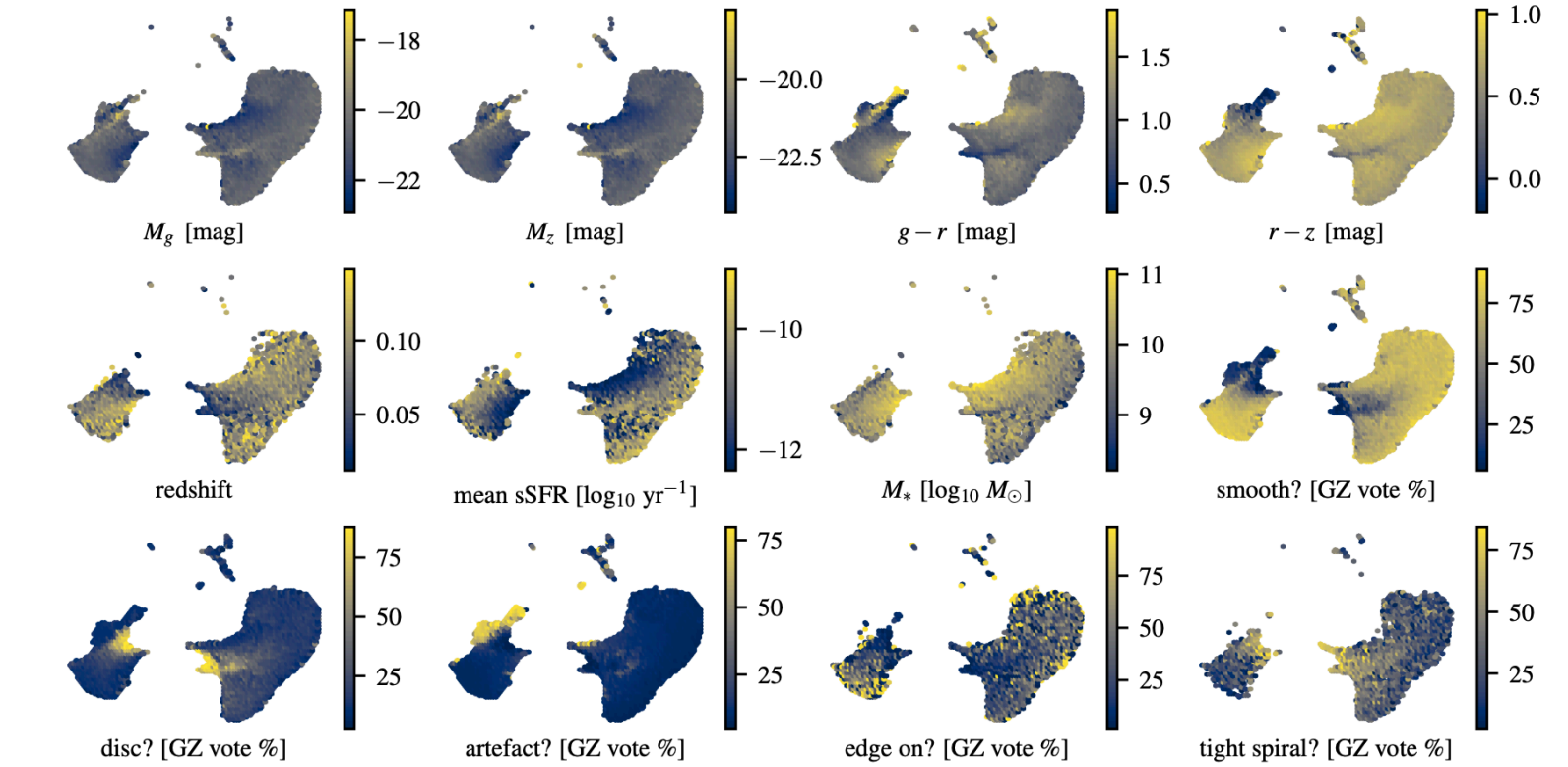


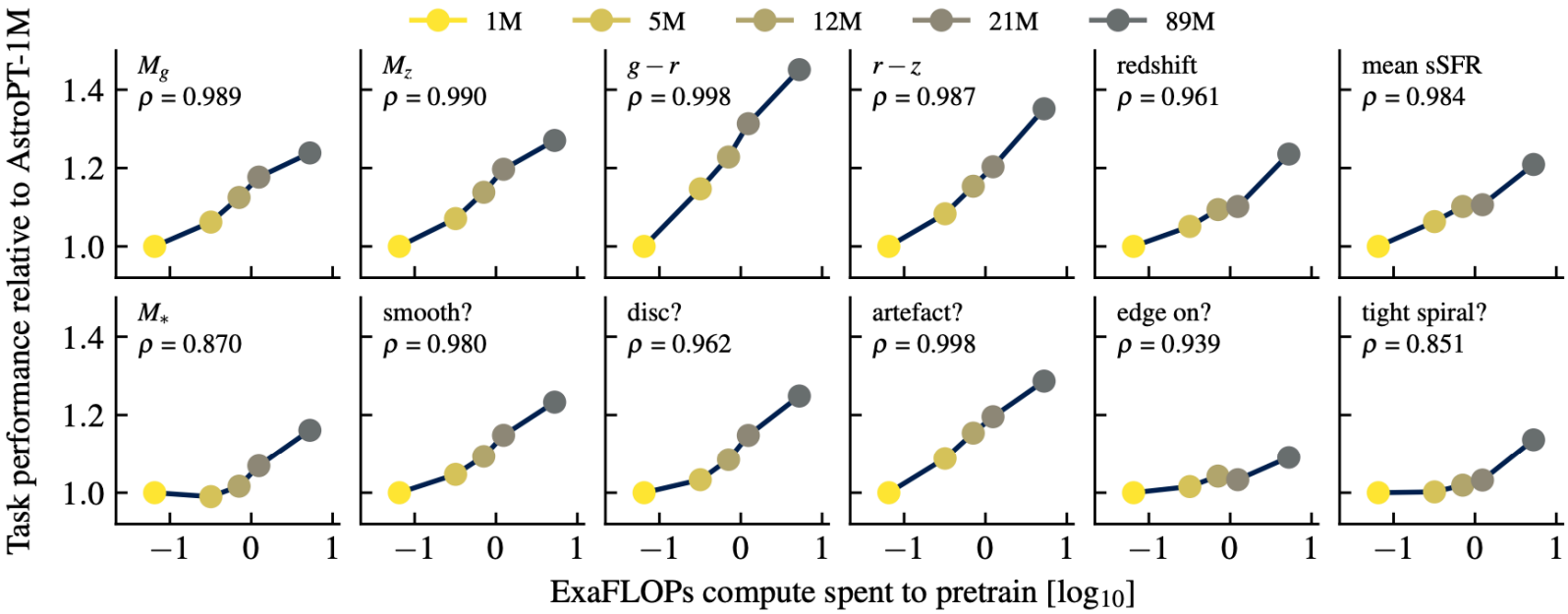


Information content of galaxy images.



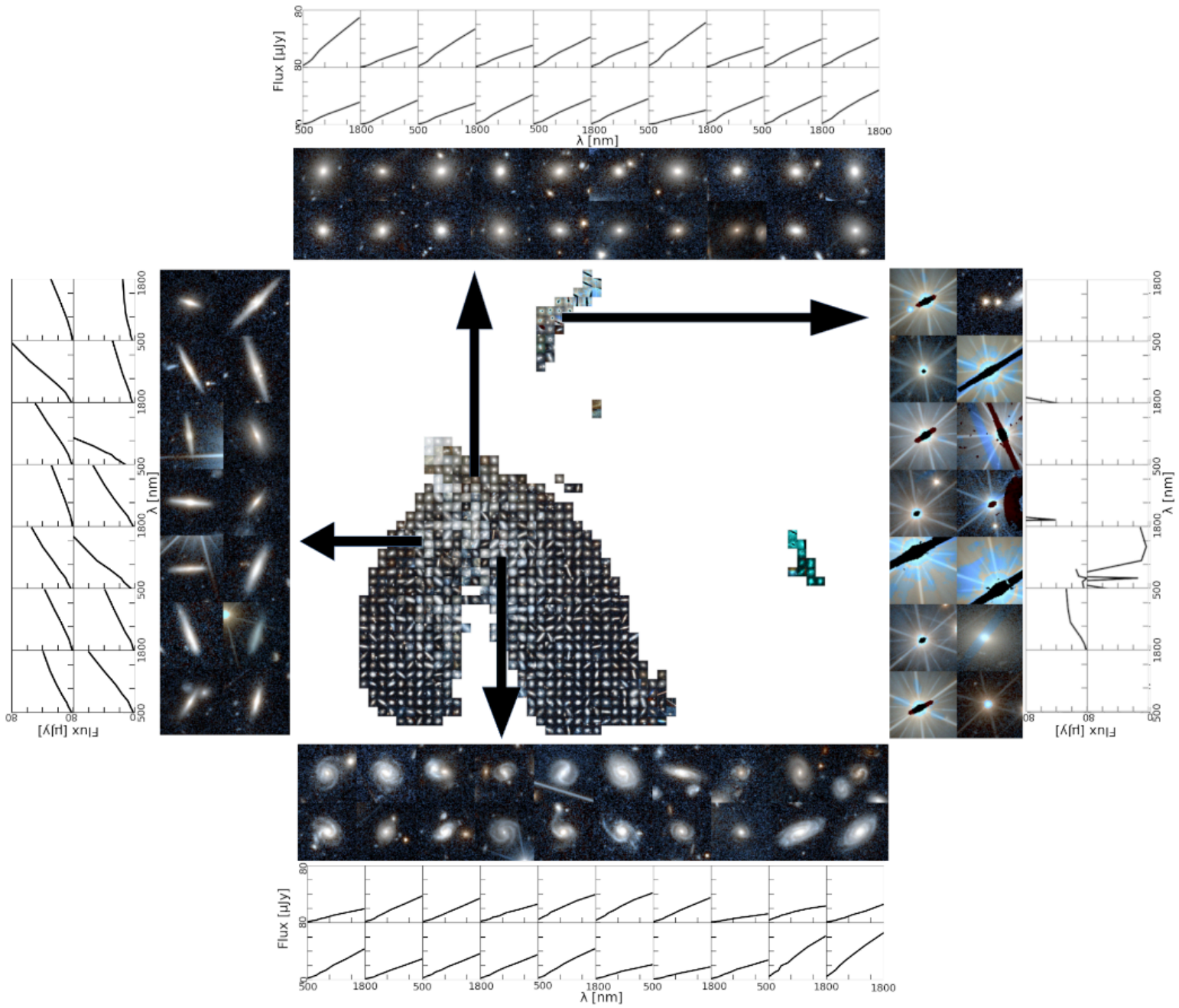
Smith+24


M_g [mag]
M_z [mag]
$g-r$ [mag]
$r-z$ [mag]
redshift
mean sSFR [$\log_{10}$ yr^{-1}]
M_* [$\log_{10} M_\odot$]
smooth? [GZ vote %]
disc? [GZ vote %]
artefact? [GZ vote %]
edge on? [GZ vote %]
tight spiral? [GZ vote %]

