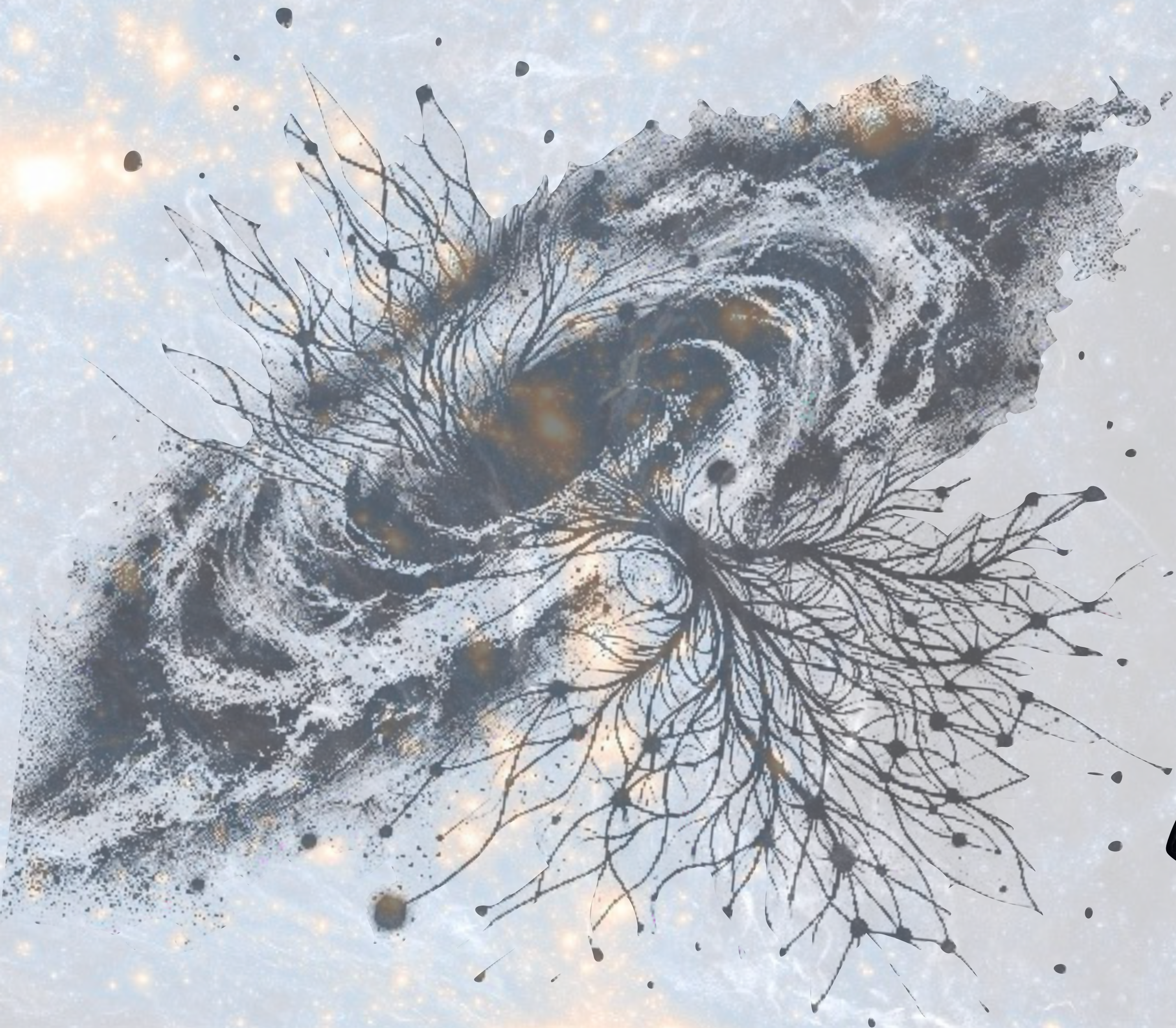
Decoding galaxy properties with machine learning and simulation based inference



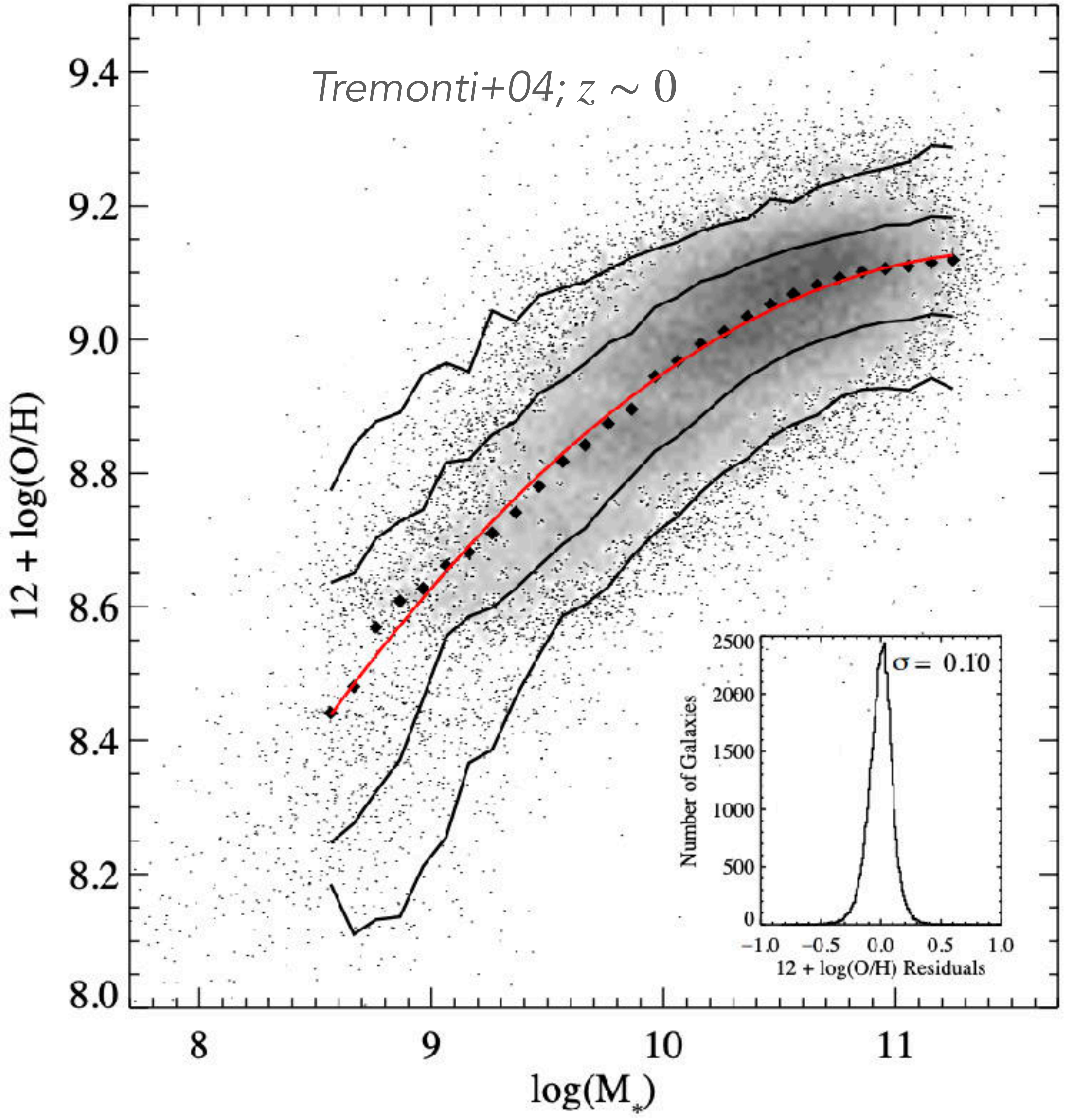
Michele Ginolfi — UniFi & INAF

[arXiv:2410.22420; arXiv:2410.16370; arXiv:2502.20448]

UNIVERSITÀ DEGLI STUDI FIRENZE

INAF ISTITUTO NAZIONALE DI ASTROFISICA

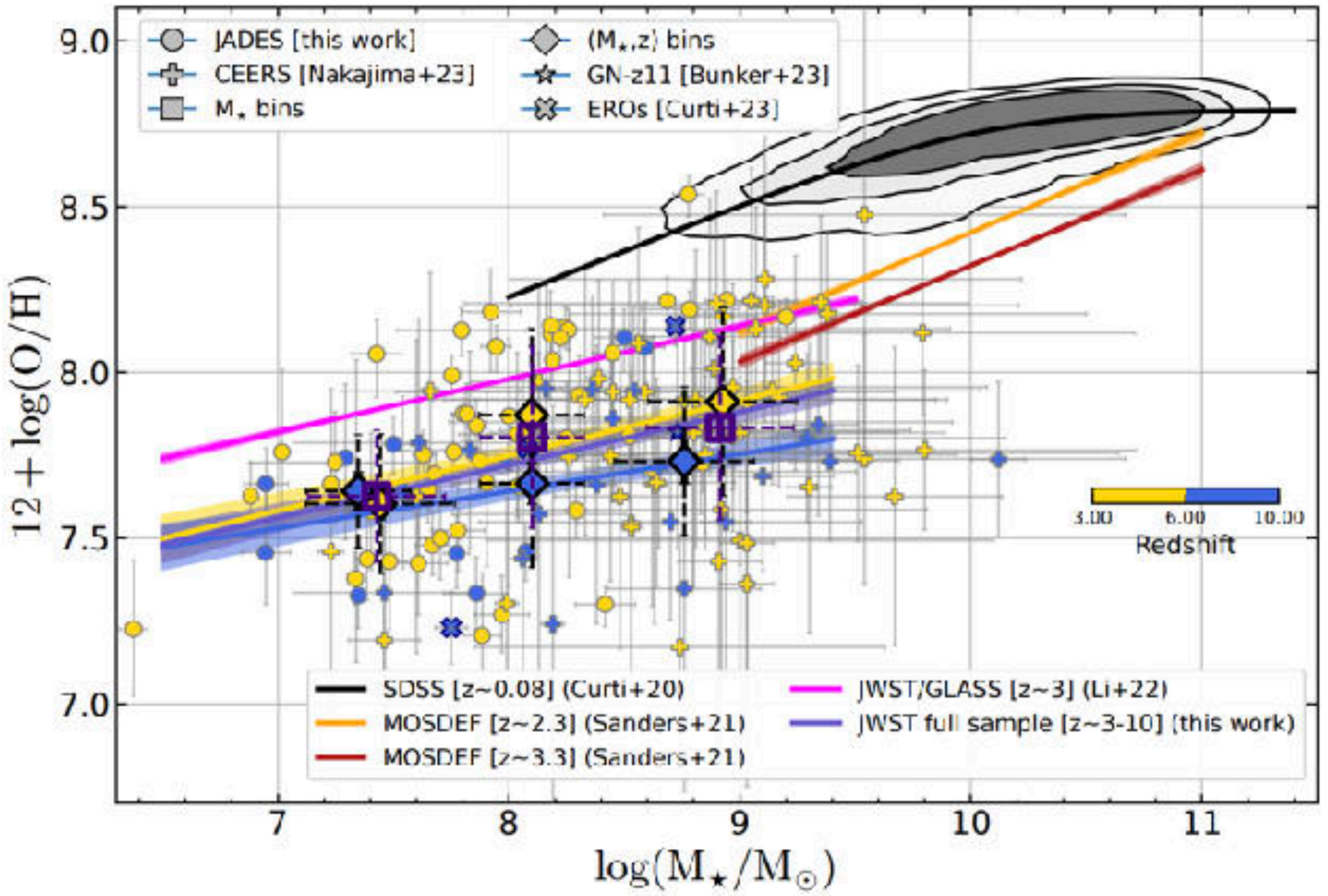
$>$ 50.000 galaxies from SDSS
(>2.000.000 galaxy spectra)



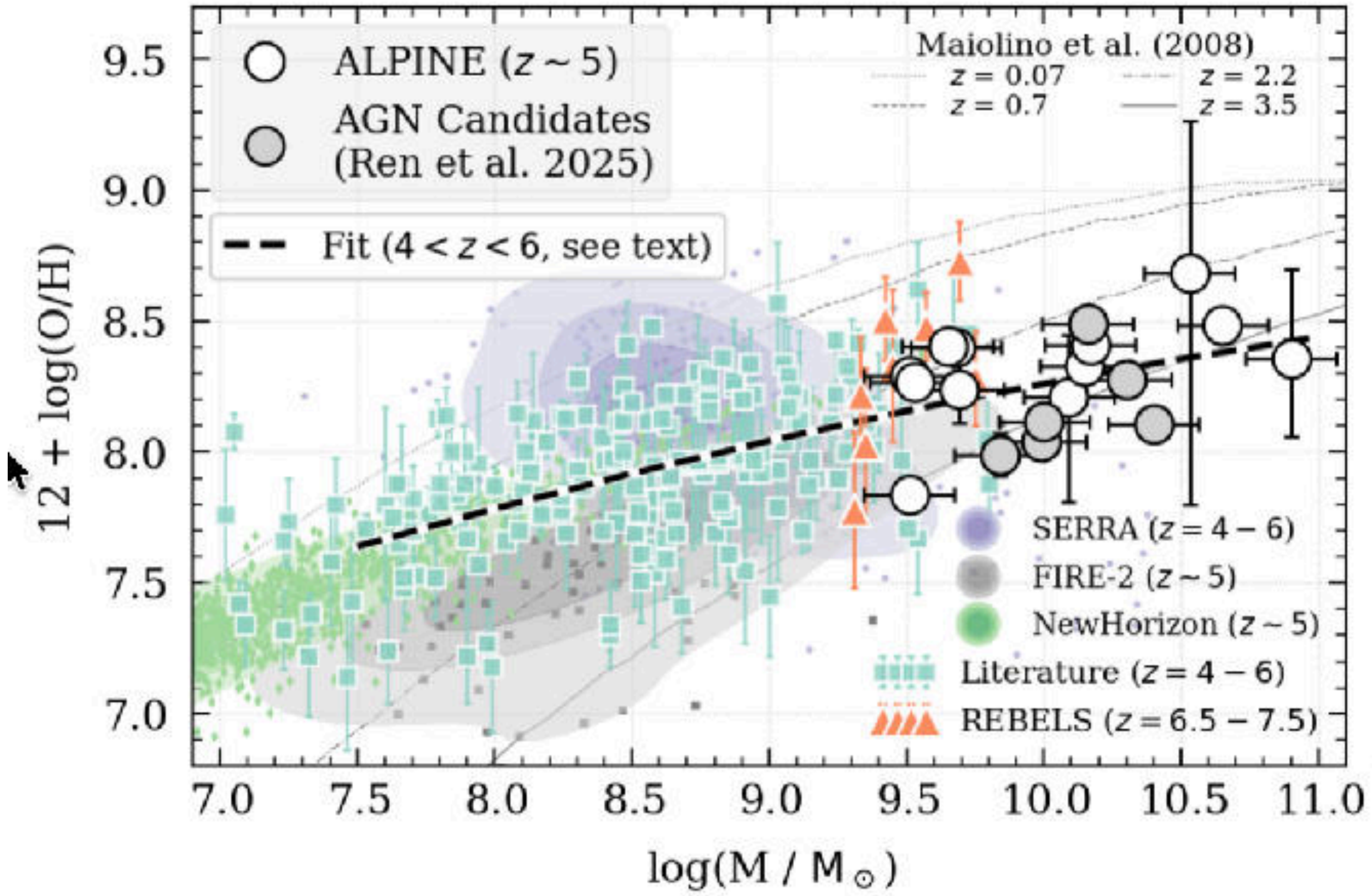
Few years ago

Today

$z \sim 0 - 10$



Curti+24



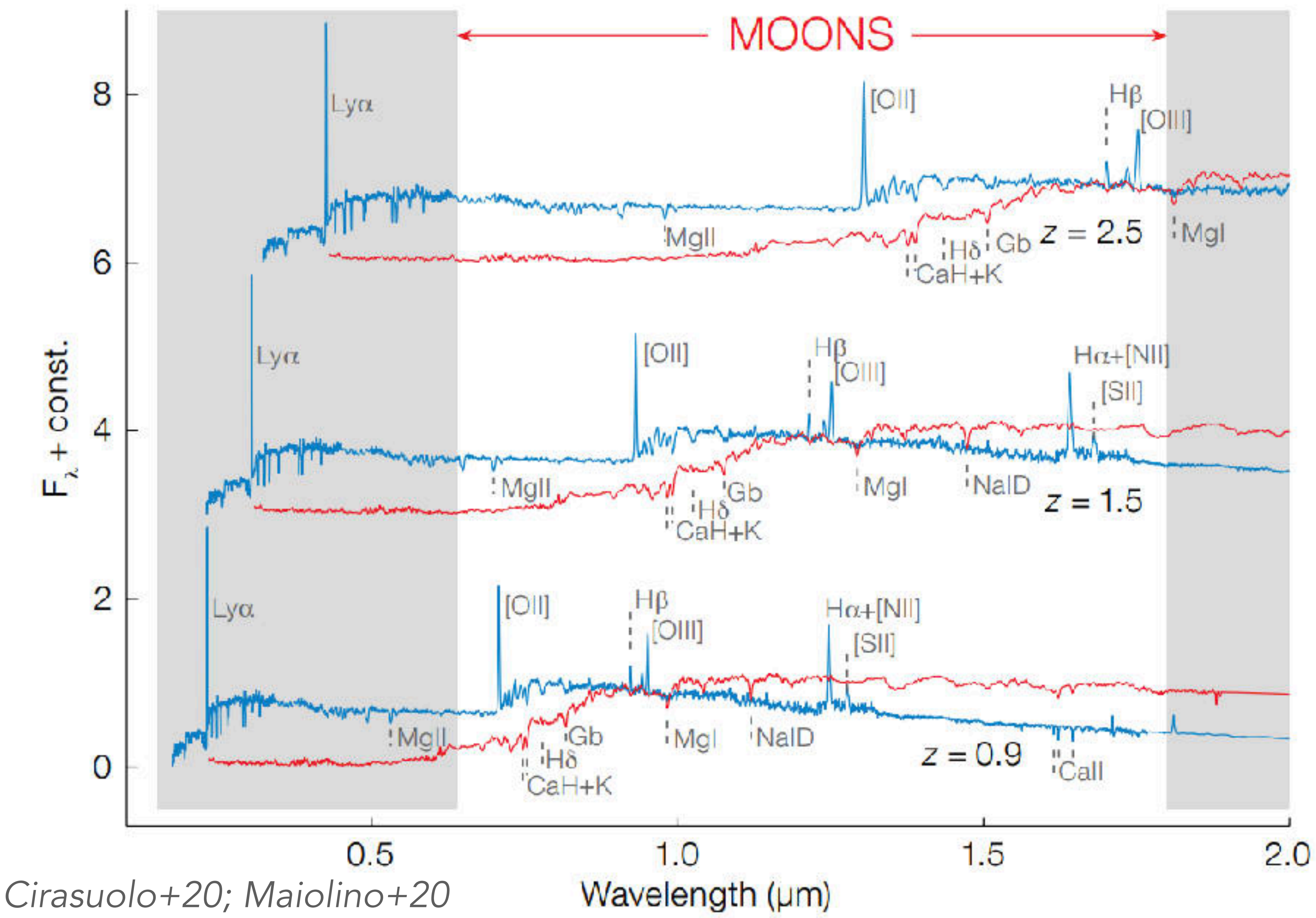
Faisst+25

Ongoing & upcoming all-sky spectroscopic surveys (DESI, 4MOST, **MOONS**)

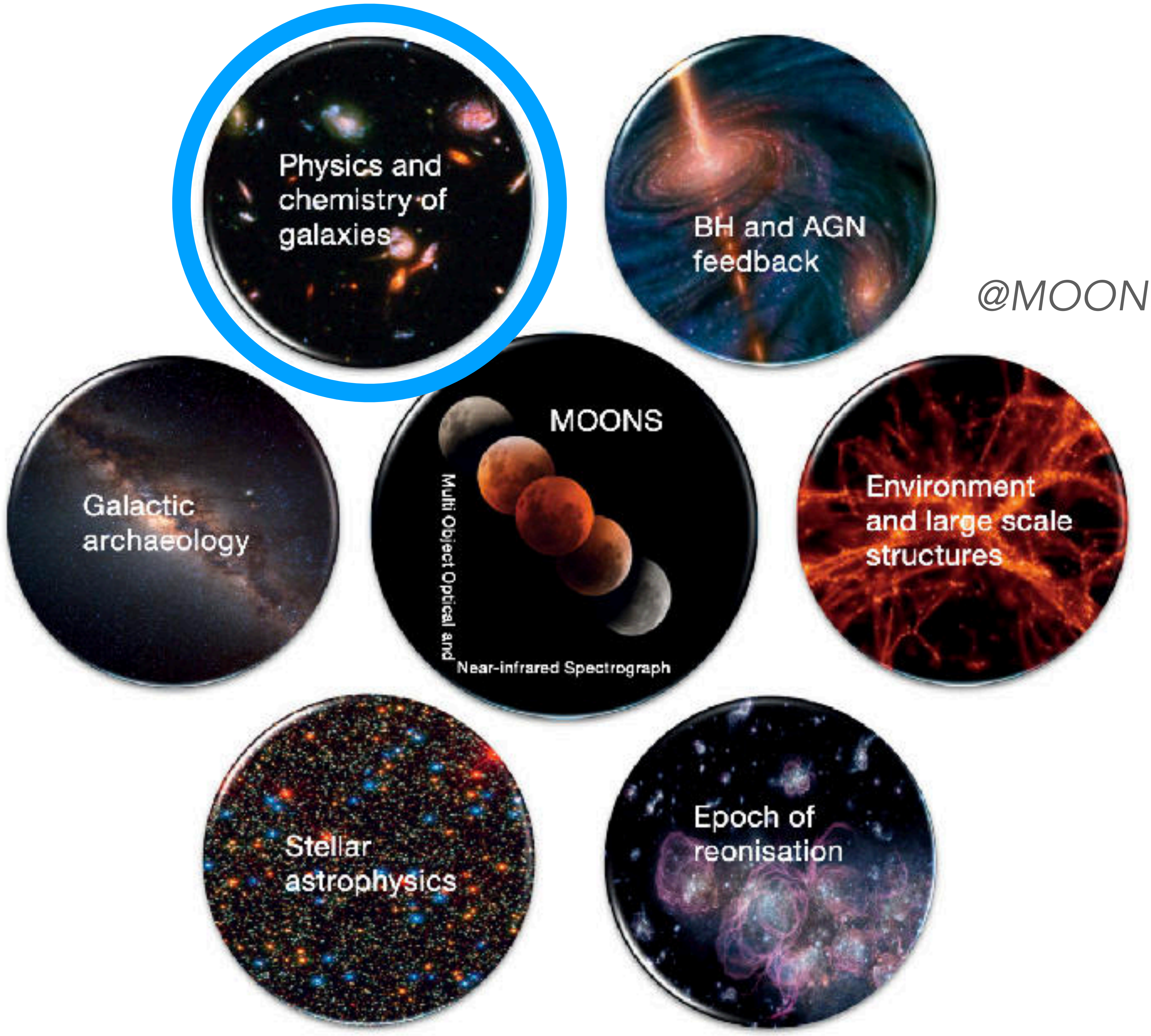
hundreds of millions of spectra will be acquired over the next half-decade!

MOONS is the new **Multi-Object Optical and Near-infrared Spectrograph**, soon to be operated @VLT, ESO

- 1000 fibres, over a field of view of $\sim 500^2$ arcmin;
- low- (R~4000–7000) / high-resolution (~19000 in *H*);
- 0.64 – 1.8 μm wavelength range.



Cirasuolo+20; Maiolino+20


Physics and chemistry of galaxies
BH and AGN feedback
Galactic archaeology
MOONS
Multi Object Optical and Near-infrared Spectrograph
Environment and large scale structures
Stellar astrophysics
Epoch of reonisation

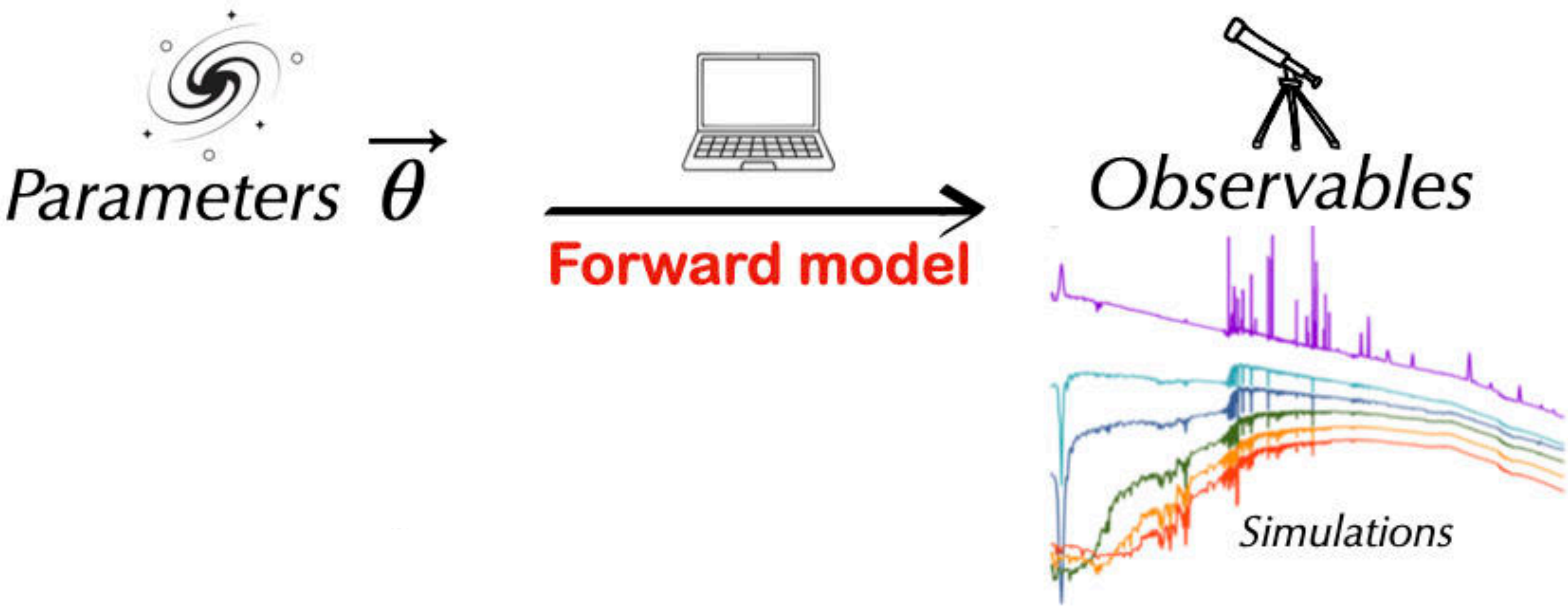
@MOONS collab

A data challenge!

- up to about **half a million galaxies** at $0.9 < z < 2.6$
- >12000 elements per spectrum in low-resolution!
- **Standard fitting methods are typically slow and often fail in weak-signal regimes.**

Mini bibliography — Taylor+18; Maiolino+20, Cirasuolo+20; Looser+21; Gonzales+22; Cabral+22; Gonzales+23; etc etc

How do we infer physical properties?



