What can we learn from the XMM source catalogue?

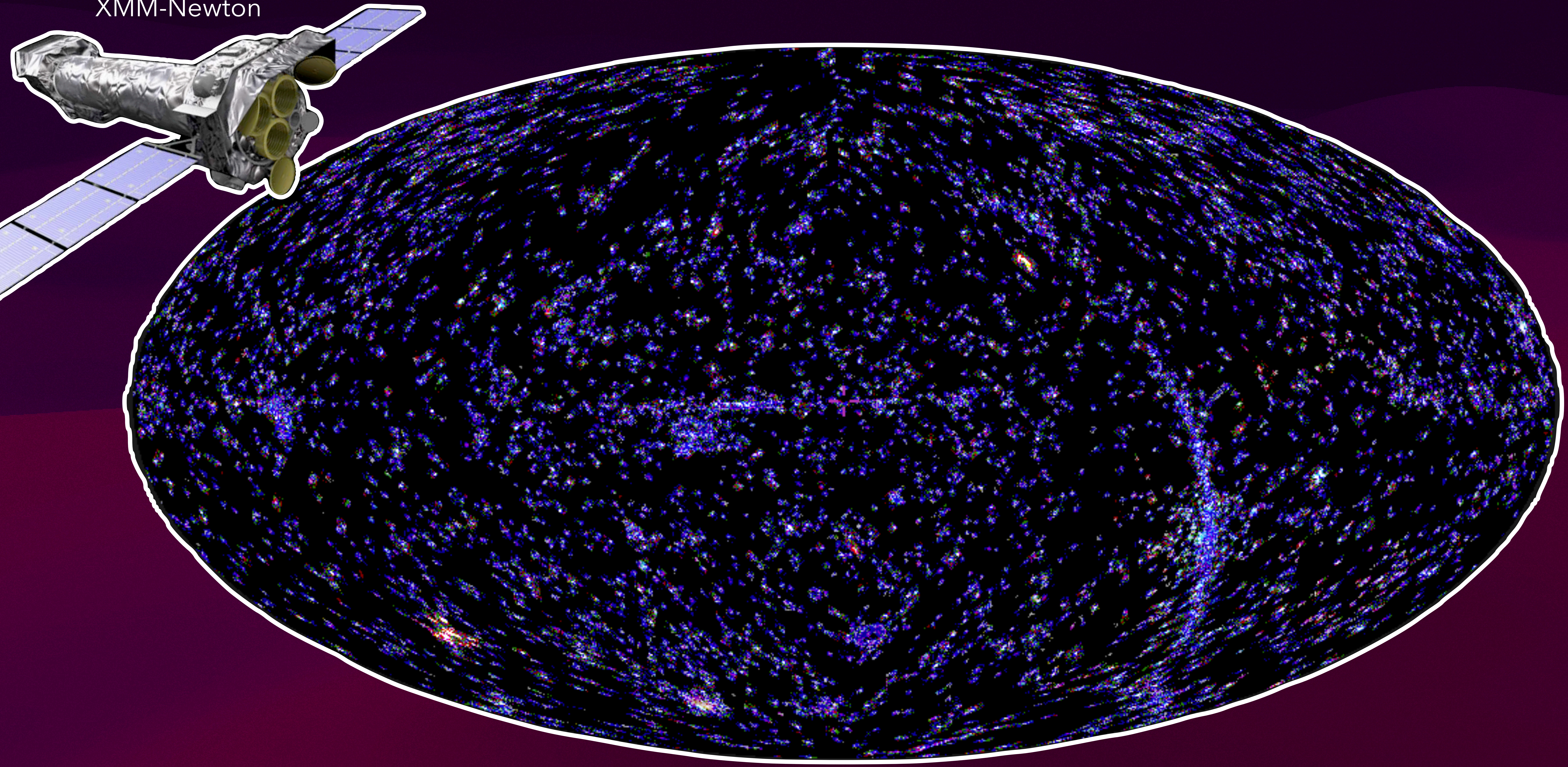
Simon Dupourqué & Erwan Quintin

AI @ IRAP (2025)

