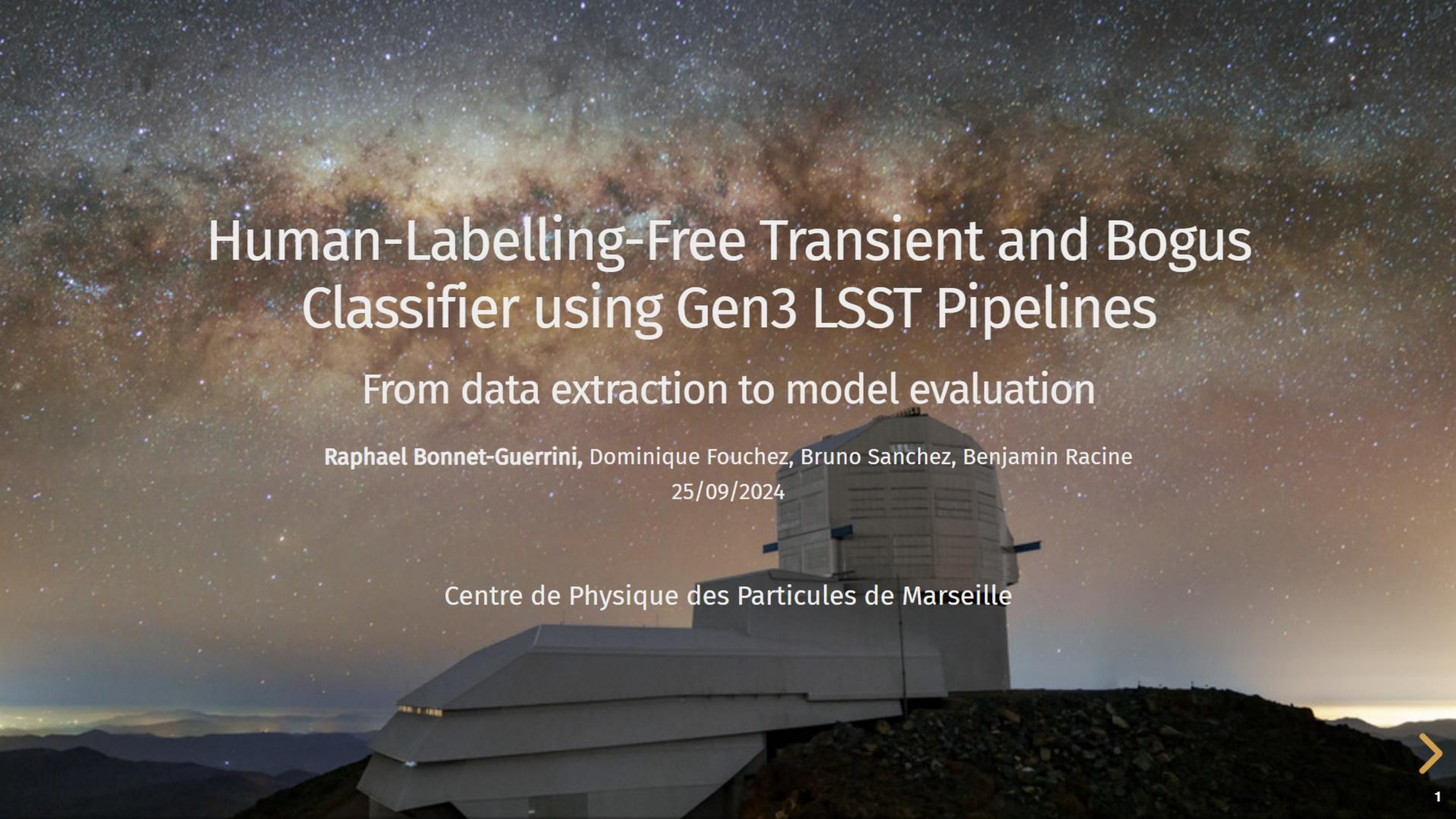


Human-Labelling-Free Transient and Bogus Classifier using Gen3 LSST Pipelines

From data extraction to model evaluation

Raphael Bonnet-Guerrini, Dominique Fouchez, Bruno Sanchez, Benjamin Racine

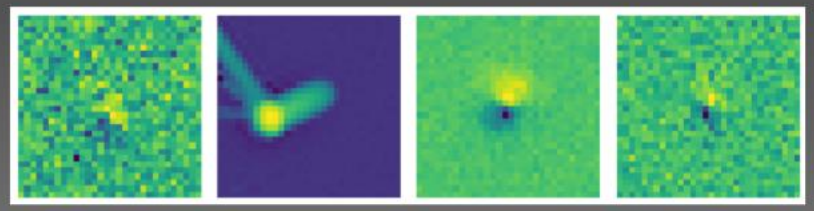
25/09/2024

Centre de Physique des Particules de Marseille

Context and of the presentation.

Context :

- Our team is interested in cosmology with Supernovae.
- Focus on difference imaging data and pipeline.
- DIA techniques produce a majority of **bogus detections** — noise or artifacts — that can result from imperfect image subtraction, cosmic rays, bad pixels, or atmospheric effects.



Presentation plan :

1. Present the **production of fake catalogs** and their **injection into real data** using the Gen3 pipeline.
2. Investigate the potential of an **ML-based Transient/Bogus classifier** using fake injections.

Intuitions behind this project

Key Intuitions:

- In **real data:** High rate of bogus, very low rate of transients.
- Possibility to **simulate transients** using fake supernova source injections.
- Assuming real data are nearly all bogus and injections are all transients, we have a (noisy) labeled dataset!

⇨ Possible Machine Learning-Based Classification Task

We train the model to classify between injections and real data, and in reality, it classifies between bogus and transient.

⇨ False Positive Predictions Are the Potential Real Transients

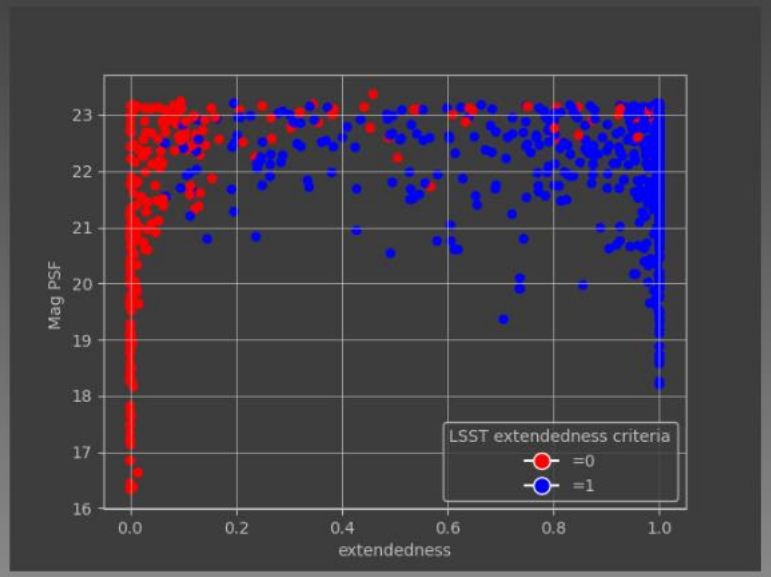
Catalog creation and pipeline production for Galaxy-Based Injection

- Working on galaxy-hosted transient injection on DRP data.
- Developing a new, enhanced catalog creation with physically motivated injections rather than random ones.
- Using the Gen3 Data Management product to ingest the injection catalog and process the dataset.