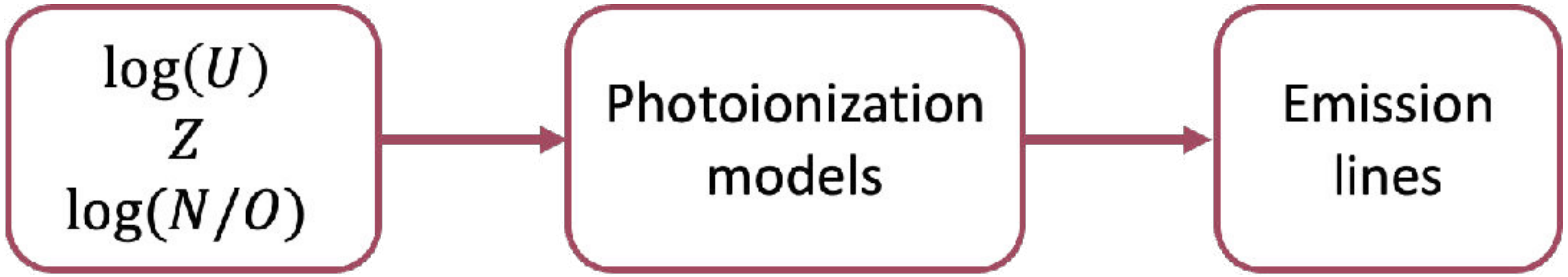
Parameters $\vec{\theta}$

Forward model

Observables

Simulations

e.g.,



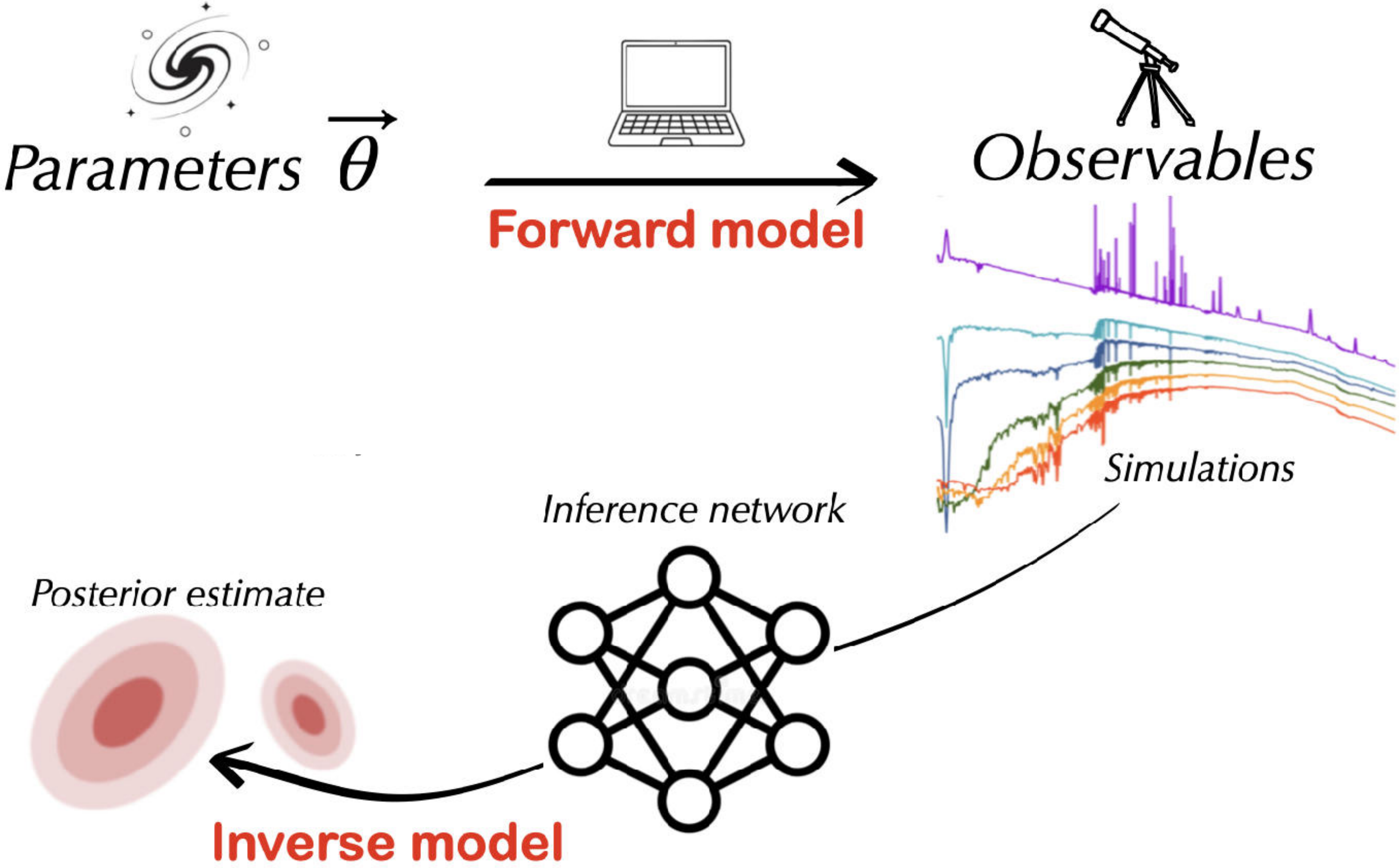
$\log(U)$
Z
$\log(N/O)$

Photoionization models

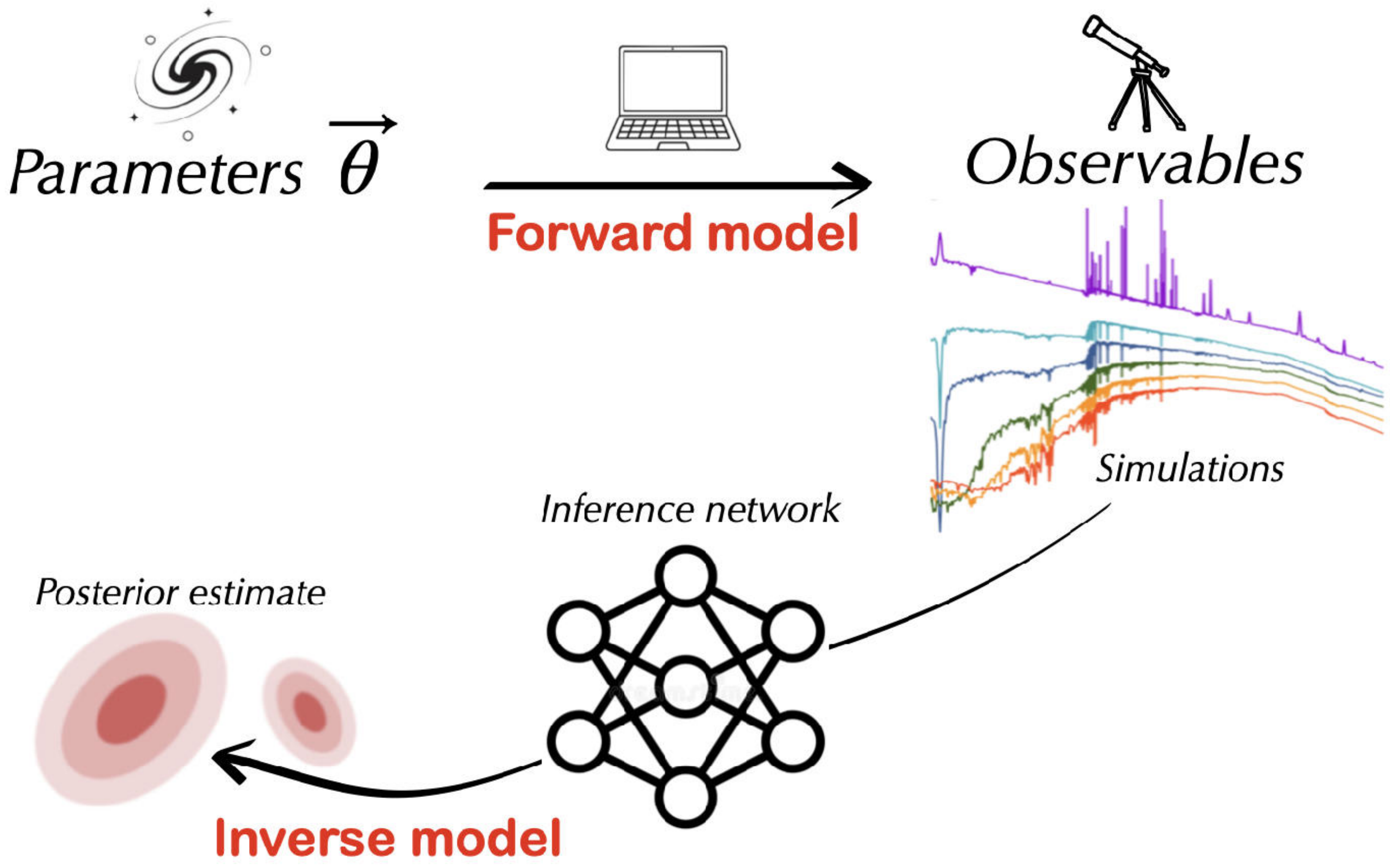
Emission lines

We need to solve the "inverse problem" ⬅️

Simulation-based inference (SBI) with neural estimators



Simulation-based inference (SBI) with neural estimators



Parameters $\vec{\theta}$

Forward model

Observables

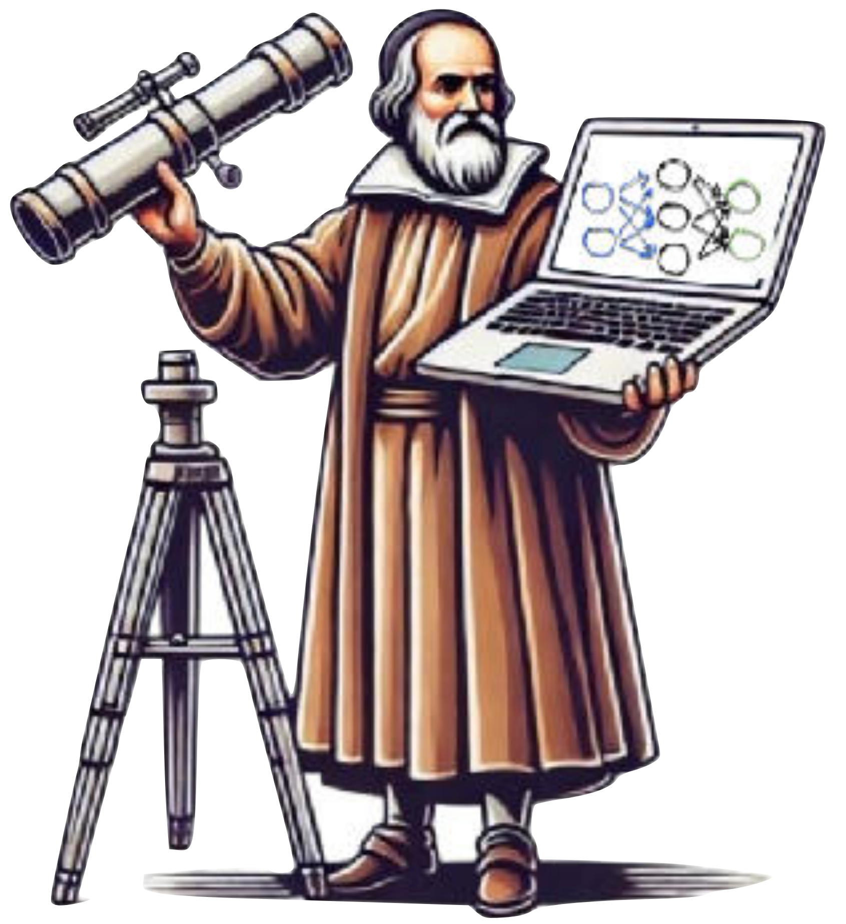
Simulations

Inference network

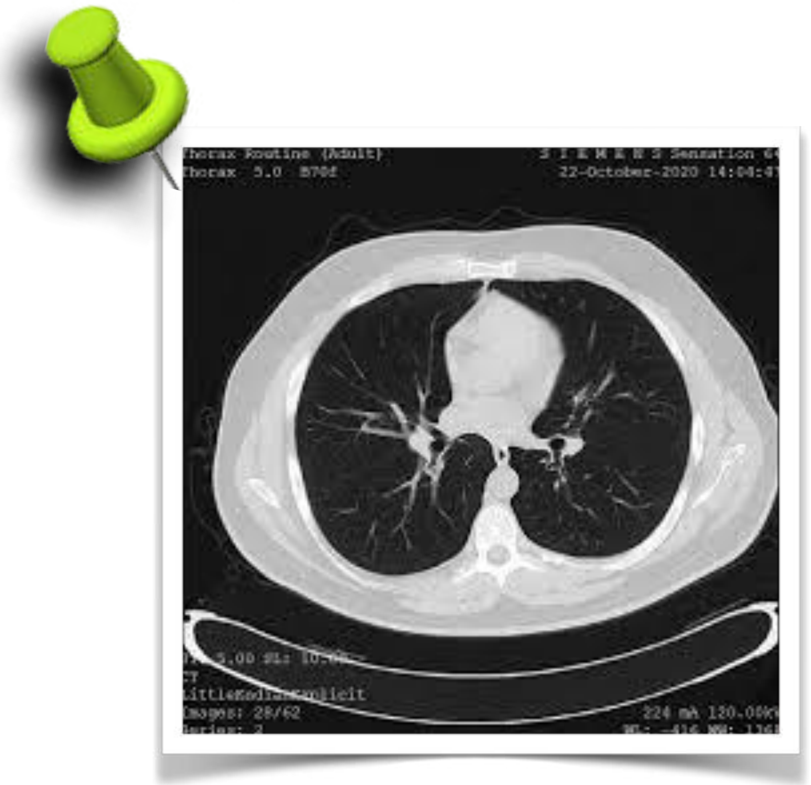
Posterior estimate

Inverse model

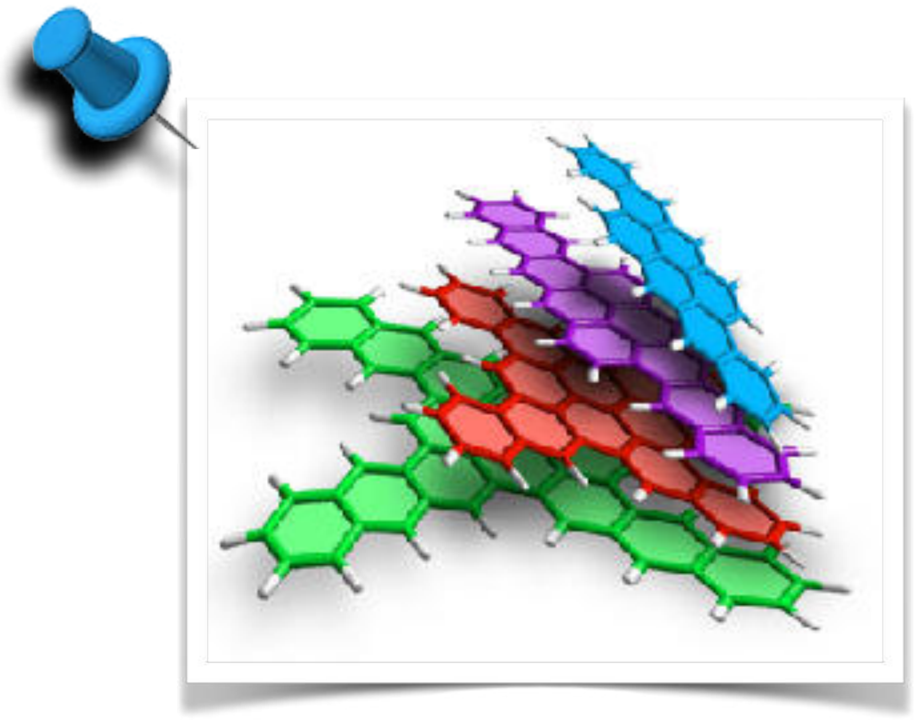
AI4phys@Florence



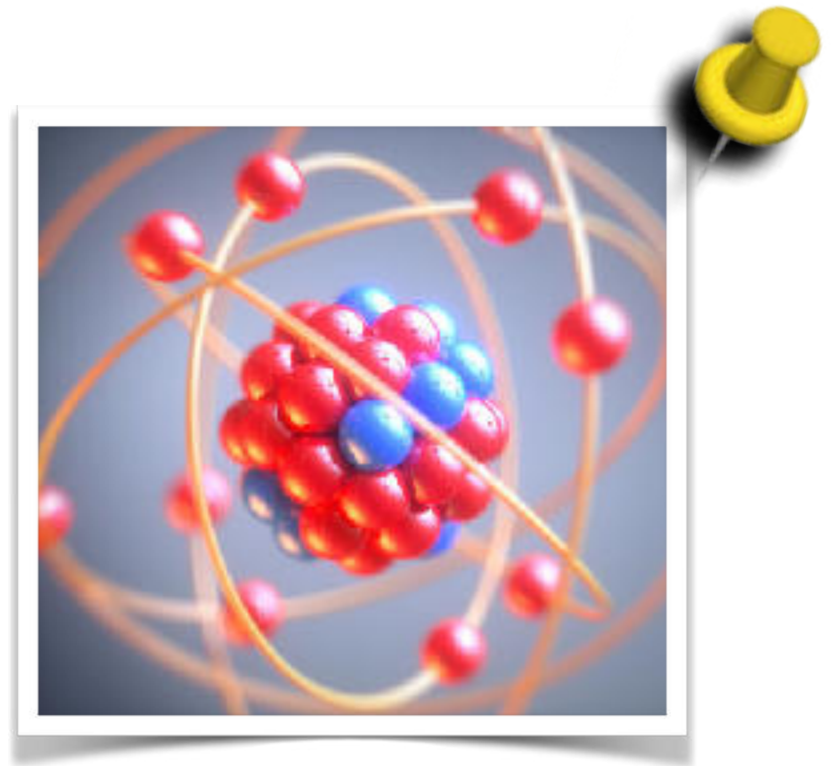
Astro



Medical imaging

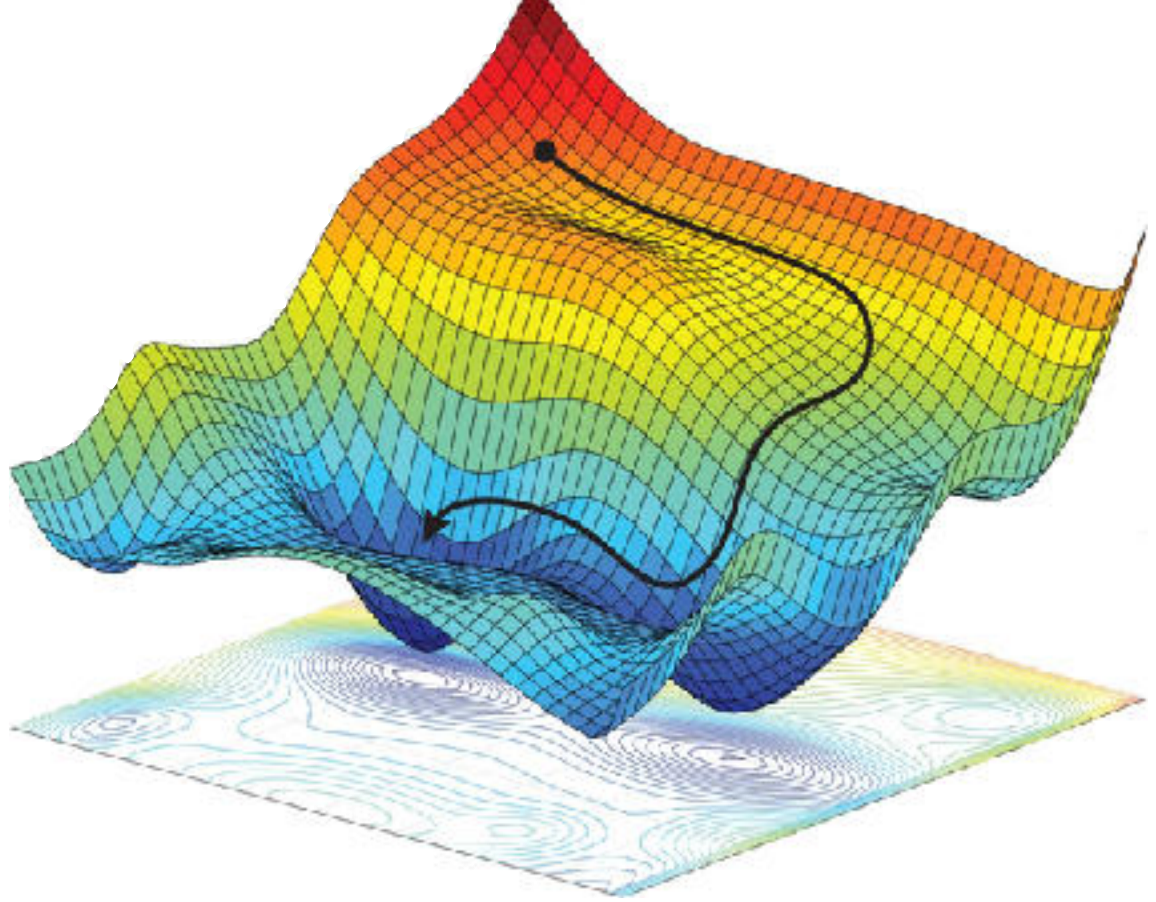


Material science



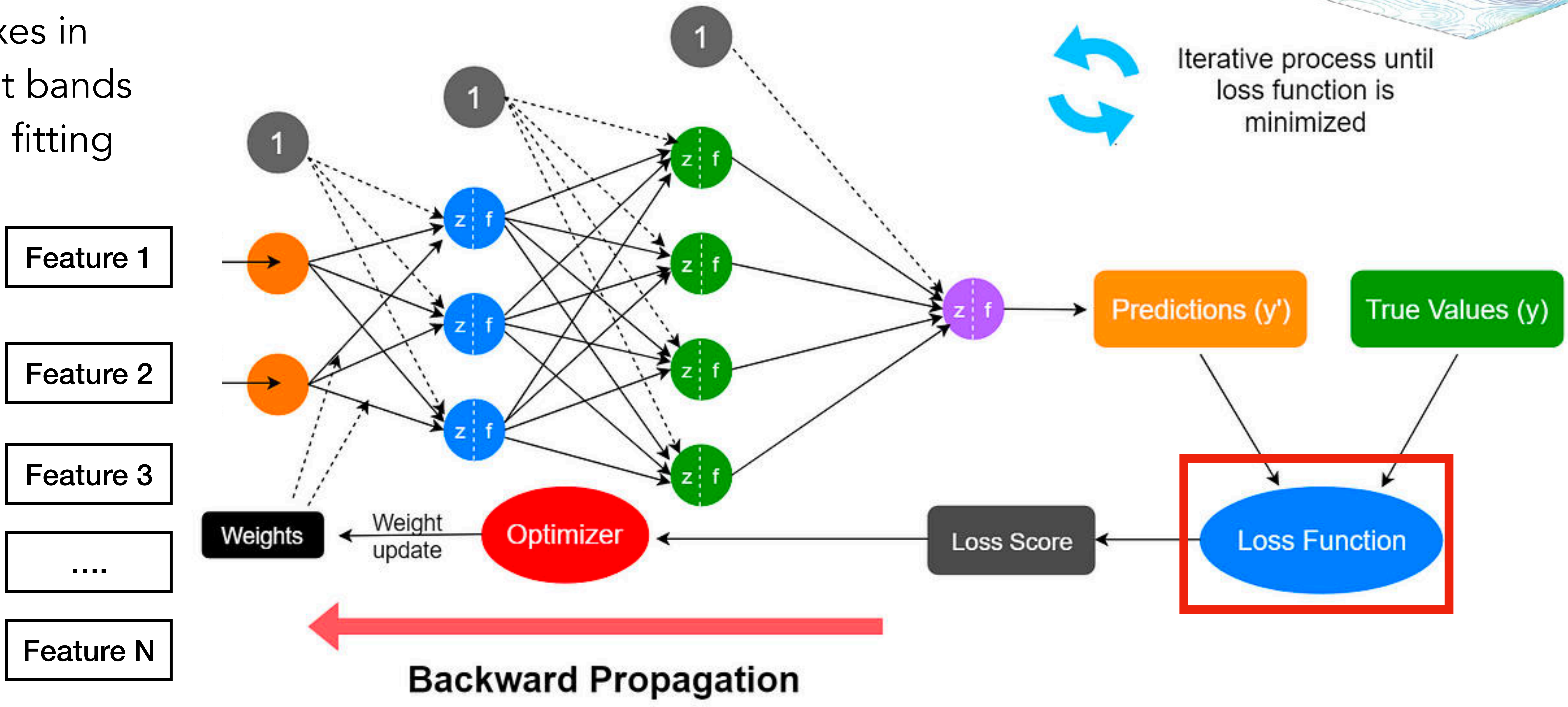
Nuclear physics

Goal: objective function — Find parameters that minimize a loss function



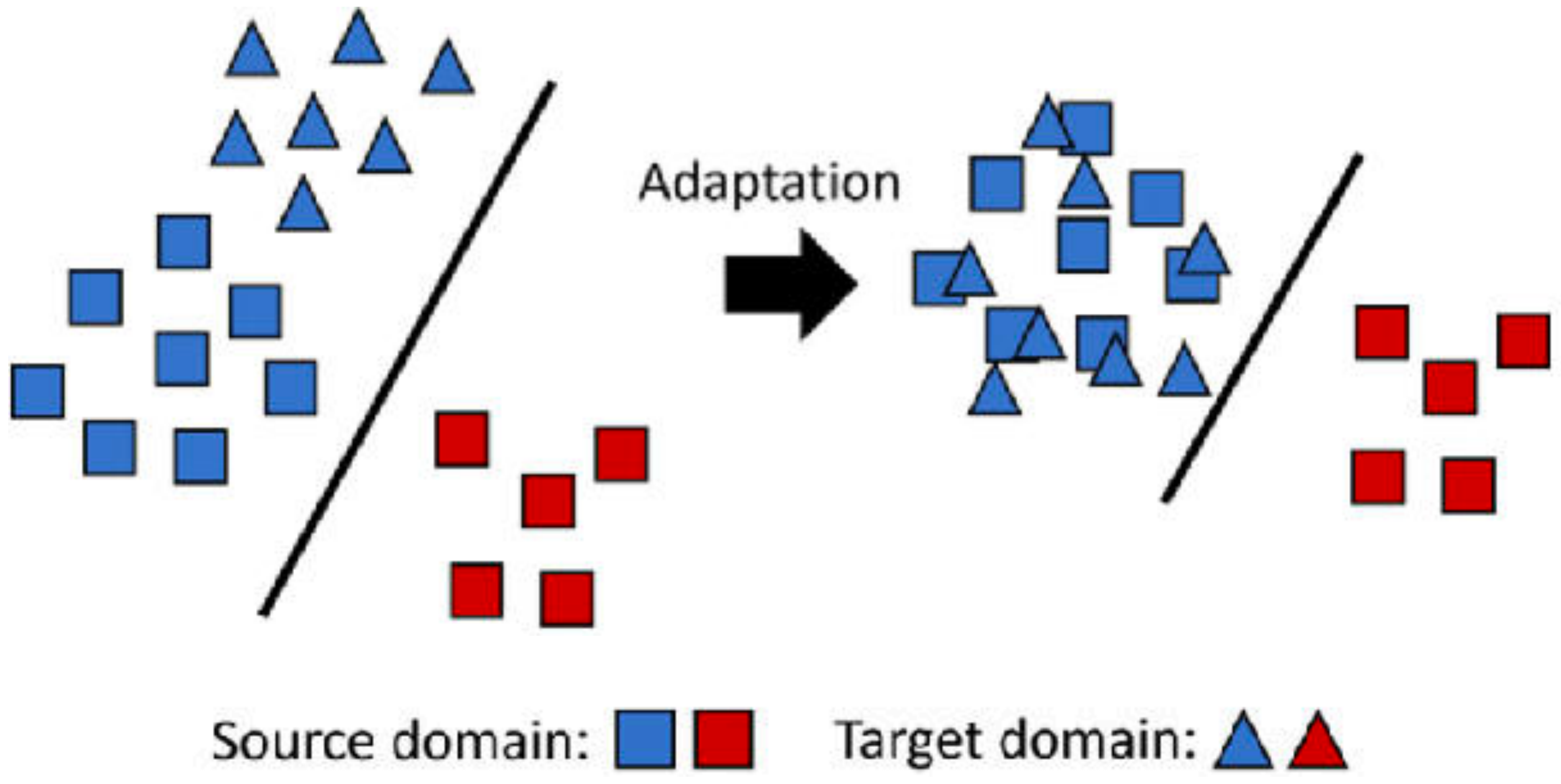
Forward Propagation

e.g, fluxes in
different bands
for SED fitting

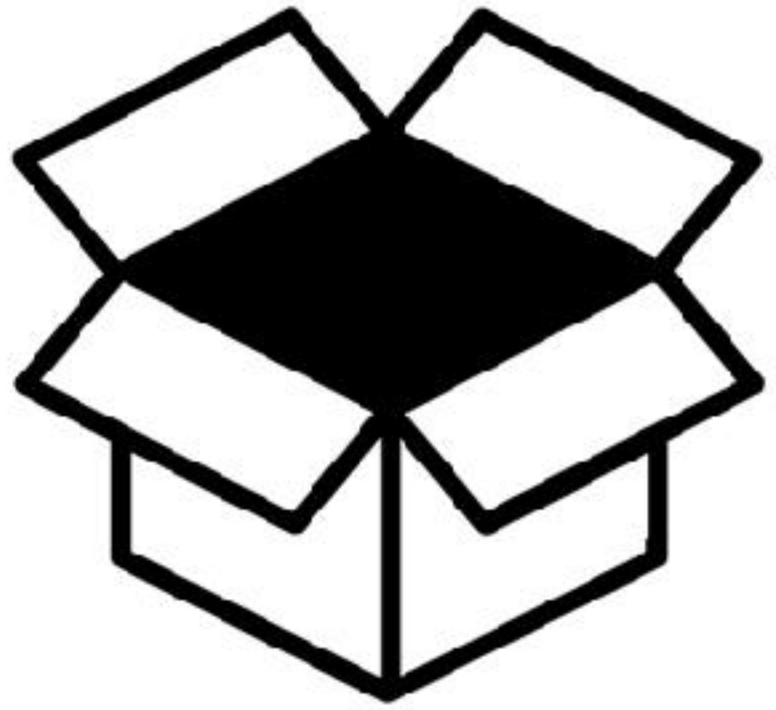


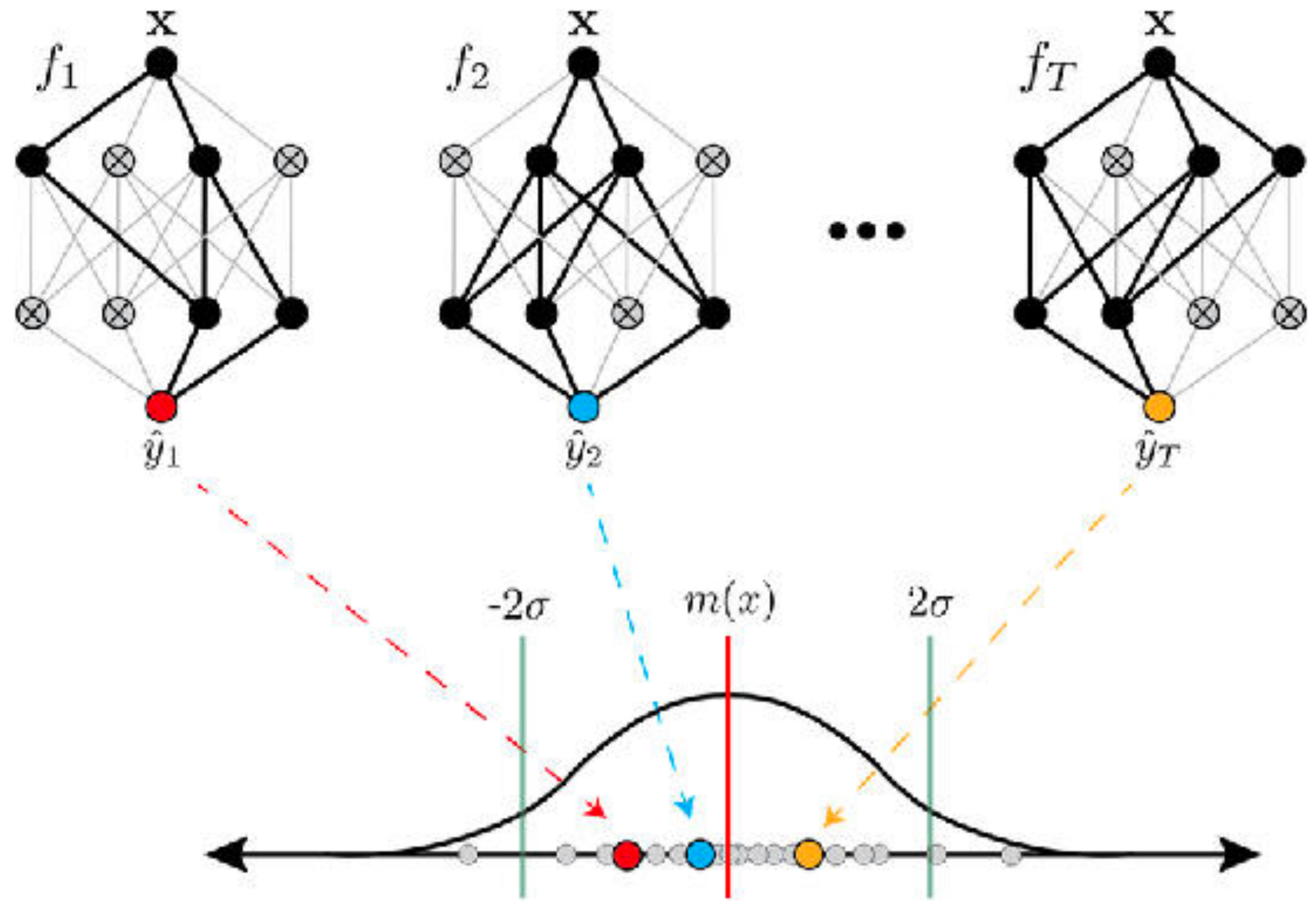
Three **big** challenges in simulation-based inference with AI

