Unsupervised exploration of the ~~4XMM~~ 5XMM catalogue with variational autoencoders

Simon Dupourqué (PostDoc @ IRAP)
Erwan Quintin (ESA Fellow)
In collaboration with N. Webb & M. Coriat (IRAP)

SF2A - 2026

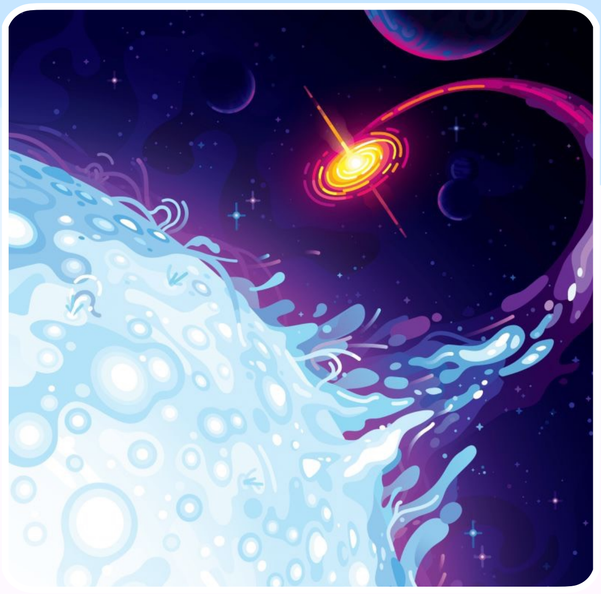
X-ray astronomy

Active galatic nuclei (AGNs)

Galaxy clusters

Neutron stars

X-ray binaries (XRBs)

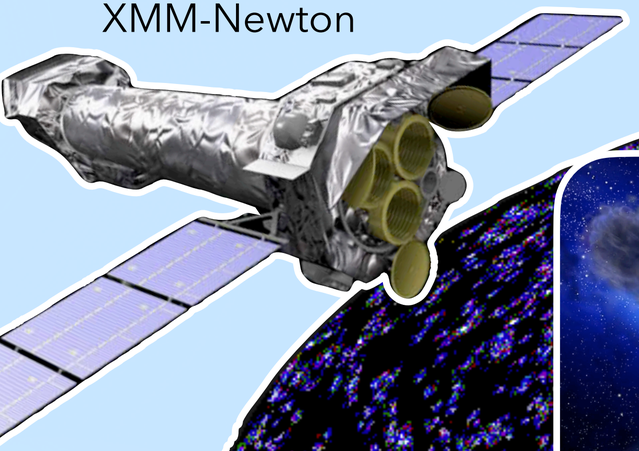


Study matter in extreme conditions

Adapted from Kurzgesagt

The XMM-Newton observatory & 5XMM catalogue

XMM-Newton



AGNs

Galaxy clusters

TDEs

Stars

X-ray binaries

CVs

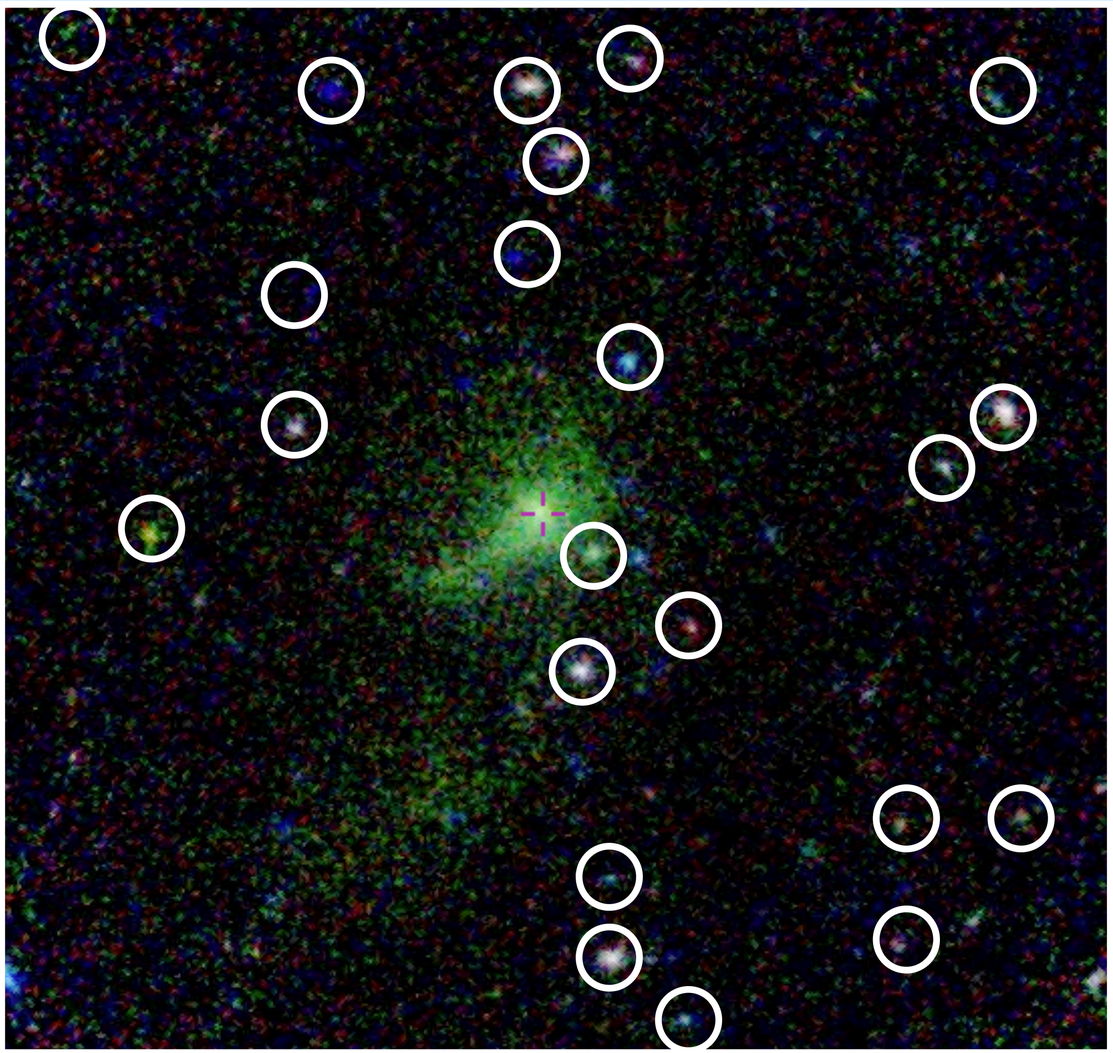
Common sources

Uncommon sources

Exotic sources

The XMM-Newton observatory & 5XMM catalogue

Automated source extraction pipeline



Each pointing brings ~ 100 serendipitous sources

3M sources reduced to **300k** (quality filters, min 200 photons, spectrum required)

See Webb & al. 2026 for more details on this catalogue

Scientific motivation

- Can we build a meaningful representation for the 5XMM sources spectra ?
- Can we do it with a physically motivated neural network architecture ?
- What could we learn from such data products ?

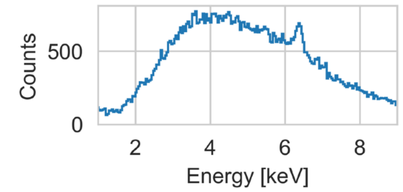
Architecture

Autoencoding the spectra

Information bottleneck

Observation space
High dimension

$\mathcal{S}$



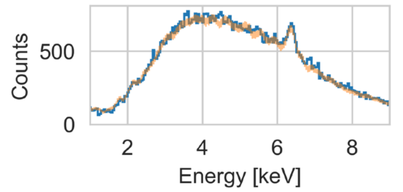
Encoder

Latent space
Low dimension

Decoder

Observation space
High dimension

$\hat{\mathcal{S}}$



Reconstruction loss between the input and output (Poisson)

$$\mathscr{L}\left(\mathcal{S}, \hat{\mathcal{S}}\right) = \hat{\mathcal{S}} - \mathcal{S} \log \hat{\mathcal{S}}$$

