CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

AstroAI

The Chandra Source Catalog: a Legacy Product for Machine Learning Discovery in High Energy Astrophysics

Rafael Martínez-Galarza
Chandra X-ray Center/AstroAI

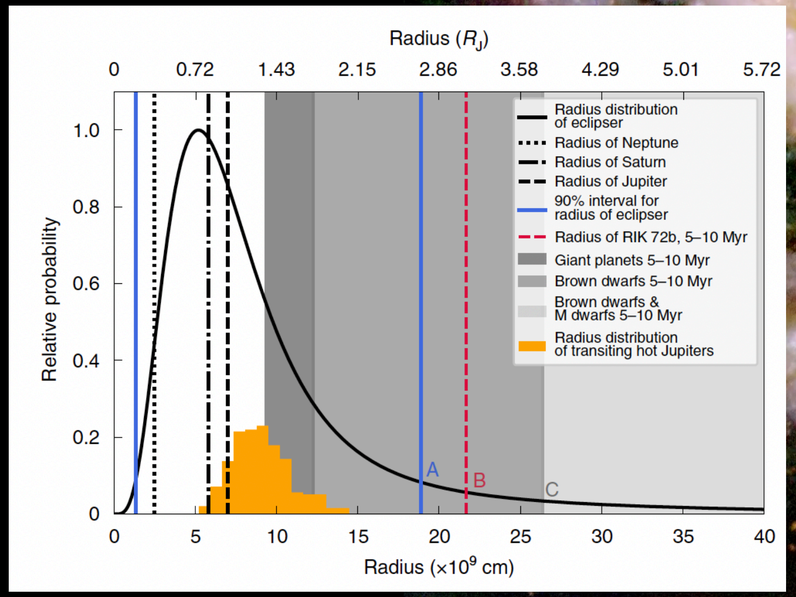
Chandra 25 Symposium, December 5, 2024

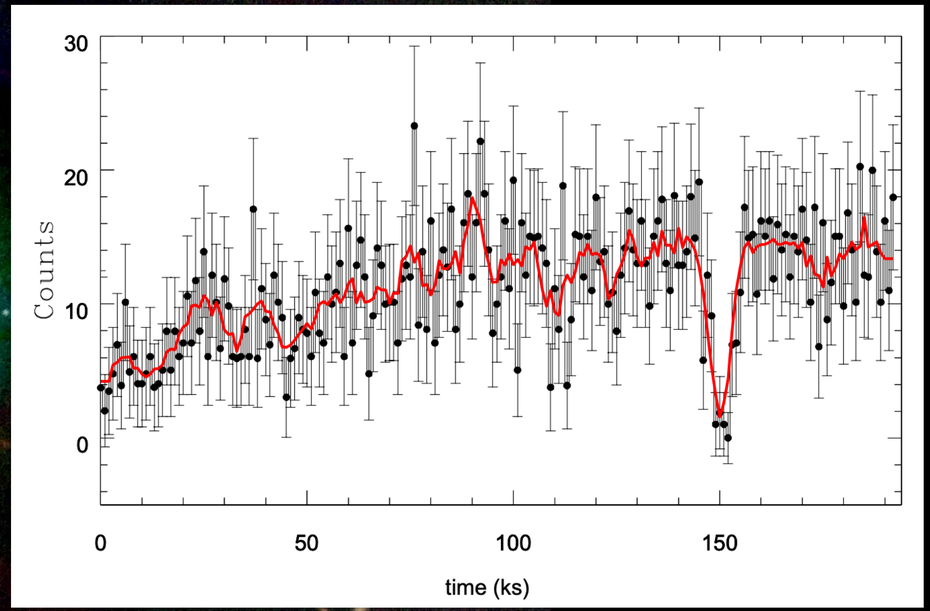
CHANDRA SOURCE CATALOG

A planet candidate in M51

Di Stefano et al.2021

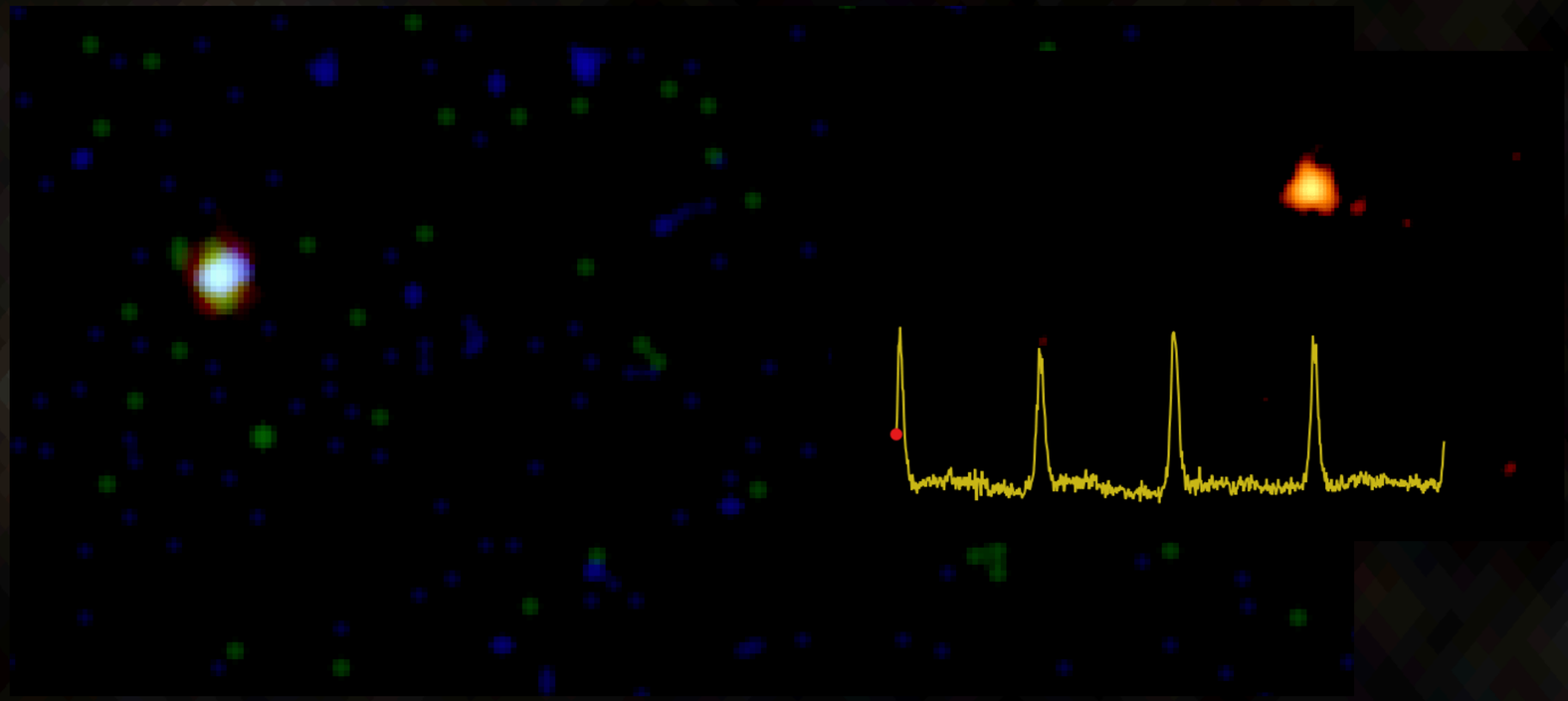
M51-ULS-1b

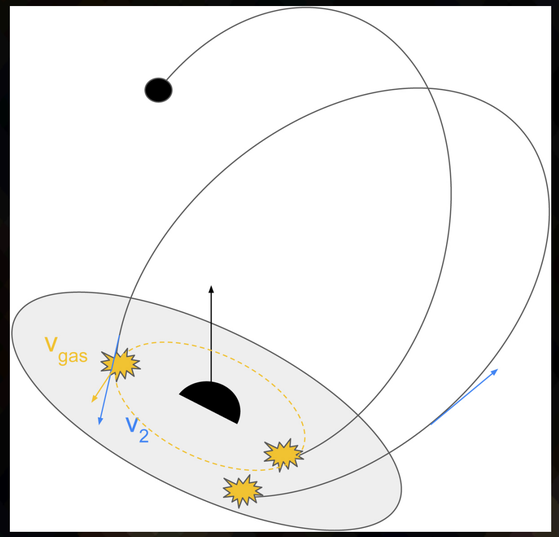




Quasi-Periodic Eruptions

Miniutti et al. 2019, Arcodia et al. 2021, Chakraborty et al. 2021





Franchini et al. 2023

- X-ray QPEs are puzzling eruptions associated to the nuclei of nearby galaxies. Their soft (thermal), repeating bursts appear associated with Tidal Disruption Events.
- The ones above happen every 9 hours. The host is a Seyfert 2 galaxy at z=0.018
- Possible explanations: Mass transfer? Gravitational Lensing? Collisions between orbiting compact object and accretion disk?

These findings have been serendipitous and have started whole new lines of research

X-ray datasets are a fertile ground for the discovery of astrophysical objects that inform models of gravitational wave emission, extrasolar planet, and the most violent explosions of the Universe.

How do we efficiently harvest X-ray catalogs to enable new science?

A Family of X-ray Catalogs

- The Chandra Source Catalog v. 2.1 contains ~408k unique sources and over 1.3 million X-ray detections at high resolution and low background, together with a **very** rich set of data products. Median flux 8.0 × 10^{-15} erg/cm/s. Spatial resolution 0.5”.

- The XMM-Newton Source Catalog v. 4XMM-DR13 contains ~657k sources, half of which have spectra and light curves. Median flux 2.2 × 10^{-14} erg/cm/s. Spatial resolution ~5”.

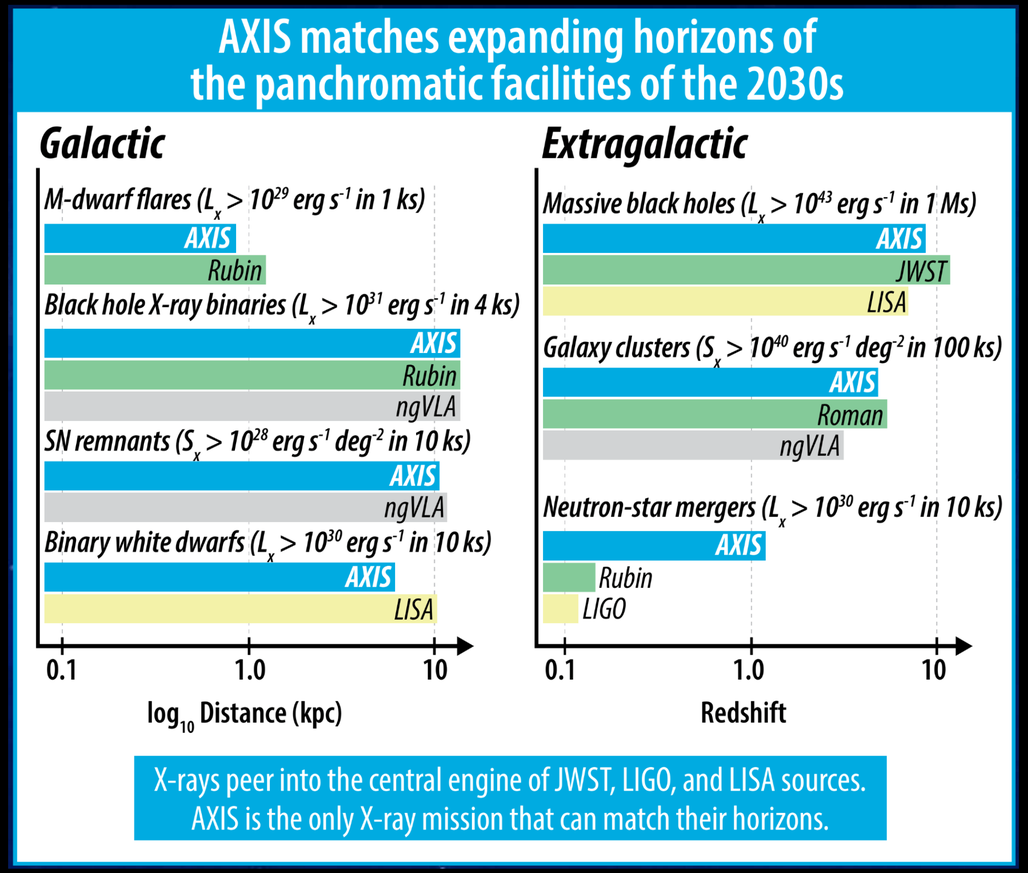
- The eROSITA eRASS catalog contains ~930k sources over a large portion over the entire sky, light curves, and spectra. Median flux 4.3 × 10^{-14} erg/cm/s. Spatial resolution ~10”.

Future facilities

ATHENA

- Compared with Chandra, AXIS will have 5 to 10 times the effective area, and better PSF performance across the field. AXIS will have fast response alert capabilities for TOOs and other transients, including BNS mergers at z ~> 1.

- The ESA-led Athena mission will operate at energies between 0.2-1.2 keV, with a resolution similar to XMM-Newton, but with a wide field imager, better collecting area, and high resolution spectroscopy capabilities.