Domain-invariant learning



Interpretable learning

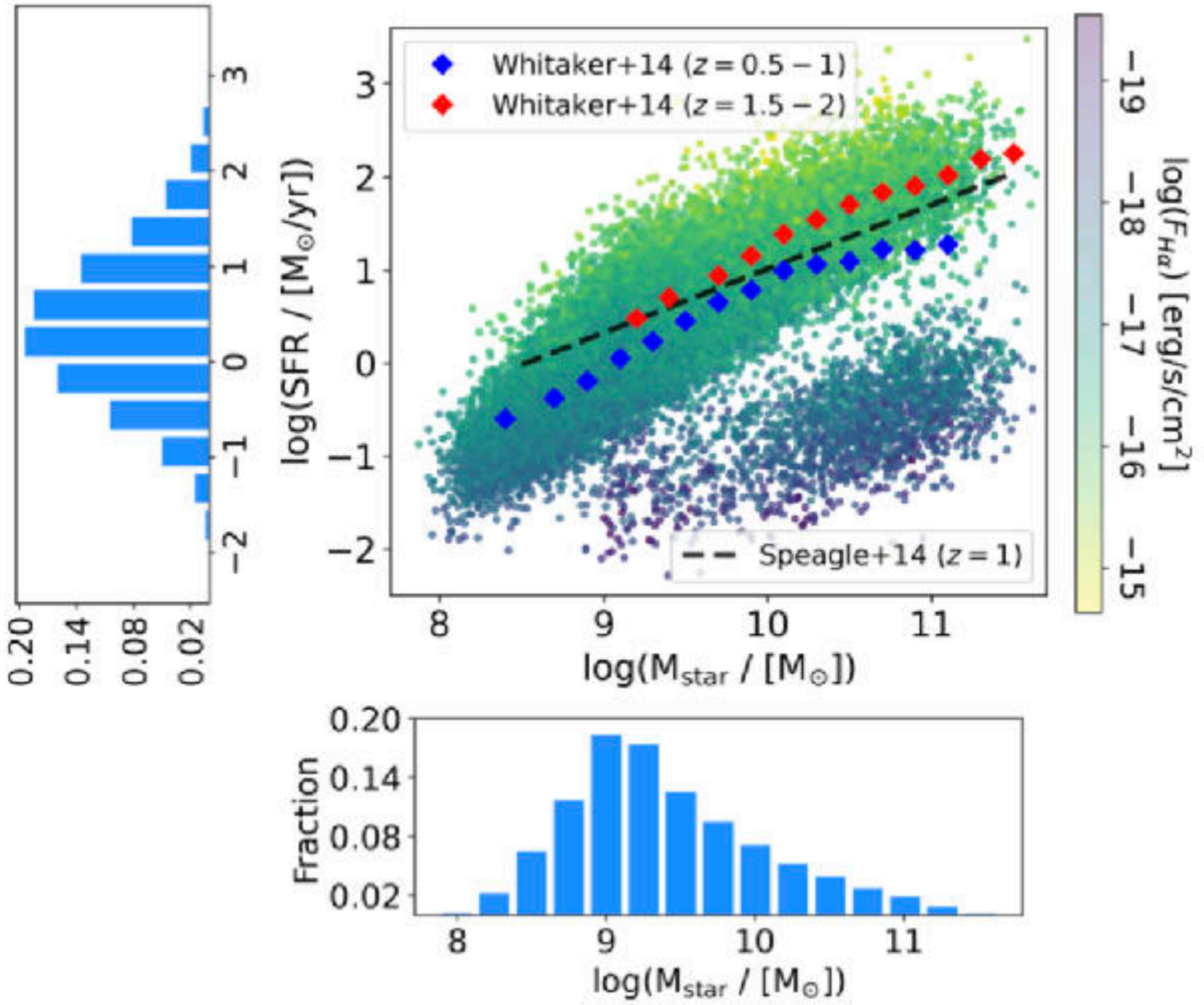


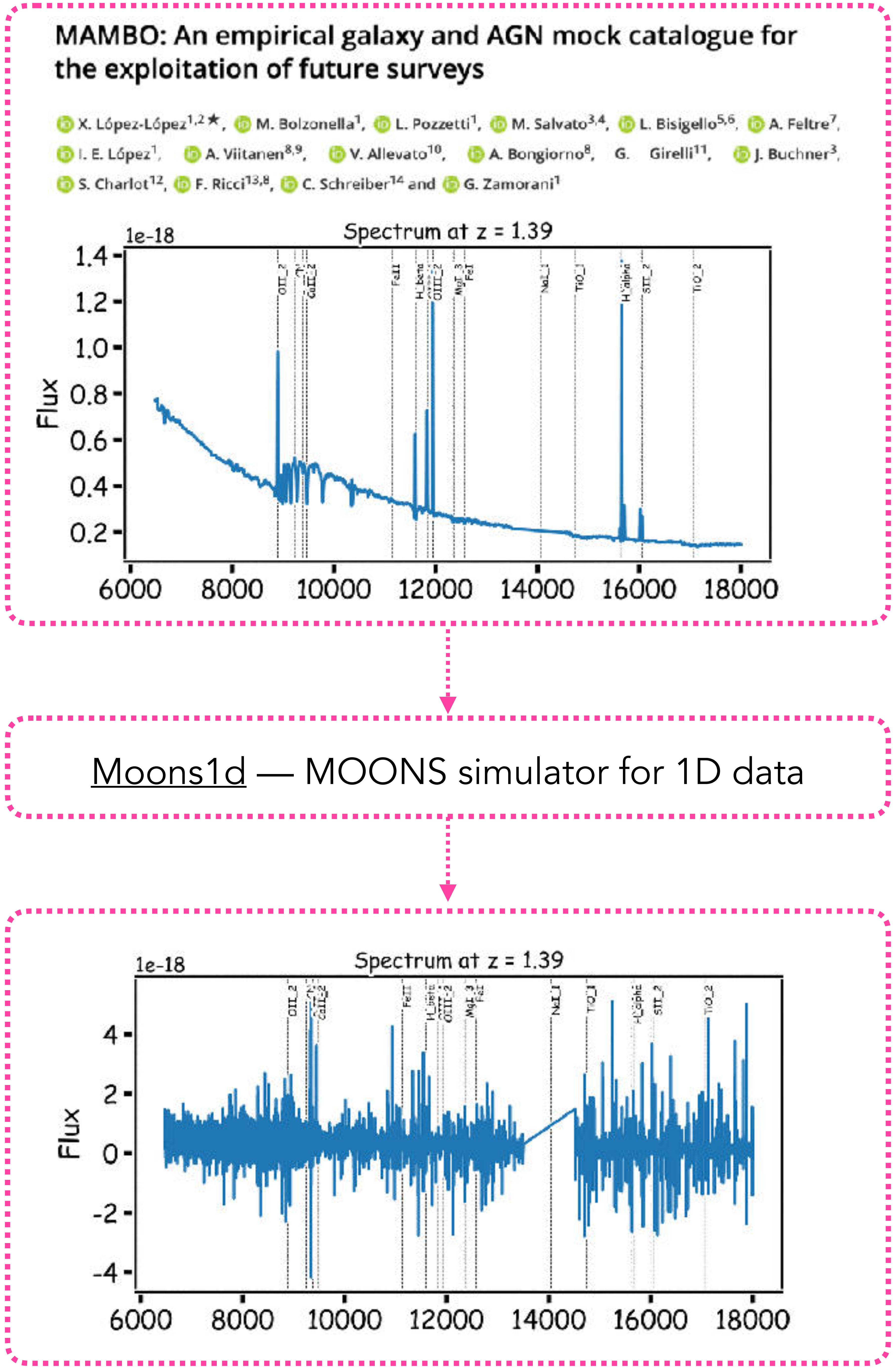


Uncertainty-aware learning

Simulated dataset

- **~120.000** spectra
- generated by running **MAMBO** templates through **moons1d**
- moons1d ran with low resolution mode for all 3 channels (RI, YJ, H), **0.64-1.8 μm**
- a seeing of 0.8'' and airmass of 1.2
- $t_{exp} = 2 - 4 - 8\ h$





MAMBO: An empirical galaxy and AGN mock catalogue for the exploitation of future surveys

X. López-López[1,2]★, M. Bolzonella[1], L. Pozzetti[1], M. Salvato[3,4], L. Bisigello[5,6], A. Feltre[7], I. E. López[1], A. Viitanen[8,9], V. Allevato[10], A. Bongiorno[8], G. Girelli[11], J. Buchner[3], S. Charlot[12], F. Ricci[13,8], C. Schreiber[14] and G. Zamorani[1]