Going variational

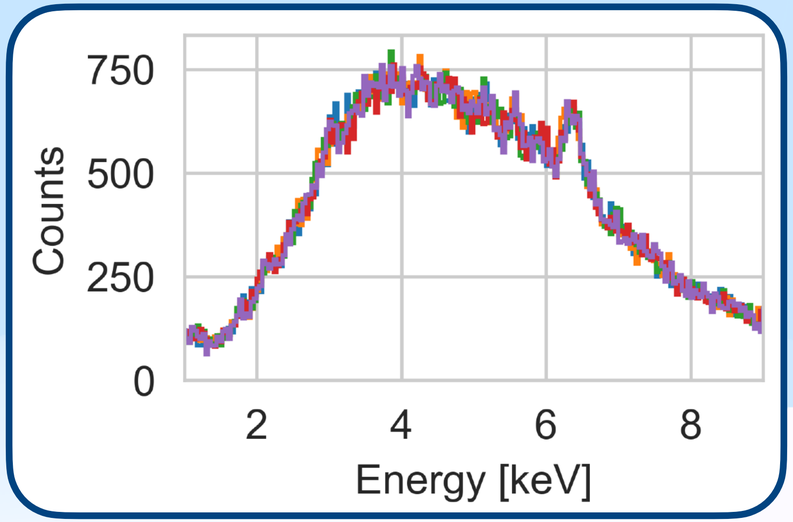
Latent space
Low dimension

μ

σ

$$z = \mu + \sigma\varepsilon$$
$$z \sim \mathcal{N}(\mu, \sigma)$$

Decoder



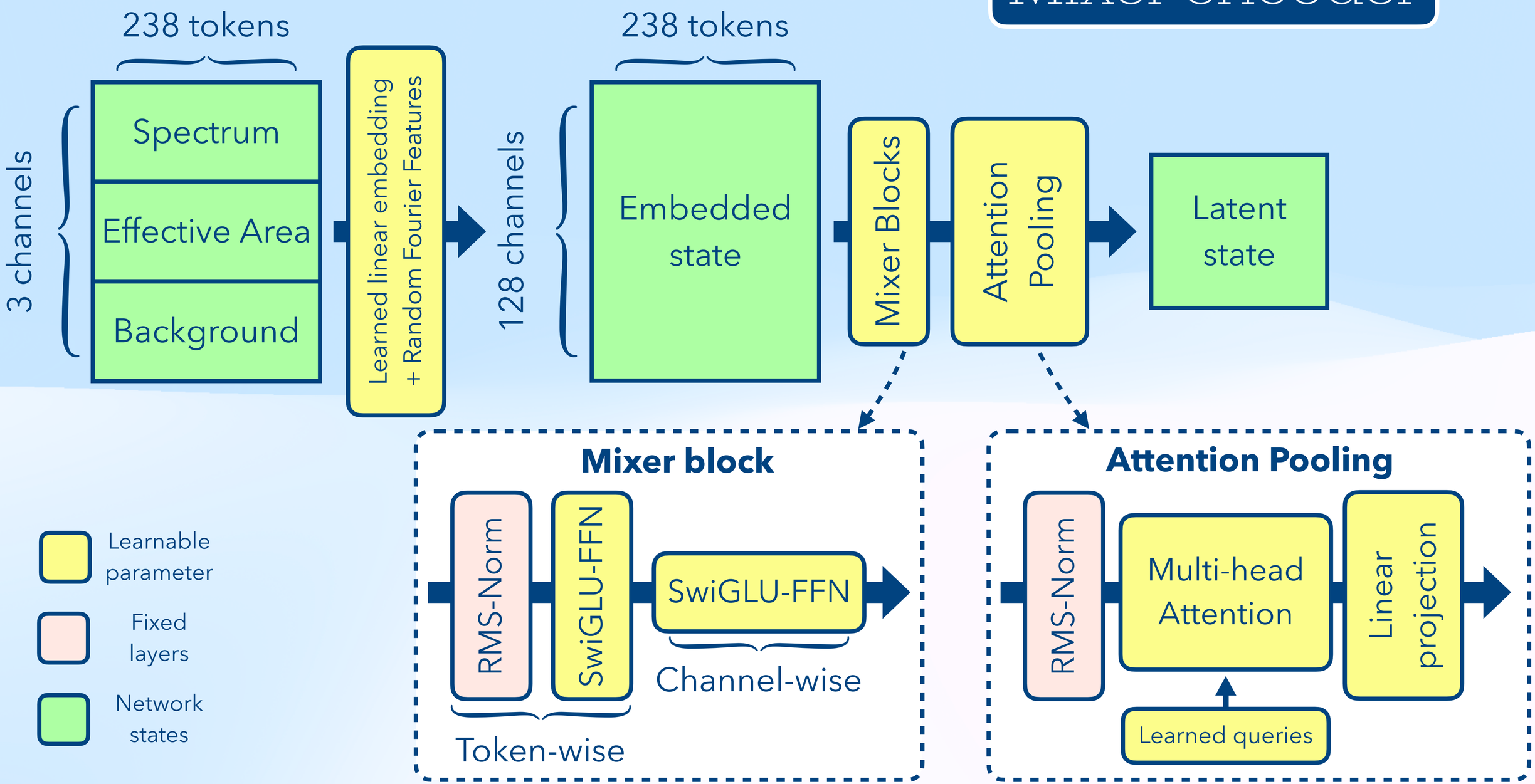
Independant Noise

ε

$$\varepsilon \sim \mathcal{N}(0,1)$$

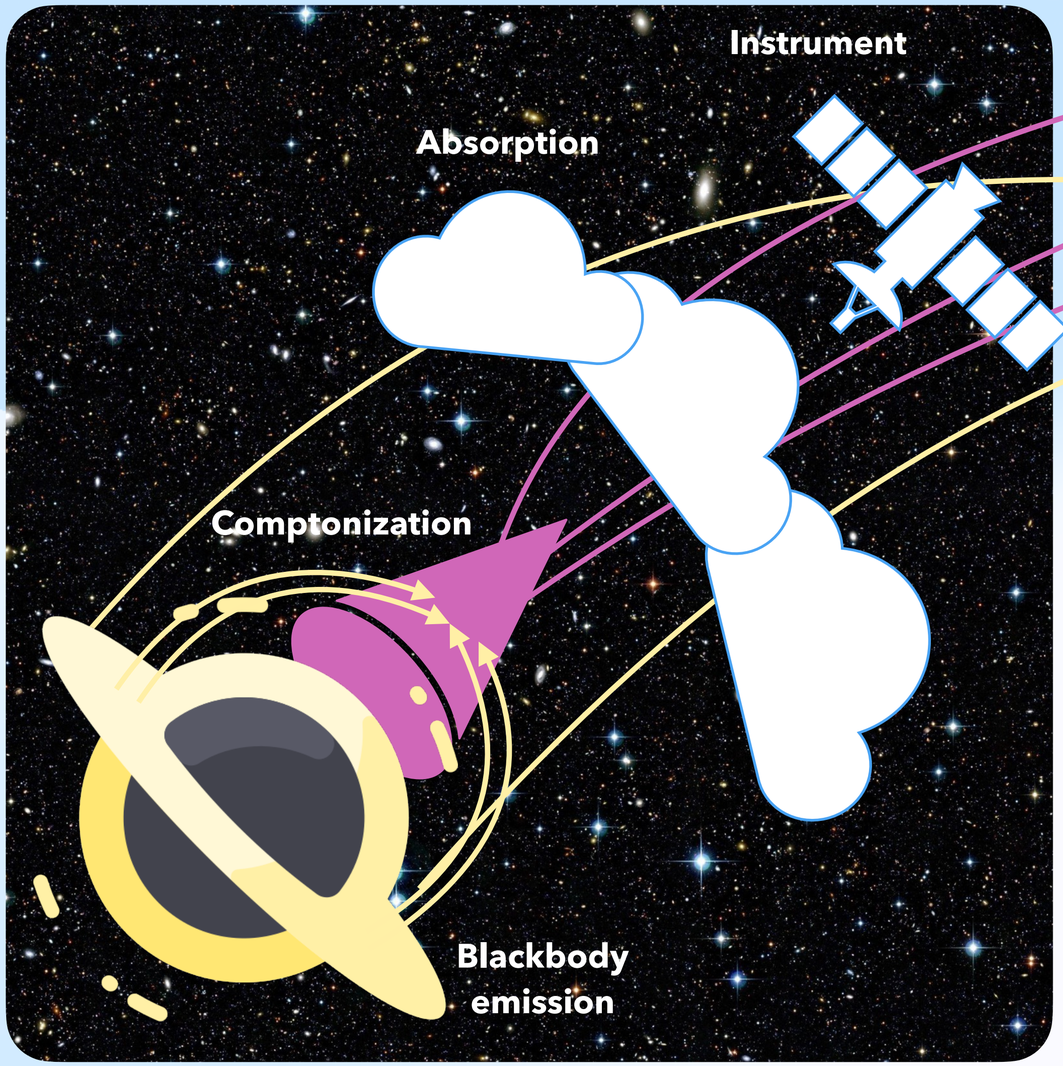
Sampling ε and feeding it to the decoder produce a **distribution** of spectra from a single spectrum

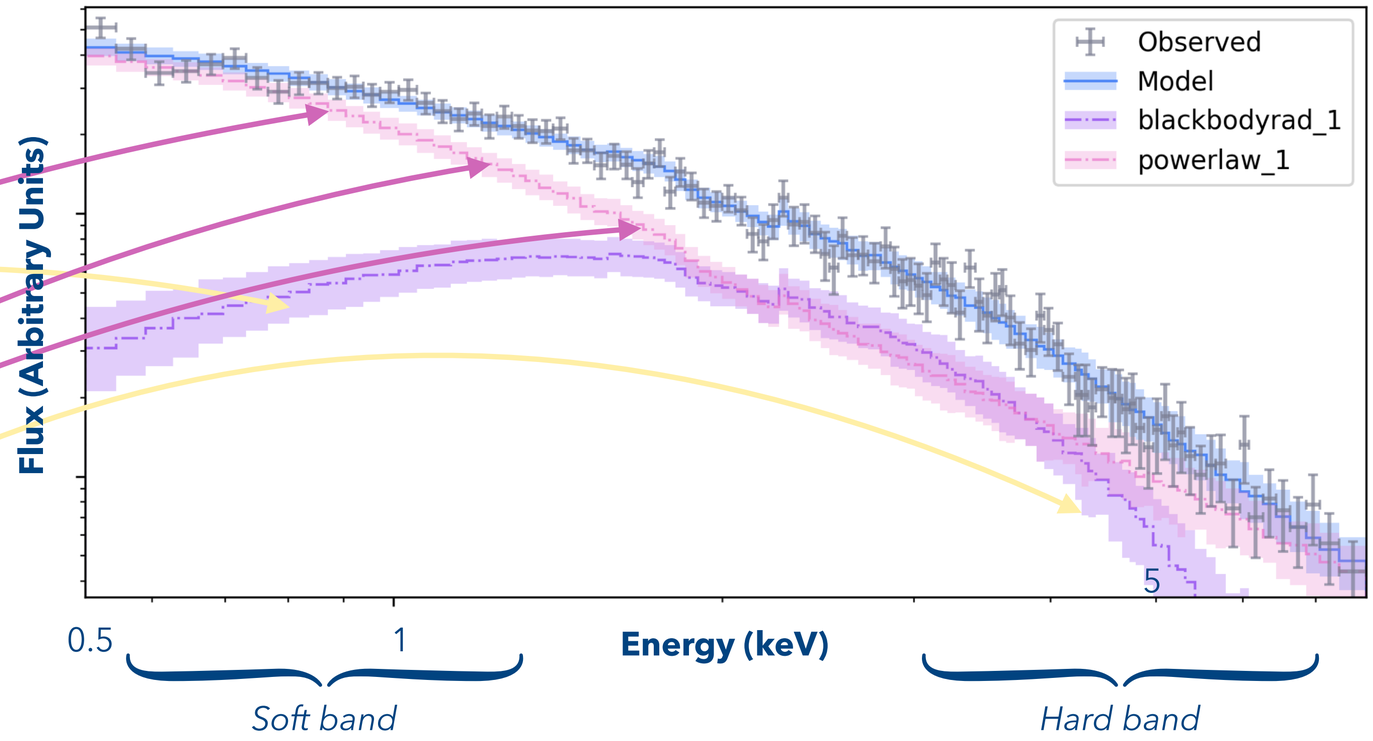
Mixer encoder



X-ray spectroscopy crash course

Example : accreting black hole





- X-ray spectra are a combination of smooth templates …
- … redshifted by the distance of the source
- … affected by ISM absorption (multiplicative template)
- … containing sharp features (emission lines)

DeepONet decoder

Latent state

Branch net

Linear projection

RMS-Norm

SwiGLU-FFN

Softmax

Weights

Split the model in two tasks :

