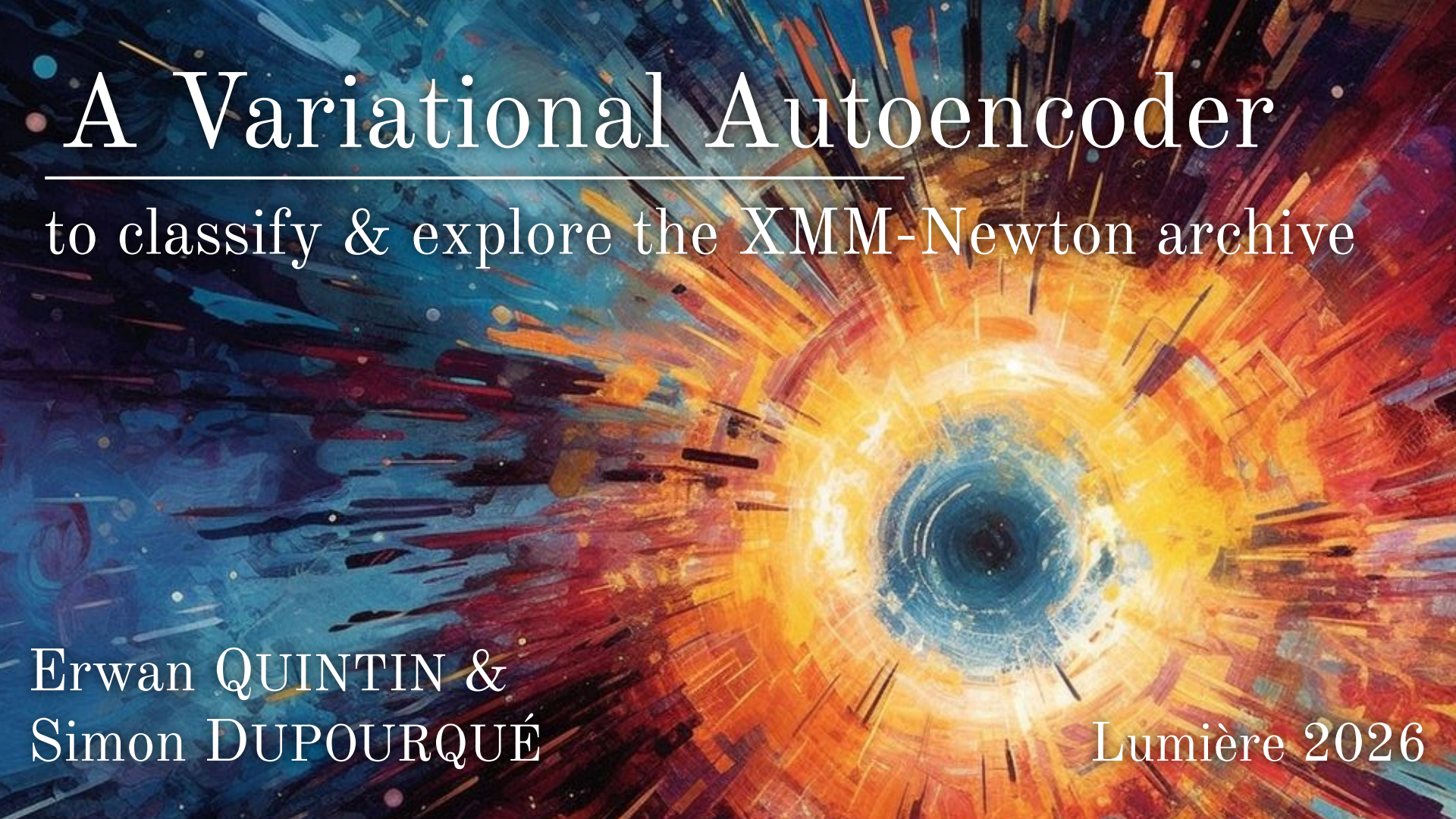
A Variational Autoencoder

to classify & explore the XMM-Newton archive



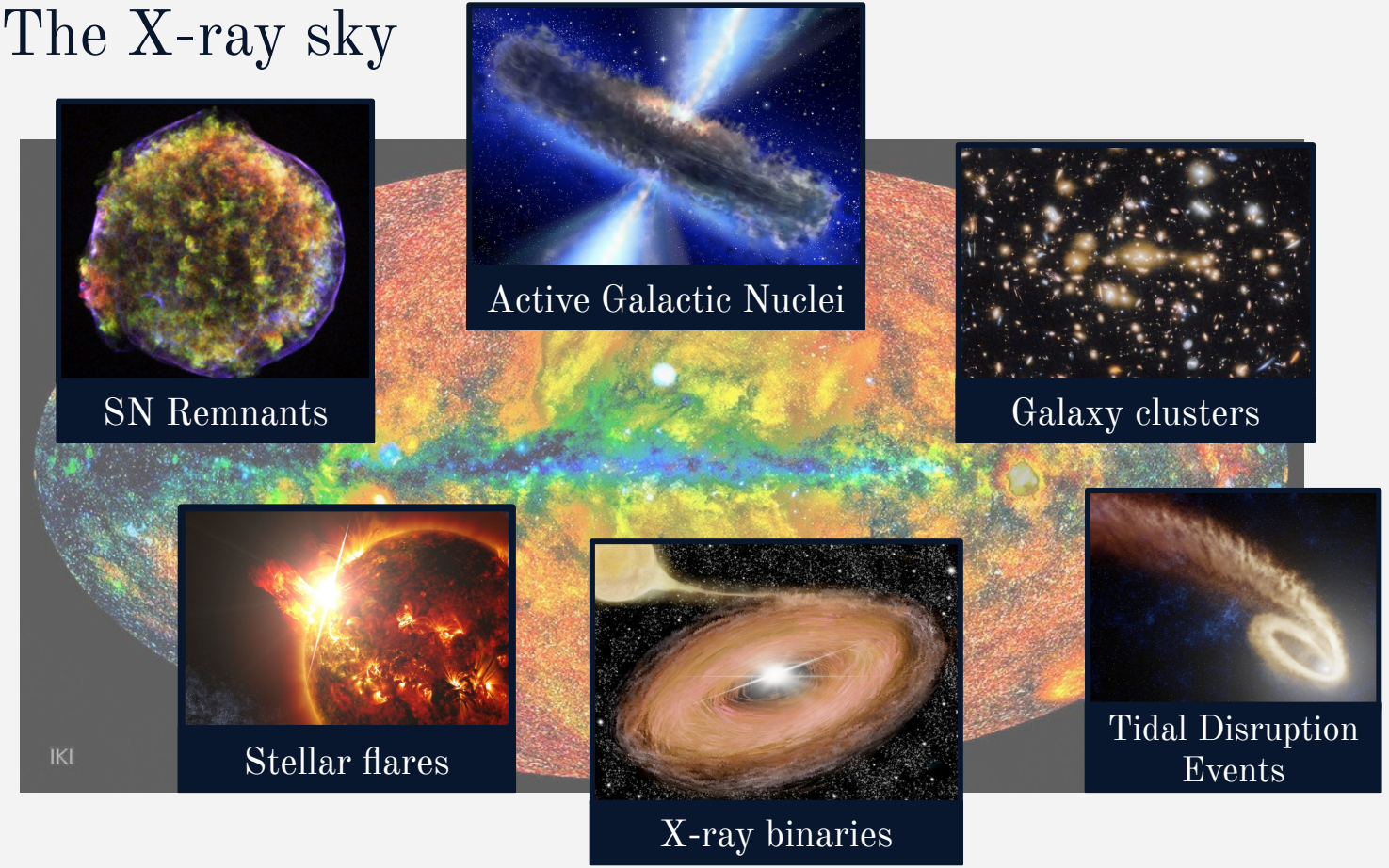
Erwan QUINTIN &
Simon DUPOURQUÉ

Lumière 2026

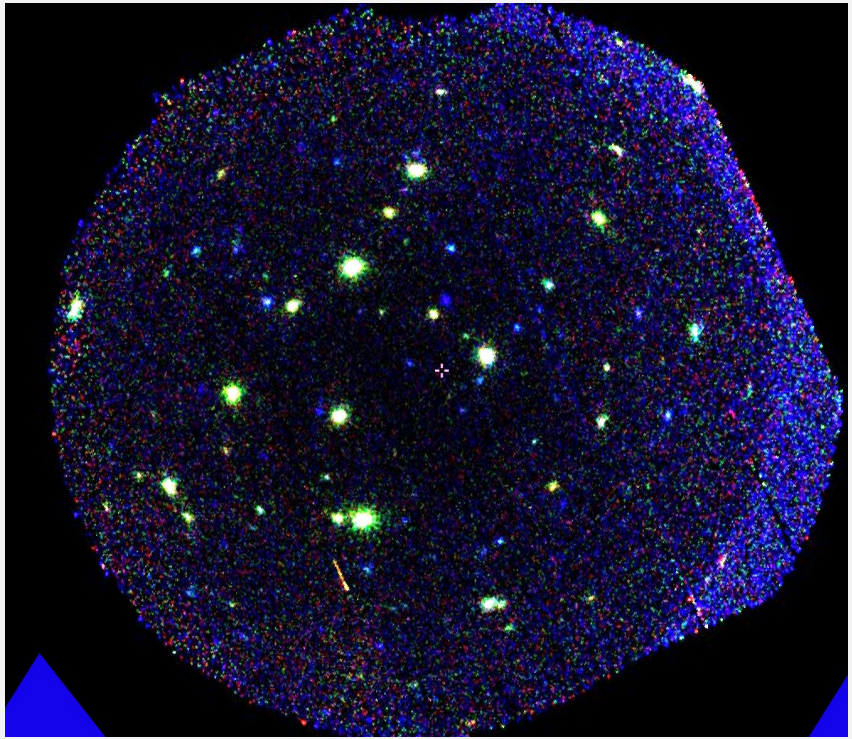
What's in the archives ?

And why we should care

The X-ray sky



XMM-Newton



~1M sources over 25 years

Need for classification

X-ray catalogs are typically ~1M sources

They can be mined to:

- Build large sample for **statistics**