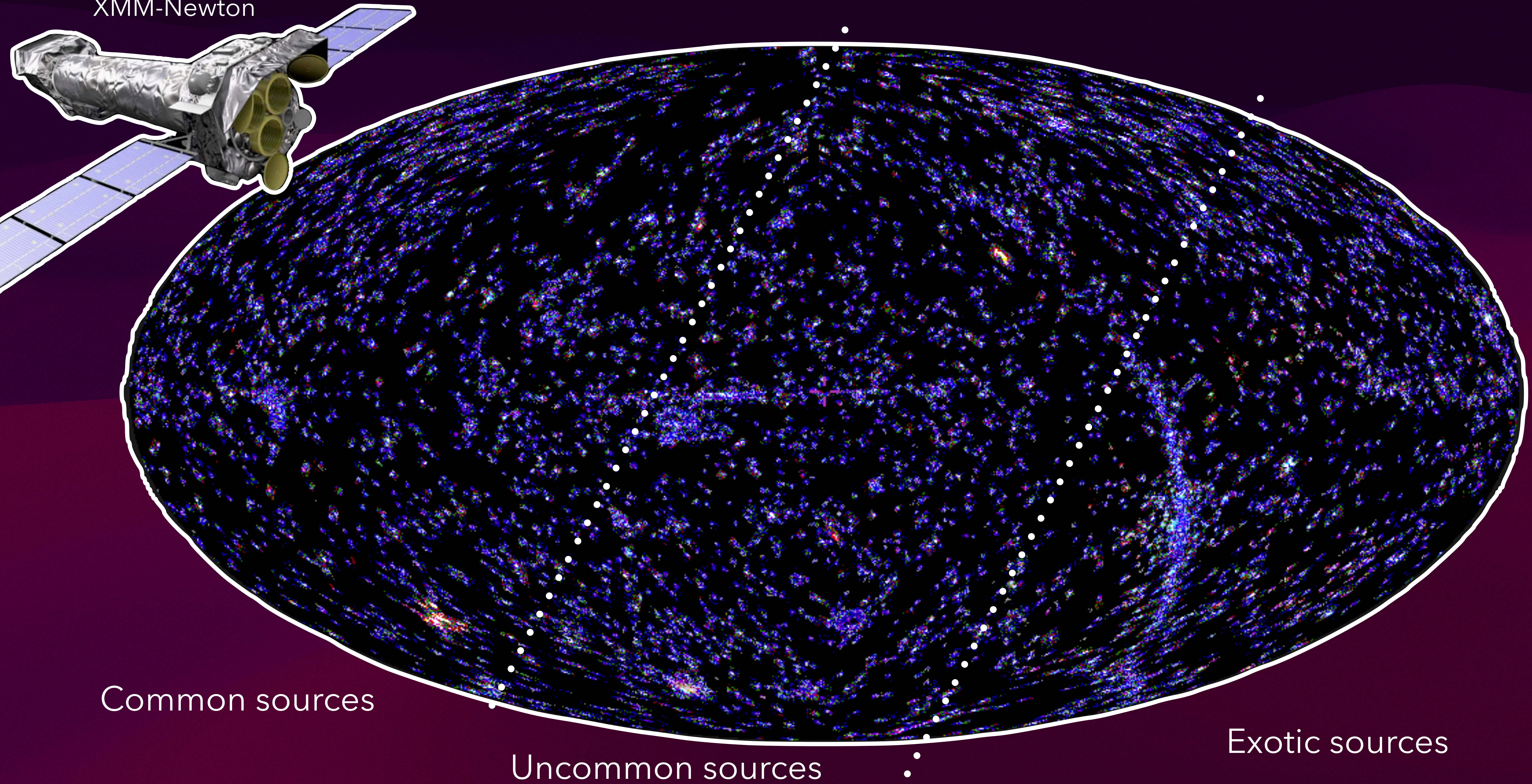
astrophysique & planétologie

XMM & the 4XMM catalogue

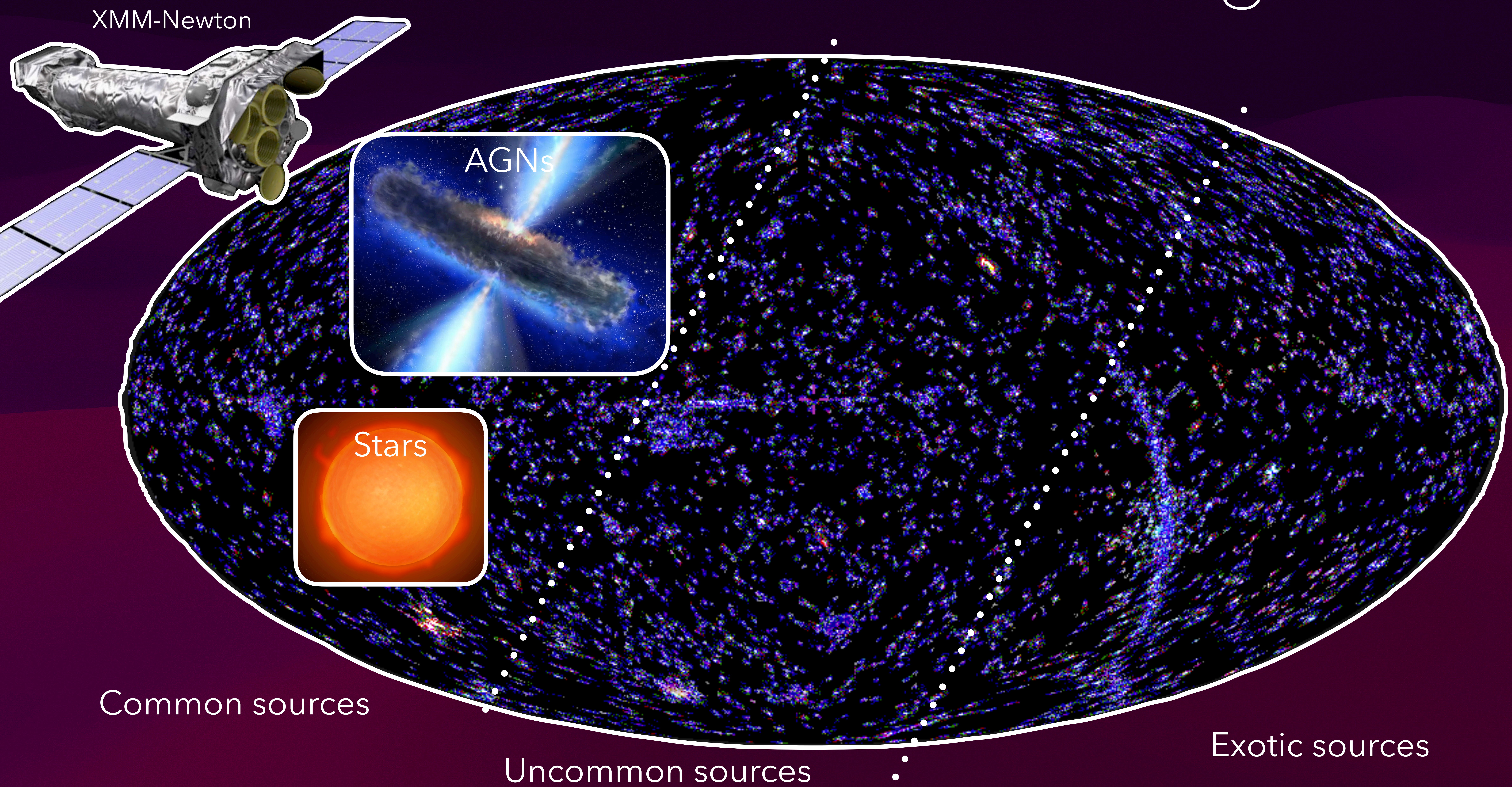
XMM-Newton



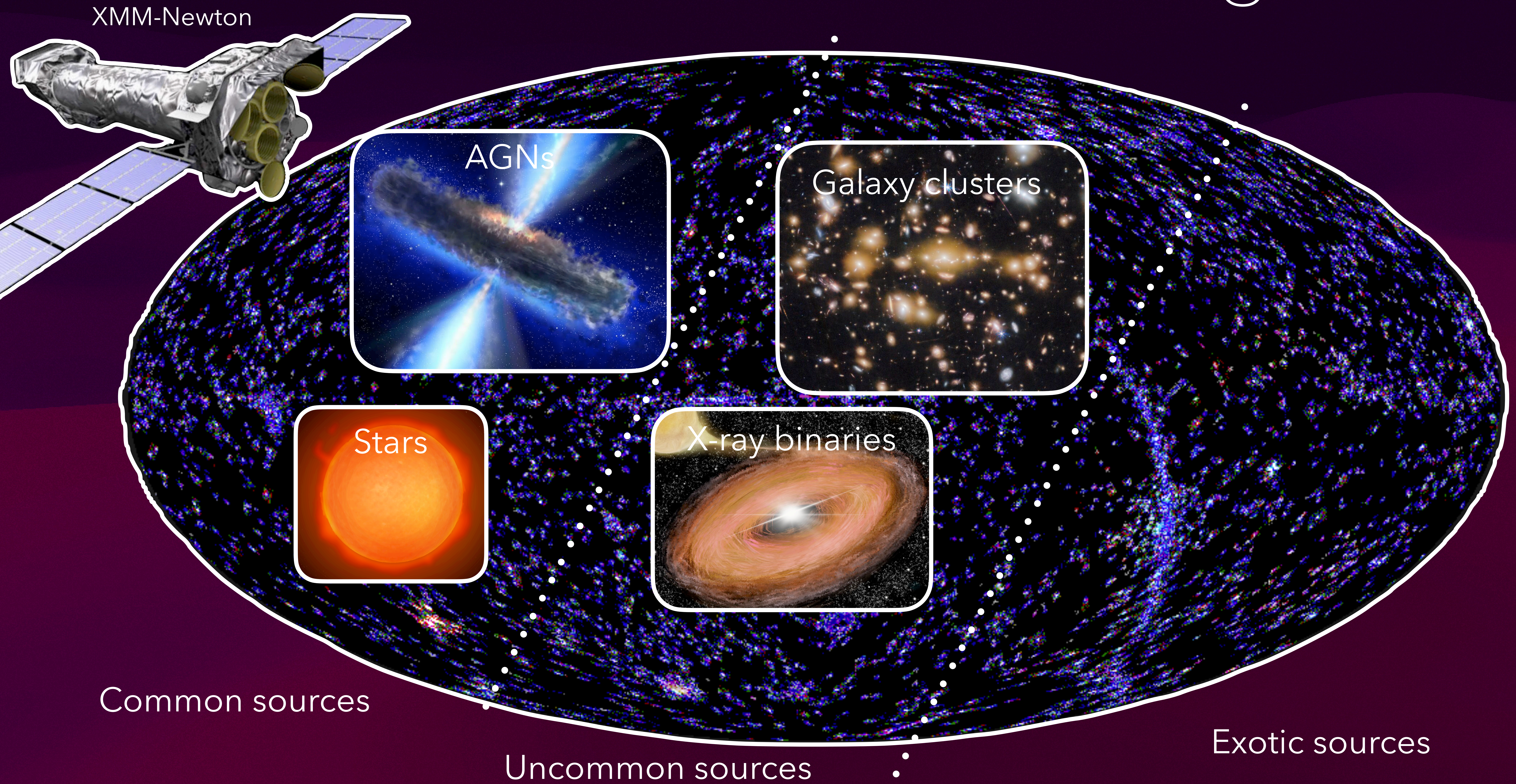
XMM & the 4XMM catalogue



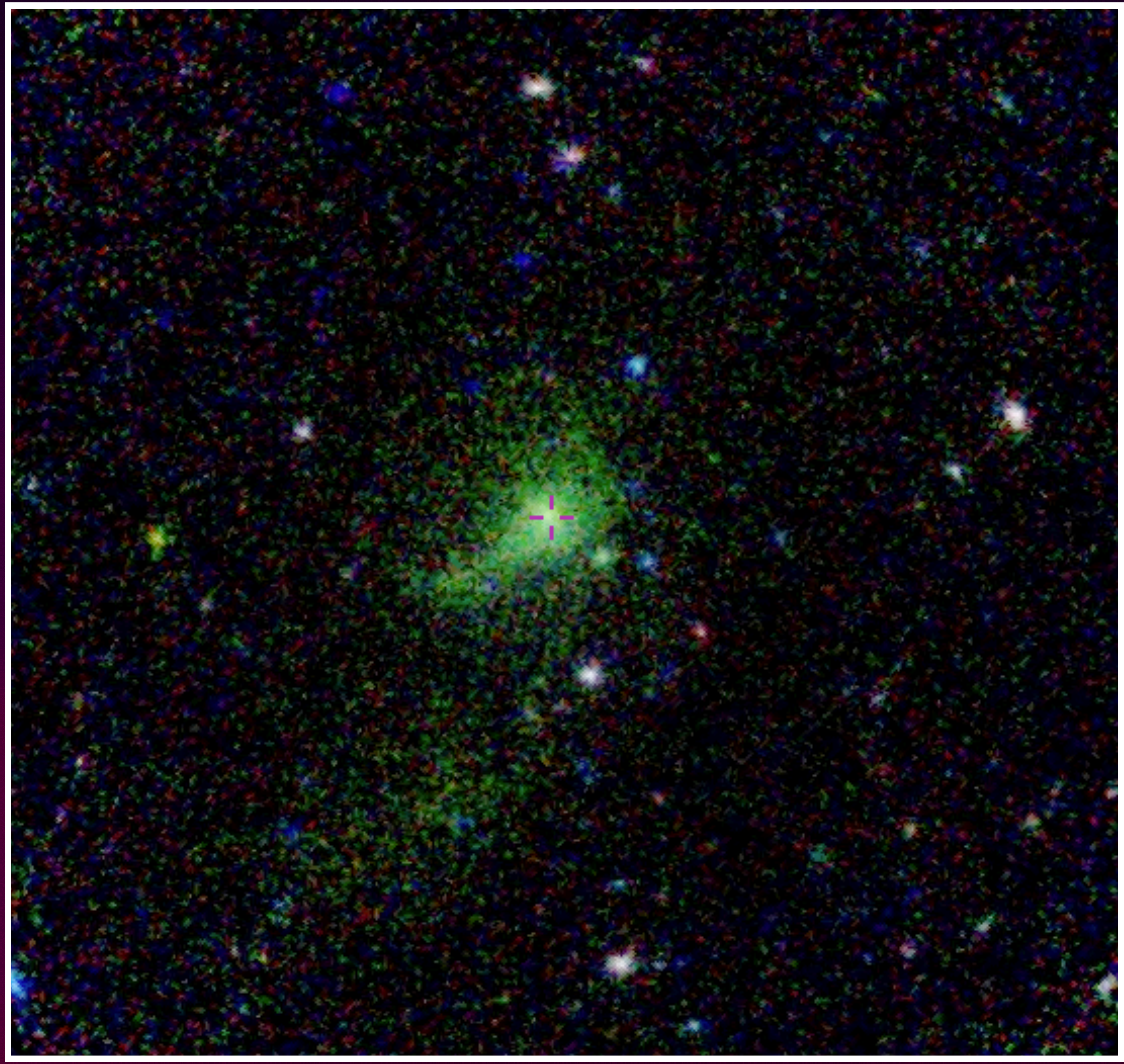
XMM & the 4XMM catalogue

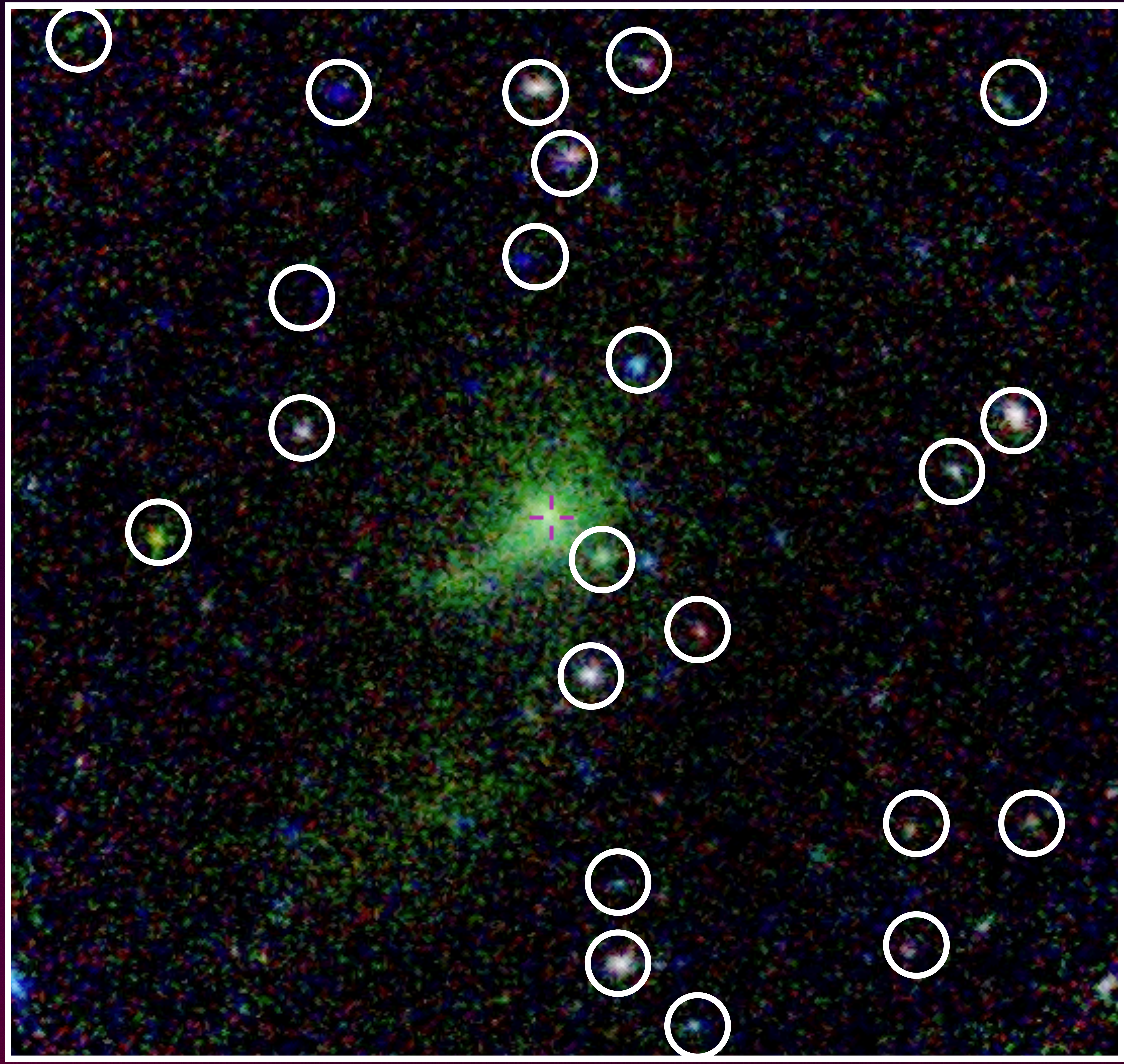


XMM & the 4XMM catalogue

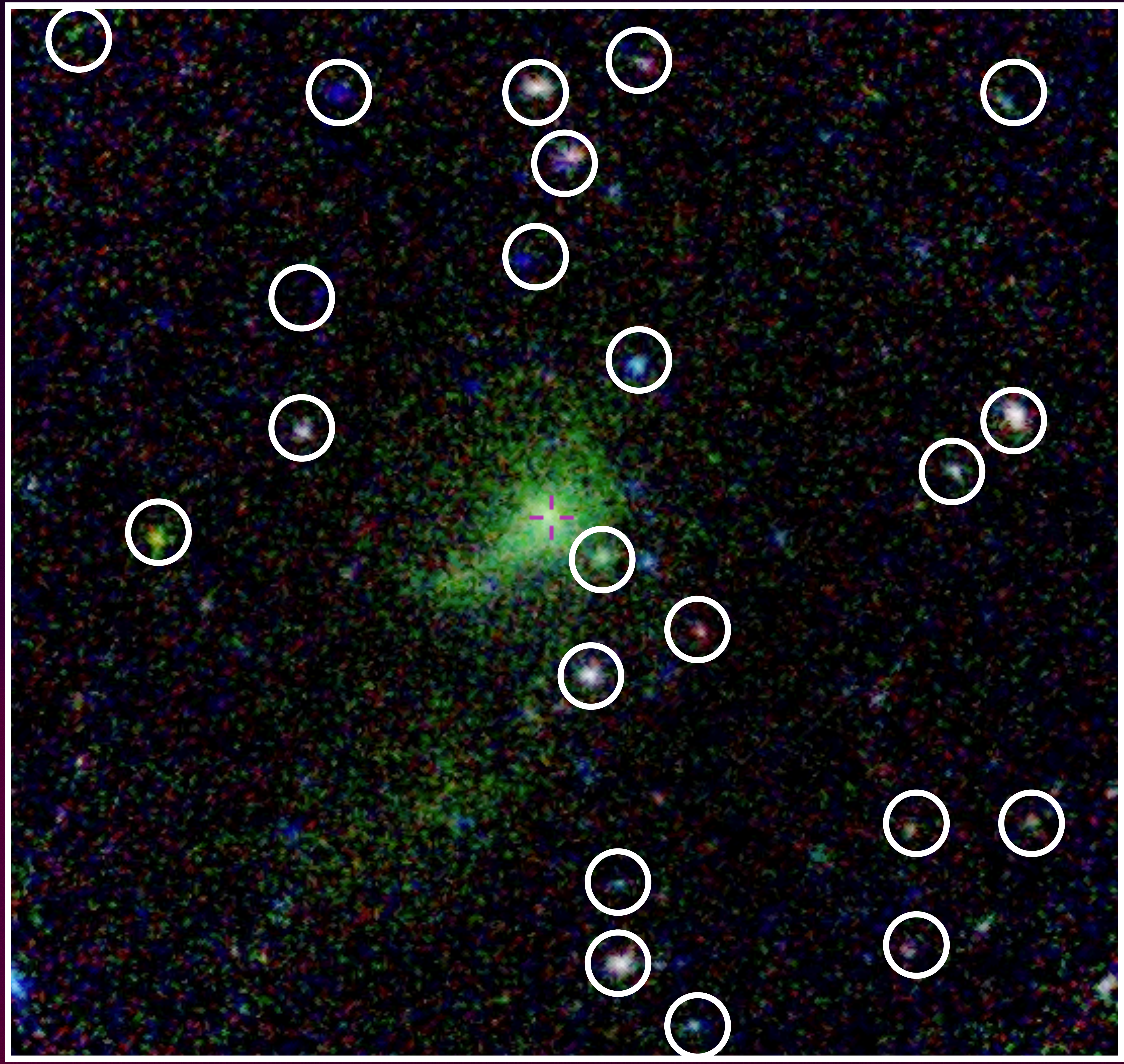


XMM & the 4XMM catalogue

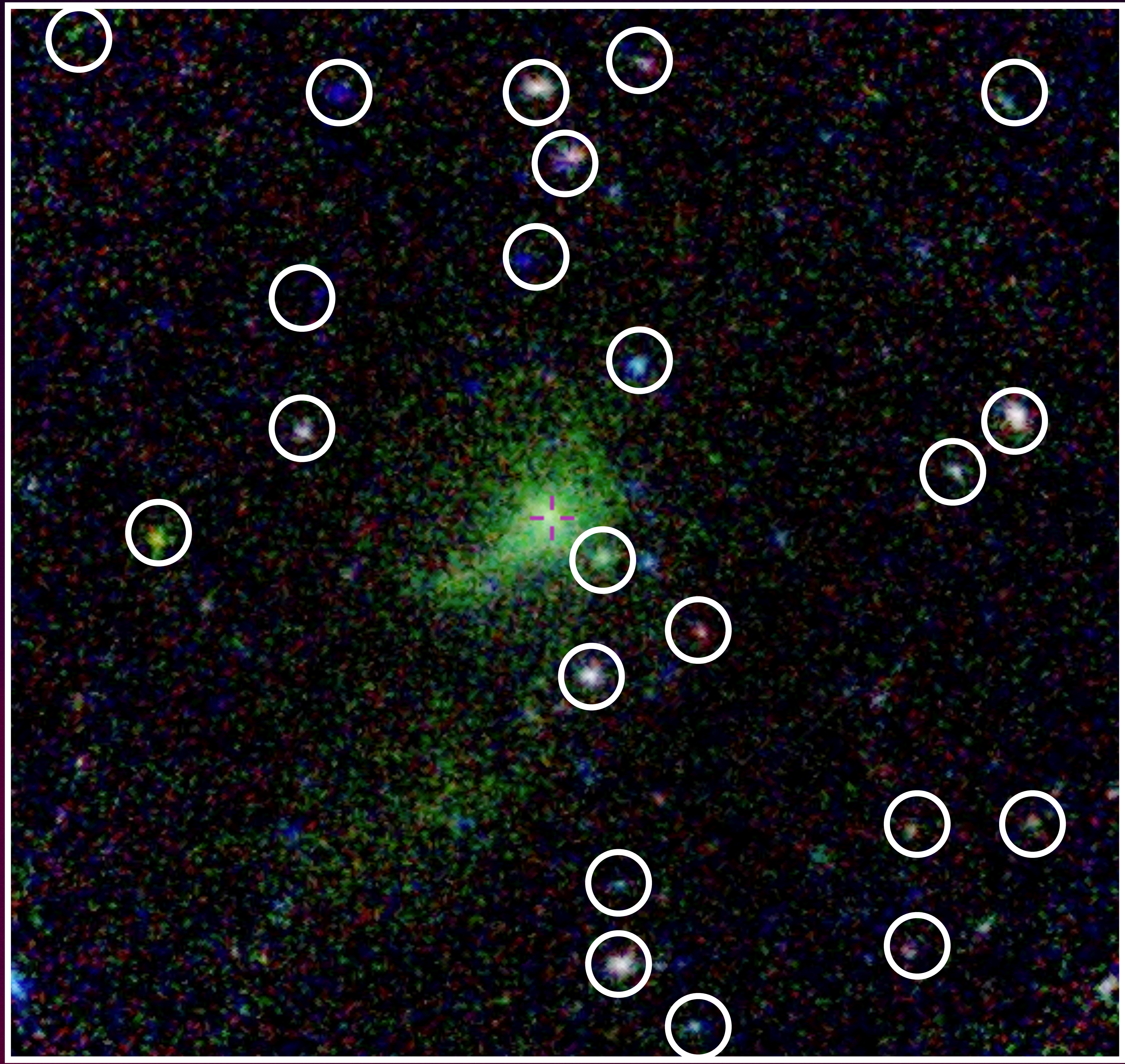




Automated source extraction pipeline

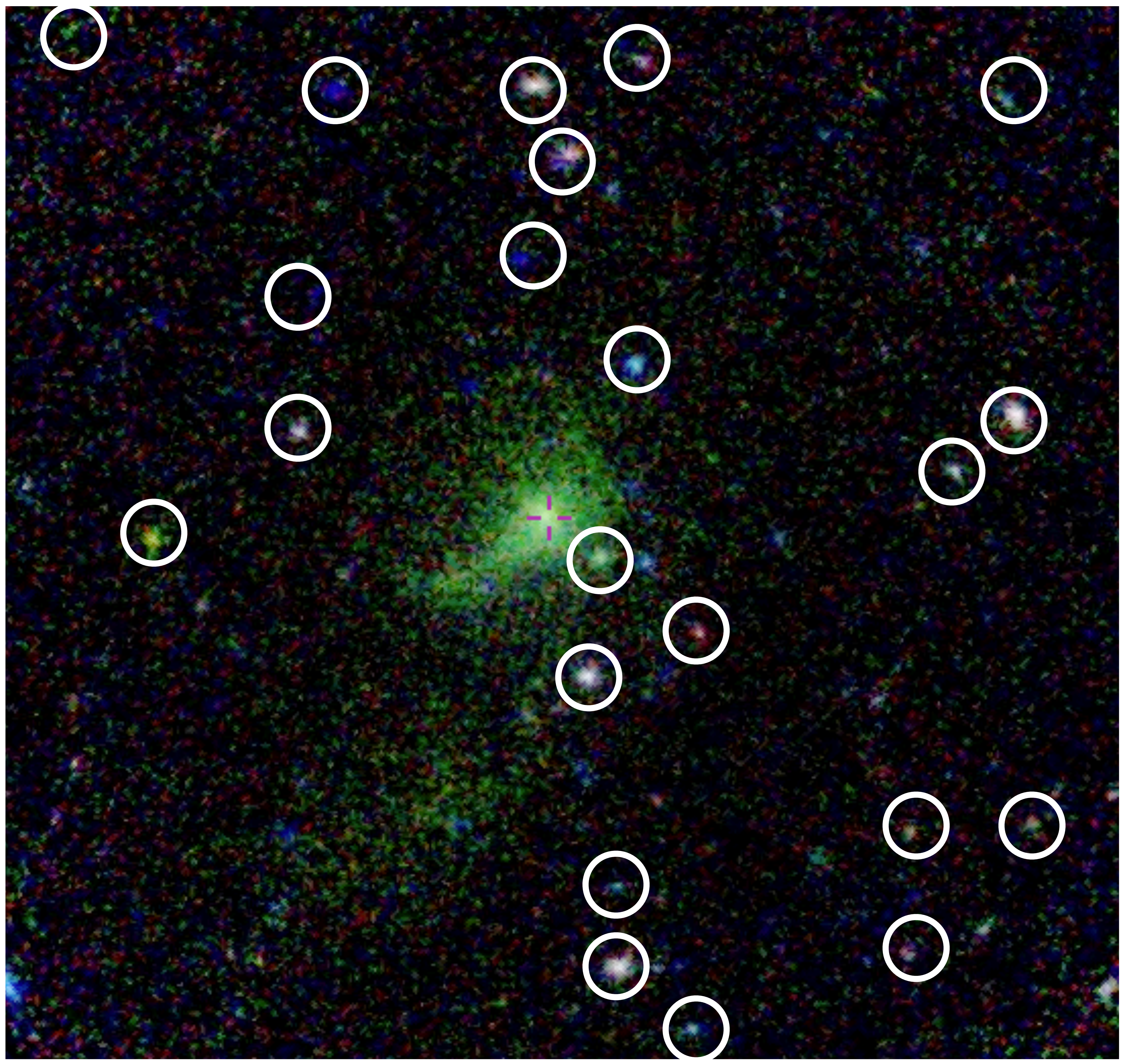


Automated source extraction pipeline



Each XMM pointing brings ~ 100 serendipitous sources

Automated source extraction pipeline

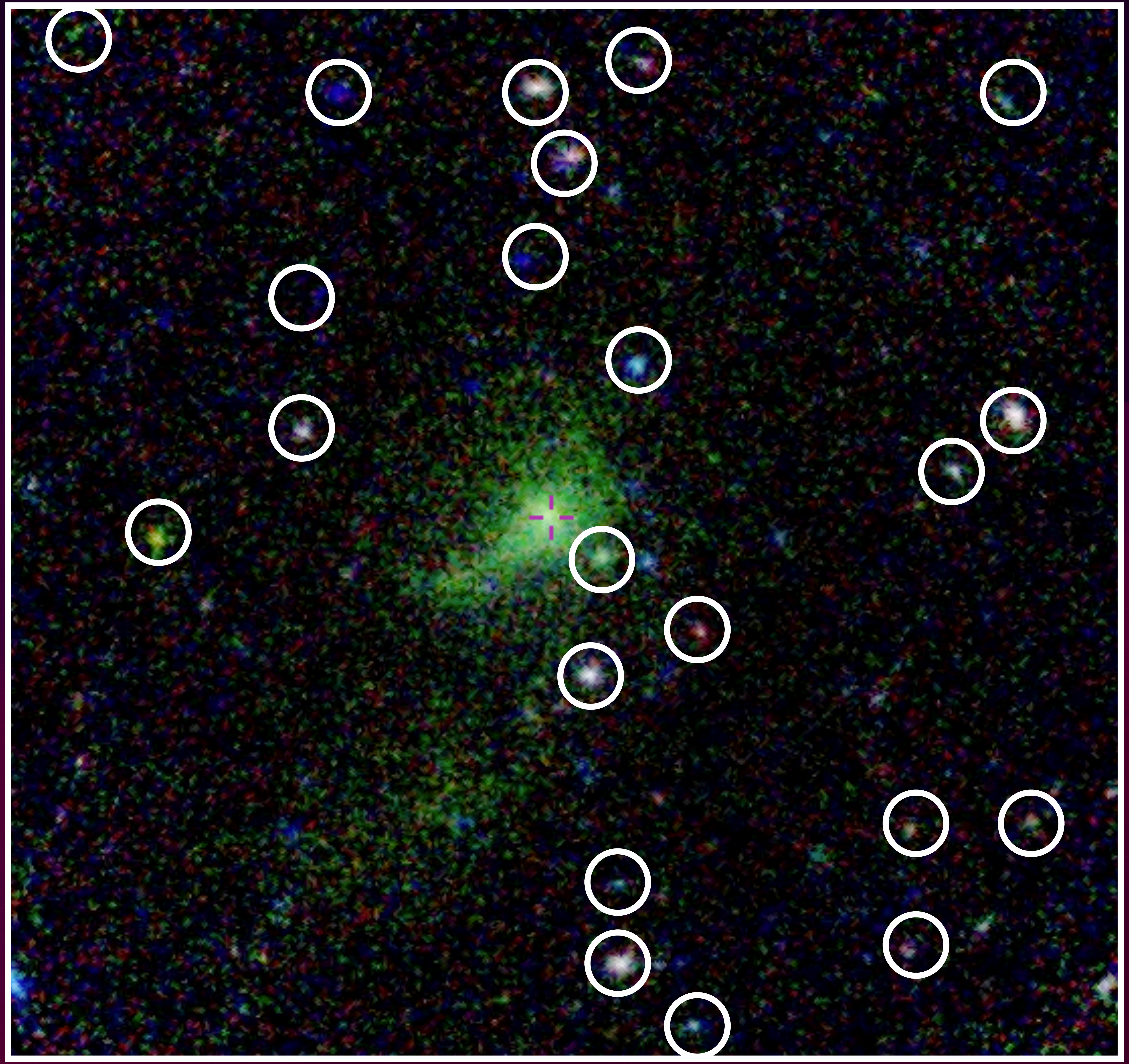


+20 years of service

Each XMM pointing brings ~ 100 serendipitous sources

Automated source extraction pipeline

Each XMM pointing brings ~ 100 serendipitous sources

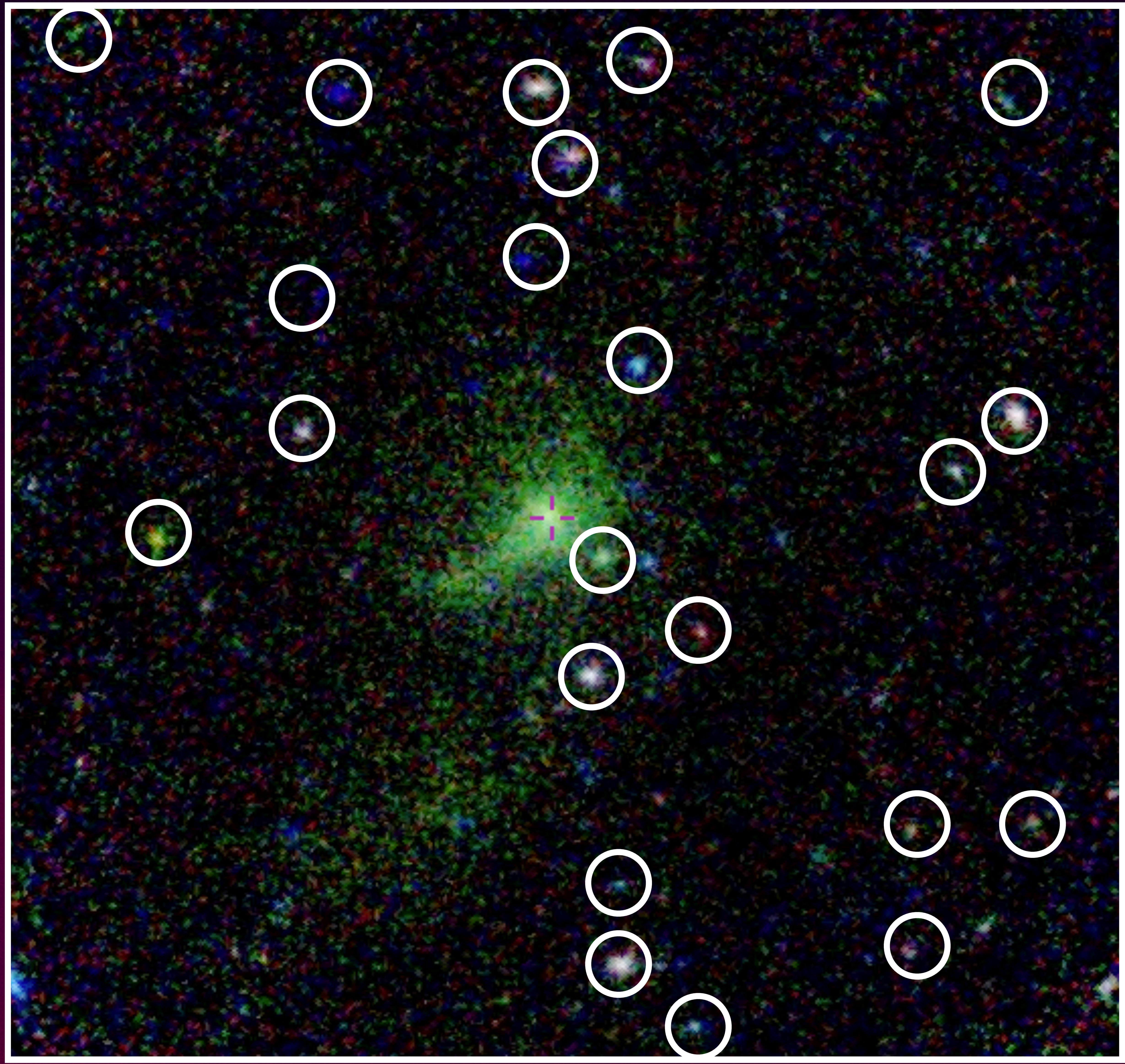


+20 years of service

+13k pointings

Automated source extraction pipeline

Each XMM pointing brings ~ 100 serendipitous sources



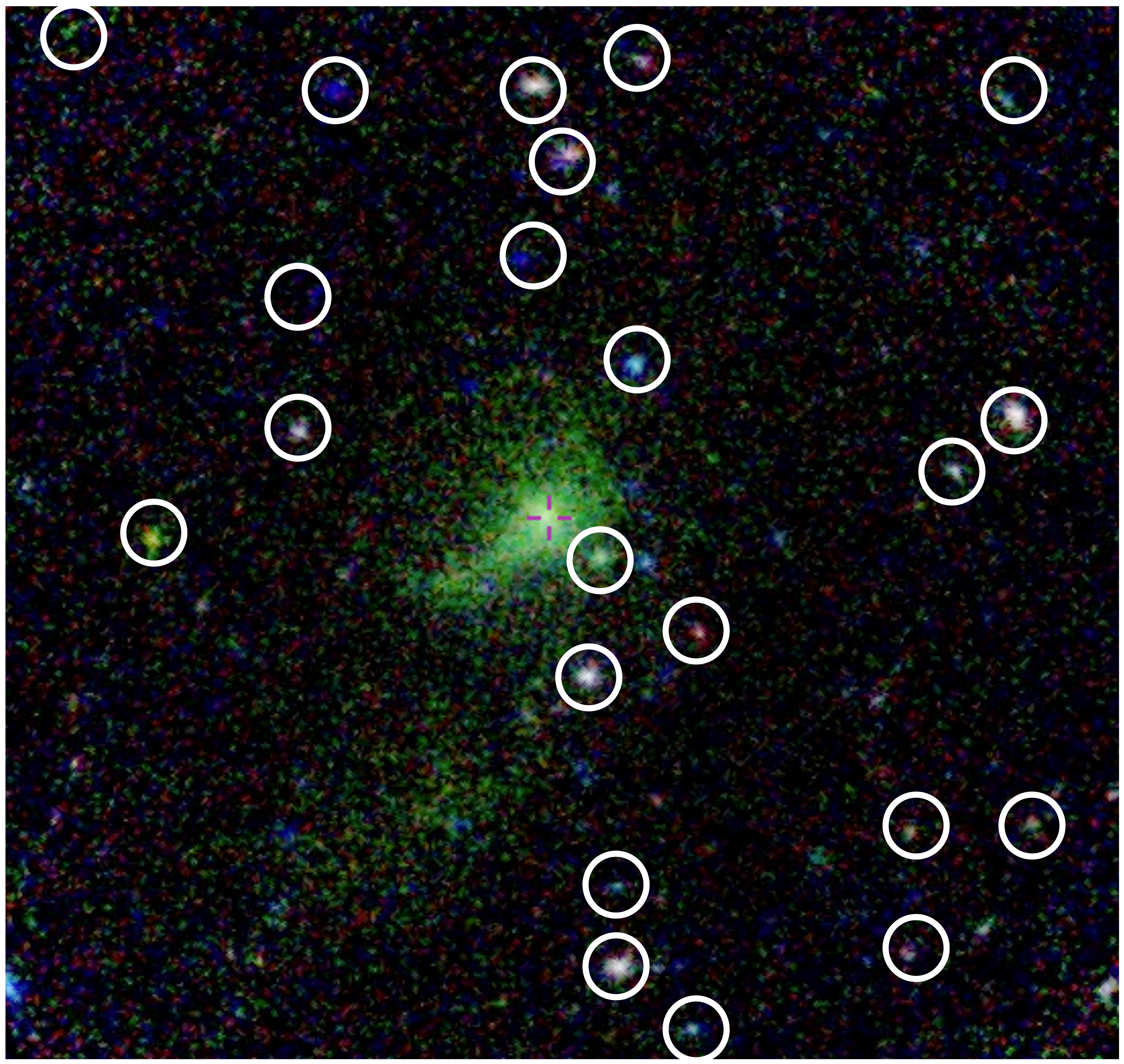
+20 years of service

+13k pointings

+ 1M sources to exploit

Automated source extraction pipeline

Each XMM pointing brings ~ 100 serendipitous sources



+20 years of service

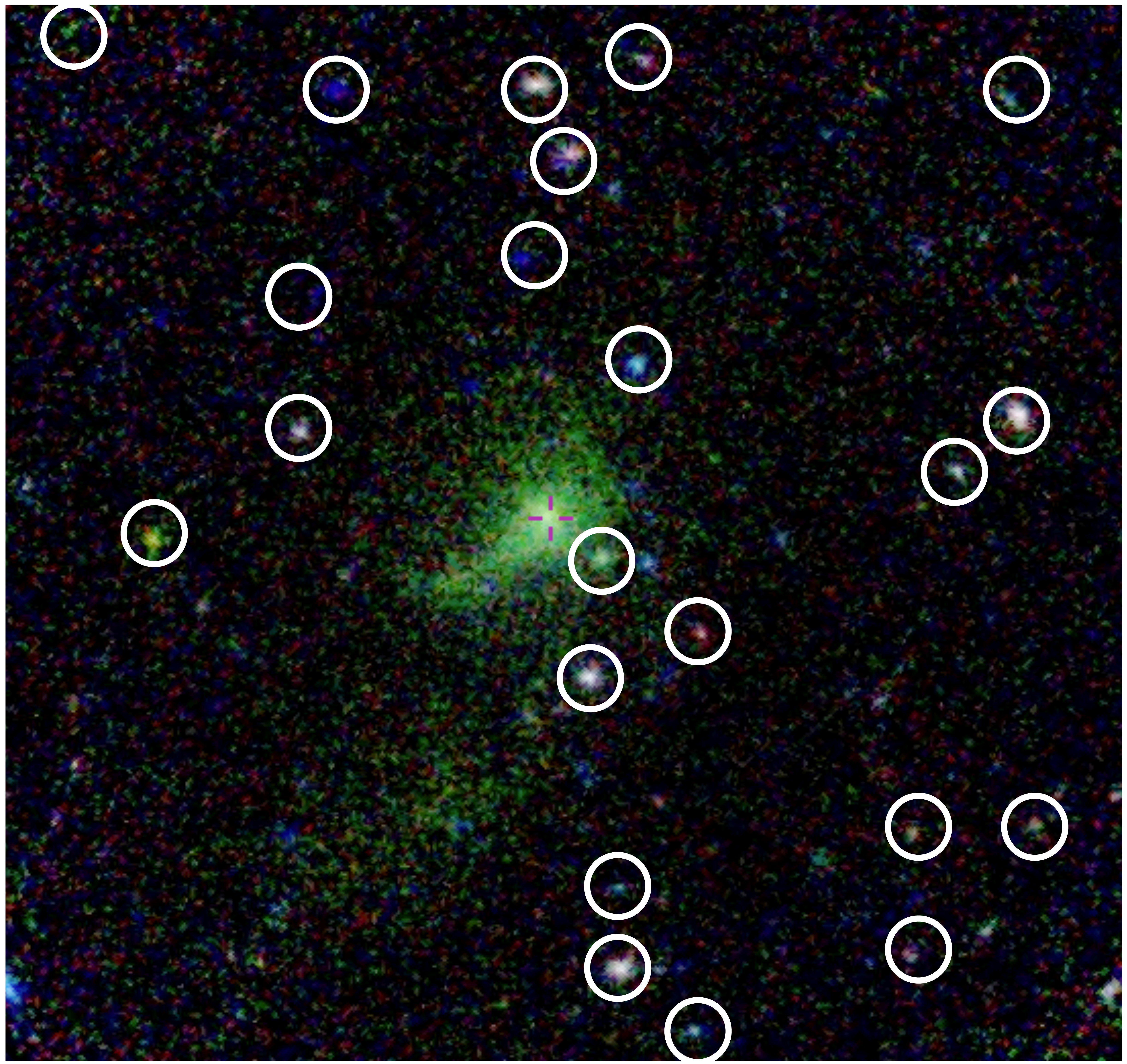
+13k pointings

+ 1M sources to exploit

Focus on spectral data

Automated source extraction pipeline

Each XMM pointing brings ~ 100 serendipitous sources



+20 years of service

+13k pointings

+ 1M sources to exploit

Focus on spectral data

What can be achieved through the self-supervised learning of spectra using the 4XMM catalogue ?

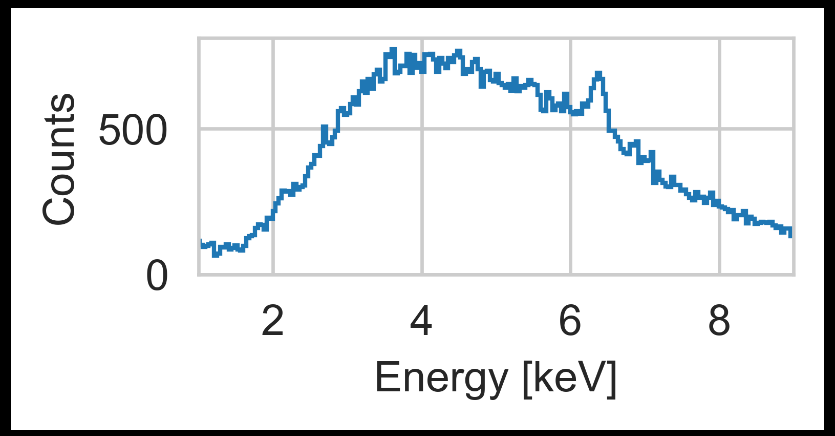
Tentative work!

Autoencoding the spectra

Observation space

High dimension

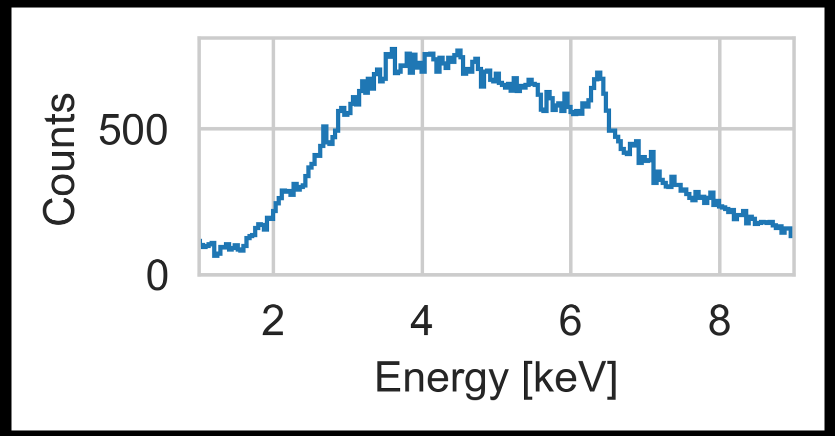
$\mathcal{S}$



Autoencoding the spectra

Observation space
High dimension

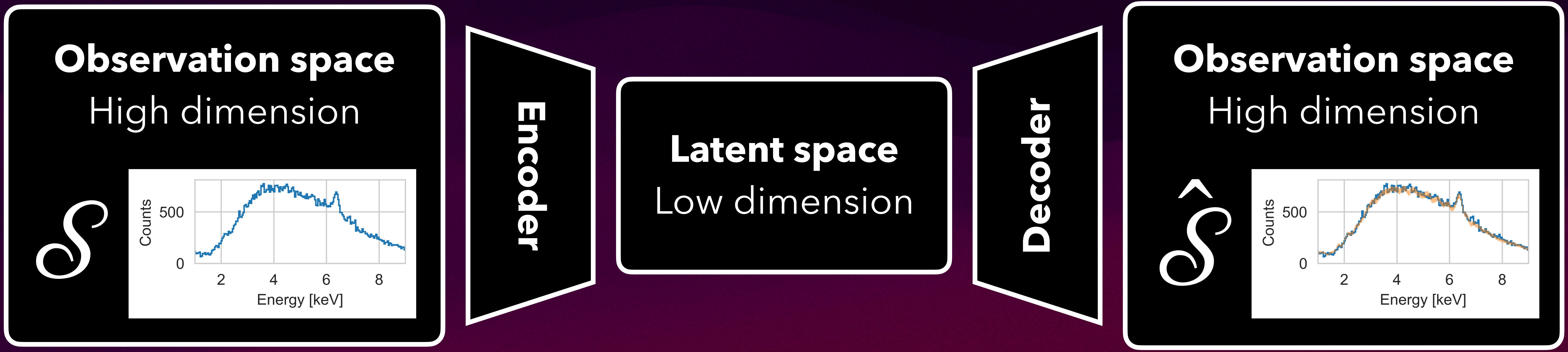
$\mathcal{S}$



Encoder

Latent space
Low dimension

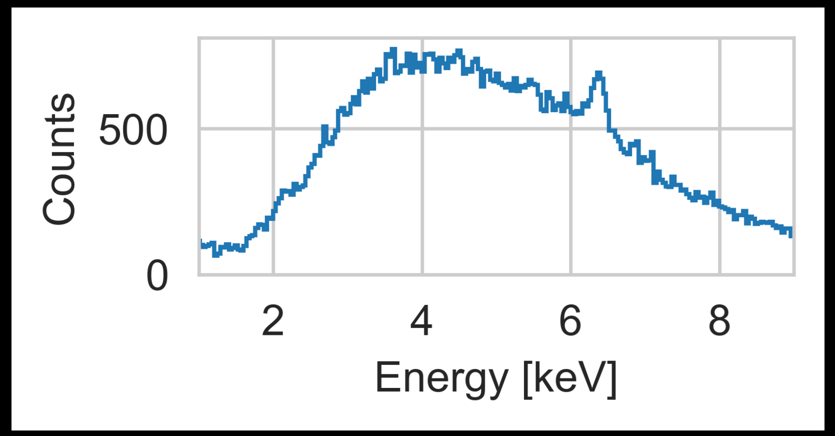
Autoencoding the spectra



Autoencoding the spectra

Observation space
High dimension

$\mathcal{S}$



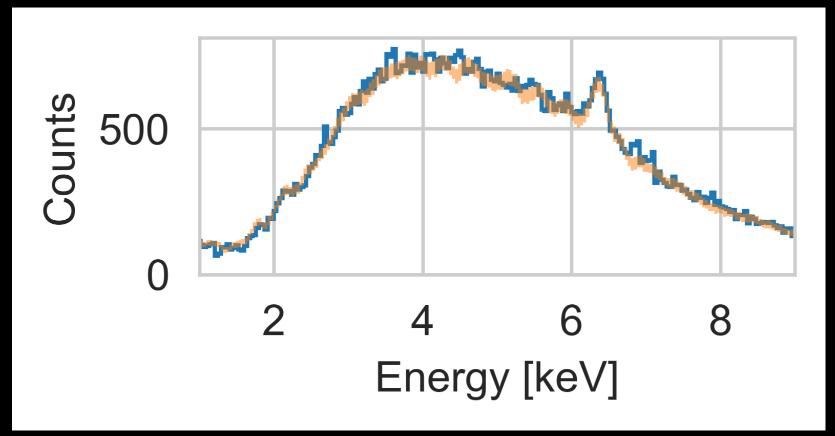
Encoder

Latent space
Low dimension

Decoder

Observation space
High dimension

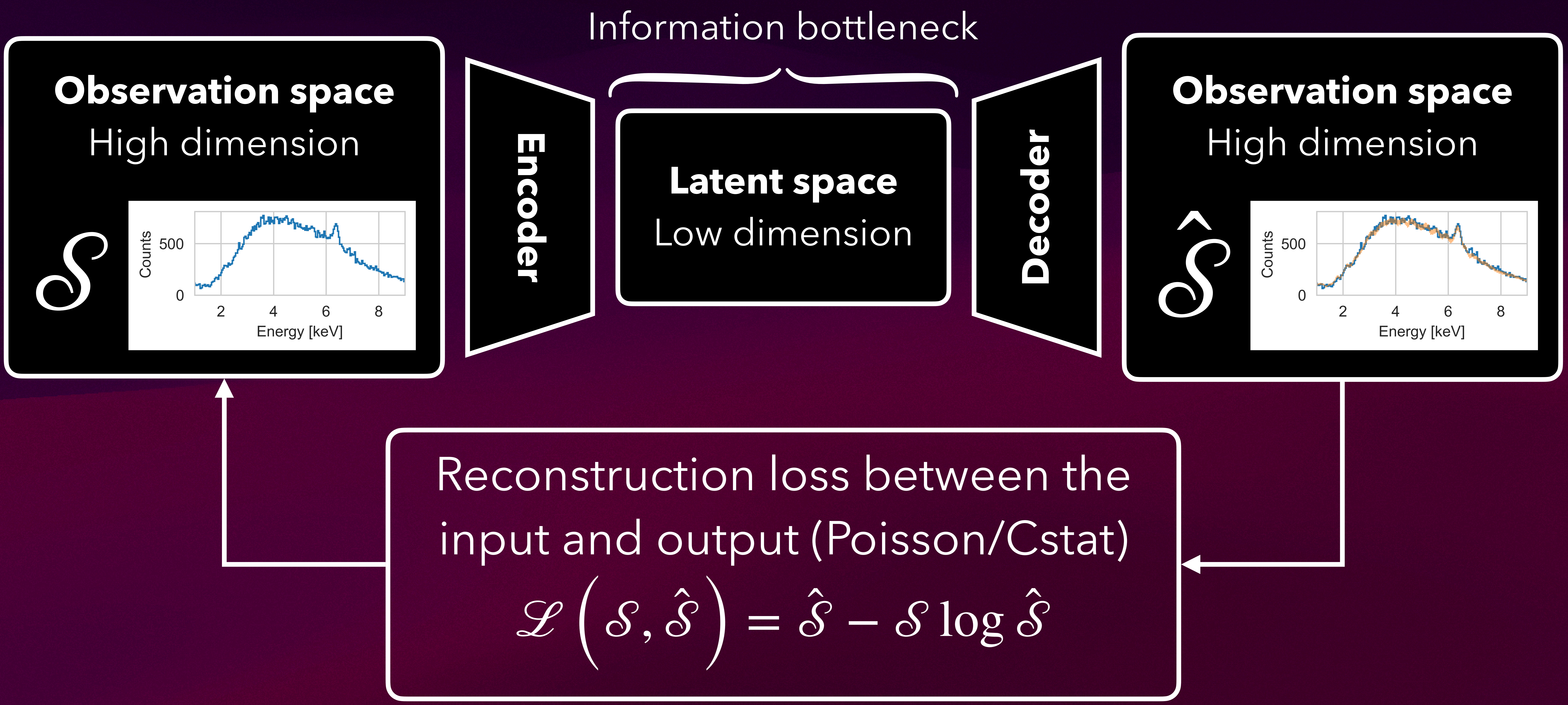
$\hat{\mathcal{S}}$



Reconstruction loss between the input and output (Poisson/Cstat)

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

Autoencoding the spectra

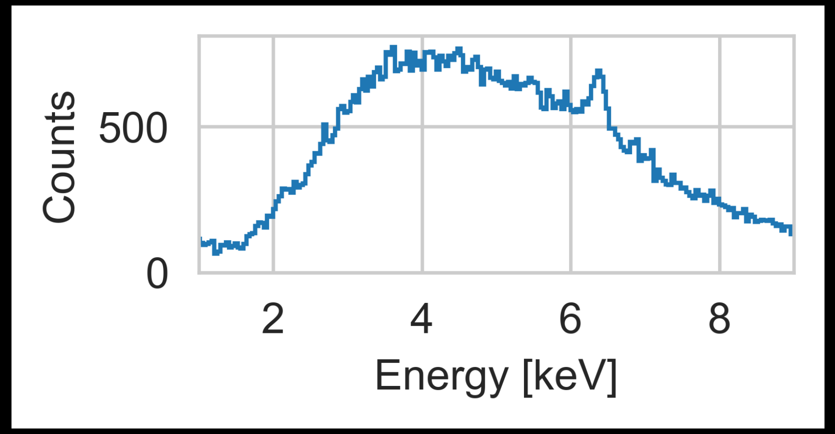


Autoencoding the spectra

Information bottleneck

Observation space
High dimension

$\mathcal{S}$



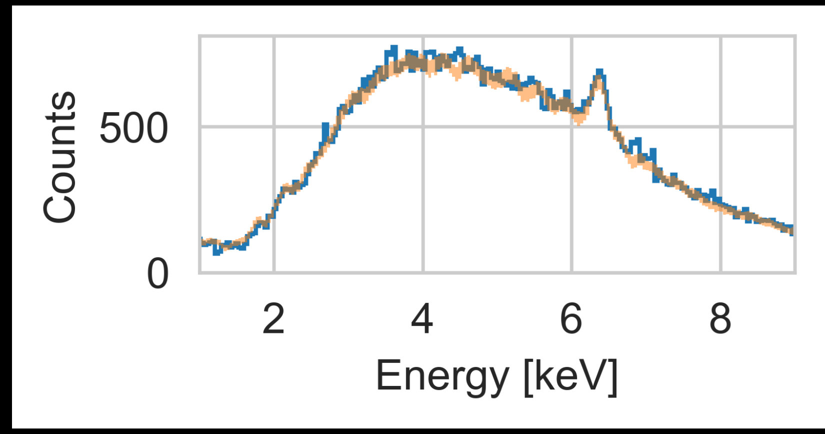
Encoder

Latent space
Low dimension

Decoder

Observation space
High dimension

$\hat{\mathcal{S}}$



Reconstruction loss between the input and output (Poisson/Cstat)

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

Denoising, Clustering, Outlier detection …

Going variational (VAE)

Latent space
Low dimension

Going variational (VAE)

Latent space
Low dimension

μ

σ

Going variational (VAE)