- Learn relevant components
- Learn how to combine them

Strong inductive bias

Output spectrum

Energies

Trunk net

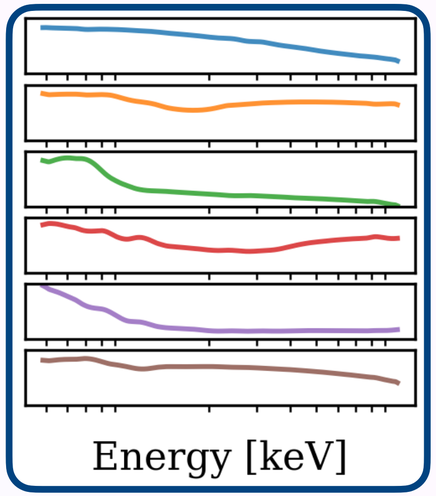
Fourier Features

RMS-Norm

SwiGLU-FFN

Softmax

Basis elements



Learnable parameter

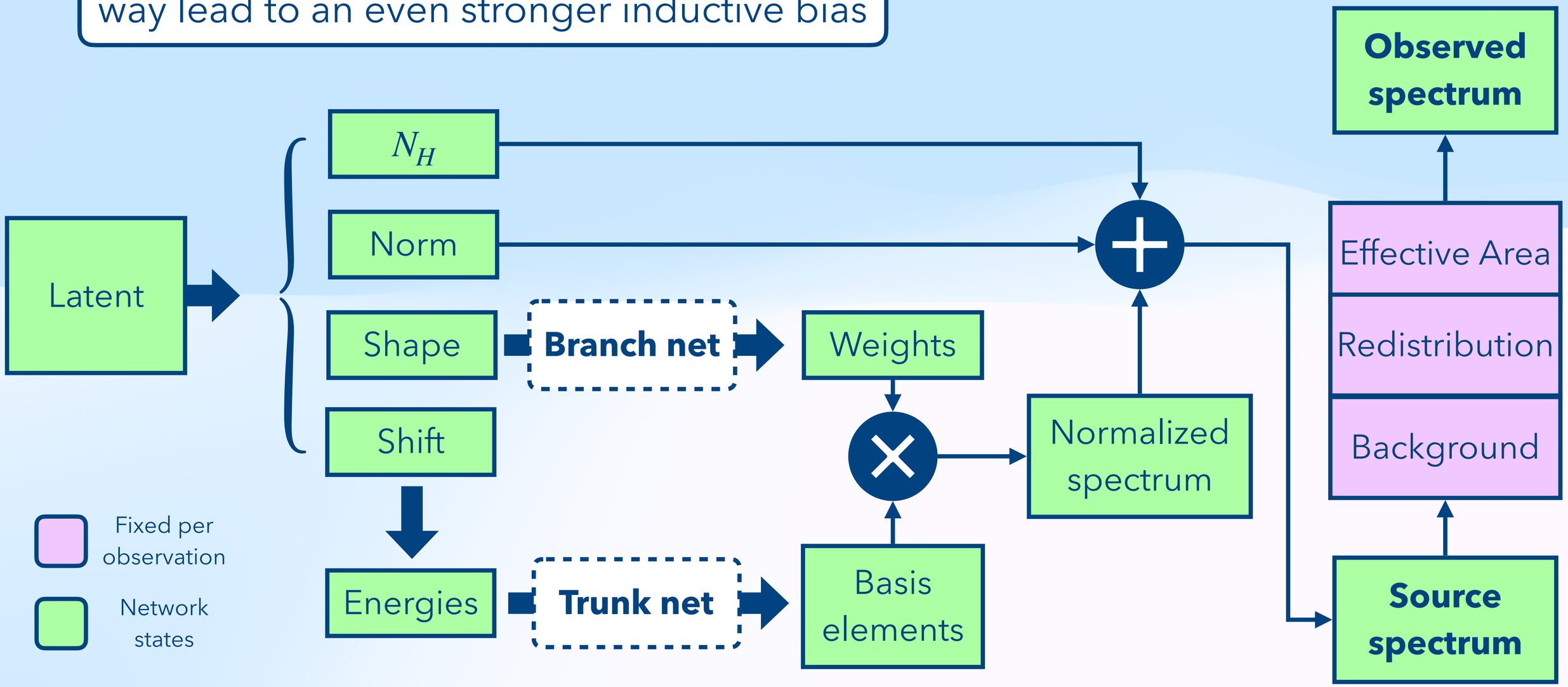
Fixed layers

Network states

Repeated blocks

Physically motivated decoding

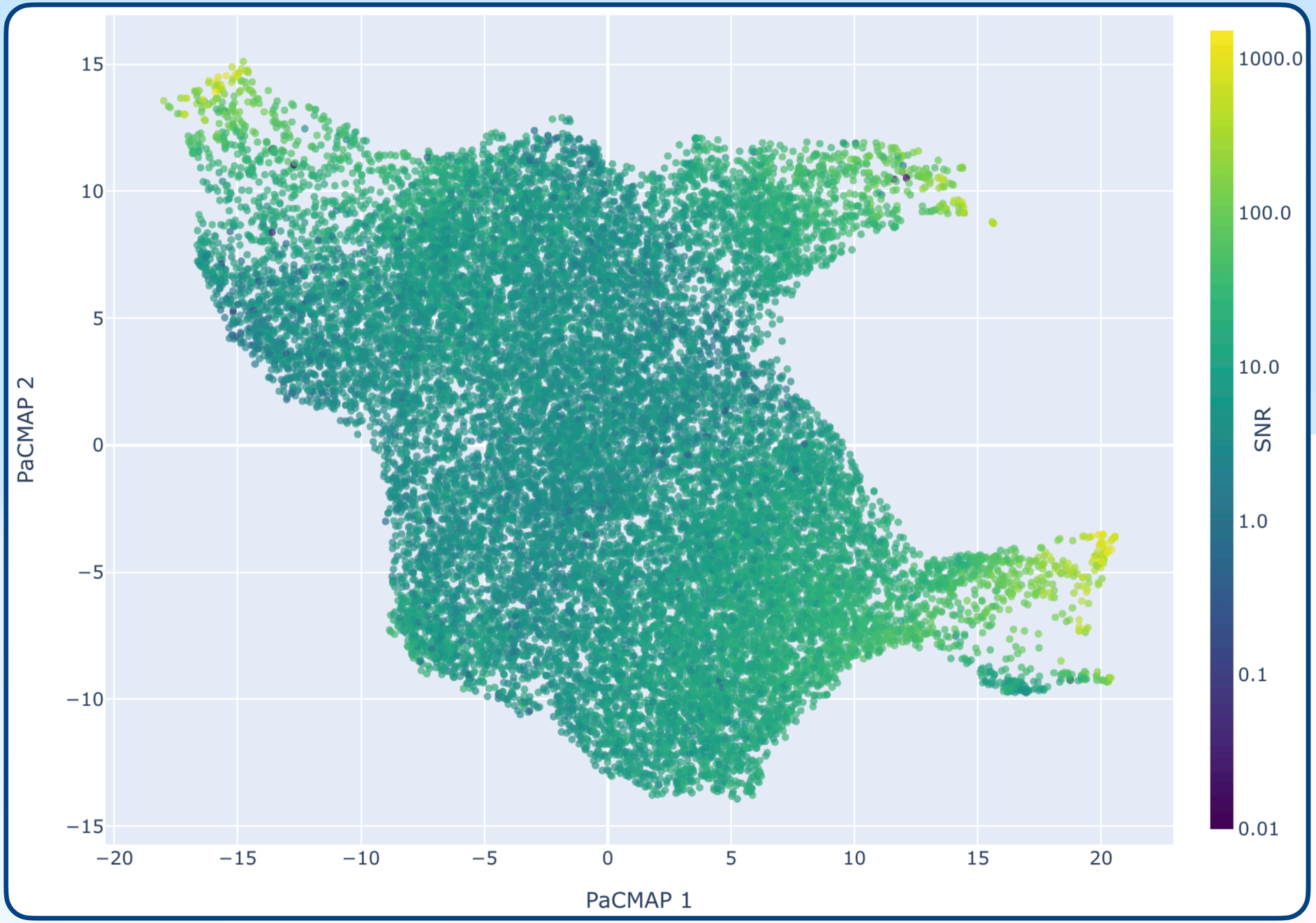
Reconstructing the spectrum in a physical way lead to an even stronger inductive bias



Exploring the latent space

Latent space (D=16)

- Projected to D=2 with PacMAP (Huang 2022)
- Subset of 20k sources
- SNR estimated with the neural network

