AI and complex networks in multimodal X-ray astrophysics

POLITECNICO
MILANO 1863

Nicolò Pinciroli

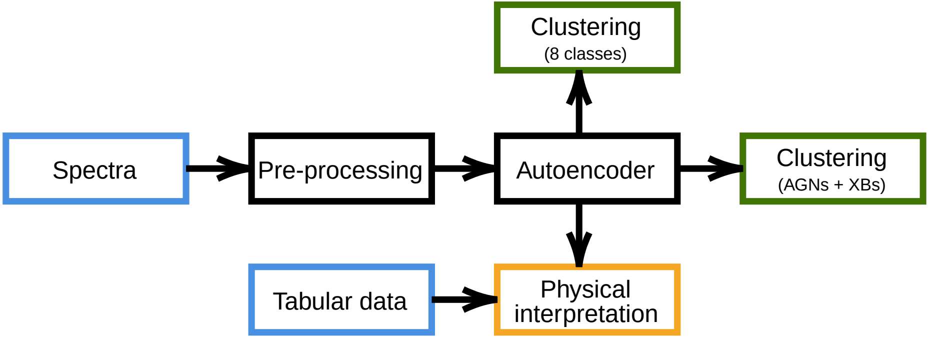
16 Apr 2025

Introduction

Goals	Data	Methods
Cluster Chandra X-ray **spectra**	Chandra spectra	Autoencoder, clustering algorithms, symbolic regression
Align multimodal data (X-ray events and scientific texts)	ADS Database, event files	Contrastive learning
Discover patterns in multimodal data	ADS Database, event files (high variability)	Complex networks

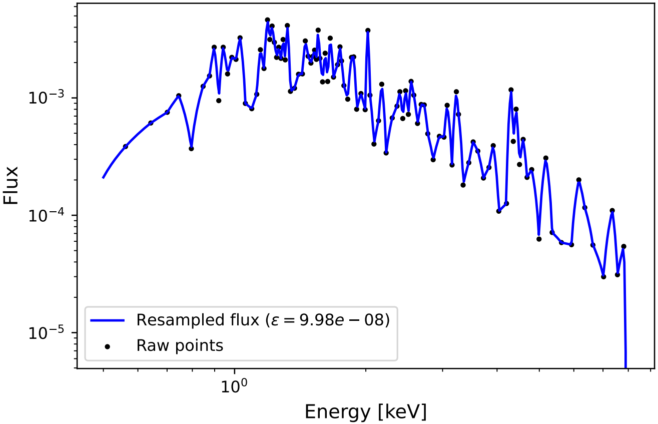
Clustering spectra

- Cluster Chandra observations based on energy spectra
- Summarize Chandra spectra in a lower-dimensional space ($\rightarrow$ *embedding space*)



Preprocessing

- Spectra have a variable number of points → resampling using 400 points with a log scale (denser sampling for low energies)
- Fluxes depend on distance, not intrinsic properties → MinMax normalization