Latent space
Low dimension

μ

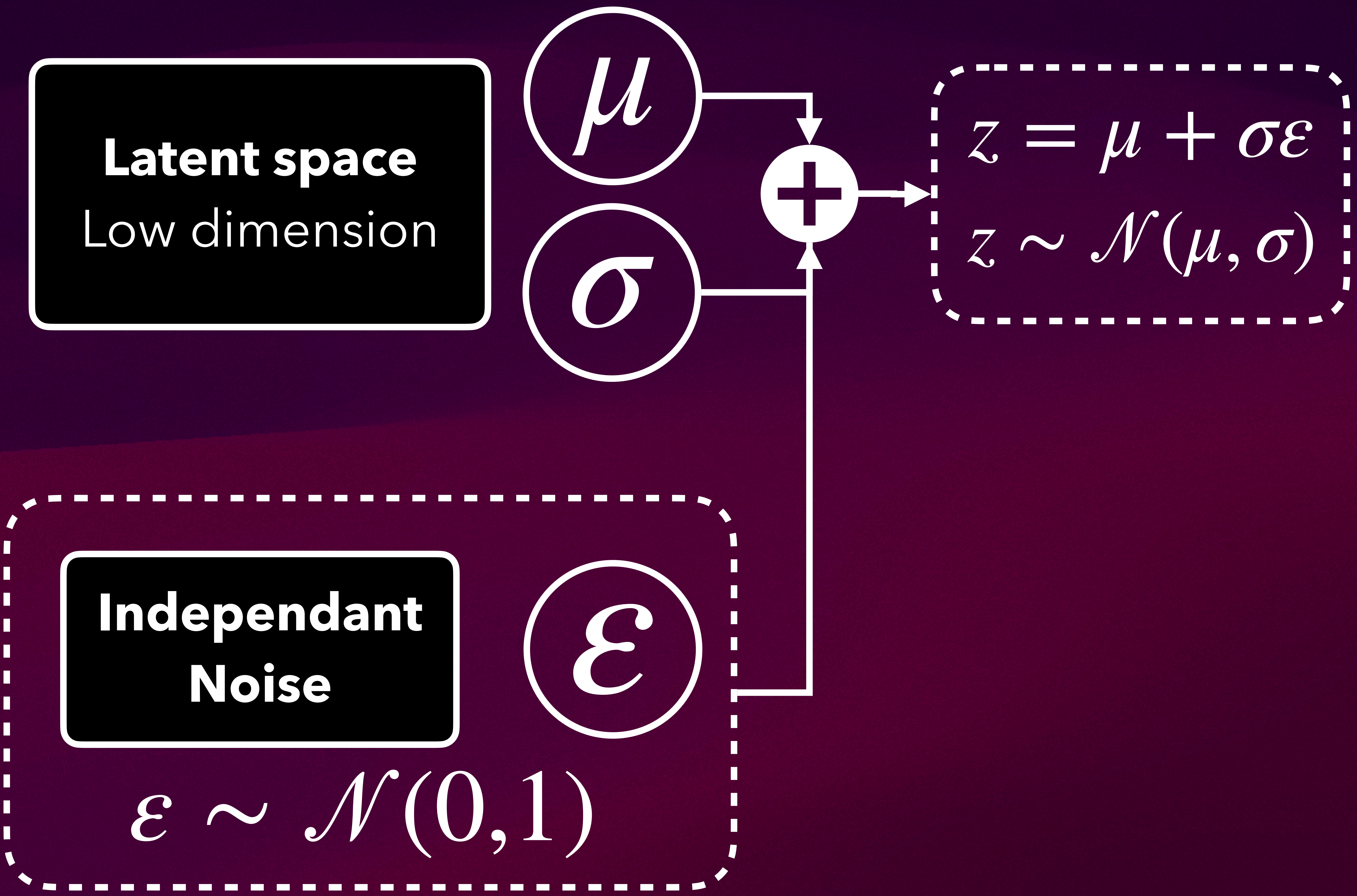
σ

Independant Noise

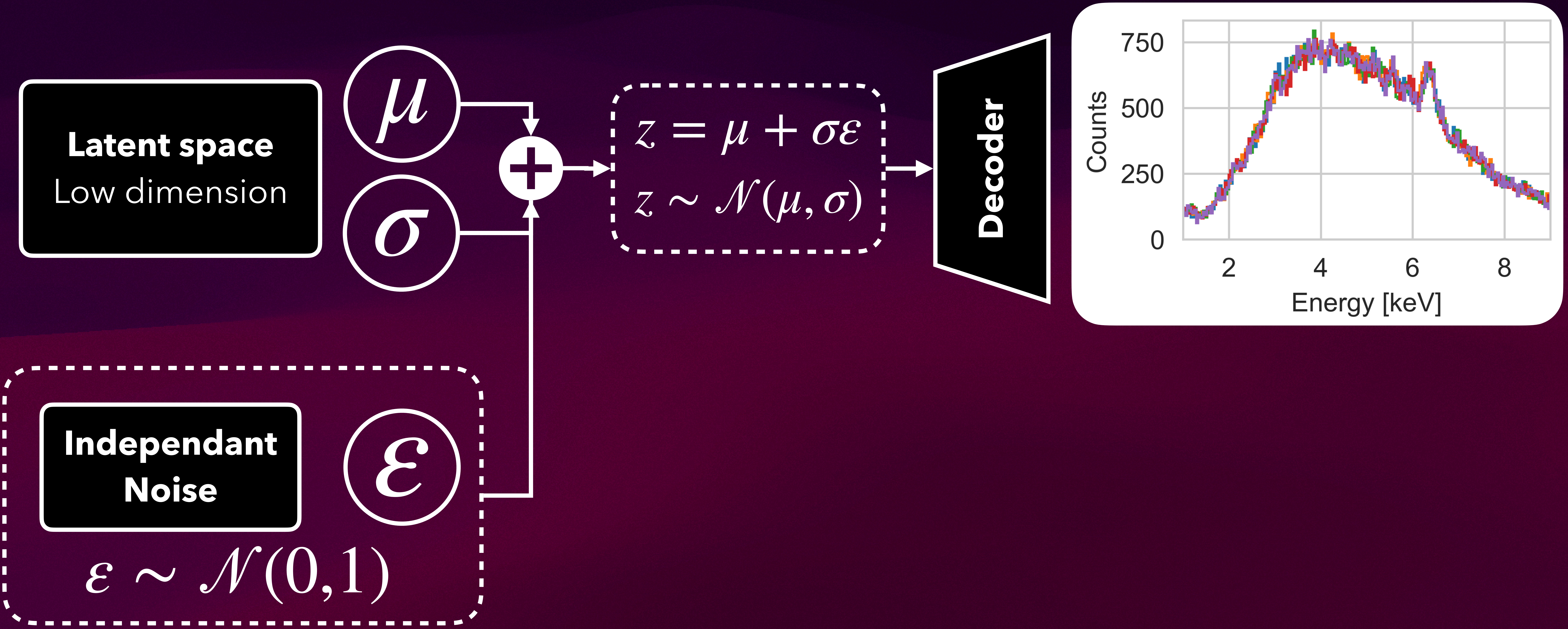
ε

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

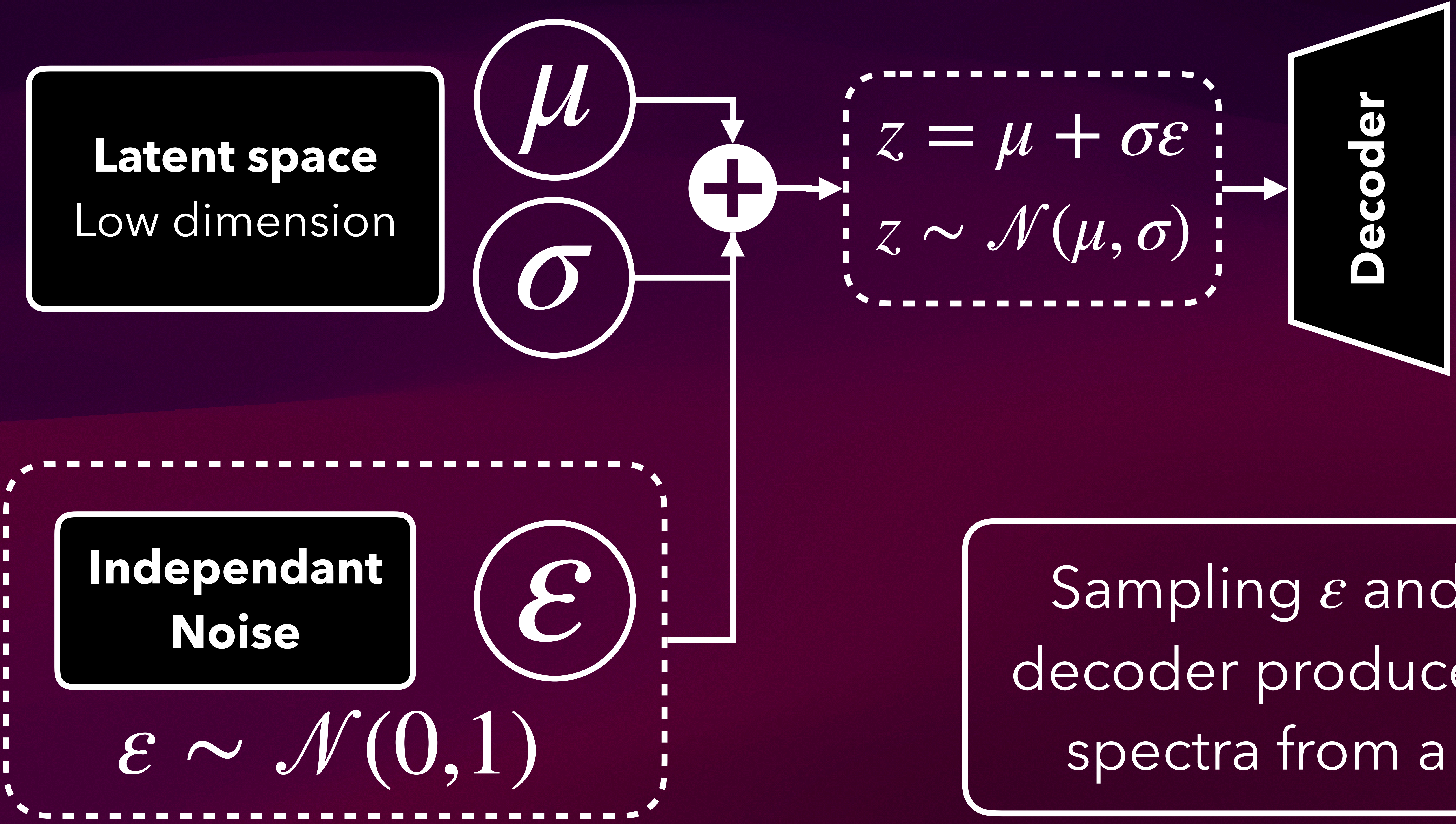
Going variational (VAE)

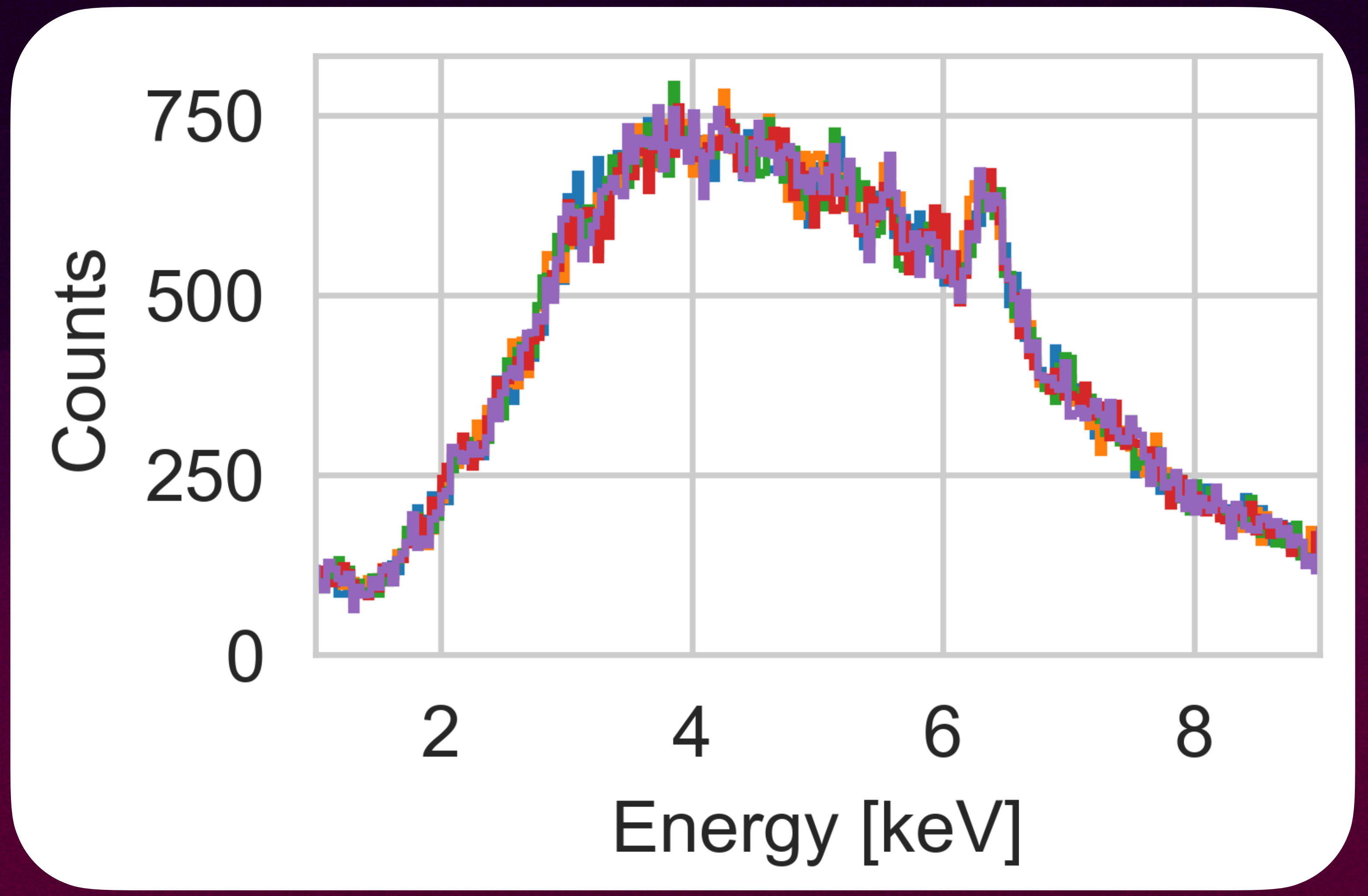


Going variational (VAE)



Going variational (VAE)





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

(Great synergy with Poisson loss)

Decoding is the hard part

MLPs are not performant at decoding

Decoding is the hard part

MLPs are not performant at decoding

Switch to a **DeepONet** architecture

Decoding is the hard part

MLPs are not performant at decoding

Switch to a **DeepONet** architecture

$$\hat{\mathcal{S}}(E,\theta) \simeq \sum_i \Phi_i(E) \times c_i(\theta)$$

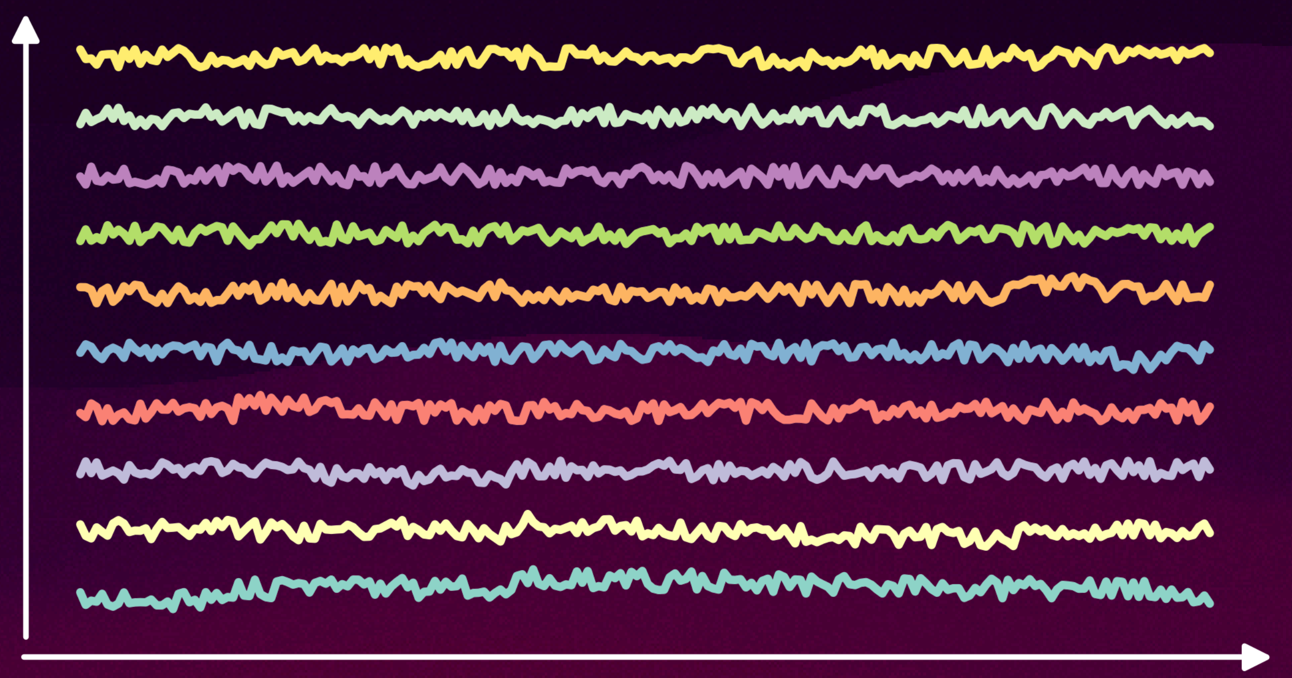
Decoding is the hard part

MLPs are not performant at decoding

Switch to a **DeepONet** architecture

Trunk Network

Predicts a set of *basis* function to combines



$$\hat{\mathcal{S}}(E,\theta) \simeq \sum_i \boxed{\Phi_i(E)} \times c_i(\theta)$$

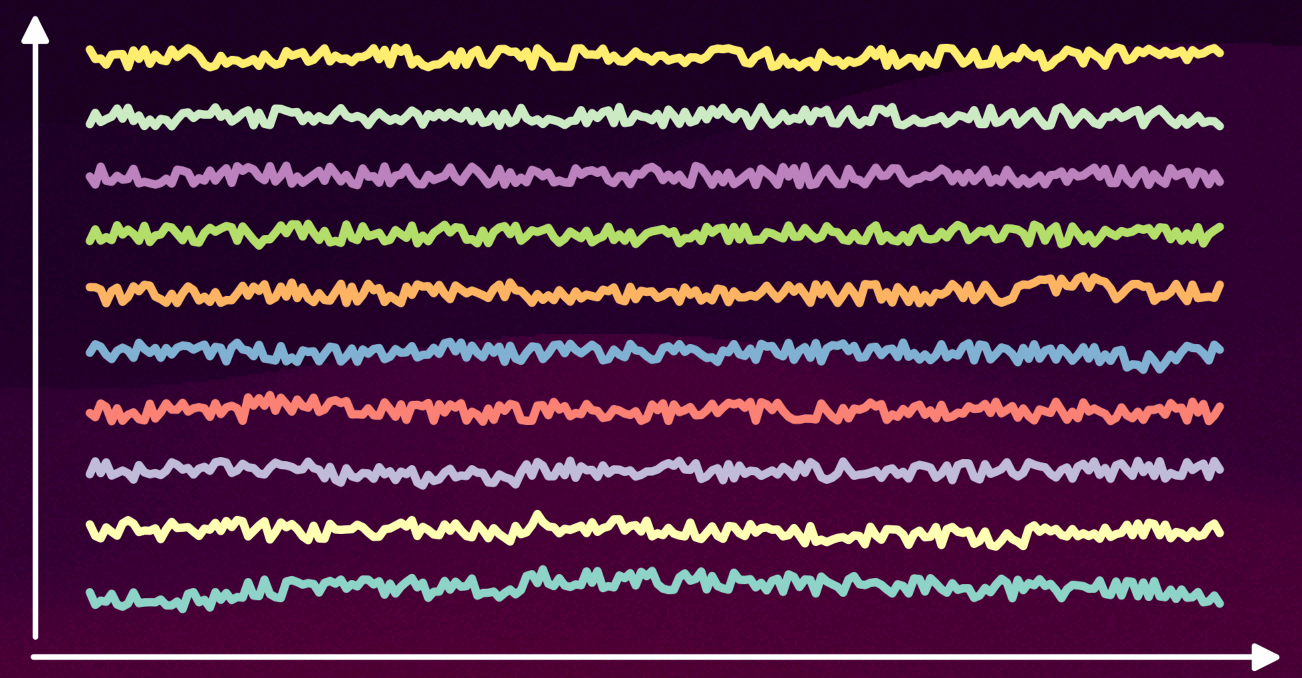
Decoding is the hard part

MLPs are not performant at decoding

Switch to a **DeepONet** architecture

Trunk Network

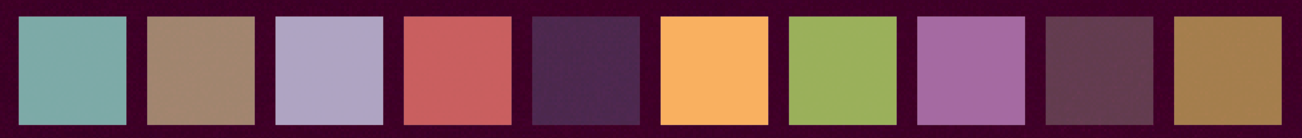
Predicts a set of *basis* function to combines



$$\hat{\mathcal{S}}(E,\theta) \simeq \sum_i \boxed{\Phi_i(E)} \times \boxed{c_i(\theta)}$$

Branch Network

Predicts a way to combine the trunks



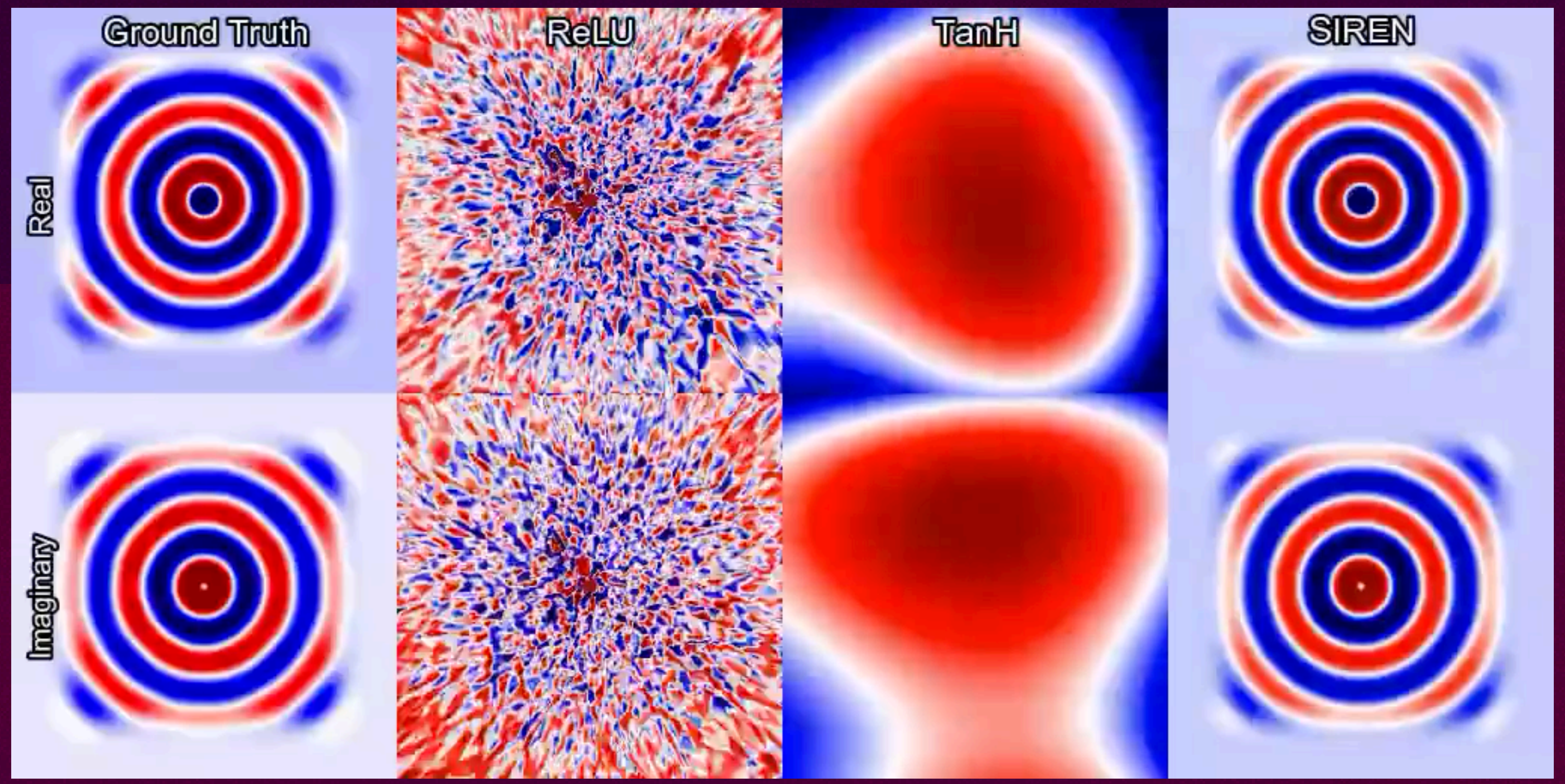
e.g. coefficients for linear combination

DeepOnet implementation

DeepOnet implementation

Trunk Network

Uses a SIREN network to learn continuous features

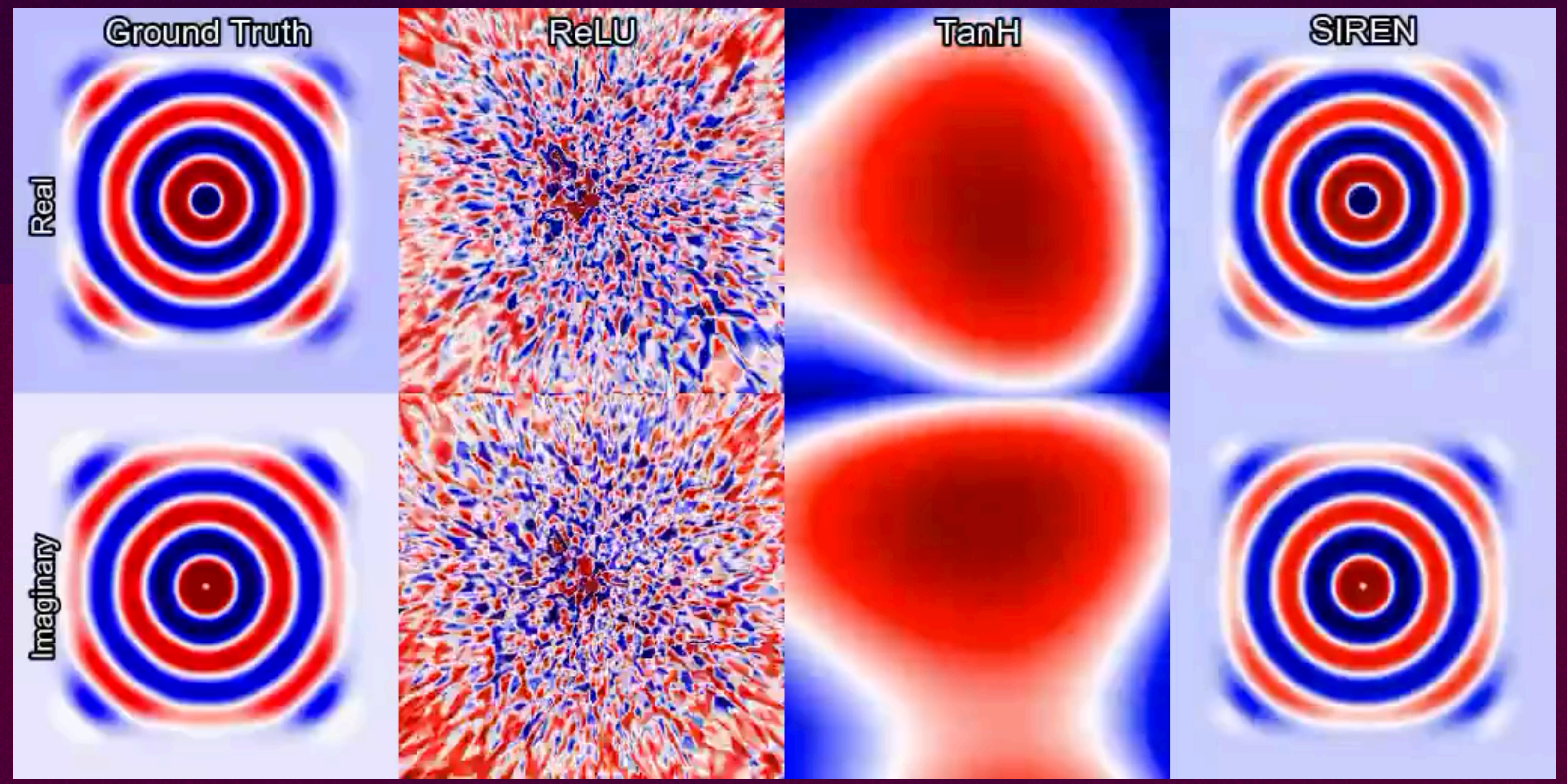


Truth ReLU TanH Sine

DeepOnet implementation

Trunk Network

Uses a SIREN network to learn continuous features



Truth ReLU TanH Sine

Branch Network

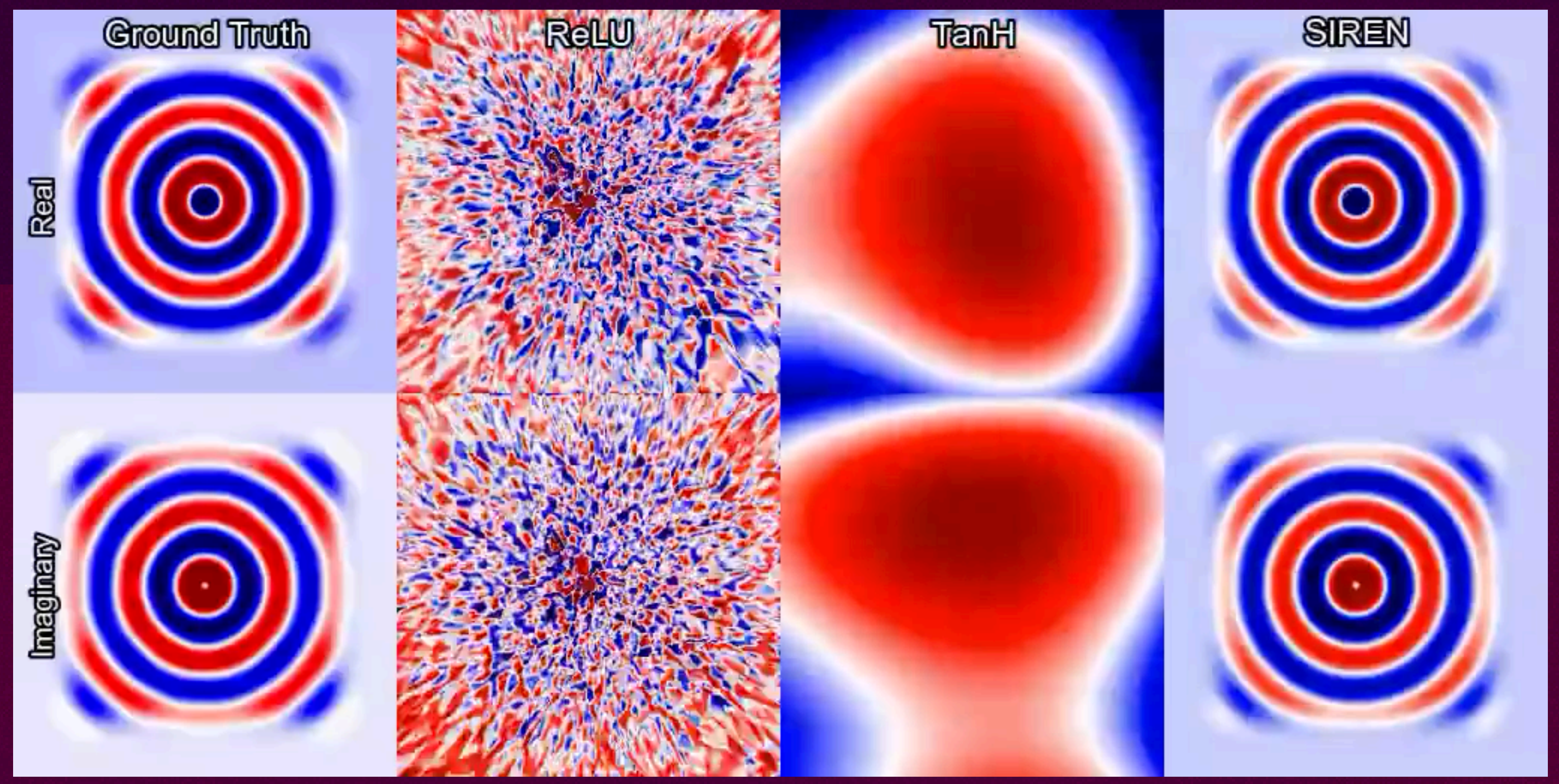
Standard MLP :