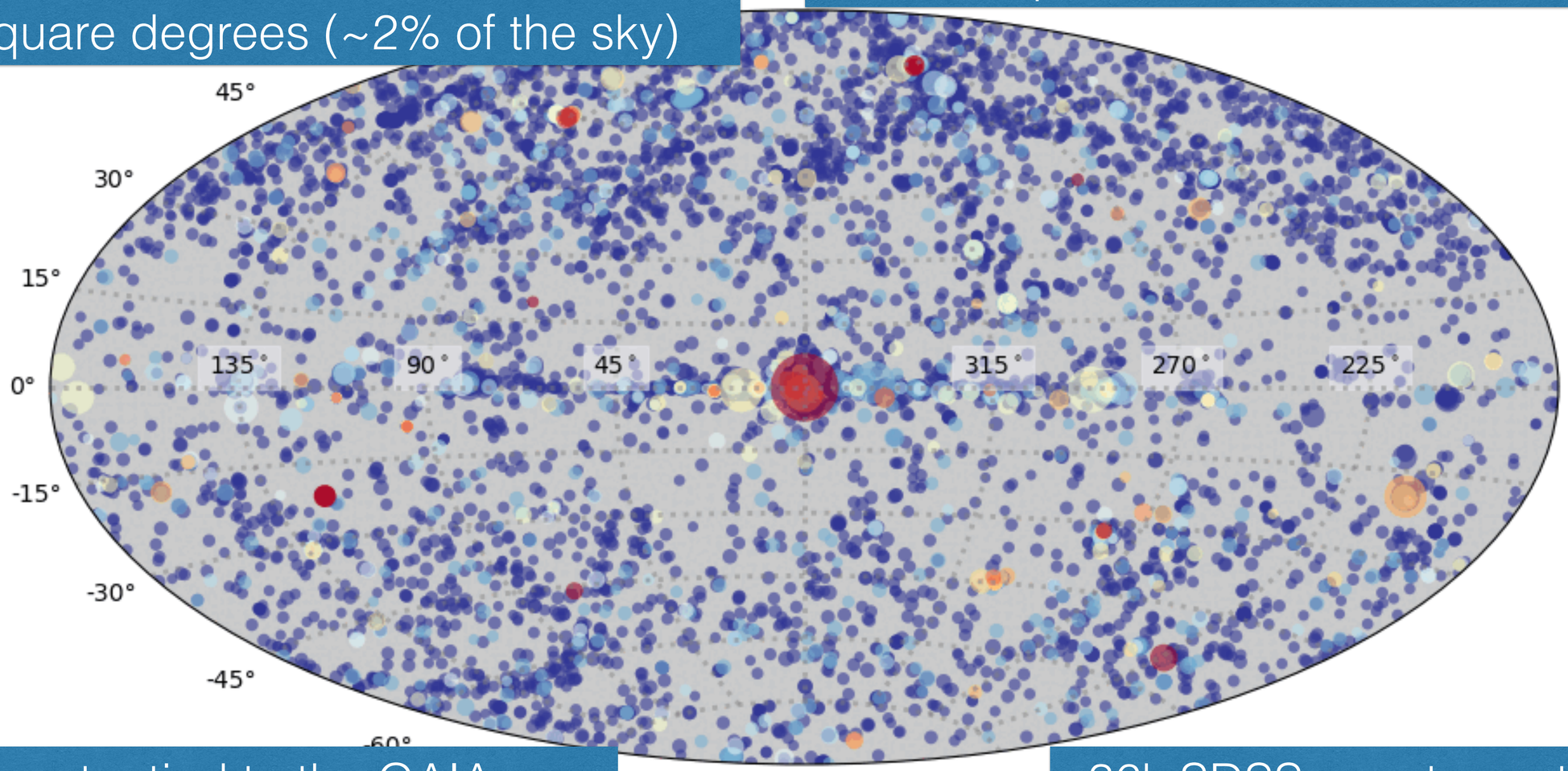


AXIS Team

The Chandra Source Catalog Version 2.1

~408k (~93k+) individual sources in the sky
~1.3 million individual detections
~730 square degrees (~2% of the sky)

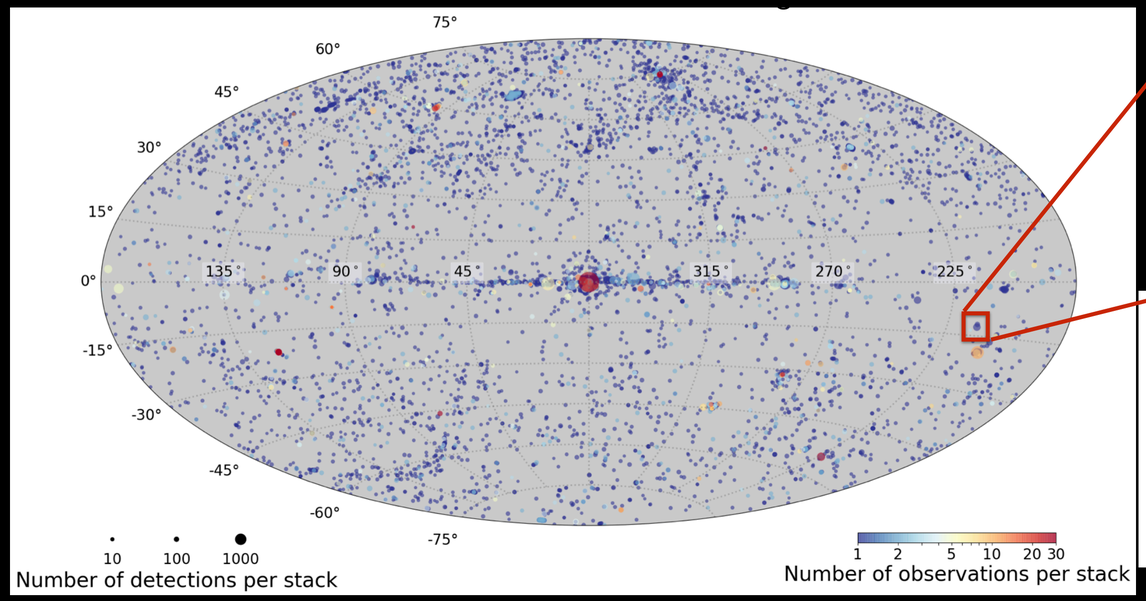
~15.5k individual Chandra Observations
Detection performed on stacked observations

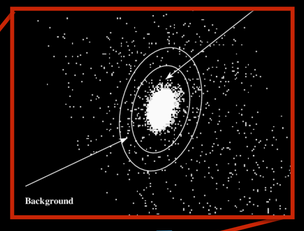


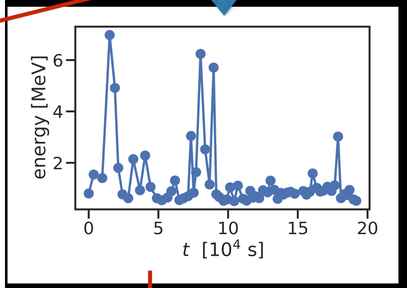
Astrometry tied to the GAIA reference frame. ε = 0.29” (95% conf.)

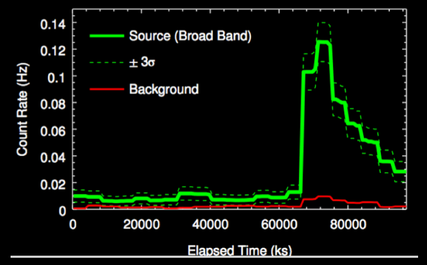
~36k SDSS counterparts provided
+ ~17k with SDSS spectra

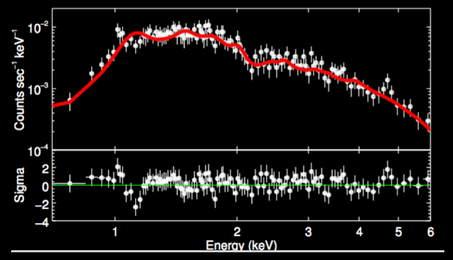
The building block of any X-ray survey are lists of individual photon detections











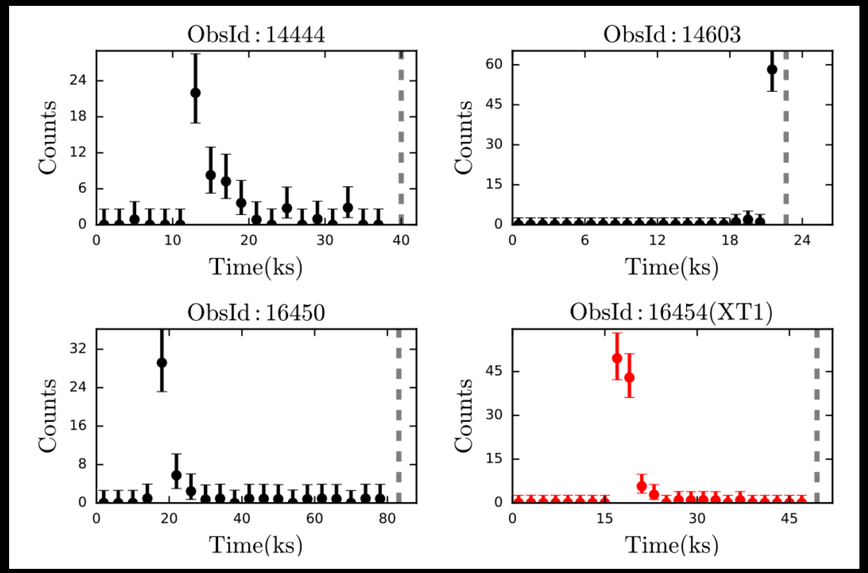
- Unlike optical data, X-ray detections are a collection of individual photon recordings of different length.
- These events effectively constitute a time series of photon energies, from which light curves and spectra are obtained.
- All relevant physical information is ultimately contained in the event list.
- No automatic alert system exists for serendipitous transients in Chandra, XMM exists.

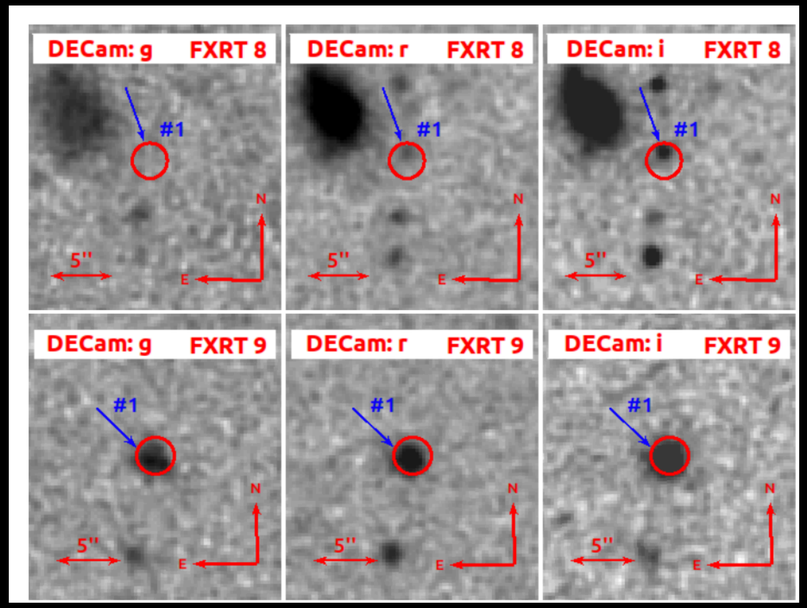
U. Of Warwick

Time-Domain and Multi-Messenger Science

Fast X-ray Transients (FXTs)

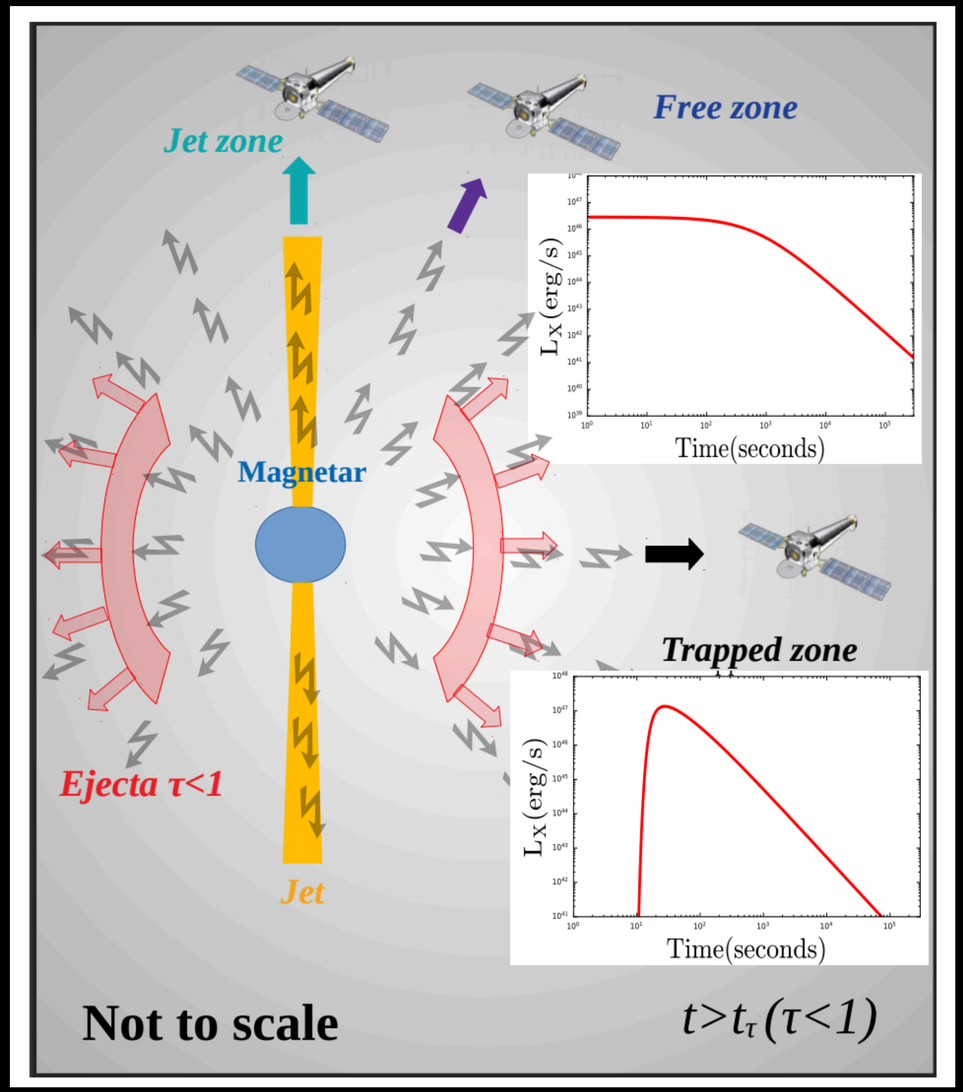
- Fast X-ray transients are rapid bursts of X-ray emission typically located in extragalactic environments.
- Spectrally soft, lasting between a few minutes to hours. Typically discovered serendipitously, days to years after the bursts. Optical counterpart are rare, but a third of the ~35 known cases are associated with host galaxies.
- What is their origin?
  - Shock breakout emission in core-collapse supernovae
  - TDE: Accretion of part of a white dwarf into an IMBH.
  - BNS mergers: fallback accretion, magnetar related?
- Dedicated search in CSC using unbind light curves. Methods are limited in time resolution.





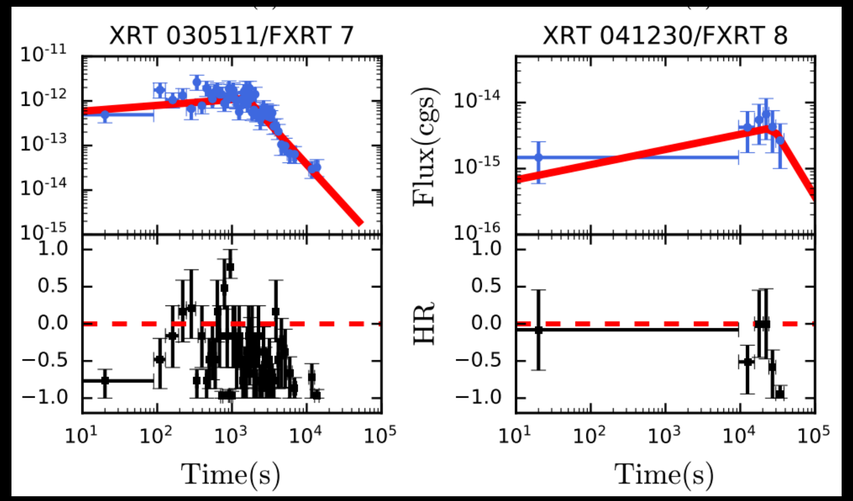
Quirola-Vásquez et al.