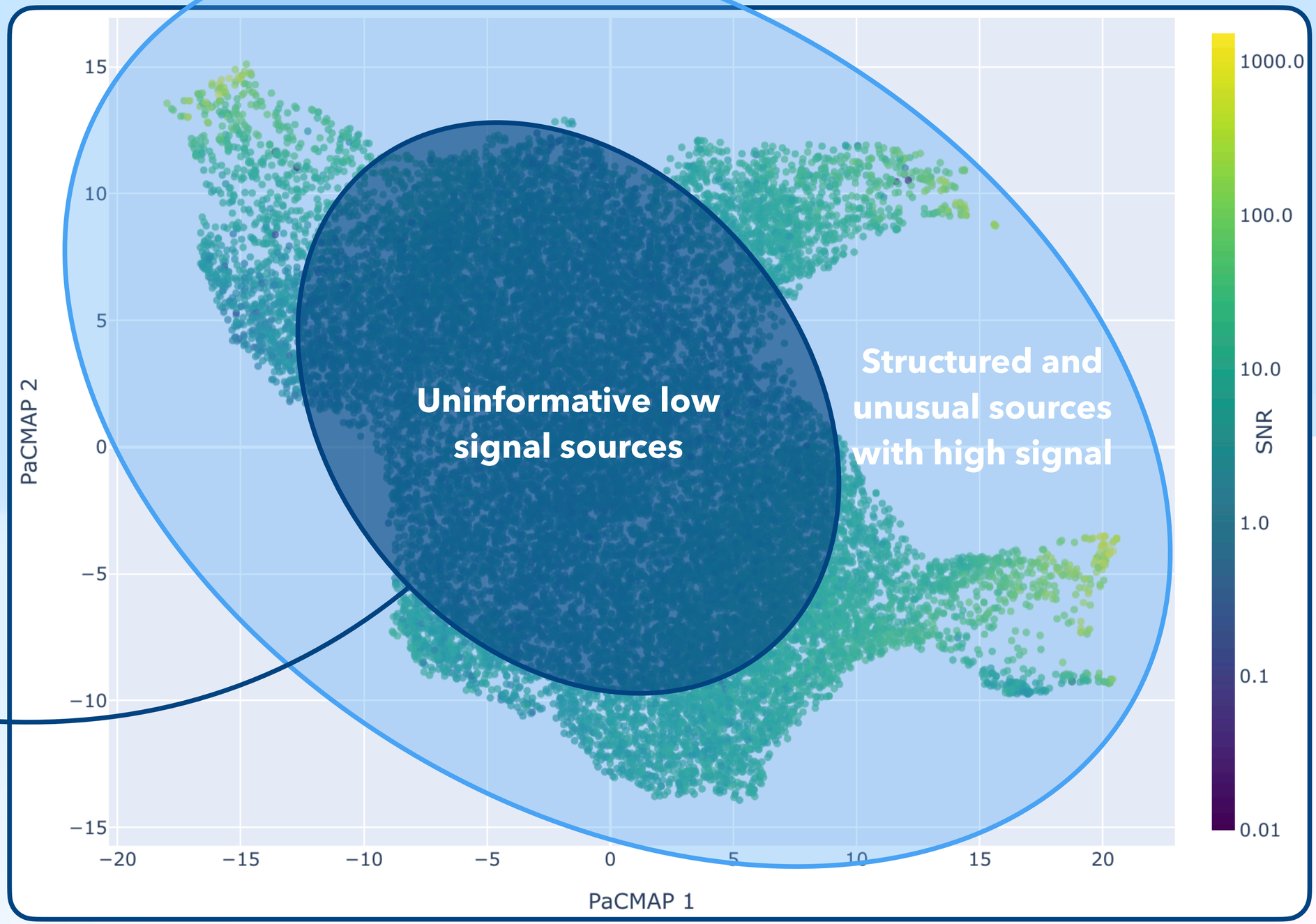


Latent space (D=16)

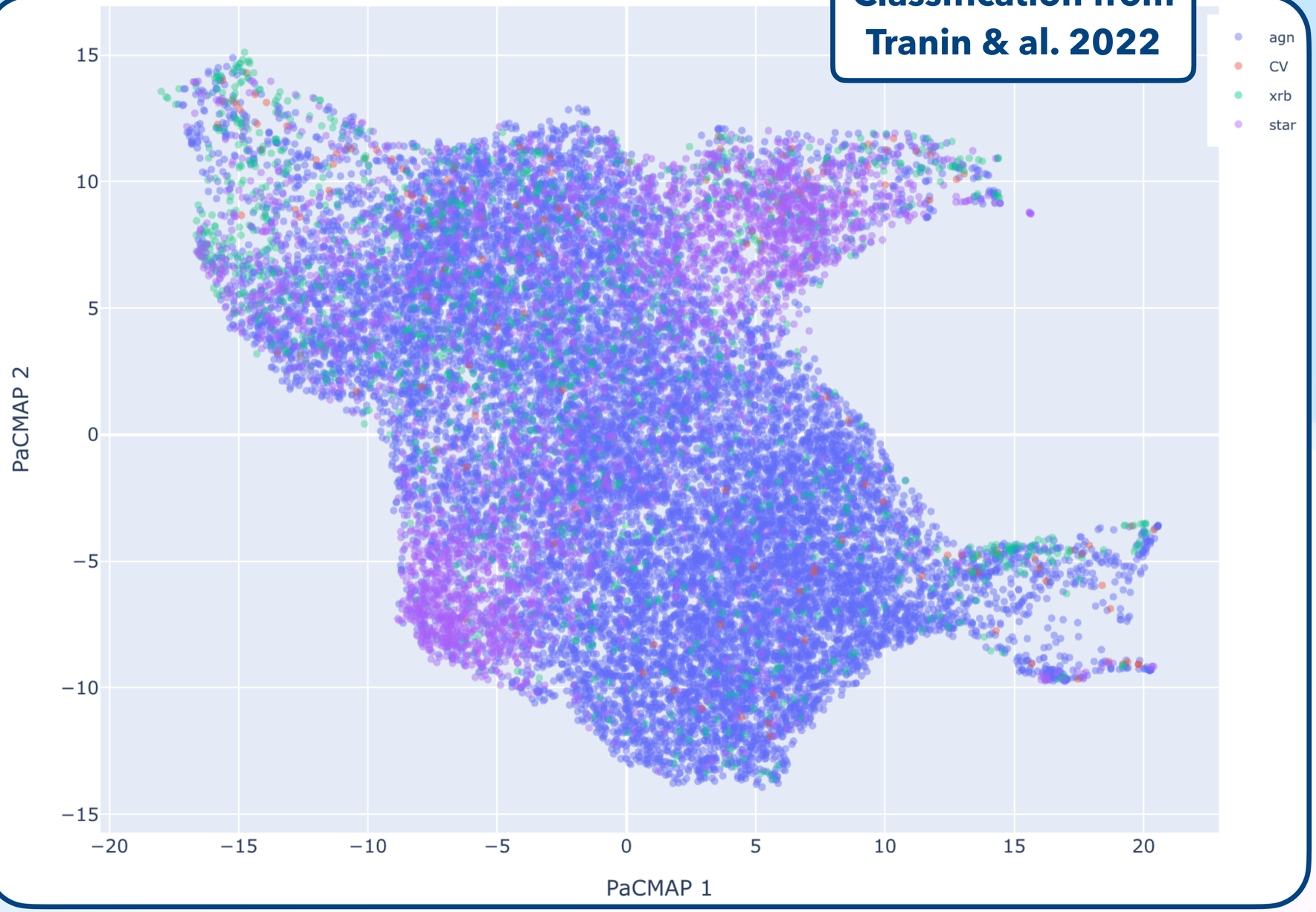
- Projected to D=2 with PacMAP (Huang 2022)
- Subset of 20k sources
- SNR estimated with the neural network