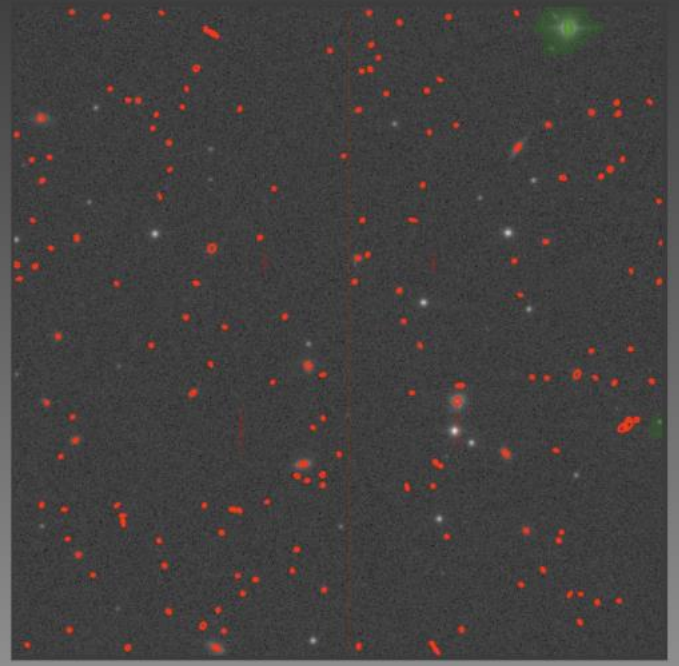
Galaxy Identification

- Selecting galaxy-type sources using the 'extendedness' criteria of the Gen3 pipeline.

- Retrieving the shape (semi-major, semi-minor axes) and magnitude to create a database of galaxies with their properties.



Extendedness measurement vs extendedness criteria.



Galaxies circled in red.

Magnitude and Positions of the Injections

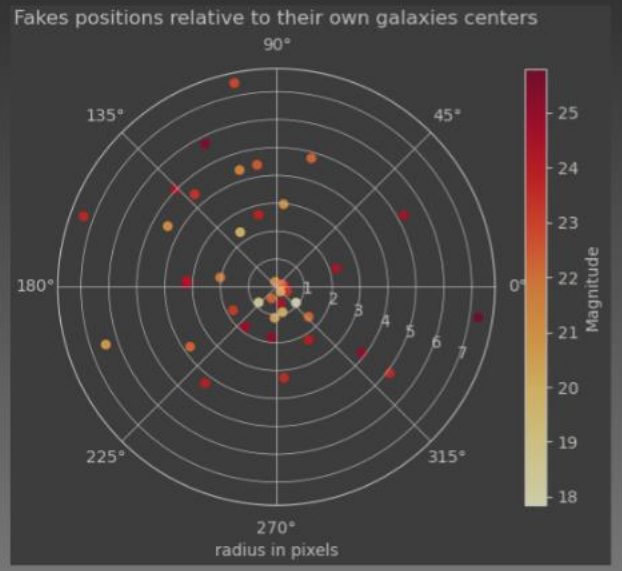
- Magnitude and distance to the host are sampled from the host galaxy properties. We are injecting in ~3% of the galaxies:

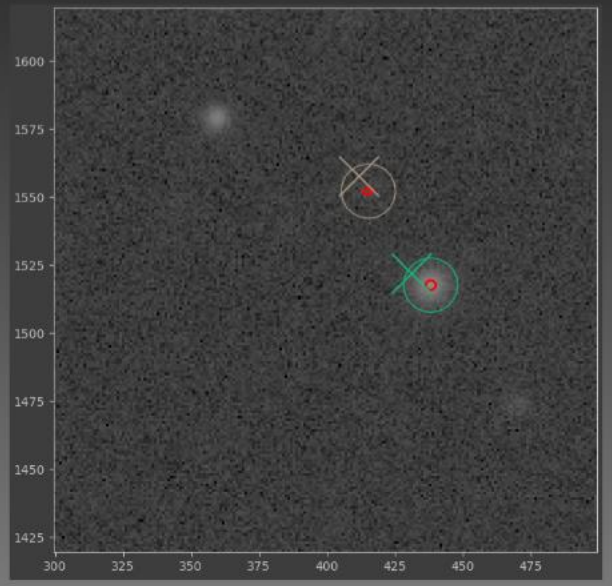
$$d_{\text{inj}} \sim \mathcal{N}(0, \text{SemiMajor}_{\text{host}})$$

$$m_{\text{inj}} \sim \text{Uniform}(m_{\text{host}} - 1, m_{\text{host}} + 3)$$

- From the host's reference frame, the positions are converted to x, y, and RA/DEC.

- We then add 5% of host-less injections.



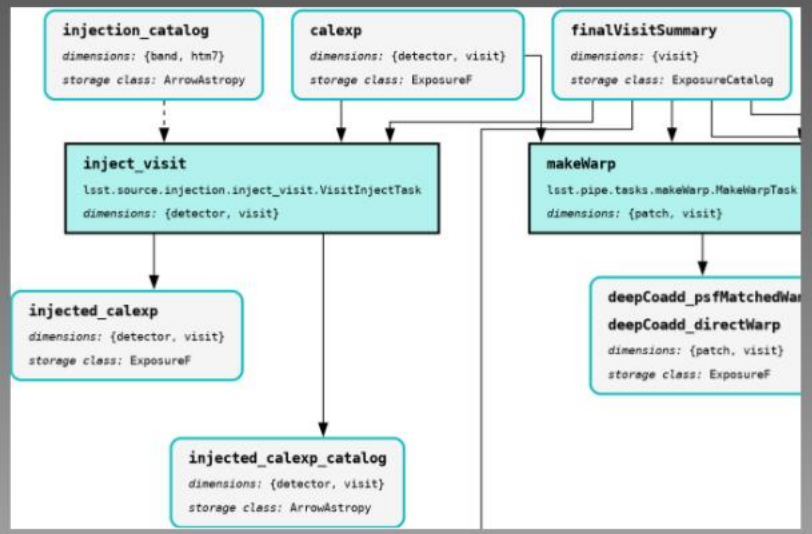


Catalog Creation, Butler Ingestion and pipeline processing:

- The injection catalogs are created for each specific visit. They are non correlated - No Light Curve.
- Ingestion is done band by band.

```
ingest_injection_catalog \
-b $BUTLER_REPO \
-i $CATALOG_REPO/g_band_catalog.csv g  \
-o u/rbonnetguerrini/inject_input_g
```

Adding a inject_visit task to the pipeline on step 3, we build a DIA object table.