A possible magnetar origin for FXTs

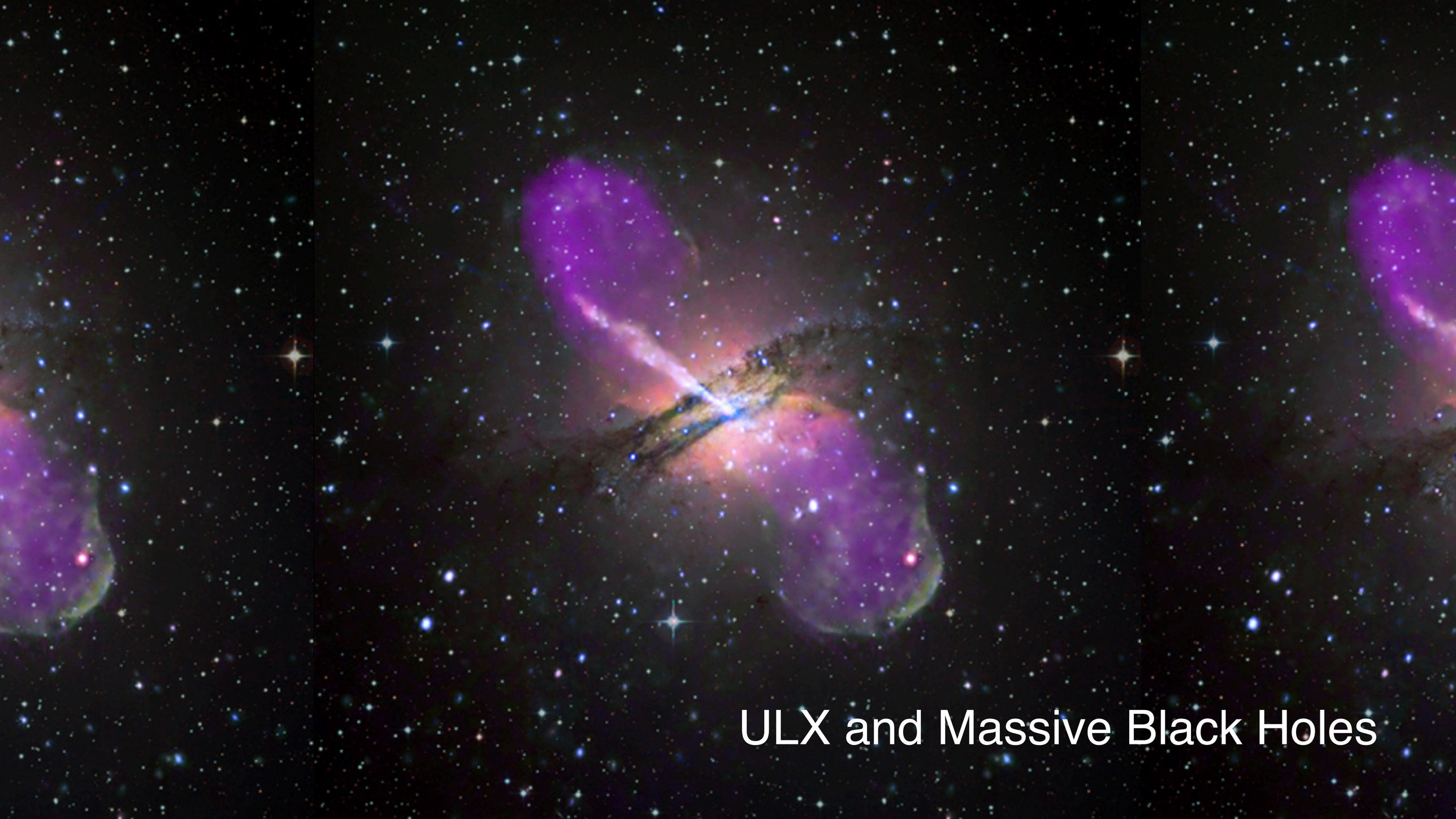


Quirola-Vásquez et al.

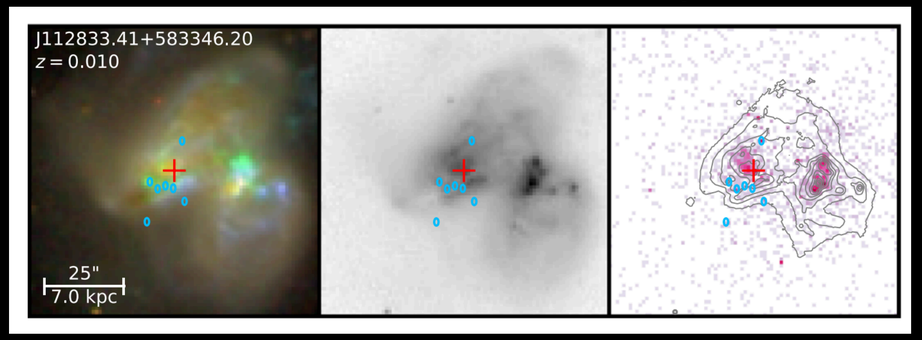
- BNS mergers resulting in a massive rapidly spinning magnetar produce X-ray afterglows with higher solid angles compared to the sGRB (Zhang 2013).
- Such afterglows can be used to probe EM counterparts to GW events that lack a γ-ray counterpart, and to search for massive millisecond magnetars.
- Light curve profile consistent with spin down luminosity of a rapidly spinning magnetar (e.g. Xue et al. 2019).

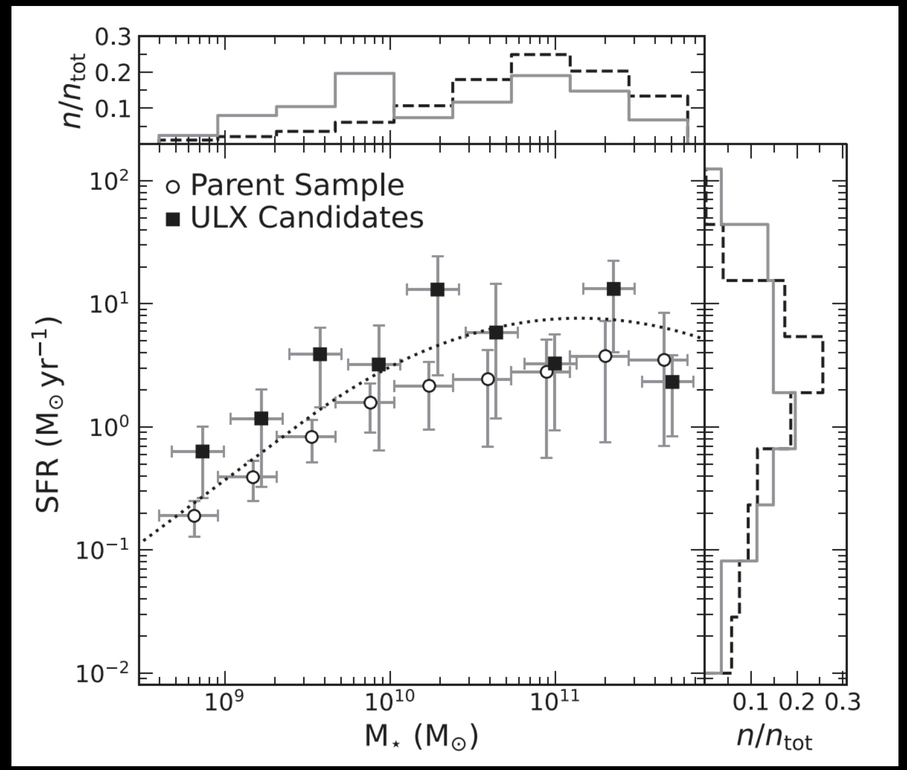


Several candidates identified in CSC dedicated searches.


ULX and Massive Black Holes

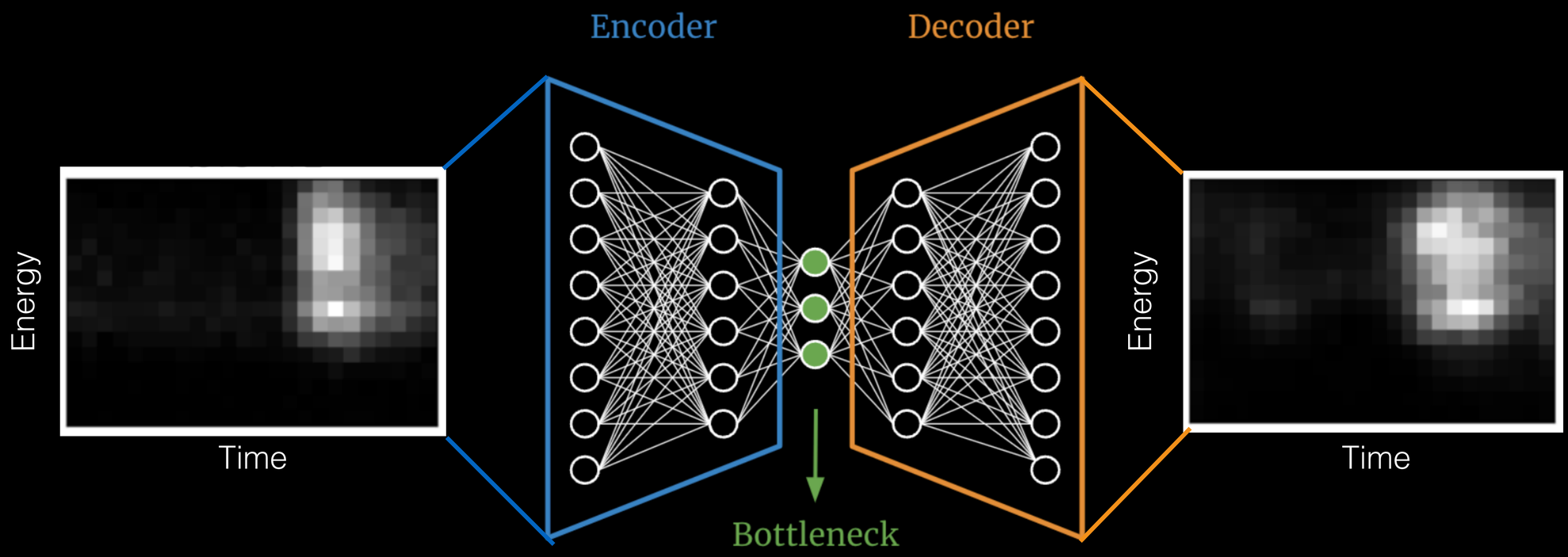
The redshift evolution of ULXs out to z~0.5 (Barrows et al. 2019, 2022)





- ULXs (Lx > 10^{39} ergs/s) and HLXs (Lx > 10^{41} ergs/s) appear to accrete above the Eddington limit.
- Uniform sample of ~260 off-nuclear ULXs with redshifts z<0.5
- Largest sample of intermediate redshift ULXs, extends to higher redshifts with respect to other catalogs.
- Systematically enhanced sSFR in ULX host galaxies compared with the parent population, suggesting an X-ray binary nature for the ULXs.
- Similar study of HLXs finds that fraction of them are consistent with IMBHs injected into galaxies through mergers (not in GCs).

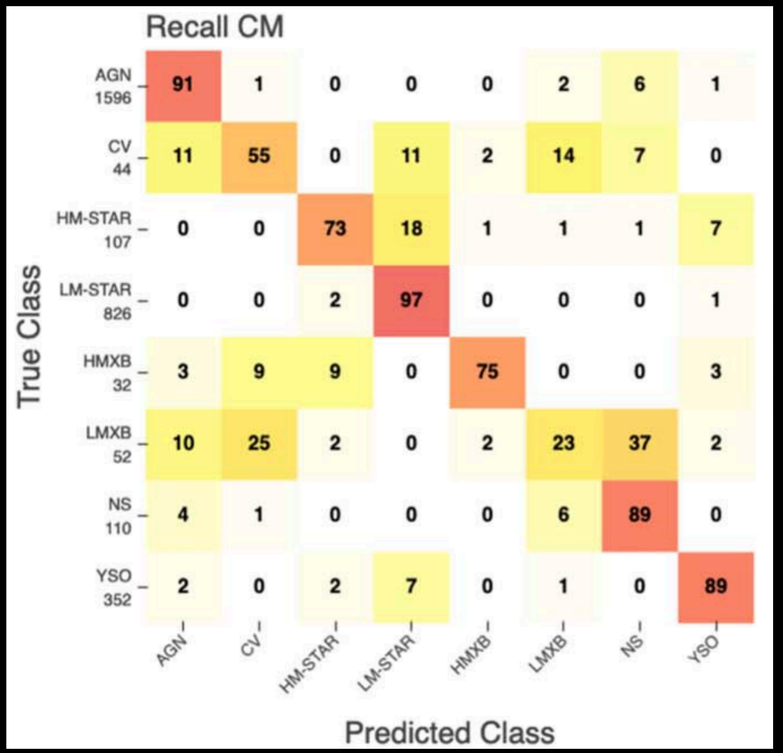
CSC enables spatially resolved studies of ULX and HLX hosts!



X-ray Datasets in the Era of Artificial Intelligence

Why machine learning in high energy astrophysics

- X-ray datasets are becoming larger and richer, but they remain unlabeled and unannotated. The vast majority of sources remain unclassified.
- Astrophysical anomalies of relevance for gravitational wave science, binary evolution, and galaxy assembly continue being identified in archival datasets, mostly serendipitously.
- Parameter inference in the presence of instrumental effects, such as pileup, remain challenges. Data-driven approach offer a way forward.



See research by the Kargaltsev group at GWU:
Yang et al. 2022, 2024, Chen et al. 2023, 2024,
See talk by Jeremy Hare

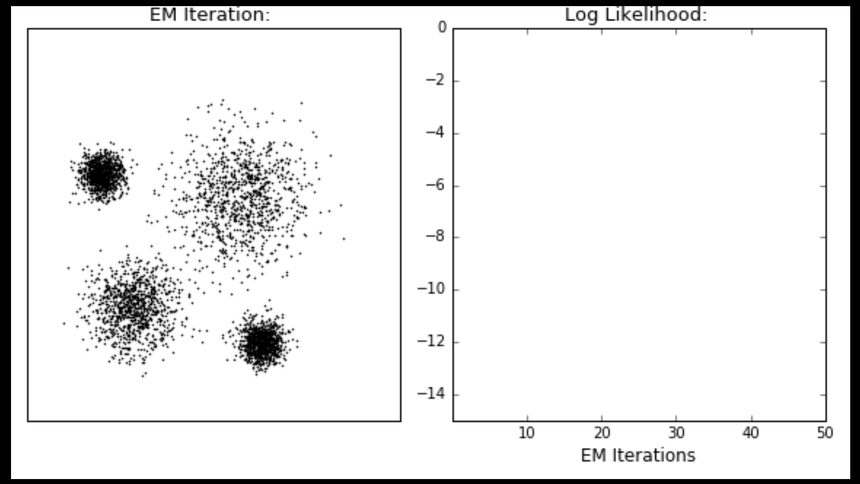
Unsupersvised Classification using Gaussian Mixtures

Pérez-Díaz et al.
2024MNRAS.528.4852P

CSC Properties:

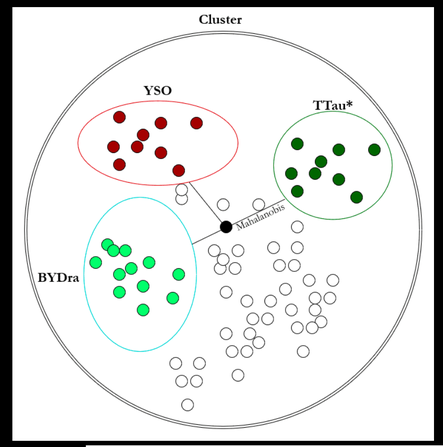
- Hardness ratios
- Variability
- Fluxes