Classes

AGN

CV

HM-STAR

HMXB

LM-STAR

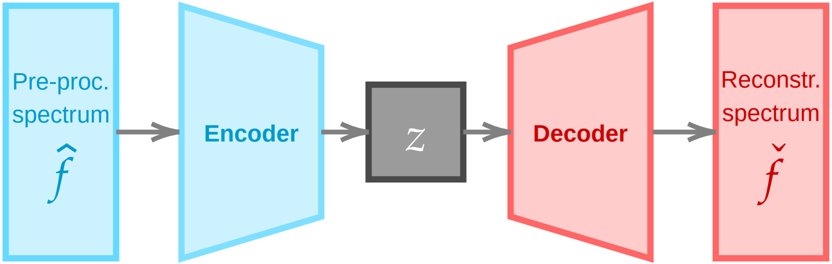
LMXB

NS

YSO

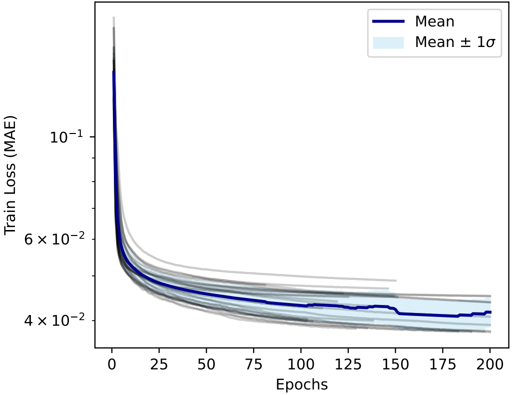
Transformer autoencoder

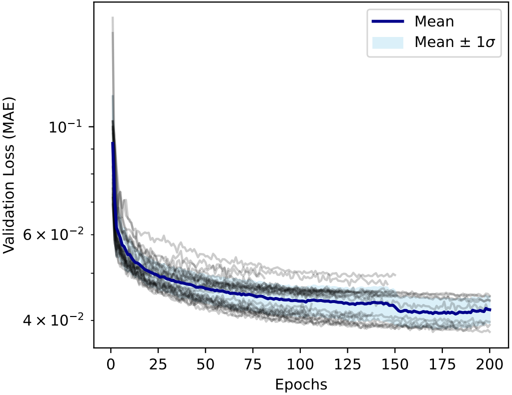
- **Goal**: reconstruction of pre-processed spectra
- RNN-based vs Transformer-based autoencoders → transformer autoencoders focus on relevant spectral regions
- **Evaluation**: reconstruction error (MAE)



Transformer autoencoder

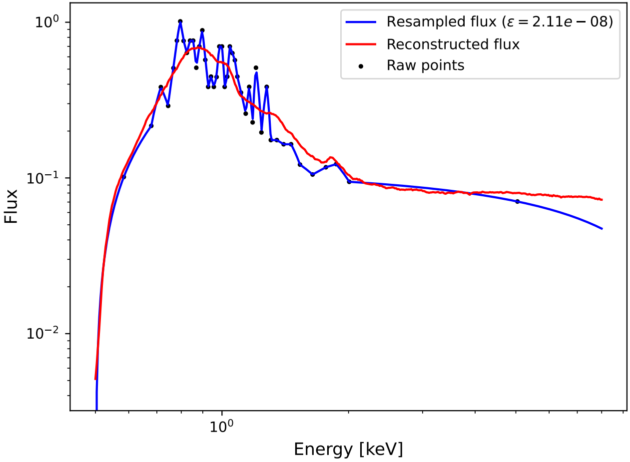
- Good reconstruction independently on the hyperparameters
- Reconstruction $\implies$ smoothing of the preprocessed spectra





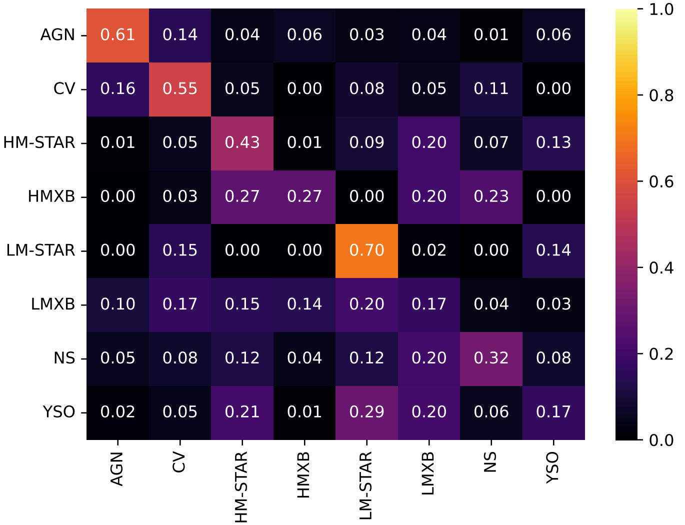
Transformer autoencoder

- Good reconstruction independently on the hyperparameters
- Reconstruction $\implies$ smoothing of the preprocessed spectra



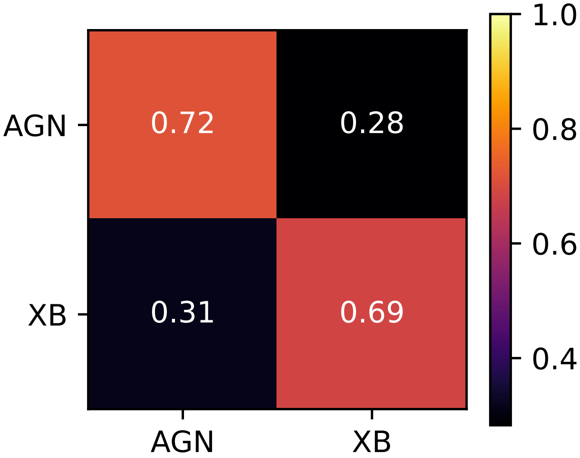
Clustering (8 classes)

- Best clustering output visualized with a **contingency matrix** (normalized confusion matrix)
- Cluster-class assignment optimized to maximize trace
- **Balanced accuracy:** $\approx 40\%$