- Build large sample for **training** purposes
- Look for objects **similar to a target**
- Look for **outliers**

Classification in X-rays

Based on X-ray data:

- Spectral data is hard to exploit fully
- Hardness ratios

In general: supervised is easier
but less likely to find exciting results

Our attempt at classification

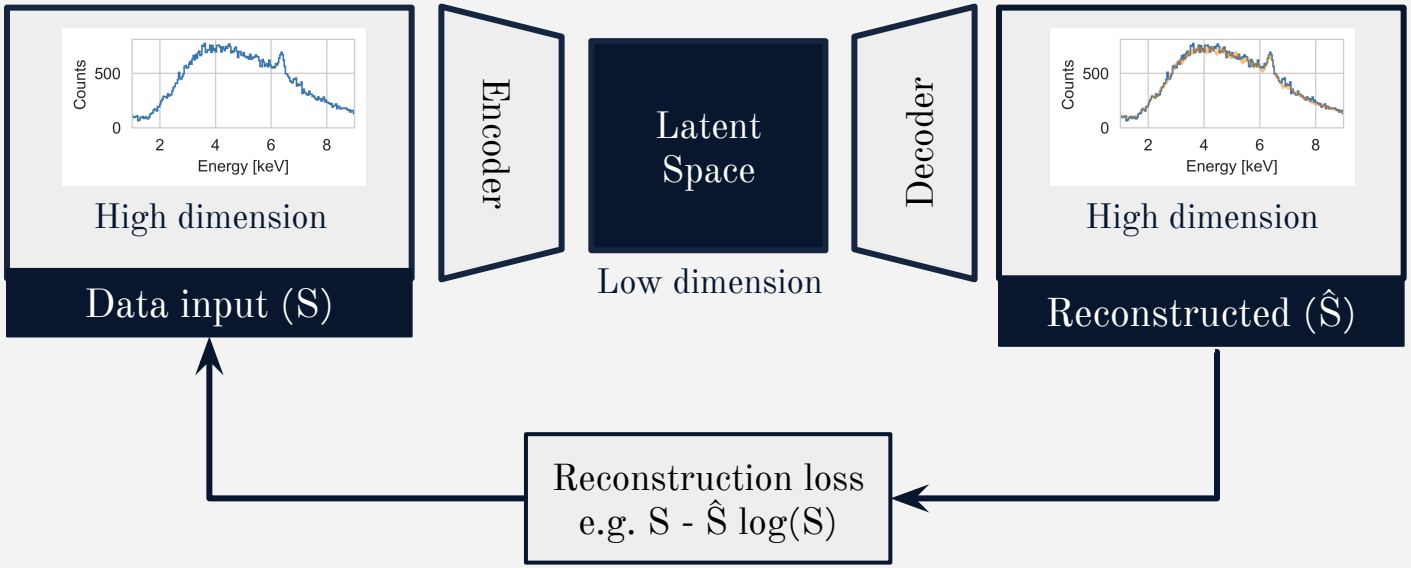
Using a VAE

Our main goal:
Spectrum-based unsupervised classification of point-like sources

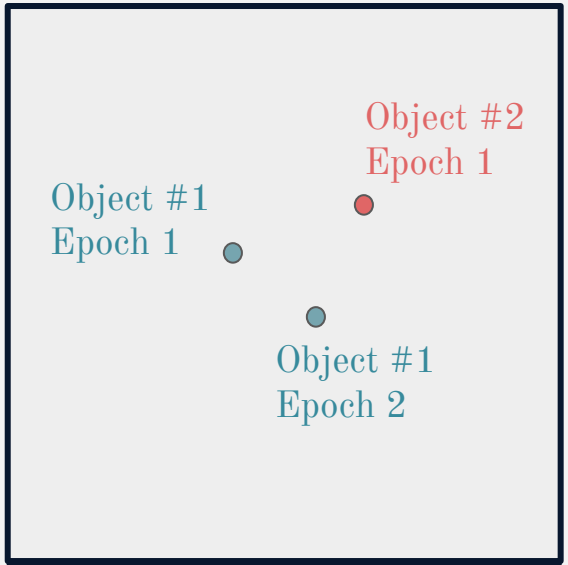
What's an Autoencoder ?

