Aligning Chandra X-ray spectra and scientific literature

Nicolò Pinciroli

POLITECNICO
MILANO 1863

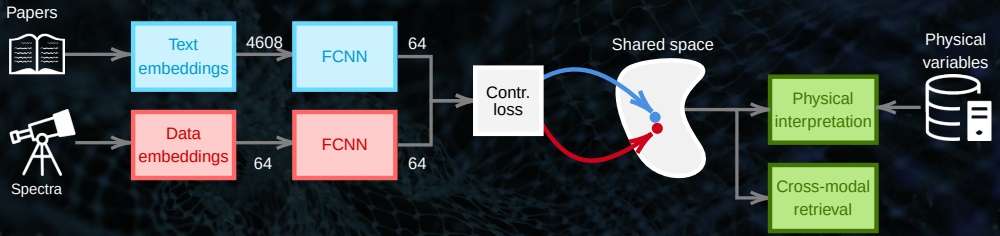
Challenges

- Rare source classes lack sufficient **labeled data** $\Longrightarrow$ unsupervised and self-supervised approaches

Challenges

- Rare source classes lack sufficient **labeled data** $\Longrightarrow$ unsupervised and self-supervised approaches
- **Scientific literature** describes astrophysical sources in natural language $\Longrightarrow$ automated cross-modal alignment between text and observations

Enhancing X-ray spectra with contrastive learning

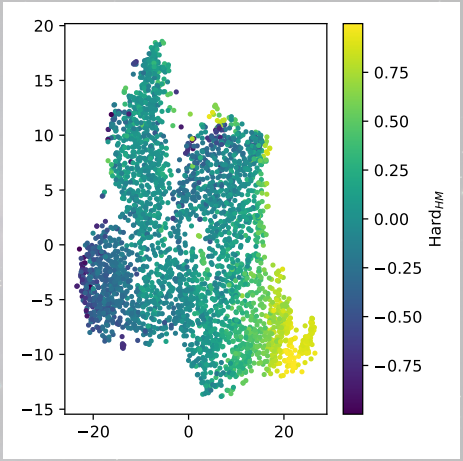


- Aligns text summaries (from literature) with X-ray spectra using InfoNCE loss
- 11,447 text-spectra pairs (1,719 test set pairs)
- The goal is to obtain a shared representational space that optimizes cross-modal retrieval

Compressing X-ray spectra with representation learning

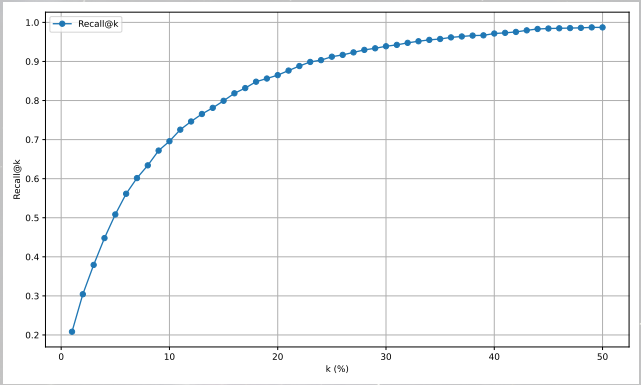
Physical explainability via regression

- Symbolic regression directly correlates embedding vectors with combinations of measurable fluxes
- Latent space accurately predicts critical Hardness Ratio parameters
- Confirms that autoencoder features are correlated with physical quantities, not noise



Enhancing X-ray spectra with contrastive learning

Contrastive retrieval

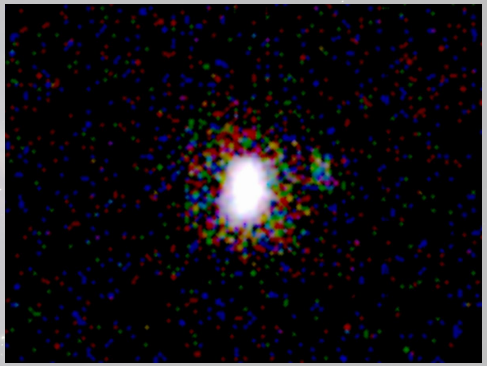


- 20% Recall@1% when pre-initialized with autoencoder data (90% Recall@23%)
- Correct textual descriptions retrieved from thousands of scientific abstracts
- Scientific texts enhance spectral data (97% reduction in latent space size)

Enhancing X-ray spectra with contrastive learning

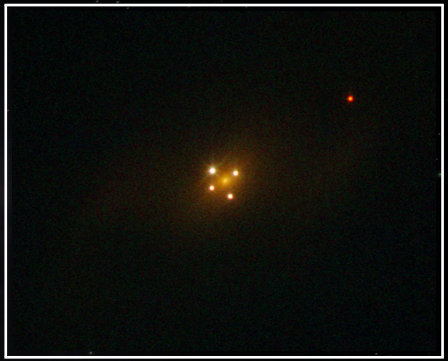
Outlier detection

- Isolation Forest on joint text/spectra embeddings identifies outliers
- Successfully flagged a known Pulsating Ultraluminous X-ray candidate
- Provides a scalable mechanism for discovering novel phenomena in petabyte surveys



Strengths

- Representation learning can transform raw data into compressed catalogs encoding physical information
- Using contrastive learning results in 97% dimensionality reduction, ensuring computational feasibility on large datasets



Quadruply-lensed quasar (ESA/Hubble)

Limitations

- LLM-generated summaries not available for rare sources $\Longrightarrow$ contrastive learning may provide descriptions of similar objects
- Contrastive learning limited to two modalities $\Longrightarrow$ apply to additional modalities (e.g., images)
- Lack of systematic outliers analysis $\Longrightarrow$ measuring deviations from *average* sources for each class

Cygnus A (Chandra)

Thank you for your attention