Model

Moons1d — MOONS simulator for 1D data

sky condition, obs strategy, etc

Simulated spectrum

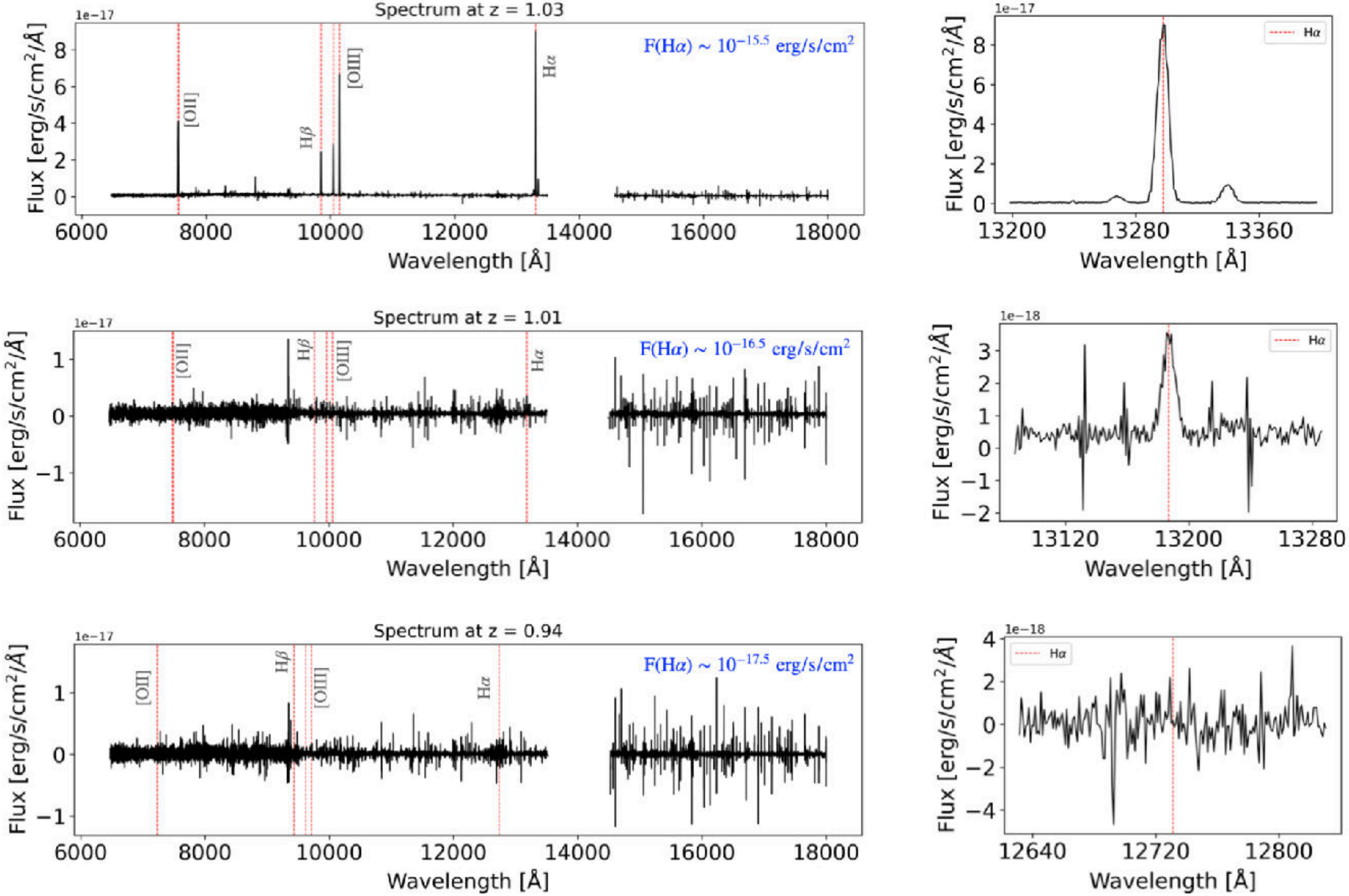
Dataset

- $t_{exp} = 2,\ 4,\ 8\ h$
- 0.64 – 1.8 μm
- 12.217 channels

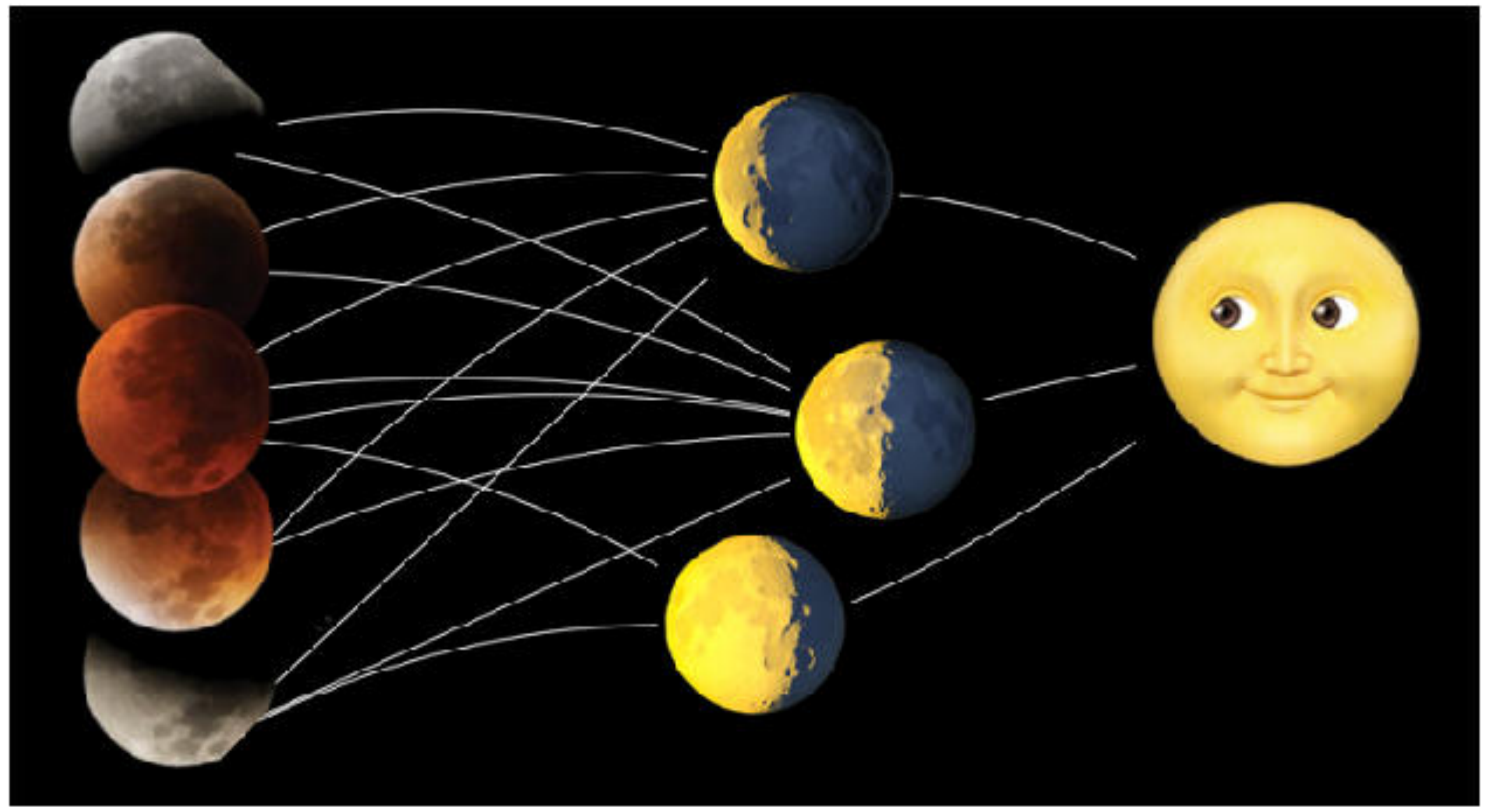
Target physics

- redshift, z
- stellar mass, $M_{\rm star}$
- star formation rate, SFR

Simulated spectra

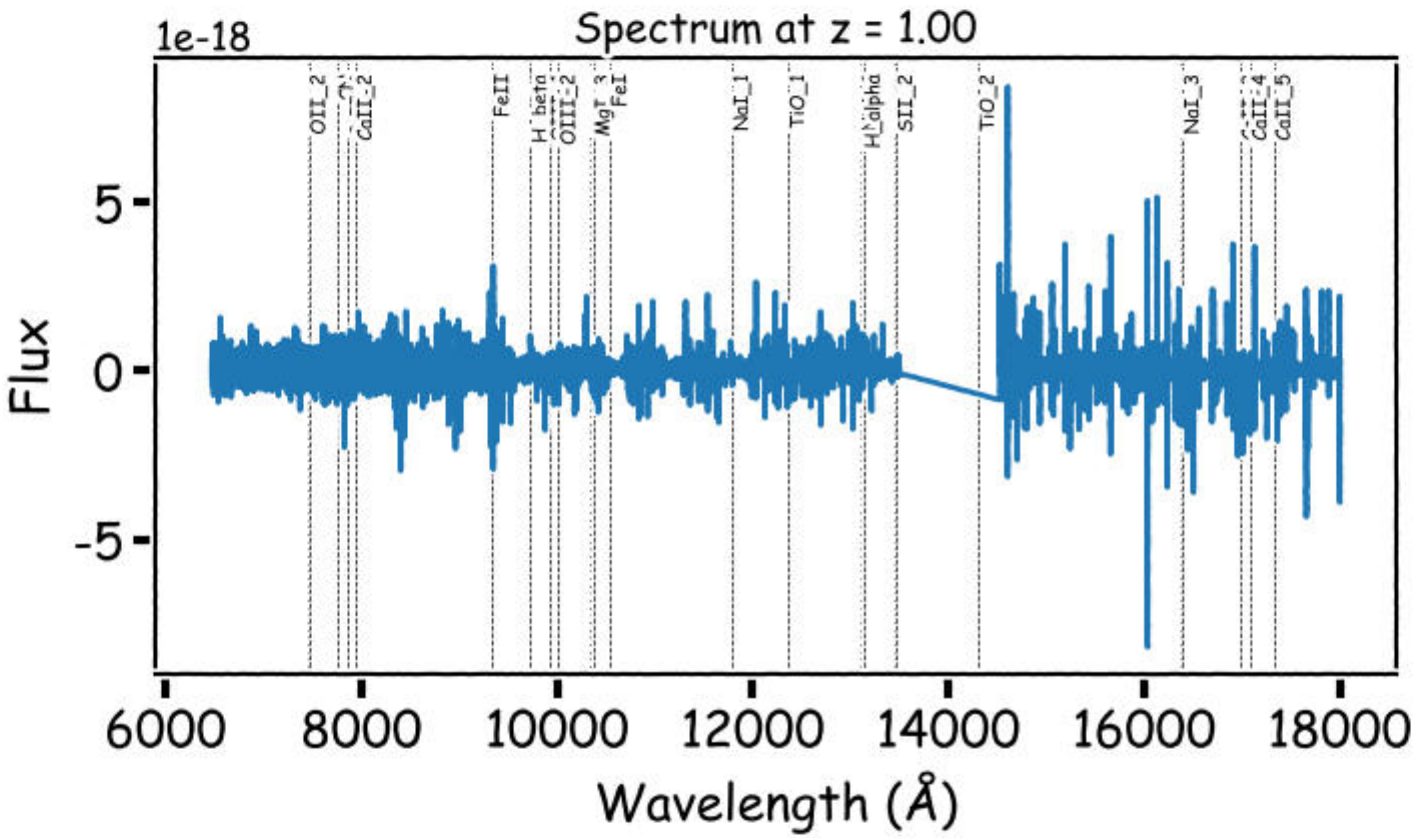


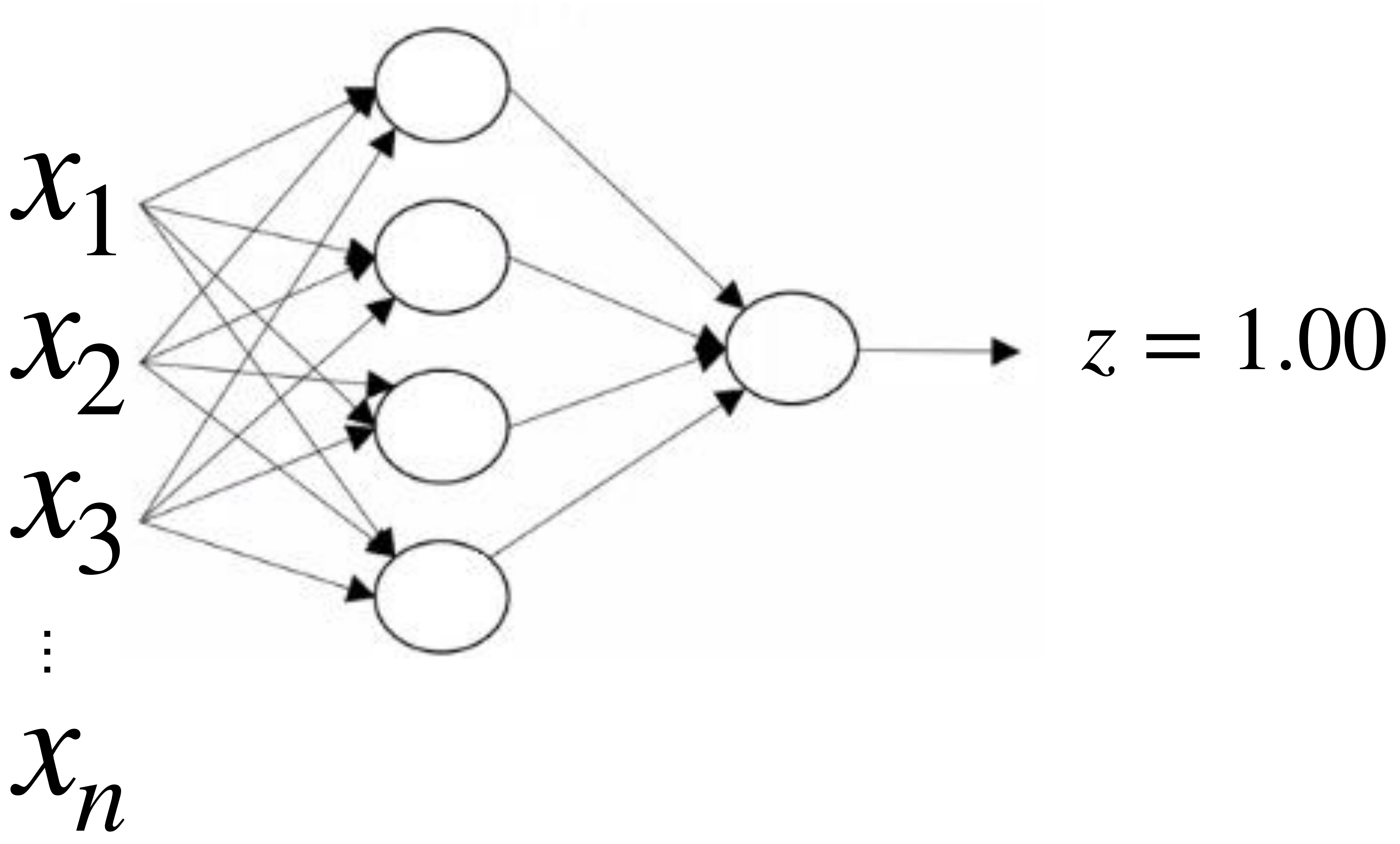
Deep learning MOONS spectra



The case of redshift

Classical scheme: a regression problem



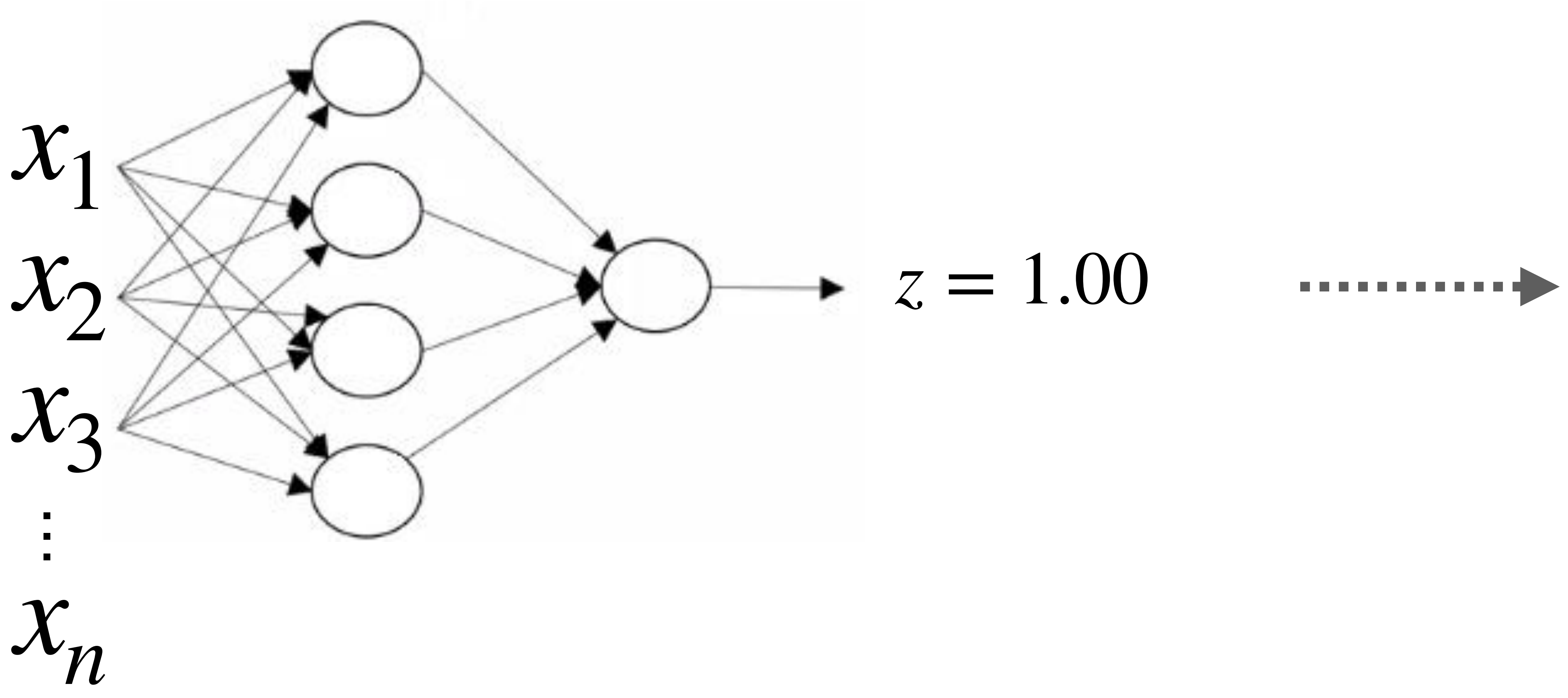


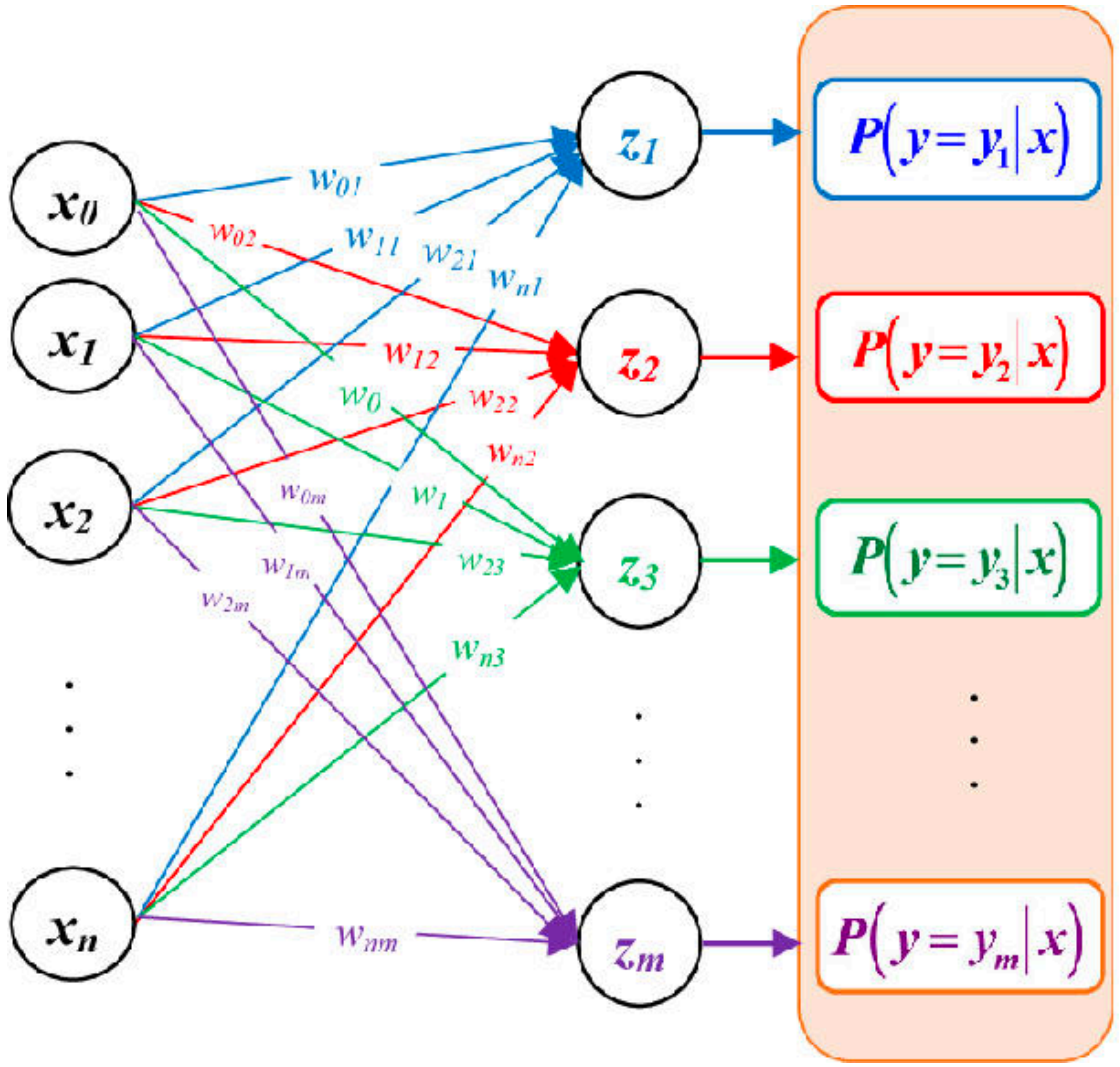
Deep learning MOONS spectra 🌒🌓🌔🌘

Why not discretise continuous redshift values into finely spaced bins? 💡

Switch to a classification task

We adopt $dz = 0.003$





$z = 0.997 \quad (P = 0.1)$

$z = 1.000 \quad (P = 0.95)$

$z = 1.003 \quad (P = 0.05)$

⋮

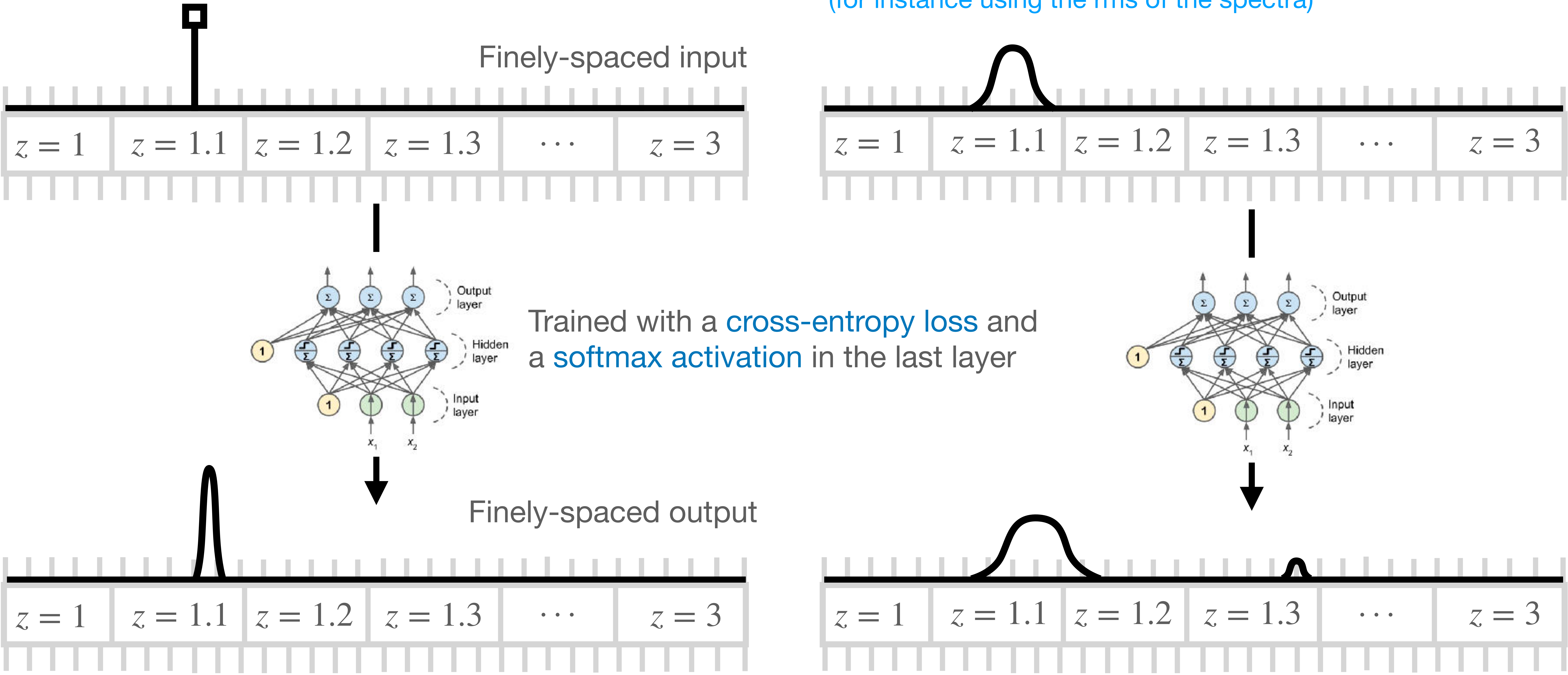
$z = 2 \quad (P = 0)$

Mini bibliography — Carrasco Kind & Brunner 2013; Stivaktakis et al. 2018; Stewart et al. 2022; Pankaj et al. 2022

Switch to a classification task

We can also account for the *quality* of spectra

(for instance using the rms of the spectra)

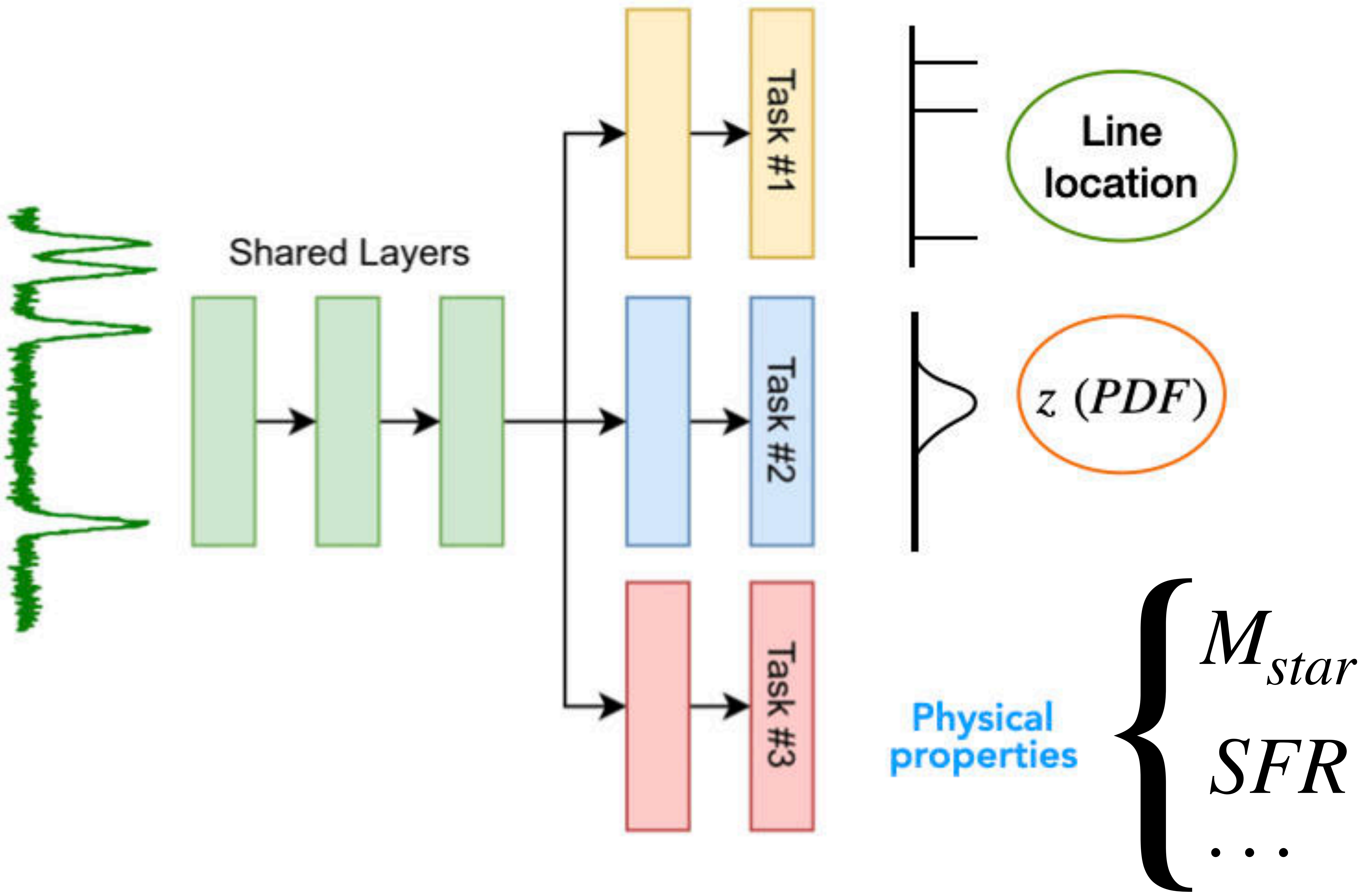


Finely-spaced input

Trained with a cross-entropy loss and a softmax activation in the last layer

Finely-spaced output

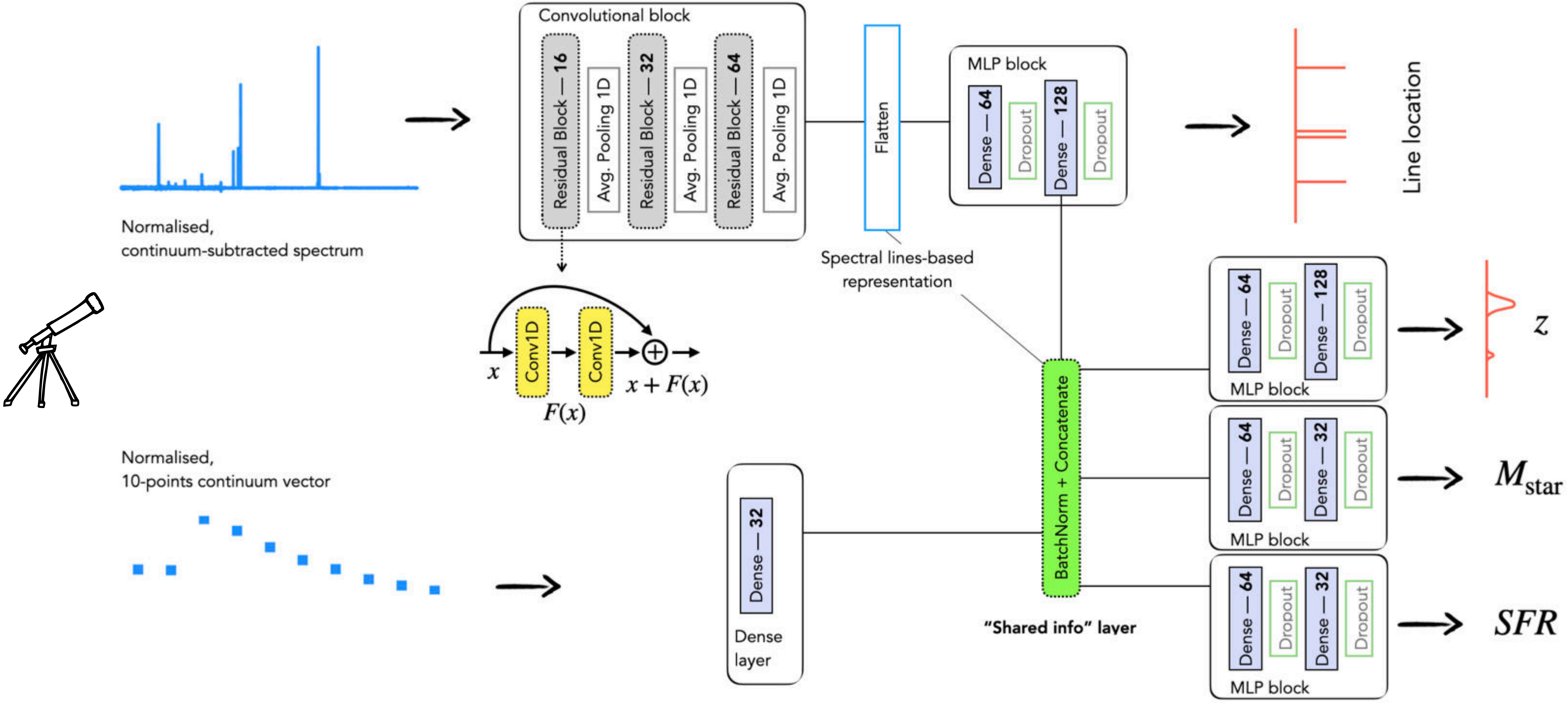
Learning through multi-task training 🤹



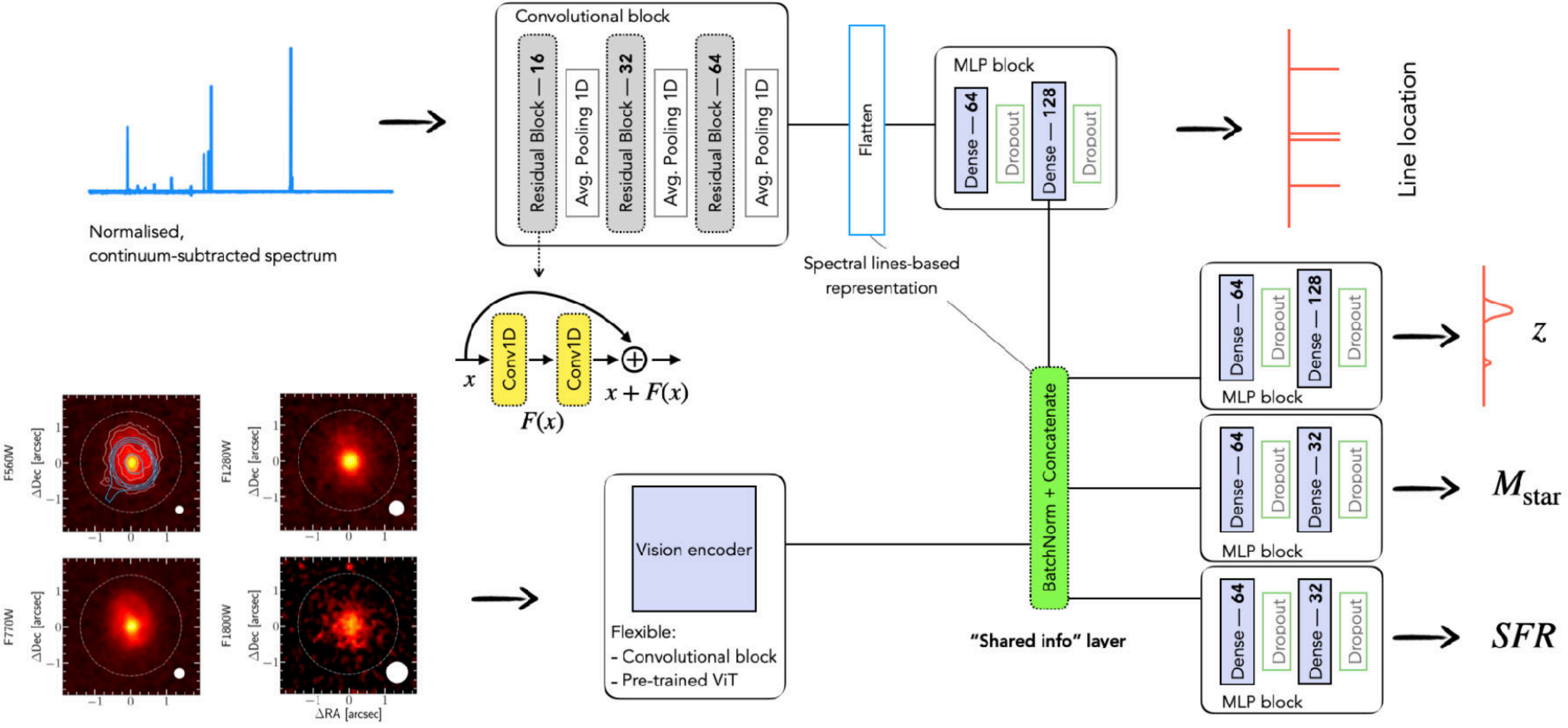
Physical properties $\left\{ \begin{array}{l} M_{star} \\ SFR \\ \ldots \end{array} \right.$

Mini bibliography — Caruana 1997; Ruder 2017; Crawshaw 2020, Hervella+24

M-TOPnet (Multi-Task network Outputting Probabilities)

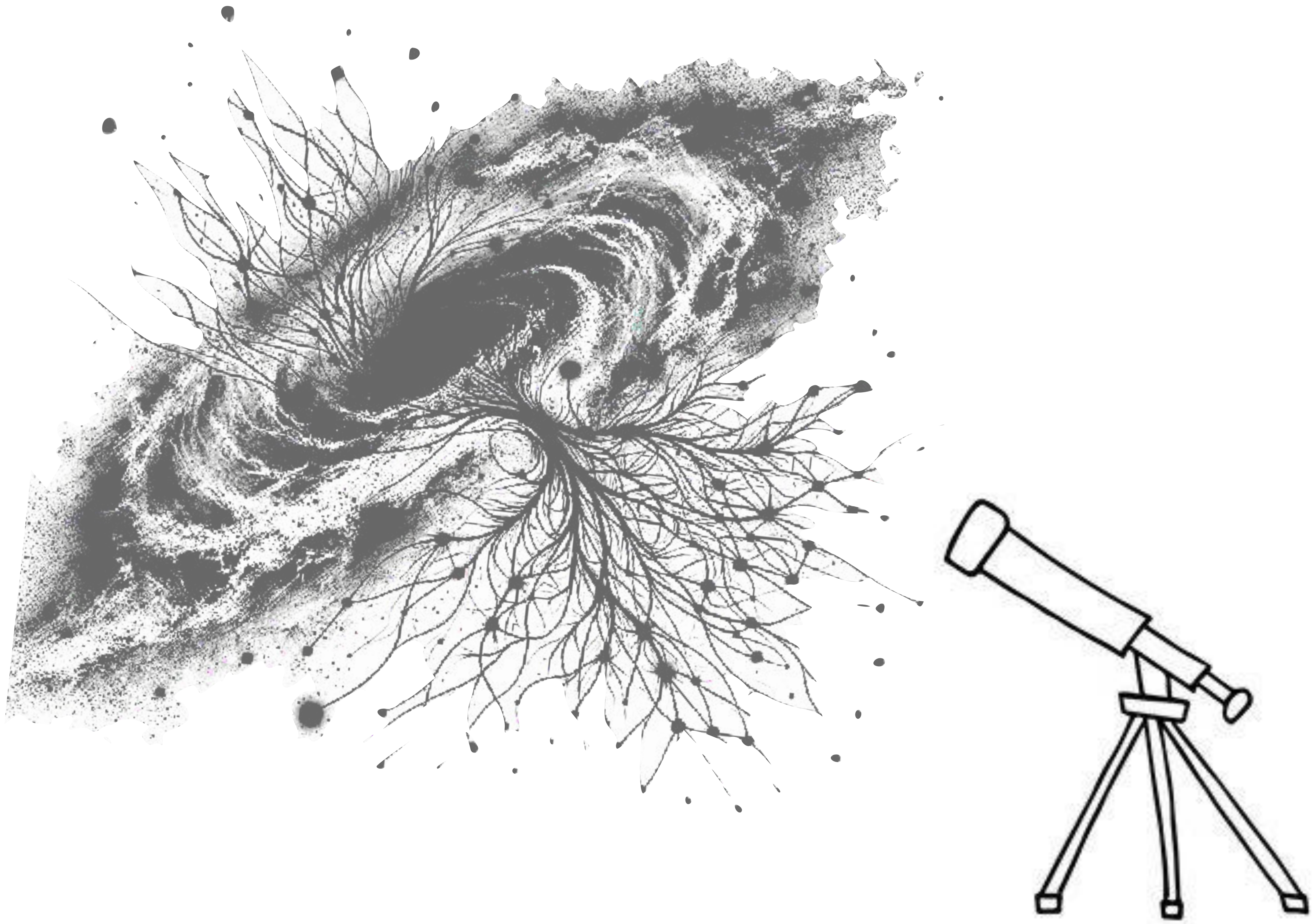


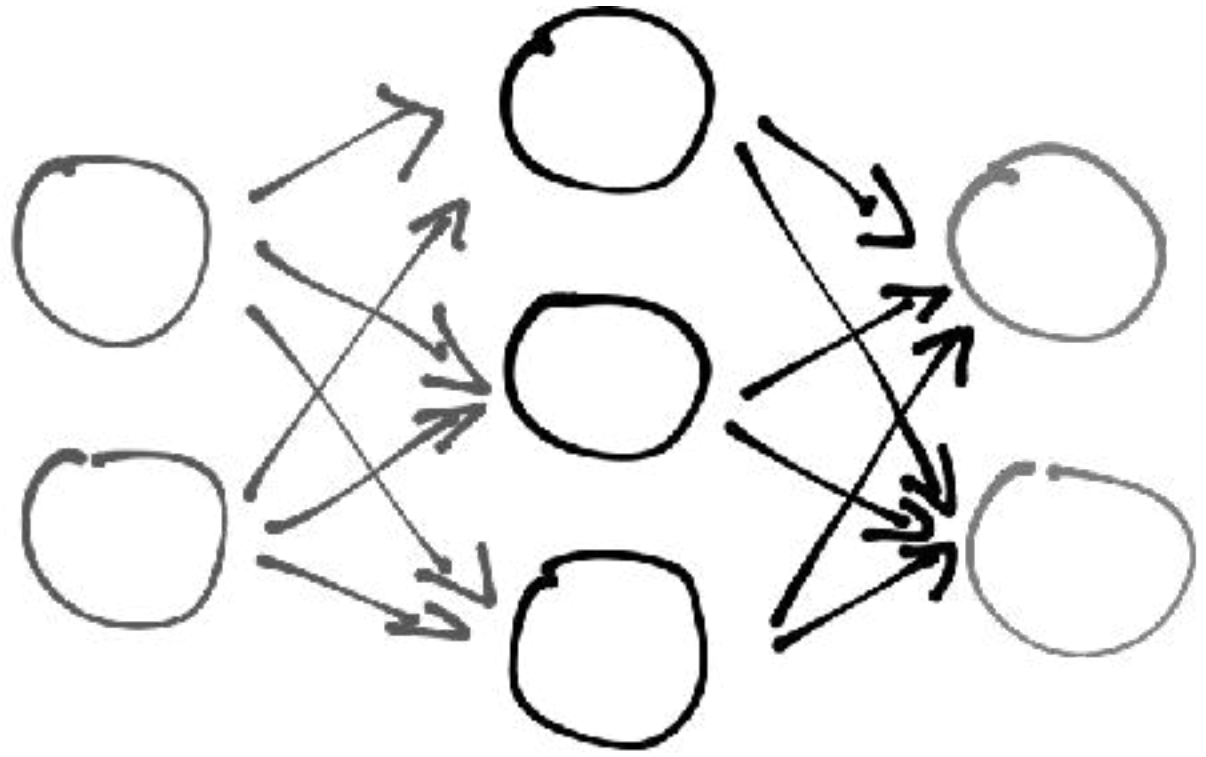
M-TOPnet (Multi-Task network Outputting Probabilities)



Predictions on the test set 🔮🪄

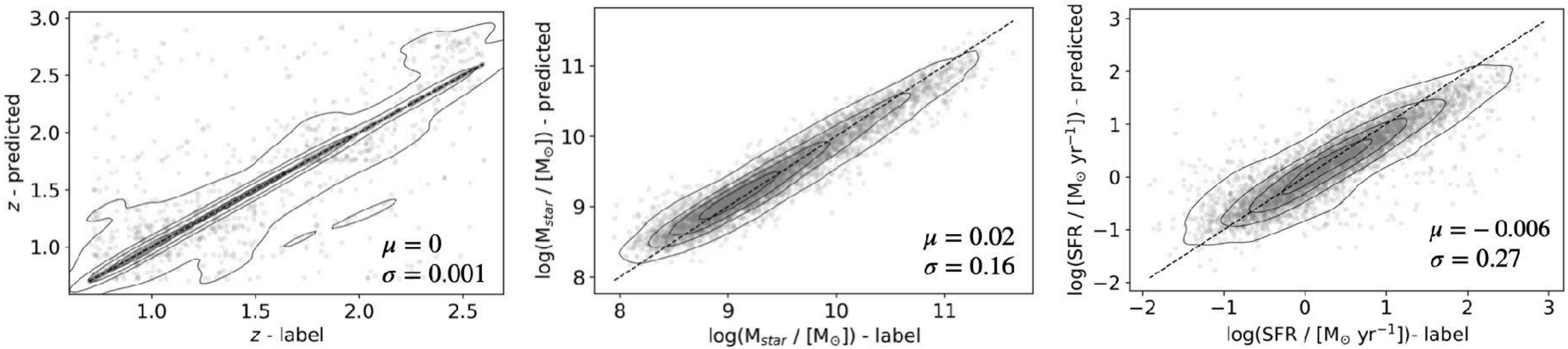
- **~18000** spectra
- Same distributions as training set