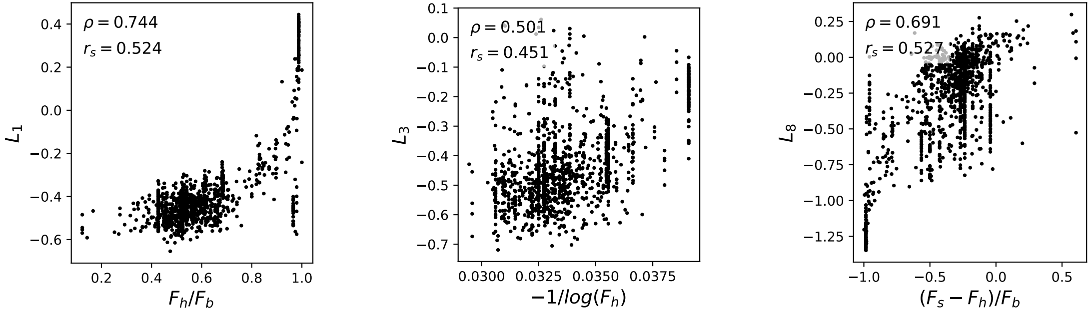
Clustering (2 classes)

- Best clustering output visualized with a **contingency matrix** (normalized confusion matrix)
- Cluster-class assignment optimized to maximize trace
- **Balanced accuracy**: $\approx 70\%$



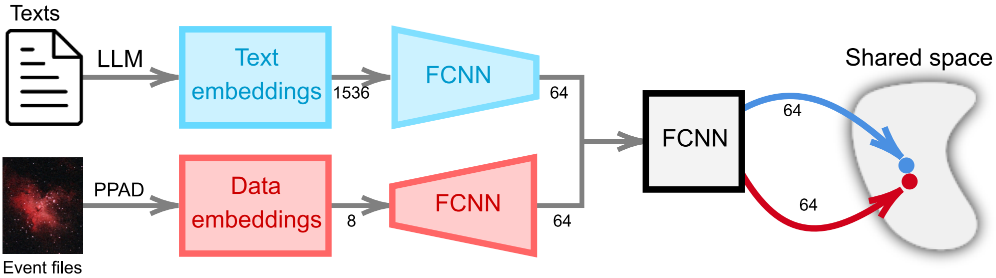
Latent space interpretation (AGNs)

Do latent dimensions have a physical meaning? → **symbolic regression**



Learning from multimodal data

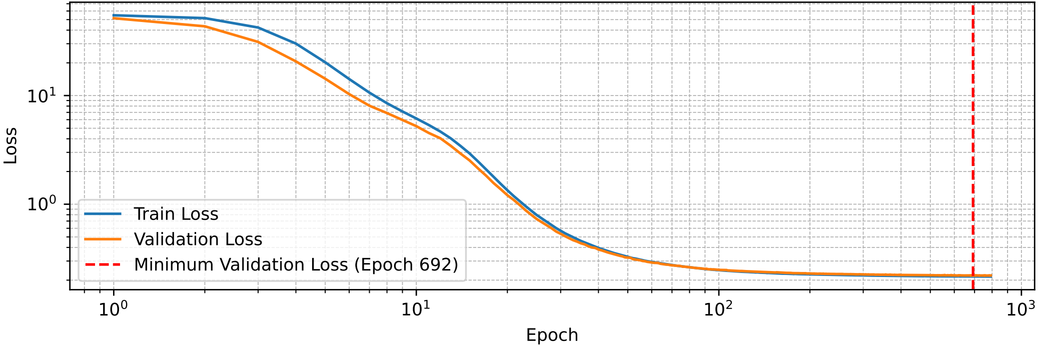
- Large multimodal datasets in astronomy → **how to learn from multiple modalities?**
- **Idea:** align X-ray data (event files) embeddings and scientific texts (from ADS) embeddings using contrastive learning



Contrastive learning

Regularized InfoNCE loss to align modalities into a 64-dimensional shared space

$$\mathcal{L} = \mathcal{L}_{InfoNCE} + \gamma \mathcal{L}_{reg}$$



Regression

- Combined modality better than individual modalities in regression
- Text information enriches light curve data for hardness ratios

Variable	Modality			Improvement	
	Text	**Data**	**Both**	**Absolute**	**%**
Hard$_{HS}$	0.40	0.29	**0.22**	0.07	24%
Hard$_{MS}$	0.32	0.27	**0.25**	0.02	7%
Hard$_{HM}$	0.27	0.22	**0.17**	0.05	23%
p_{var}	0.23	0.22	**0.21**	0.01	5%
F_{sig}	5.15	2.43	**2.27**	0.16	7%