Half of the sources have a SNR < 2

Large bulk of unstructured sources



Classification from Tranin & al. 2022

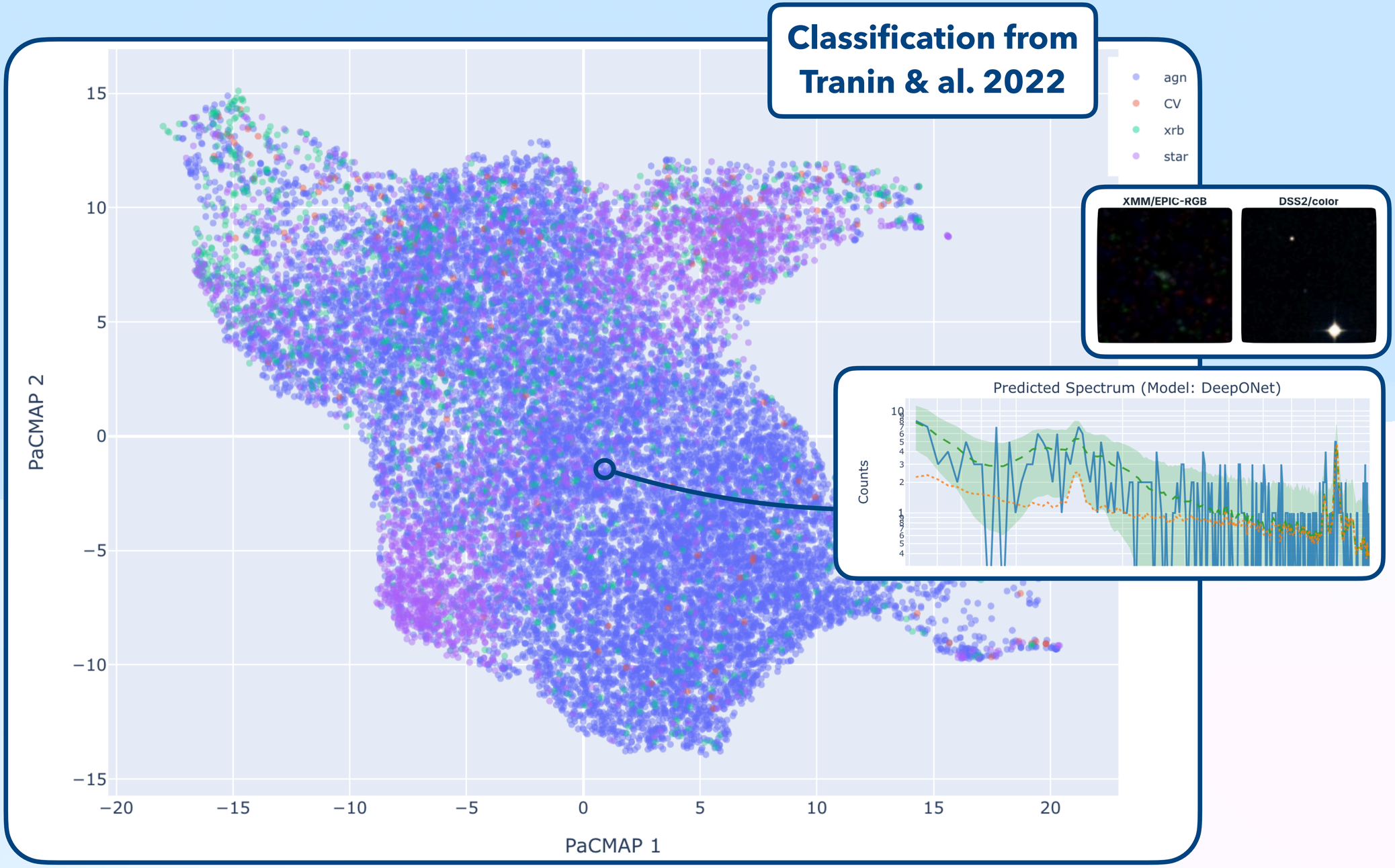


Raw spectrum

Reconstructed spectrum

Background

Classification from Tranin & al. 2022

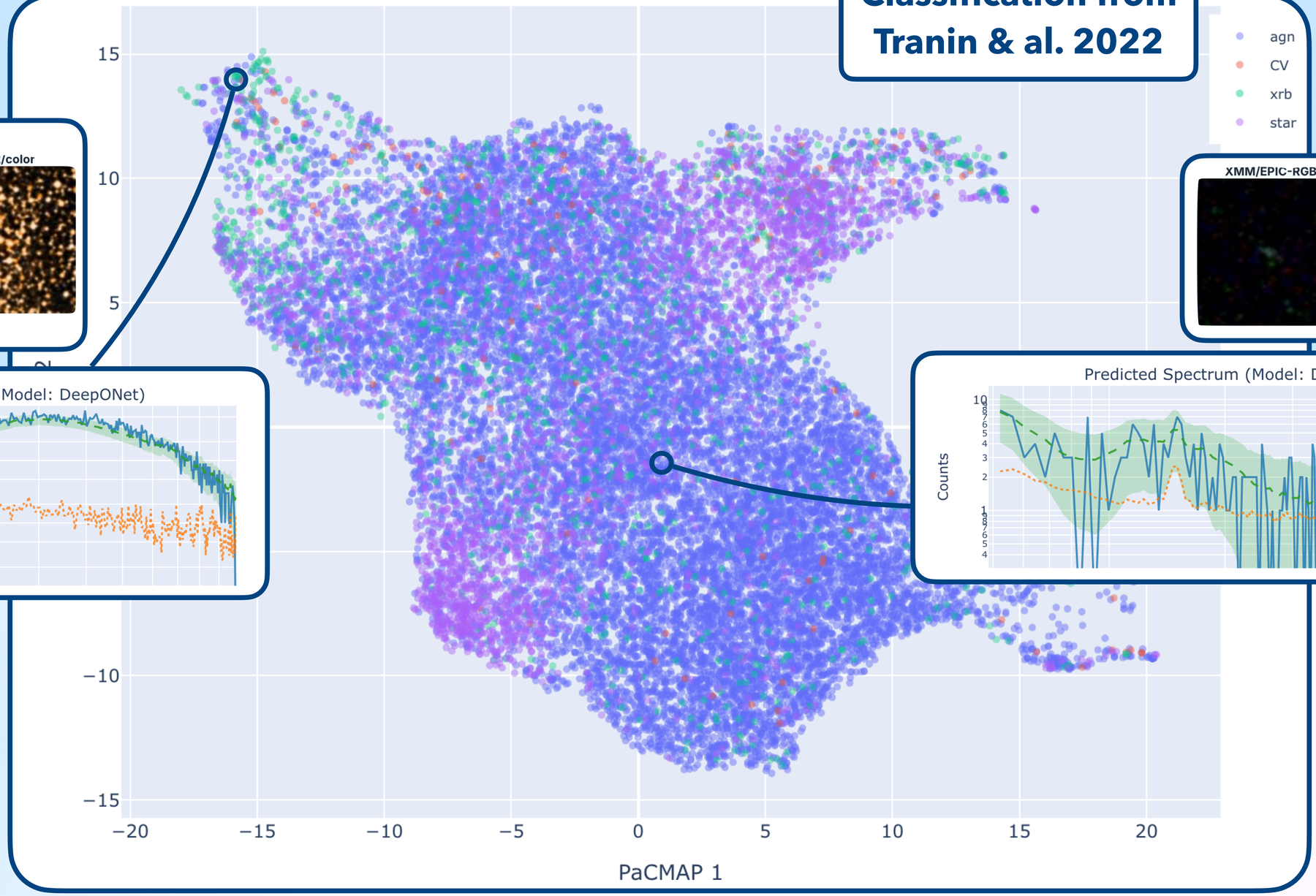


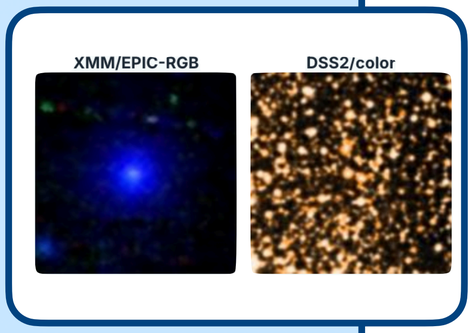
Raw spectrum

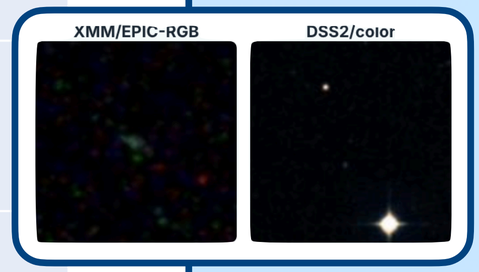
Reconstructed spectrum

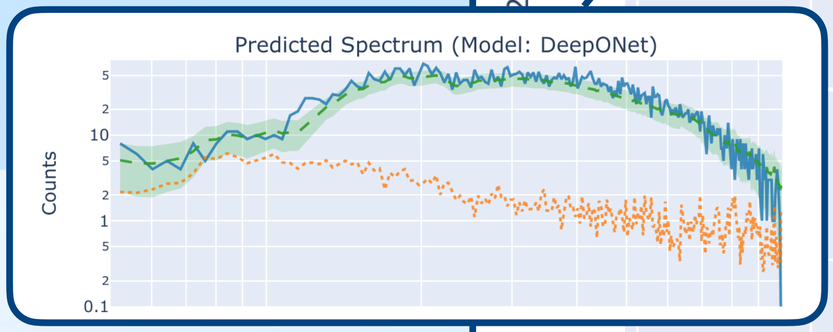
Background

Classification from Tranin & al. 2022







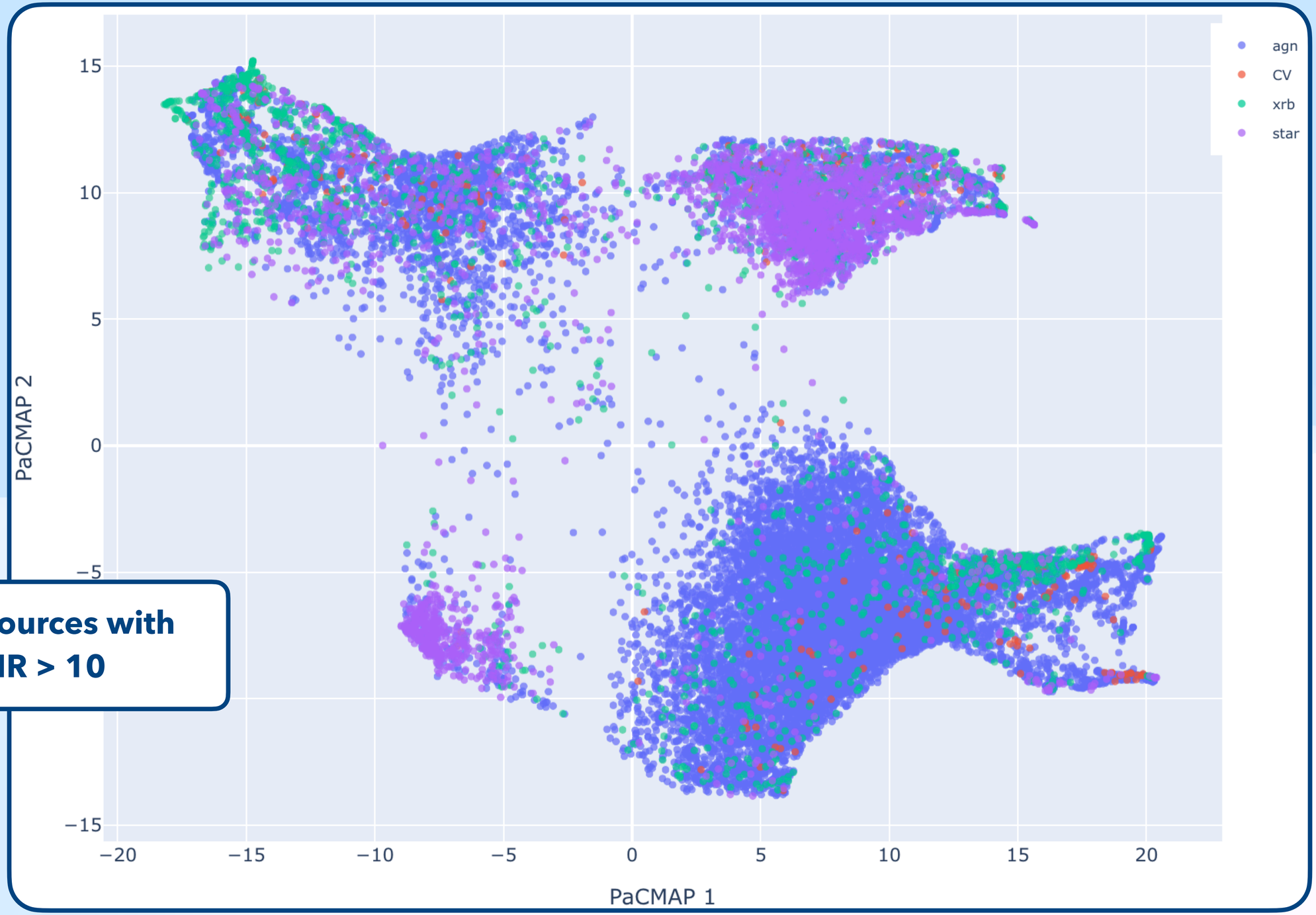


Raw spectrum

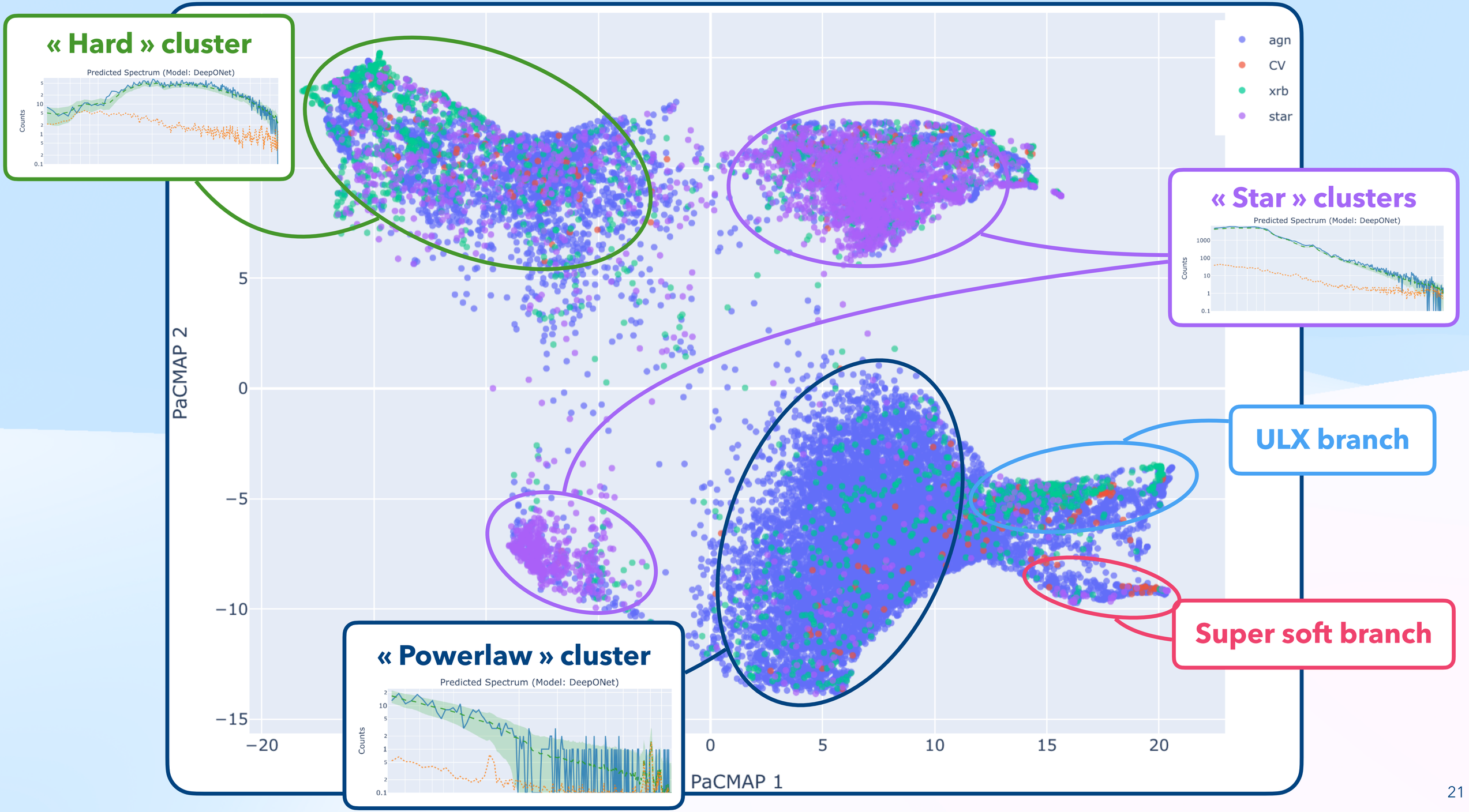
Reconstructed spectrum

Background