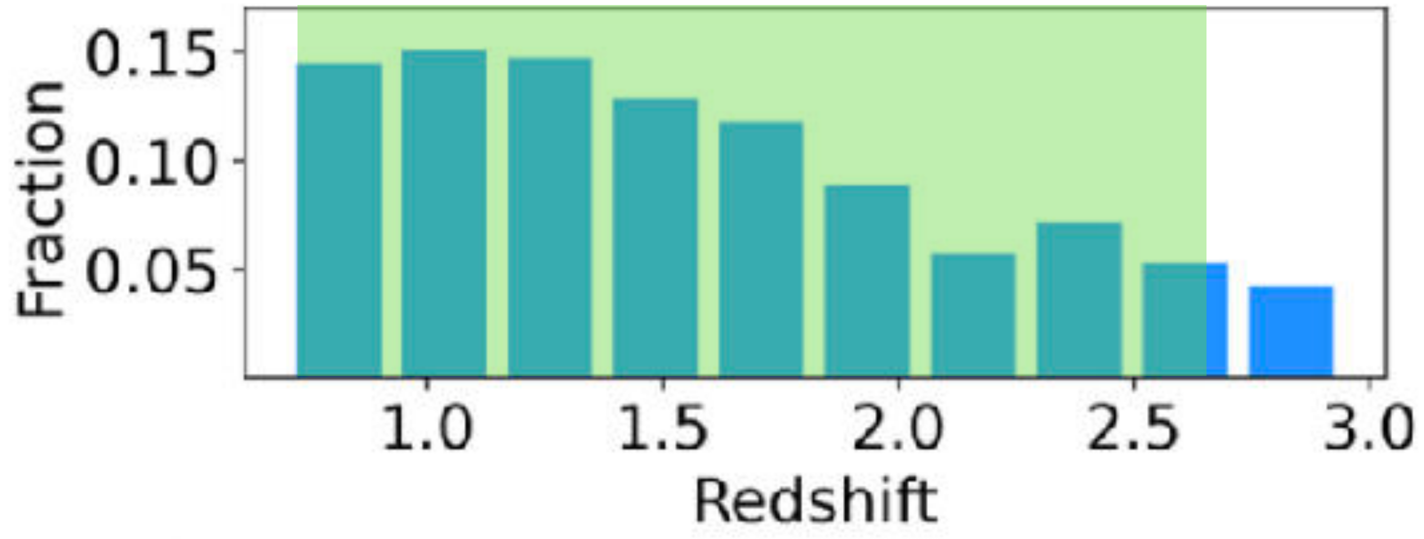


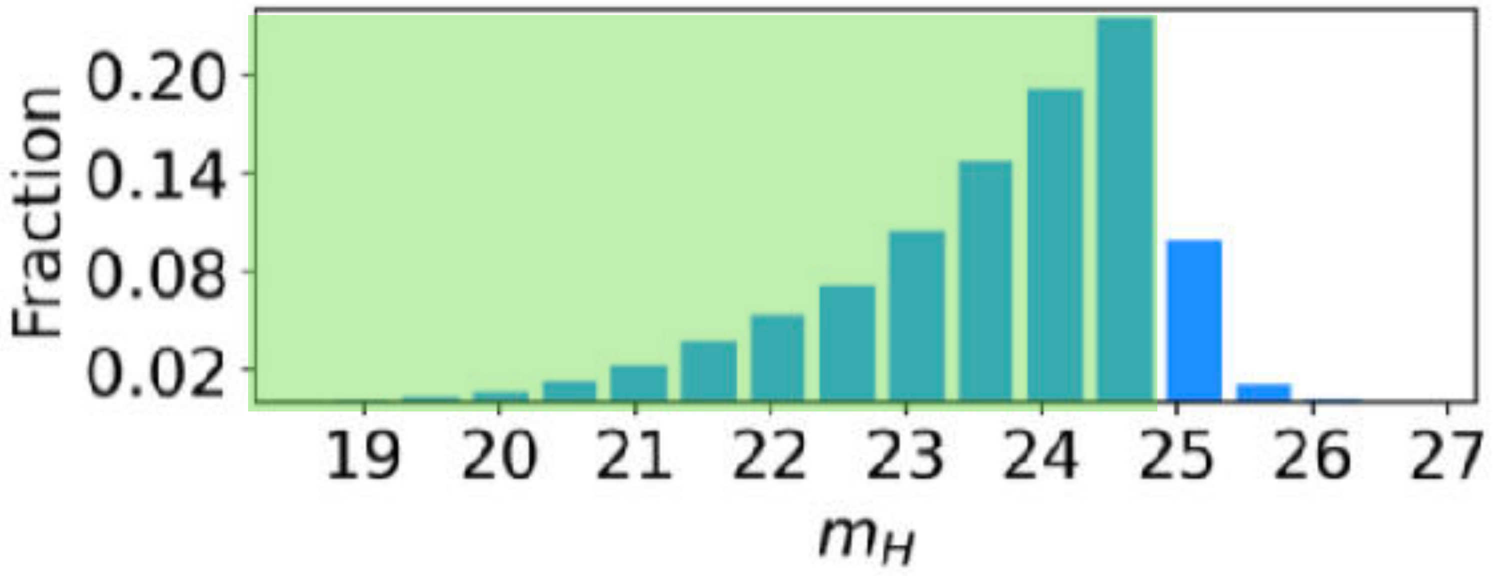


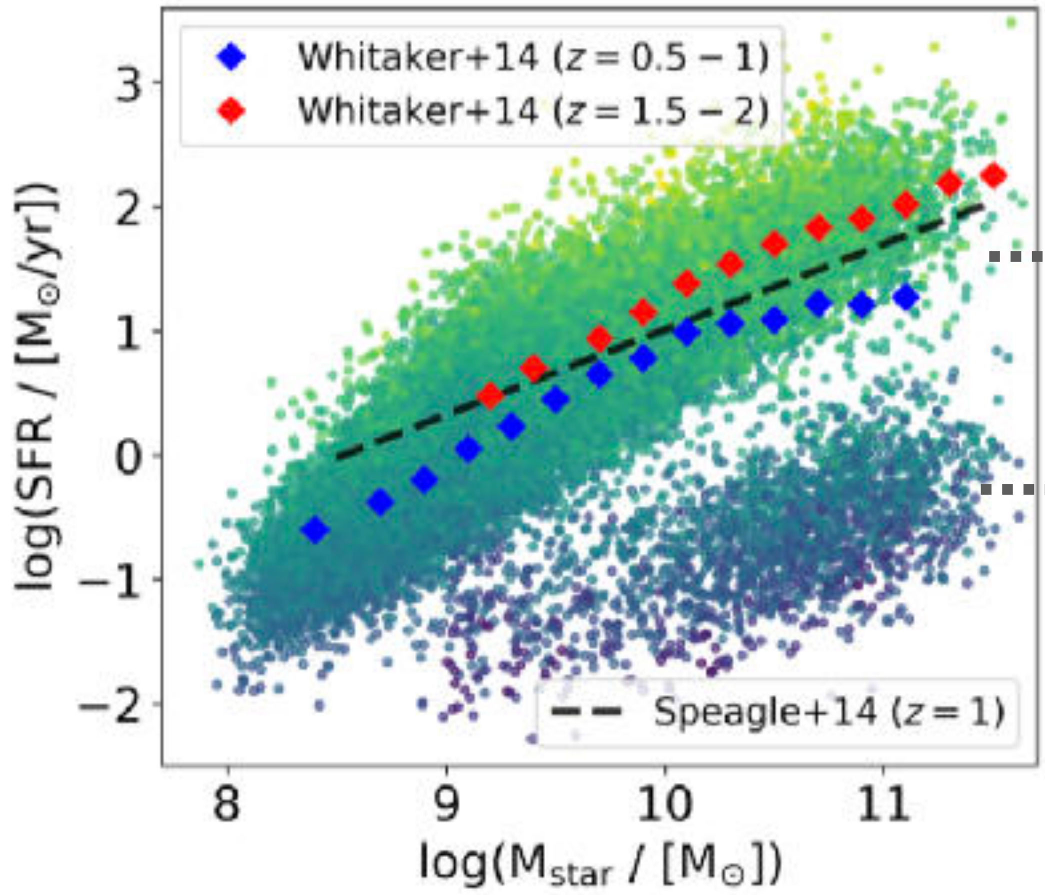
A physically-motivated test sub-set 🔭

MOONRISE-like selection: $m_H < 25$; $z < 2.6$; $t_{exp} = 8\ h$ for passive galaxies; $t_{exp} = 2\ h$ for star-forming galaxies





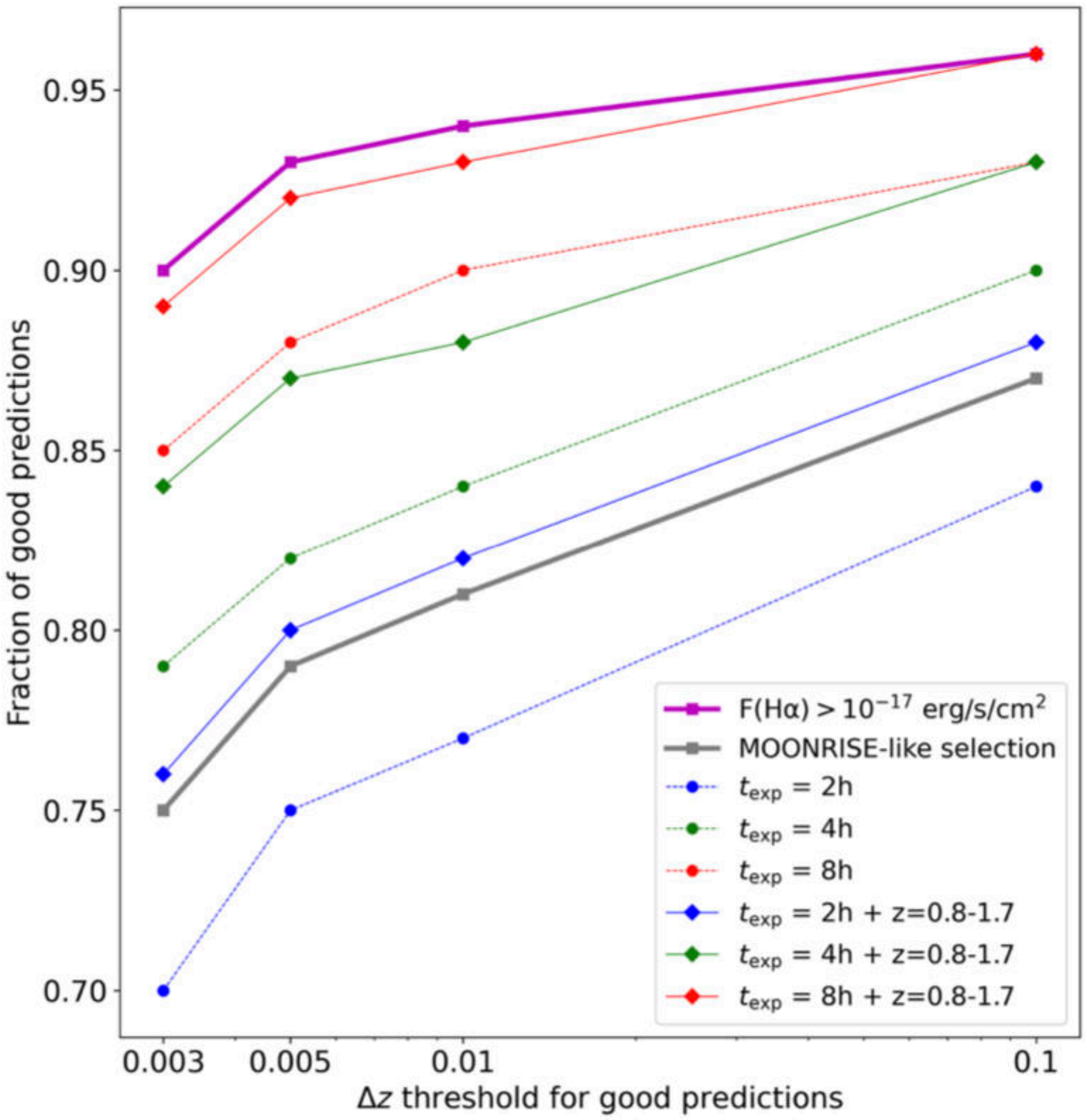


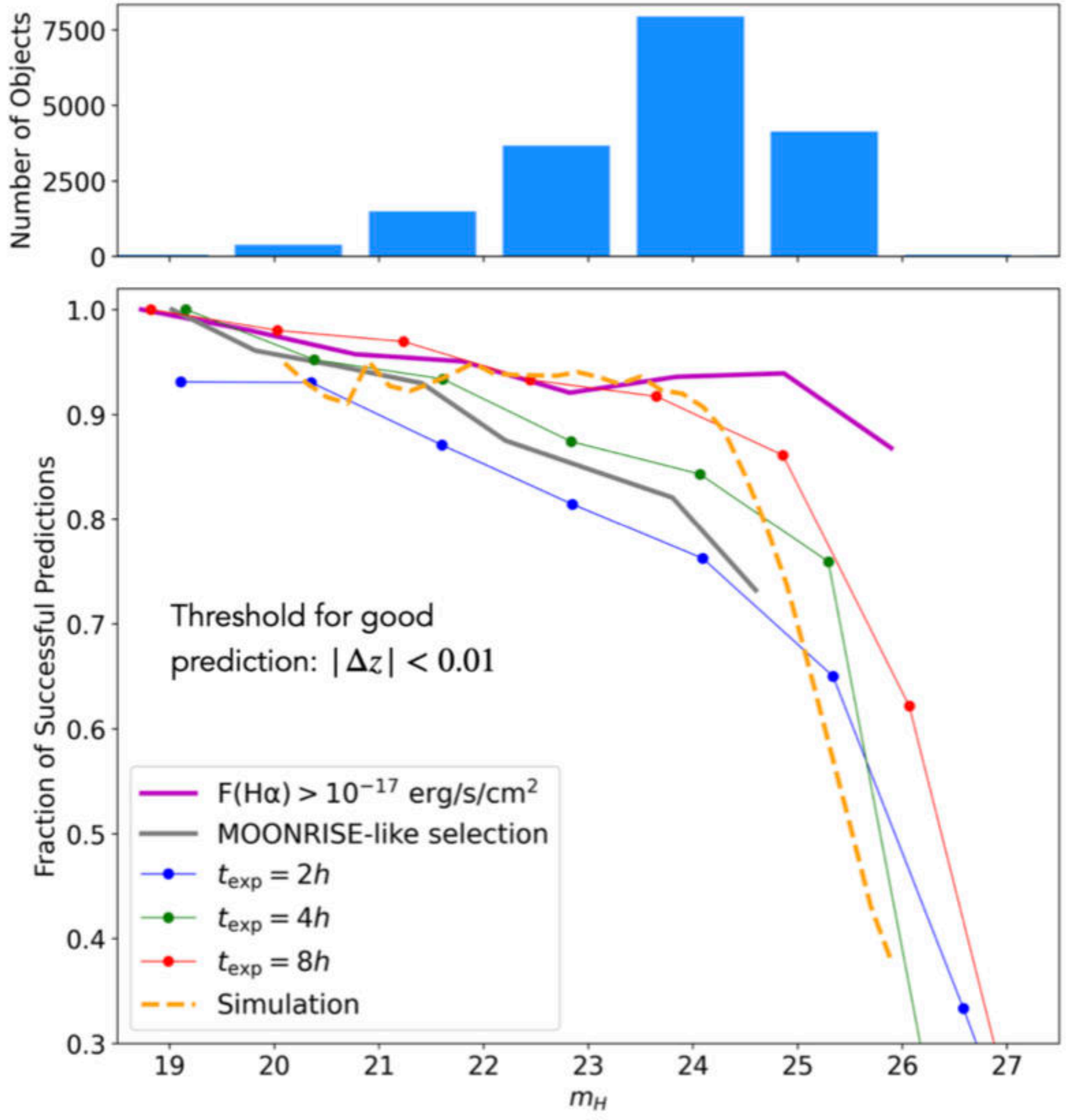


$t_{exp} = 2\ h$

$t_{exp} = 8\ h$

Metrics: let's help design observational strategies 🧐

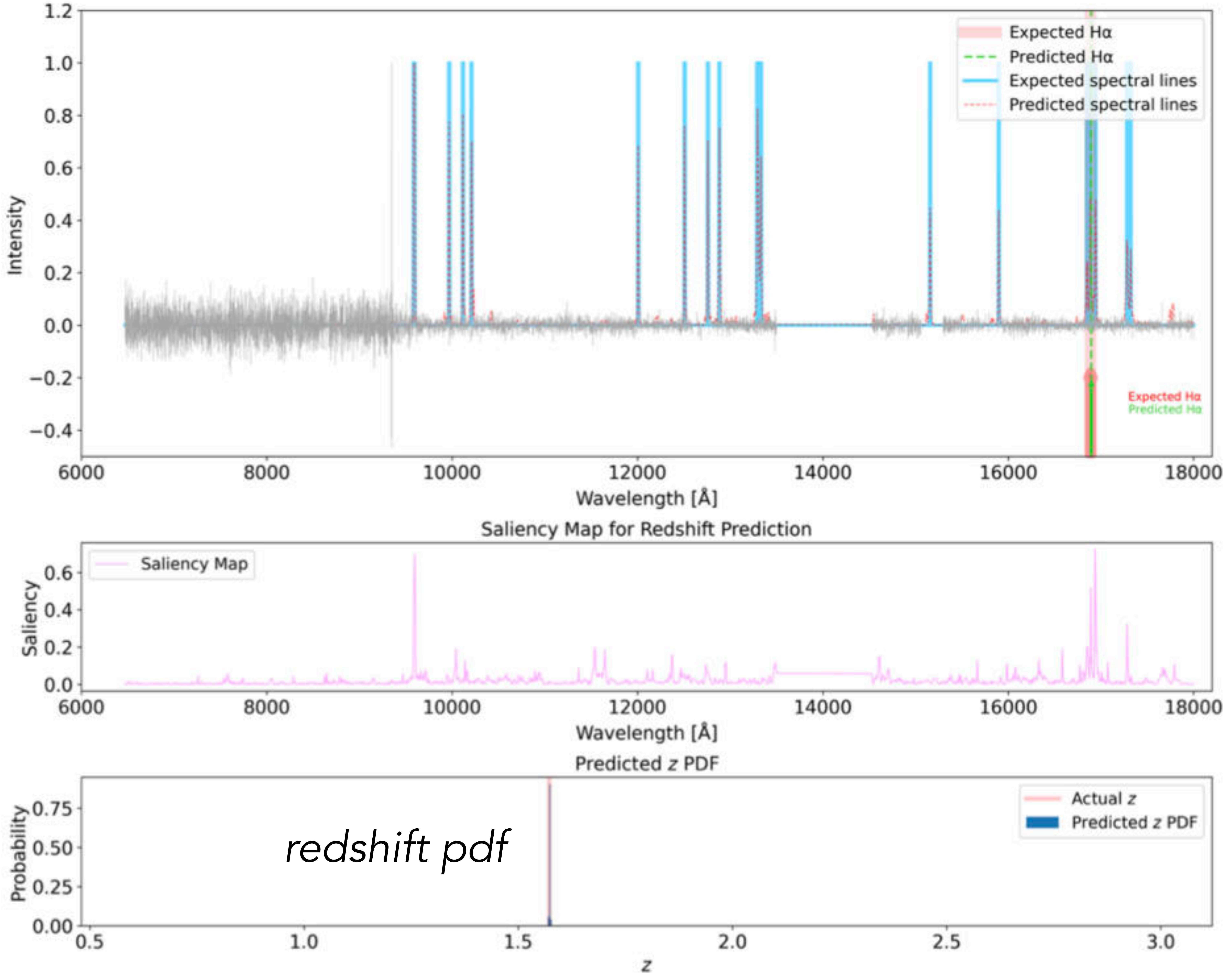




Threshold for good prediction: $|\Delta z| < 0.01$

Let's have a look at the predictions 🧙‍♂️

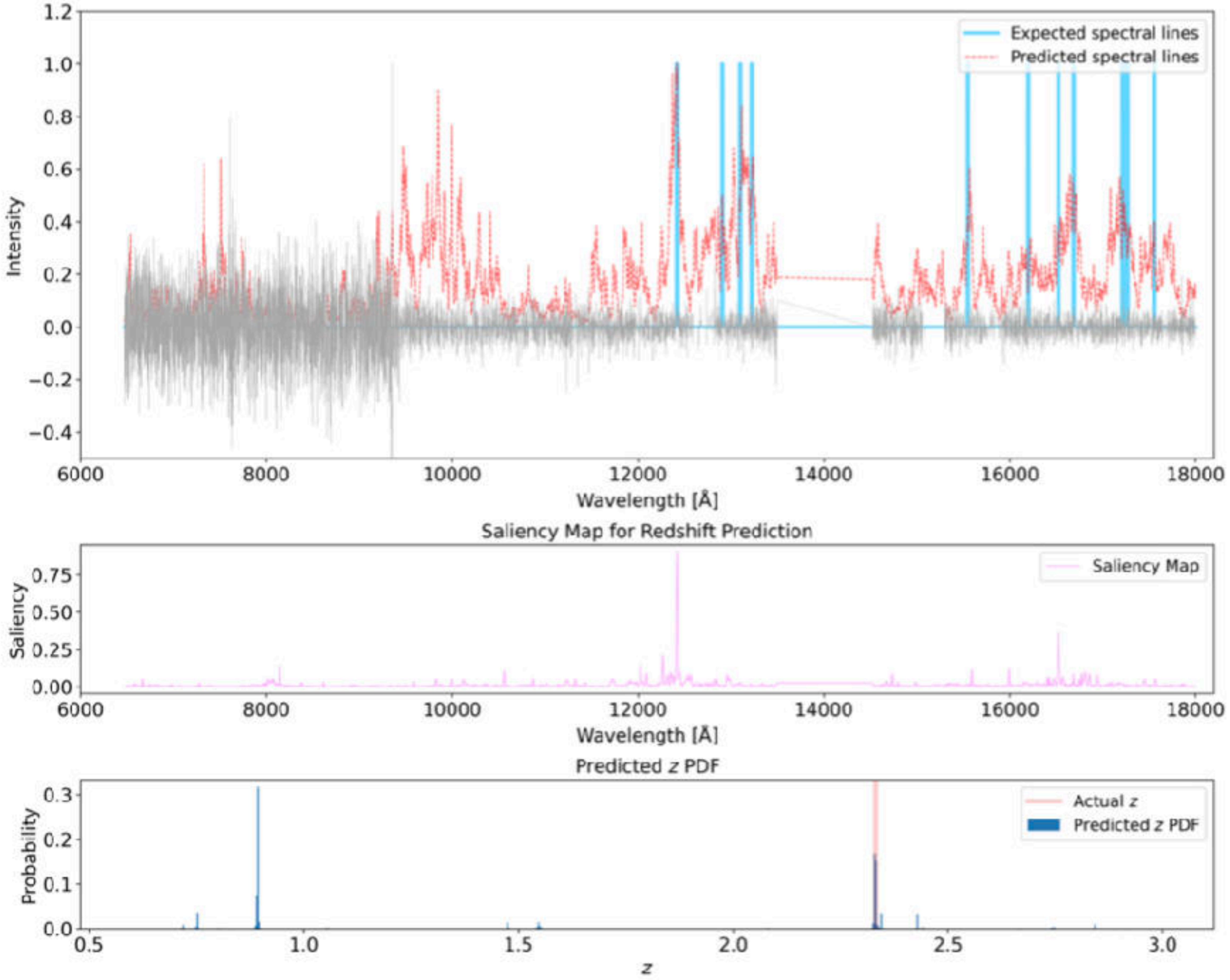
Good prediction

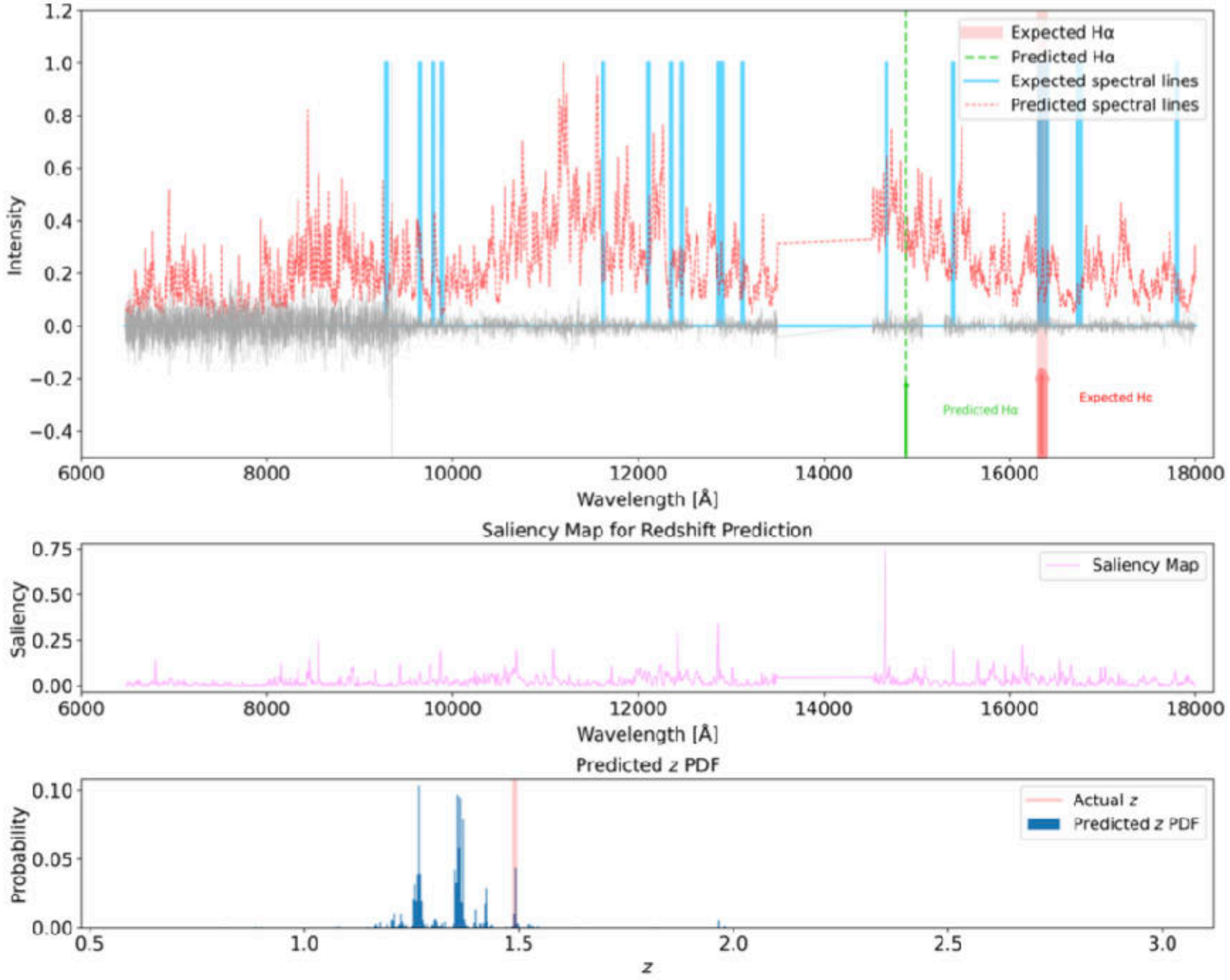


redshift pdf

Let’s have a look at the predictions 🧙‍♂️

Bad predictions

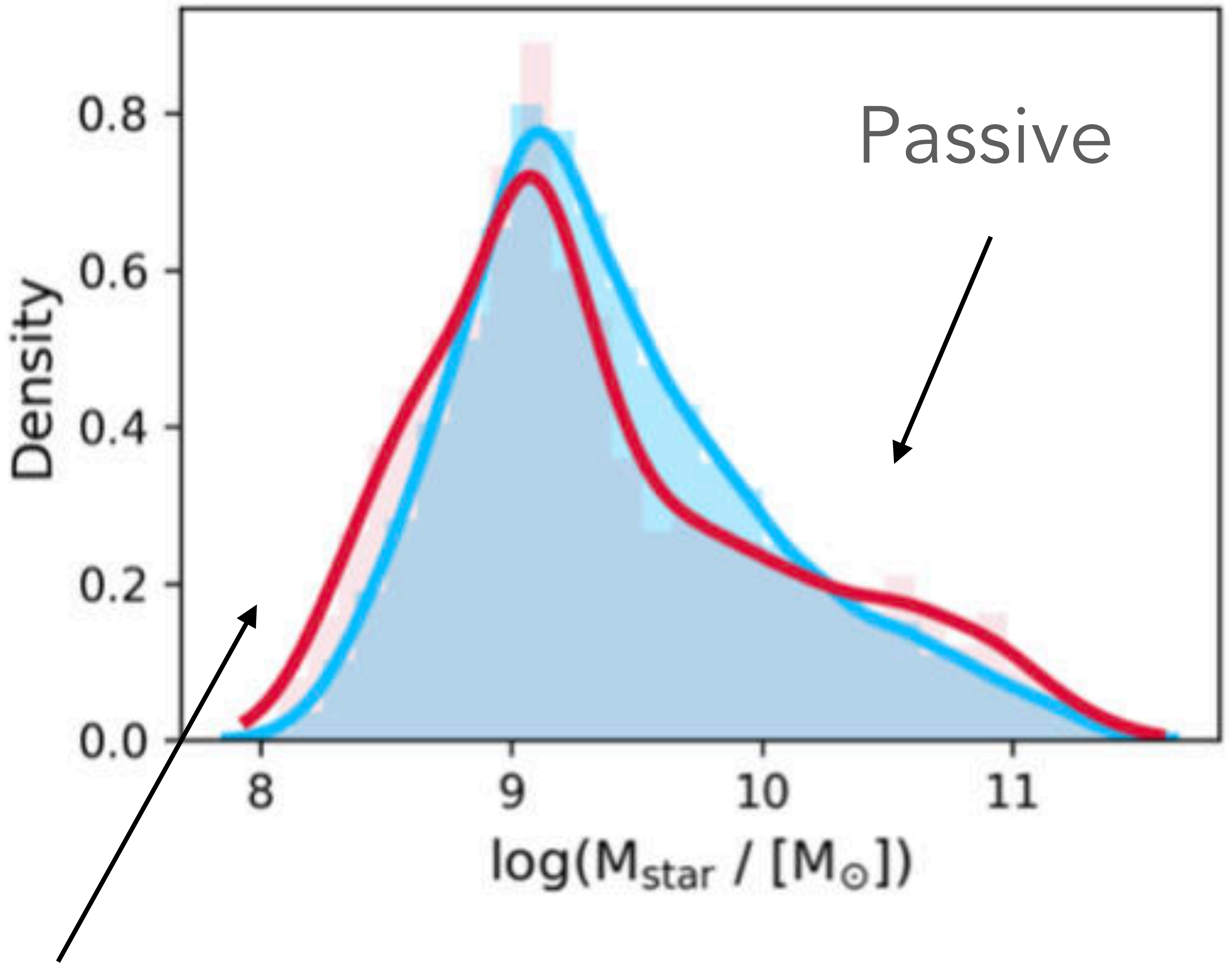




Good vs bad predictions 🎯

$$|\Delta z| < 0.01$$

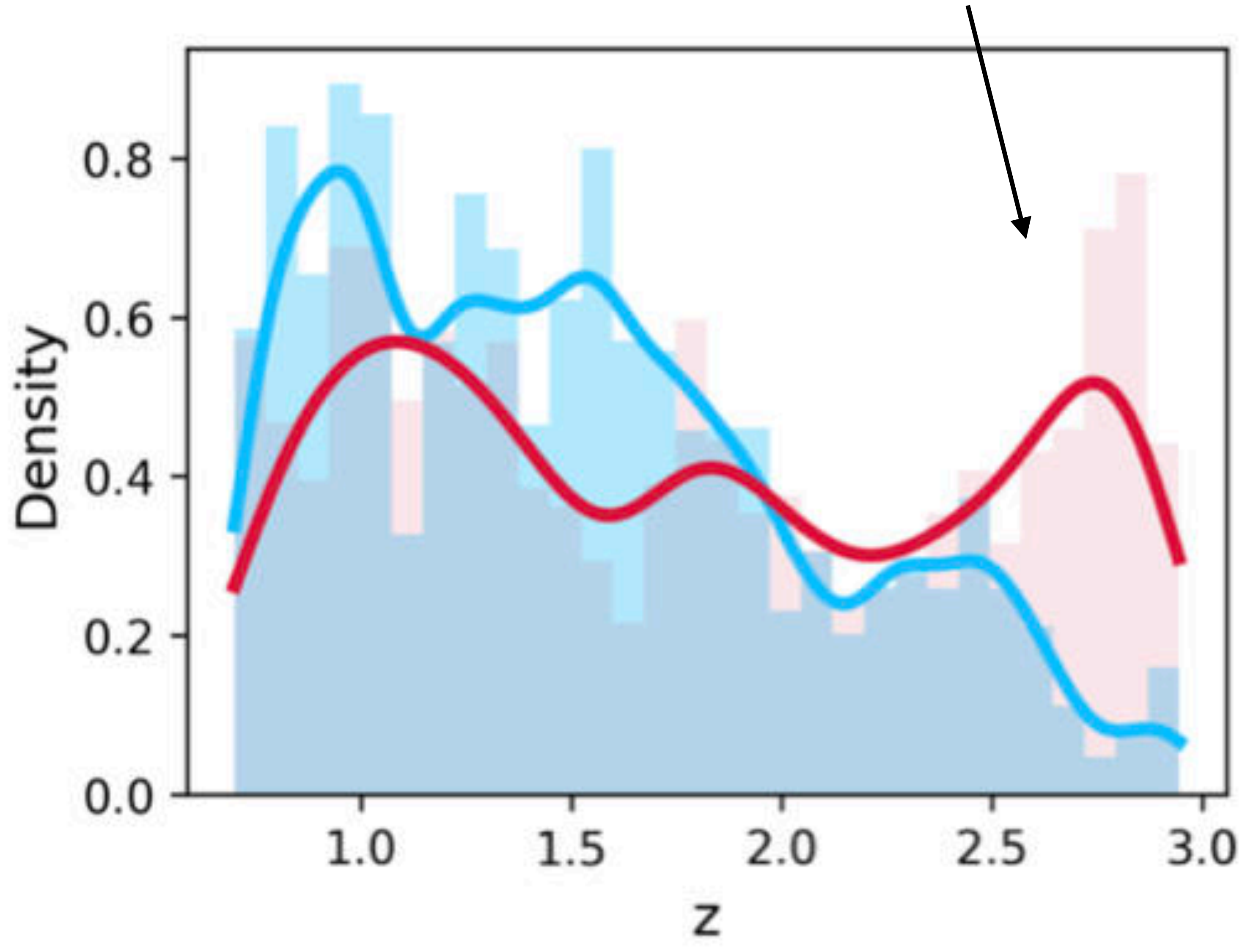
- 🟦 Well Classified
- 🟥 Poorly Classified