Nearest Neighbor search and Anomaly Detection

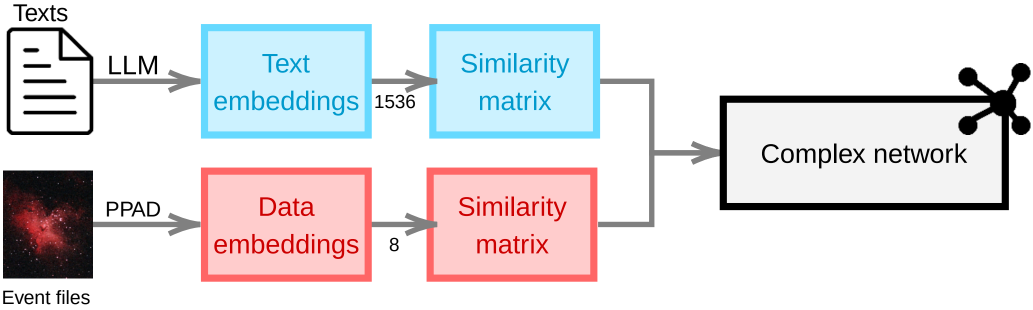
- Aligned representations show semantically narrower domain than pre-alignment
- Aligned neighbors exhibit smaller variability in spectral properties
- Anomaly detection effectively isolates unique sources (e.g., ULXs)

Name	**ObsID**	**Hard**$_{\mathrm{HS}}$	**Type**
2CXO J095550.1+694046	10542	0.92	ULX
2CXO J223940.2+751321	8588	-0.13	YSO
2CXO J201536.9+371123	11092	0.64	Cataclysmic

Focus on highly variable X-ray sources

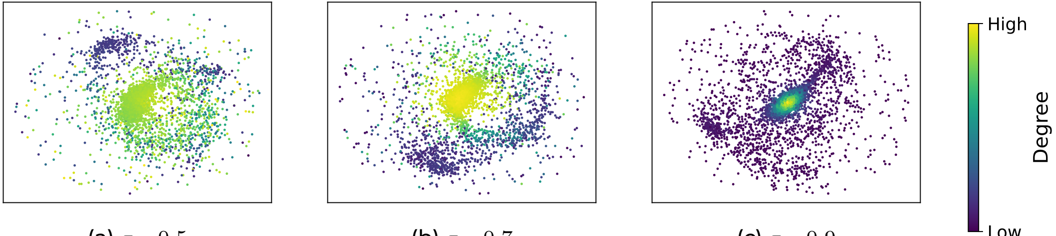
- **Focus** on observations with high variability (`var_prob_b` > 0.9)
- Information from **scientific texts** and **event files**
- **Goal**: Identification of patterns and anomalous observations, including rare objects
- **Idea**: model embeddings as complex networks (graphs with a non-trivial topology)

- Extract embeddings from both observational data and scientific texts.
- Compute similarity matrices and construct a weighted undirected graph.
- Apply different thresholds ($\tau = \{0.5, 0.7, 0.9\}$) to analyze network structure.



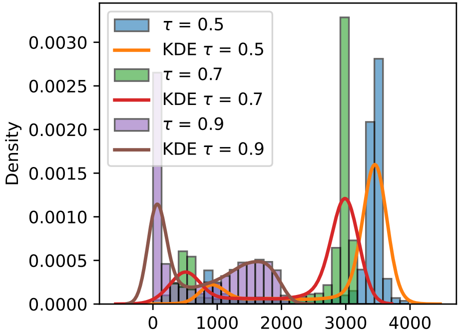
Network visualization

- **Degree**: number of edges connected with a node ($\implies$ number of observations similar to a given one)
- **Core-periphery** structure (core: similar observations; periphery: anomalies)
- Low threshold (τ) $\implies$ isolation of very anomalous observations



(a) $\tau = 0.5$ (b) $\tau = 0.7$ (c) $\tau = 0.9$

Degrees Distribution



Object type	Degree		High degree %
	Low (<1500)	High (≥ 1500)	
YSO	214	1042	83%
YSO_Candidate	34	149	81%
Star	159	571	78%
X	128	380	75%
HighMassXBin	58	80	58%
...	...	...	...
ULX	9	0	0%

Conclusions

Transformer-based autoencoders are **effective** for spectral reconstruction

Small latent vectors allow **clustering**

The **latent space** has a physical **meaning**

Conclusions

Alignment of **multimodal** data allows better regression and anomaly detection

Complex networks highlight **anomalous** observations within highly variable sources

Complex networks exhibit a **core-periphery** topology