- Amplitude $\mathscr{A}$
- Shift ϕ
- Smoothing α

DeepOnet implementation

Trunk Network

Uses a SIREN network to learn continuous features



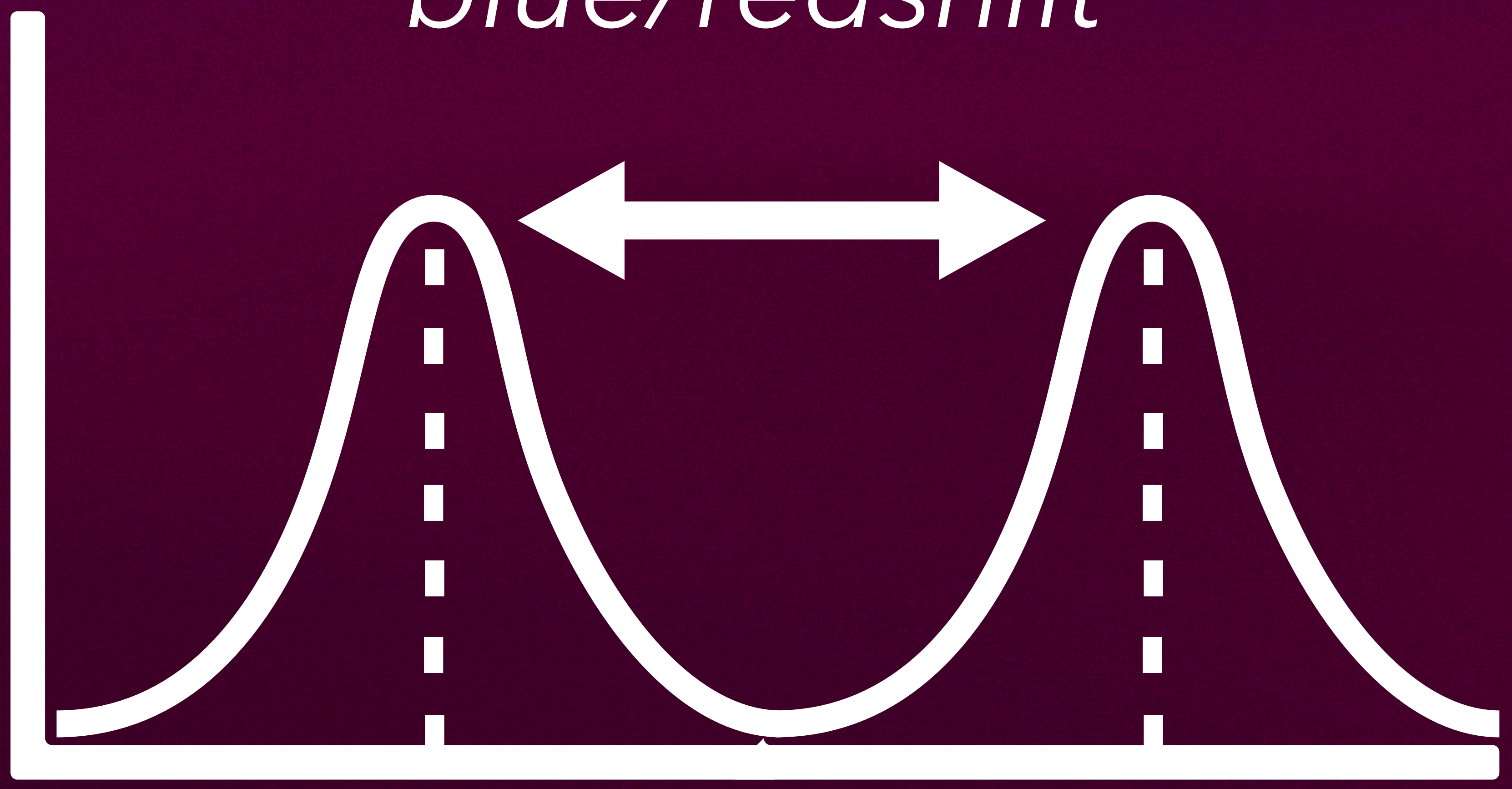
Truth ReLU TanH Sine

Branch Network

Standard MLP :

- Amplitude $\mathscr{A}$
- Shift ϕ
- Smoothing α

Suited to reproduce phenomena such as *blue/redshift*



Count spectrum
[238 dims]

Count spectrum
[238 dims]

↓

Preprocessing

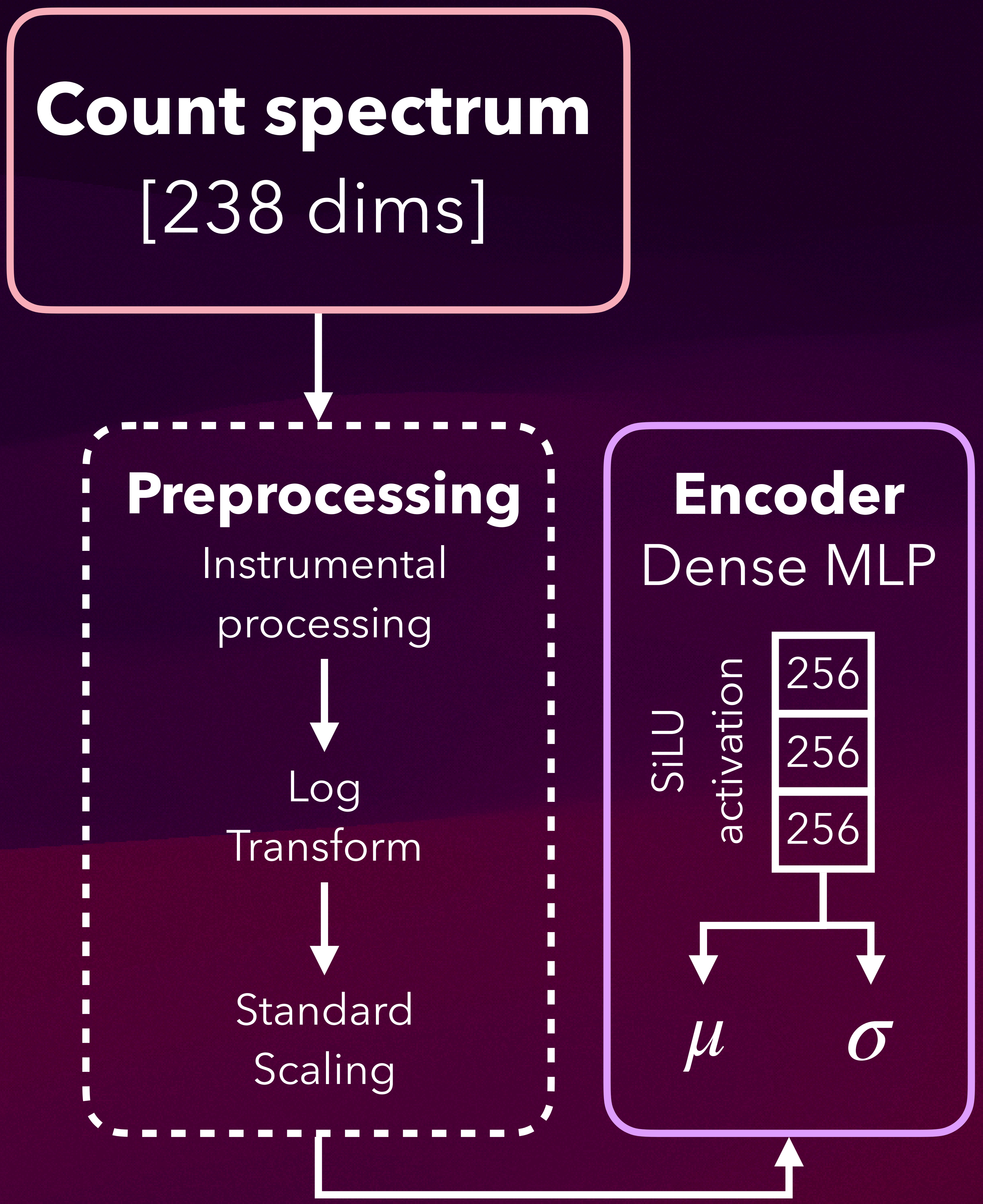
Instrumental processing

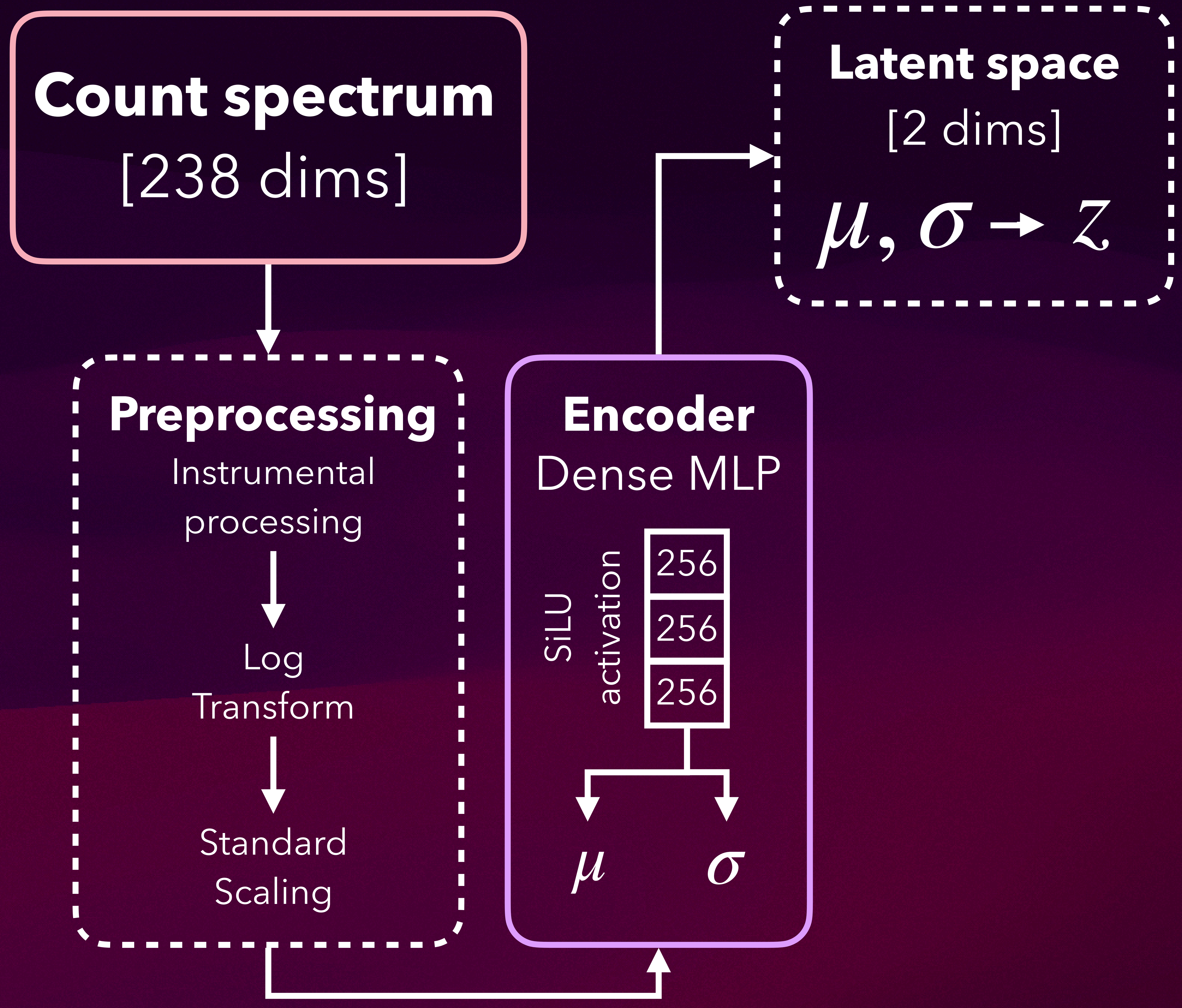
↓

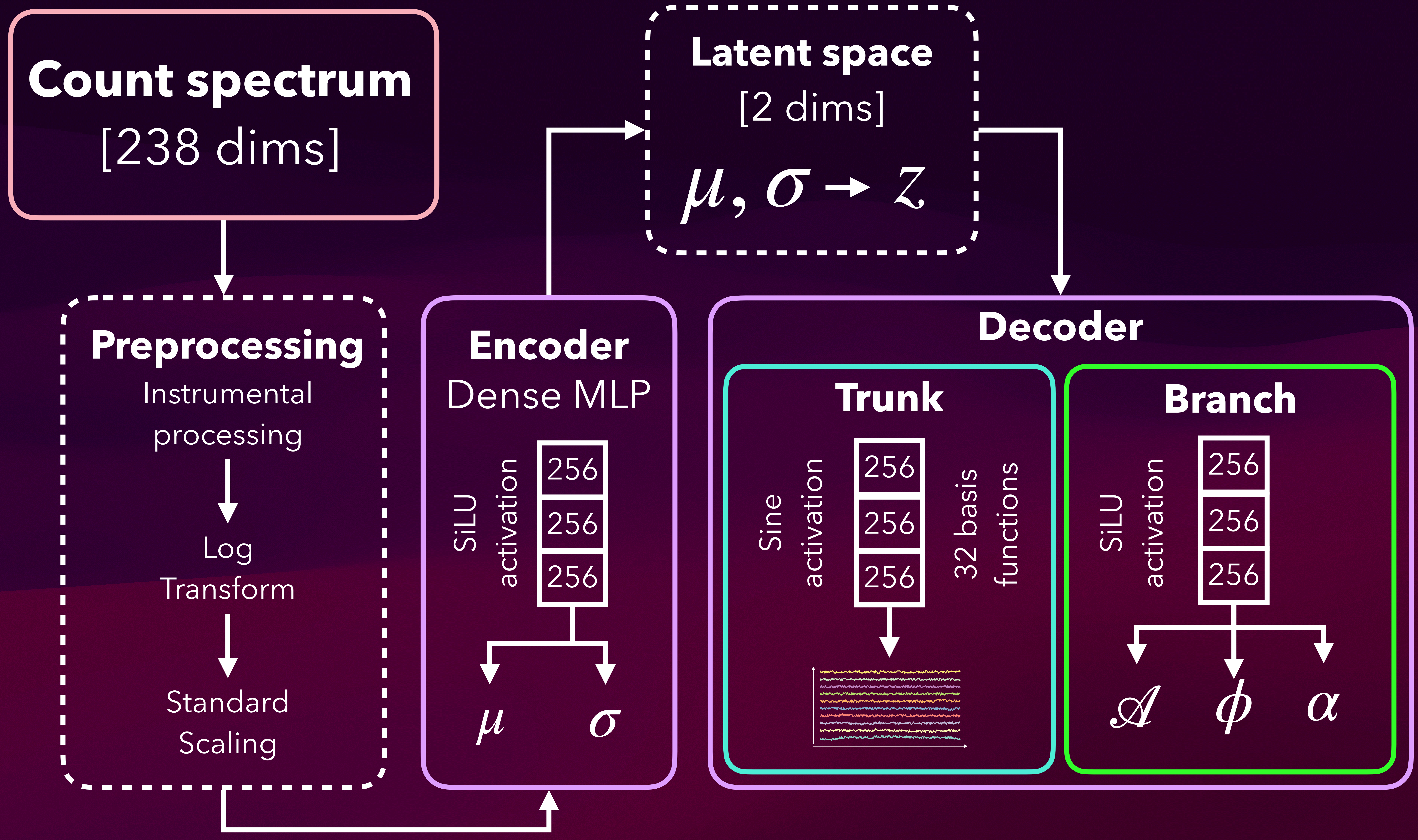
Log Transform

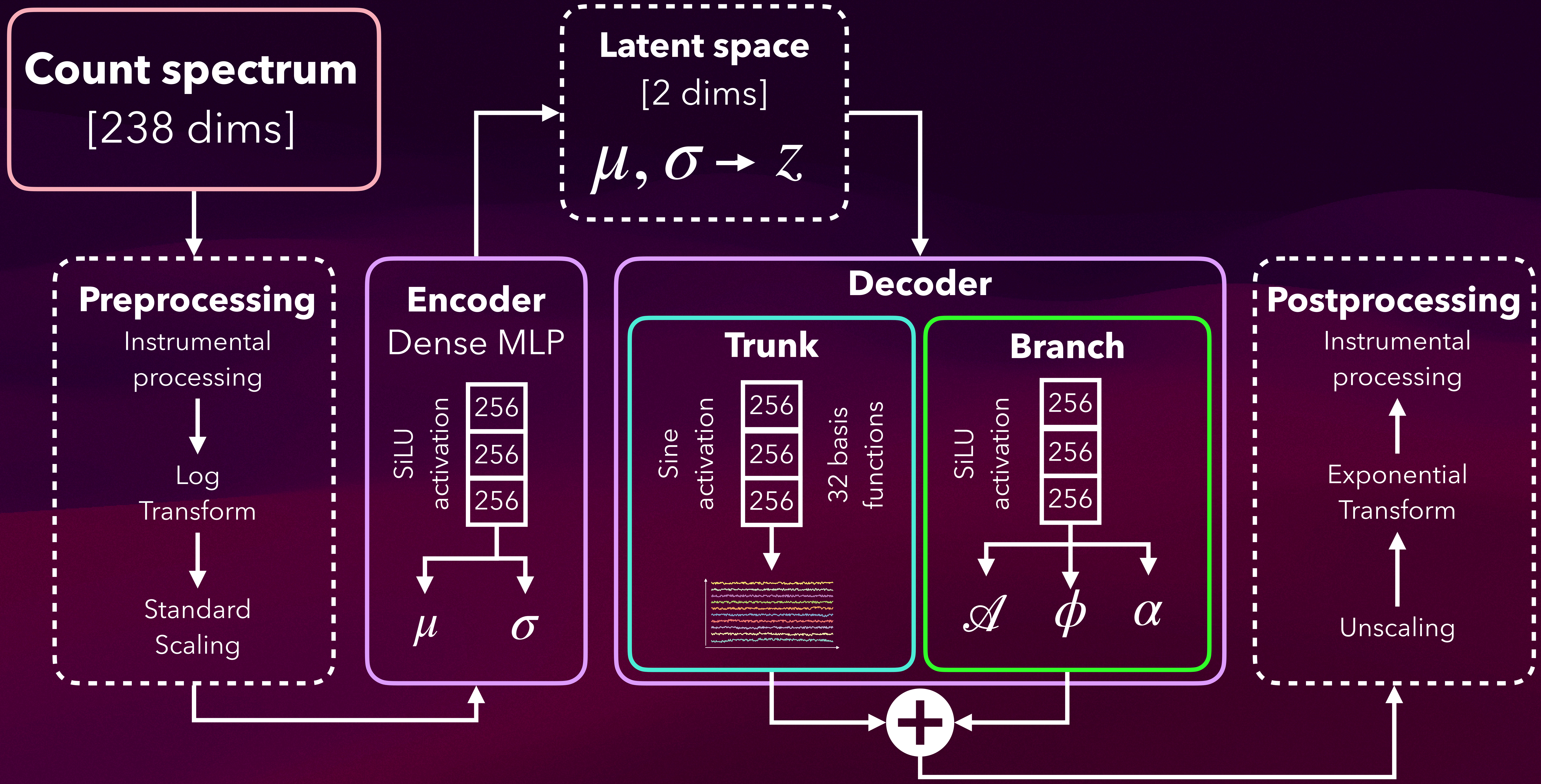
↓

Standard Scaling

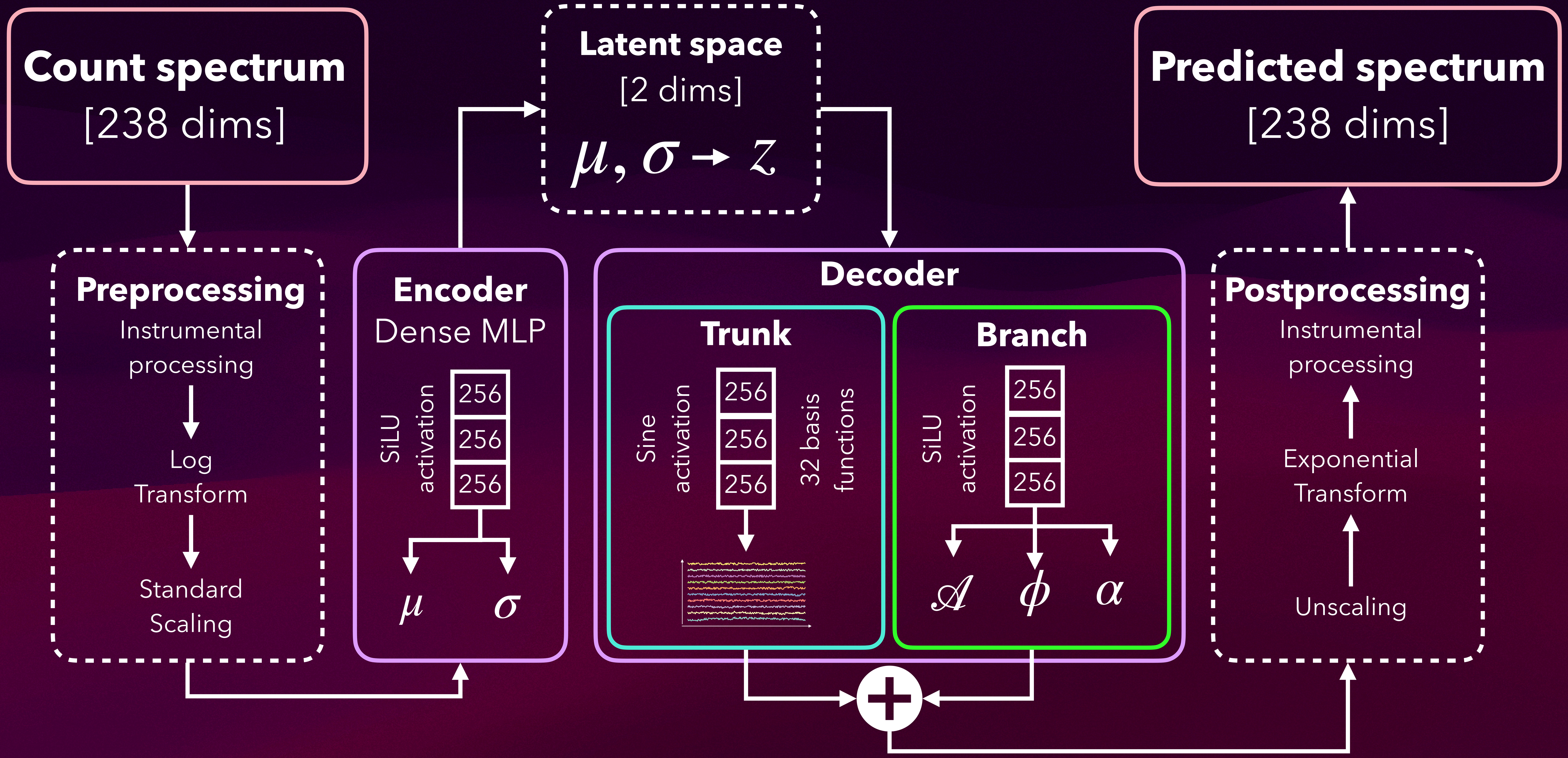

Count spectrum
[238 dims]
Preprocessing
Instrumental processing
Log Transform
Standard Scaling
Encoder
Dense MLP
SiLU activation
256
256
256
μ
σ

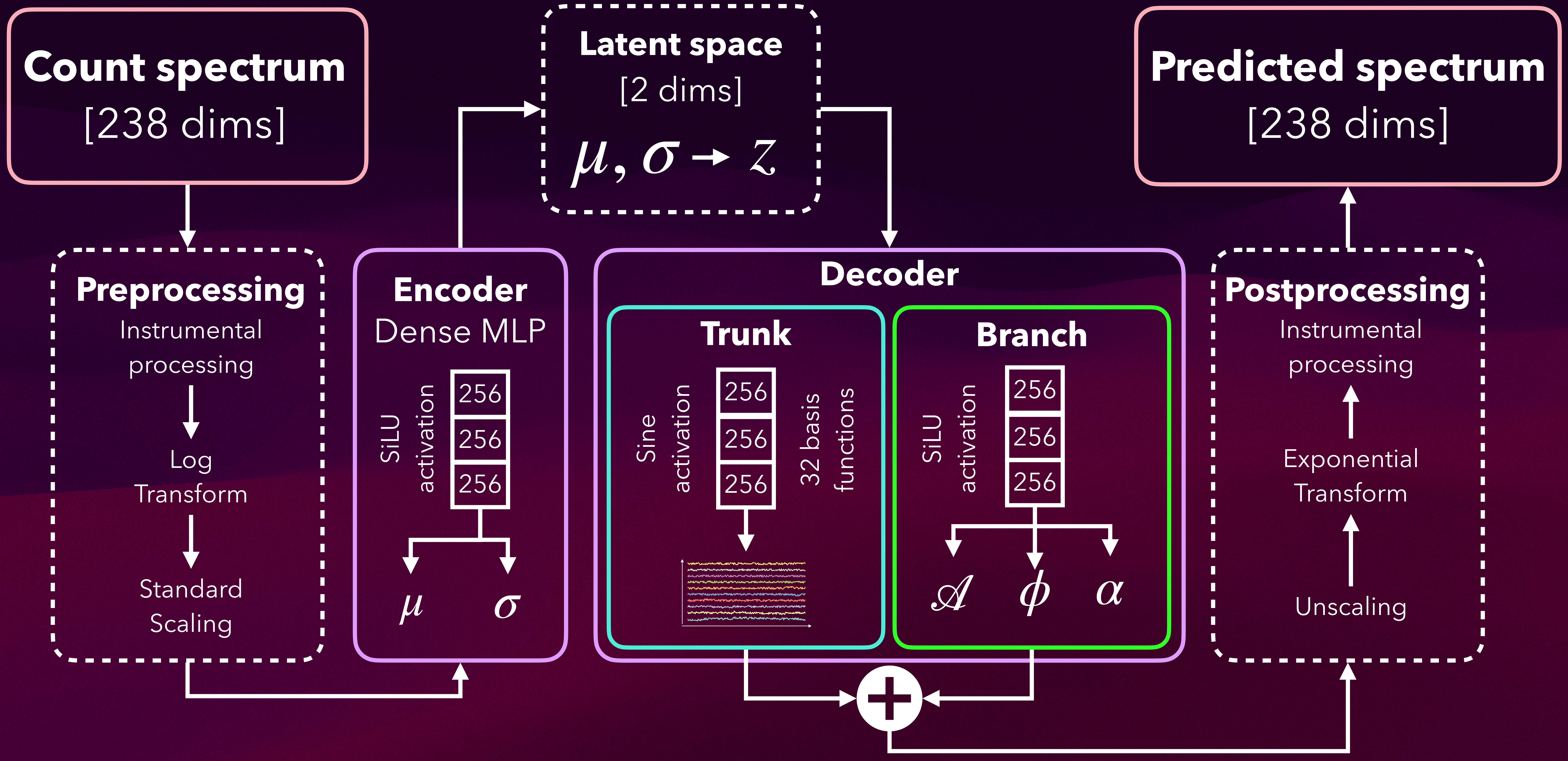

Count spectrum
[238 dims]
Preprocessing
Instrumental processing
Log Transform
Standard Scaling
Encoder
Dense MLP
SiLU activation
256
256
256
μ
σ
Latent space
[2 dims]
$\mu, \sigma \rightarrow z$


Count spectrum
[238 dims]
Latent space
[2 dims]
$\mu, \sigma \rightarrow z$
Preprocessing
Instrumental processing
Log Transform
Standard Scaling
Encoder
Dense MLP
SiLU activation
256
256
256
μ
σ
Decoder
Trunk
Sine activation
256
256
256
32 basis functions
Branch
SiLU activation
256
256
256
$\mathscr{A}$
ϕ
α



Count spectrum
[238 dims]
Latent space
[2 dims]
$\mu, \sigma \rightarrow z$
Preprocessing
Instrumental processing
Log Transform
Standard Scaling
Encoder
Dense MLP
SiLU activation
256
256
256
μ
σ
Decoder
Trunk
Sine activation
256
256
256
32 basis functions
Branch
SiLU activation
256
256
256
$\mathscr{A}$
ϕ
α
Postprocessing
Instrumental processing
Exponential Transform
Unscaling

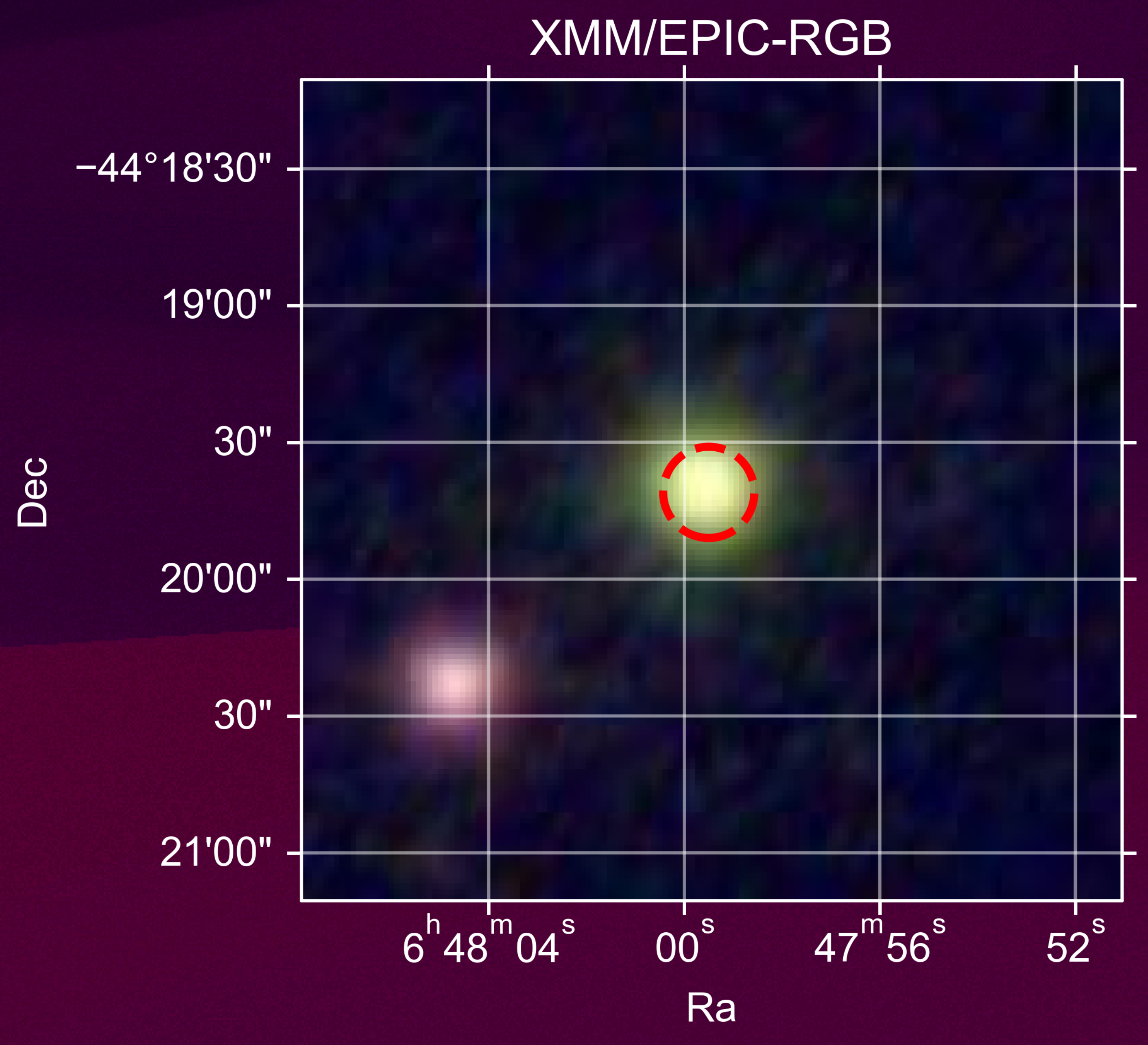

Count spectrum
[238 dims]
Latent space
[2 dims]
$\mu, \sigma \rightarrow z$
Predicted spectrum
[238 dims]
Preprocessing
Instrumental processing
Log Transform
Standard Scaling
Encoder
Dense MLP
SiLU activation
256
256
256
μ
σ
Decoder
Trunk
Sine activation
256
256
256
32 basis functions
Branch
SiLU activation
256
256
256
$\mathscr{A}$
ϕ
α
Postprocessing
Instrumental processing
Exponential Transform
Unscaling