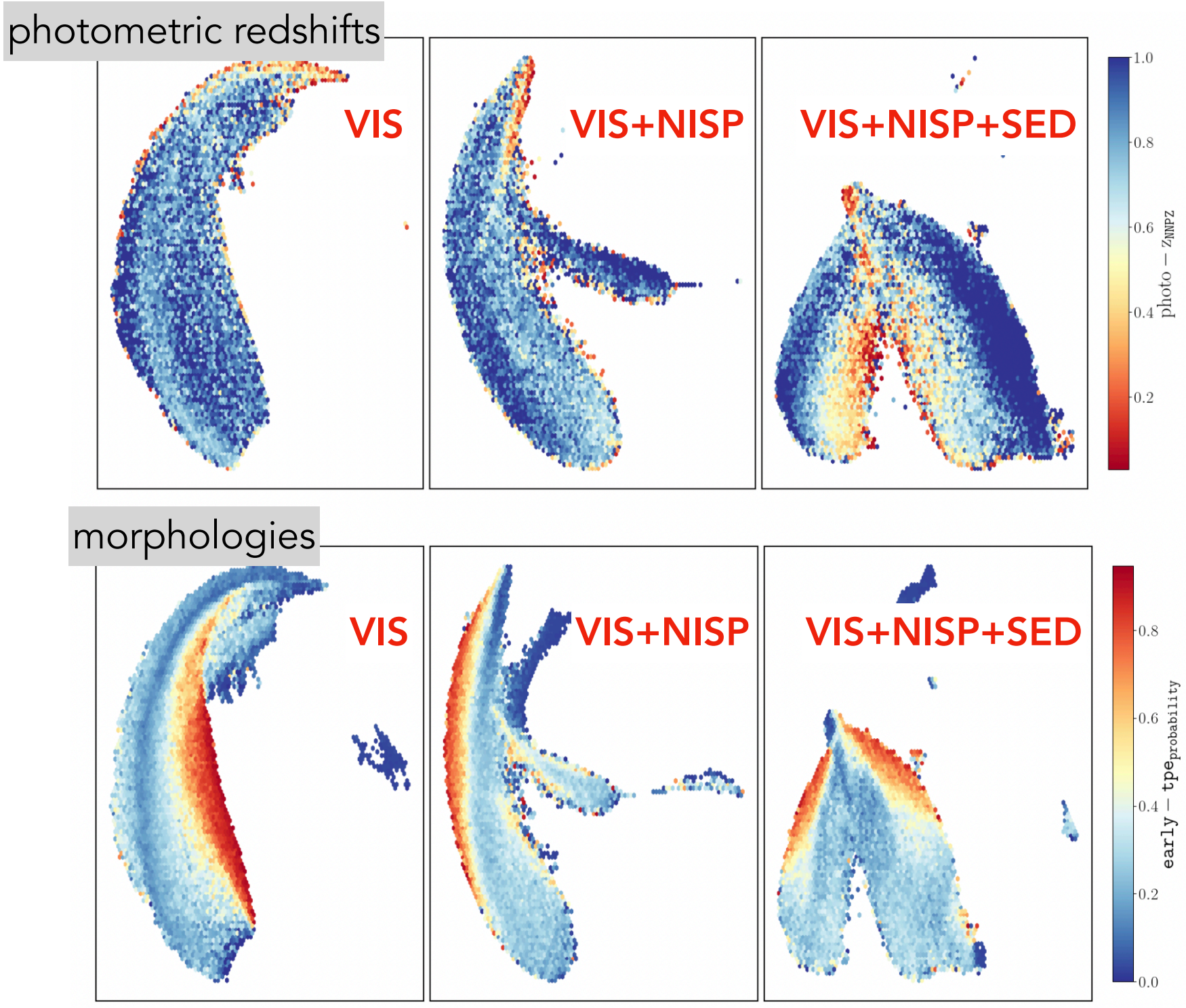

1M
5M
12M
21M
89M
M_g $\rho = 0.989$
M_z $\rho = 0.990$
$g-r$ $\rho = 0.998$
$r-z$ $\rho = 0.987$
redshift $\rho = 0.961$
mean sSFR $\rho = 0.984$
M_* $\rho = 0.870$
smooth? $\rho = 0.980$
disc? $\rho = 0.962$
artefact? $\rho = 0.998$
edge on? $\rho = 0.939$
tight spiral? $\rho = 0.851$
Task performance relative to AstroPT-1M
ExaFLOPs compute spent to pretrain [$\log_{10}$]