Data Presentation

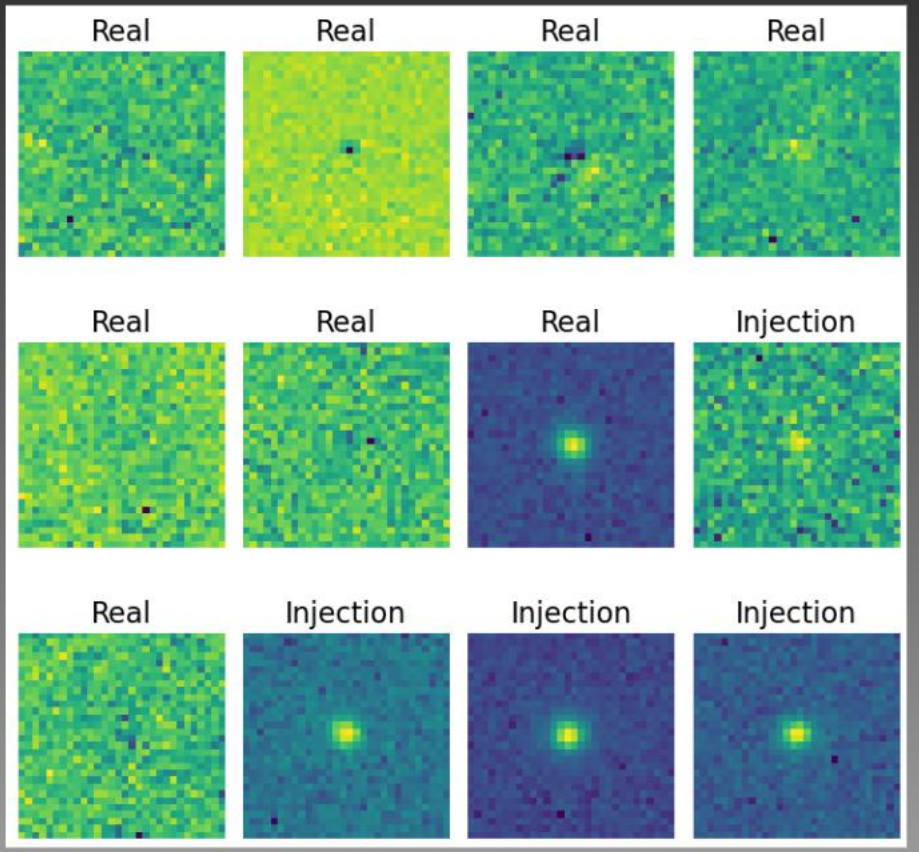
The HSC RC2 subset is composed of 6 detectors, with 8 visits per filter. UDEEP COSMOS foreseen.

Producing the Cutouts:

- Cutout coordinates are extracted from the DIA source tables and produced from the Calexps.
- Final format: (30x30), normalized grayscale.

Classes and Labels:

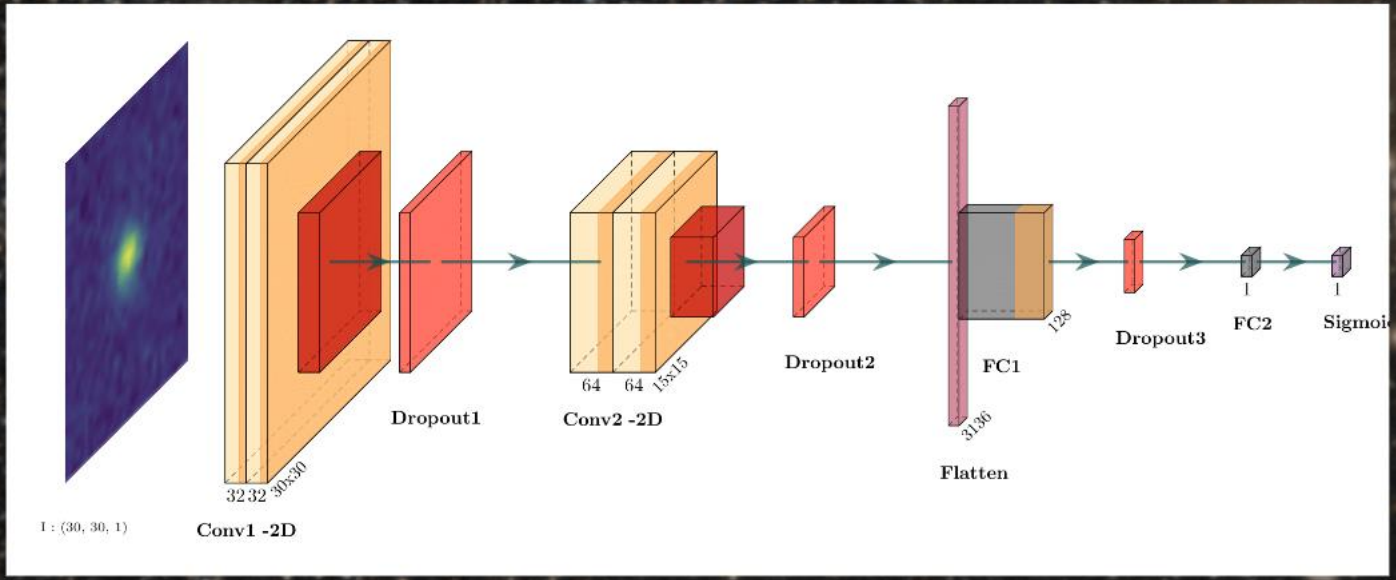
- Inferring the injected data from the matchDiaSrc.



Network Architecture

A simple CNN architecture:

- 2 2D-convolutional layers with ReLU activation and max-pooling.
- 2 fully connected layers (FCC) for high-level abstractions and classification output.
- Dropout layers to avoid overfitting.

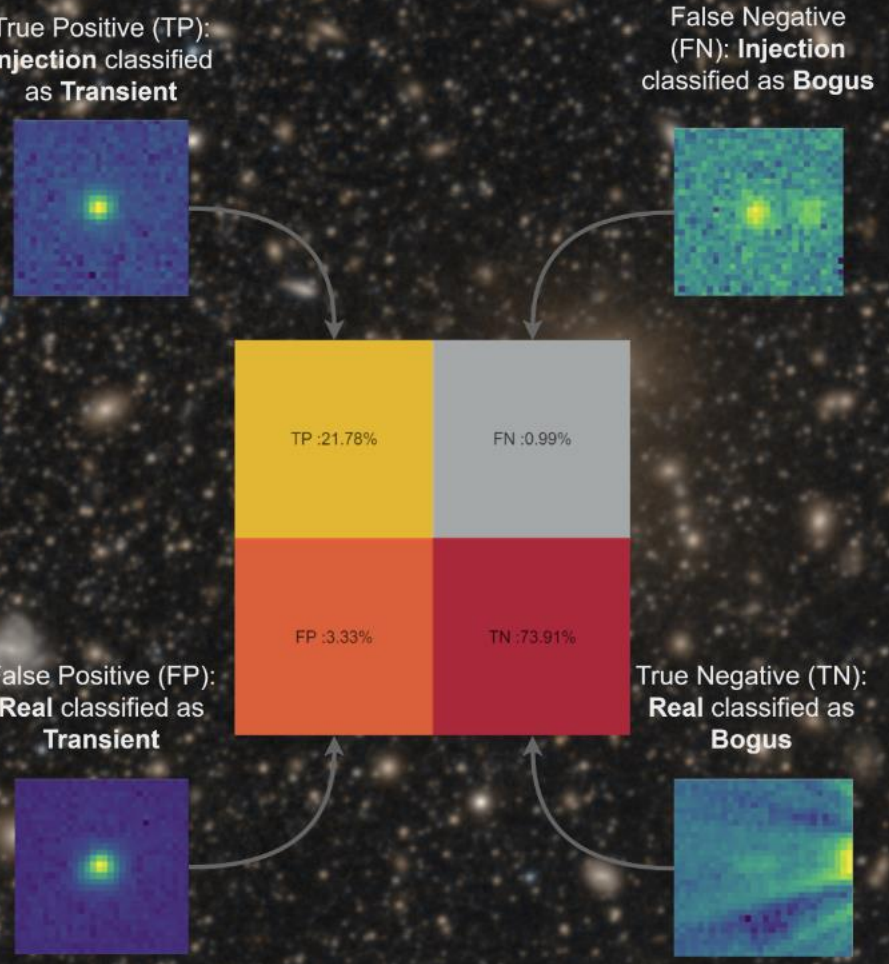


Confusion Matrix

Output Classes and Their Interpretation

Focus:

- Minimize False Negatives (FN) and maximize Sensitivity.
- Monitor False Positives (FP) and Precision.