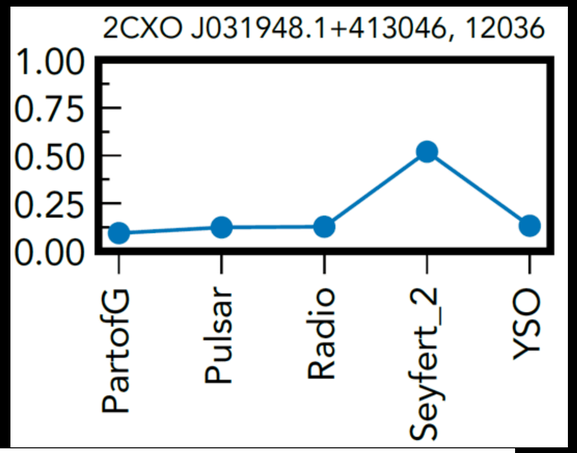
Published Catalog of >15k probabilistic Classifications

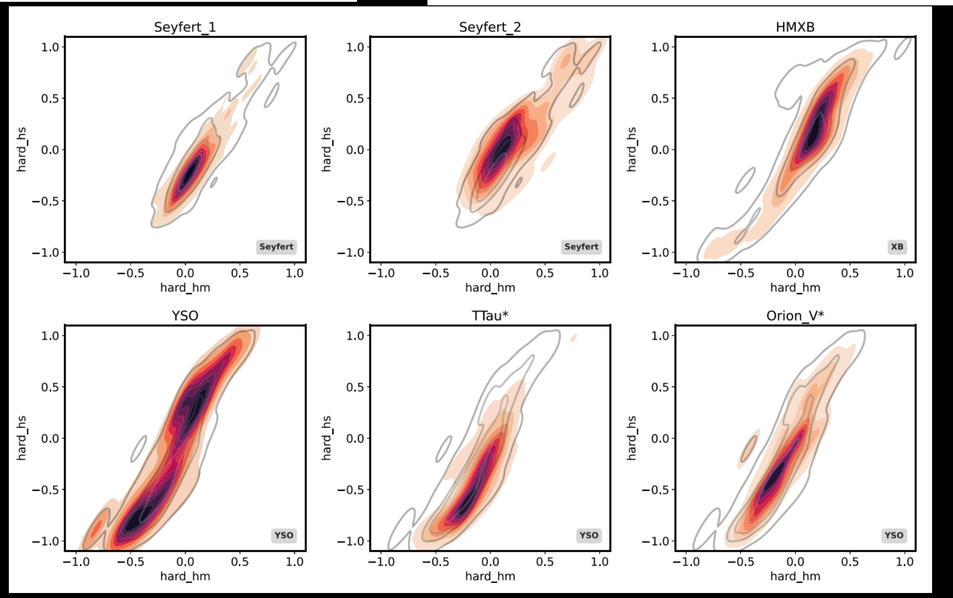


CLUSTERS

By crossmatching the clusters with existing catalogs of independent classifications, we assign probabilistic classes based on distance





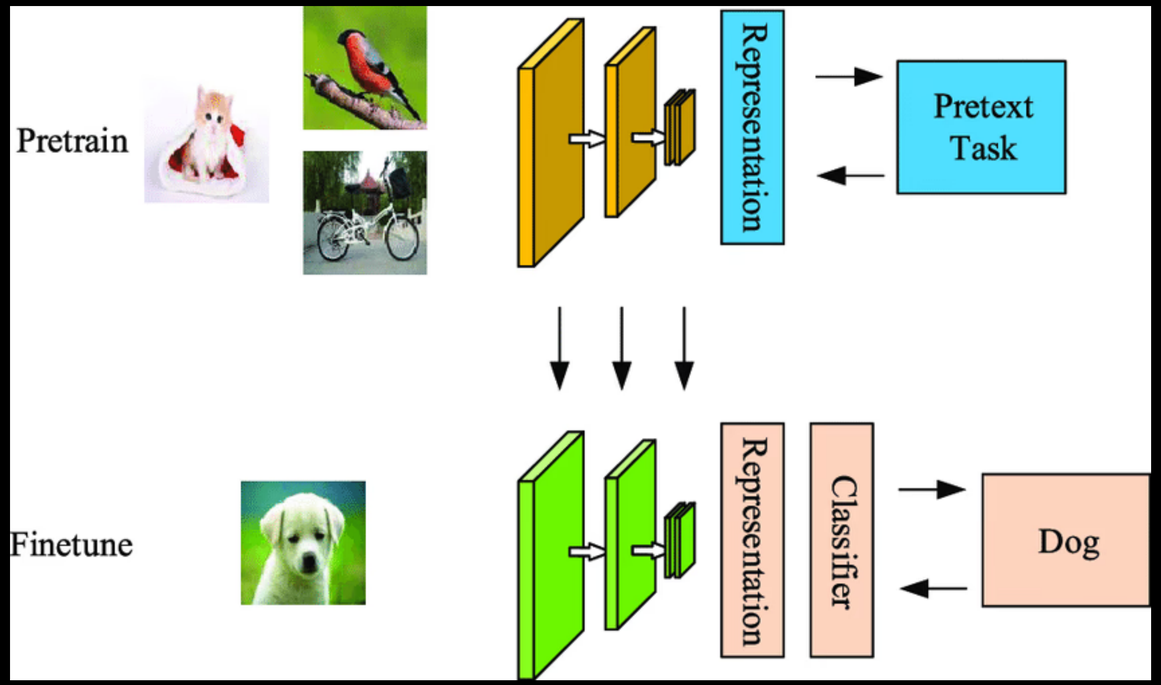


https://umlcaxs-playground.streamlit.app/

See also Yang et al. 2023

Self-supervised (representation) learning

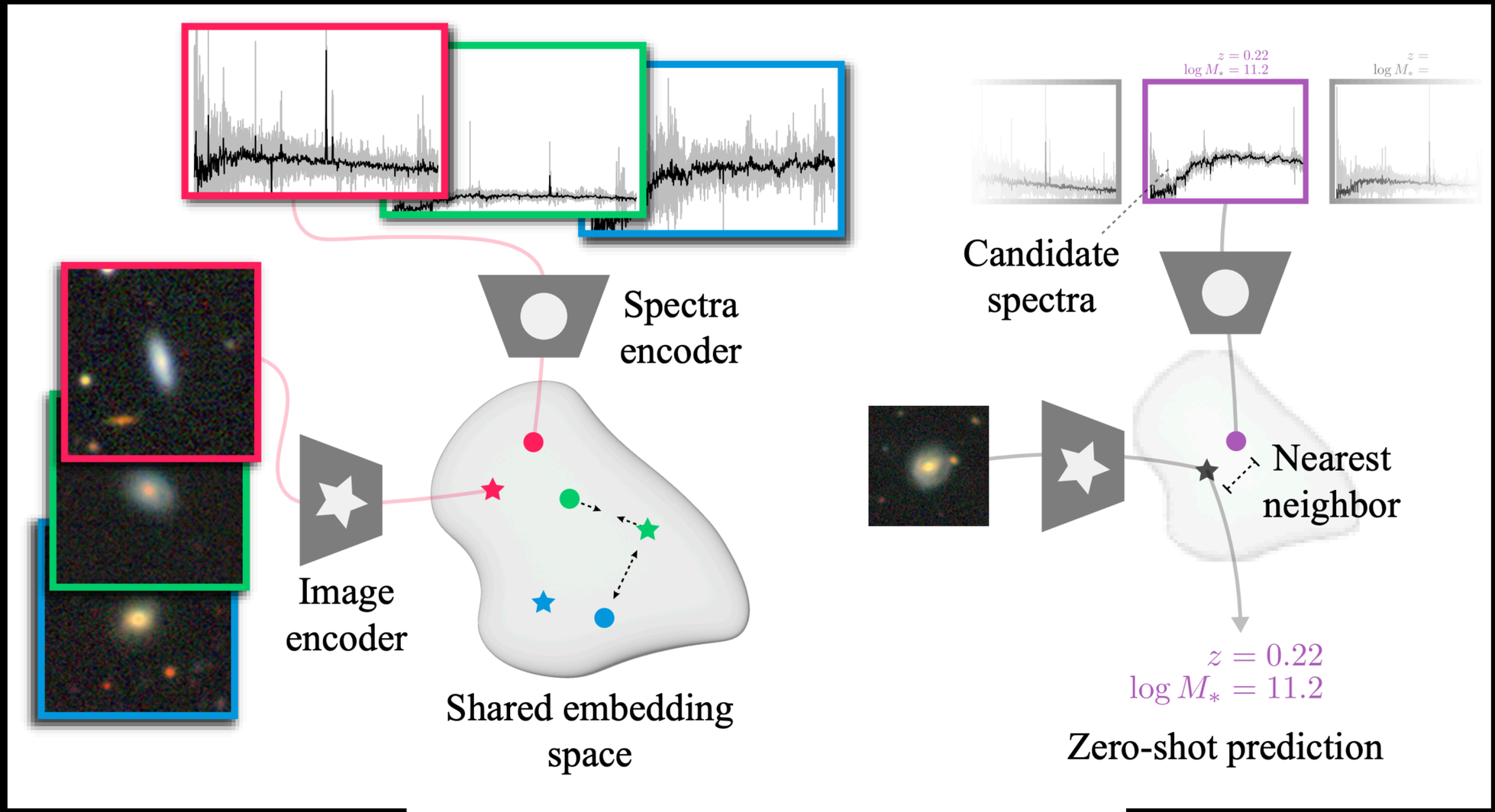
- Self-supervised learning is a type of machine learning where the system learns representations of data without the need for explicit human-provided labels.
- Instead, labels are generated by the system to solve a “pretext” task.
- If trained on lots of data, the representations can be used for downstream tasks in which they were not trained (e.g. Classification)



Rohit Kundu

I grew up in France, I speak fluent…French

Contrastive Learning: AstroCLIP

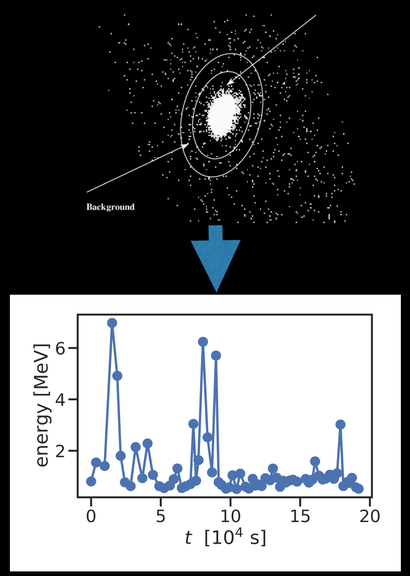


$$\mathcal{L}_{InfoNCE}(\mathbf{X}, \mathbf{Y}) = -\frac{1}{K}\sum_{i=1}^{K} \log \frac{\exp(S_C(\mathbf{x}_i, \mathbf{y}_i)/\tau)}{\sum_j^K \exp(S_C(\mathbf{x}_i, \mathbf{y}_j)/\tau)}$$

Parker, Lanusse, et al. 2024

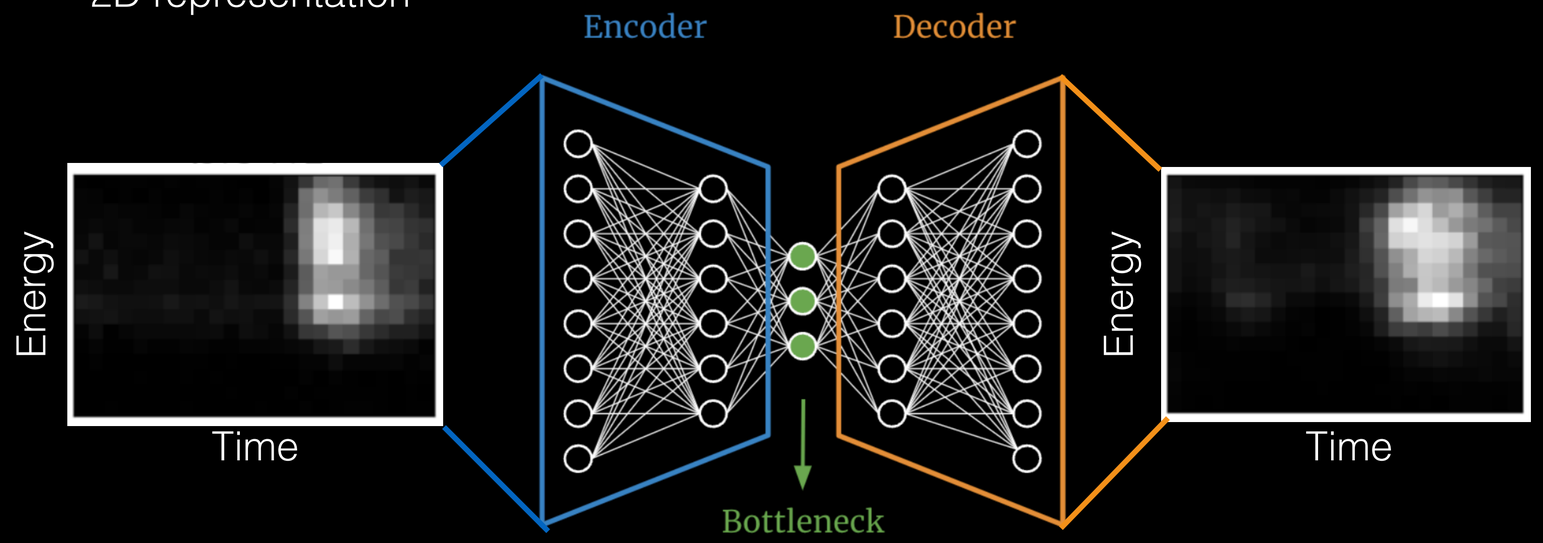
Representation Learning in X-ray Datasets

Input: events



Pre-processing:
2D representation

Architecture:

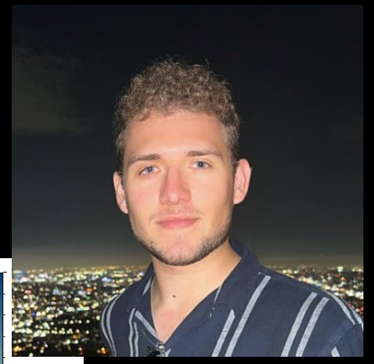


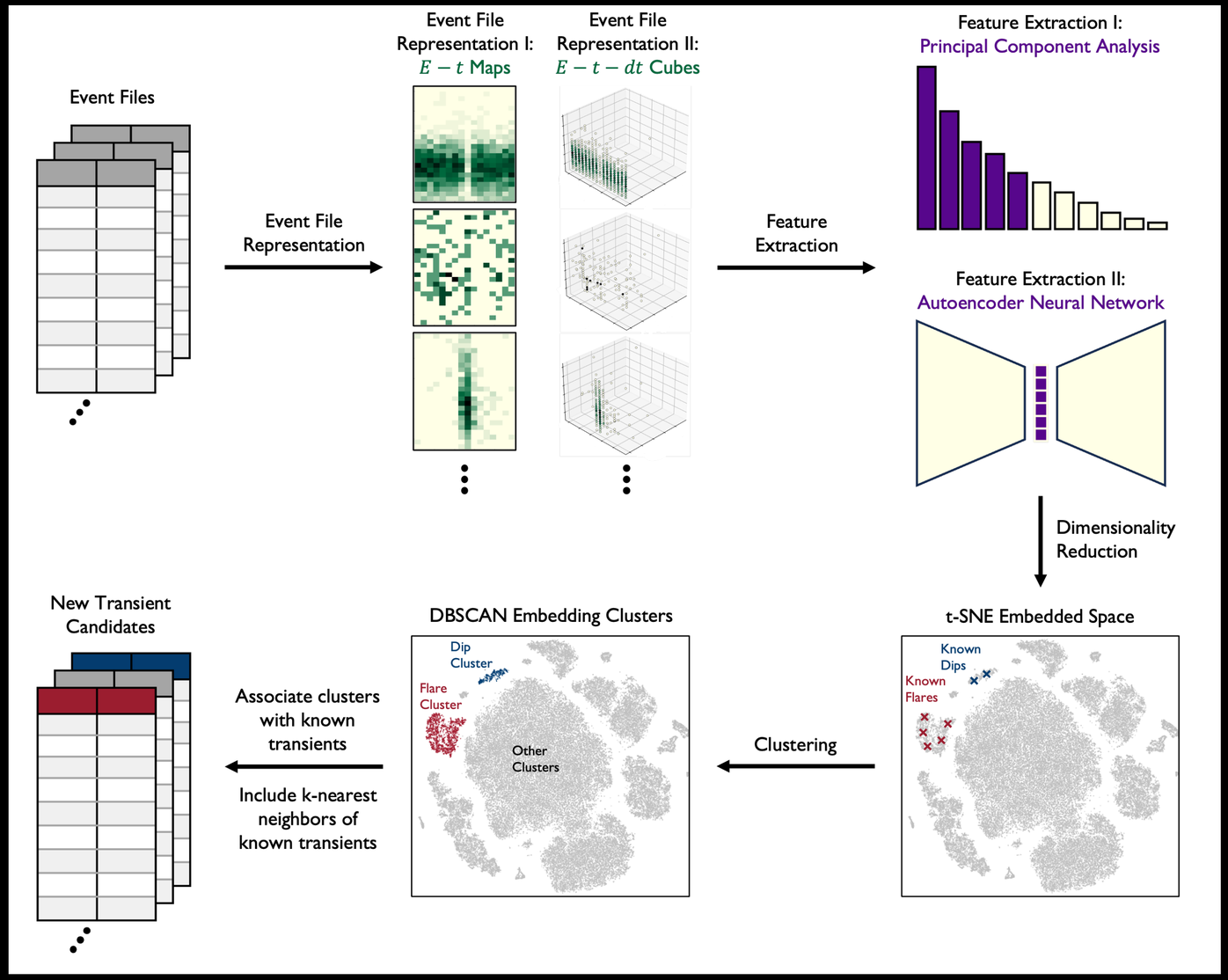
Pretext task:
Image reconstruction

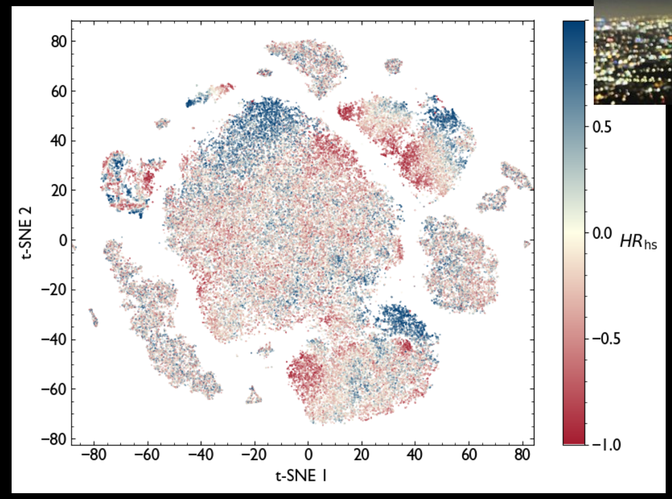
- Trained with all CSC detections with S/N ratio larger than 5.
- Size of the latent space: 24
- Spectral/spatial resolution optimized over all examples, but same for all examples

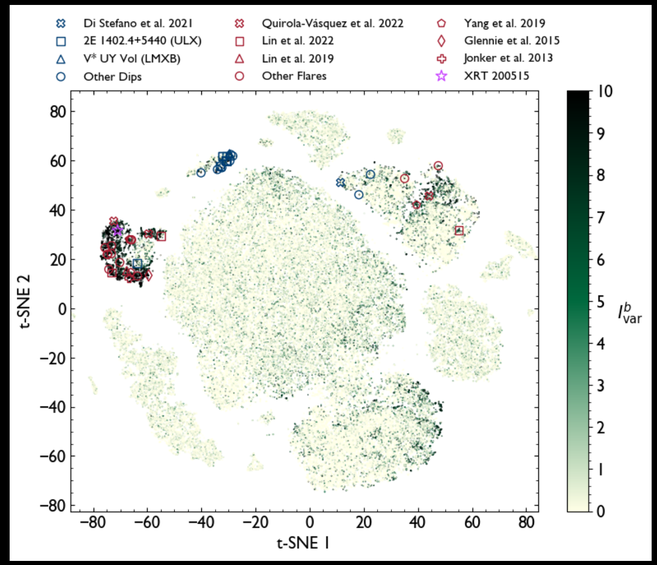
Dillmann et al., submitted

Representation Learning in X-ray Datasets







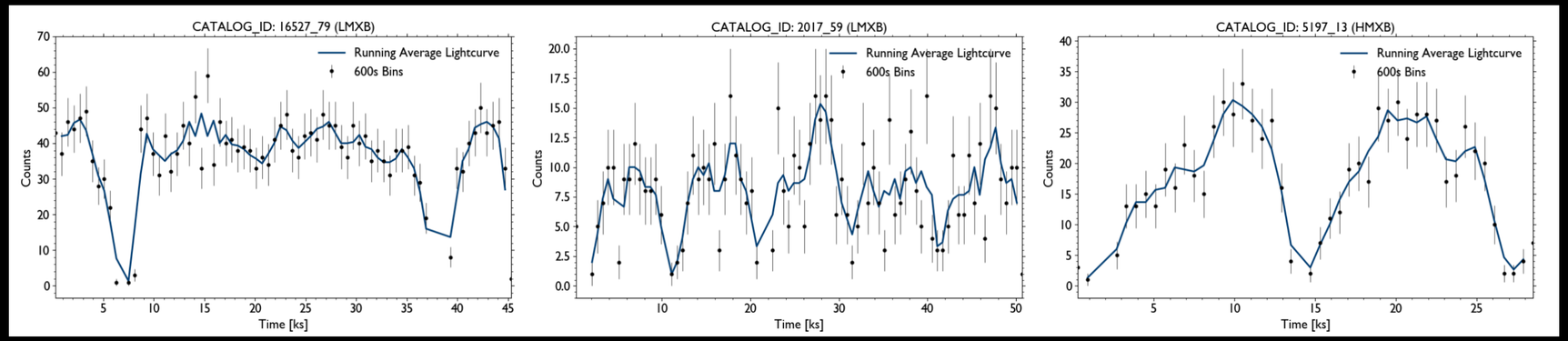


- Catalog of ~3000 dip and flare candidates in CSC

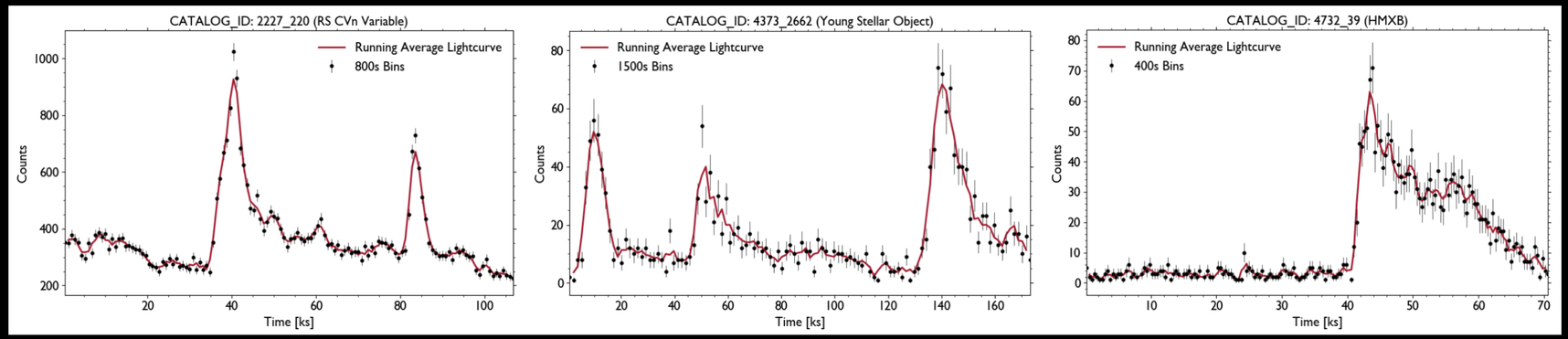
Dillmann et al., https://arxiv.org/abs/2412.01150

Examples of catalog objets

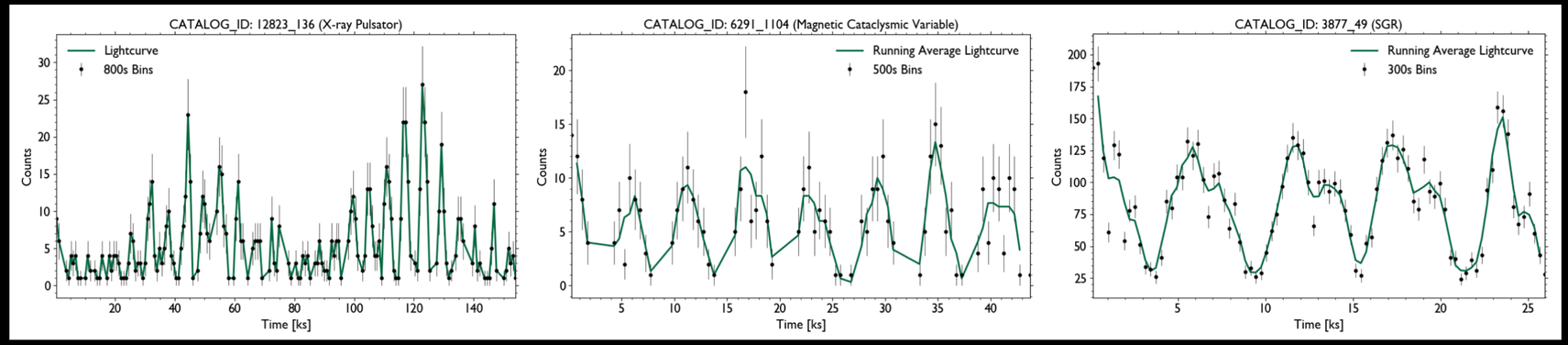
Dips



Flares

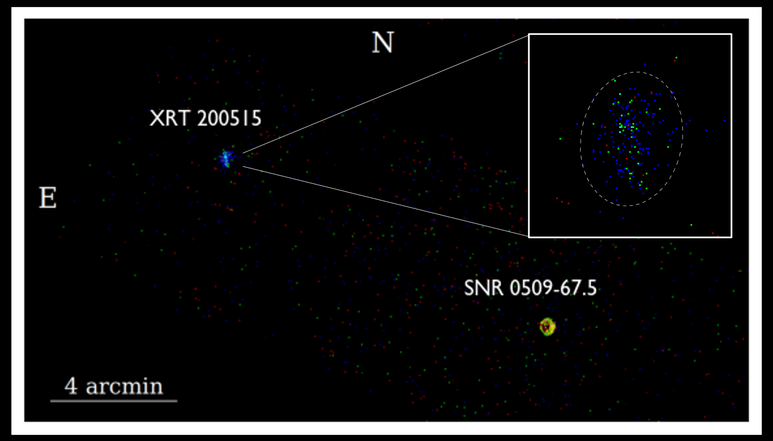


Periodic

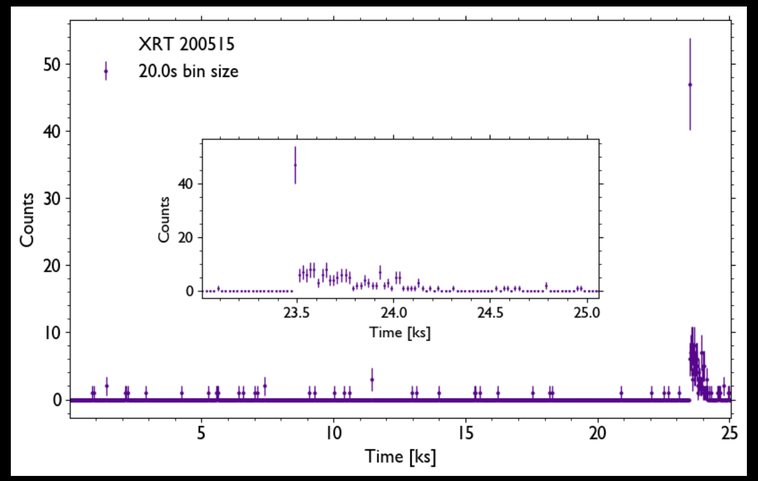


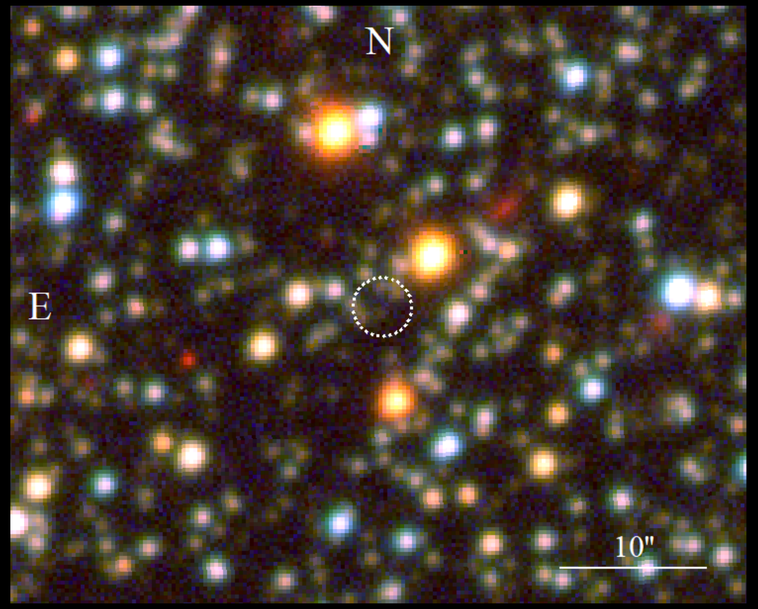
Dillmann et al., submitted

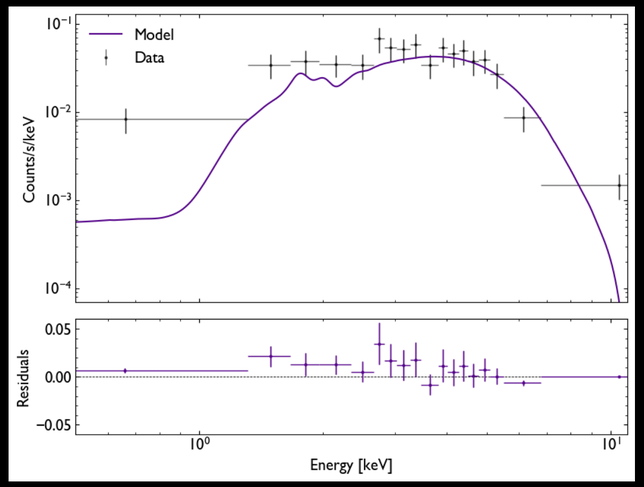
X200515: A Fast X-ray Transient in the LMC



- Fast (~10s) rise of the flux, followed by long tail with spectral softening after the burst.
- Peak luminosity: Lx ~ 2×10^{38} erg/s
- Potential counterparts are consistent with the old stellar population in this region of the LMC.
- Harder that other FXT reported, no plateau.