- Dimension-reduction machine
- Other example: Principal Component Analysis

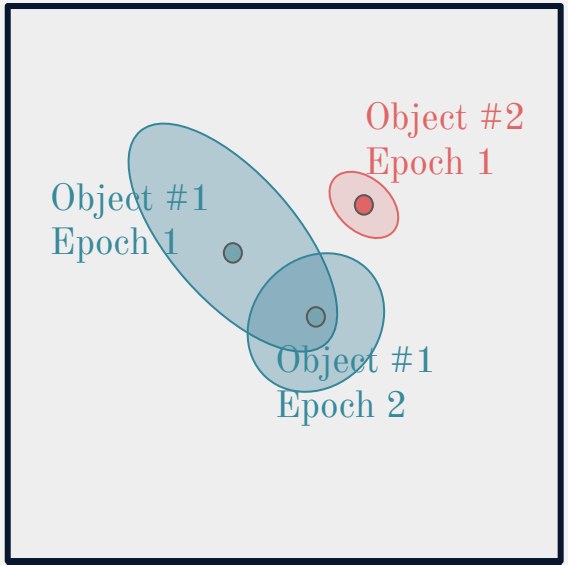
What's an Autoencoder ?



What's a Variational Autoencoder ?



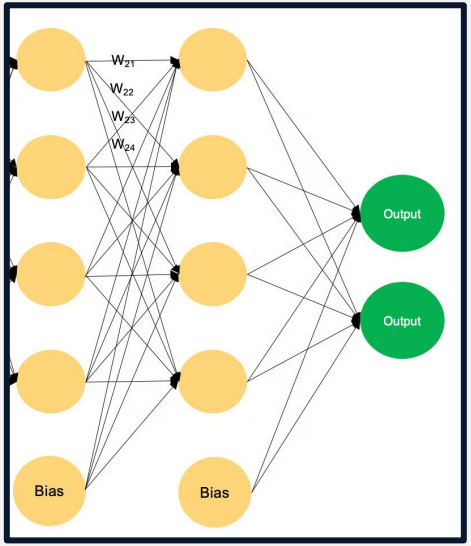
Standard latent space



Variational latent space

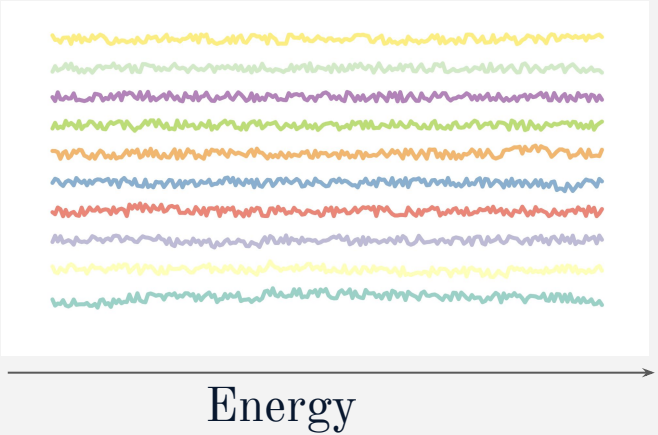
How to encode / decode

- Typically both are simple neural networks (Multi-Layer Perceptron, MLP)
- This works OK for encoding, but struggles at decoding (to catch sharp features)
- All physical info is lost (energy bins are not independent)



How to encode / decode

- Solution to decode: Deep Operational Networks (**DeepONet**)
- Two parallel networks:
  - The **Trunk network** learns “templates” spectra
  - The **Branch network** learns how to combine these templates



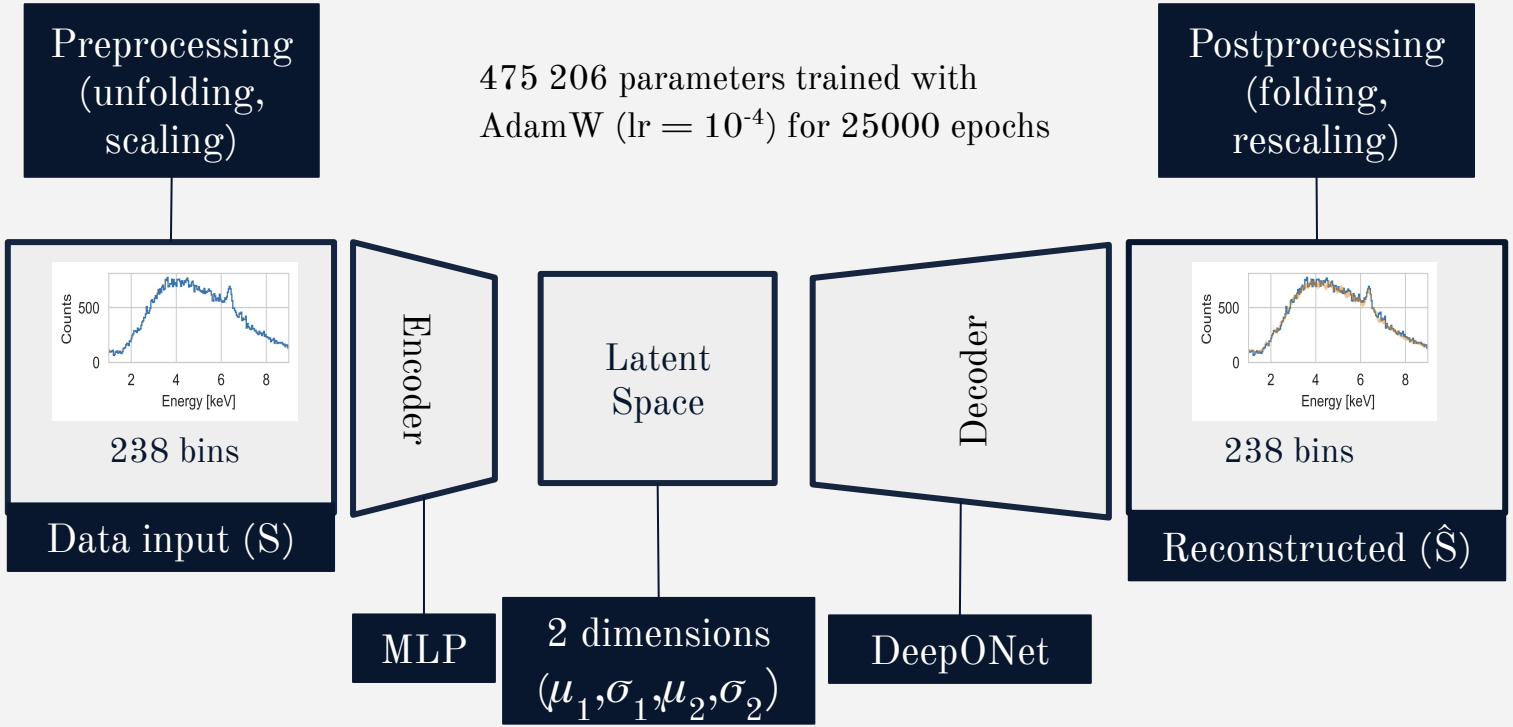
About the data itself

- We take the **EPIC pn spectra** from the 4XMM DR14 catalog
- Keep those above **300** counts, SUMFLAG$<$3
- Should account for observational aspects
  (i.e. work on **unfolded spectra**)
- Bin them in 238 bins in the 0.3-10 keV band
- Scaled