Confusion Matrix

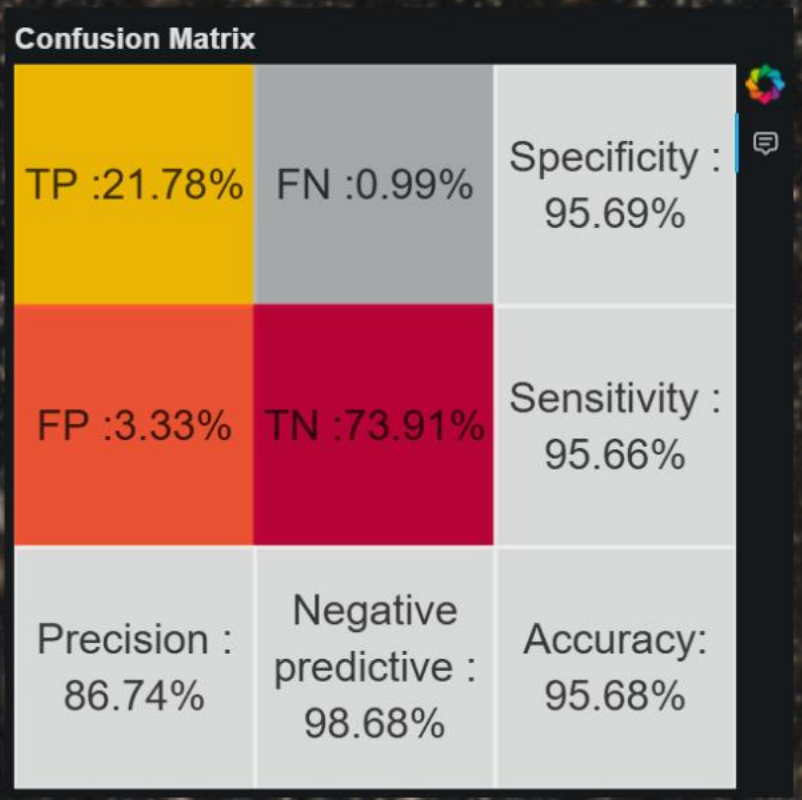
Trained on all different available visits:

Focus:

- Minimize False Negatives (FN) and maximize Sensitivity.
- Monitor False Positives (FP) and Precision.

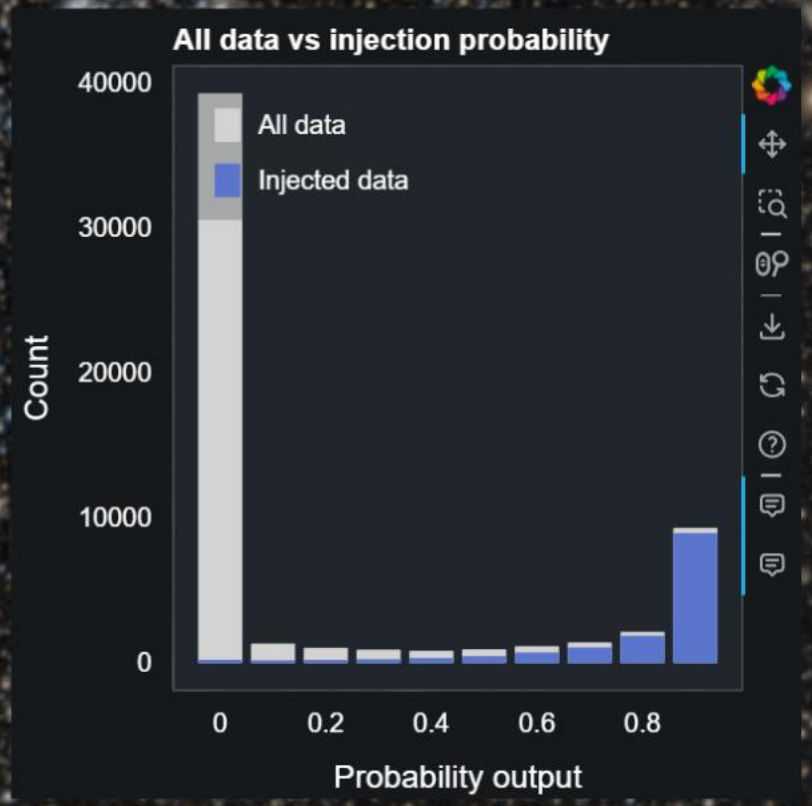
Analysis:

- Need more insight on how the network is predicting these classes.



Injected and Real Data Output Probability Comparison

Trained on all different available visits:



We are looking at the output probability comparison between the full dataset and the injection.

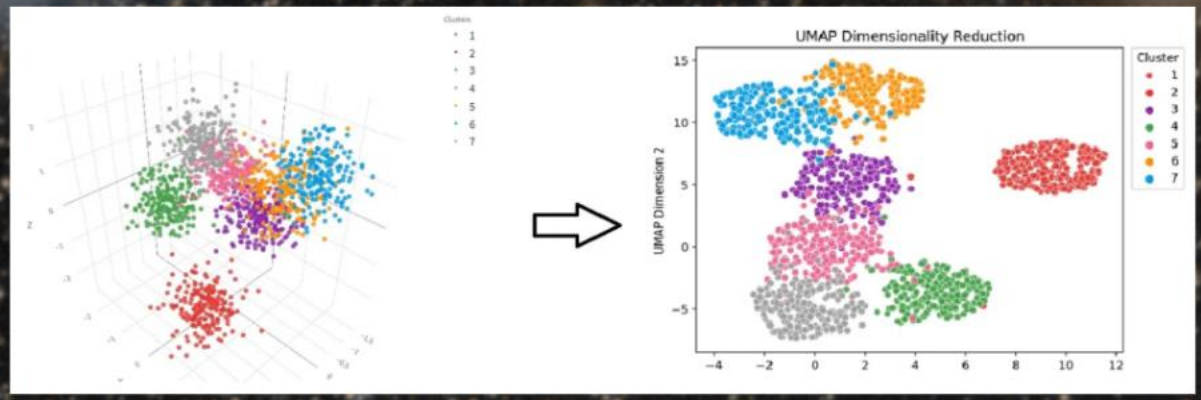
- Injections should be predicted around 1.
- The additional data predicted around 1 are our potential transients.
- We wish to see a clear split with fewer 'in-between' predictions.
- We also want to reduce or explain the injections predicted around 0.