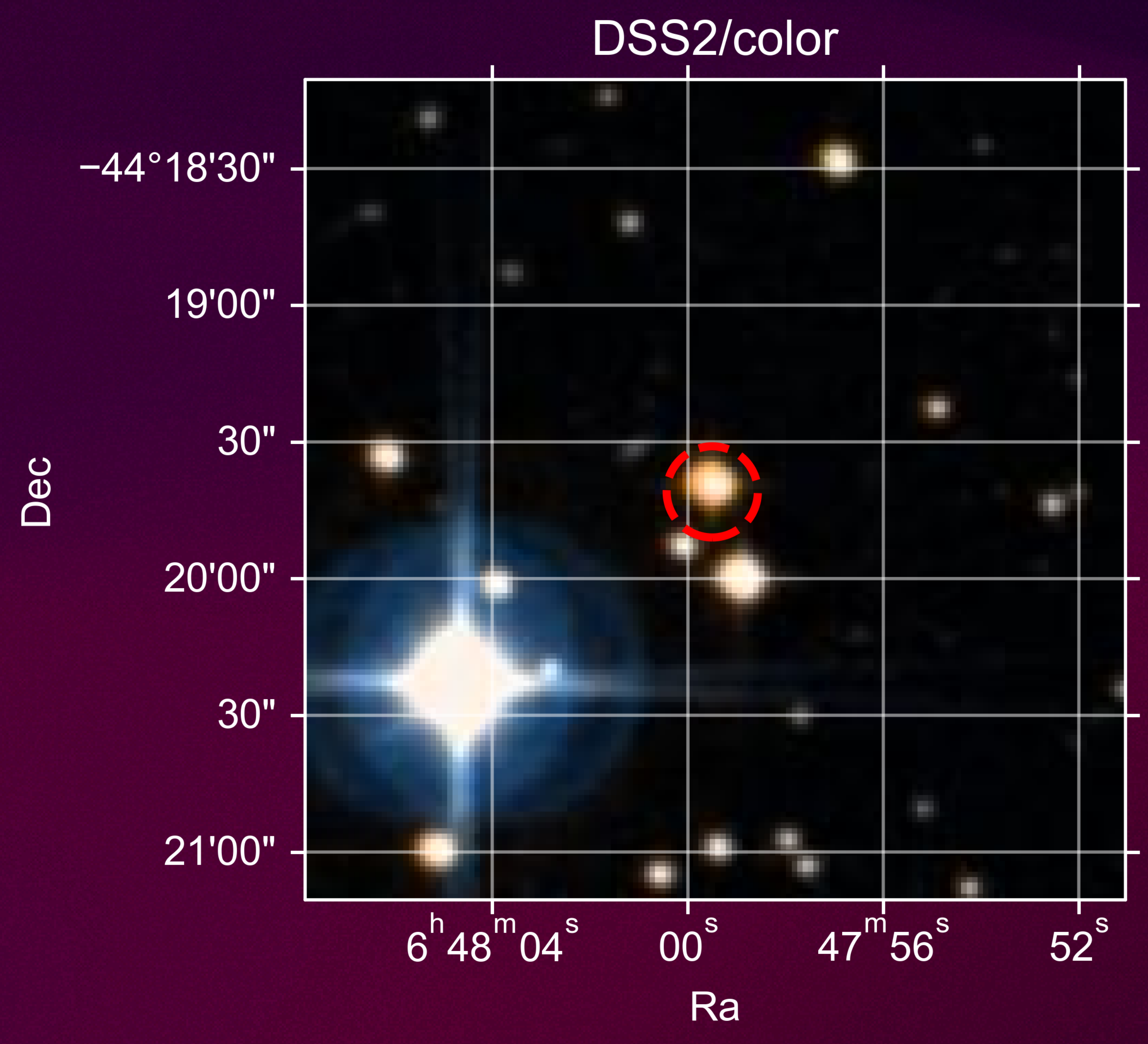


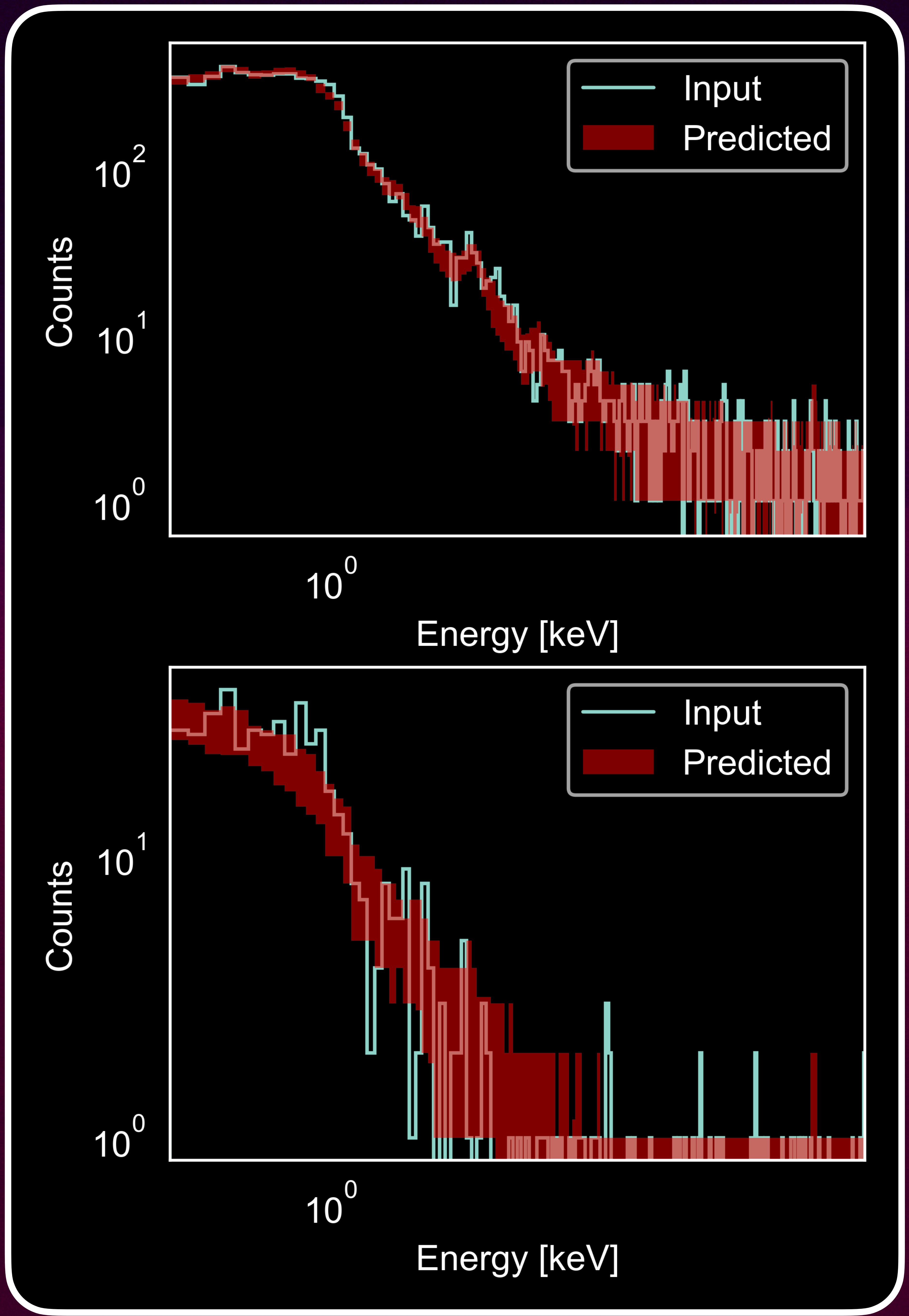
475 206 parameters (1.9 MB) trained with AdamW (lr $= 10^{-4}$) for 25000 epochs

And now, some cherry picked results!

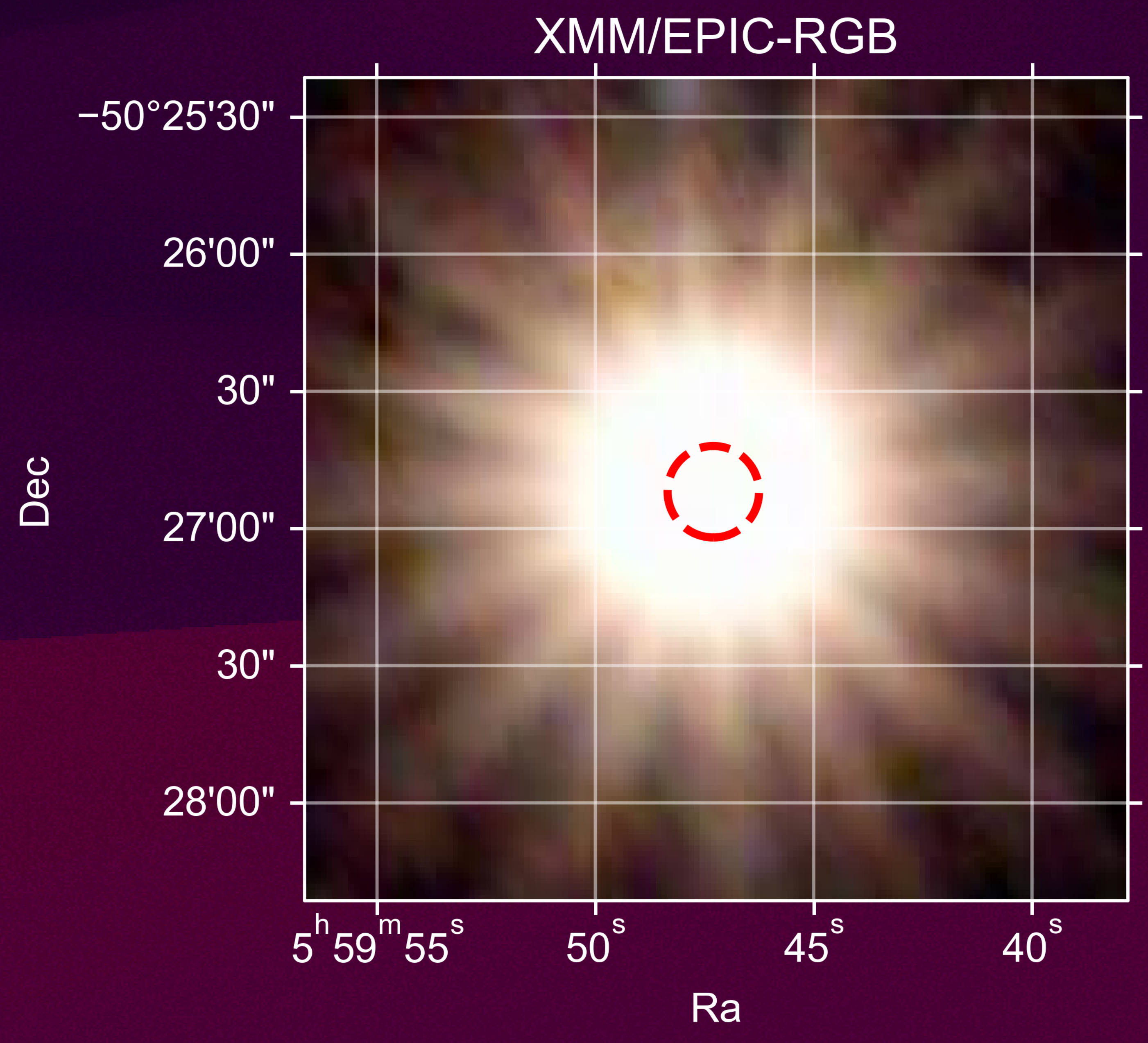
4XMM J064759.5-441941 (Star)

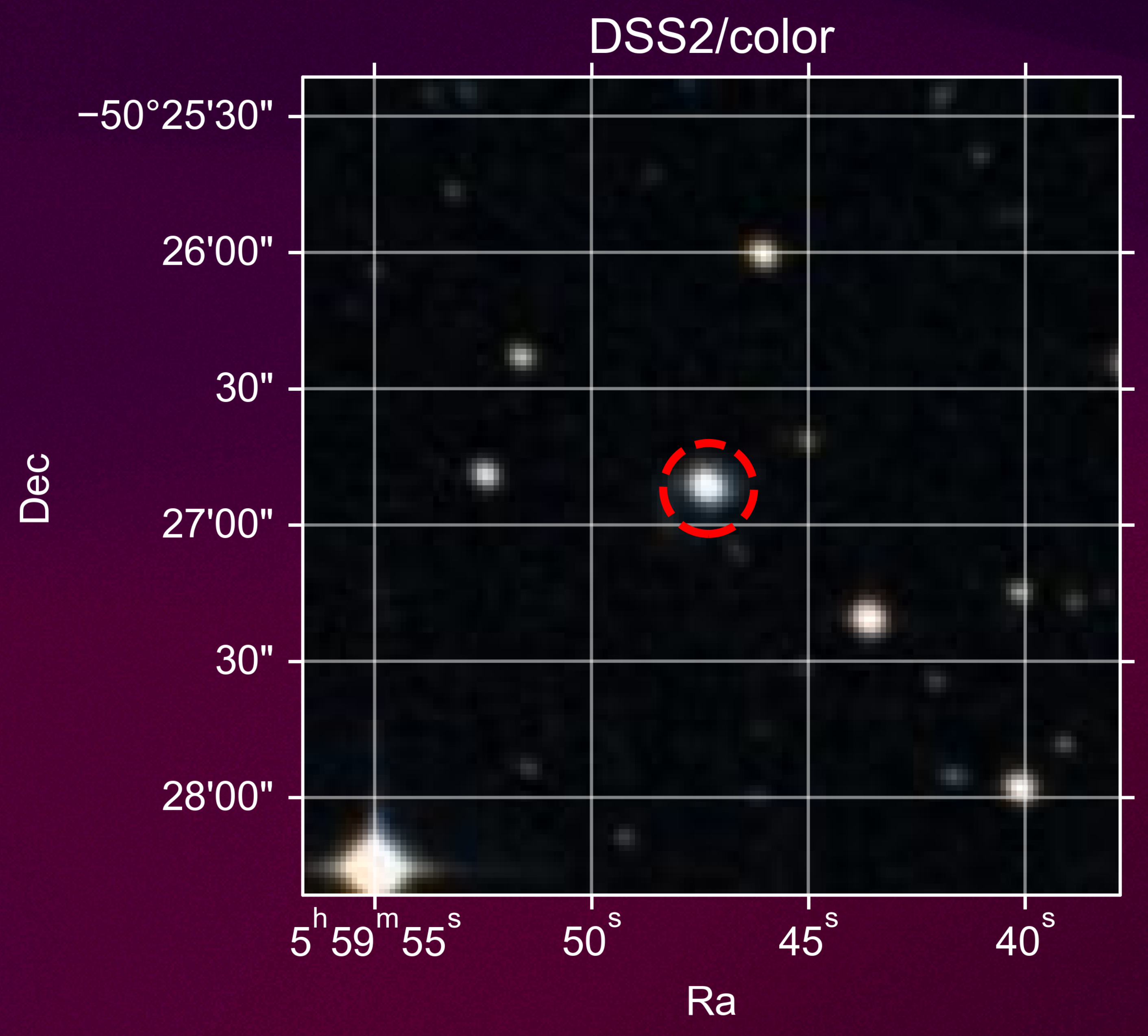


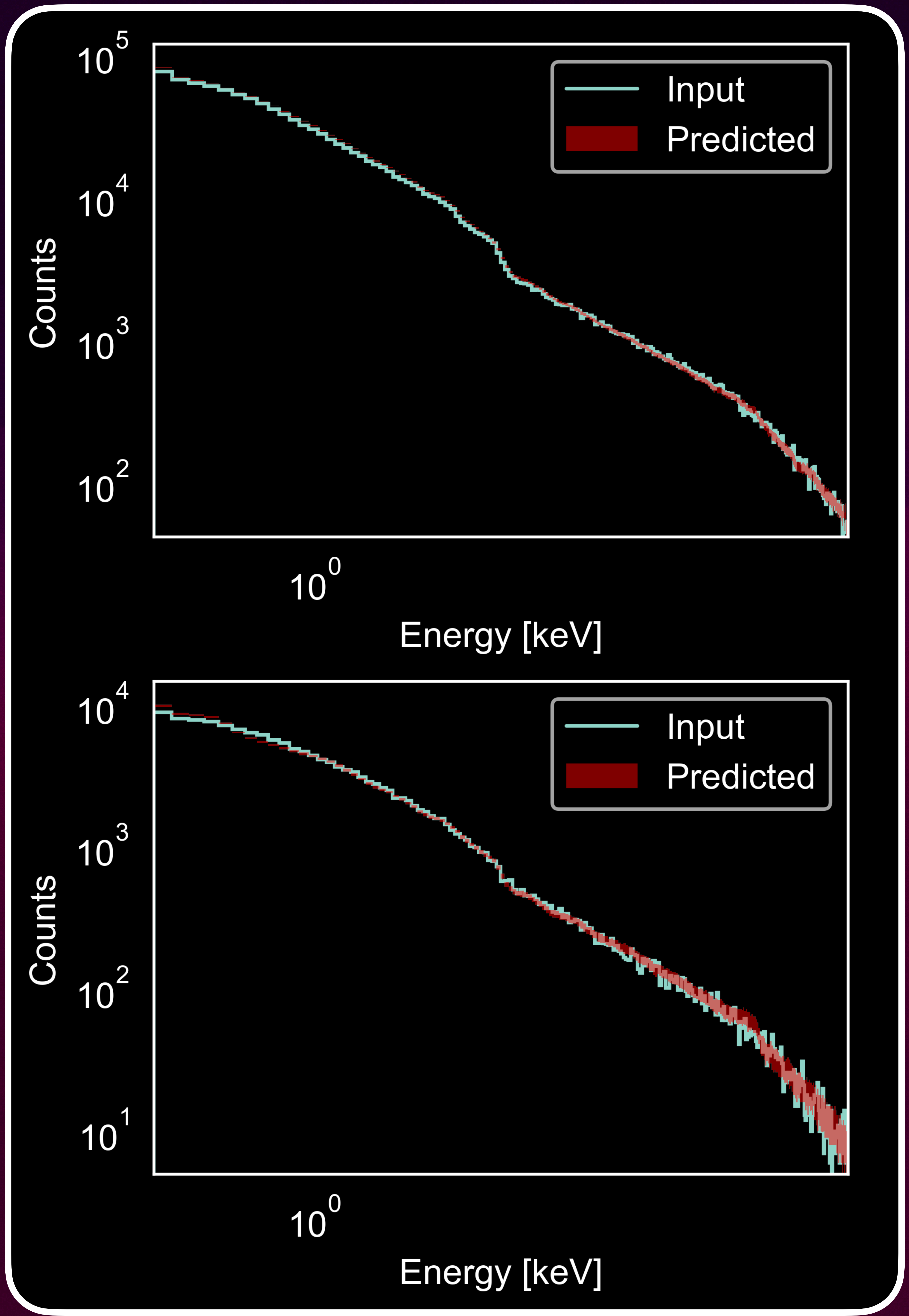




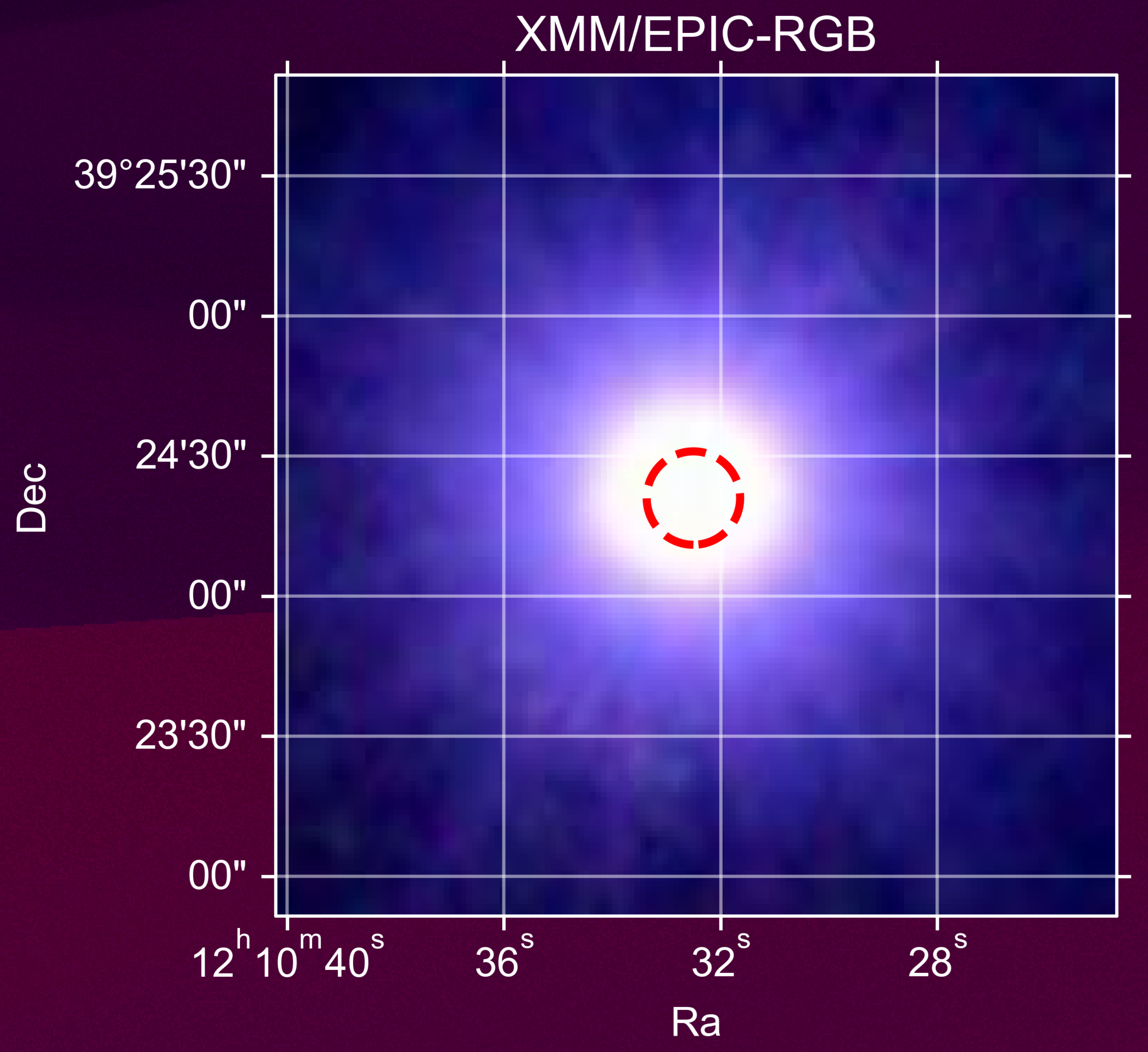
4XMM J055947.3-502652 (AGN)

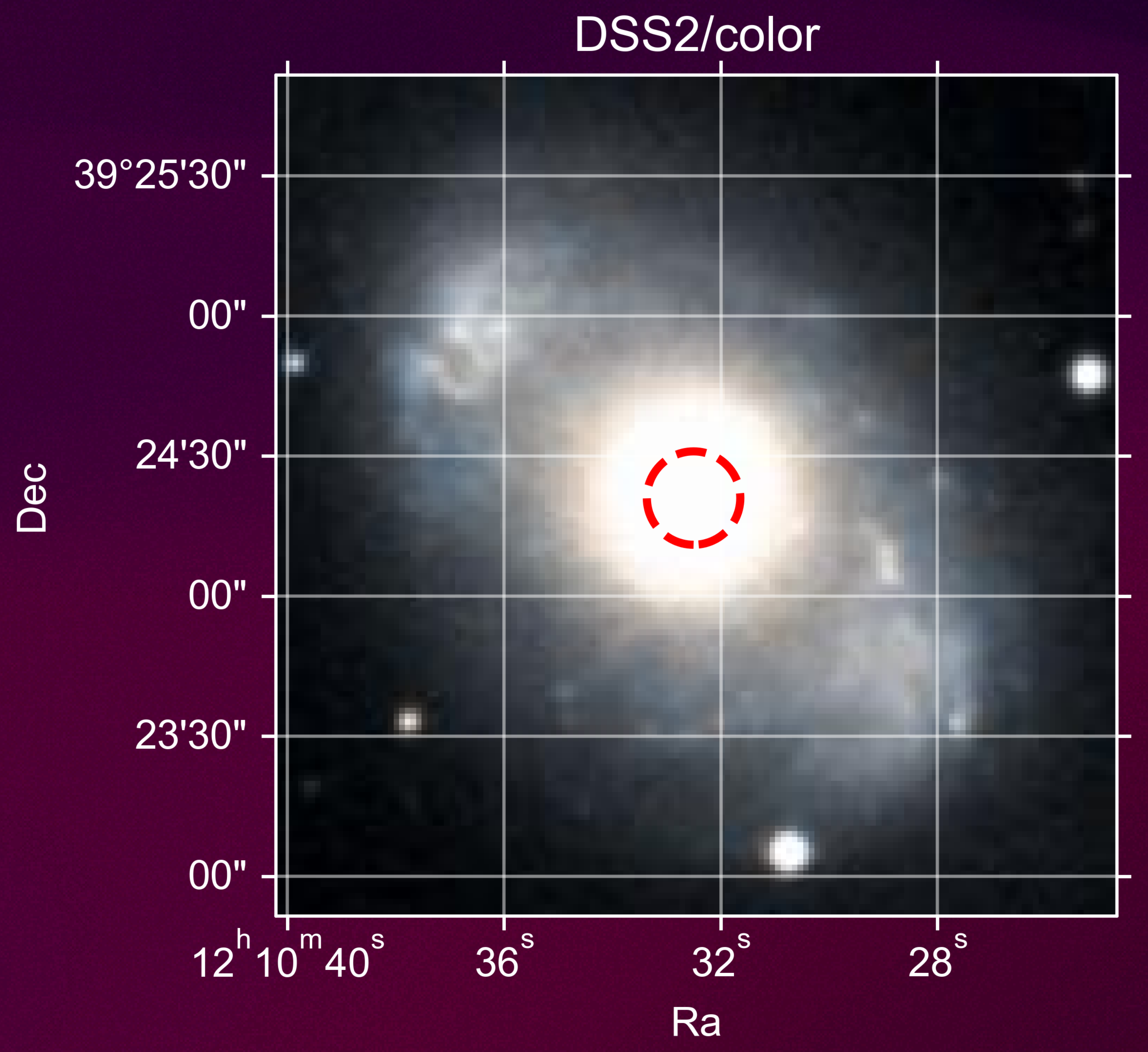


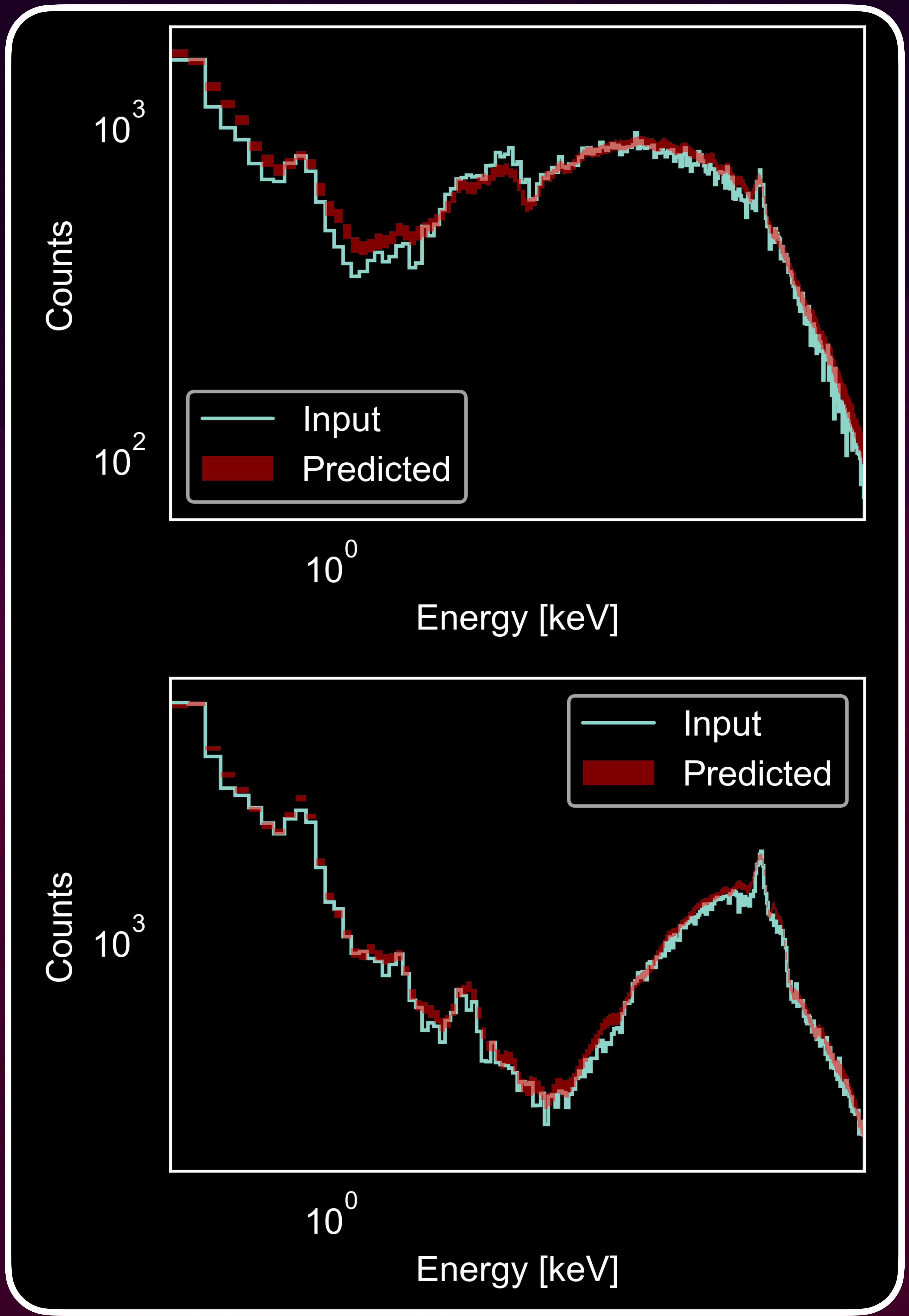




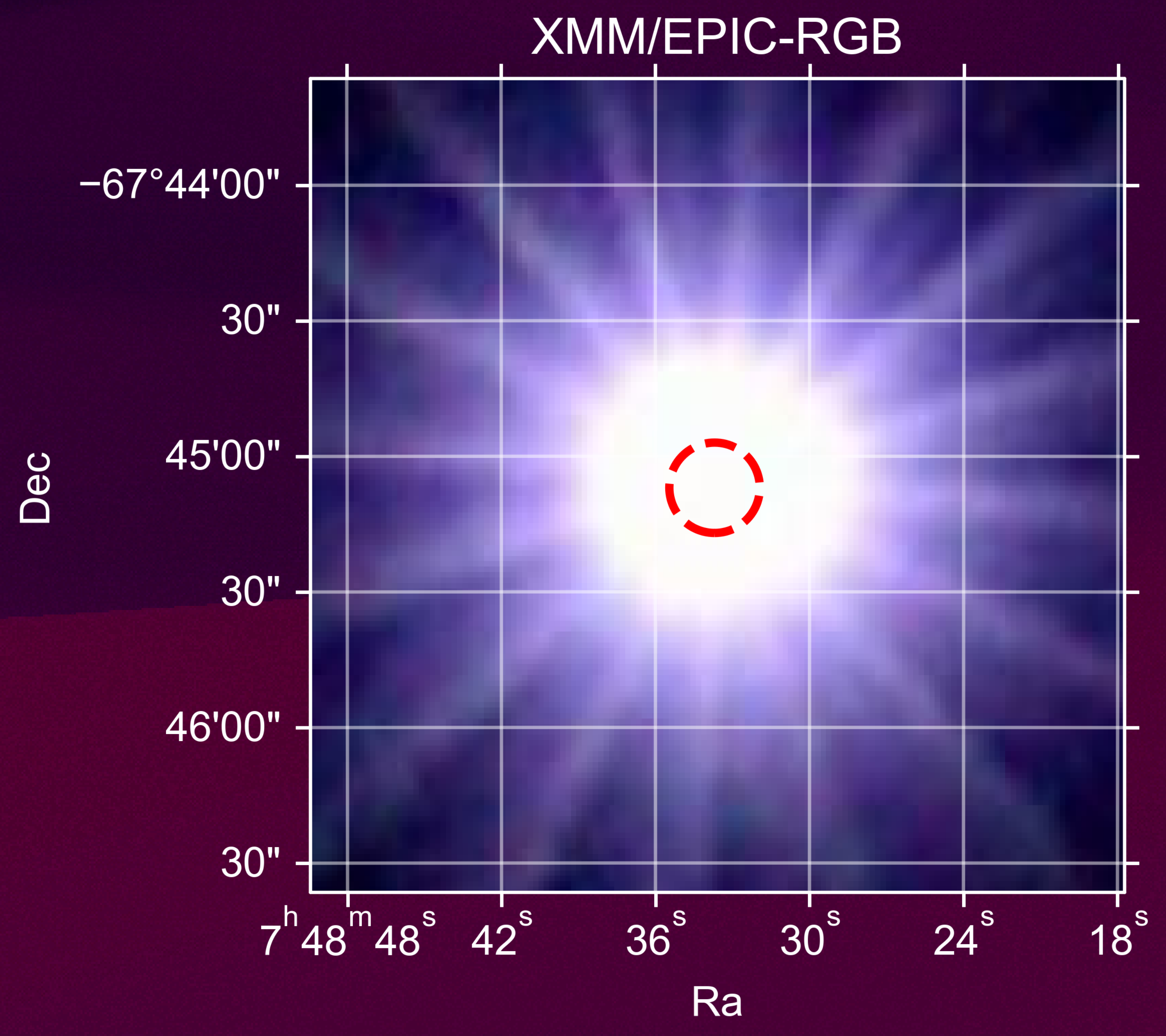
4XMM J121032.5+392421 (AGN)