$>$ One spectrum per observation per object

Final architecture

- The archives
- Method
- Early results
- Conclusion

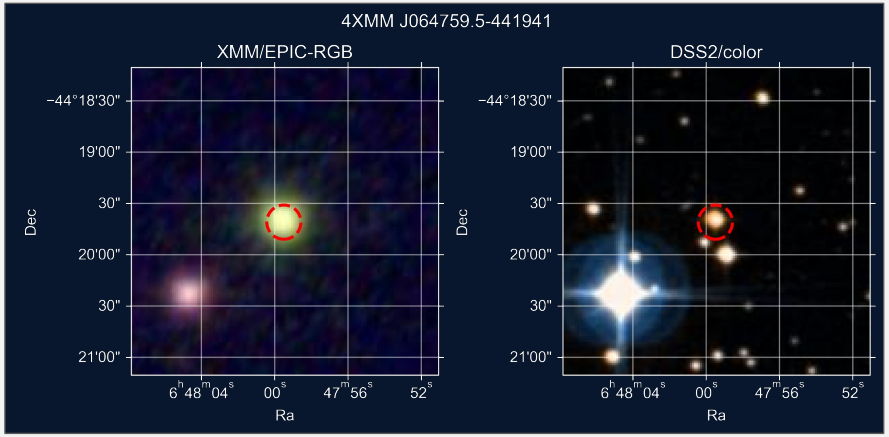


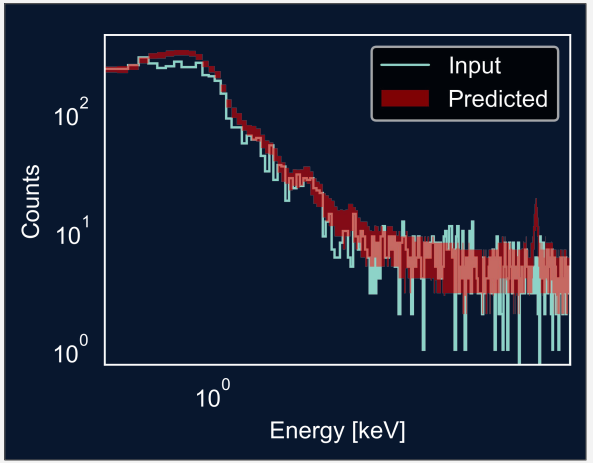
Some early results

Some reconstructed spectra

- The archives
- Method
- Early results
- Conclusion

Star

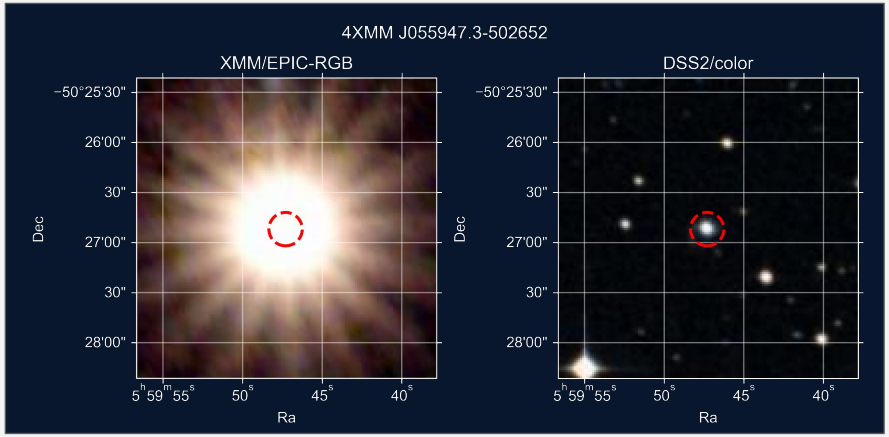


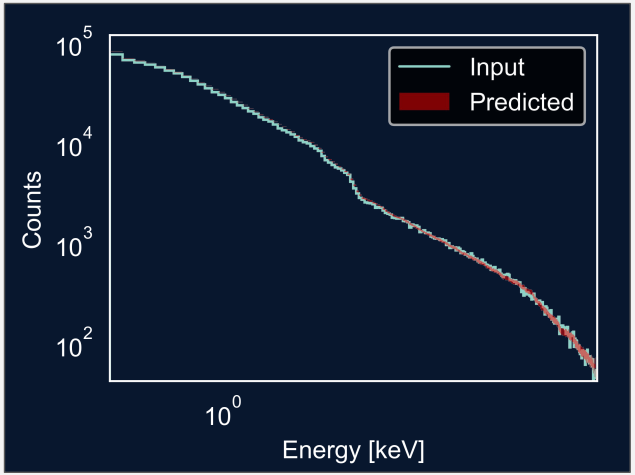


Some reconstructed spectra

- The archives
- Method
- Early results
- Conclusion

AGN

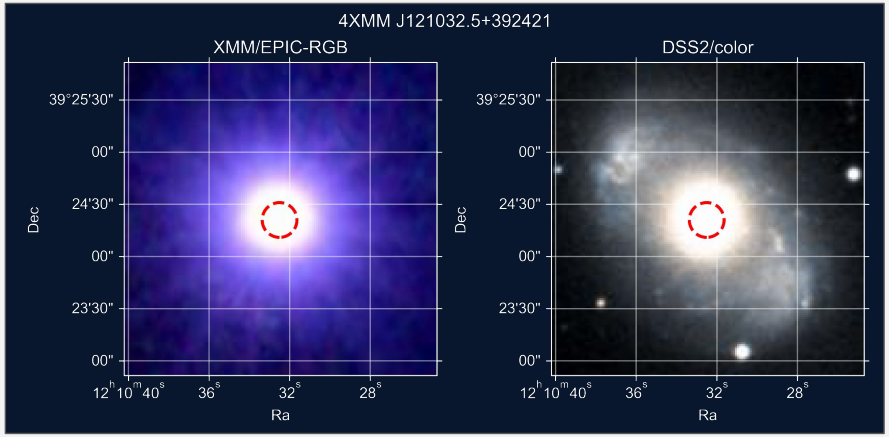


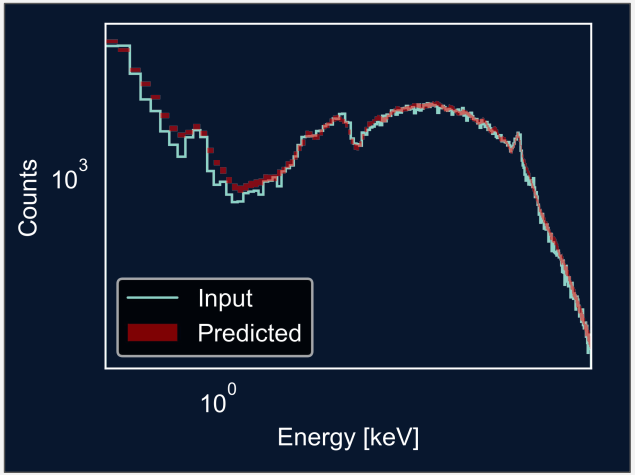


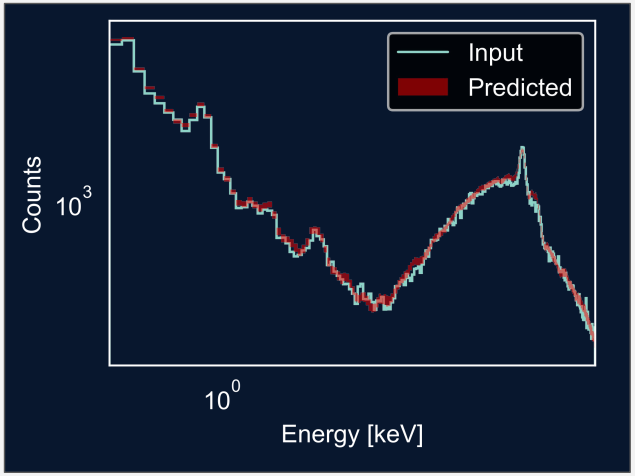
- The archives
- Method
- Early results
- Conclusion