UMAP: A Visualization Tool for Neural Network Latent Space

cs arXiv.1802.03426

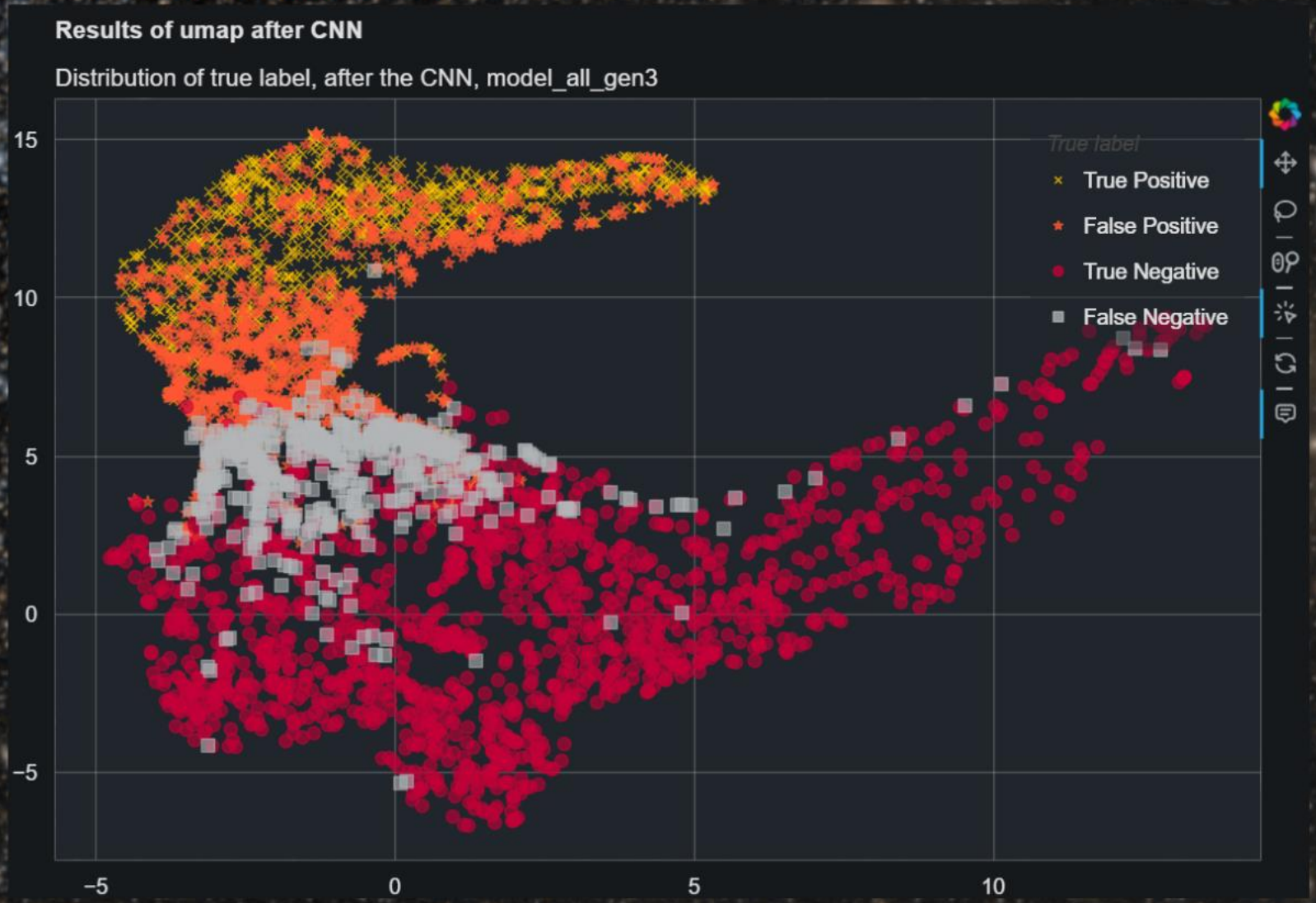
UMAP Overview:

- Helps visualize high-dimensional data in 2D.
- Preserves both local and global structure using non-linear dimensionality reduction.
- Builds a nearest-neighbors graph and optimizes it for lower dimensions.
- False Negative: Injection data misclassified as Bogus.



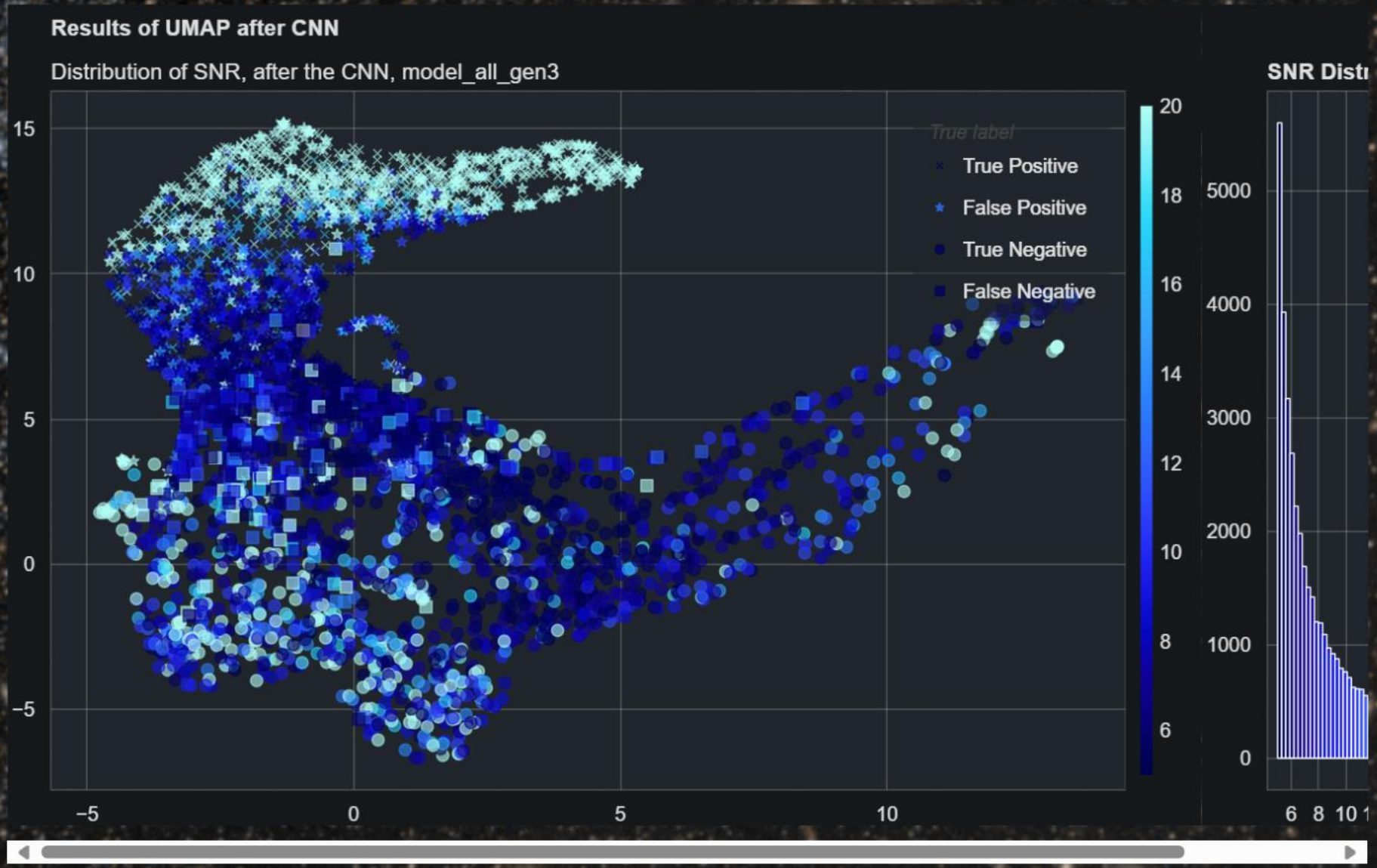
UMAP with Data Class Predictions

Trained on all available visits:



UMAP with Data Class Predictions and SNR

Trained on all available visits:



Confusion Matrix

Trained on all different available visits, evaluating only high SNR:

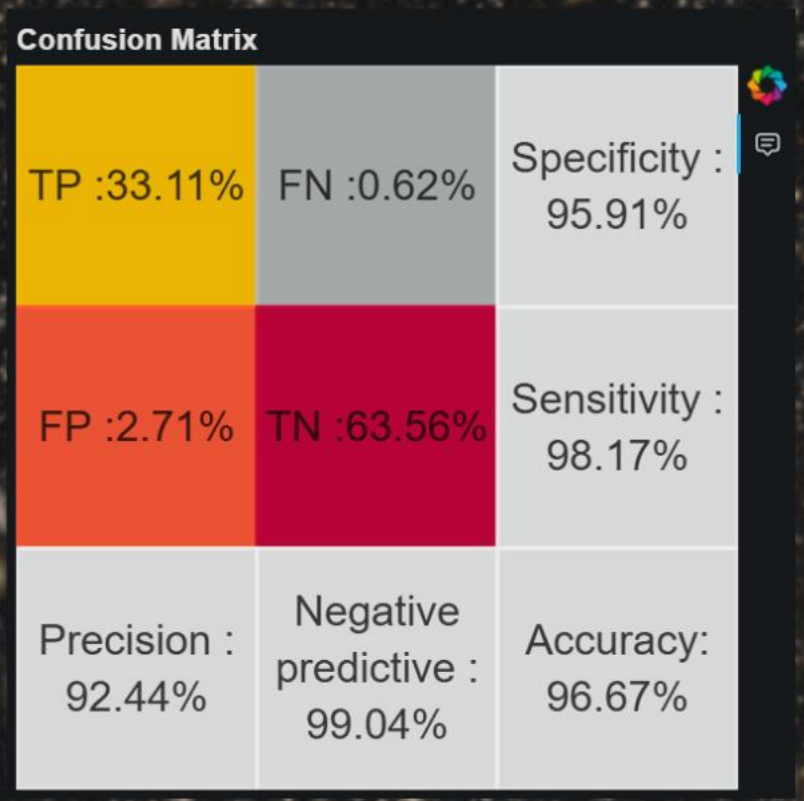
$$\mathrm{SNR} \notin [0, 8]$$

Focus:

- Minimize False Negatives (FN) and maximize Sensitivity.
- Monitor False Positives (FP) and Precision.

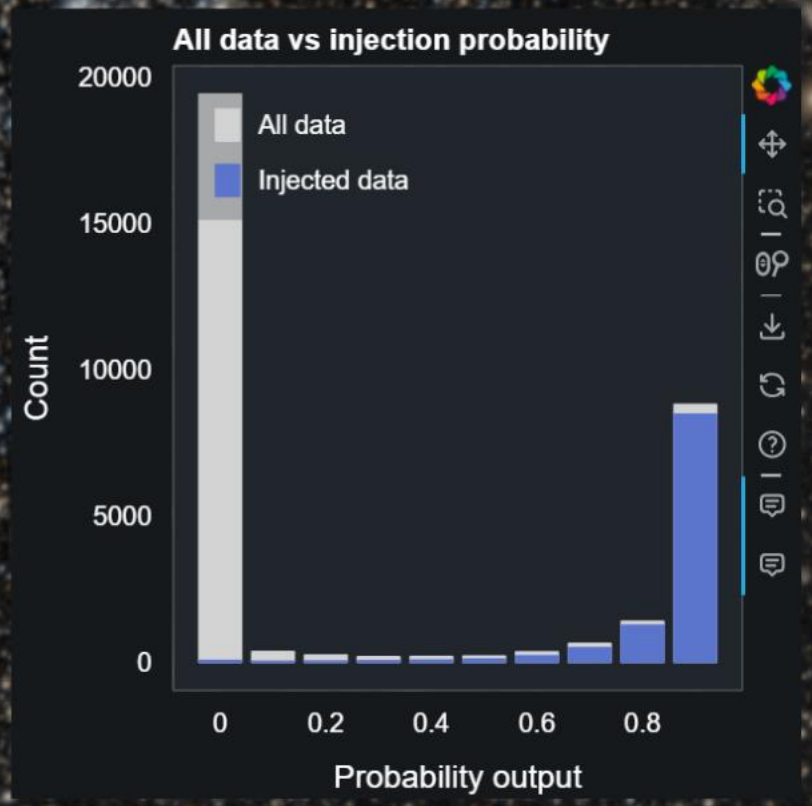
Analysis:

- Reduction of False Negatives (FN).
- Stable False Positive (FP) detections.



Injected and Real Data Output Probability Comparison

Trained on all different available visits, evaluating only high SNR:



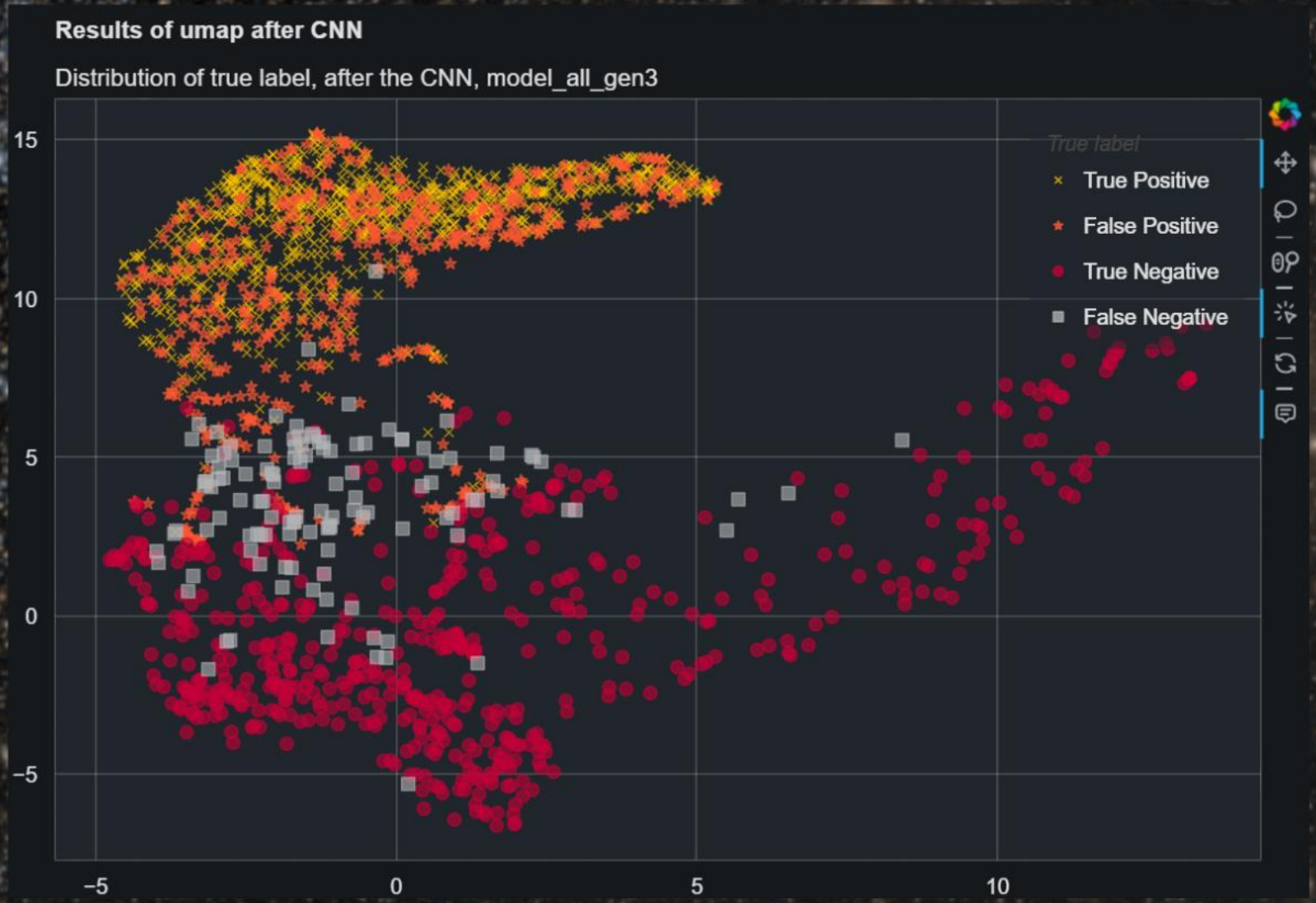
$$\mathrm{SNR} \notin [0, 8]$$

We are looking at the output probability comparison between the full dataset and the injection.