Smith+24

astro-PT joint embedding of SEDs and images

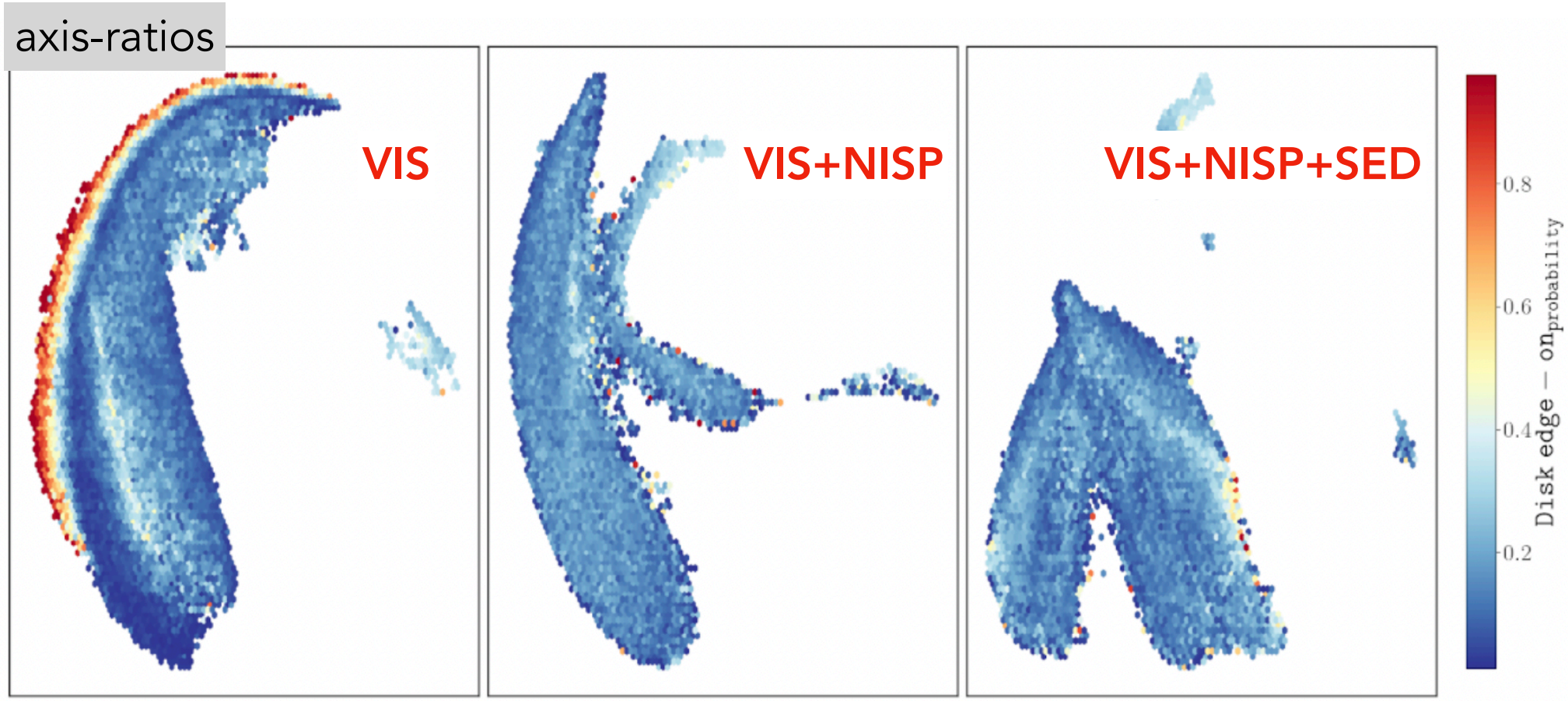


EC: Siudek, MHC+25



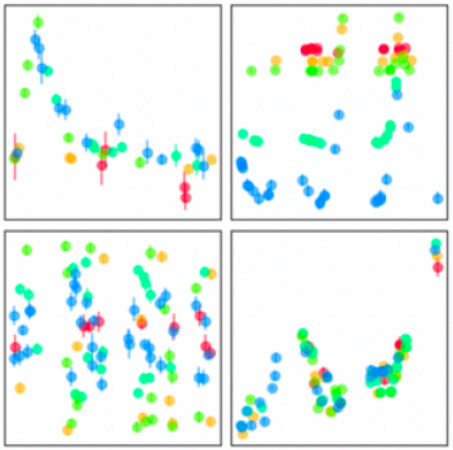
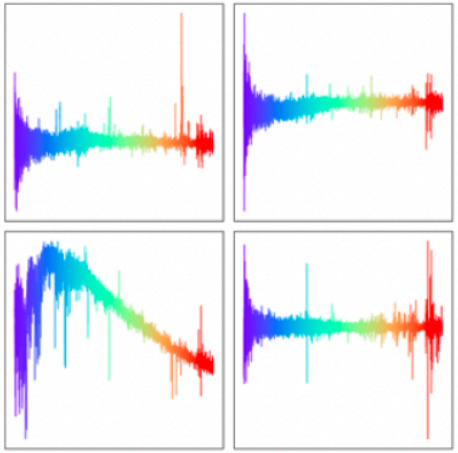
EC: Siudek, MHC+25