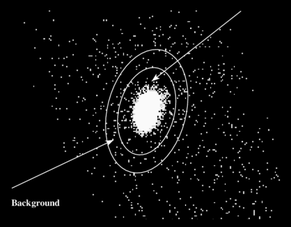


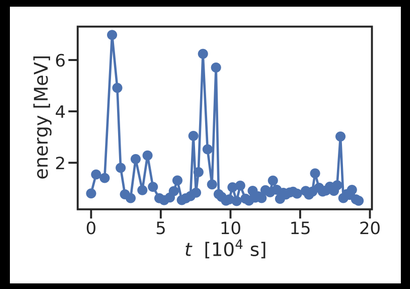
- Too fast to be consistent with a stellar flare.
- No known flares in the Milky Way halo, that resemble XRT 200515.
- Observed time scale and hard spectrum suggest connection to magnetic fields or relativistic jets.
- Giant Magnetar Flare? Oscillating tale. GRB from more distant merger?

Dillmann et al., submitted

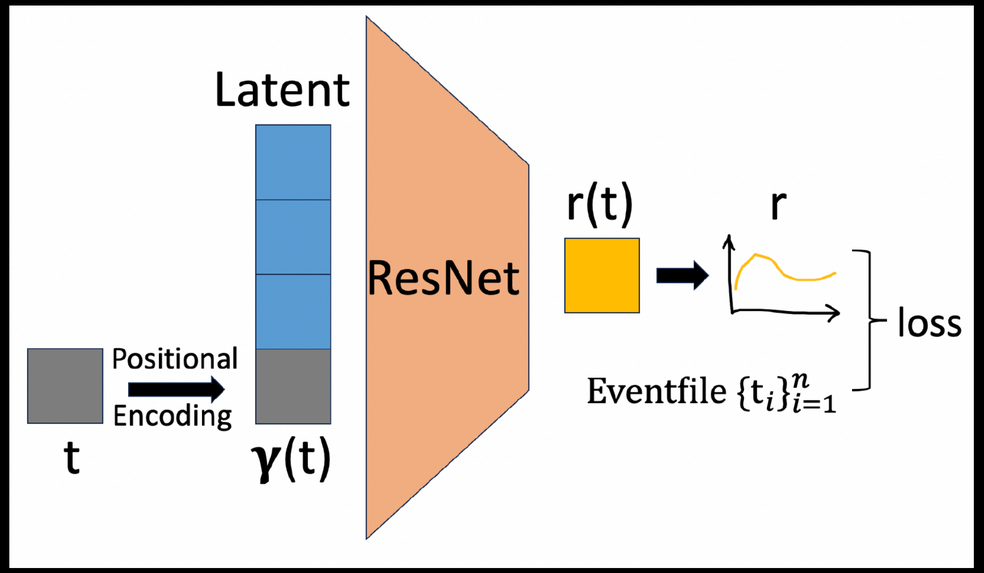
Accounting for every photon: a Poisson Process Autodecoder

Input: events

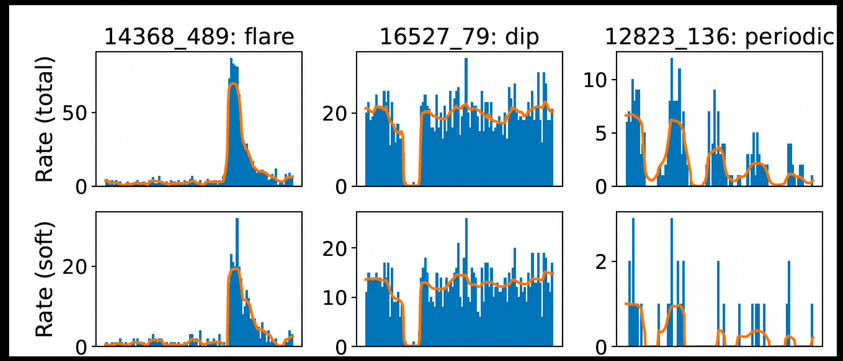




Architecture: PPAD



Pretext task:
Poisson r(t) reconstruction



- We train to maximize the likelihood of a set of events, given the Poisson rate r(t)
- Positional encoding helps for faster convergence
- At training, both ResNet weights and latent vectors are optimized. At inference, ResNet is frozen.
- Light curve can be reconstructed at any resolution, for given energy bands.

Loss function

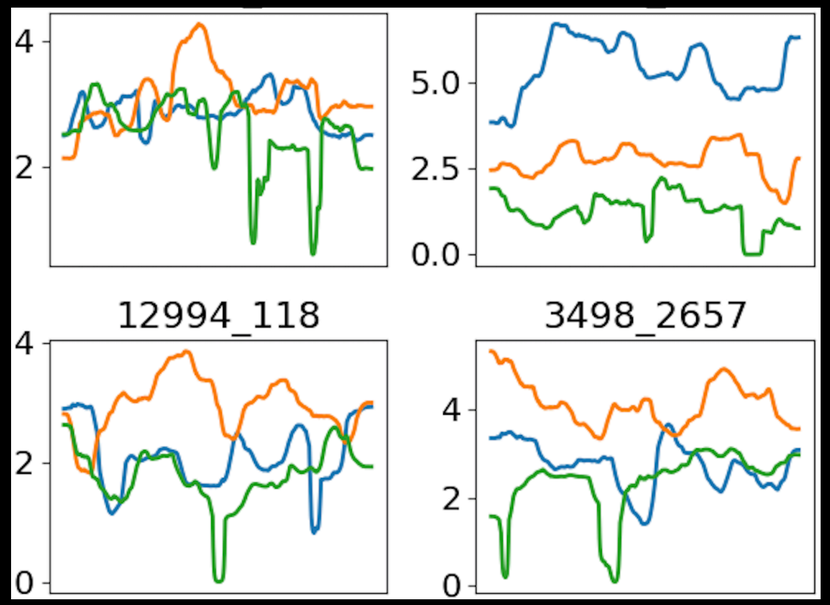
Given the output neural field r and an eventfile $\{t_i\}_{i=1}^n$,

$$L(r_e; \{t_{e,i}\}_{i=1}^{n_e}) := \underbrace{-\sum_{i=1}^{n_e} \log r(t_{e,i}) + \int_0^T r_e(t)dt}_{\text{negative log likelihood}} + \underbrace{\lambda_{TV} \cdot \sum_{i=1}^{n-1} |r_e(t_{e,i}) - r_e(t_{e,i-1})|}_{\text{smoothness penalty}}$$

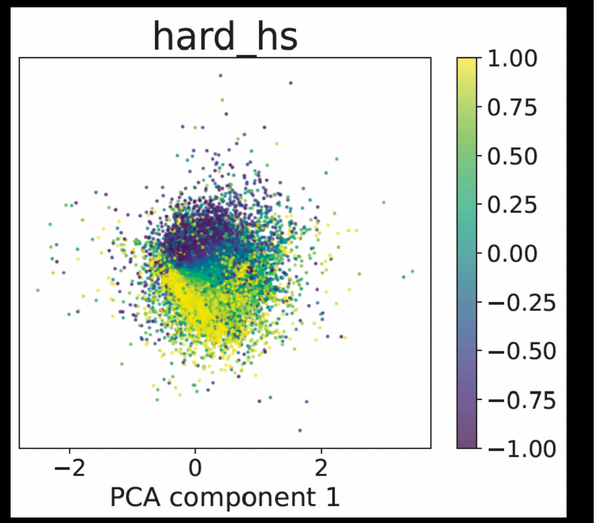
Yang et al., accepted to NeurIPS

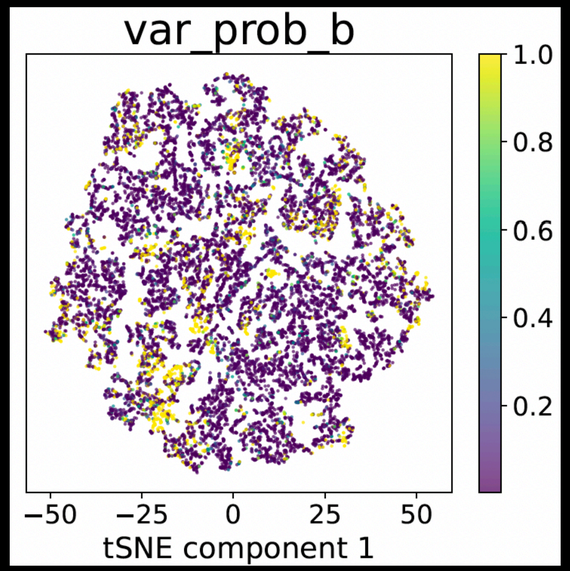
Testing on unrelated downstream tasks

- We test the representation power of our self-learned latent features by using them as the input for regression and classification.
- Performance in classification is comparable with multi-wavelength supervised approaches.
- Nearest neighbor anomaly search:



Regression Traget	MSE	R^2
hard_ms	0.02	0.87
hard_hm	0.01	0.88
hard_hs	0.02	0.93
Classification Target	**Accuracy**	**F1 Score**
var_index_b > 0.5?	0.92	0.63
source type	0.62	0.25
YSO vs AGN	0.75	0.70





Song et al., accepted to NeurIPS

Inferring physical parameters while accounting for instrumental effects

J. Yang et al., in prep

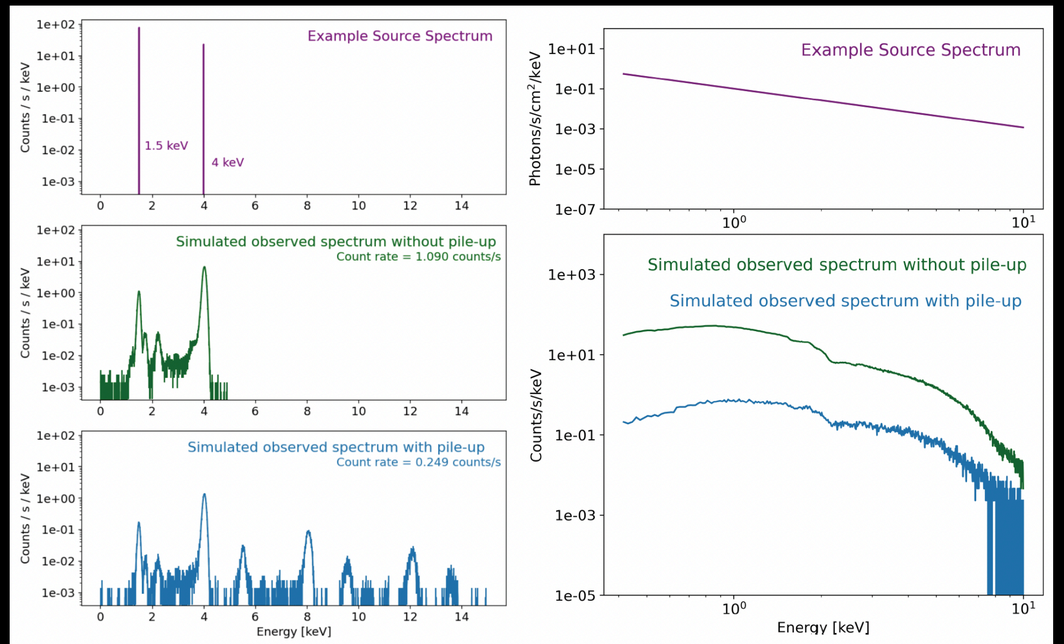
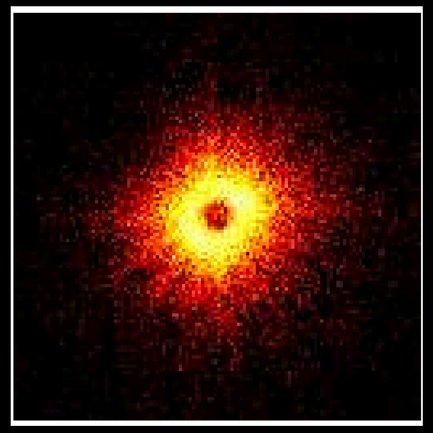


Fig. 1: simulations of how *Chandra*/ACIS-S would observe different source spectra with no pile-up vs. with a large amount of pile-up. Photons of low energies are observed as events at higher energy.

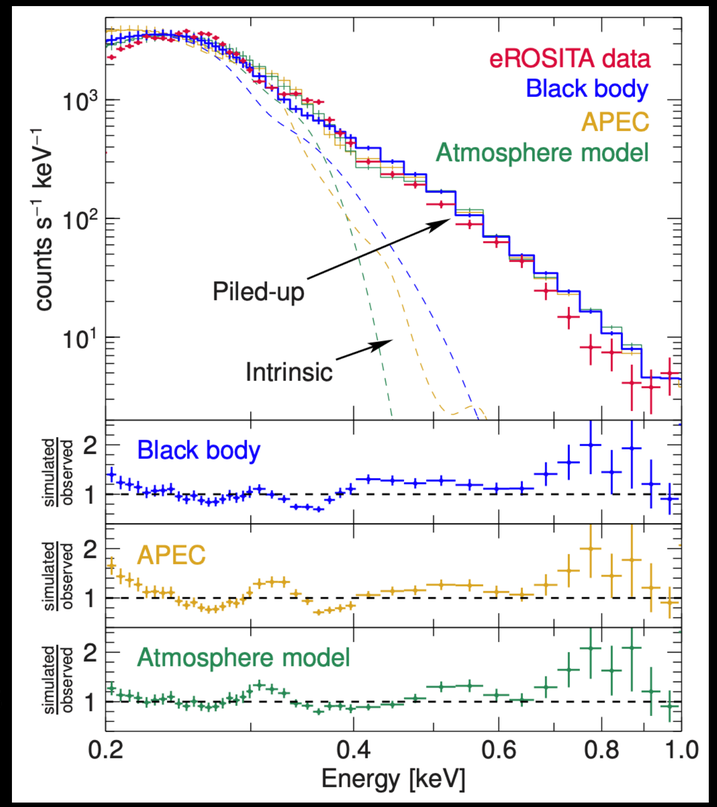
When 2 or more photons strike a detector in nearby pixels and within a single readout time, they are interpreted as a single “piled-up” event with aggregated properties

YZ Reticuli - eROSITA



- 1000 counts pers second. Severely piled-up.
- Pile-up correction depends on instrument simulation.
- Pile-up itself strongly depends on the actual spectra shape and charge cloud. Systematic error is 10%-20%/

König et al.2022, Nature



Non-linear detector effects will remain an issue in the AXIS and Athena era

Inferring physical parameters while accounting for instrumental effects

J. Yang et al., in prep

A neural network can emulate simulated pile-up in relativistic reflection spectra

Incident spectrum with instrumental response

Chandra-specific simulator, MARX, including pile-up

Neural Network
3 hidden layers
256 neurons each
ReLU activations
dropout(p=0.1)

Training data

Neural network prediction for piled-up counts/s spectrum

Minimizing a loss function
(MSE of normalized counts/s spectra)
adjusts the neural network parameters

Trained On ~22,000 simulations of reflexion models

García & Dauser's relxill model (2014)