- Injections should be predicted around 1.
- The additional data predicted around 1 are our potential transients.
- Reduction of the 'in-between' predictions.
- By removing low SNR, we target the uncertain classifications of the network.

UMAP with Data Class Predictions

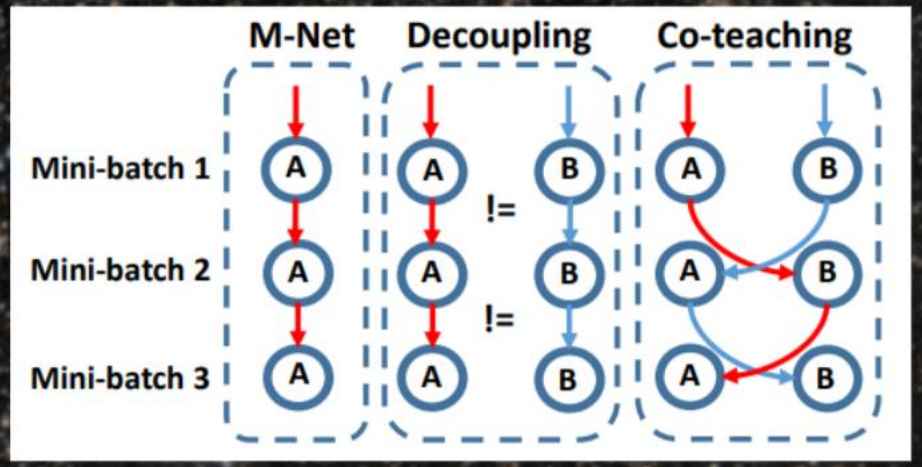
Trained on all available visits, evaluated only for high SNR:



What Are the Potential Improvements for the Model?

Co-Teaching: A Self-Event-Selection Strategy for Weakly Supervised Learning

Two models are trained simultaneously with different views on the same dataset.



In each batch, each model selects the datum with the smallest loss (most confident predictions).

Avoid training on the wrong labels.

Pros:
- Effective for noisy datasets.
- Increases computational cost.

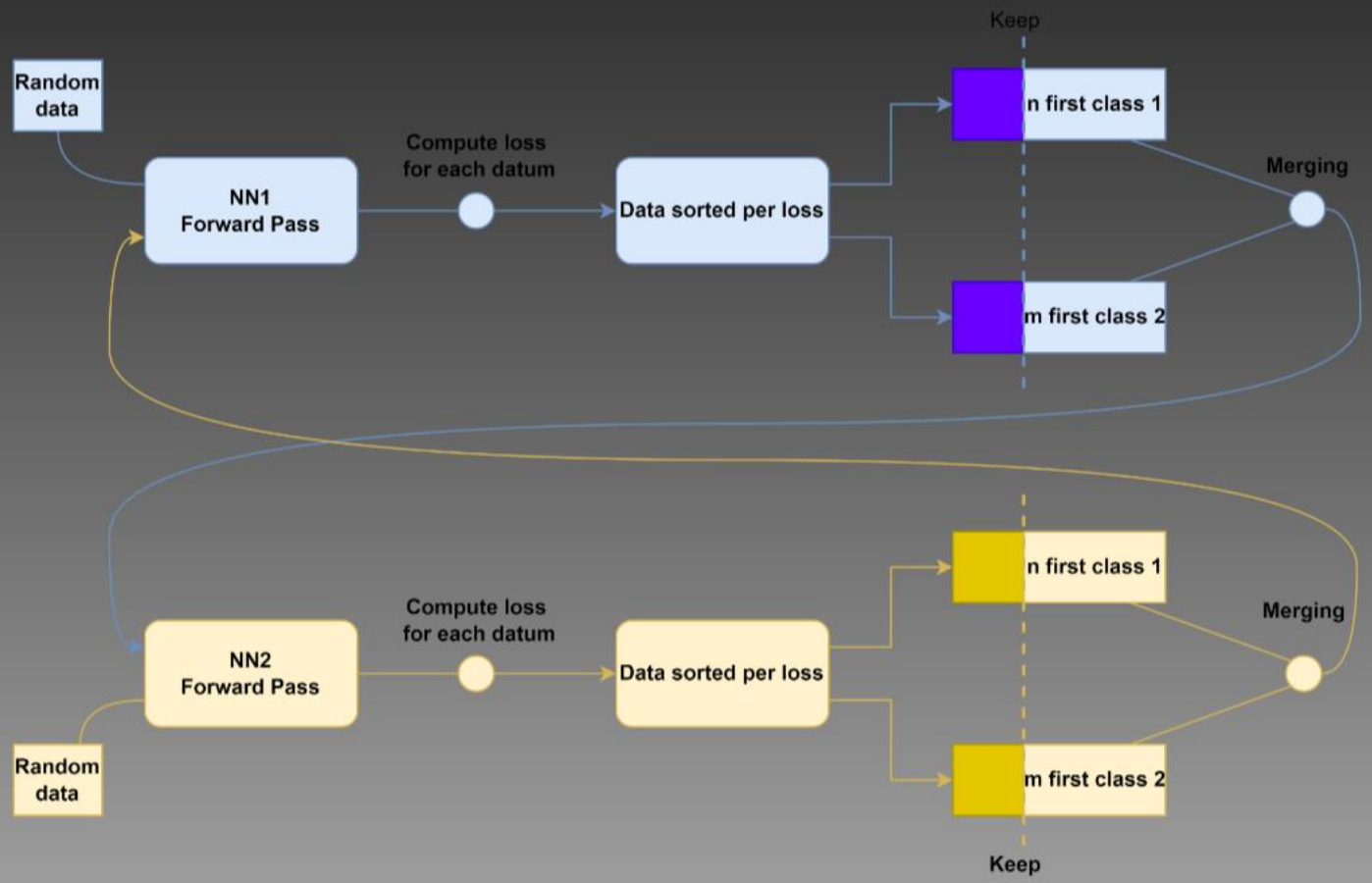
Cons:
- Assumes symmetrical noise.

cs arXiv.1804.06872

Asymmetrical Co-Teaching

Key Changes:

- Implements different remembering rates for each class.
- Better fits the needs of our asymmetrically weakly supervised dataset.



Confusion Matrix

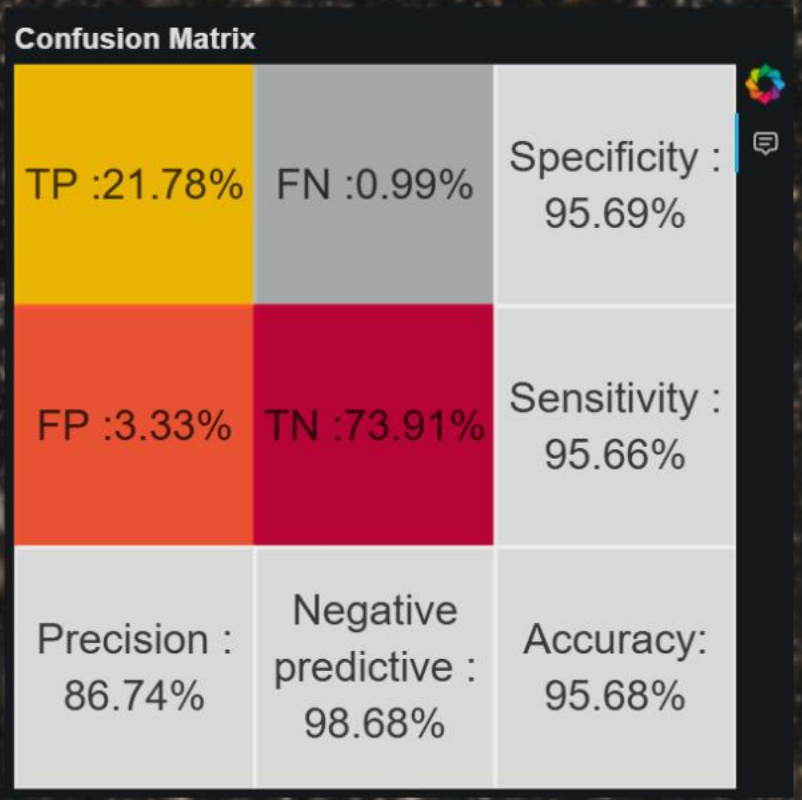
Trained on all available visits:

Focus:

- Minimize false negatives and improve sensitivity.
- Monitor false positives and precision.