Adding information does not always result in more informative embedding spaces



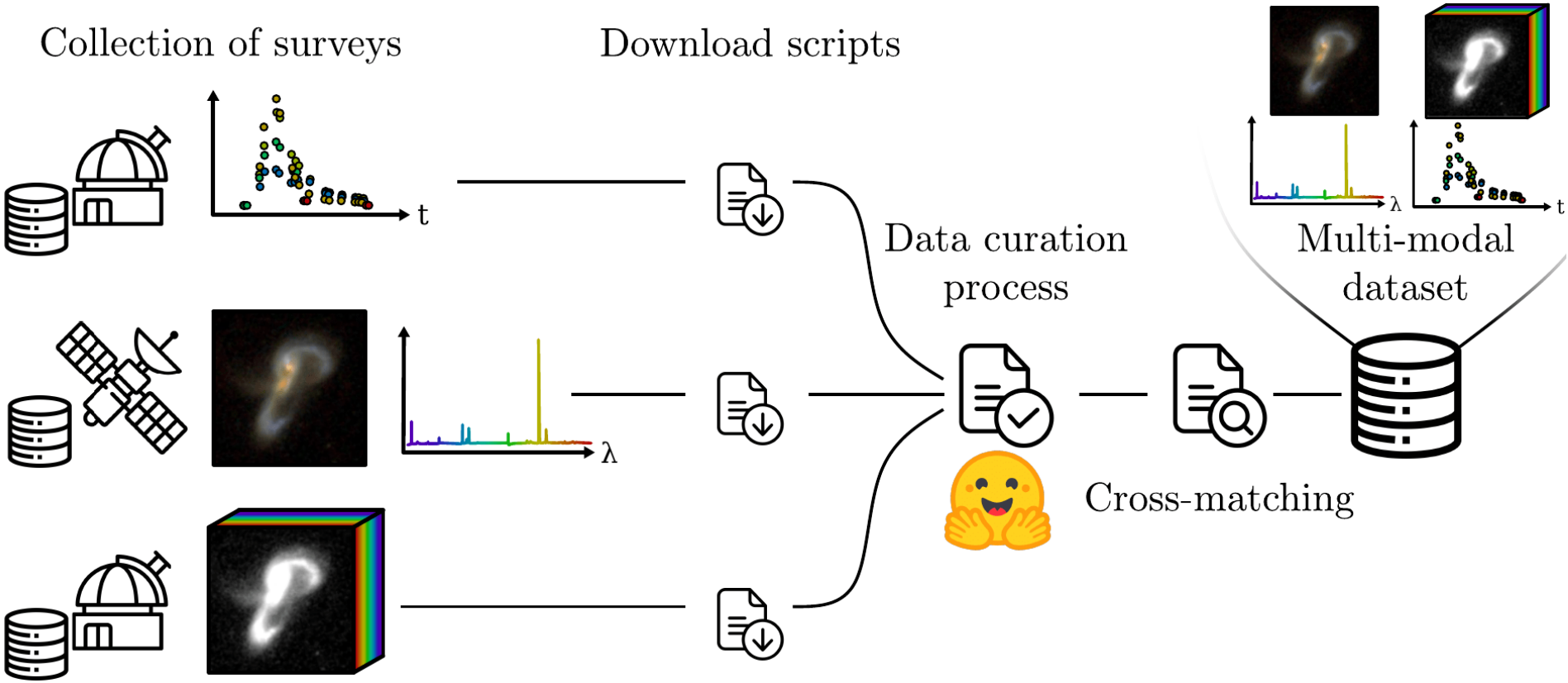
EC: Siudek, MHC+25

The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data

	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			

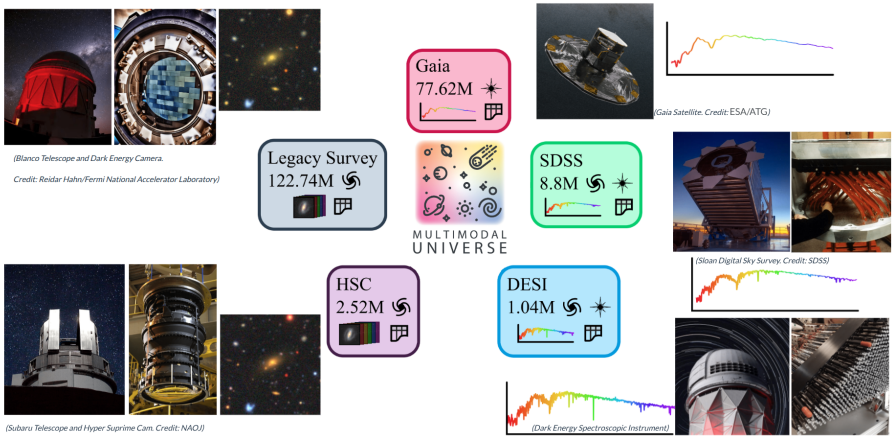
MMU collaboration et al. (2024), NeurIPS dataset track

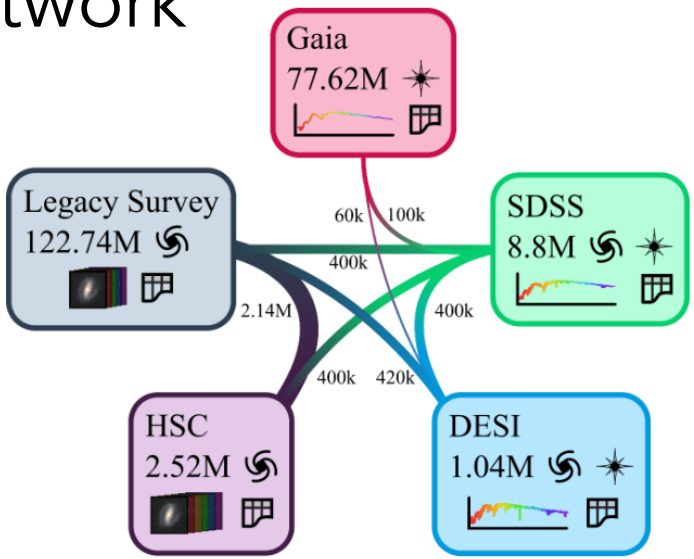
The Multimodal Universe: Enabling Large-Scale Machine Learning with 100 TB of Astronomical Scientific Data

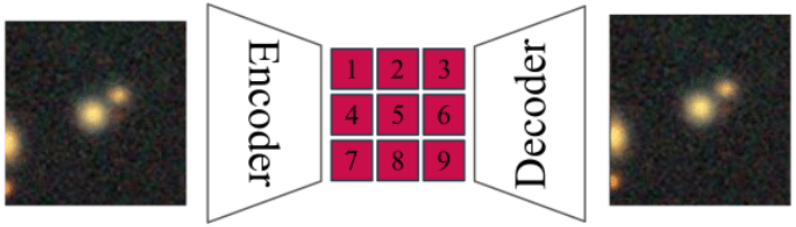


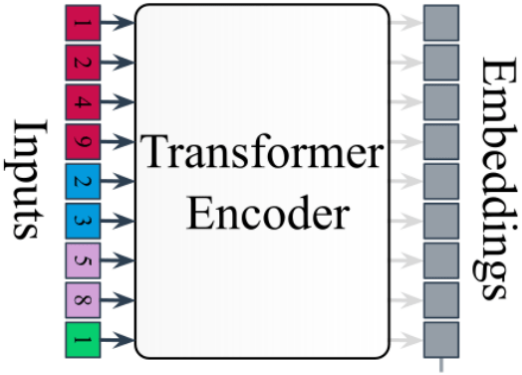
MMU collaboration et al. (2024), NeurIPS dataset track

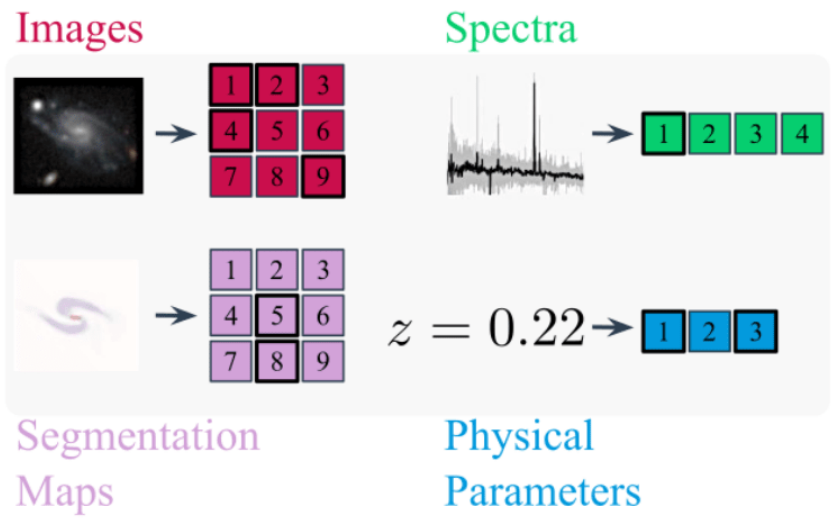
AION-1: Astronomical **Omnimodal** Network (*polymathic team)