Some reconstructed spectra

AGN



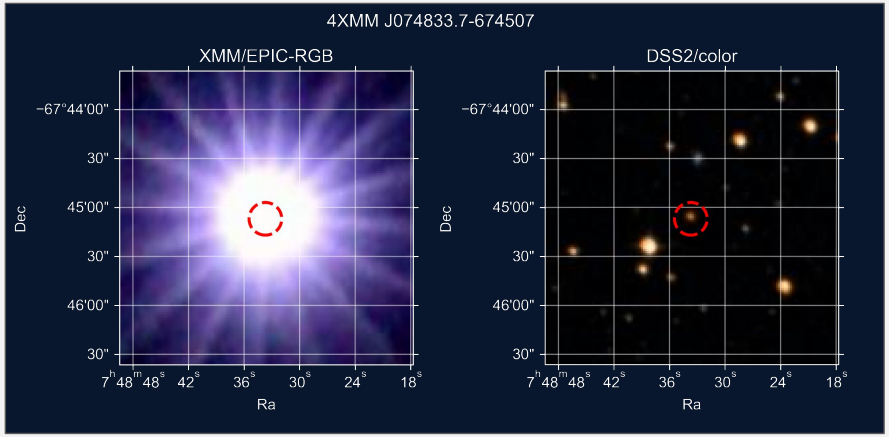


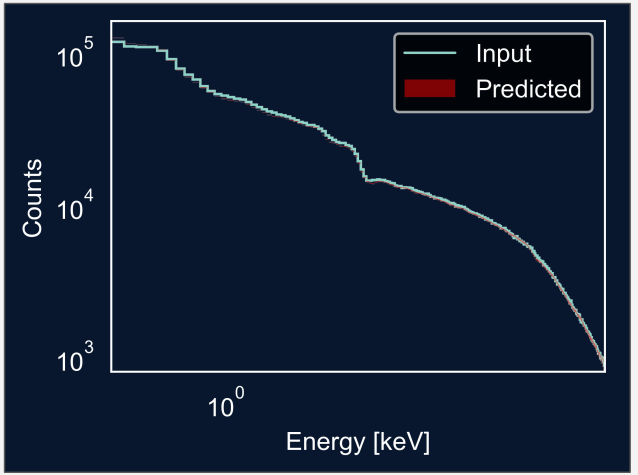


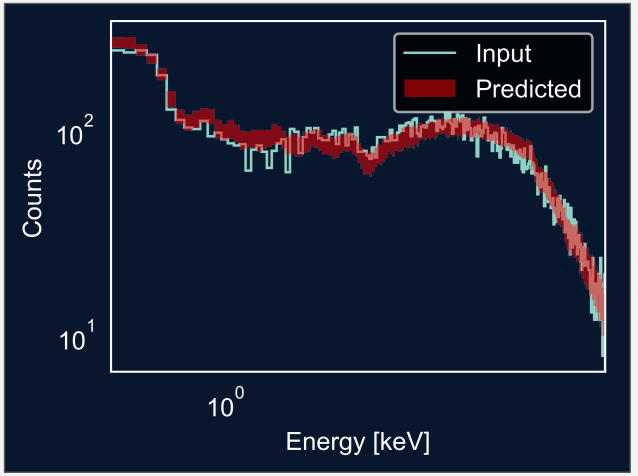
Some reconstructed spectra

- The archives
- Method
- Early results
- Conclusion

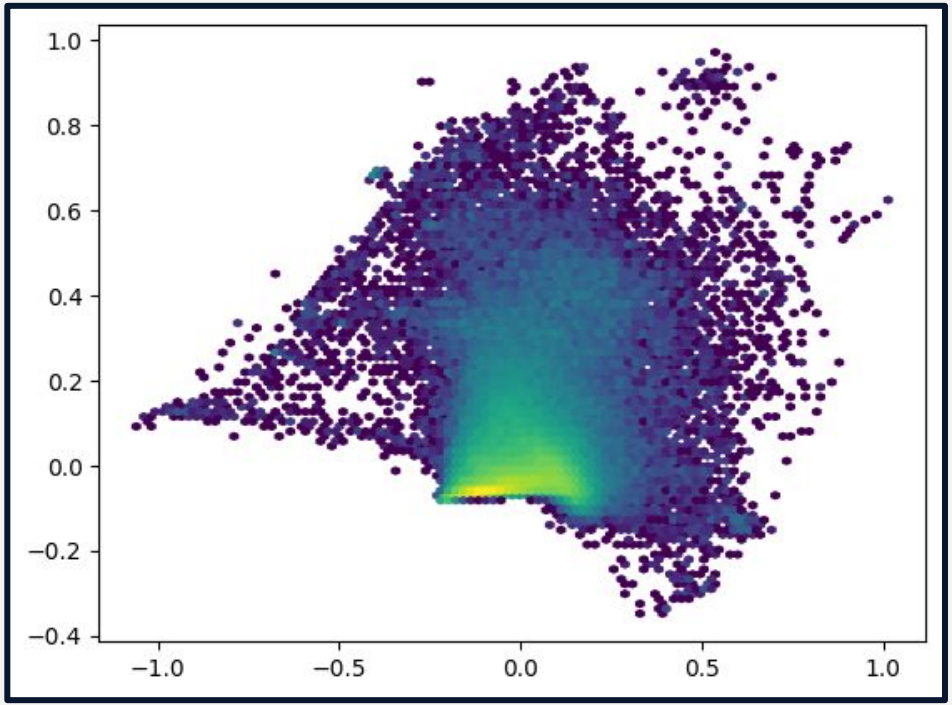
XRB





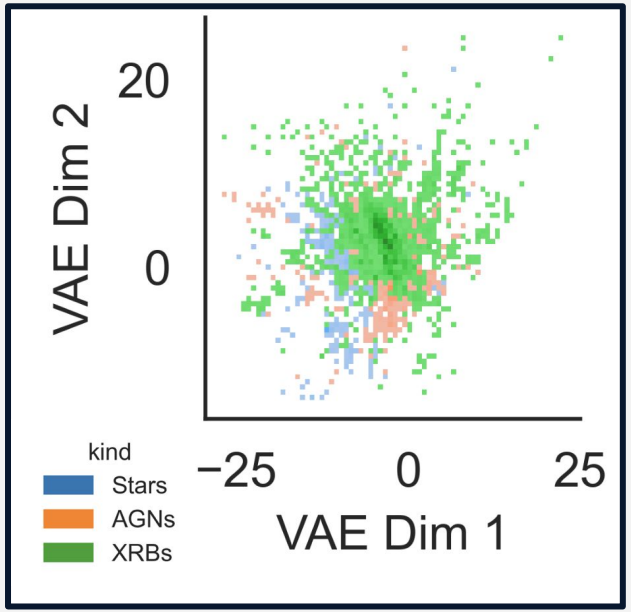


Latent space



Heatmap of latent space
Some clusters

Latent space



Simbad match to assess clustering

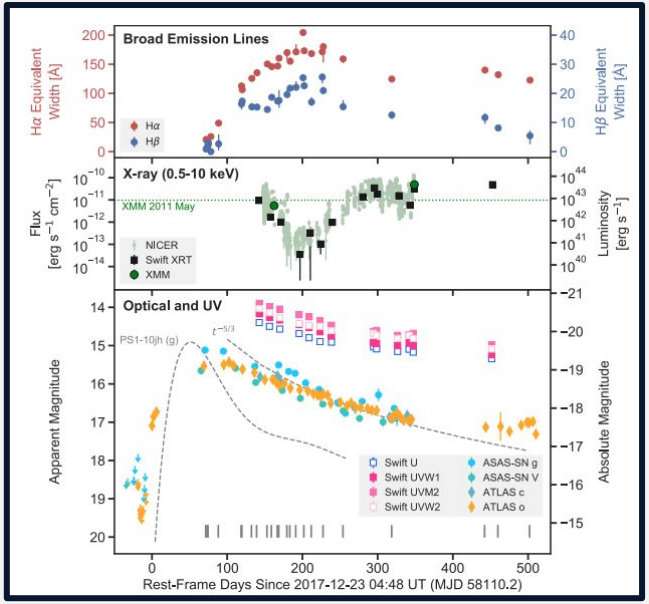