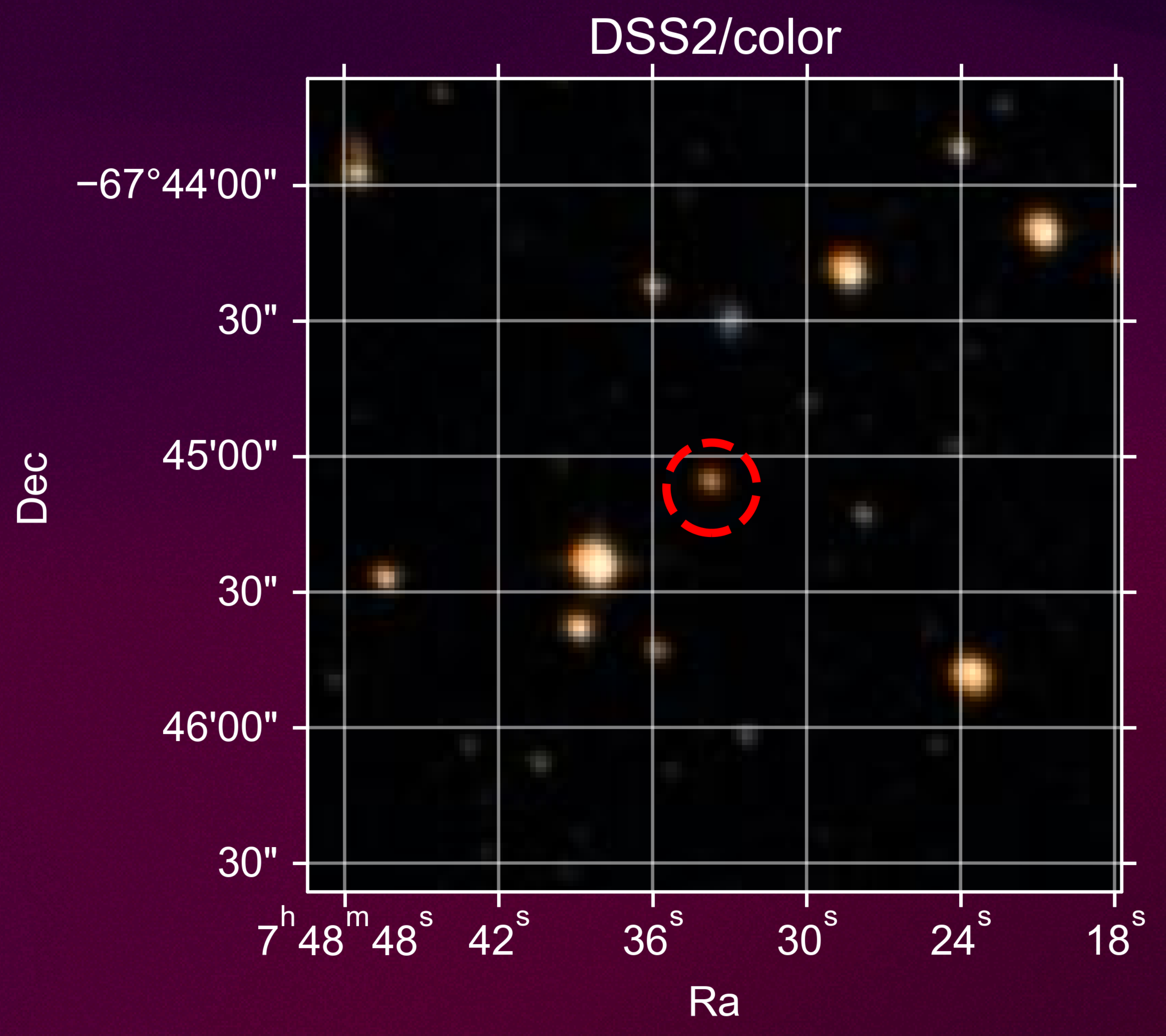


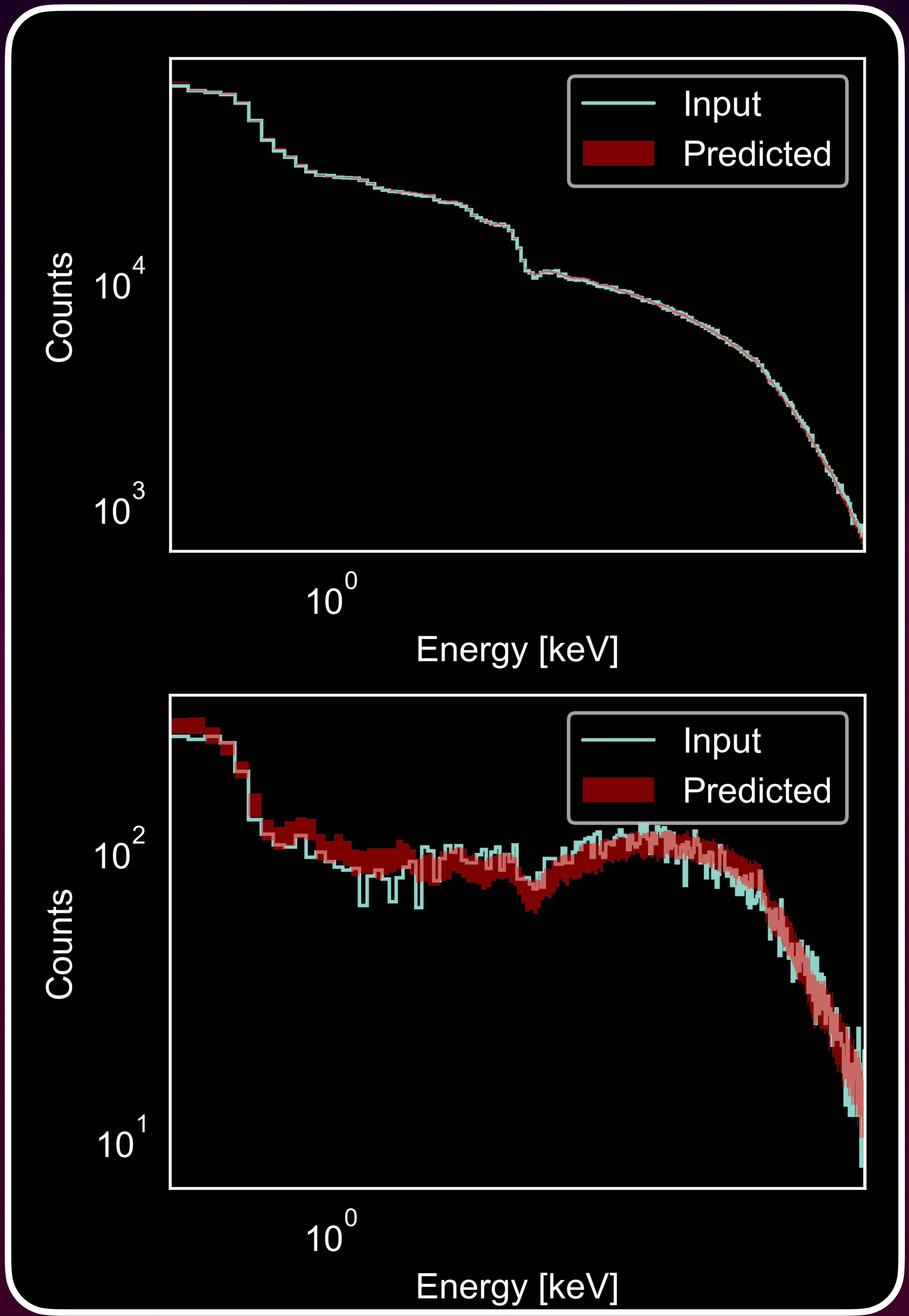




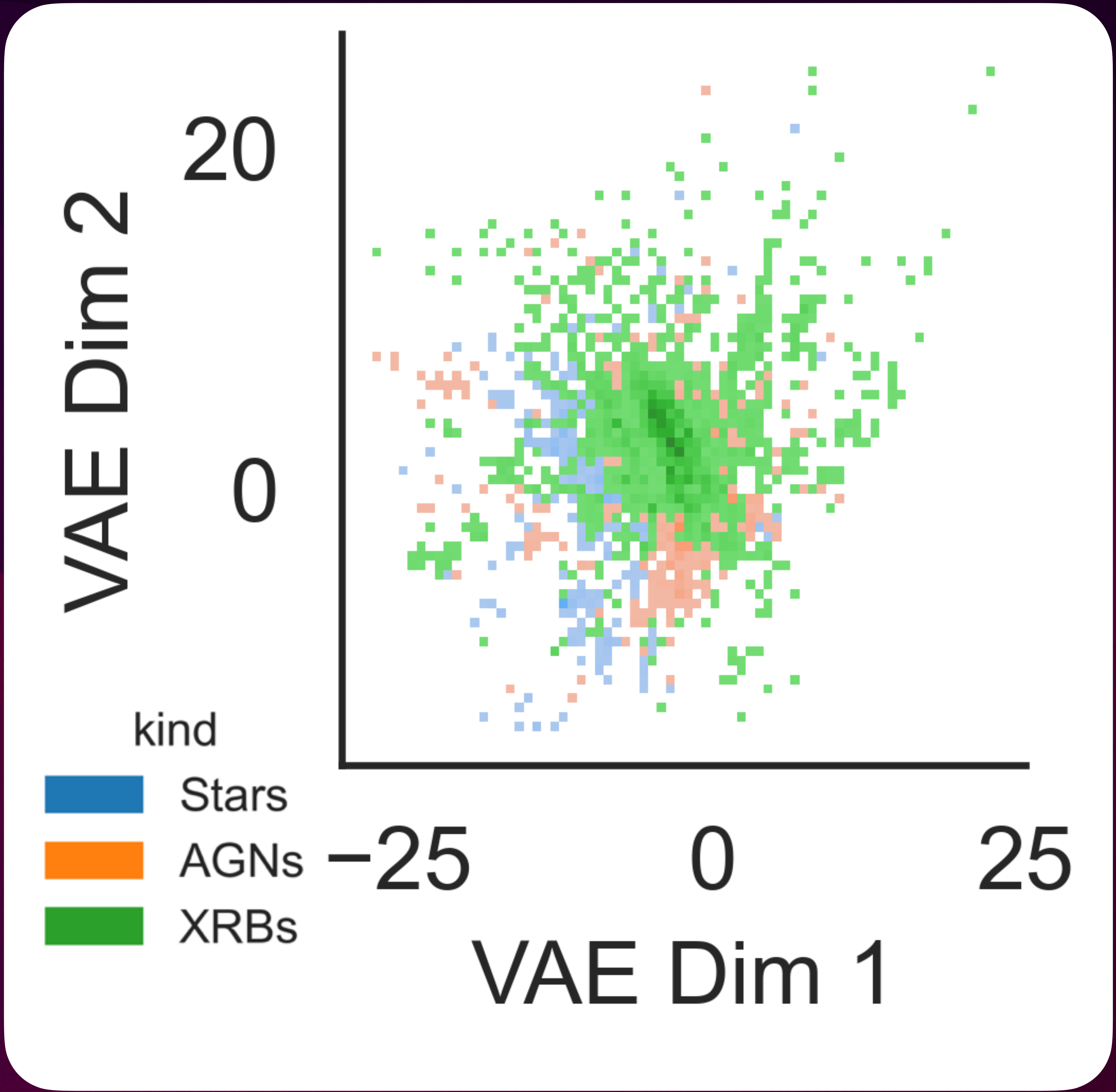
4XMM 074833.7-674507 (XRB)





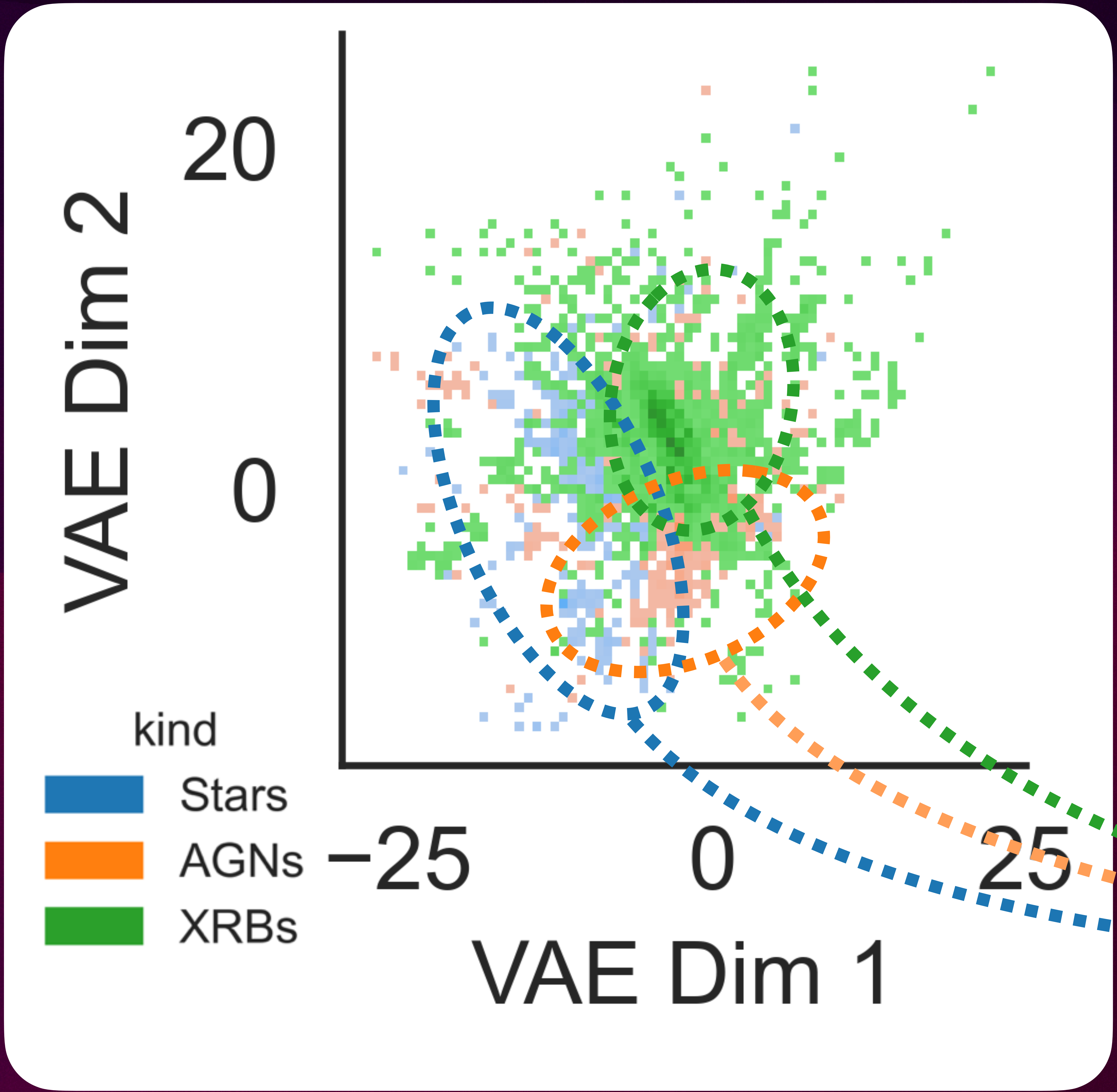


What's in the latent space ?



- Encode a bunch of sources with more than 5k counts
- Compare it with a crude classification (SIMBAD match)
- Corner plot of μ for all the sources

What's in the latent space ?

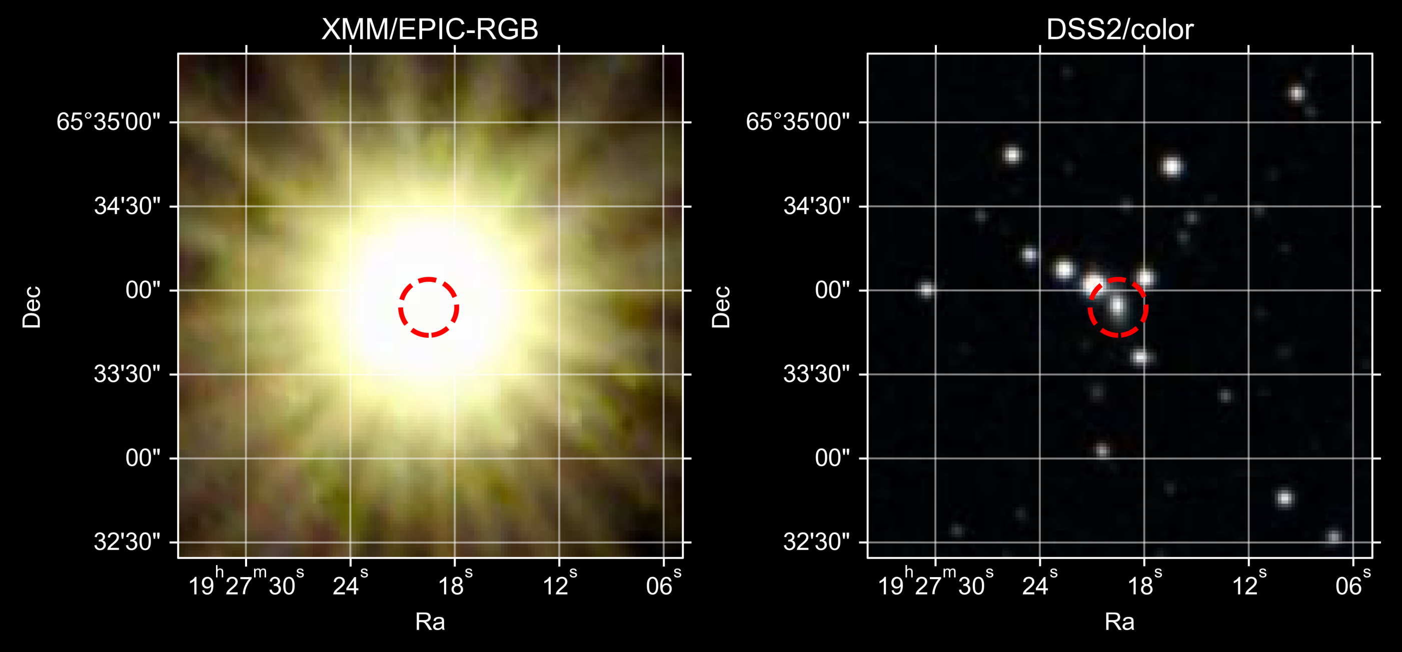


- Encode a bunch of sources with more than 5k counts
- Compare it with a crude classification (SIMBAD match)
- Corner plot of μ for all the sources

Do we see **clusters** in the latent space ?

Looking for similar sources

4XMM J192719.5+653354

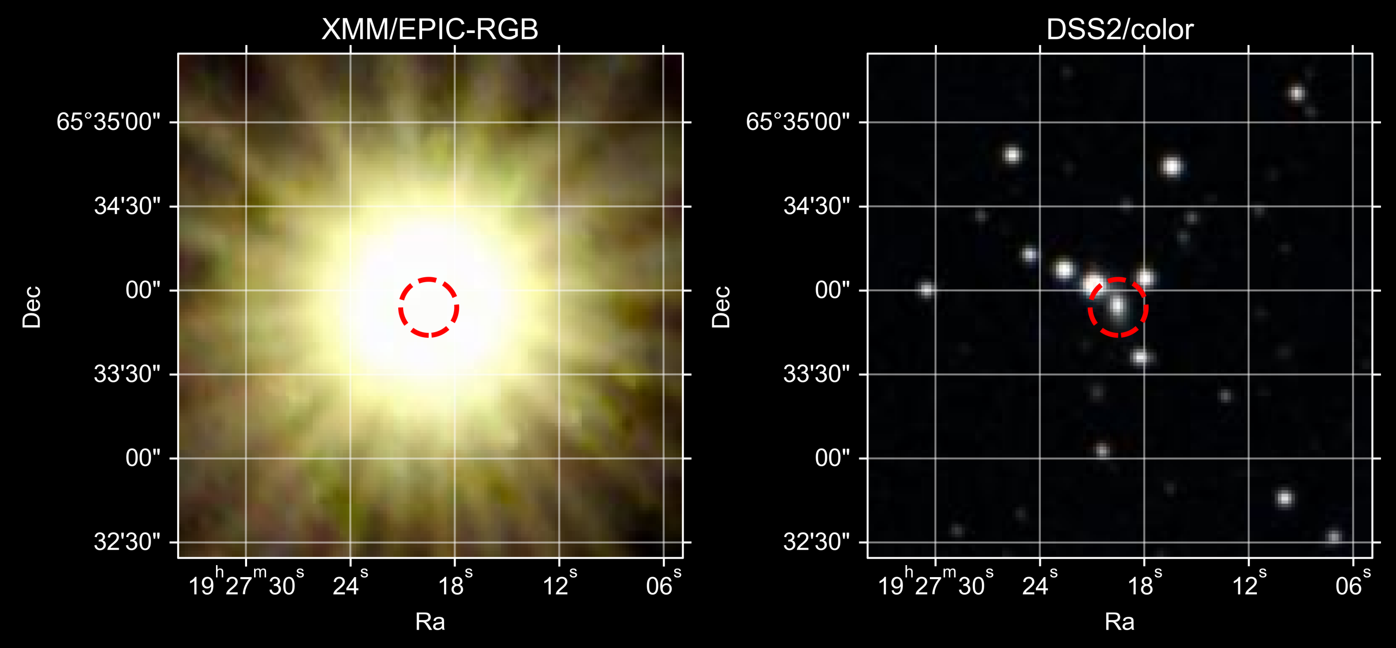


Super weird AGN with spectral variability

At some point, accretion was quenched by a TDE event

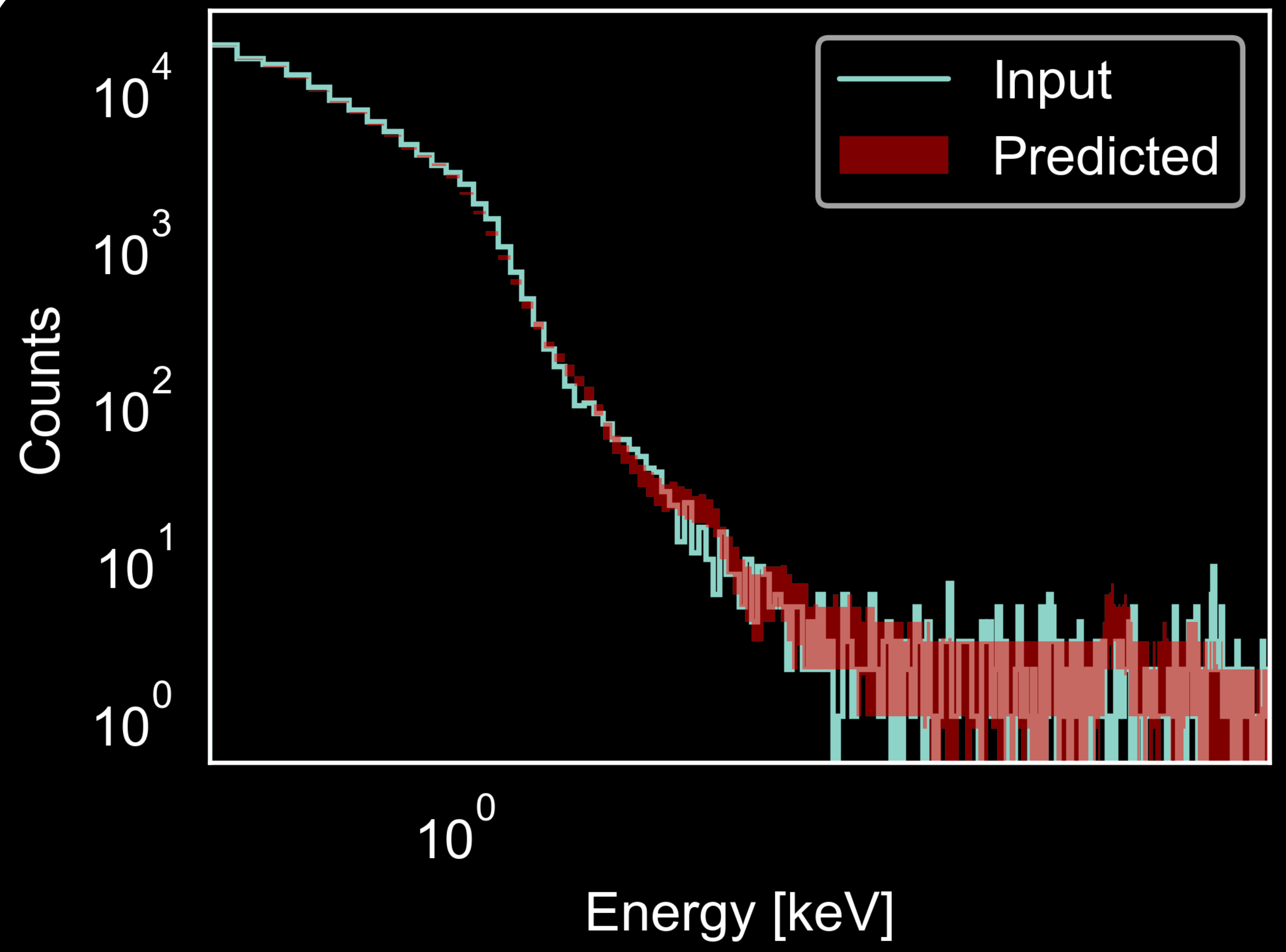
Looking for similar sources

4XMM J192719.5+653354



Super weird AGN with spectral variability

At some point, accretion was quenched by a TDE event

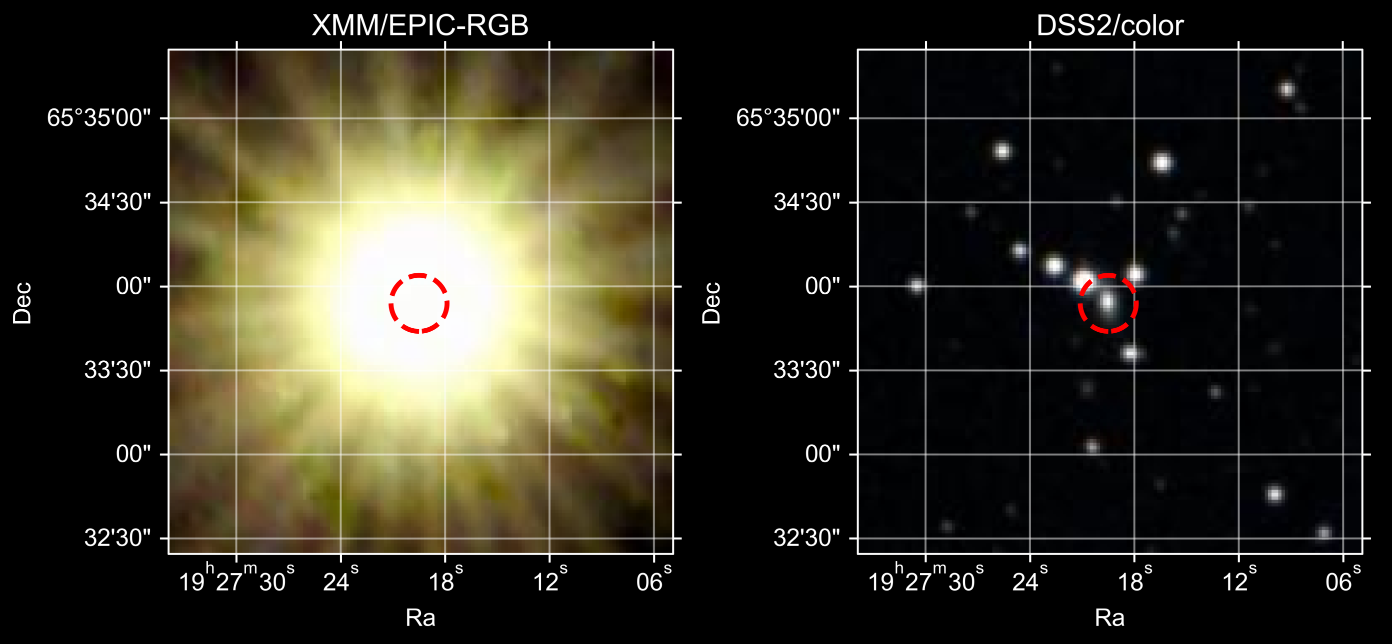


Spectra similar to this one ?