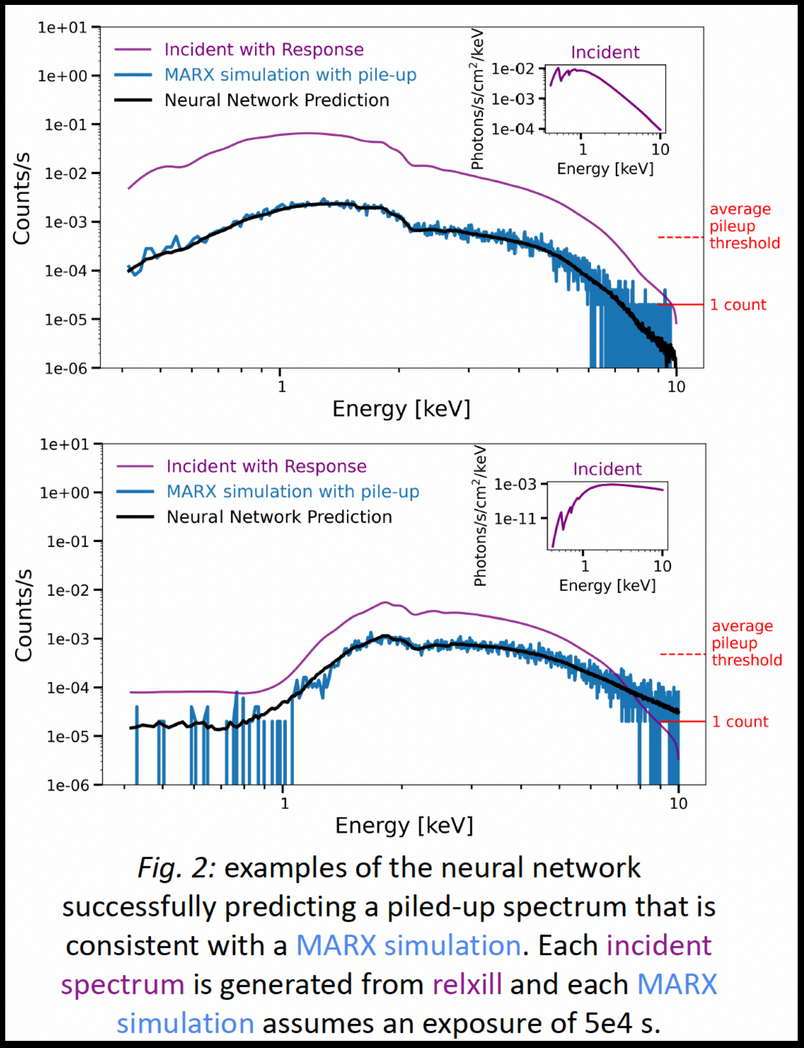
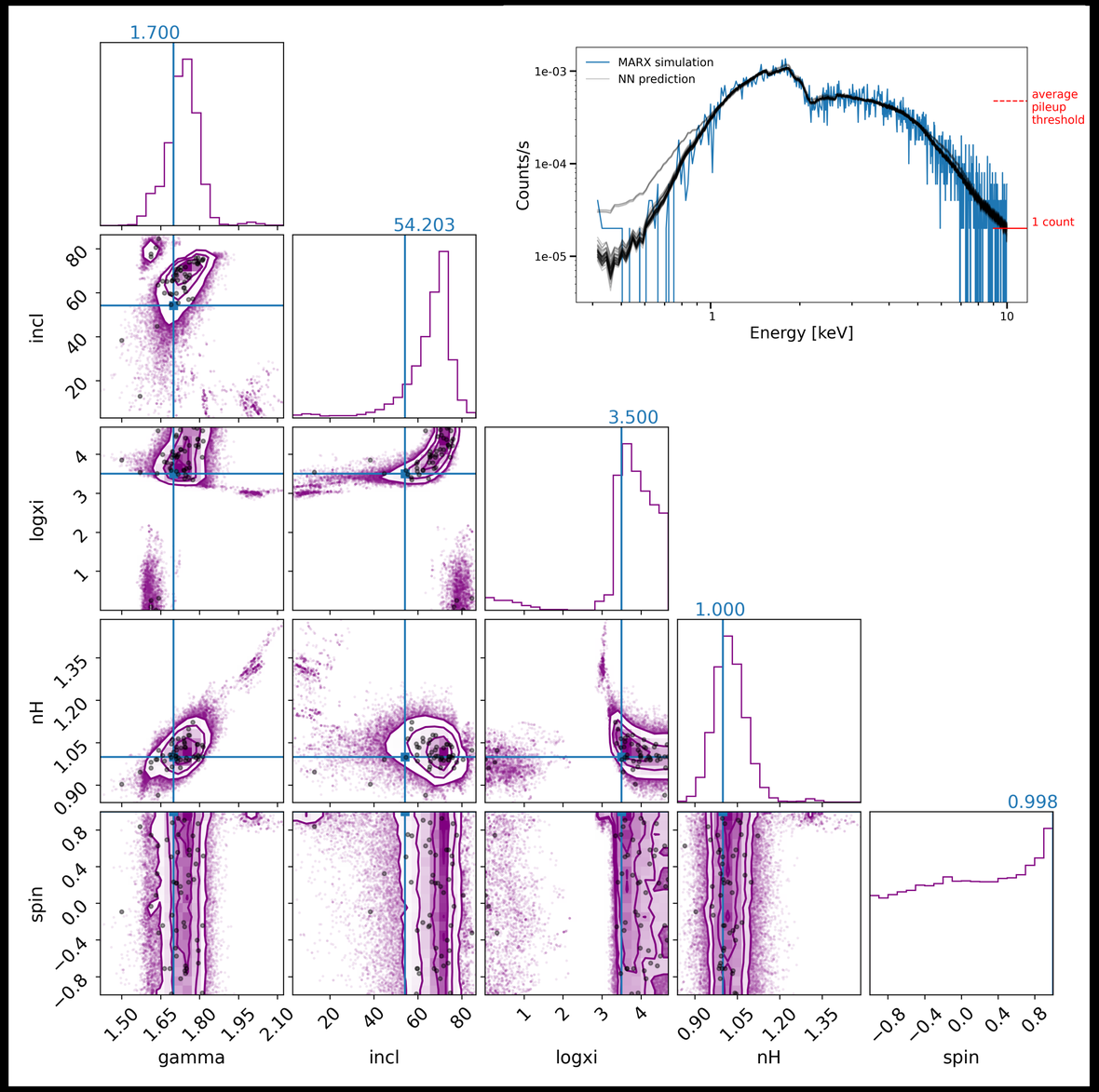
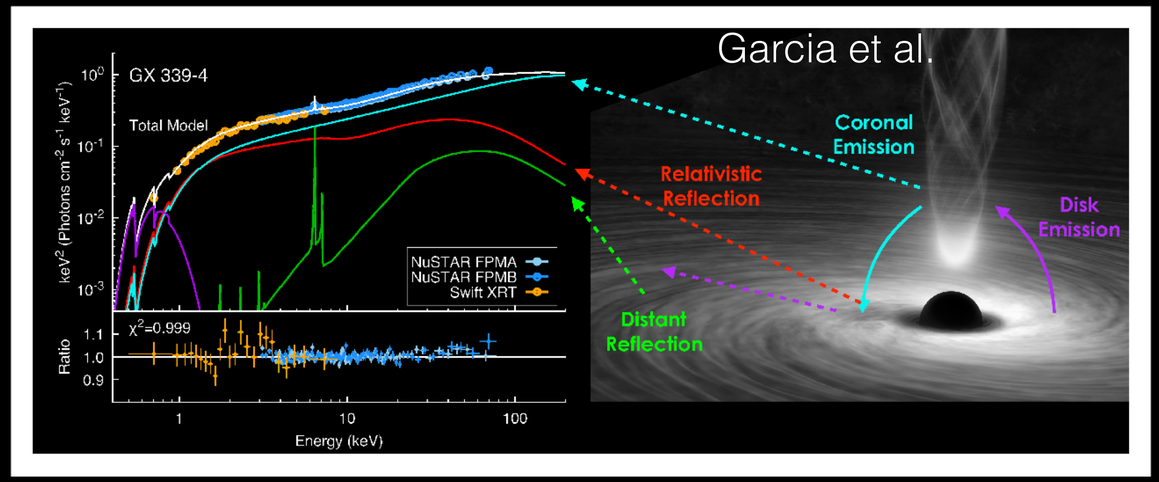


Fig. 2: examples of the neural network successfully predicting a piled-up spectrum that is consistent with a MARX simulation. Each incident spectrum is generated from relxill and each MARX simulation assumes an exposure of 5e4 s.

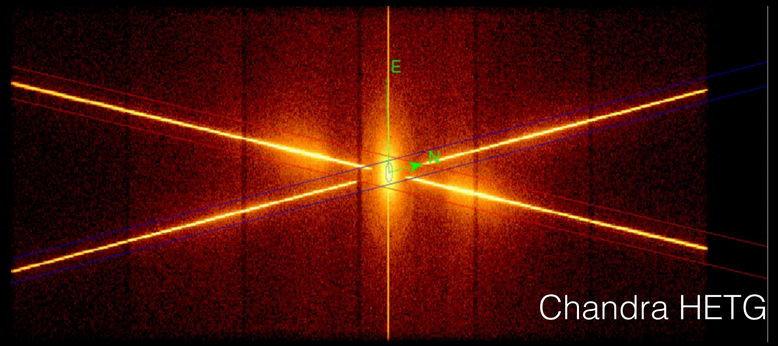
Inferring physical parameters while accounting for instrumental effects



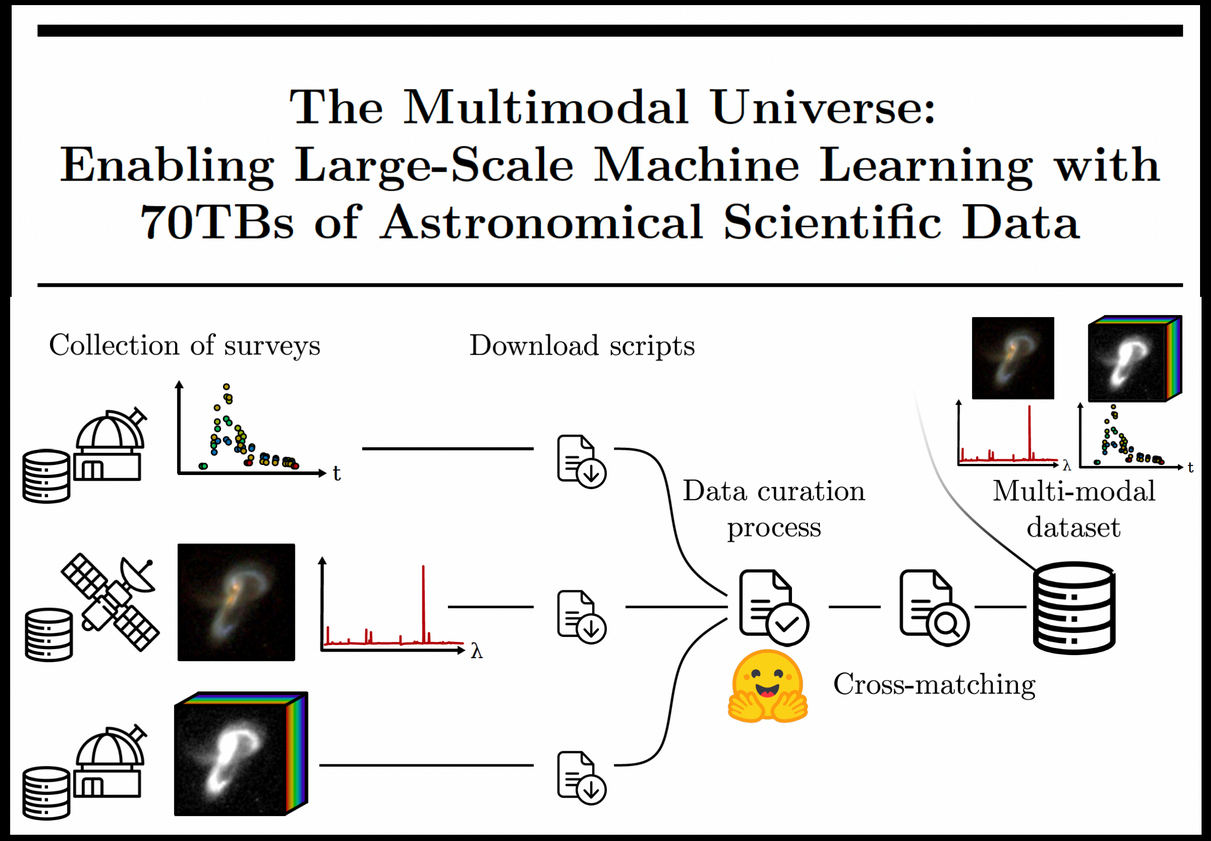
J. Yang et al., in prep



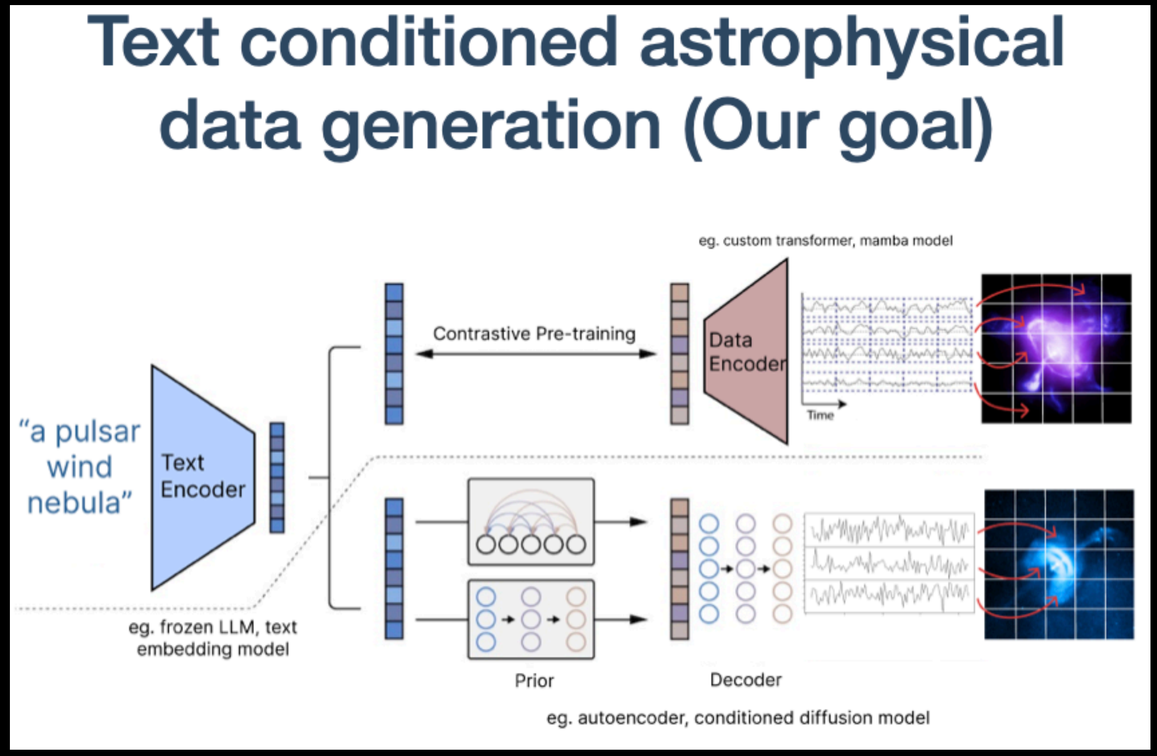
- The Bayesian MCMC emulator has been incorporated in the fitting process for parameter inference.
- But, we don't want to learn the simulation prescription of pile-up. We want to learn from the data.
- Domain adaptation using spectral grating data.