Sort of works ?

Progress to be made here

Similarity search

- The archives
- Method
- Early results
- Conclusion

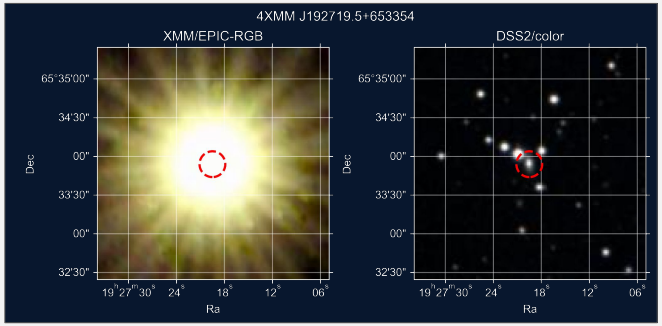
1ES 1927

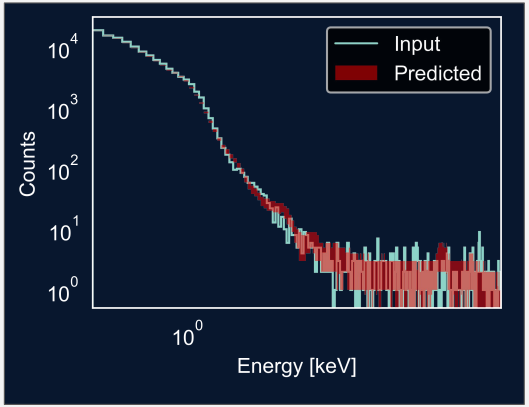


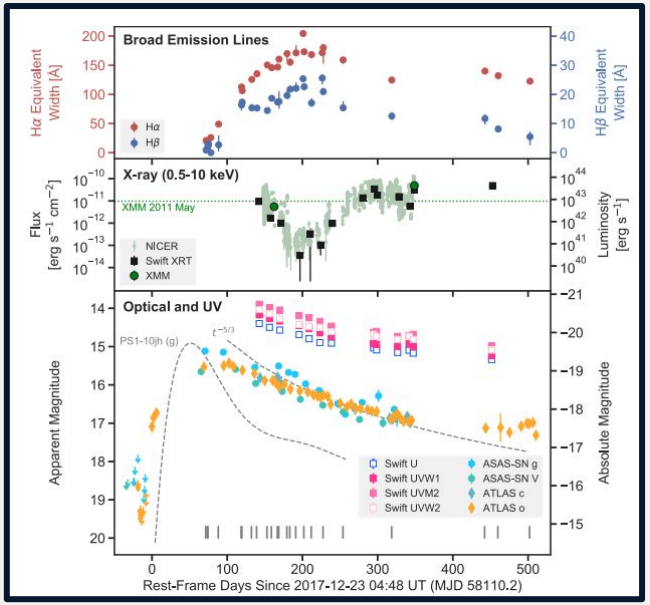
Similarity search

- The archives
- Method
- Early results
- Conclusion

1ES 1927





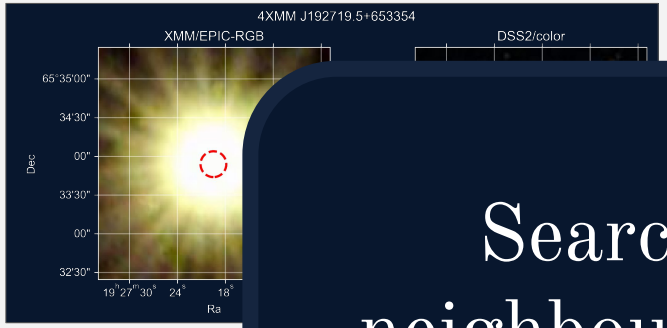


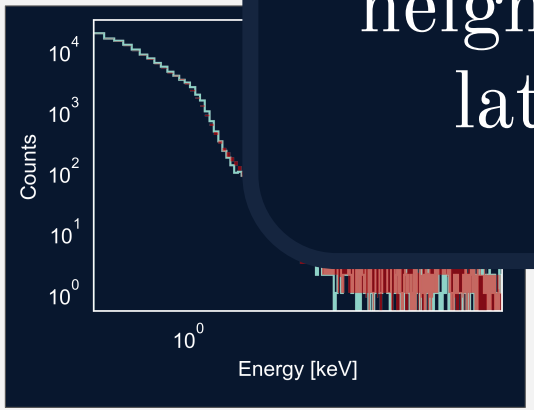
Very red spectrum:
look for similar ones ?