Search for sources close in the latent representation

Looking for similar sources

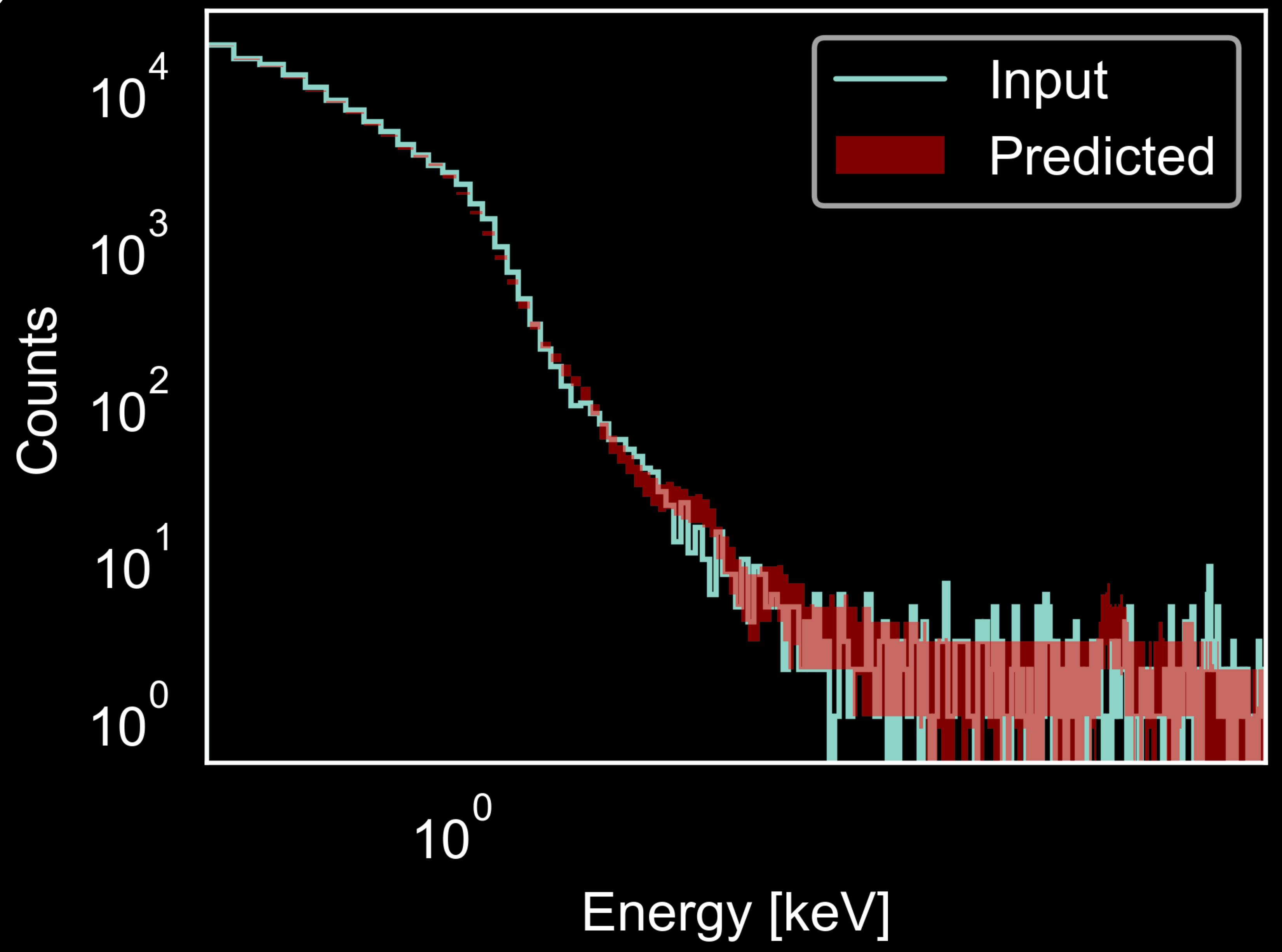
4XMM J192719.5+653354



Super weird AGN with spectral variability

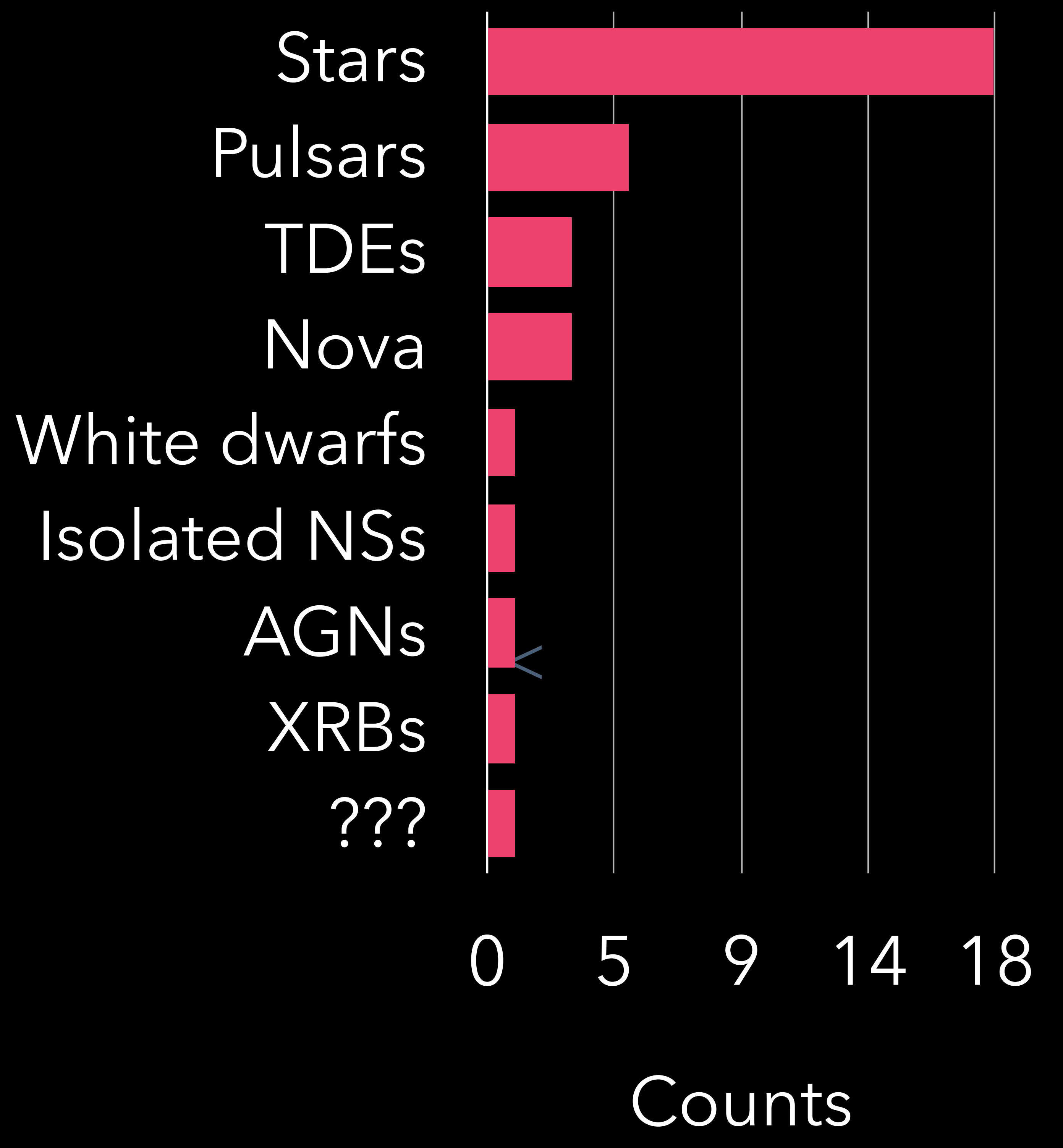
At some point, accretion was quenched by a TDE event

Minimize $\mathbb{E}\left(\|z - z_{ref}\|\right)$



Spectra similar to this one ?

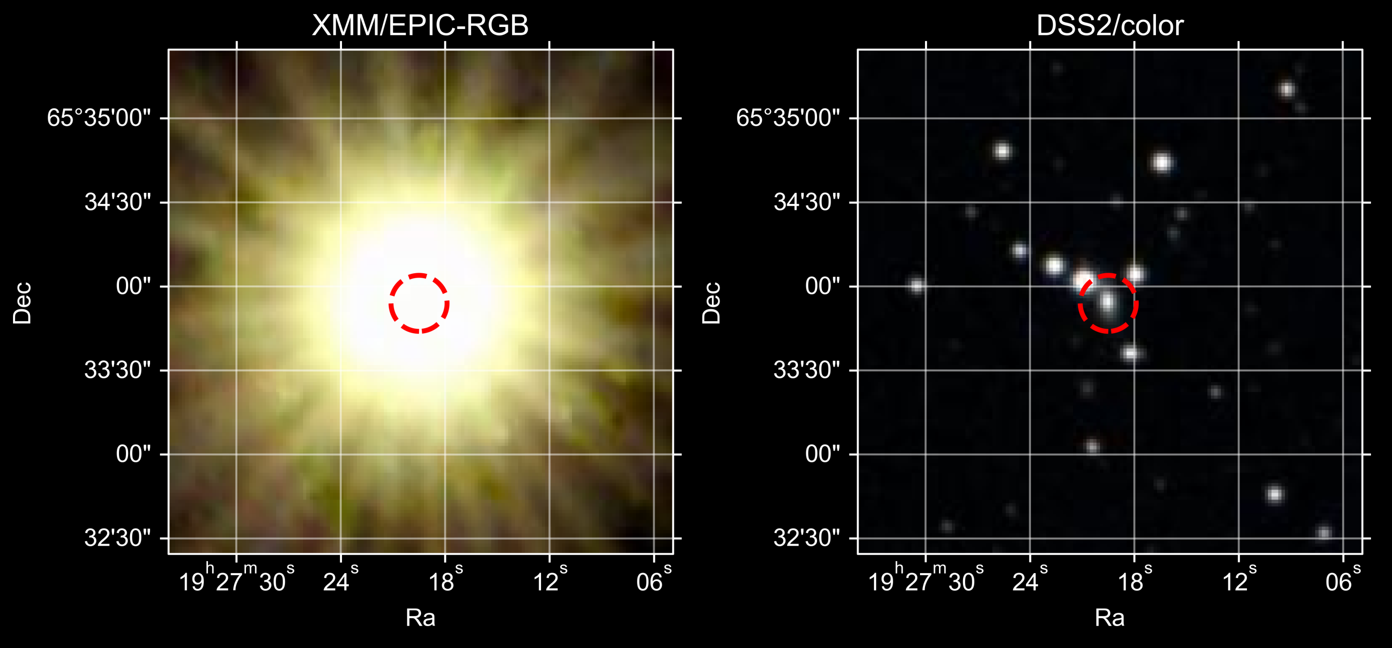
Search for sources close in the latent representation



~50% of interesting sources in the 34 closest matches

Looking for similar sources

4XMM J192719.5+653354

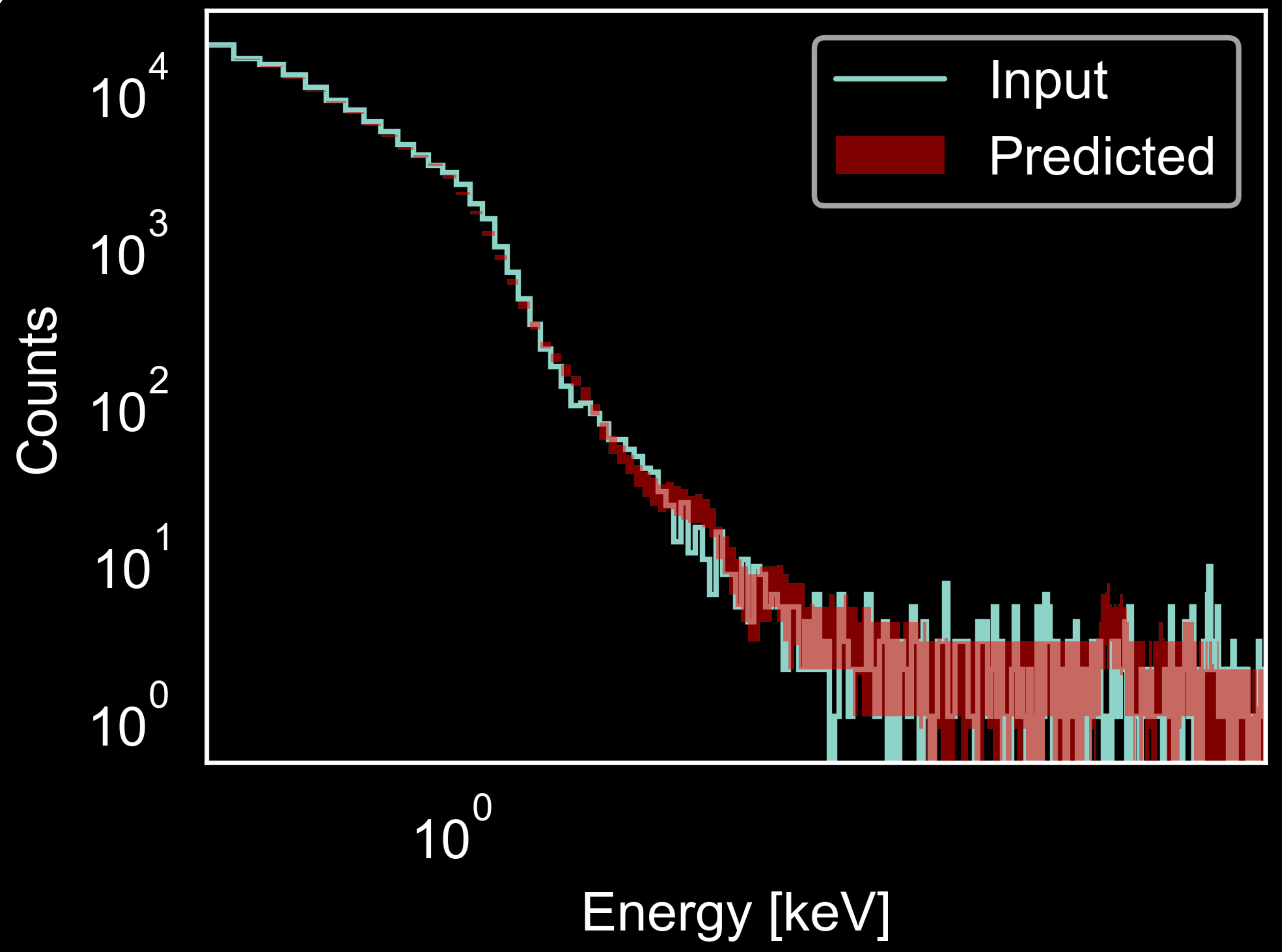


Super weird AGN with spectral variability

At some point, accretion was quenched by a TDE event

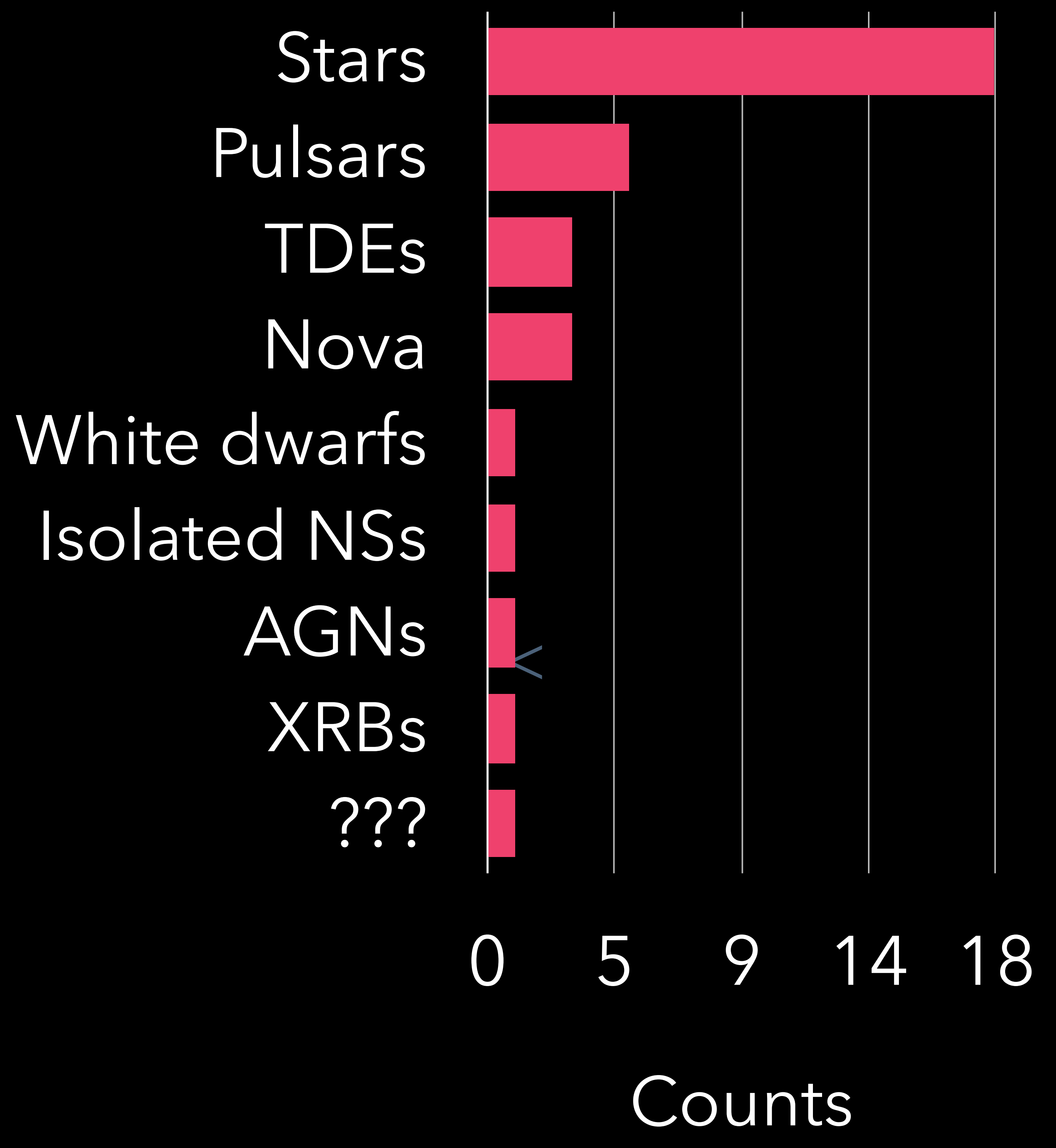
Same can be done with any source in the catalogue!

Minimize $\mathbb{E}\left(\|z - z_{ref}\| \right)$



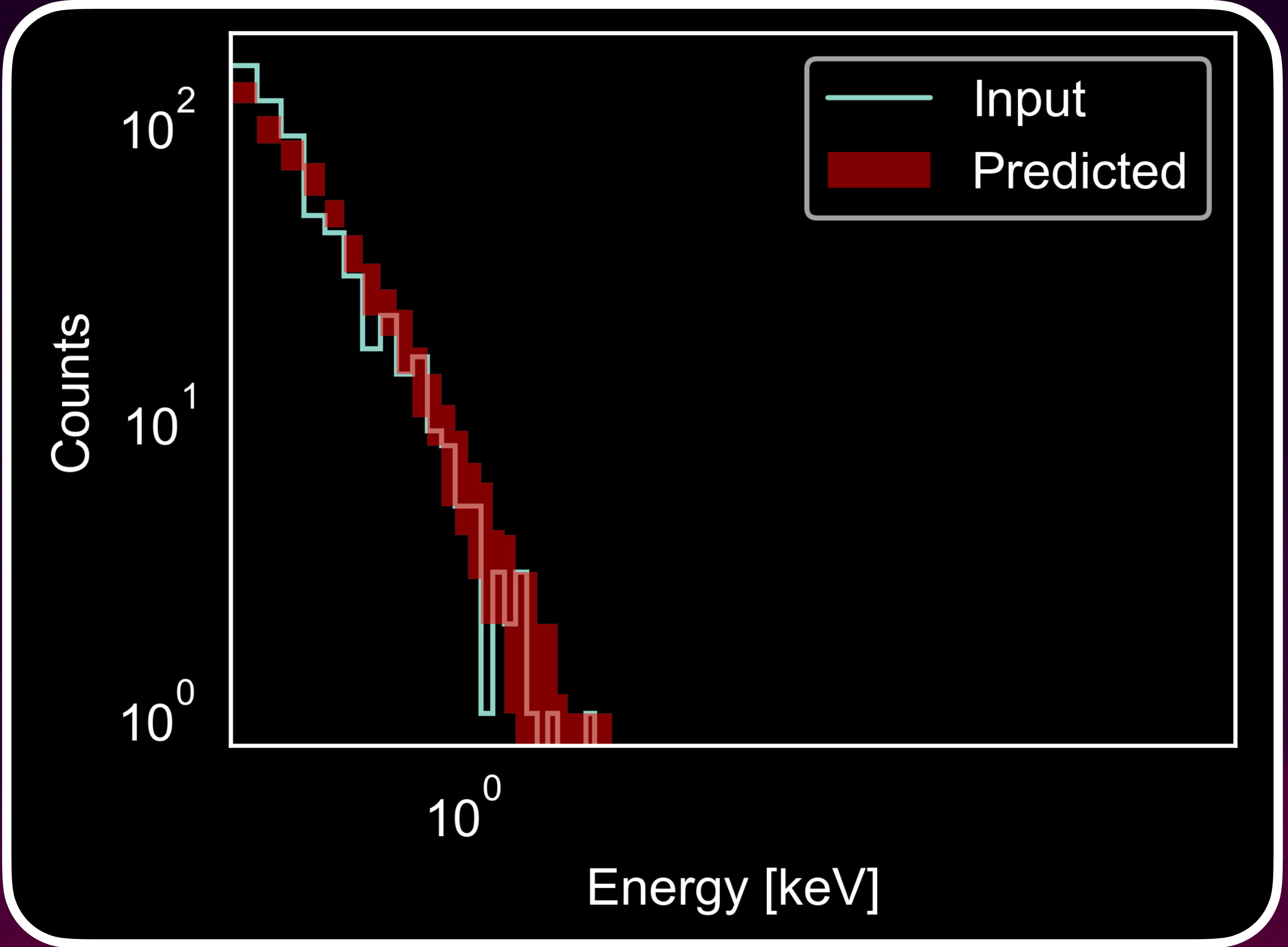
Spectra similar to this one ?

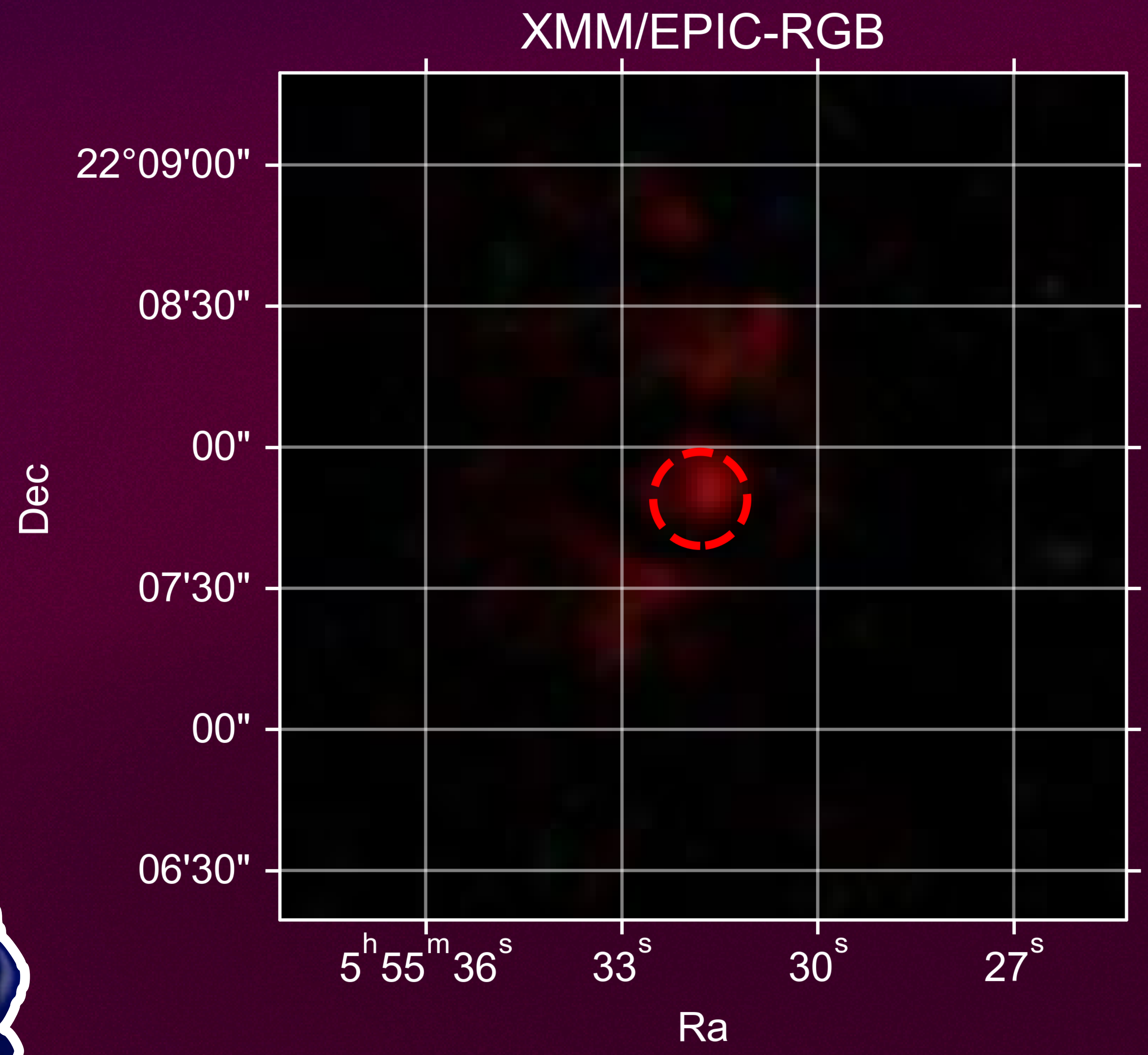
Search for sources close in the latent representation

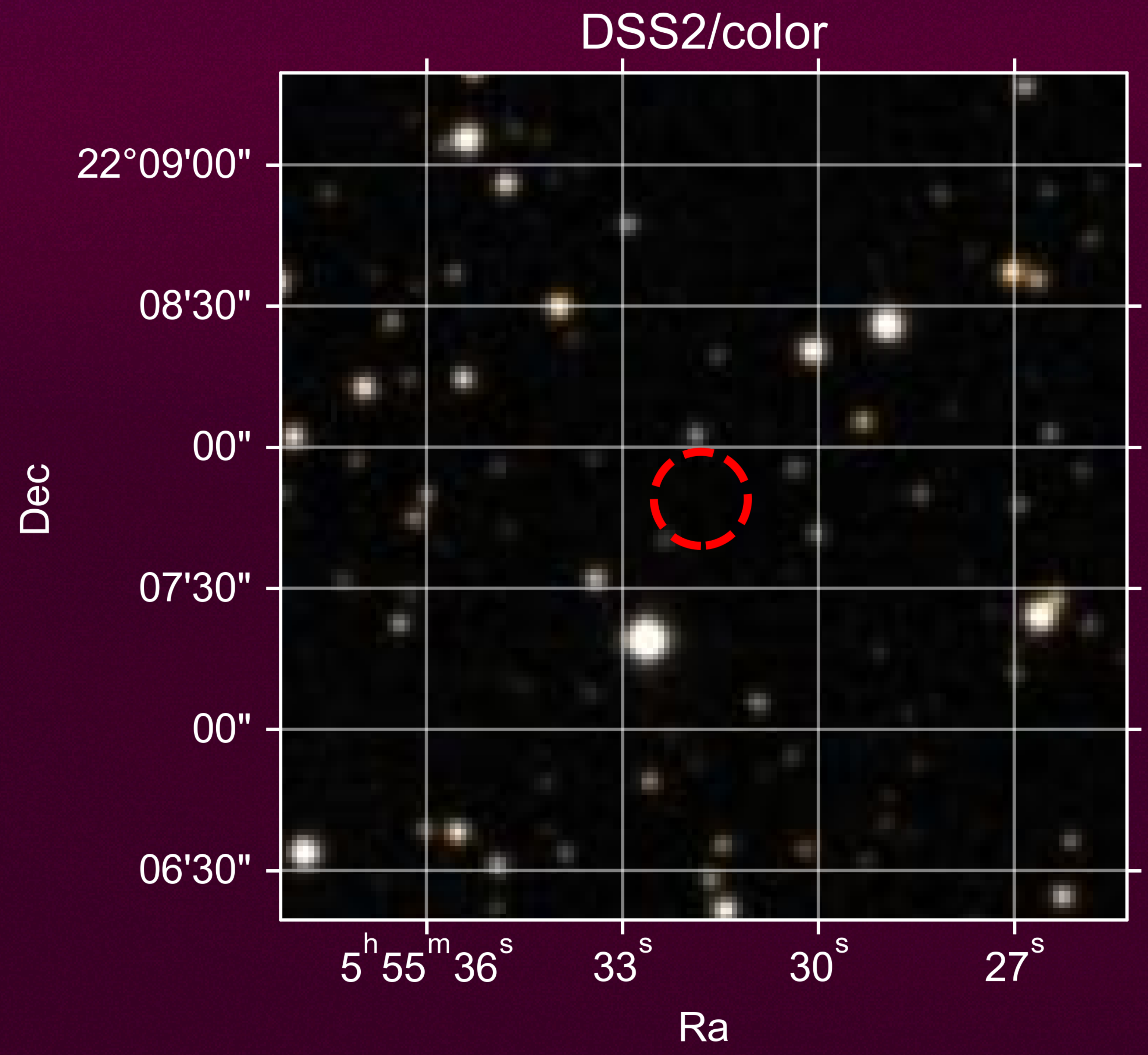


~50% of interesting sources in the 34 closest matches

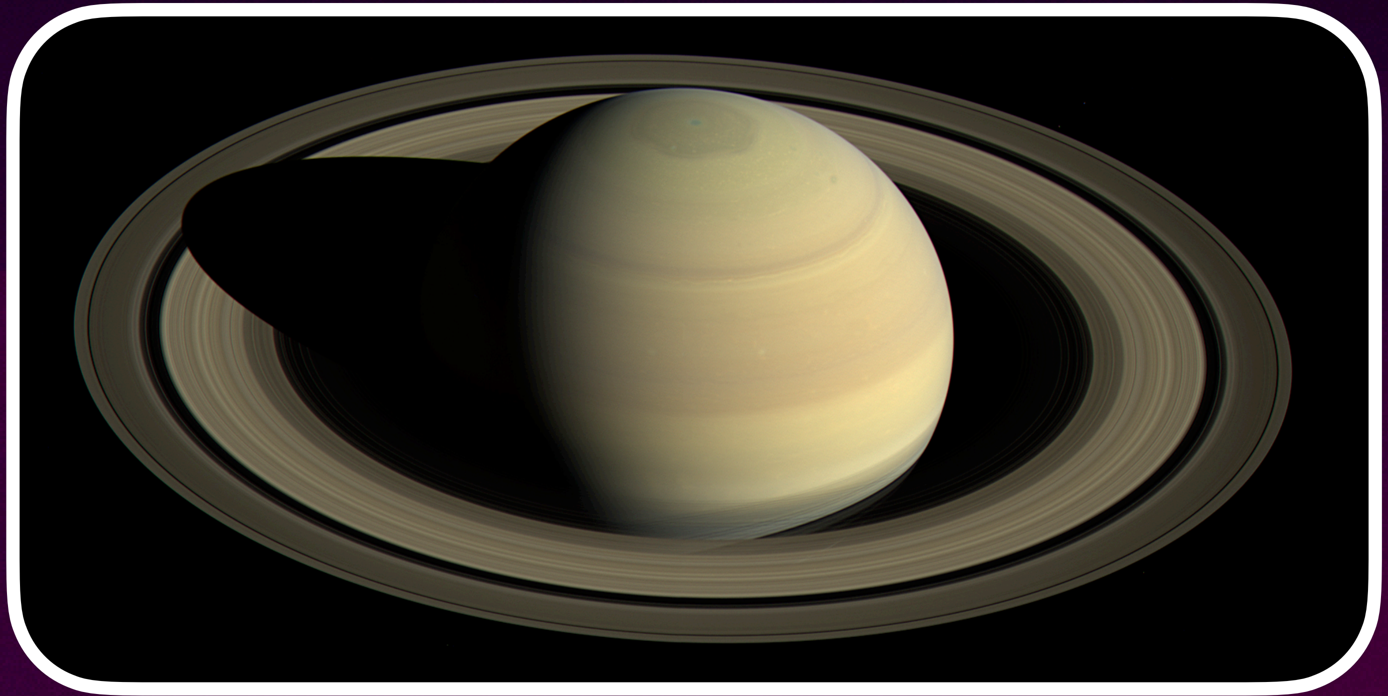
Mystery source ?