Where do we go next? X-ray datasets in the era of foundation models



Accepted to NeurIPS, Angeloudi et al. 2024



Martinez-Galarza et al. in prep

A few final remarks

- X-ray astronomy was always the science of “a few photons”. We now have enough photons to learn relevant patterns from them that inform physical models.

- X-ray catalogs and archival research have enabled an era of data-driven discovery in high energy astrophysics, in particular transient science and multi-messenger themes. The moment is ripe for profiting from cutting-edge machine learning methods.

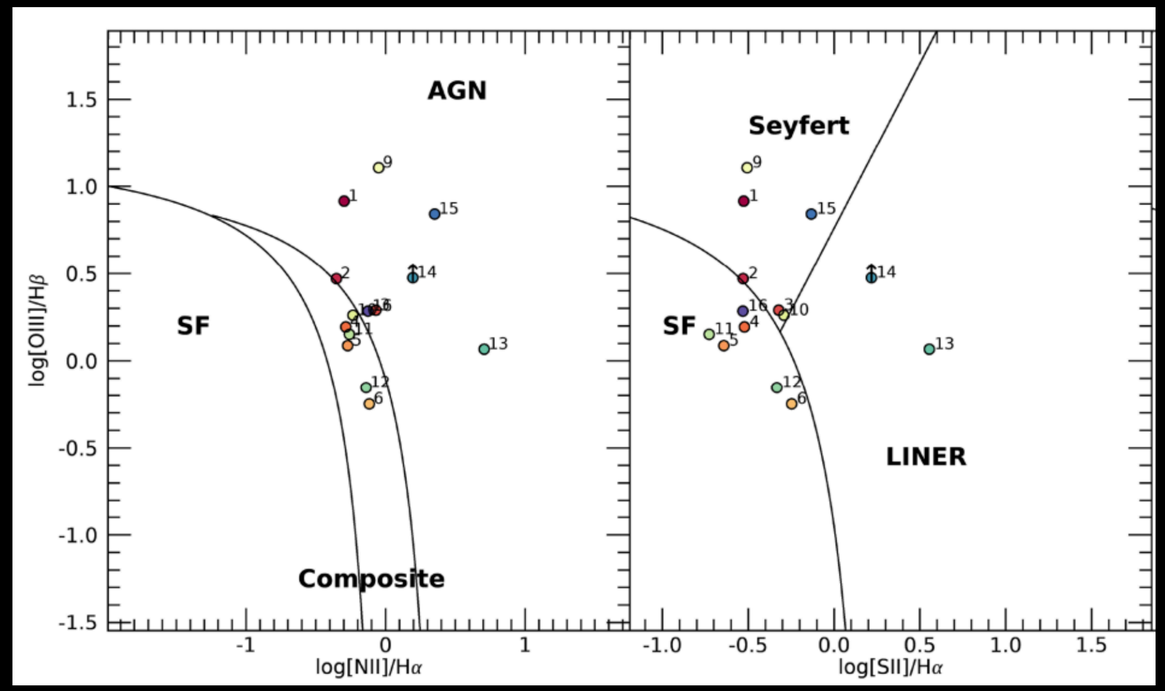
- Self-supervised learning can result in meaningful representations of X-ray event data that can be effectively used for downstream tasks such as classification, regression, and anomaly detection, without having to rely on features or labels created by humans.

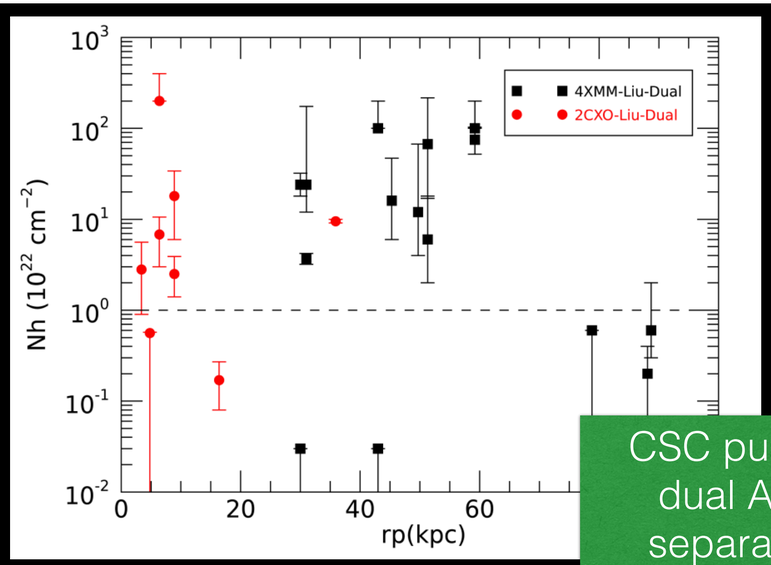
- As we move to the next generation of X-ray facilities, data complexity will continue growing, but the non-linear instrumental effects such as pileup will also be there. ML emulation offers an opportunity to learn those effects from data.

Your Science Here!

https://cxc.cfa.harvard.edu/csc/

X-ray view of optically selected dual AGNs



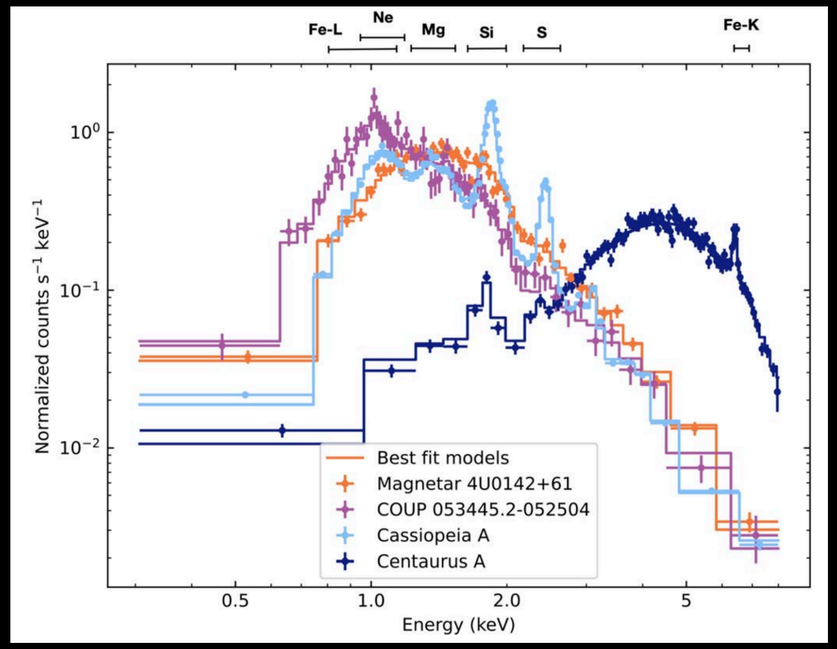


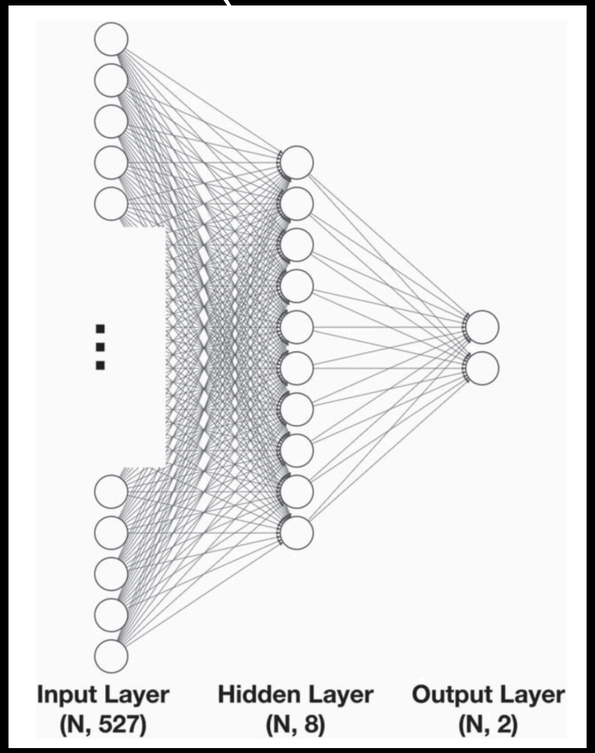
- Cross-matching of CSC sources with optical and IR catalogs allows for multi-wavelength identification and classification of dual AGNs.
- >80% of the targets identified in pairs have confirmed AGN emission.
- X-ray luminosity increases with decreasing pair separation, suggesting that mergers contribute to more luminous AGNs.
- From X/mid-IR ratio vs. HR, evidence that dual AGNs are more obscured that isolated ones, and less obscured than more evolved mergers.

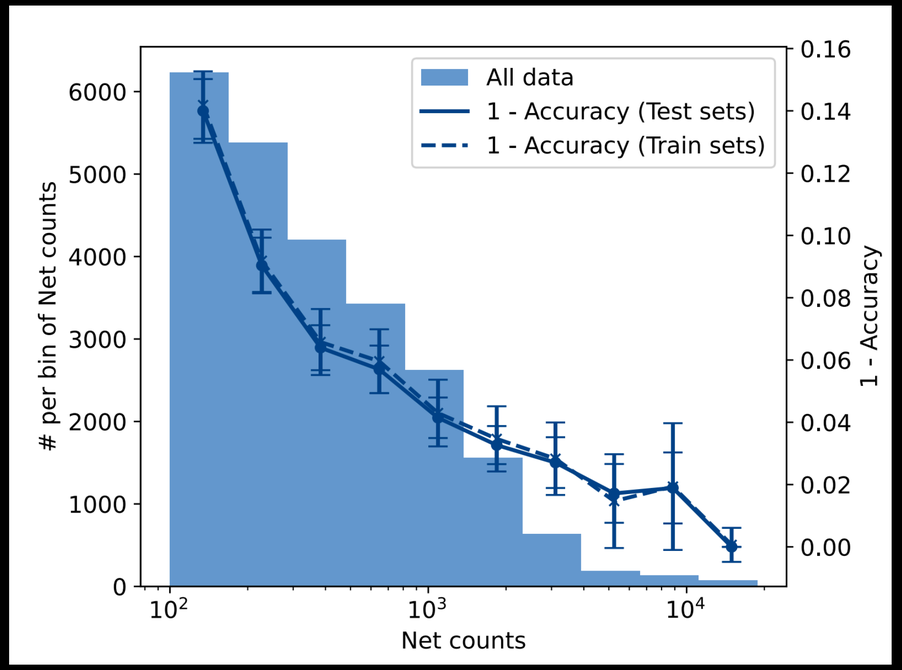
CSC pushes the limit of dual AGN studies to separations of <5kpc

DeRosa et al. 2023

Classification of CSC spectra using Neural Nets (Hebbar et al. 2023)



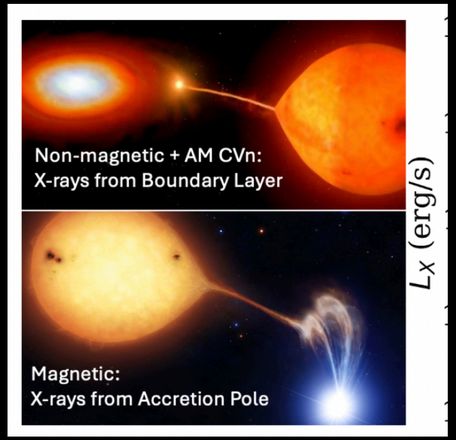


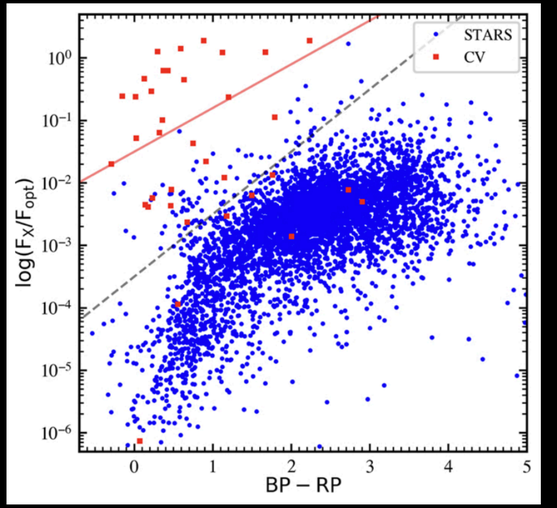


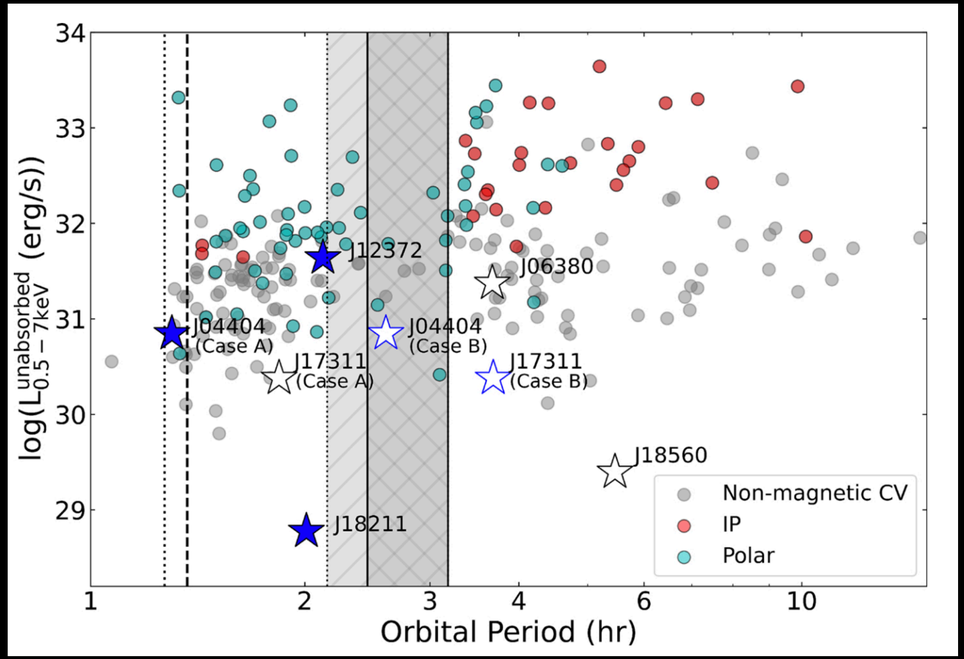
- A Neural Network trained to classify stars and AGNs, using the simulated spectra computed using priors derived from CSC spectra.
- Trained model applied to both the simulated and observed CSC spectra, and achieve accuracies over 90%
- See also Yang et al. 2022, Chen et al. 2023, Kumaran et al. 2023, Perez et al 2023

Nearby Cataclysmic variables and Accreting WD

- Accreting WDs are the more abundant interacting compact binaries. 14 newly identified using the CSC+ Gaia
- Ultra-compact versions of WD in binaries are probes of general relativity and are expected to emit GWs in the LISA band.
- The study of their formation channels is relevant for binary evolution and subsequent multi-messenger science.
- Joint X-ray and optical searches efficiently find CVs. Chandra: magnetic and low accretion rate CVs, which could be missed by purely optical surveys.







Galiullim et al. 2024, Rodriguez et al.