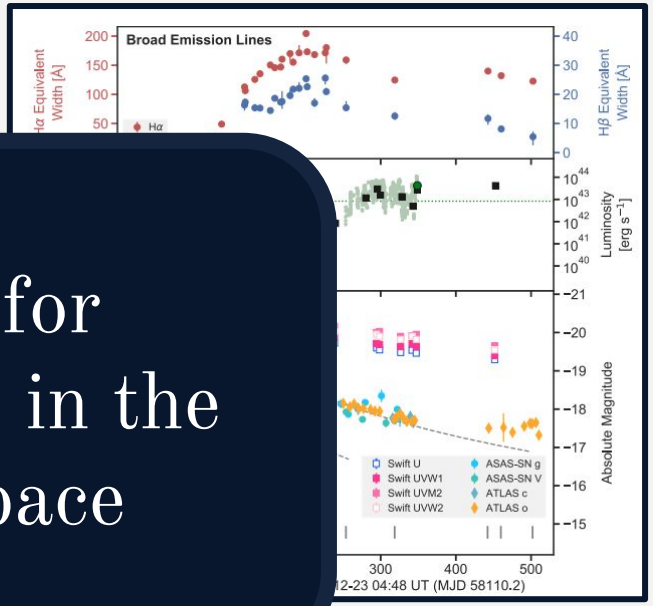
Similarity search

- The archives
- Method
- Early results
- Conclusion

1ES 1927







Search for neighbours in the latent space

Very red spectrum:
look for similar ones ?

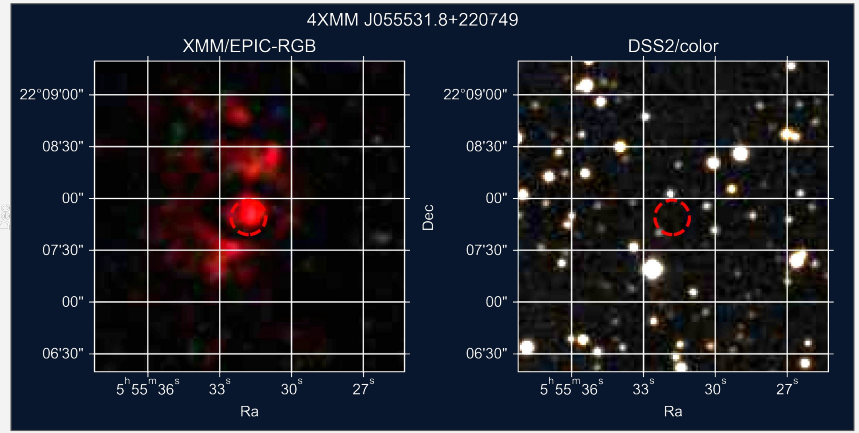
Similarity search

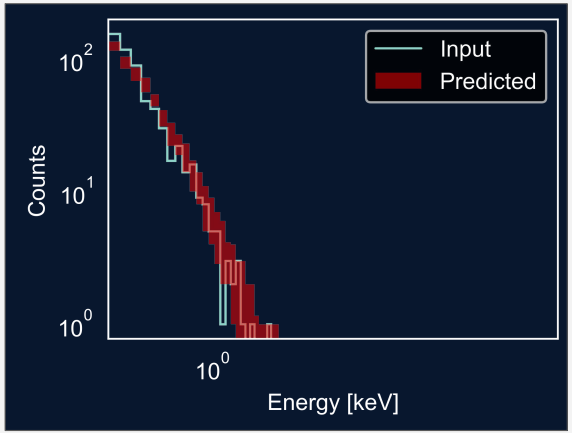
- A lot of TDEs / X-ray soft objects
-

Similarity search

- The archives
- Method
- Early results
- Conclusion

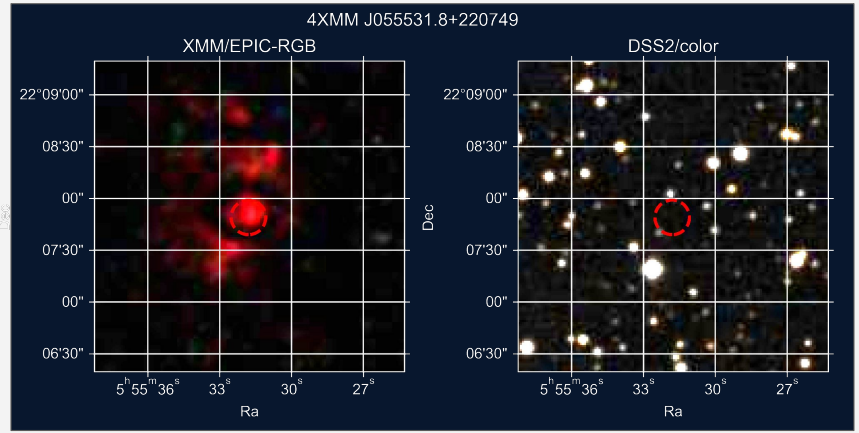
- A lot of TDEs / X-ray soft objects
- A mysterious object