Classification from Tranin & al. 2022
agn
CV
xrb
star
XMM/EPIC-RGB
DSS2/color
Predicted Spectrum (Model: DeepONet)
Counts
Raw spectrum
Reconstructed spectrum
Background

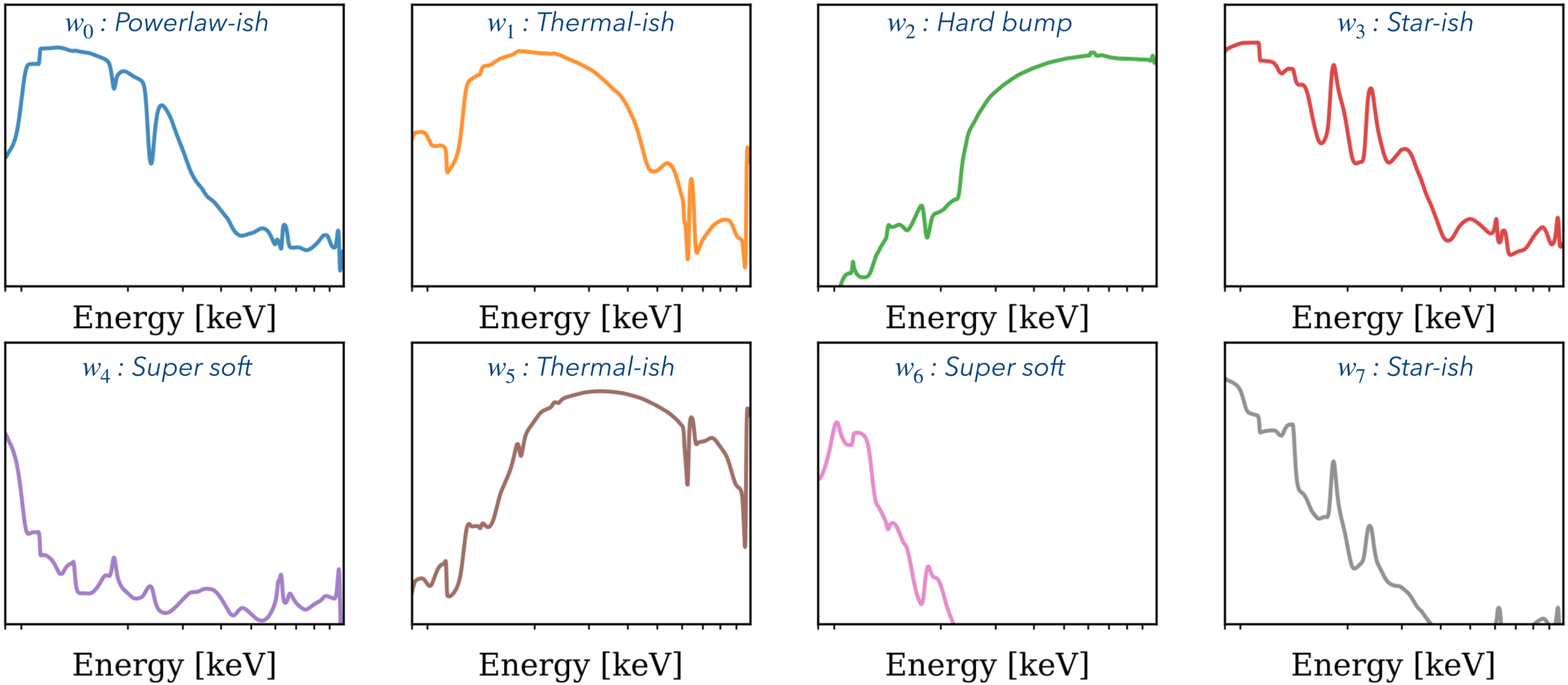


Filter sources with SNR > 10


« Hard » cluster
Predicted Spectrum (Model: DeepONet)
Counts
« Star » clusters
Predicted Spectrum (Model: DeepONet)
Counts
« Powerlaw » cluster
Predicted Spectrum (Model: DeepONet)
Counts
agn
CV
xrb
star
ULX branch
Super soft branch
PaCMAP 1
PaCMAP 2

DeepONet components [0.8 keV - 10 keV]

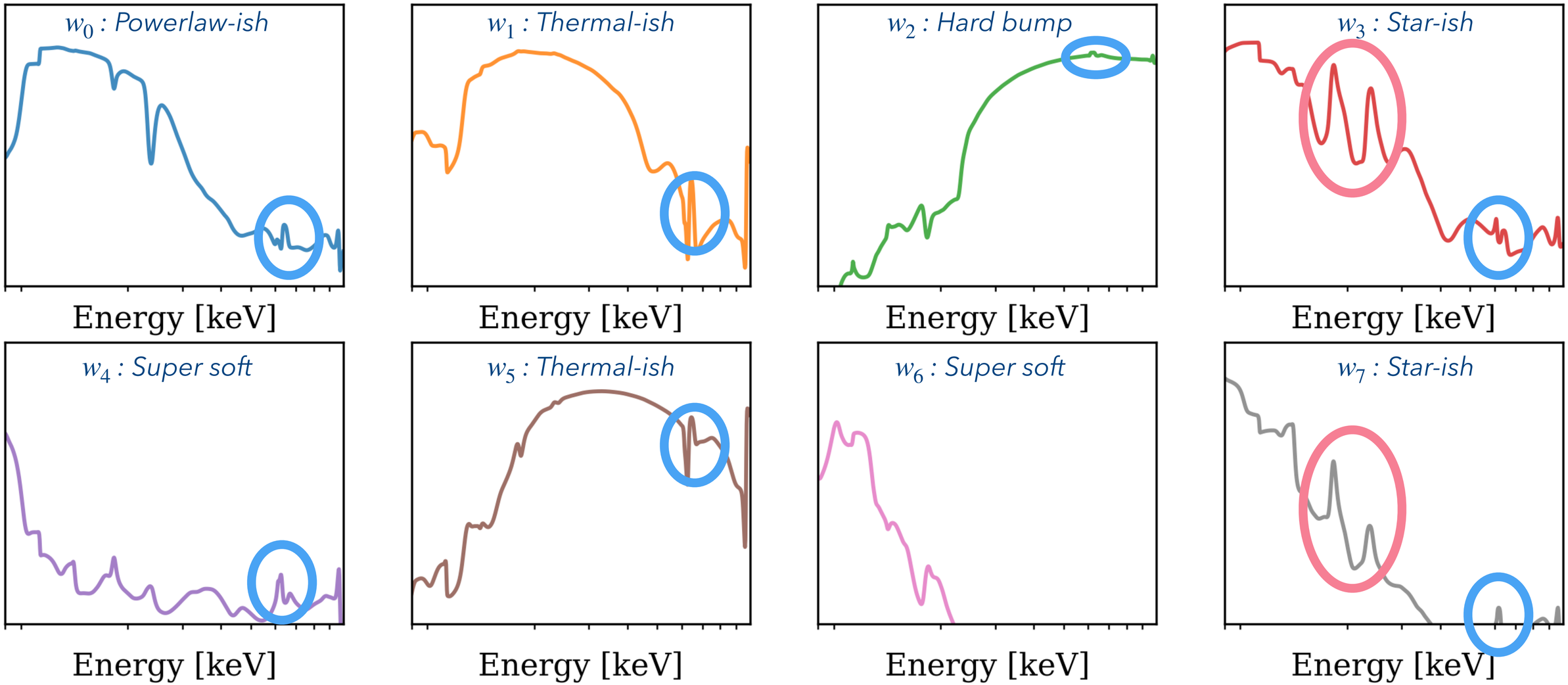
- Additive templates learnt from the data
- « Quite » physical but too expressive

