# Visualizing three years of STIX X-ray flare observations using self-supervised learning

Mariia Drozdova, Vitaliy Kinakh, Francesco Ramunno, Erica Lastufka, Paolo Massa and Slava Voloshynovskiy

UNIVERSITÉ DE GENÈVE

Exploring the Sun's X-ray emissions, the Solar Orbiter's Spectrometer/Telescope for Imaging X-rays (STIX) has documented three years of flare activity. We've employed advanced self-supervised learning models to visualize this data, facilitating easier exploration and analysis. In addition to filtering and clustering pipeline-generated images, we demonstrate the effectiveness of a small manually labeled dataset in facilitating classification and anomaly detection.

## Data

From STIX's database, containing more than 43,000 events, we filtered 9,000 flares for analysis, specifically those with energies ranging from 4-10 KeV and at 16-28 KeV. A small labeled set of images consists of 156 images with one source, 5 images with two sources, 1 image with three sources, and 1 image capturing four sources.

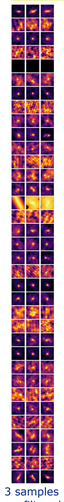
STIX dataset center

## Methodology

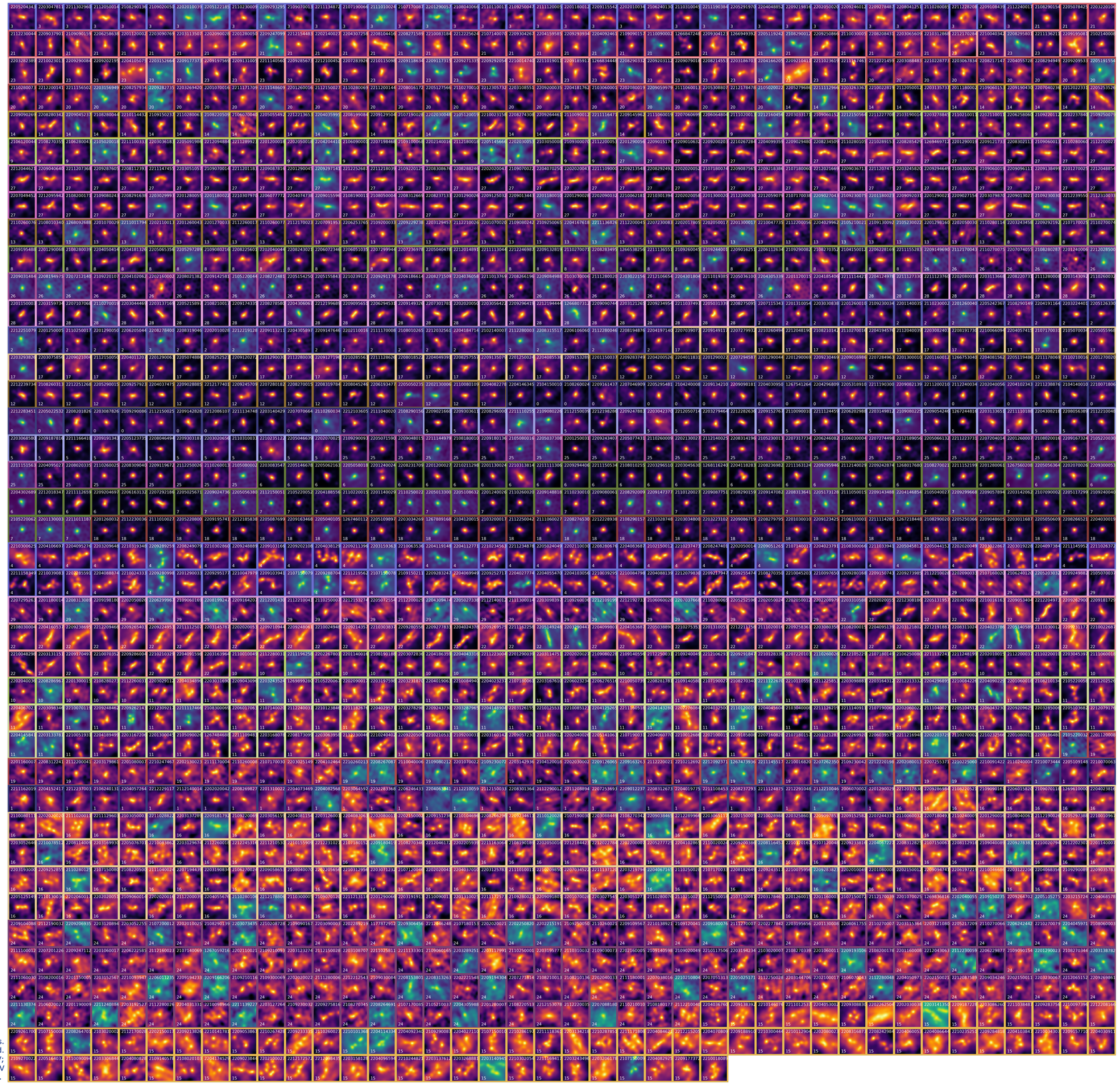
We experimented with diverse encoders to analyze STIX's solar flare images, capitalizing on their ability to learn from unlabeled data through self-supervised (SSL) and weakly supervised learning. This approach is key in fields where annotated datasets are limited. We experimented with encoders, all leveraging the Vision Transformer (ViT) architecture, including DINO, MSN, CLIP, I-JEPA, pretrained on natural images without pretraining on STIX data.

## Filtering

Visually the best clustering was achieved with CLIP. Then, we conducted Principal Component Analysis (PCA) in the feature space followed by K-Means clustering with k=30. The Figure on the left shows clusters hand-selected for removal due to many visual artifacts. The central figure displays the 4758 high-quality images organized via clustering that are kept for further analysis.



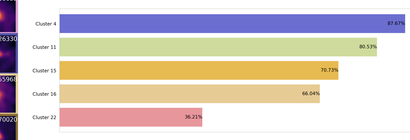
3 samples per filtered out cluster



Selected images grouped by clusters.
Each image has a corresponding uid.
Viridis colormap corresponds to 16-28 KeV;
plasma colormap to 4-10 keV
Only 1586 images are visualized due to printing restrictions.

## Multiple X-ray sources

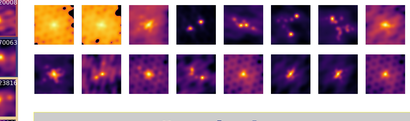
For our study, we train a classifier to distinguish between images containing one compact X-ray source and those with multiple sources, using the small labeled set for initial training. This classifier is then applied to our unlabeled, filtered dataset. Several of our previously identified clusters contain high proportions of images classified as containing more than a simple compact source.




The outcome of the supervised classifier supports the results of unsupervised clustering.

## Anomaly Detection

Using Isolation Forest on CLIP extracted features, we identified a group containing all labeled images with 2, 3, and 4 sources, alongside some distinct one-source images. This highlights the method's potential in isolating rare patterns, though it also indicates the complexity of distinguishing between common and unusual flare characteristics. Images are ordered by the anomaly score (from high to low).



## Conclusions

Our use of pretrained encoders efficiently groups STIX solar flare images by distinct patterns, without necessitating model retraining. Visually, CLIP produces best results. The impact of data variability and artifacts on clustering underscores the importance of further investigation.

https://github.com/MariiaDrozdova/STIX-clustering