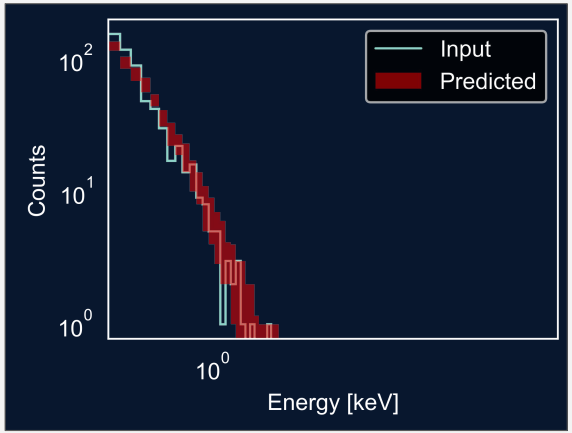




Similarity search

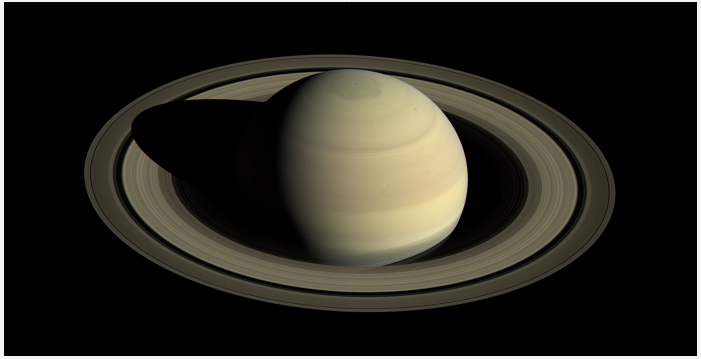
- A: AGN
- B: XRB
- C: Star
- D: Pulsar

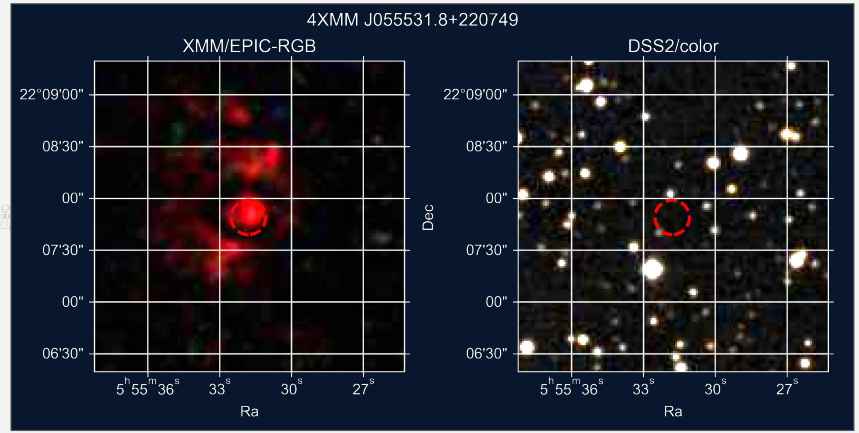


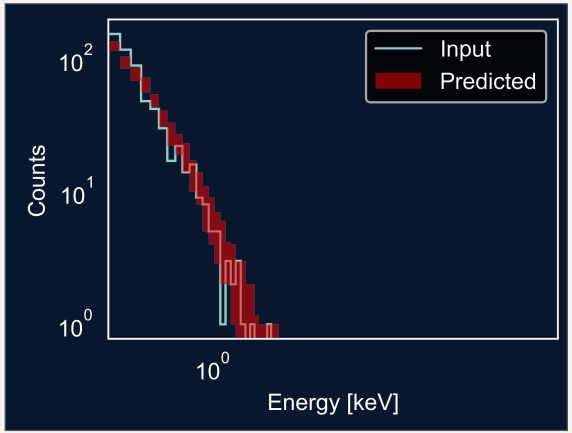


Similarity search

- A: AGN
- B: XRB
- C: Star
- D: Pulsar

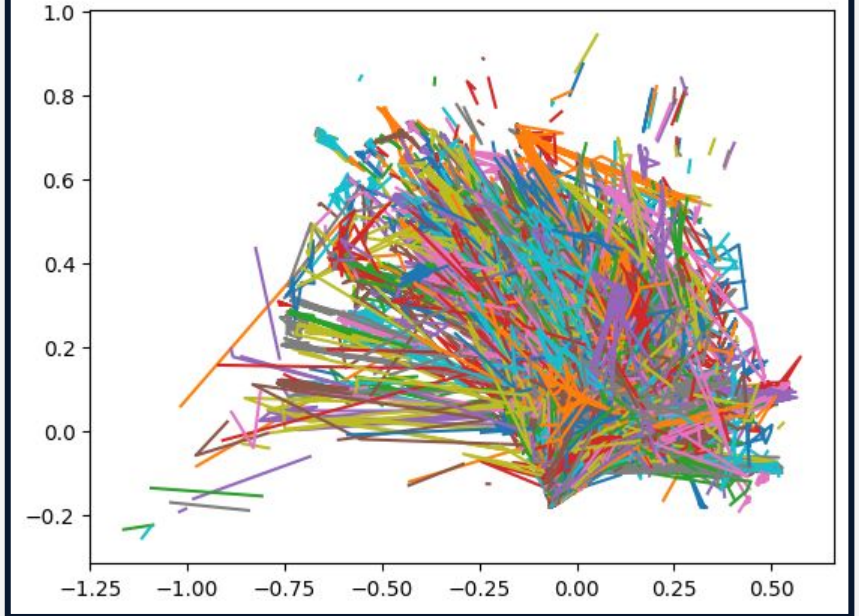






Conclusions / Perspectives

Important features to implement

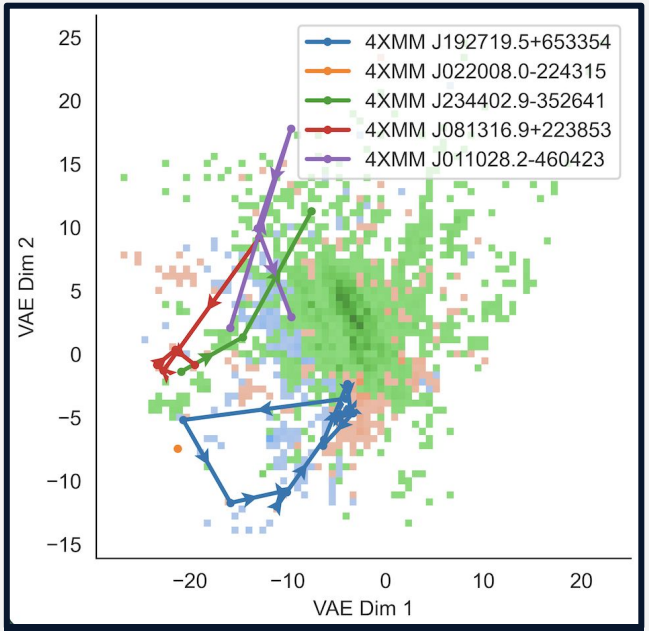


Latent space trajectories

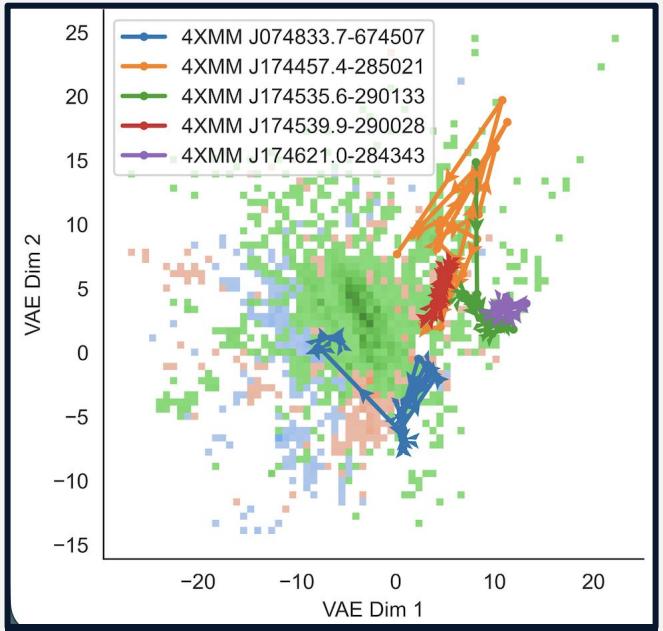
- Temporal evolution interpreted as **trajectories** in latent space

Important features to implement

- The archives
- Method
- Early results
- Conclusion



Latent TDE trajectories



Latent XRB trajectories

Other improvements

- Latent space trajectories
- Hyperparameter tuning (e.g. dimension)
- Clustering / regularisation in latent space ?

Summary

- Autoencoding XMM spectra to represent X-ray detections in a **low-dimension space**
- Then use this to classify / look for neighbors / outliers
- Still a lot to improve / develop, but promising

Questions ? erwan.quintin@esa.int


Thank you !

What's a Variational Autoencoder ?