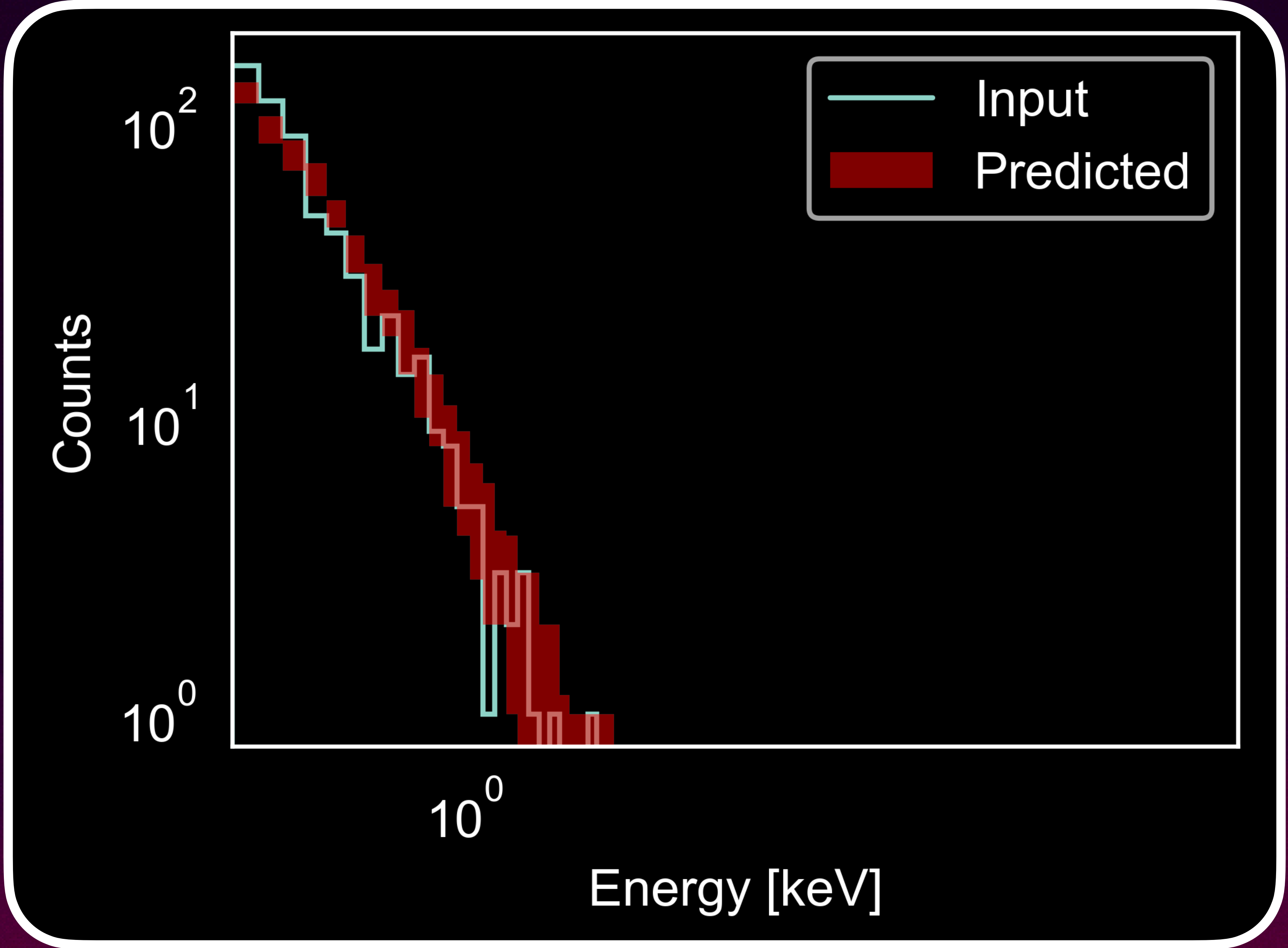


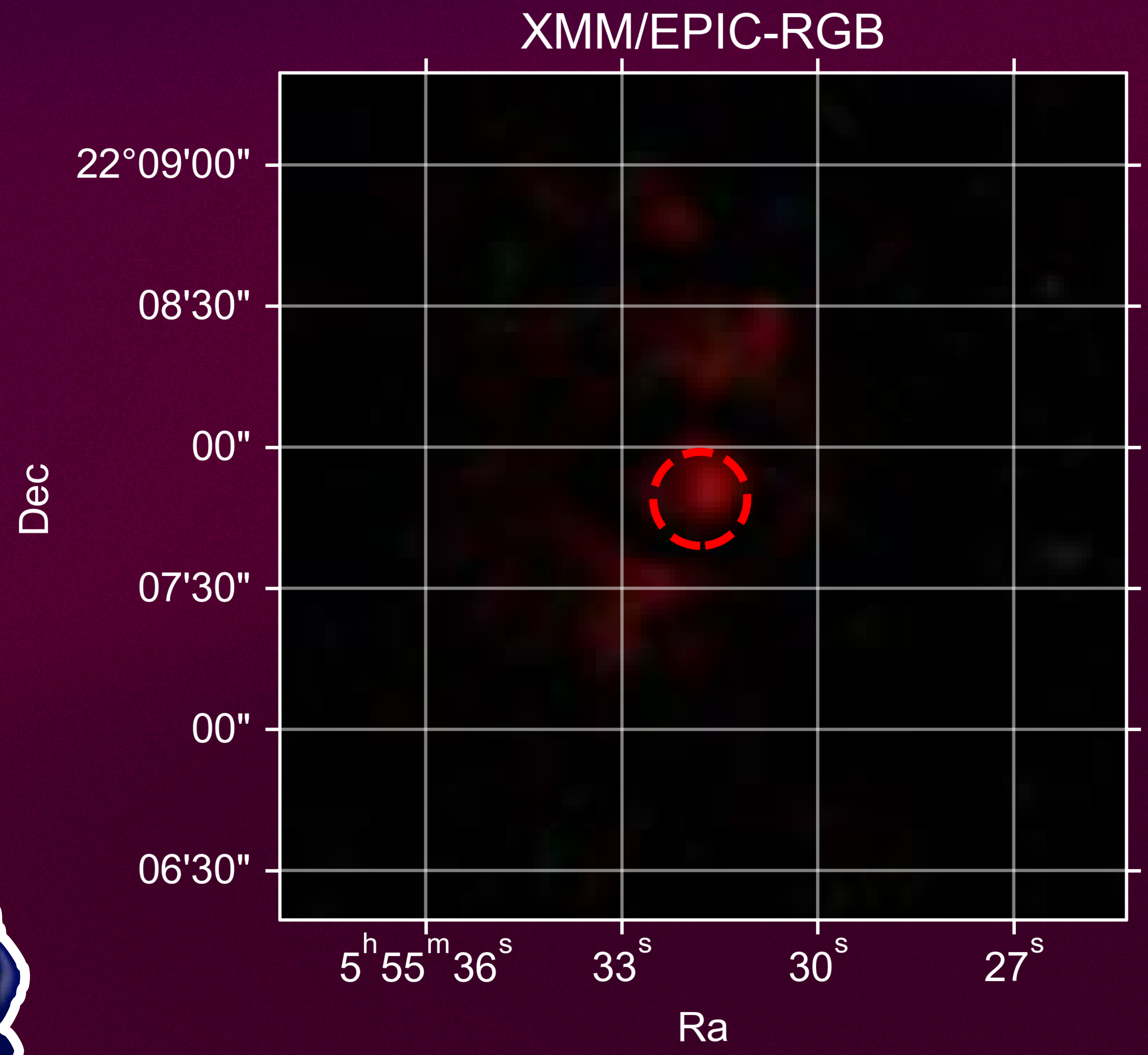


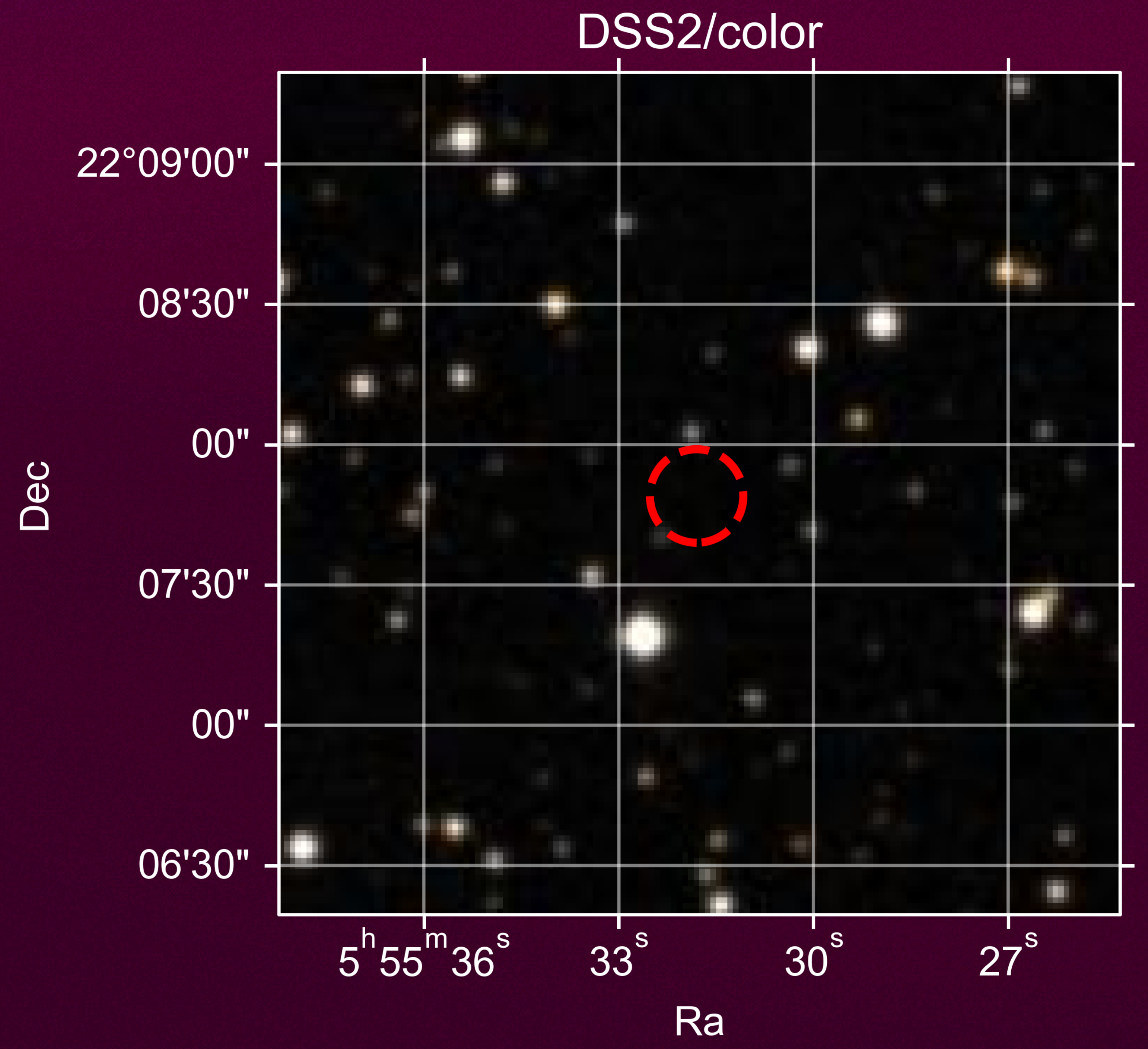


Mystery source ?

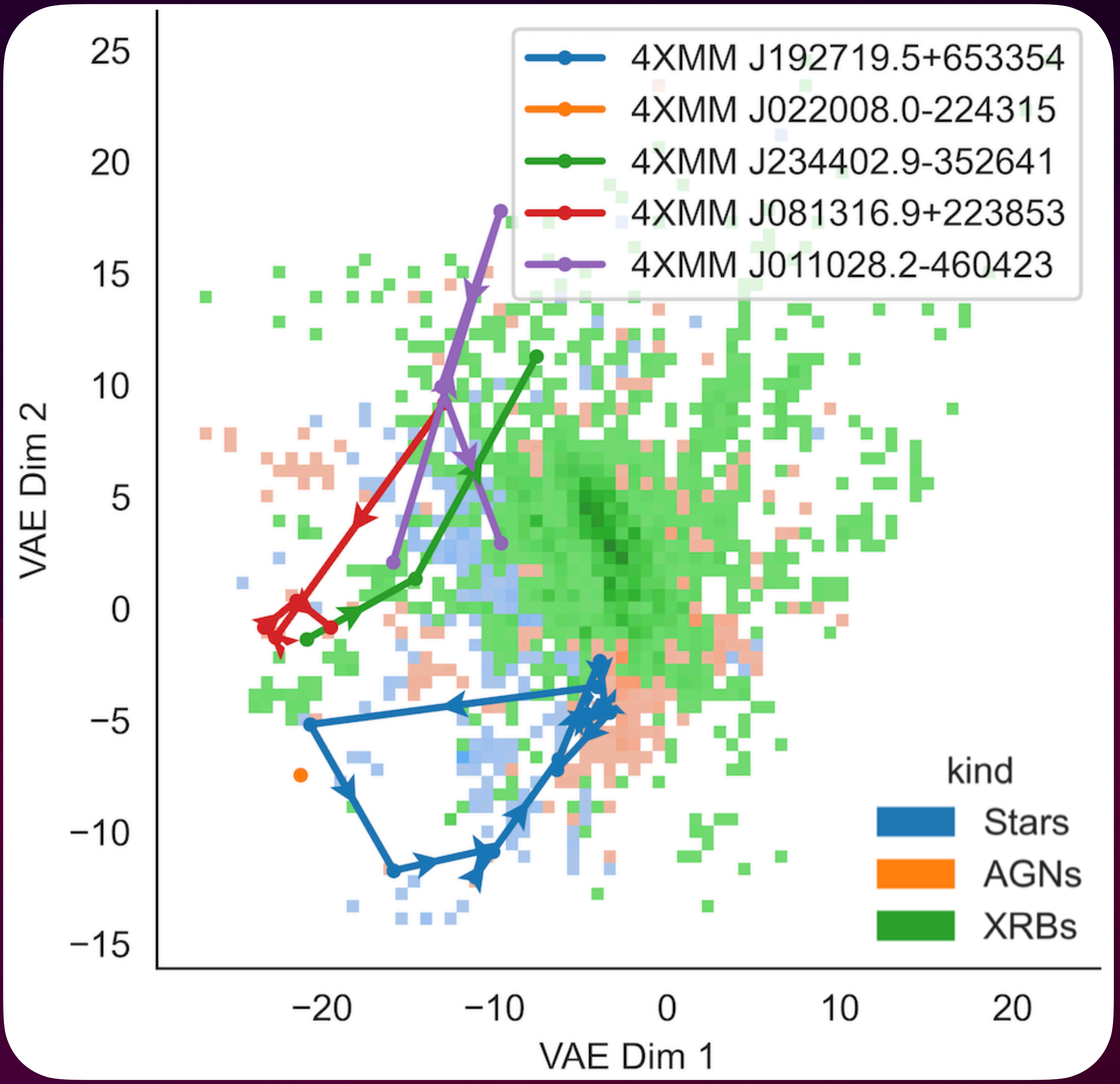






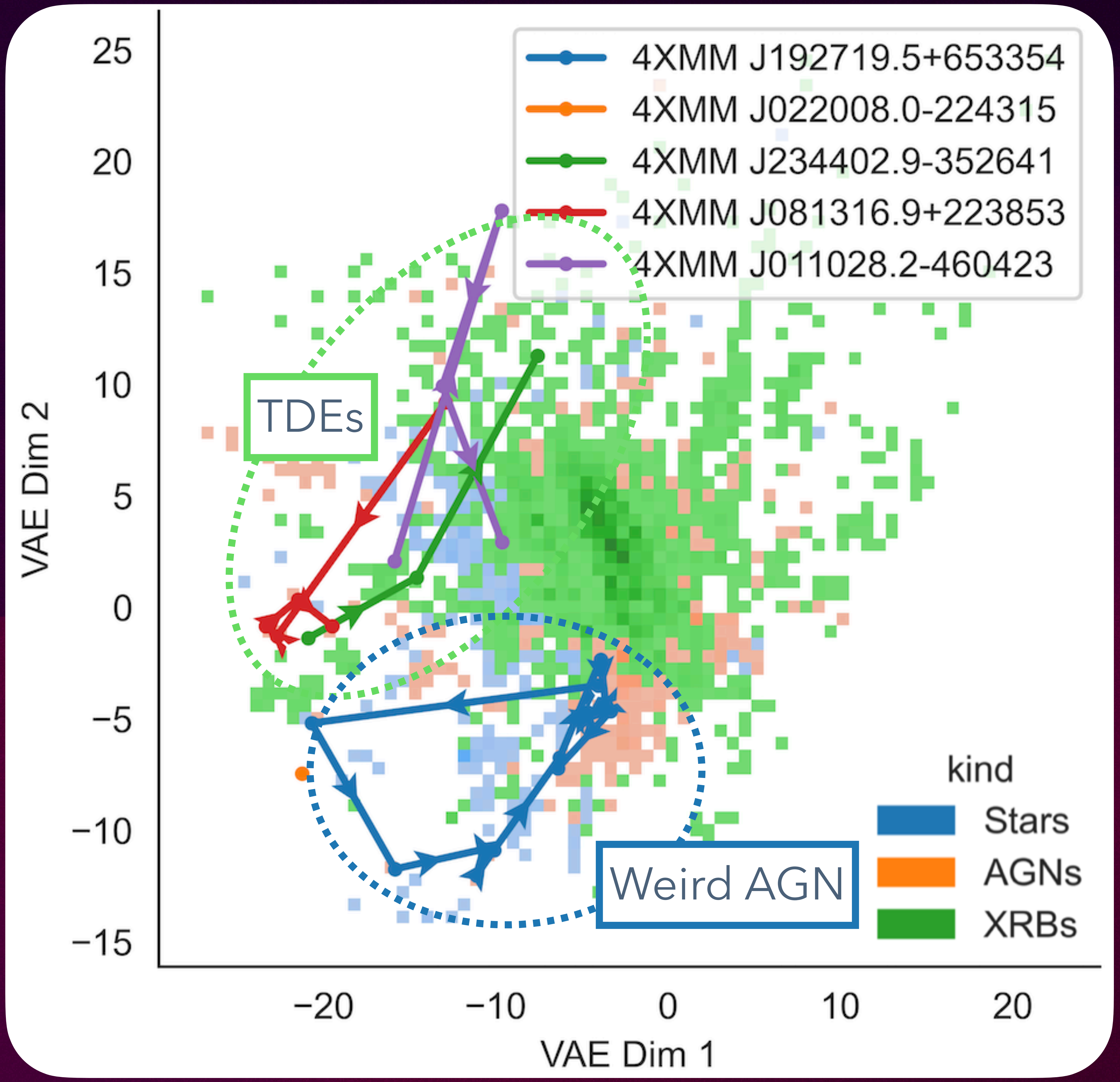


How do varying sources evolve?



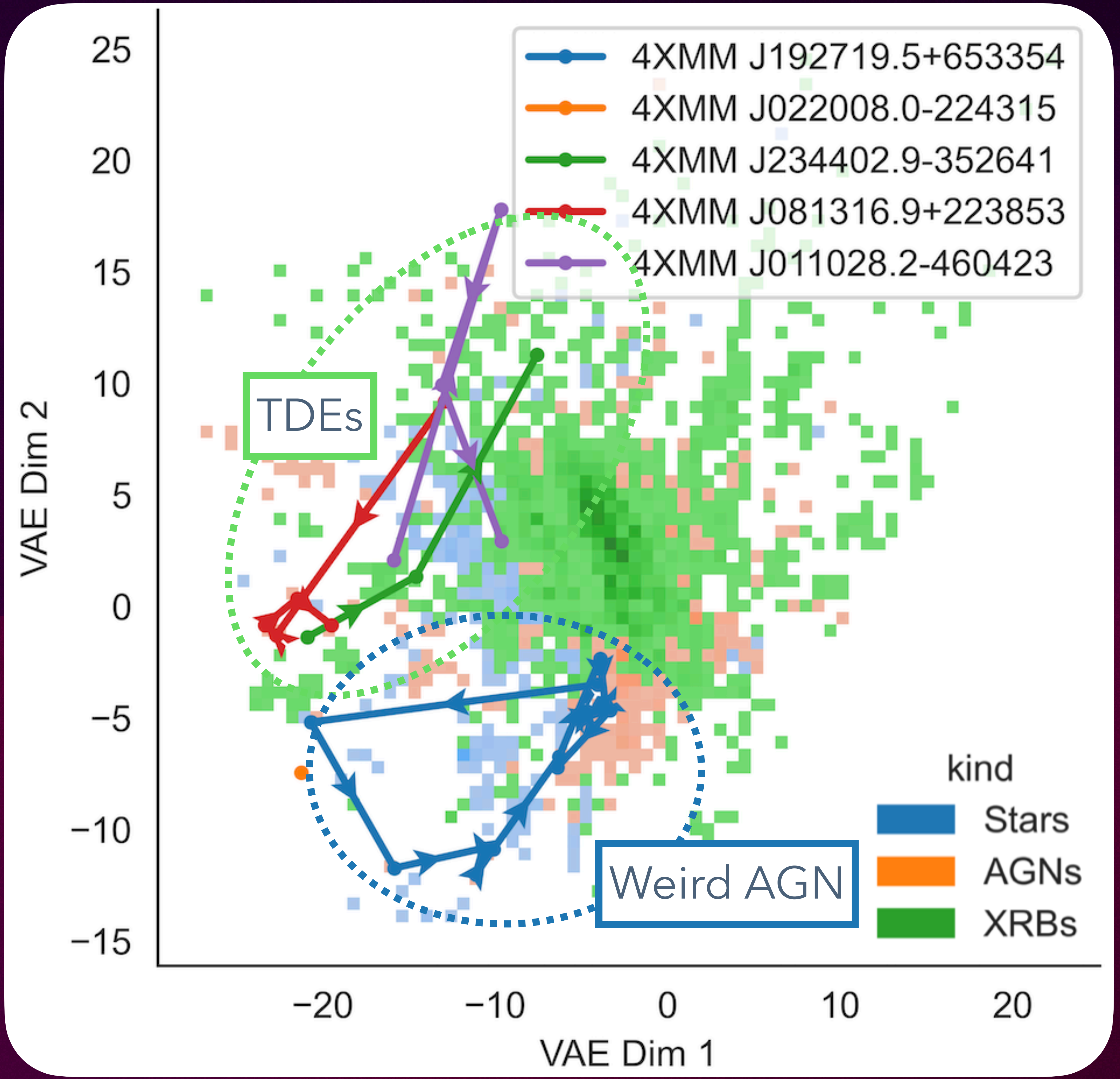
Trends in TDEs

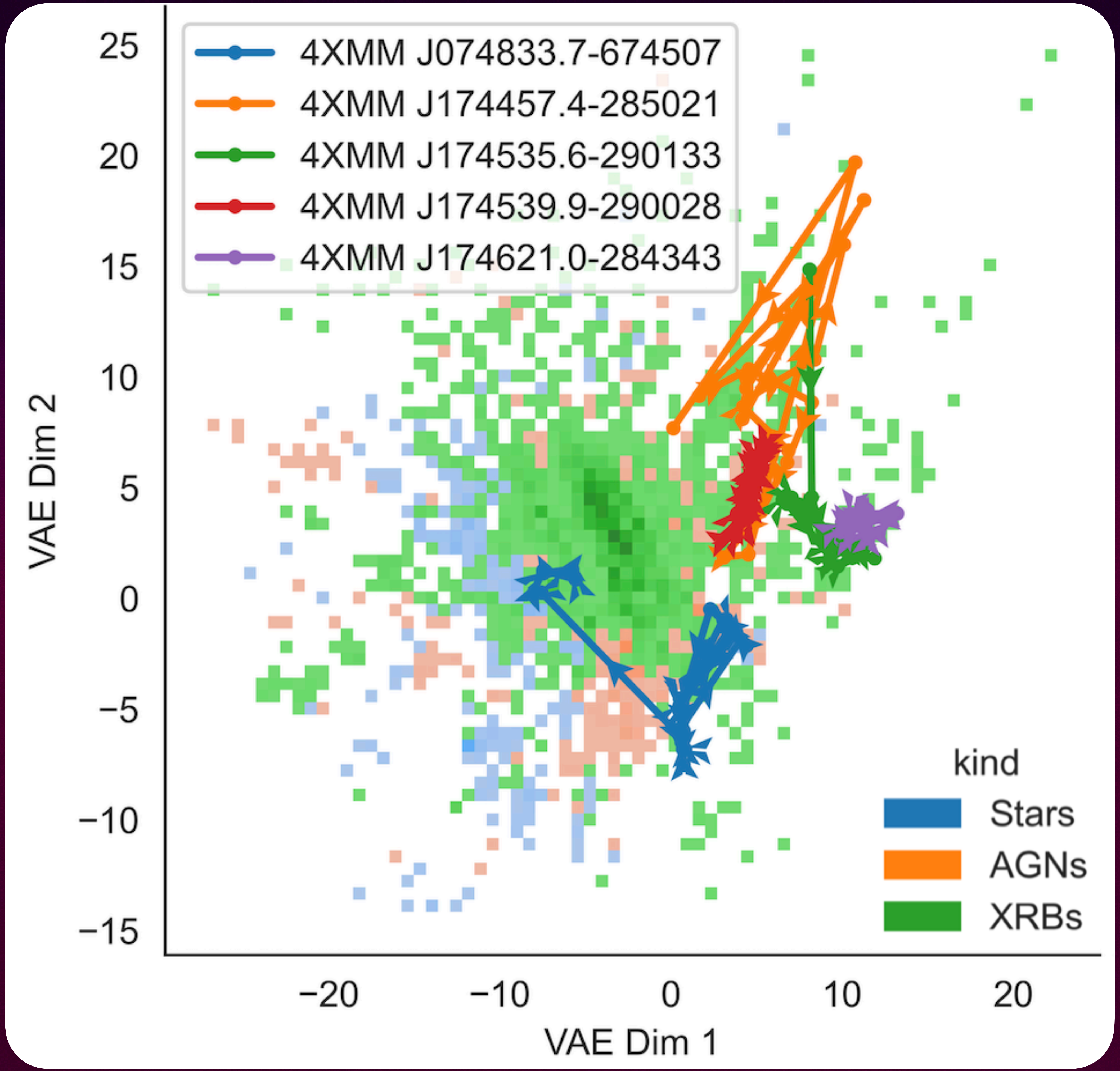
How do varying sources evolve?



Trends in TDEs

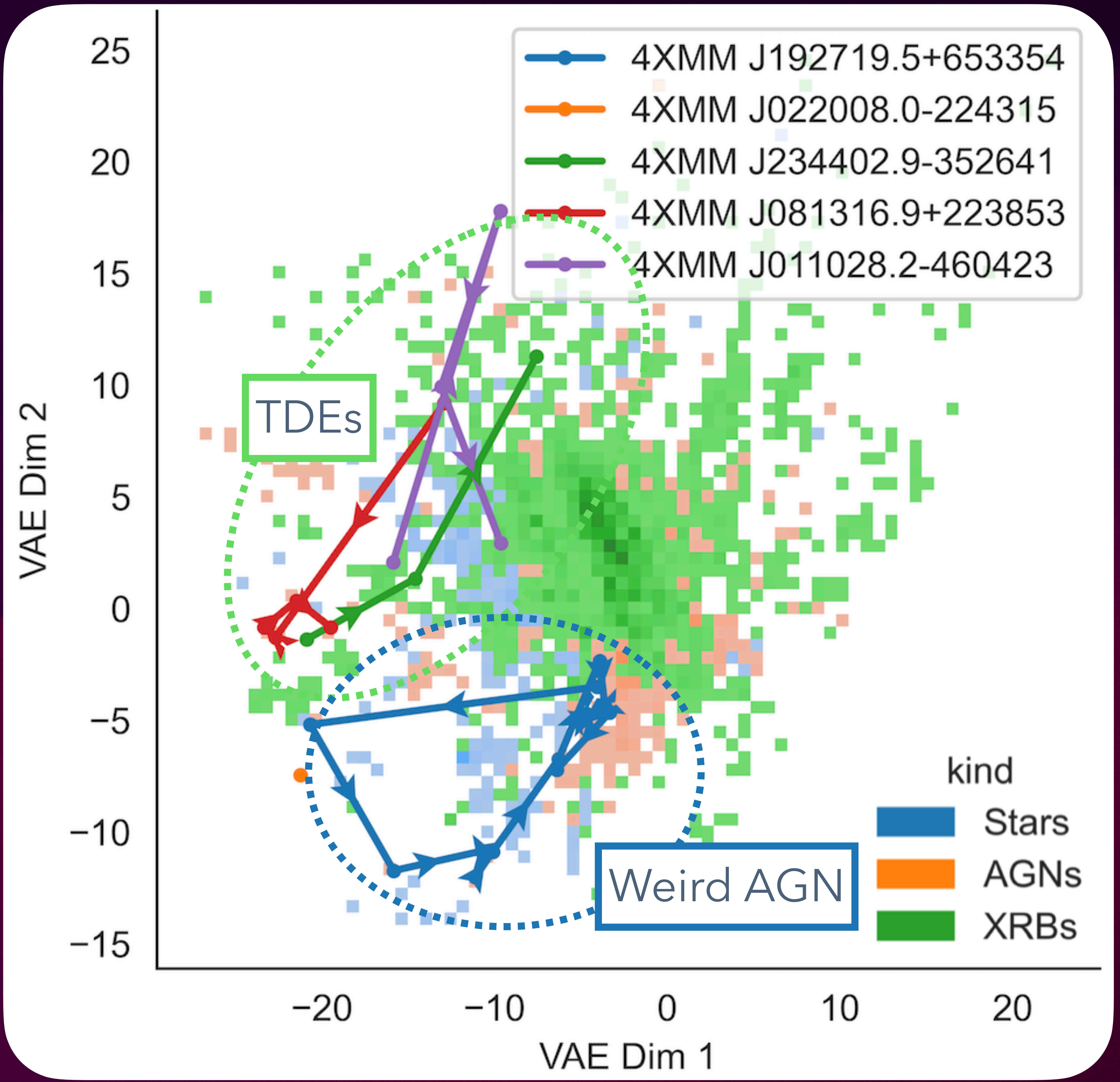
How do varying sources evolve?



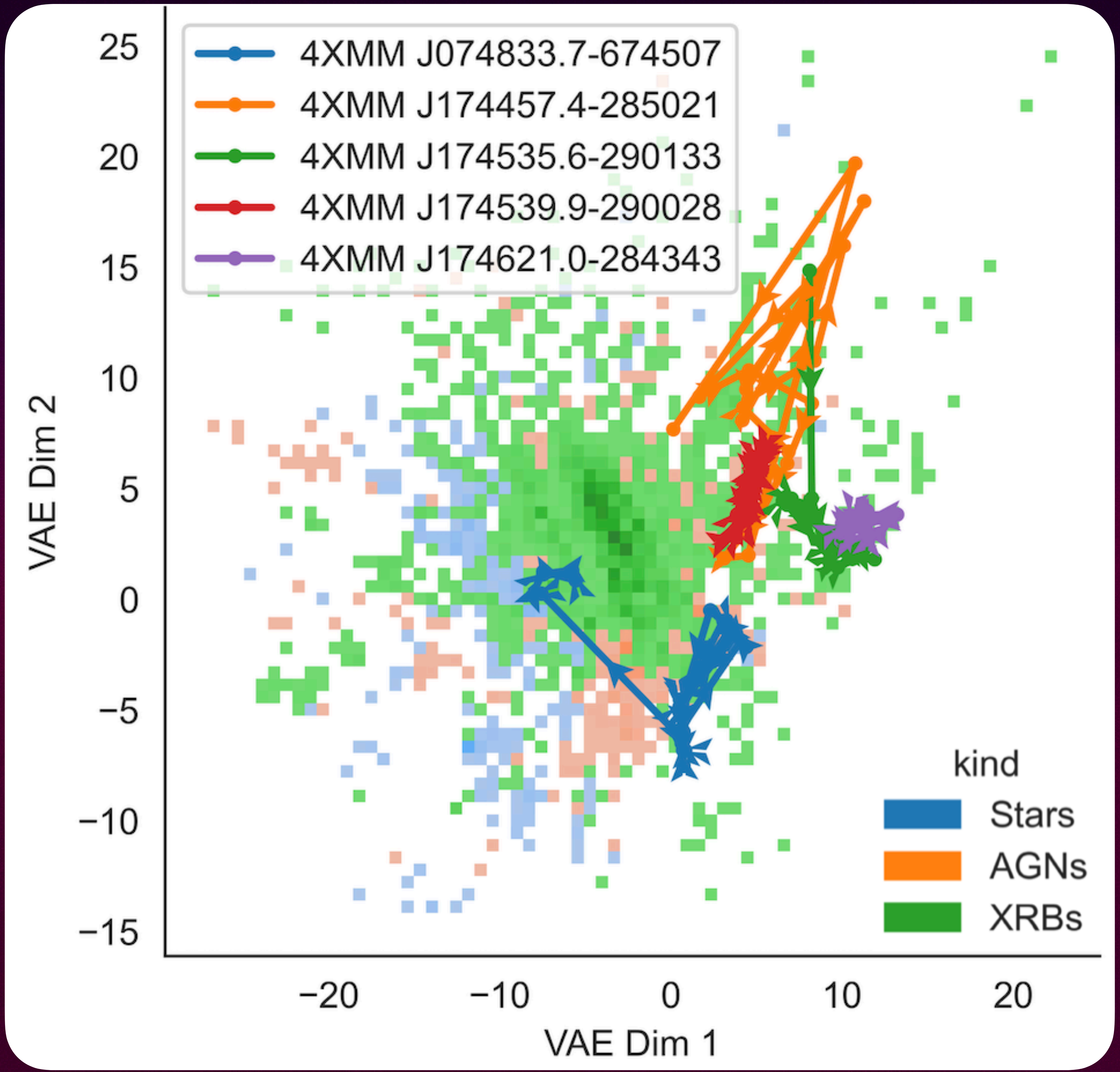


Trends in TDEs

How do varying sources evolve?



Trends in TDEs



Change of states in XRBs

Conclusion

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects
- Looking for similarities between sources in the latent space can help us in finding interesting objects in the catalogue haystack

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects
- Looking for similarities between sources in the latent space can help us in finding interesting objects in the catalogue haystack

What to do next ?

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects
- Looking for similarities between sources in the latent space can help us in finding interesting objects in the catalogue haystack

What to do next ?

- Hyperparameter tuning

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects
- Looking for similarities between sources in the latent space can help us in finding interesting objects in the catalogue haystack

What to do next ?

- Hyperparameter tuning
- Clustering in the latent space

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects
- Looking for similarities between sources in the latent space can help us in finding interesting objects in the catalogue haystack

What to do next ?

- Hyperparameter tuning
- Clustering in the latent space
- Exploration of latent trajectories

Conclusion

- Auto-encoding the XMM catalogue demonstrates the amazing science that is feasible with Machine Learning applied to large scale surveys
- Having a compressed & meaningful representation of the sources spectra help in studying the nature & the variability of these objects
- Looking for similarities between sources in the latent space can help us in finding interesting objects in the catalogue haystack

What to do next ?

- Hyperparameter tuning
- Clustering in the latent space
- Exploration of latent trajectories

Still we have a lot of work to do before we can publish this
Stay tuned for Dupourqué, Quintin + 202?