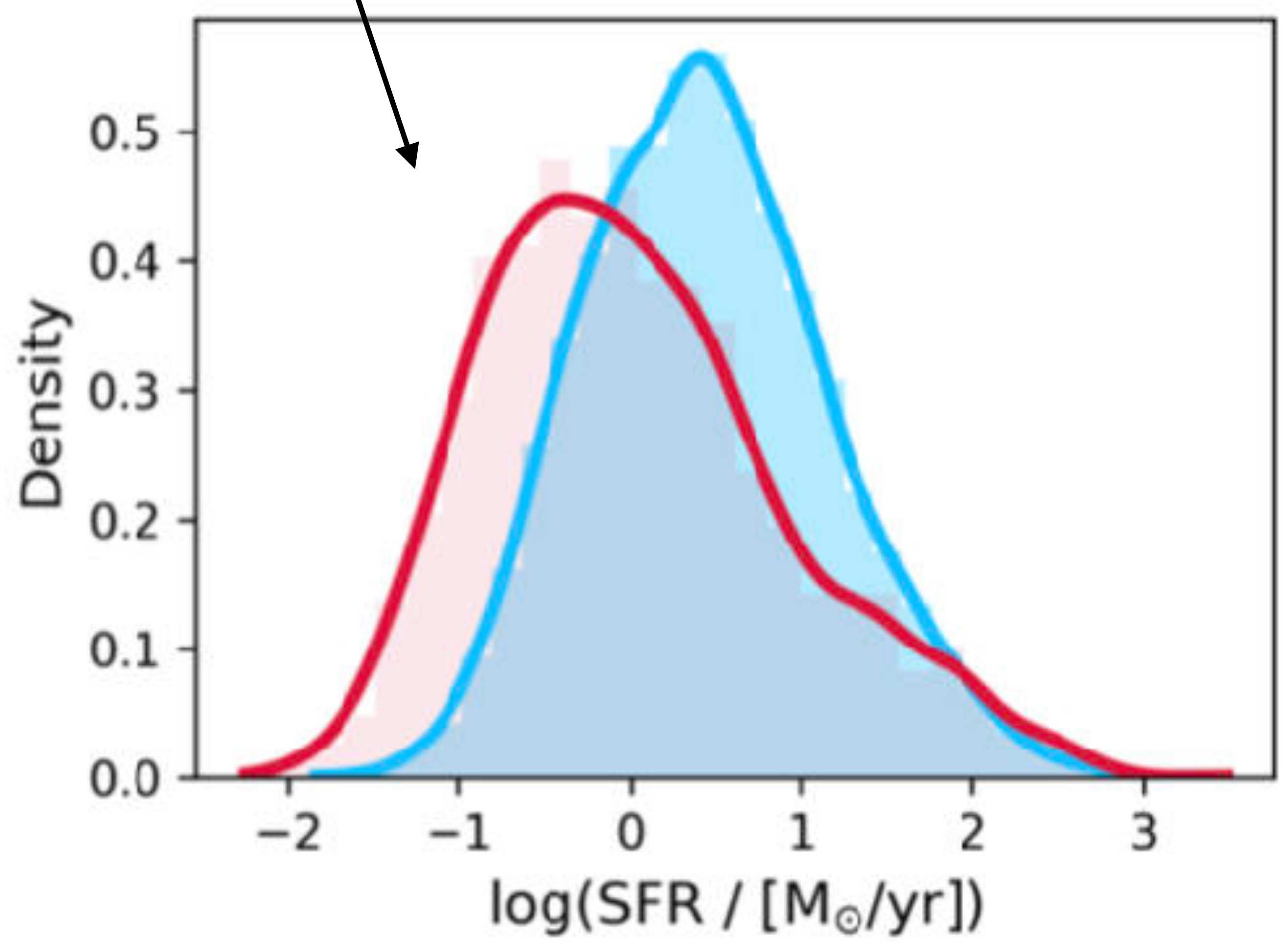
Passive

No ongoing SF

No bright emission lines

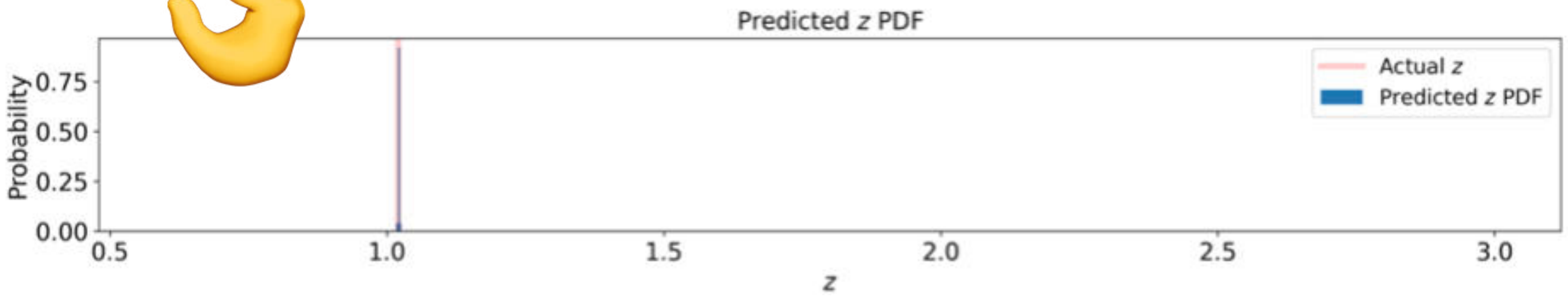


No ongoing SF

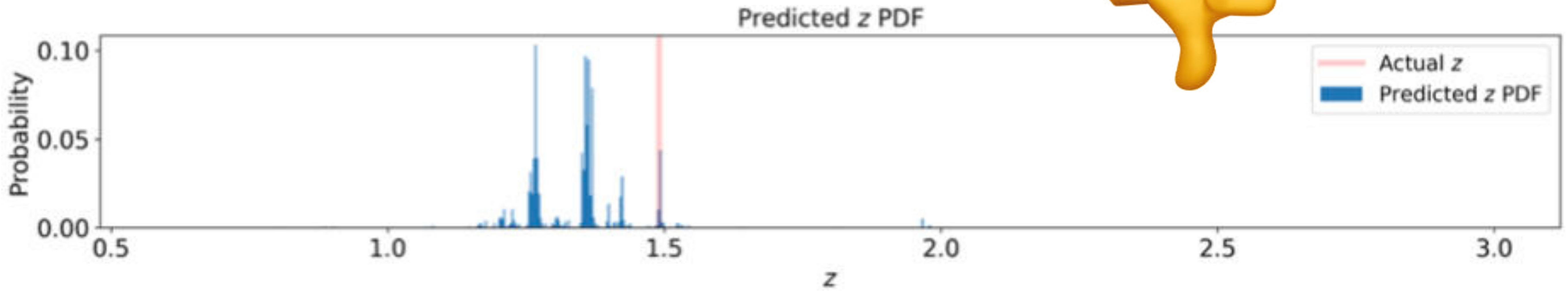


Analysis and screening of the output redshift PDFs 🤔

👌



👎

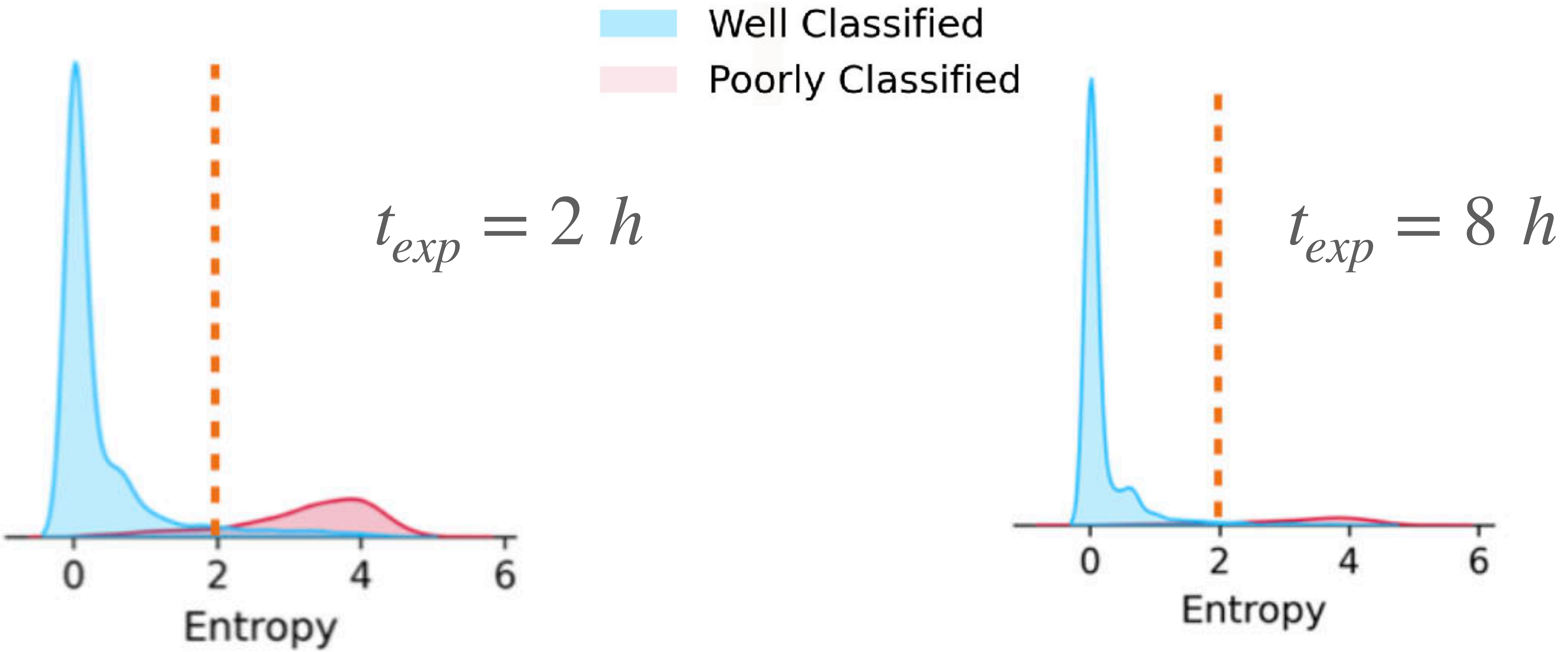


They look different, right?

Intuitively, one might expect that galaxy spectra with unsuccessfully predicted redshifts (for instance, due to the reasons discussed before) would have broader or more dispersed PDFs with multiple peaks.

Can this intuition be quantified objectively, potentially enabling further a posteriori screening of the output?

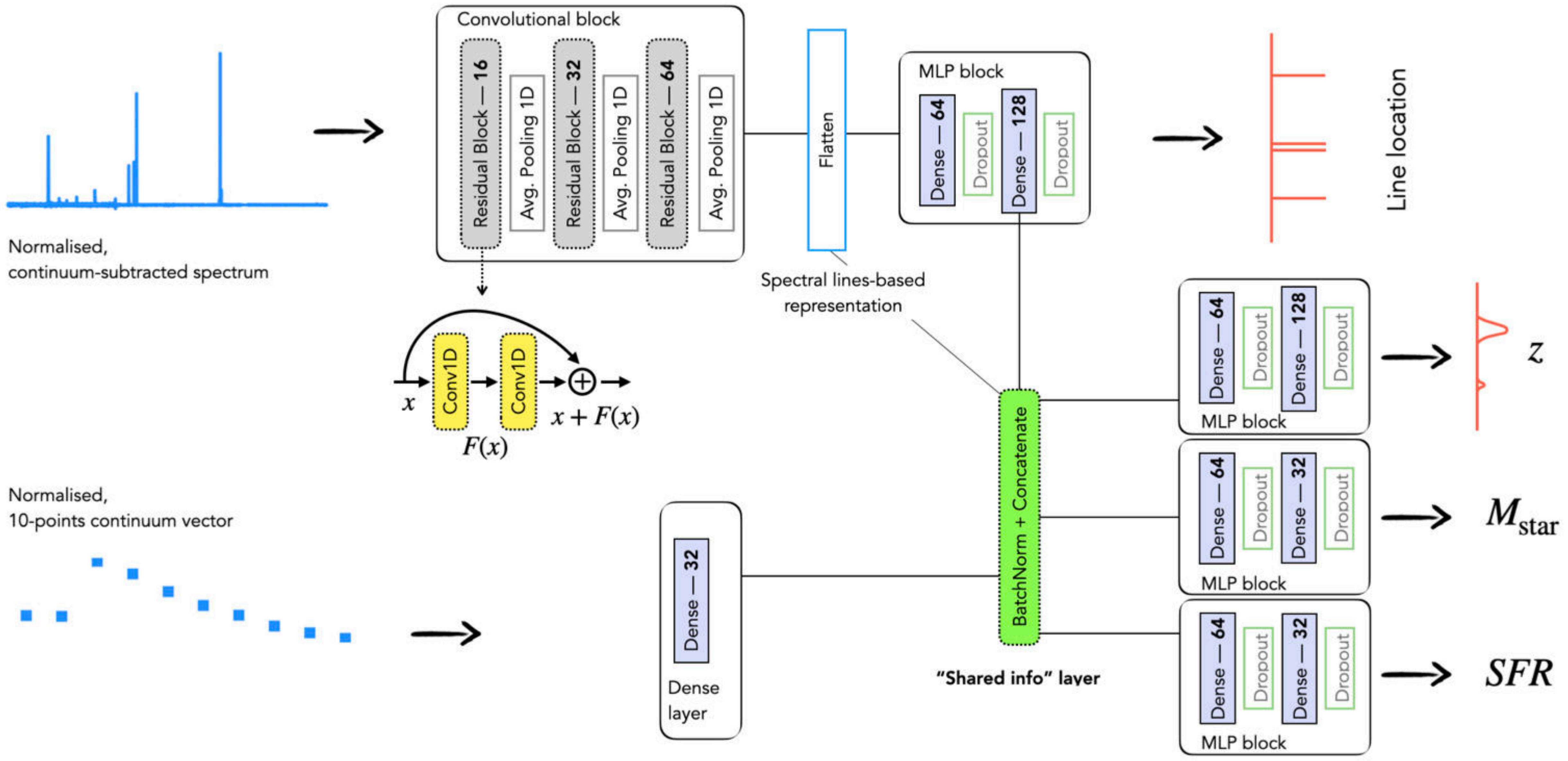
Analysis and screening of the output redshift PDFs 🤔



accuracy before filtering: 0.77;
accuracy after filtering: 0.97
Filtered-out spectra ~27%

accuracy before filtering: 0.9;
accuracy after filtering: 0.99 🤯
Filtered-out spectra ~12%

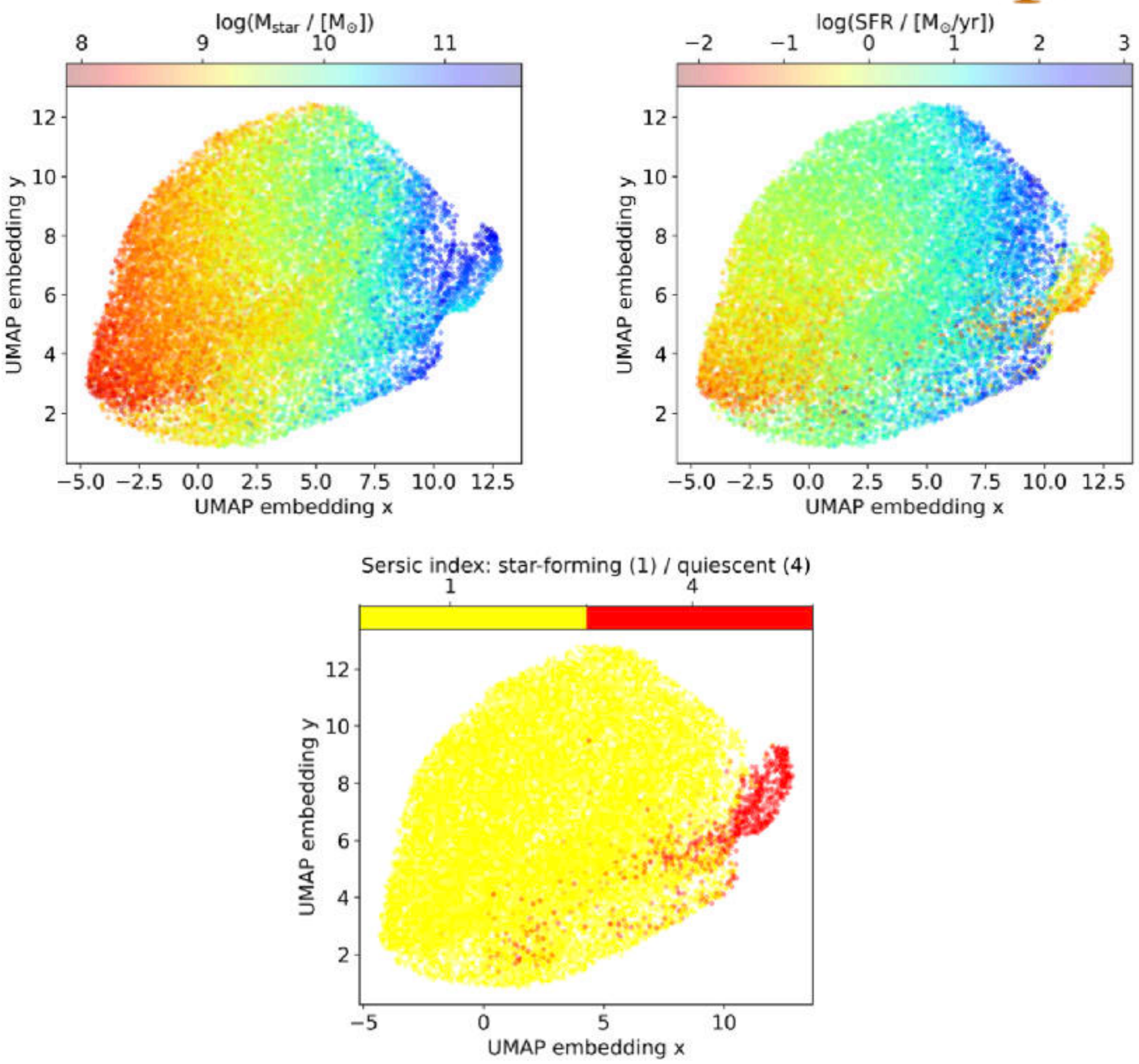
Information encoded in the last embedding layers



~100 D space

UMAP dimensionality reduction

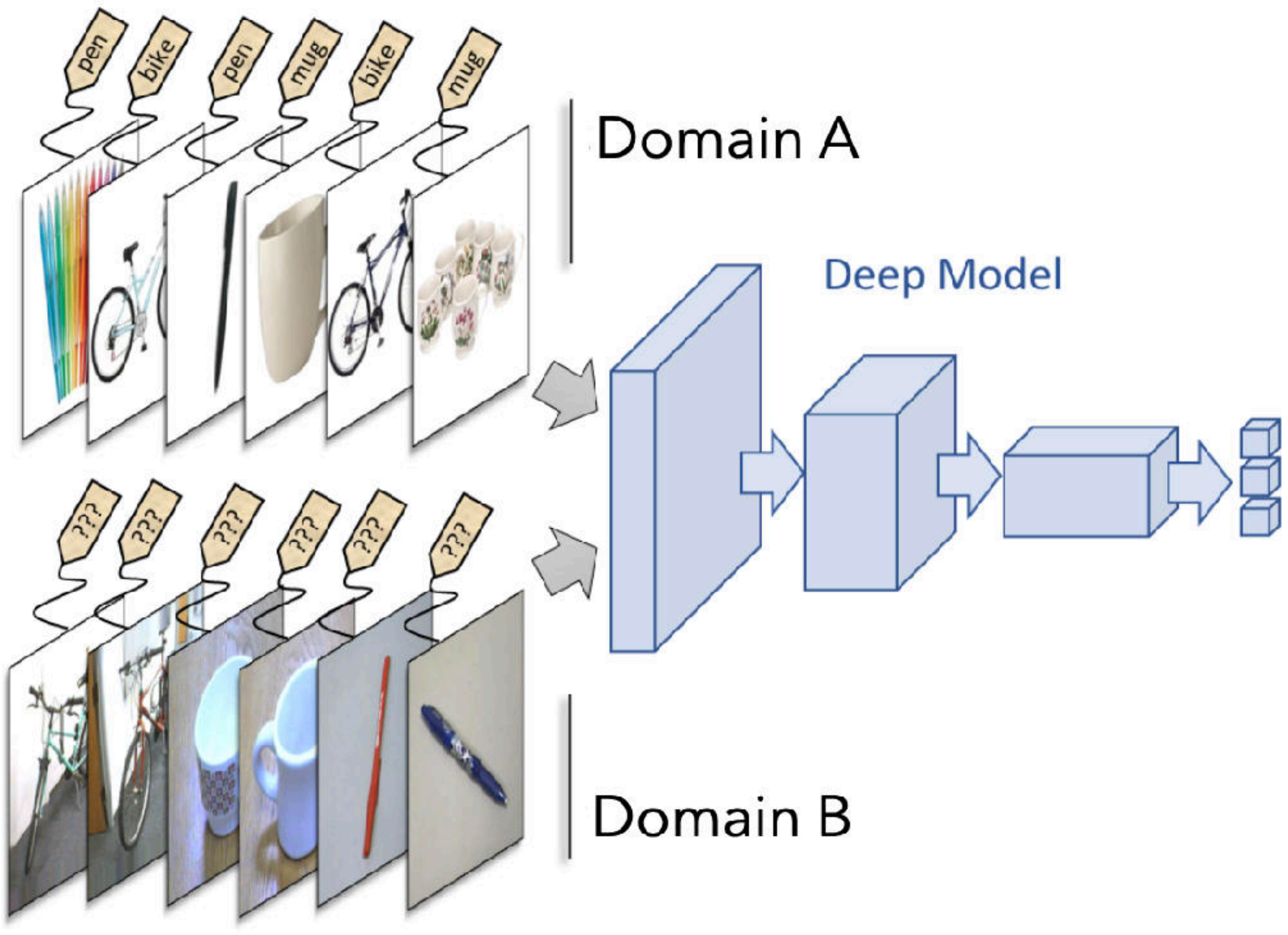
2-D projection



The model encodes information that can be described through classes and quantities that "make sense" and that the model has not seen in training — explainable learning.

Mini bibliography — Portillo et al. 2020; Pat et al. 2020; Liang et al. 2023; Stoppa et al. 2023; Sarmiento et al. 2021; Melchior et al. 2023; Huertas-Company & Lanusse 2023

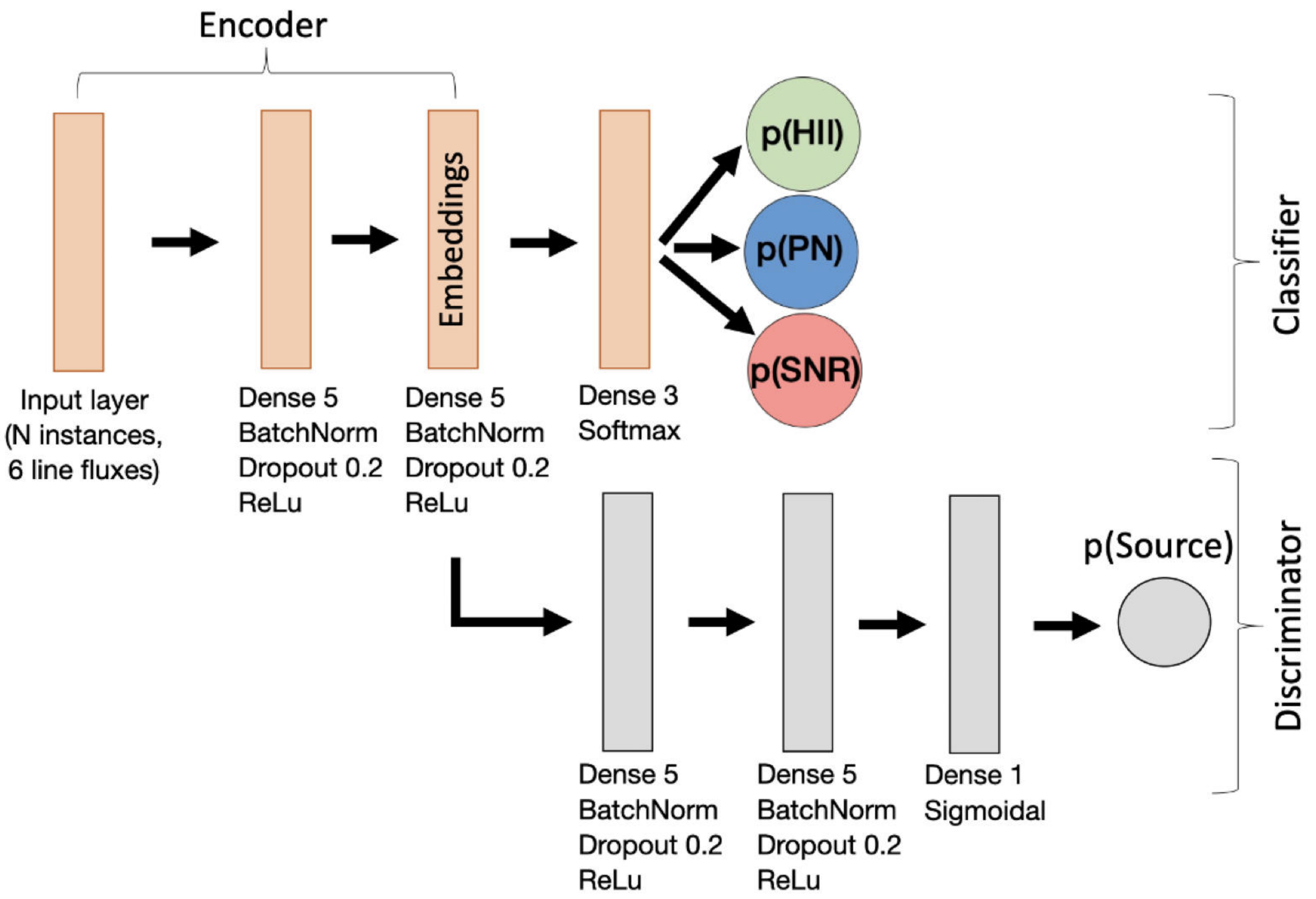
Domain Gap



Neural networks learn by optimising their performance on the training set.

They tend to converge to domain-specific solutions.