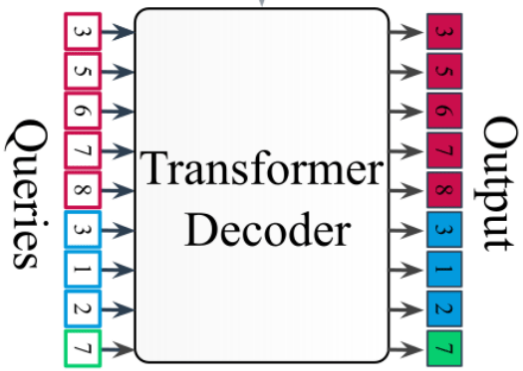






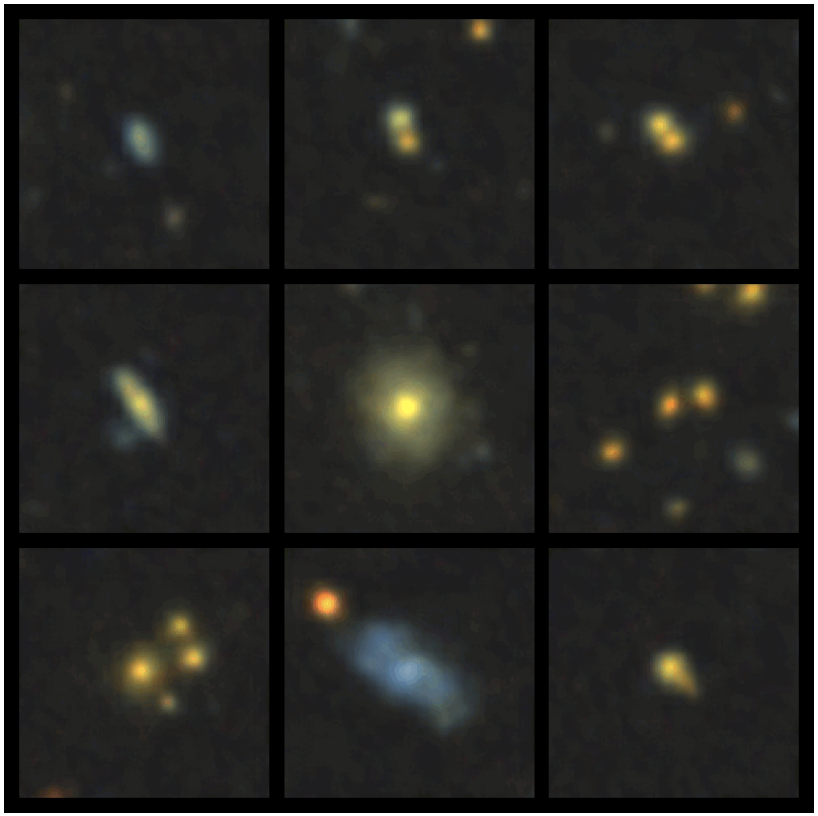


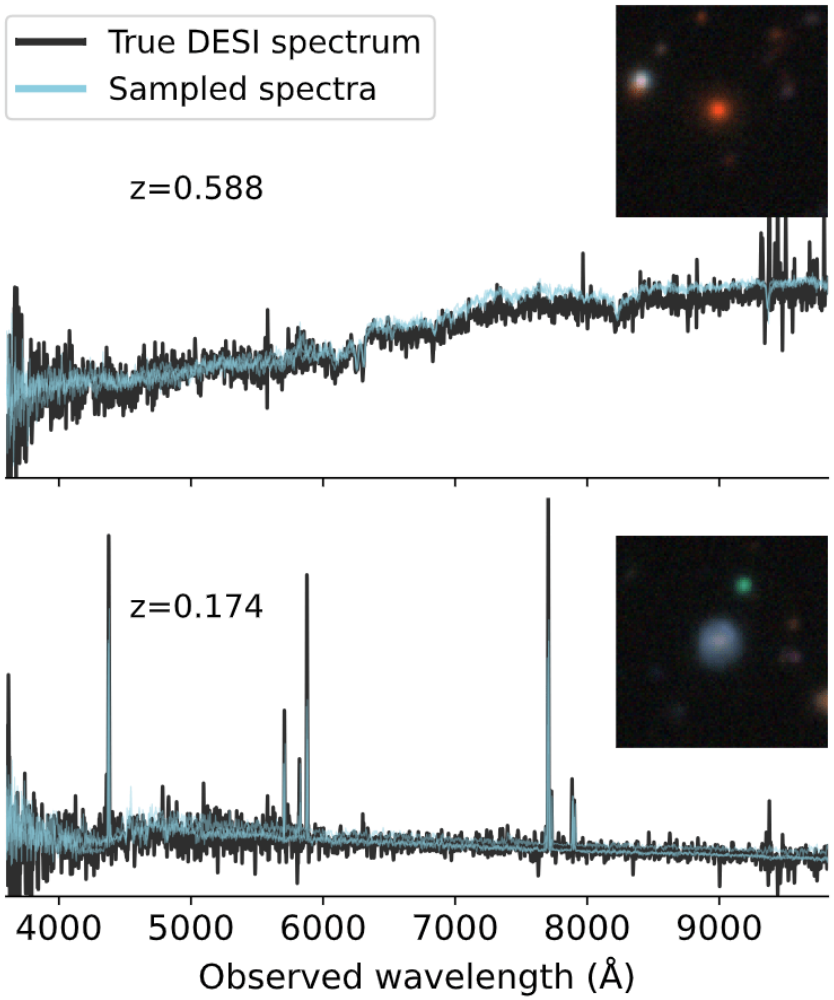




AION-1: Astronomical **Omnimodal** Network

(*polymathic team)