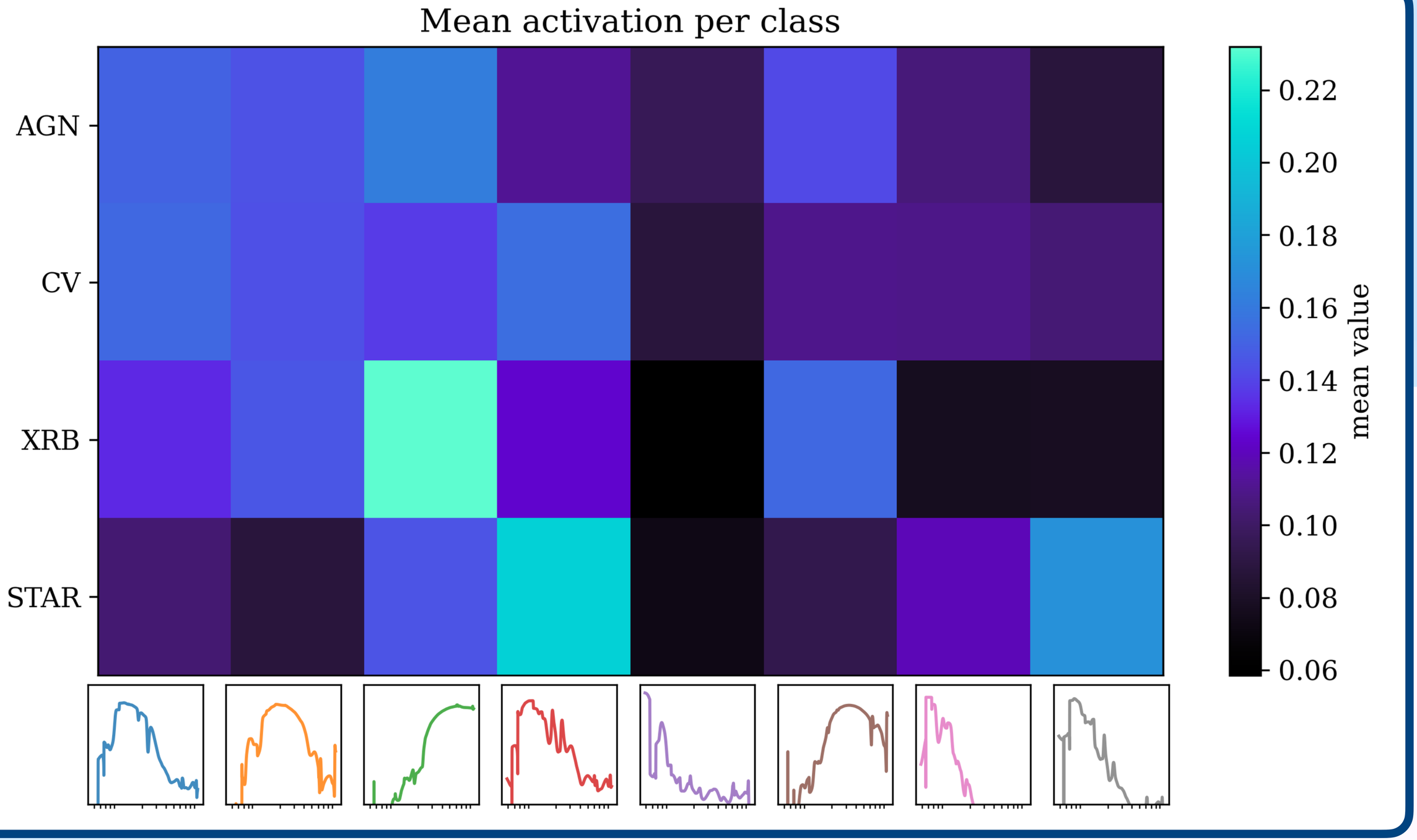


DeepONet components [0.8 keV - 10 keV]

- Additive templates learnt from the data
- « Quite » physical but too expressive

- **6.4 keV feature** : Iron K-alpha line
- **1.8 keV / 2.4 keV features** : Si / S lines




Mean activation per class
AGN
CV
XRB
STAR
0.22
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
mean value

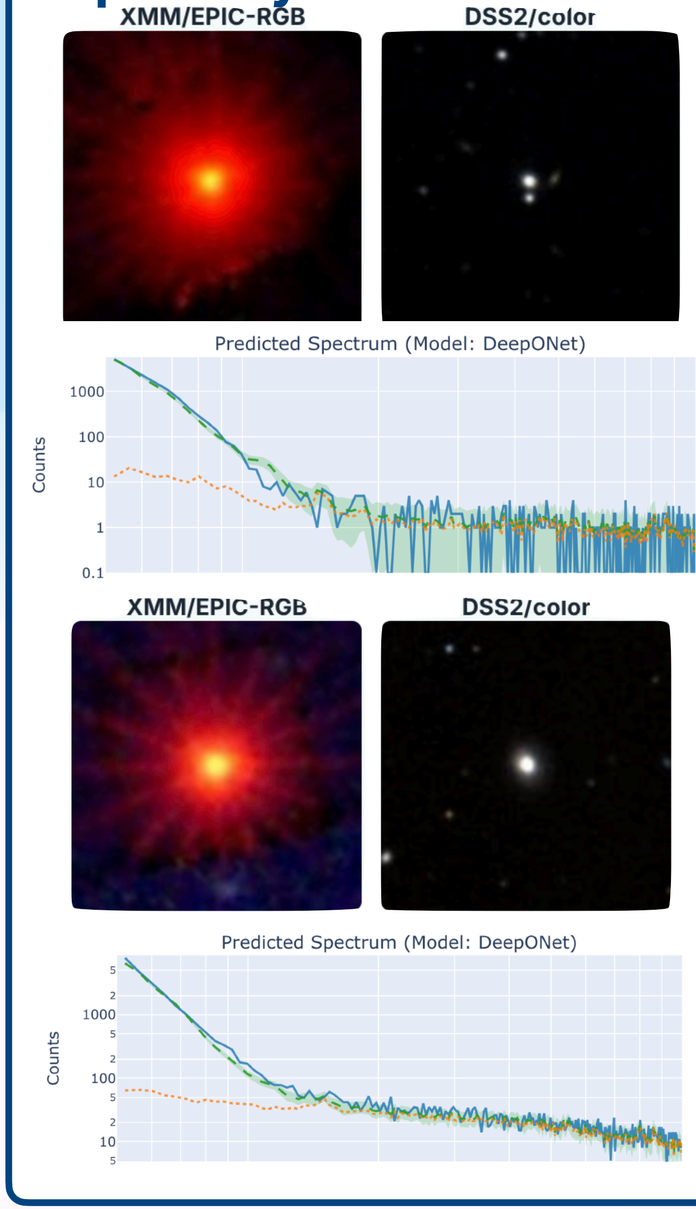
Similarity search

Known source

Statistical distance with all latent encoded spectra

- Wasserstein distance
- Kullback-Leibler divergence
- Bhattacharyya distance
- *Your favorite one*

Spectrally similar sources

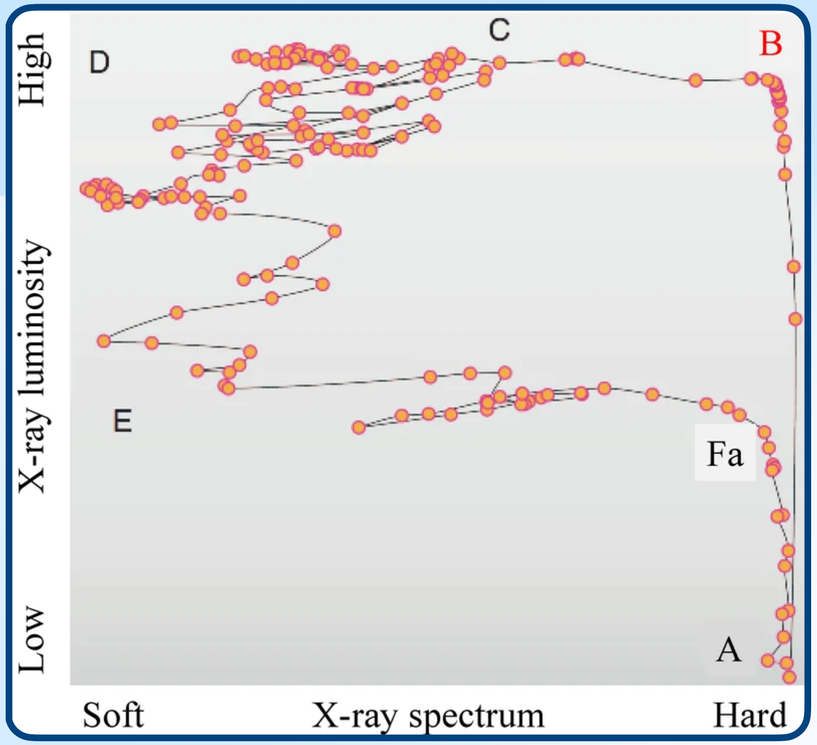


Samples produced this way have a **higher purity** than simpler spectral criterions such as Harness Ratios

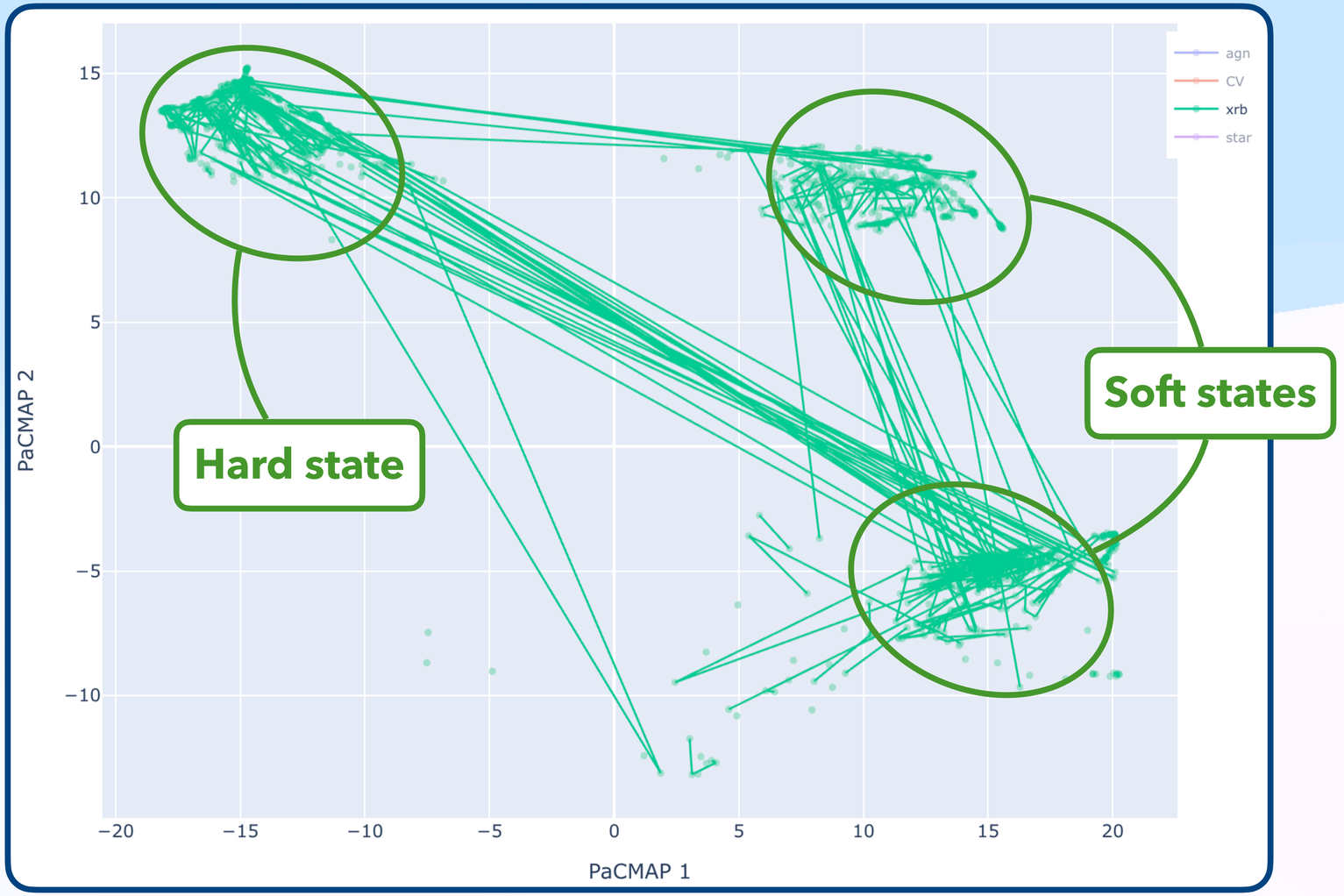
Latent trajectories

Source moving in the latent space between observations → **Spectral variability**