A domain adversarial neural net to classify nebulae in the ISM



Encoder

Input layer
(N instances,
6 line fluxes)

Dense 5
BatchNorm
Dropout 0.2
ReLu

Embeddings

Dense 5
BatchNorm
Dropout 0.2
ReLu

Dense 3
Softmax

p(HII)

p(PN)

p(SNR)

Classifier

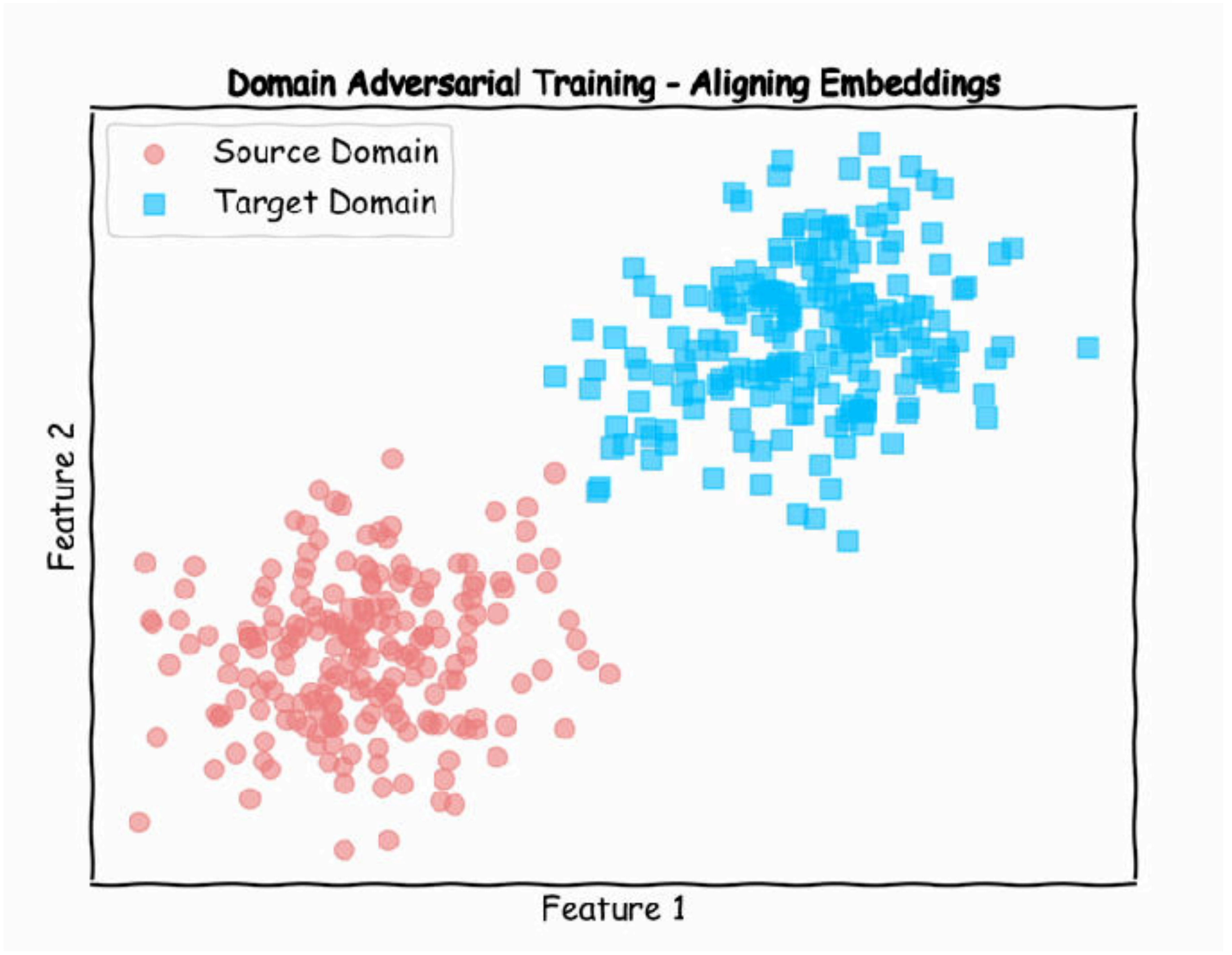
p(Source)

Dense 5
BatchNorm
Dropout 0.2
ReLu

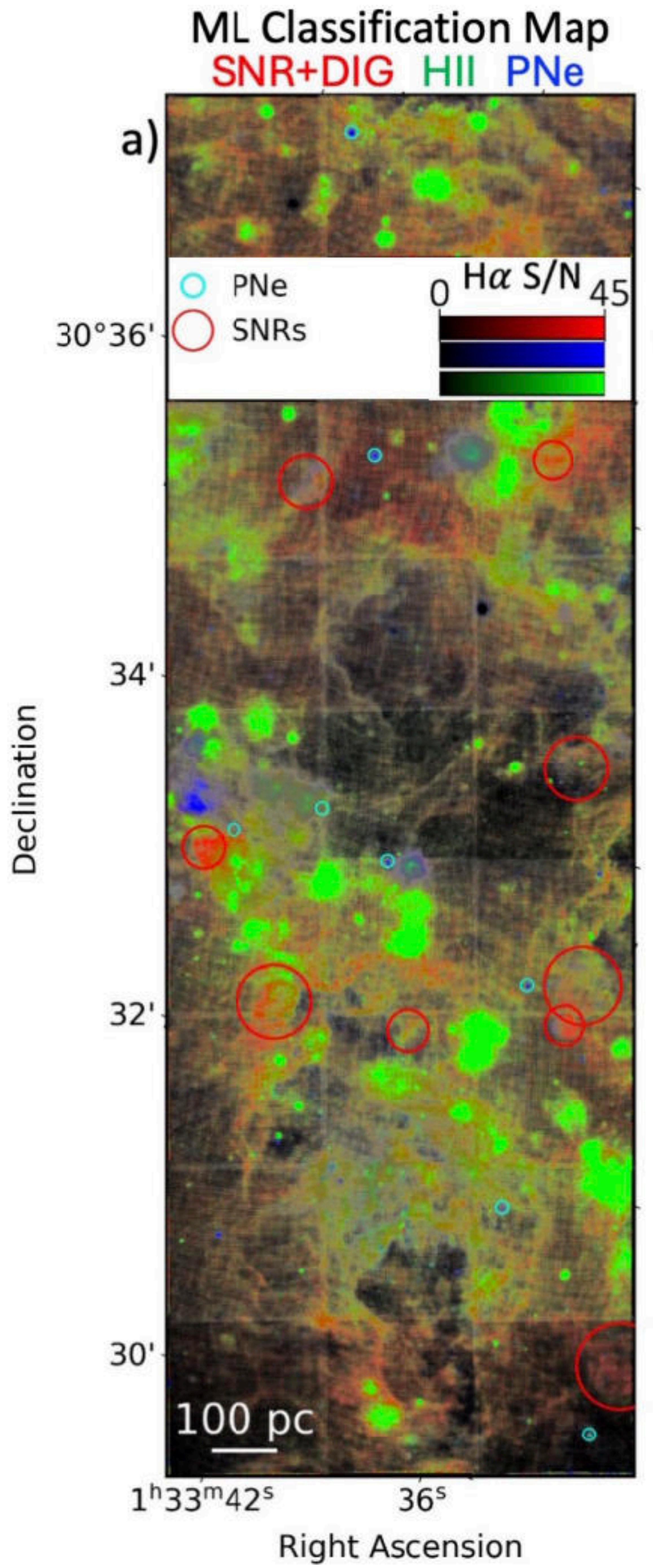
Dense 5
BatchNorm
Dropout 0.2
ReLu

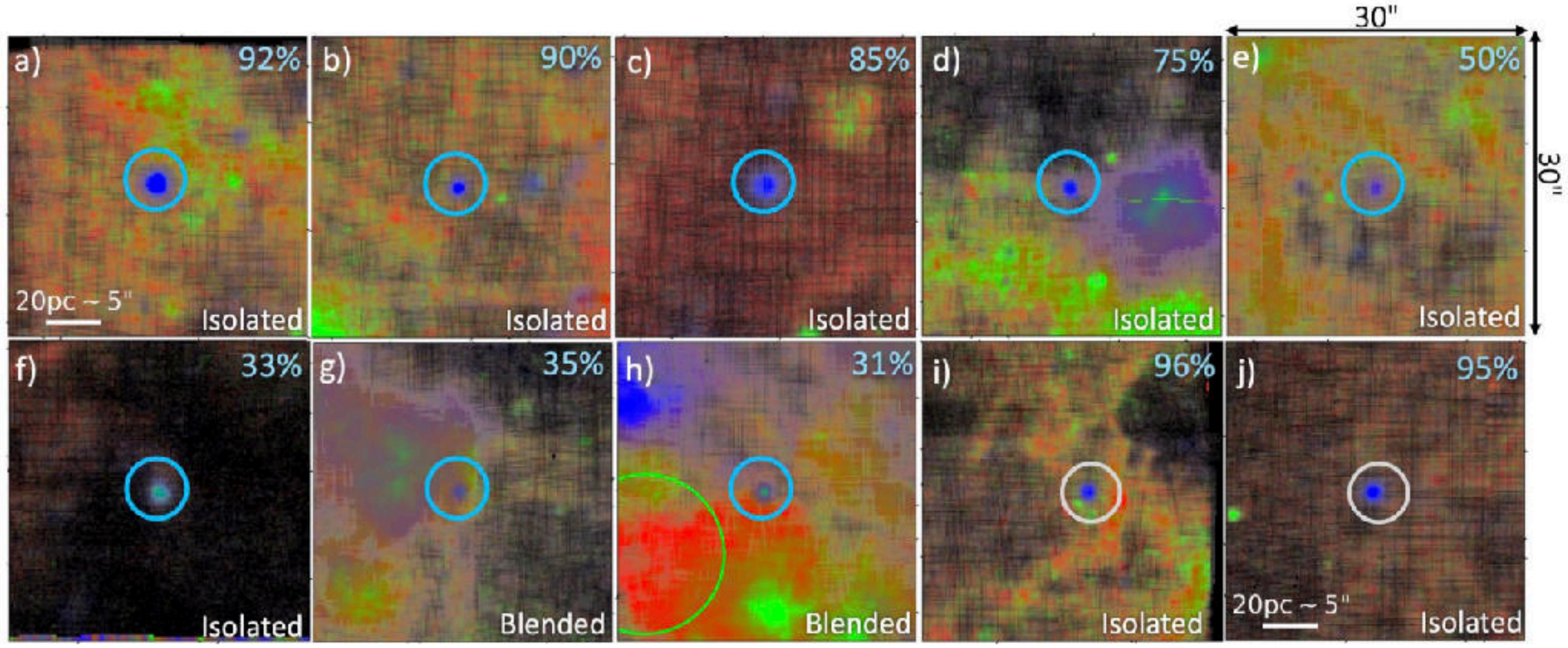
Dense 1
Sigmoidal

Discriminator



Belfiore, Ginolfi+ 25

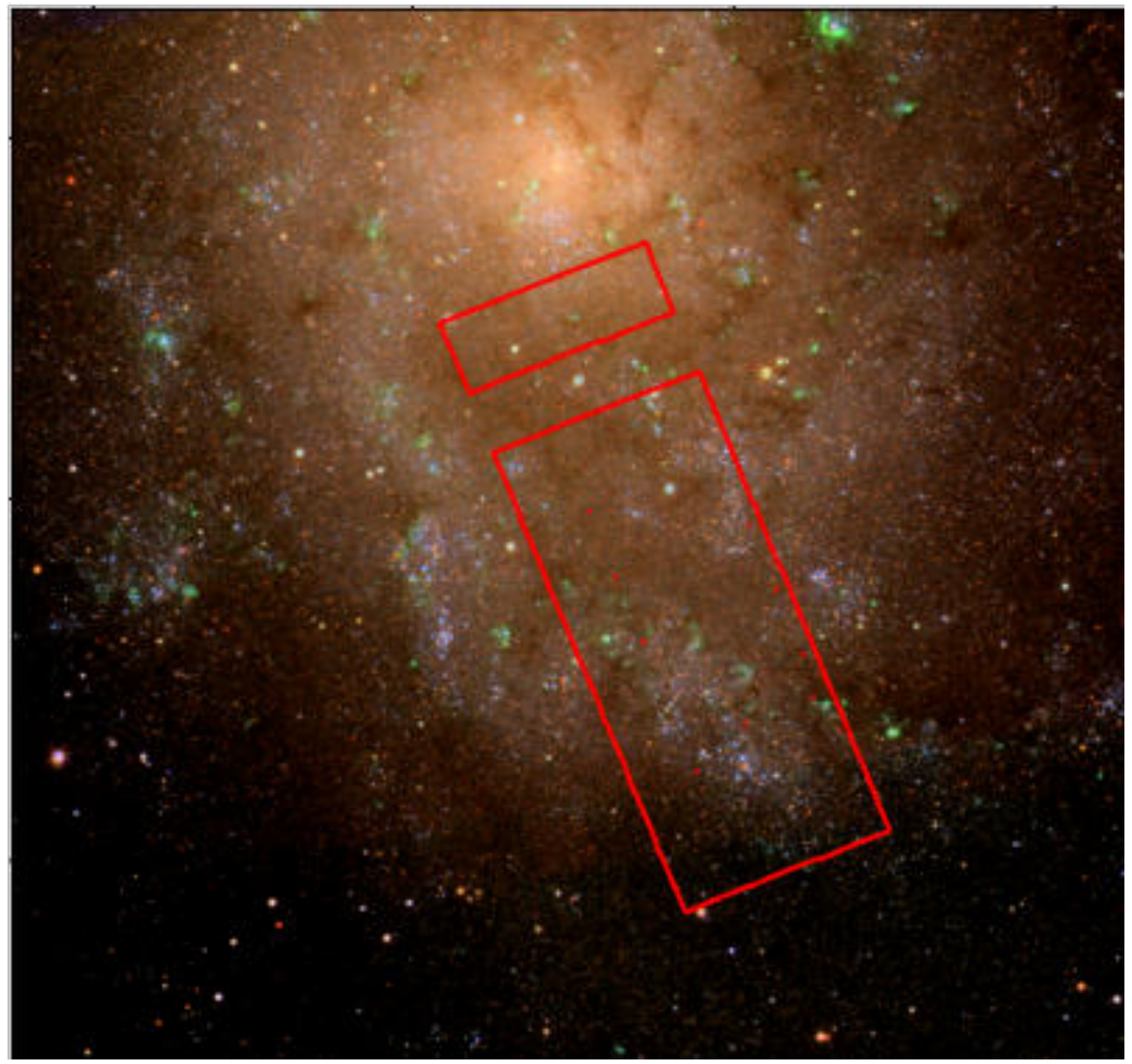




HII region

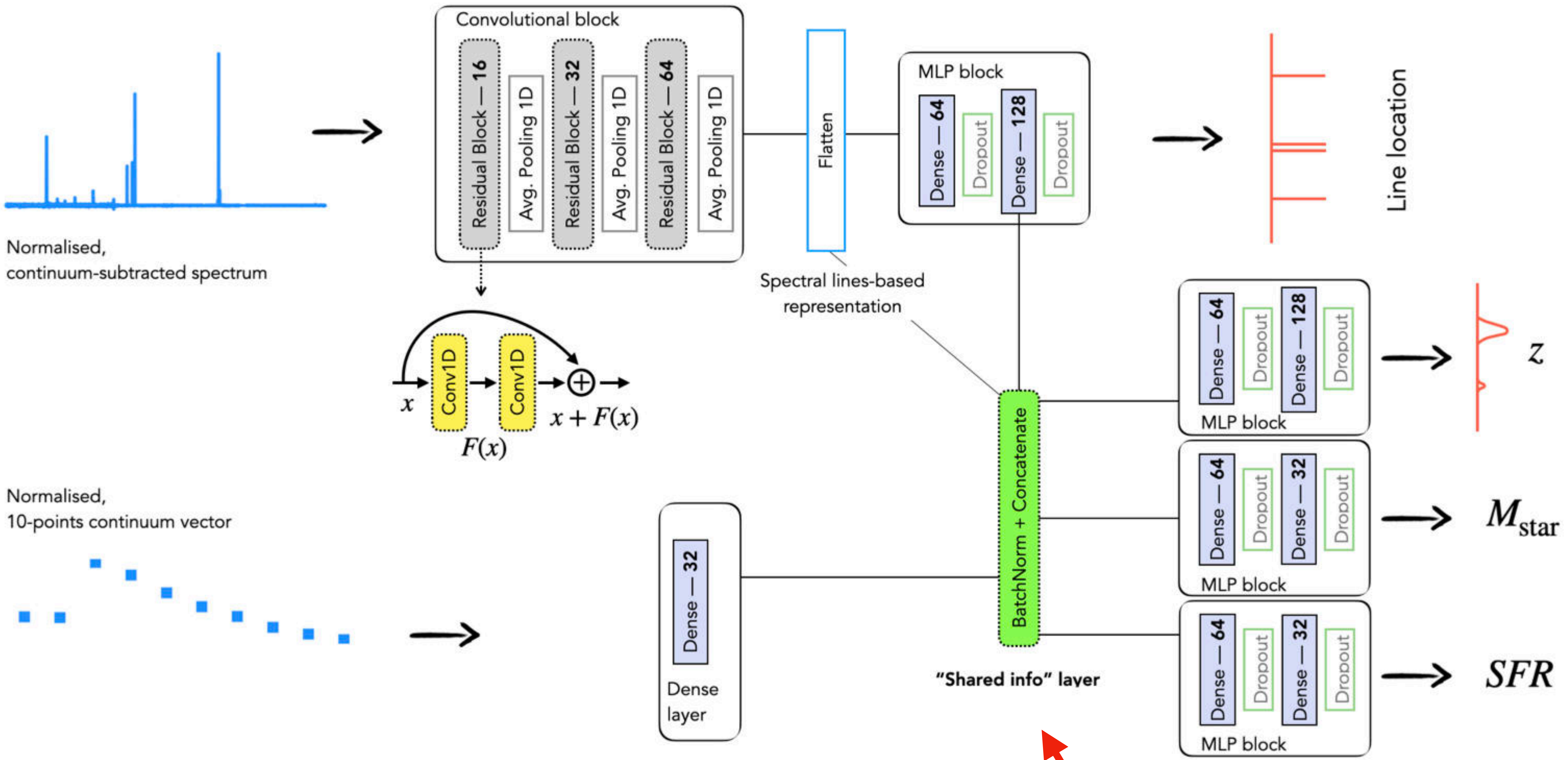
PNe

SNR + DIG



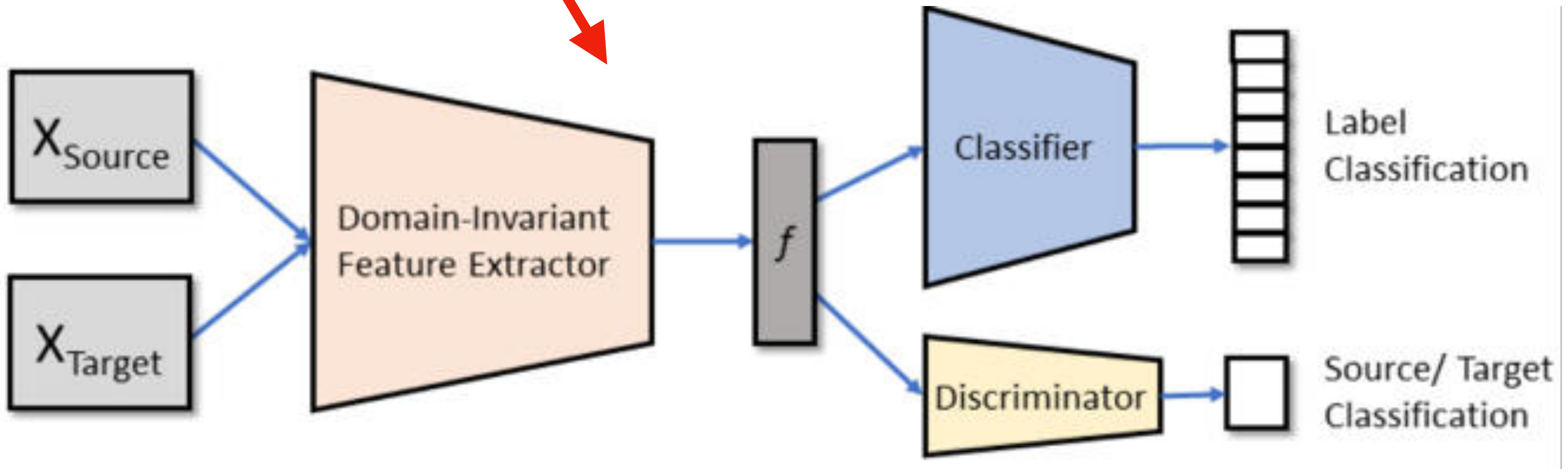
- 24 MUSE pointings 🔭
- 8 billion pixel (spatial + spectral)
- Needs to work well on other IFU data too!

Bracci+25, PhD @florence

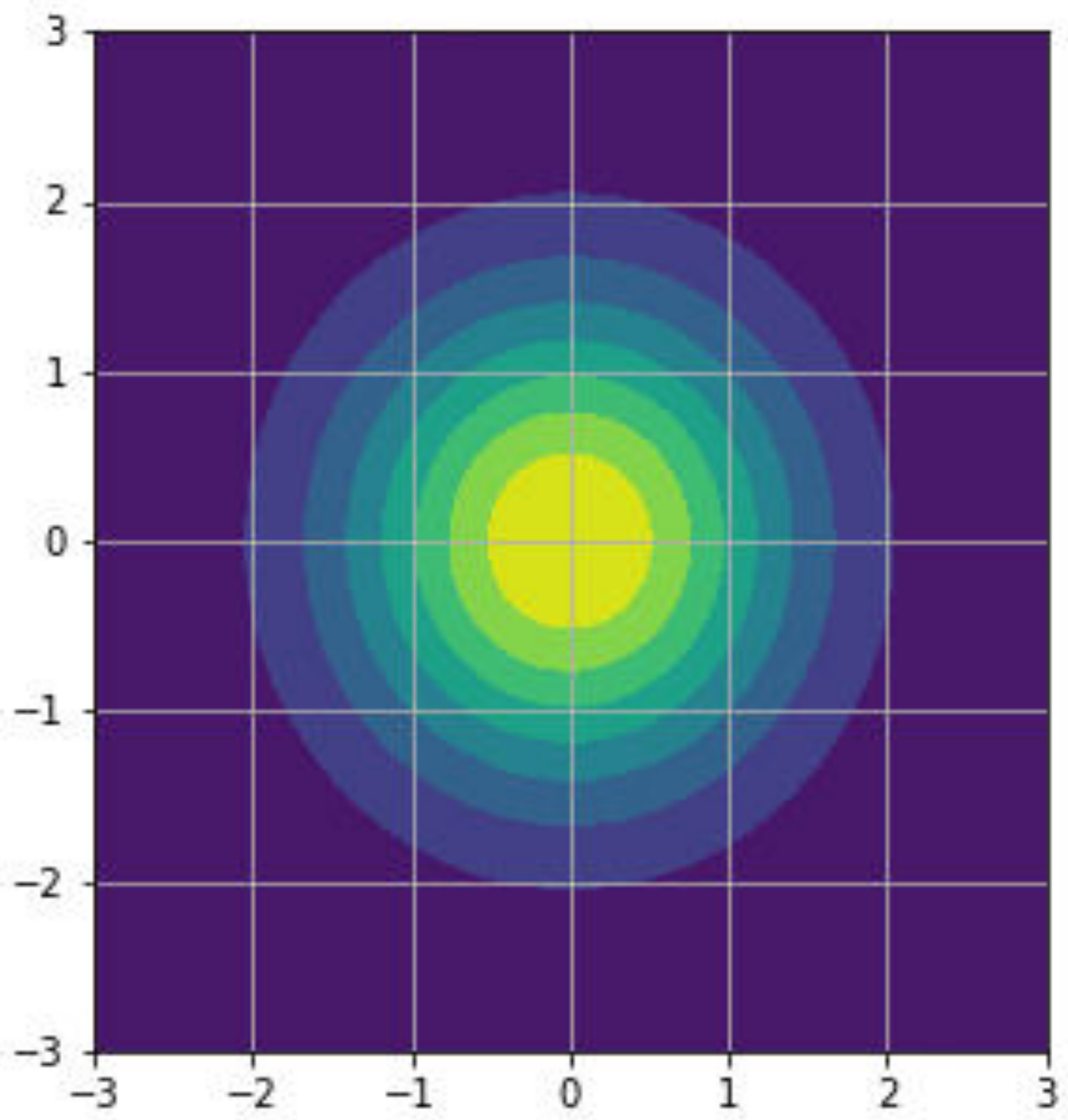


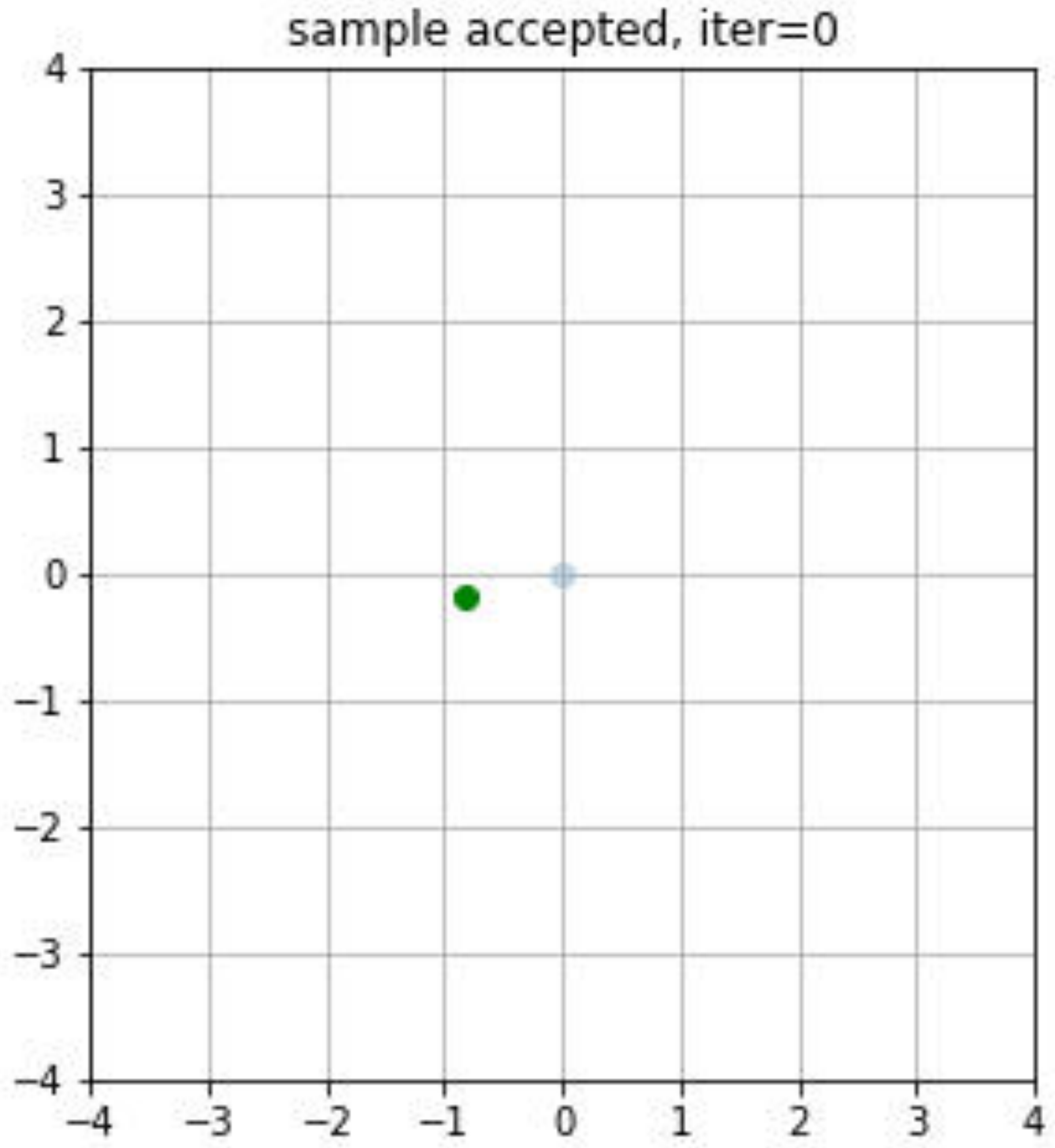
Our plan 👀

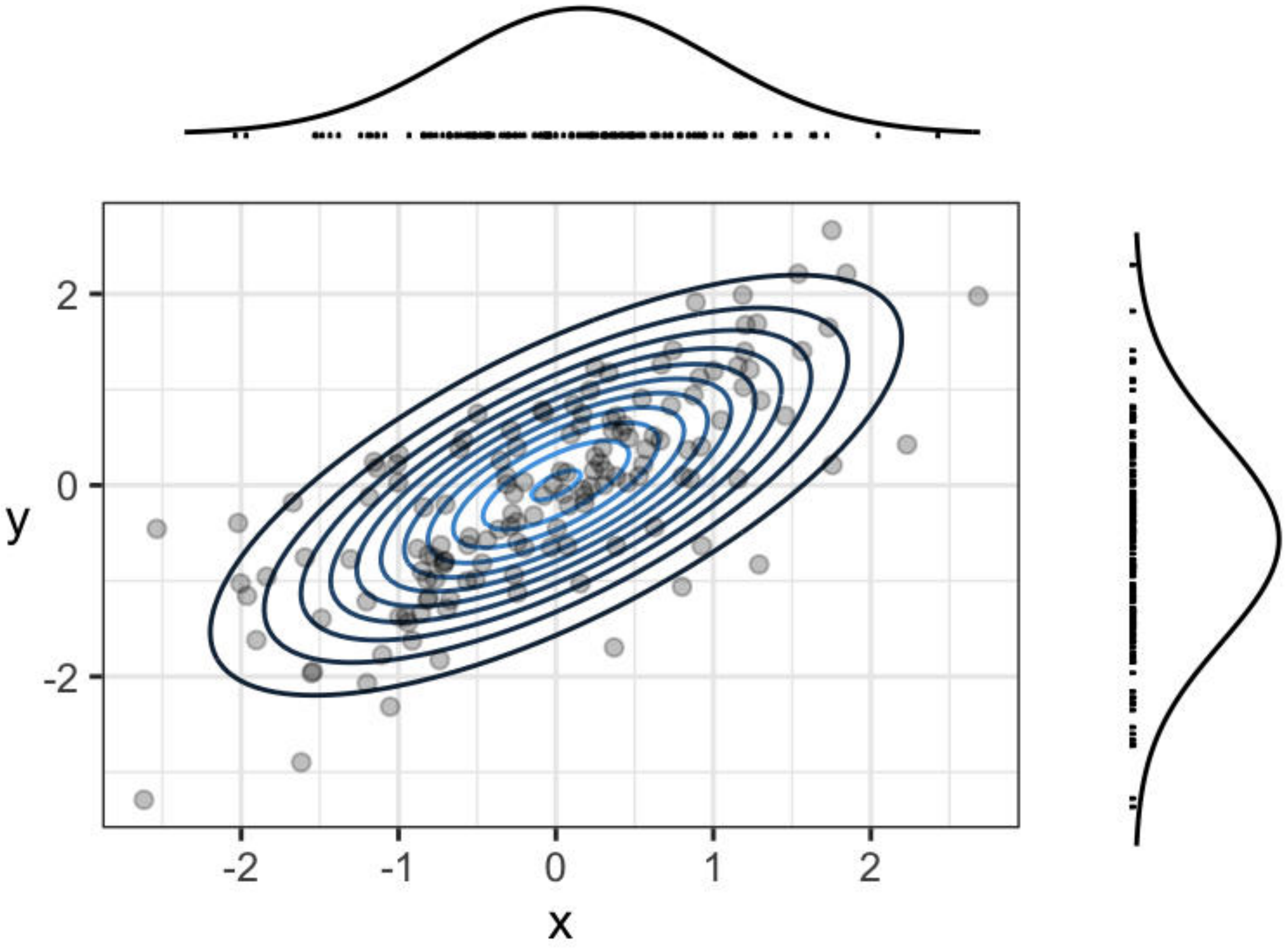
a tool ready-to-use with any data



Degenerate problems → Bayesian approach (e.g., MCMC sampling)