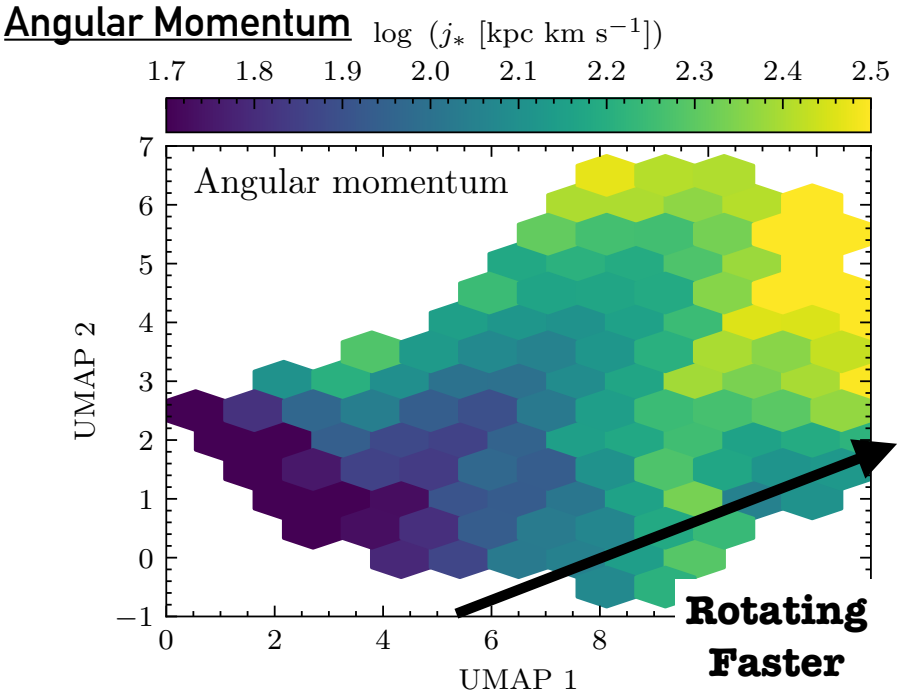




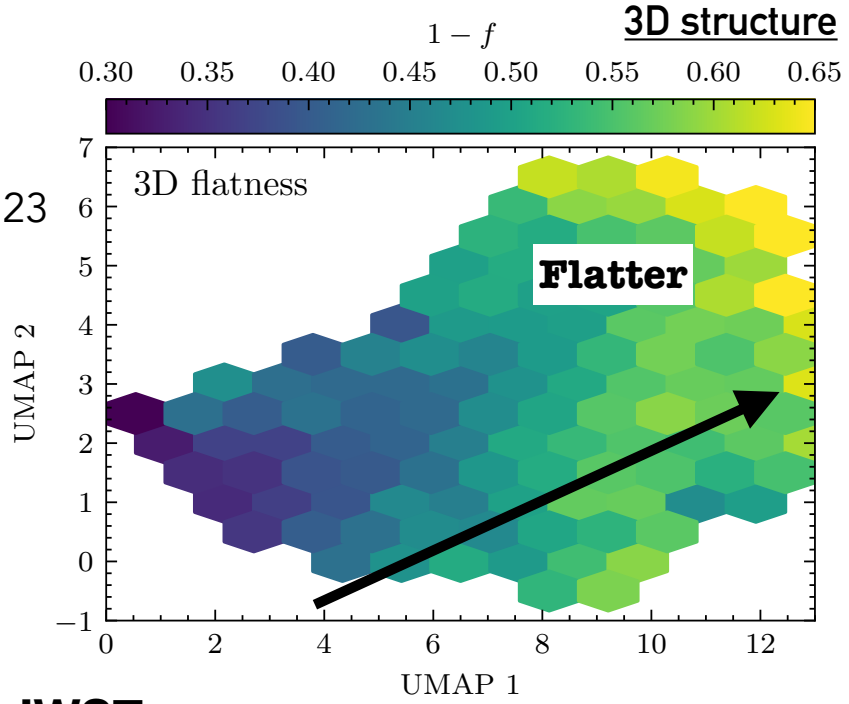
$$p(\boldsymbol{x}_{HSC}|\boldsymbol{x}_{DES})$$

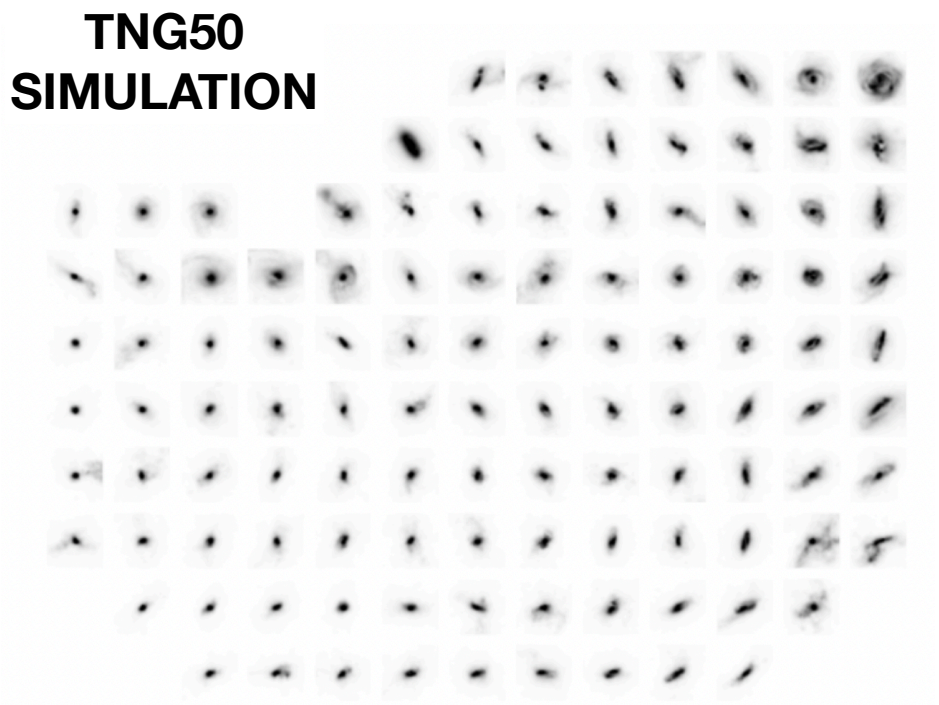
$$p(\boldsymbol{x}_{DESI}|\boldsymbol{x}_{HSC})$$