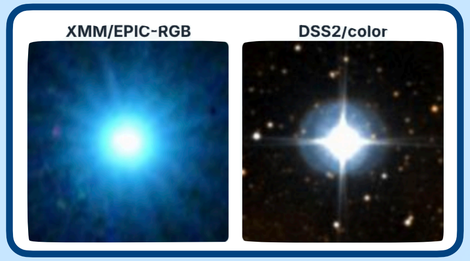
Proxy of the XRB Q-diagram at catalogue level



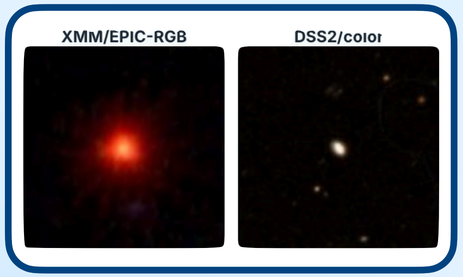
Adapted from Contopoulos + 2019

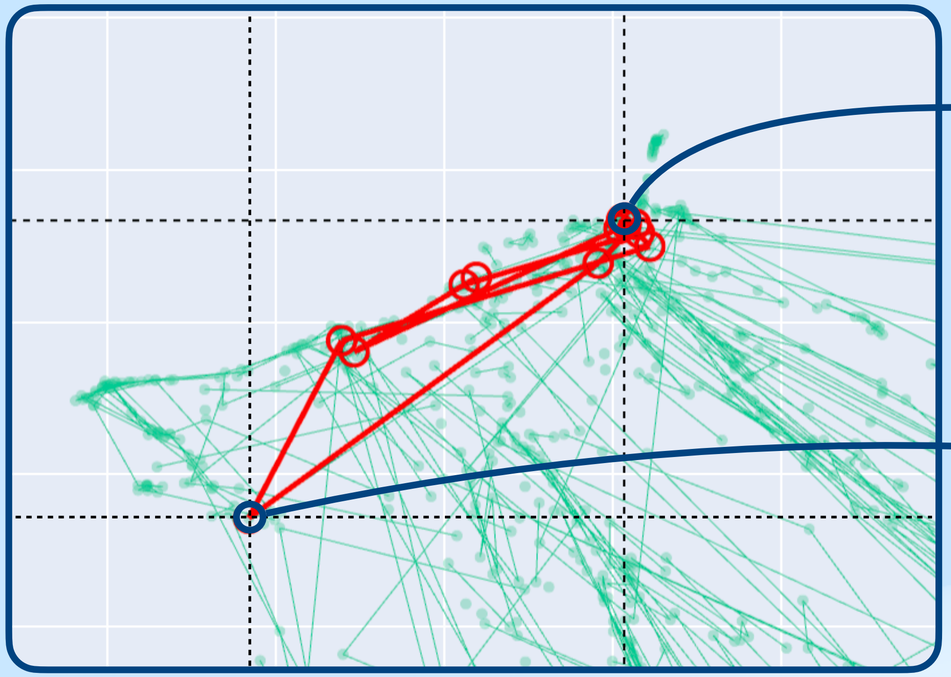


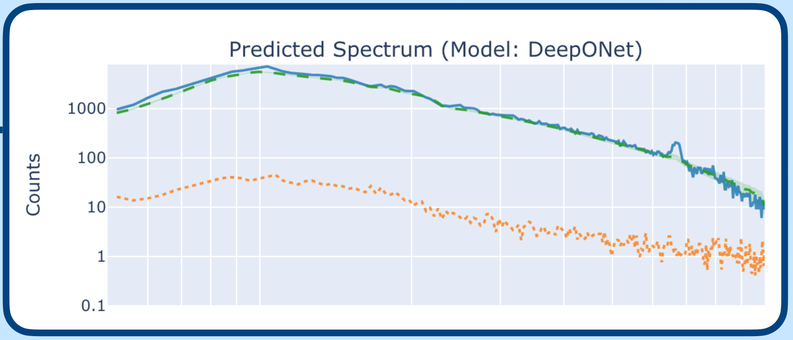
Orbital variability of WR140 (Colliding wind binary) during its orbital motion

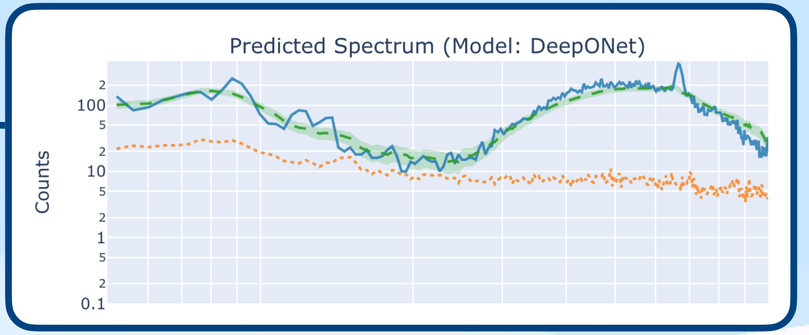


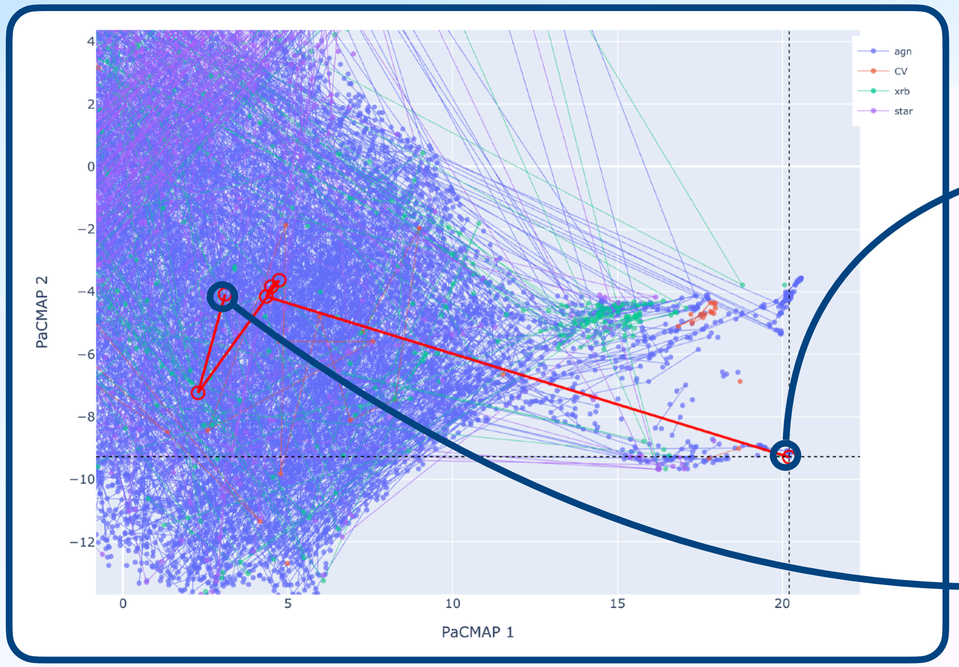
A TDE shutting down

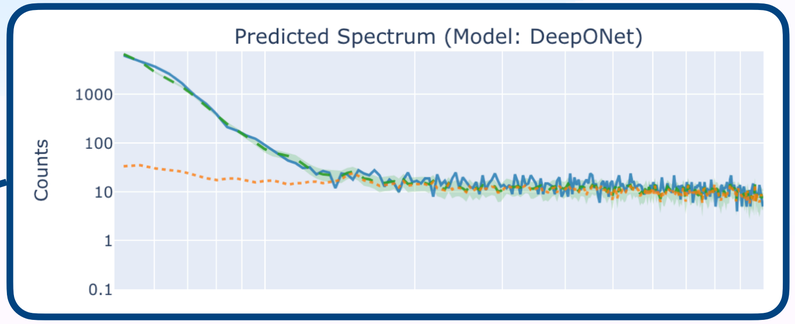


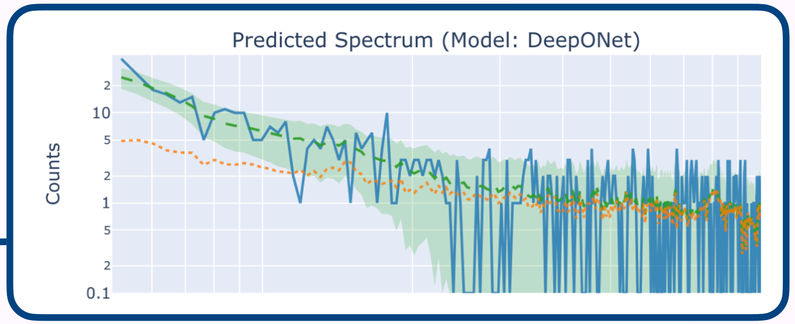












Conclusion

- We trained a variational auto-encoder to reconstruct the spectra gathered by XMM-Newton over 30 years of service
- Including the instrument response, the observational setup, and enforcing a physically motivated scheme using a DeepONet architecture yield meaningful reconstruction
- The network was able to learn basic components of X-ray spectroscopy, provided a meaningful latent space and clustering
- This representation opens the door to new ways to dig in the catalogue e.g. observing the latent trajectory or getting the similar sources