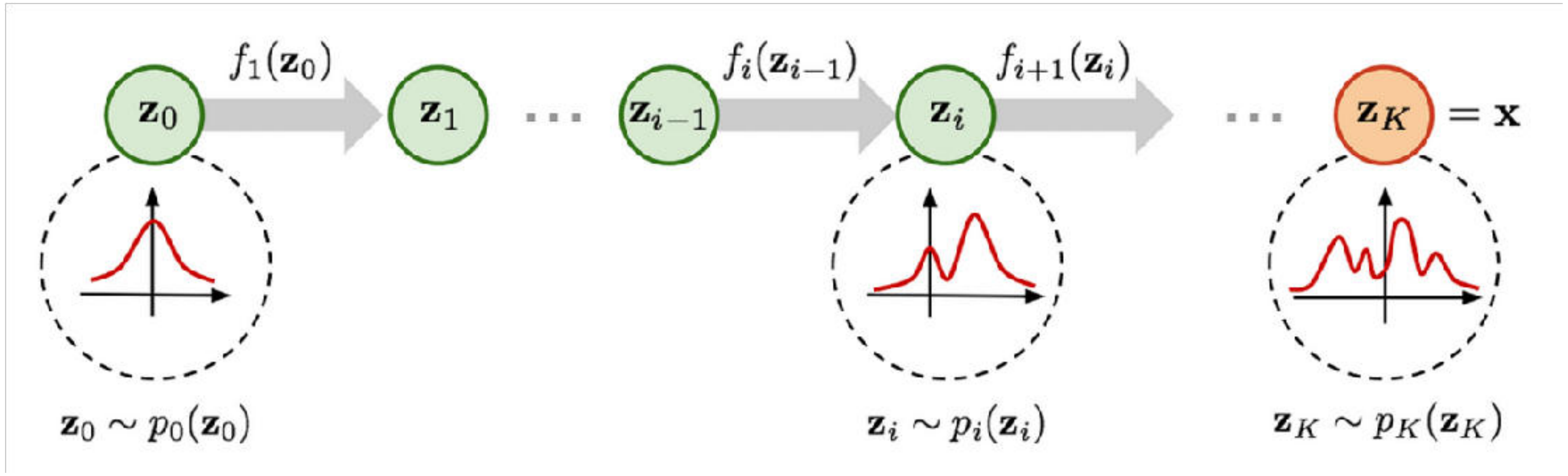


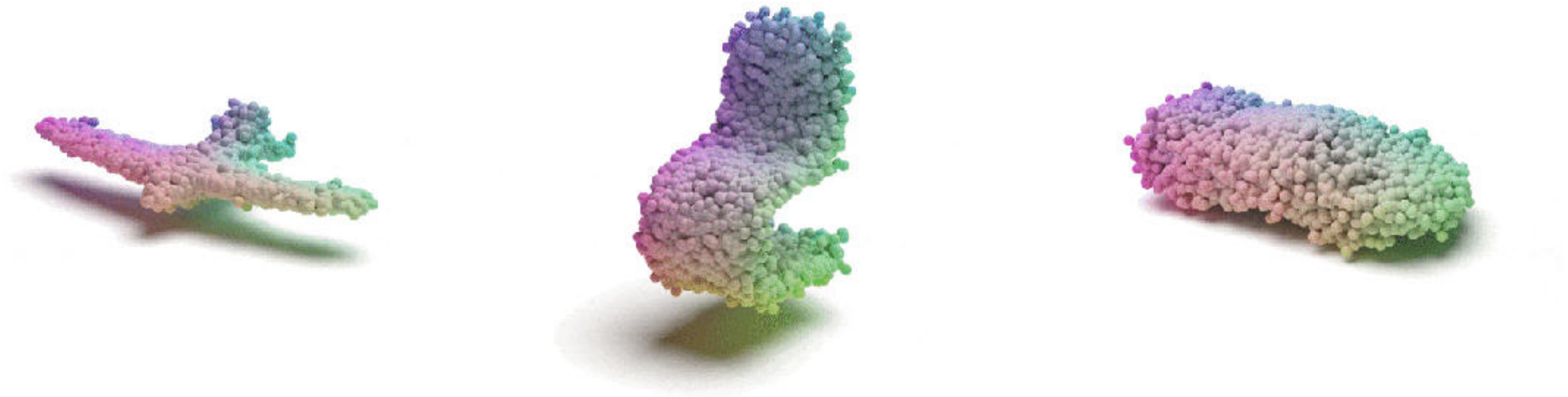




$$P(\theta|data) = \frac{P(data|\theta)P(\theta)}{\int P(data|\theta)P(\theta)\,\mathrm{d}\theta}$$

Simulation-based inference with conditional flows


$f_1(\mathbf{z}_0)$
$\mathbf{z}_0$
$\mathbf{z}_1$
$f_i(\mathbf{z}_{i-1})$
$\mathbf{z}_{i-1}$
$\mathbf{z}_i$
$f_{i+1}(\mathbf{z}_i)$
$\mathbf{z}_K = \mathbf{x}$
$\mathbf{z}_0 \sim p_0(\mathbf{z}_0)$
$\mathbf{z}_i \sim p_i(\mathbf{z}_i)$
$\mathbf{z}_K \sim p_K(\mathbf{z}_K)$

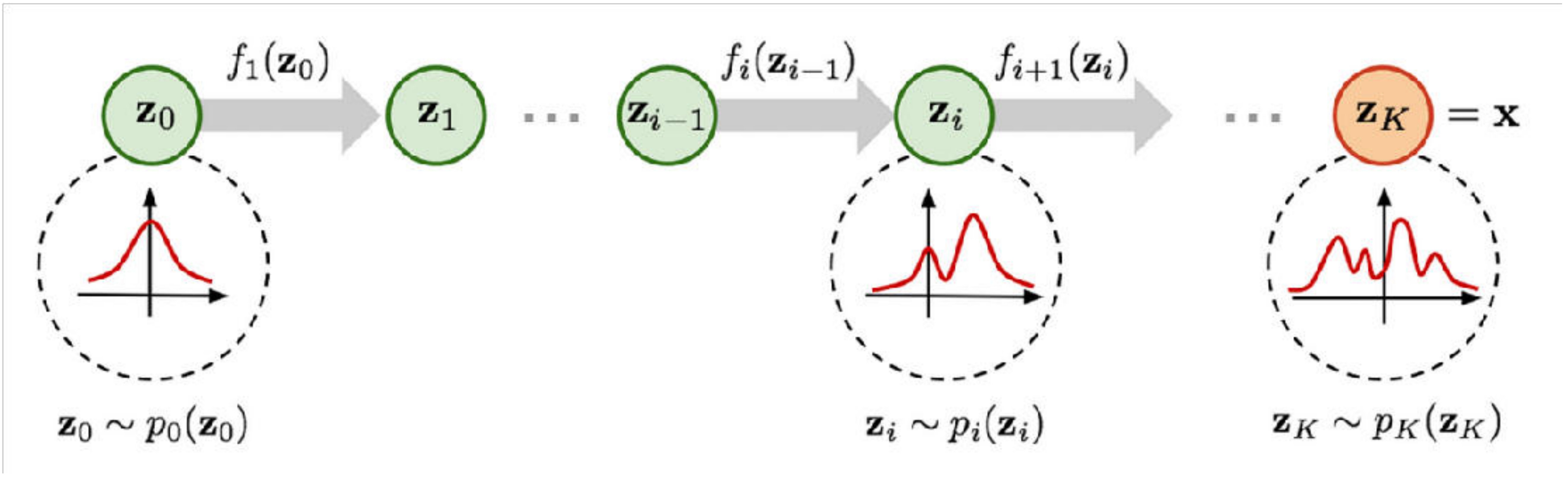


Yang +19

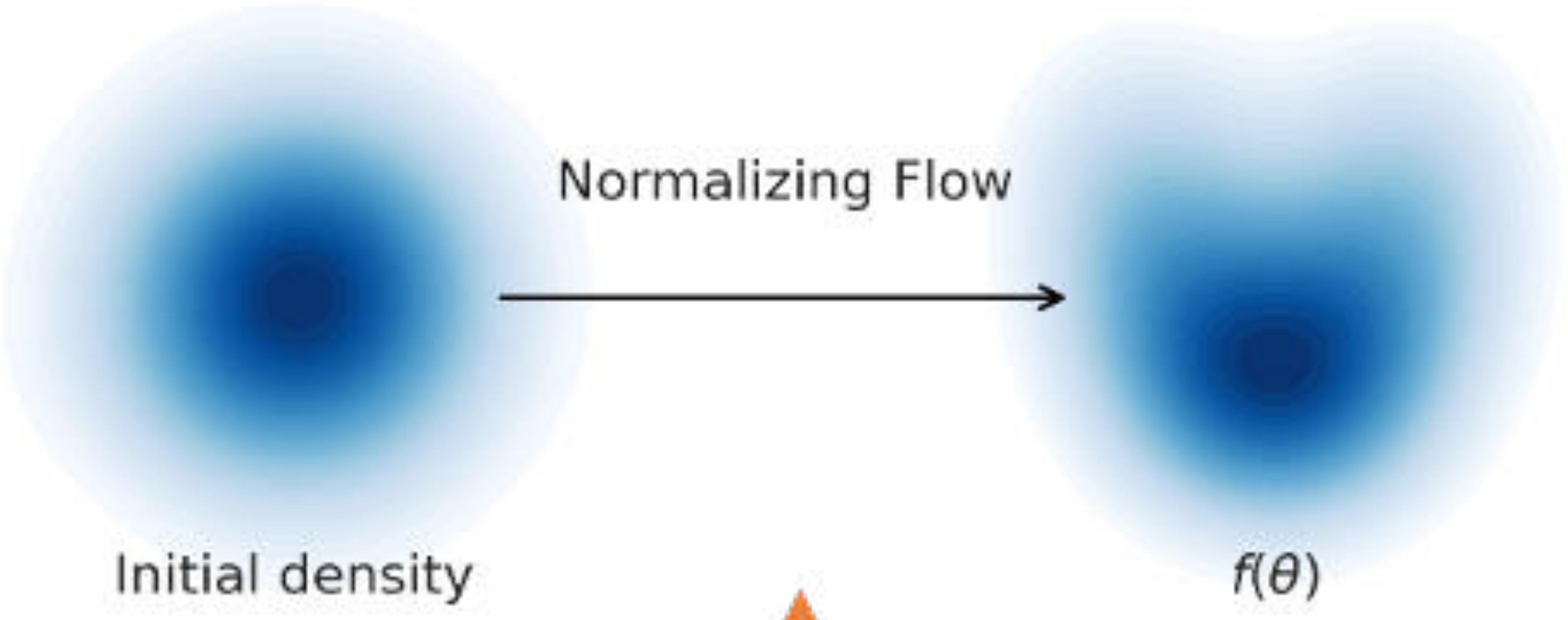
Normalising flows —> they learn invertible transformation from simple to complex distribution

Mini bibliography for simulation based inference in astro using conditional flows — List et al. 2021; Mishra-Sharma & Cranmer 2021; Hahn et al. 2023a,b; Lemos et al. 2023; Bhardwaj et al. 2023; Alvey et al. 2023; Aubin et al. 2023; Hahn et al. 2024; Massara et al. 2024; Candebat et al. 2024; Angeloudi et al. 2024; Iglesias-Navarro et al. 2024; Barret & Dupourqué 2024 etc etc

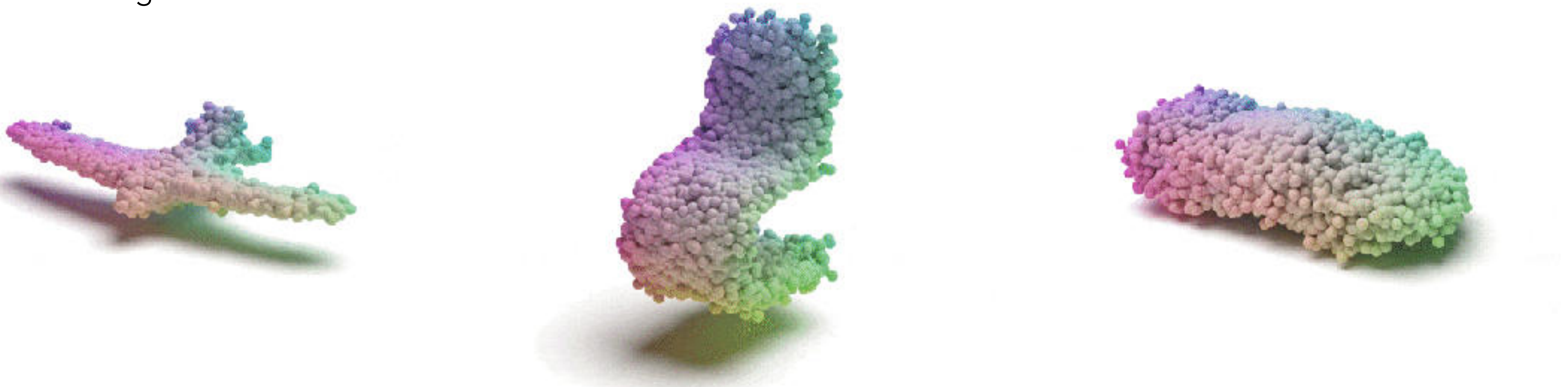
Simulation-based inference with conditional flows



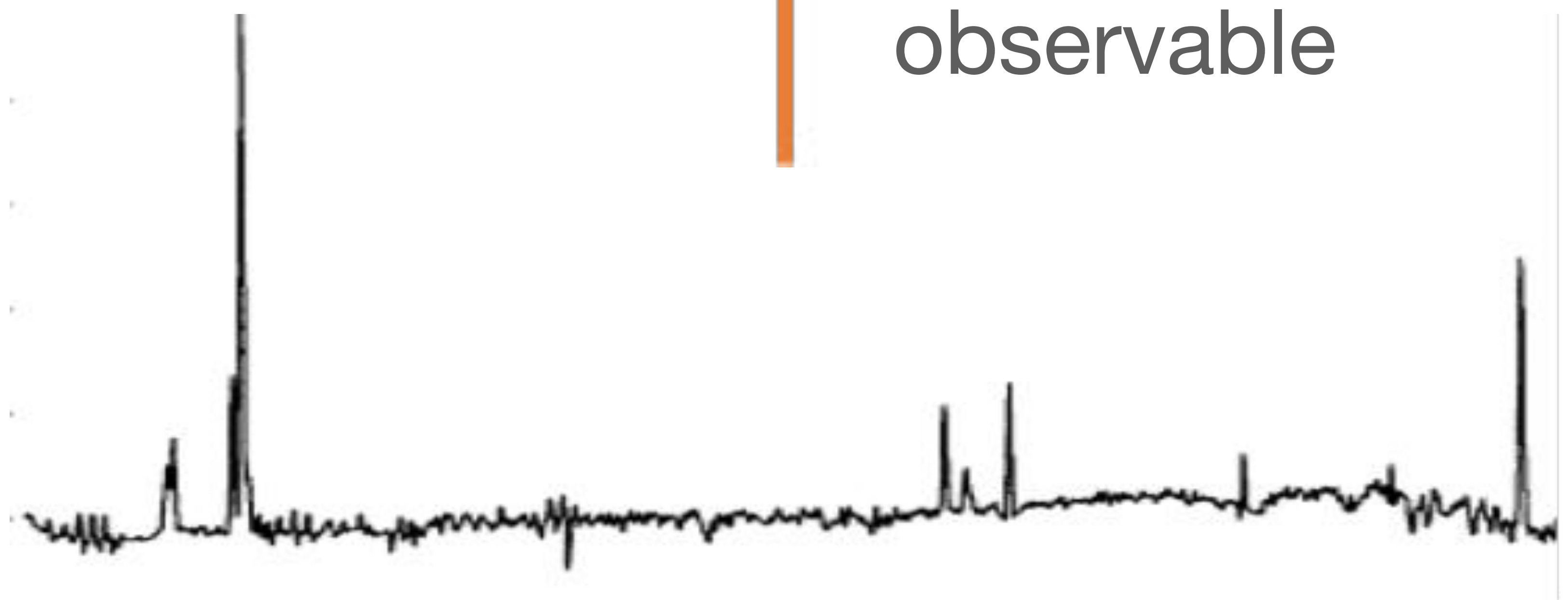
$f_1(\mathbf{z}_0)$
$\mathbf{z}_0$
$\mathbf{z}_1$
$f_i(\mathbf{z}_{i-1})$
$\mathbf{z}_{i-1}$
$\mathbf{z}_i$
$f_{i+1}(\mathbf{z}_i)$
$\mathbf{z}_K = \mathbf{x}$
$\mathbf{z}_0 \sim p_0(\mathbf{z}_0)$
$\mathbf{z}_i \sim p_i(\mathbf{z}_i)$
$\mathbf{z}_K \sim p_K(\mathbf{z}_K)$



Yang +19



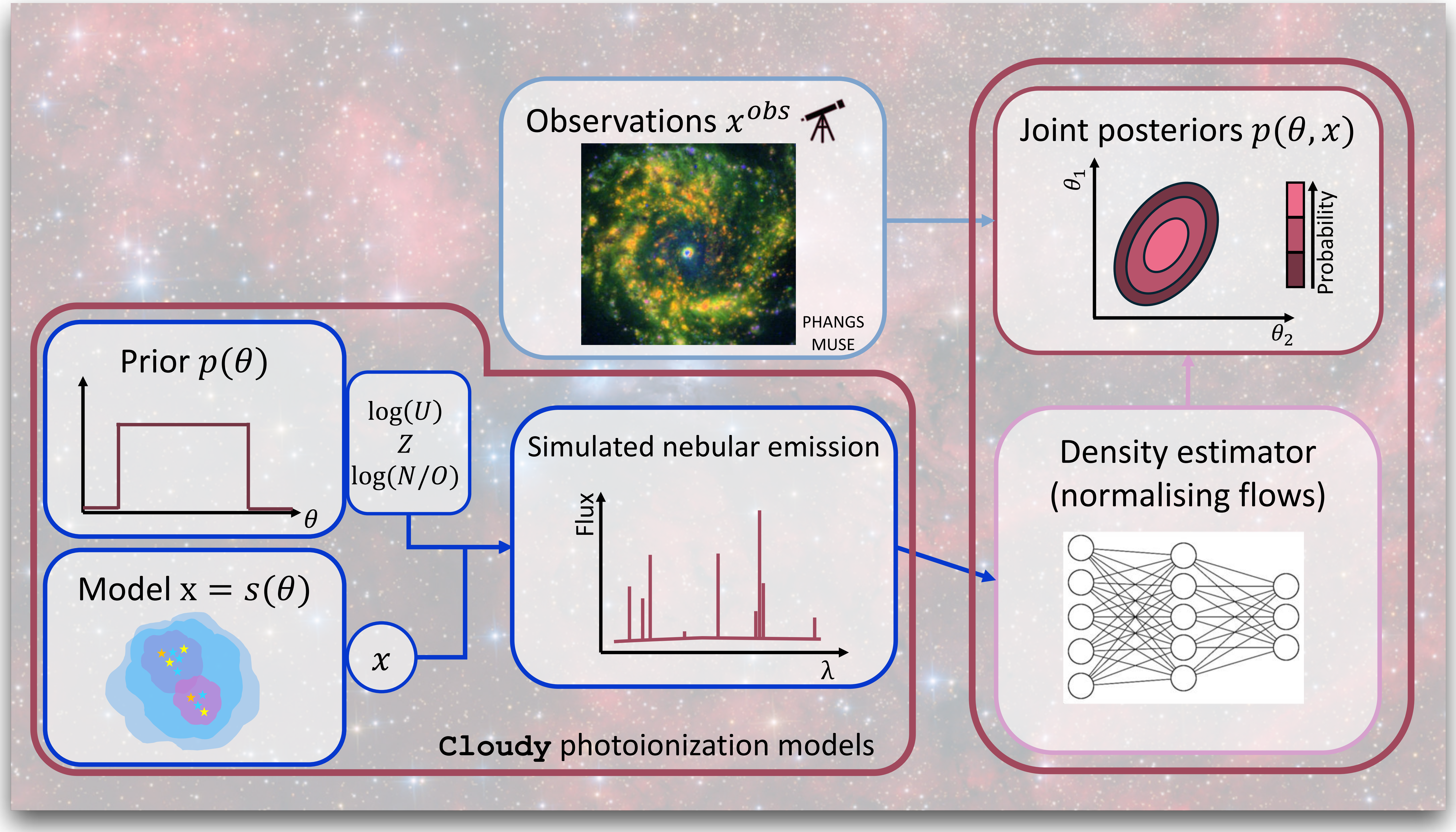
Transformation conditioned to an observable



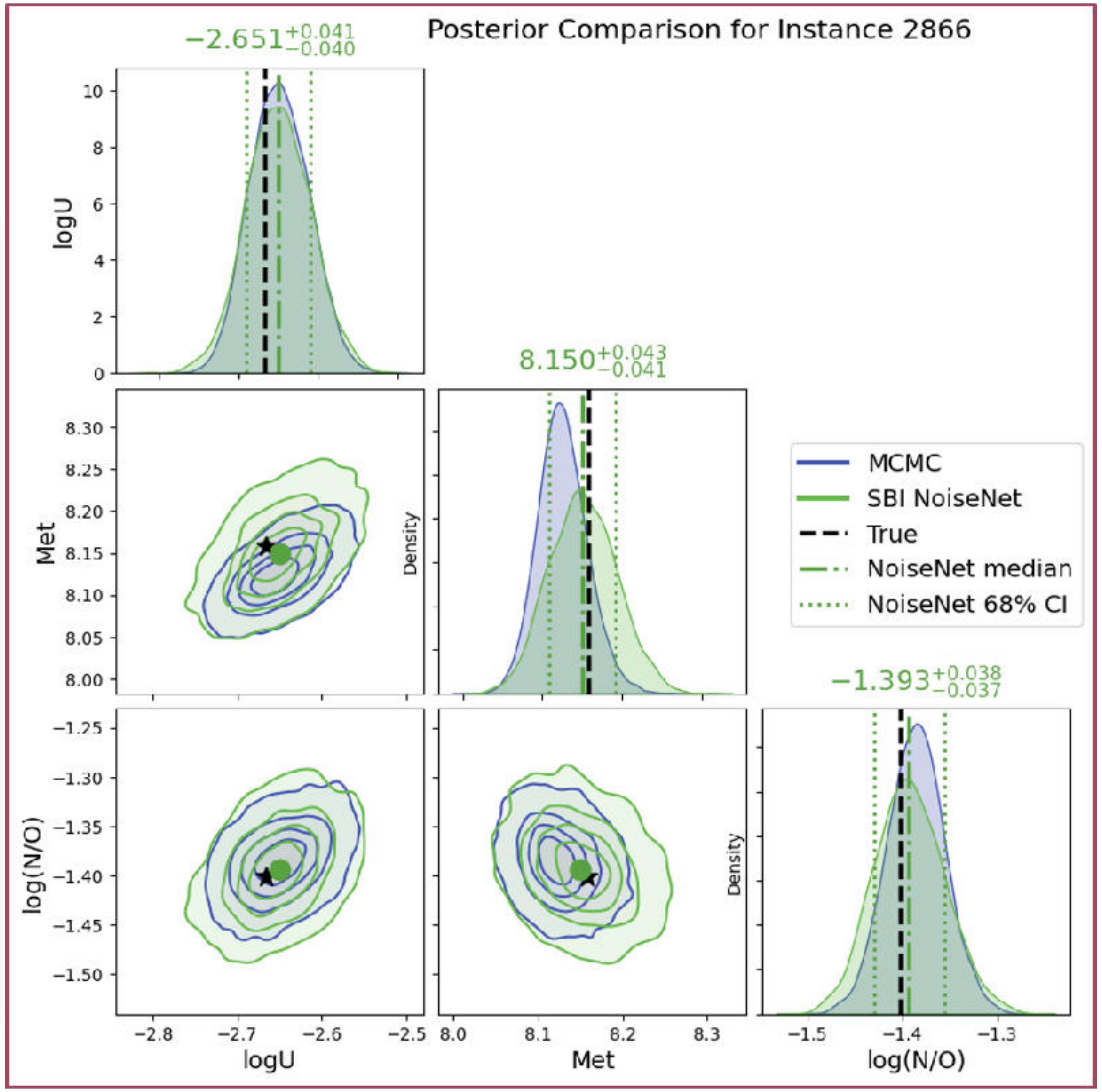
Normalising flows —> they learn invertible transformation from simple to complex distribution

Mini bibliography for simulation based inference in astro using conditional flows — List et al. 2021; Mishra-Sharma & Cranmer 2021; Hahn et al. 2023a,b; Lemos et al. 2023; Bhardwaj et al. 2023; Alvey et al. 2023; Aubin et al. 2023; Hahn et al. 2024; Massara et al. 2024; Candebat et al. 2024; Angeloudi et al. 2024; Iglesias-Navarro et al. 2024; Barret & Dupourqué 2024 etc etc

$\text{Cloudy}^{-1}_{\text{SBI}}$



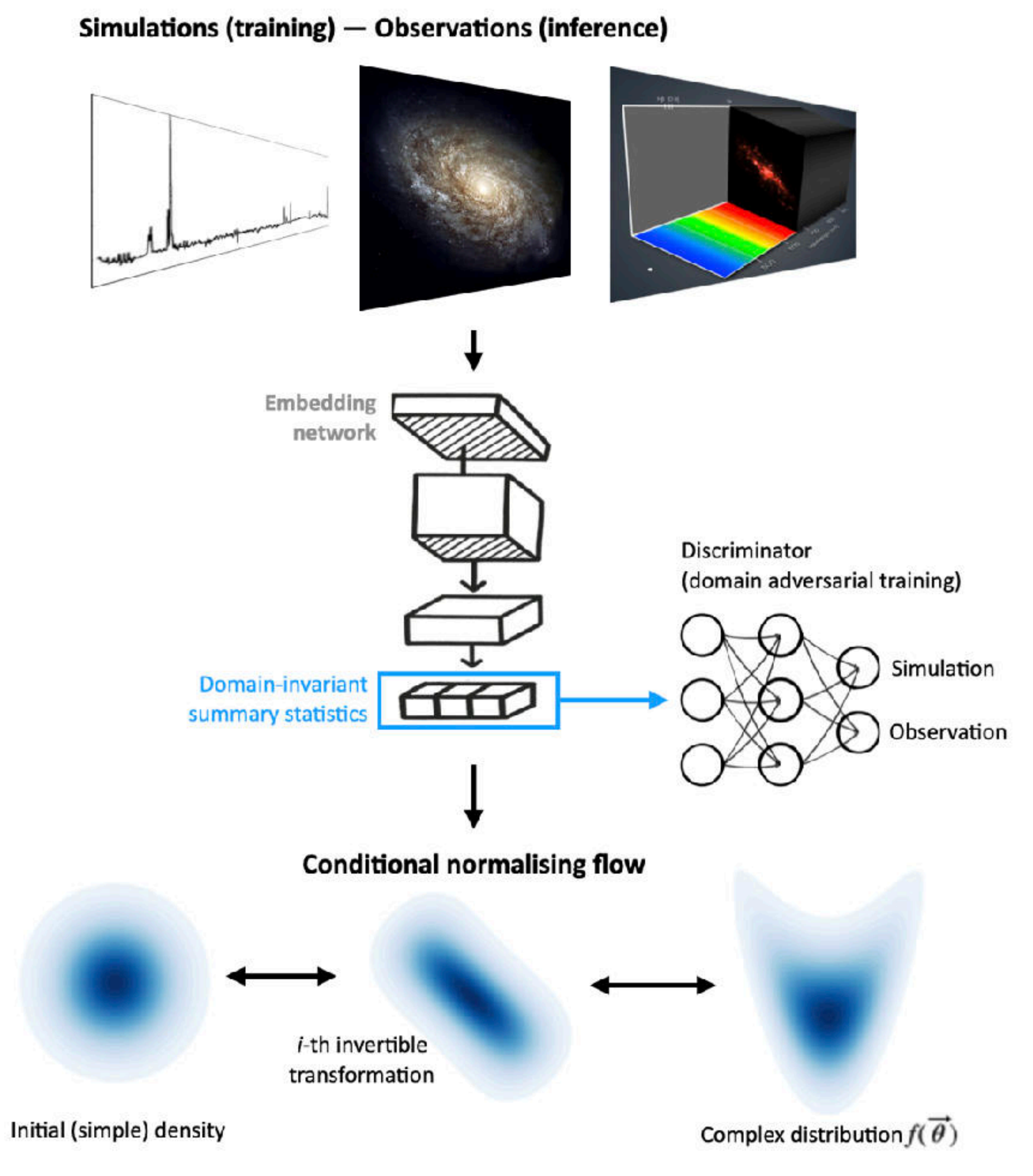
Example: individual instance..



Bracci+, in prep

Advantages: accurate & amortized

A proposed framework: SBI w. domain-invariant neural flows


Simulations (training) — Observations (inference)
Embedding network
Domain-invariant summary statistics
Discriminator (domain adversarial training)
Simulation
Observation
Conditional normalising flow
i-th invertible transformation
Initial (simple) density
Complex distribution $f(\vec{\theta})$

- Use virtual data from state-of-the-art simulators
- Works with any type of data & ensures that summary statistics are domain-invariant
- Handles uncertainties effectively
- Currently being tested…

Take away messages

- The upcoming volume and complexity of spectral data, especially around cosmic noon, call for the help of deep learning.
- The integration of simulation-based inference (SBI) with machine learning has emerged as a transformative approach across scientific disciplines, and it’s starting to impact galaxy spectroscopy.
- It is crucial to address the domain gap between simulations and real data; domain adaptation techniques can help reduce this gap.
- Methods such as conditional neural flows can make models uncertainty-aware, effectively mimicking Bayesian approaches.