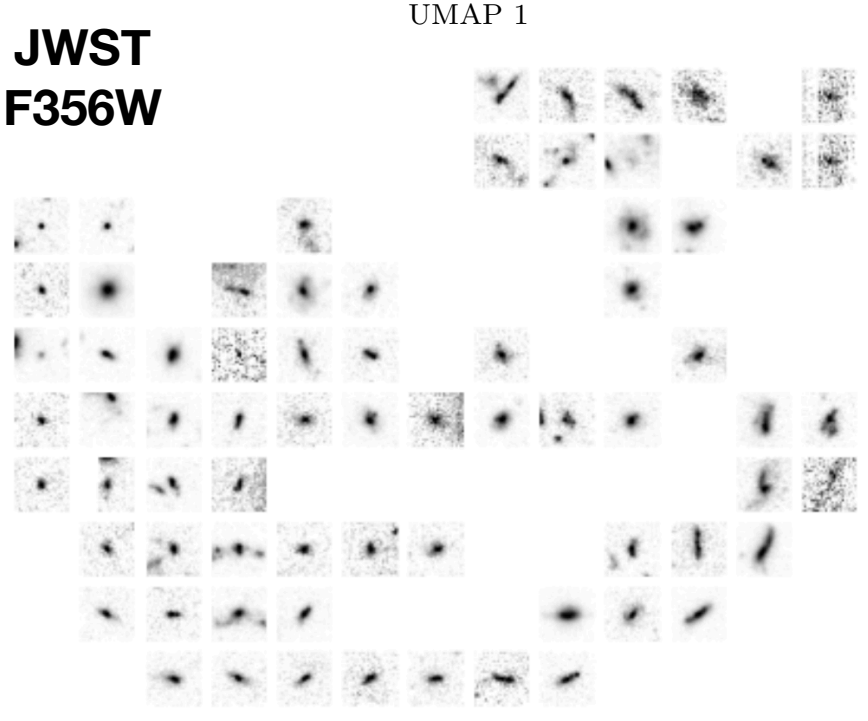
Can we interpret these embedding spaces to learn something about the physics?



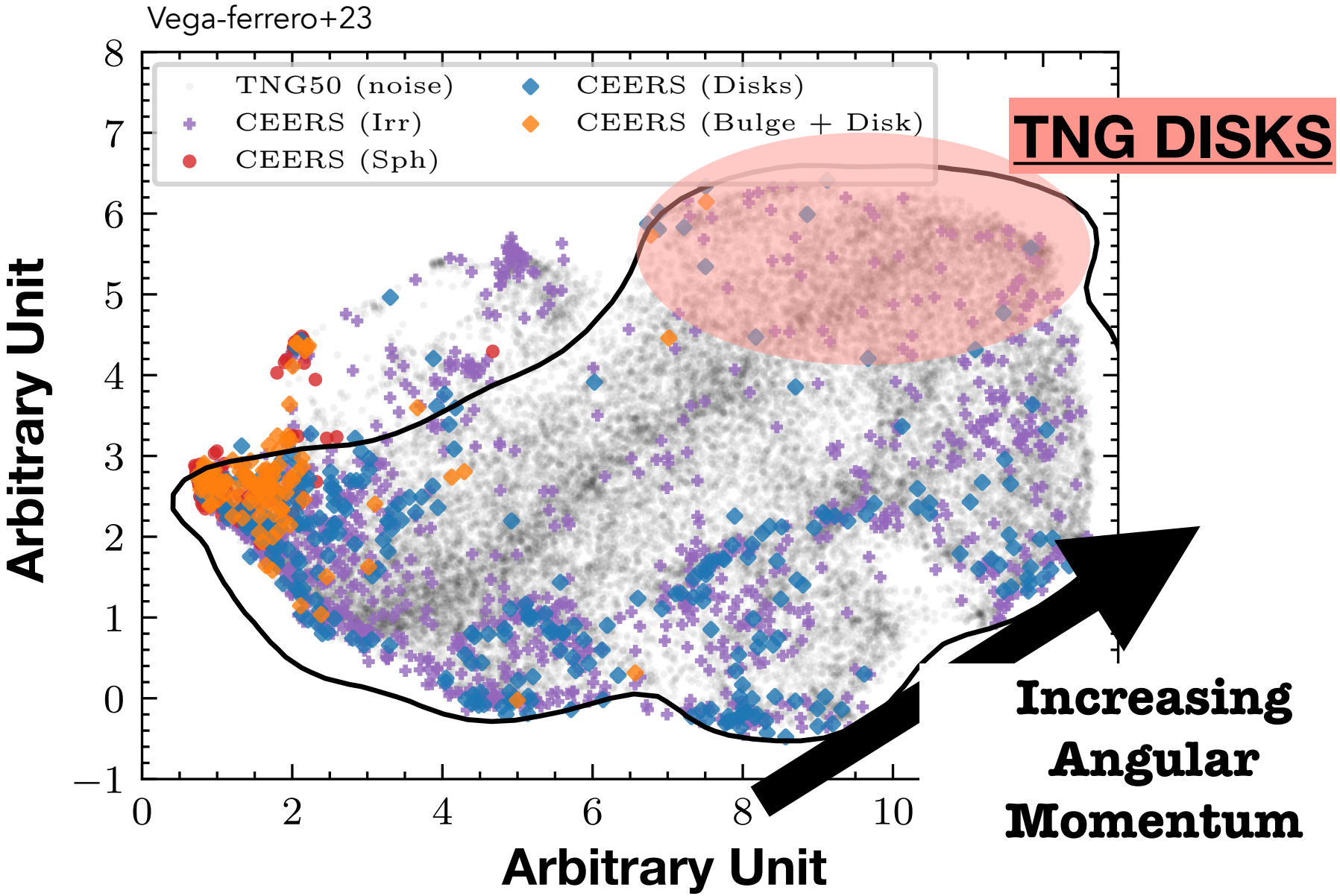
Vega-ferrero+23







projecting simulations and observations in the same embedding space



Color points: JWST observations

Gray points: simulations