Analysis:

- More insight needed into how the network predicts these classes.



Confusion Matrix

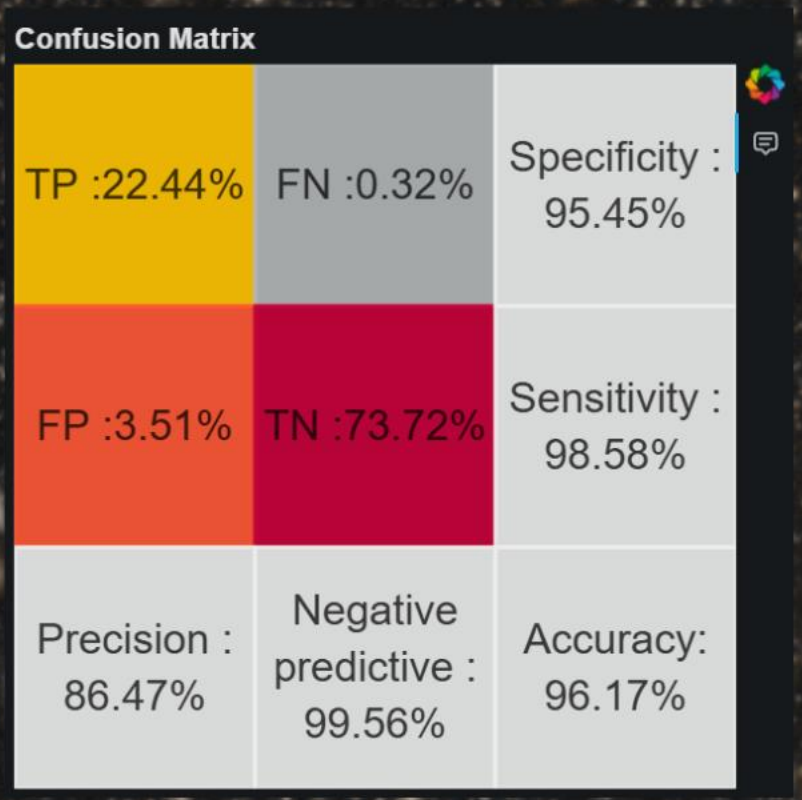
Trained on all available visits using the co-teaching method:

Focus:

- Minimize false negatives and improve sensitivity.
- Monitor false positives and precision.

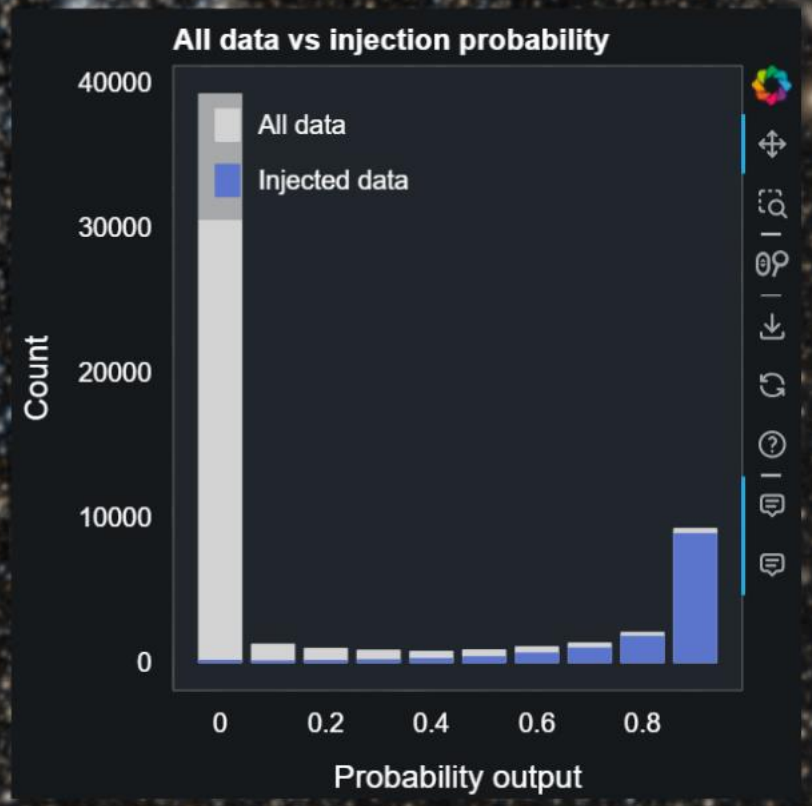
Analysis:

- Further reduction in false negatives.
- Maintained constant false positives.



Injected and Real data output probability comparison

Trained on all different available visits :

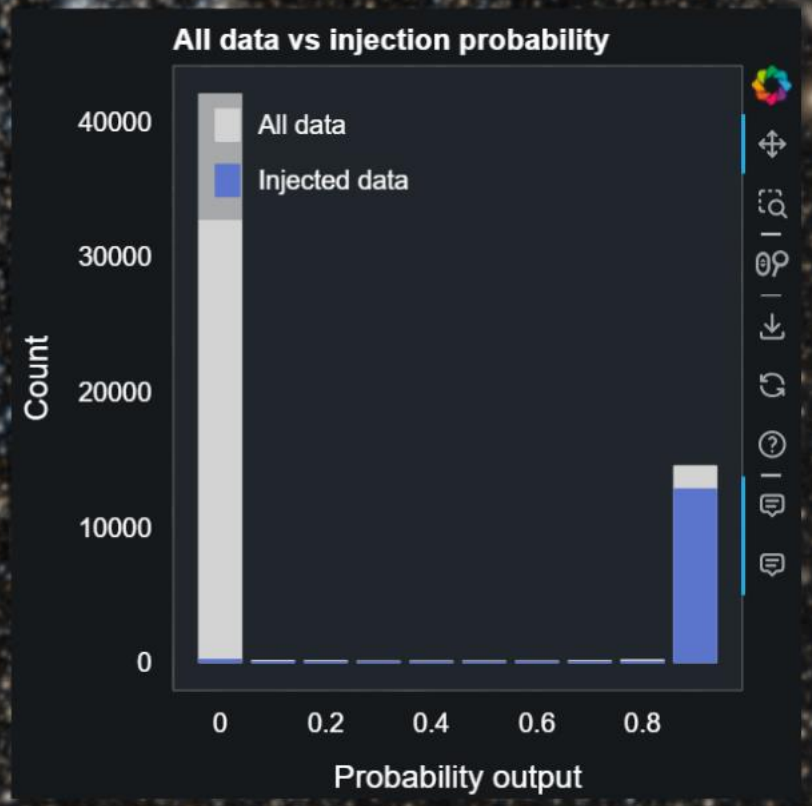


We are looking at the output probability comparison between the full data set and the injection.

- Injection should be predicted around 1.
- The additional data predicted around 1 are our potential transient.
- We wish to see a clear split with the less 'in-between' predictions.
- We also want to reduce or explain the injection predicted around 0.

Injected and Real data output probability comparison

Trained on all different available visits, using co teaching method.

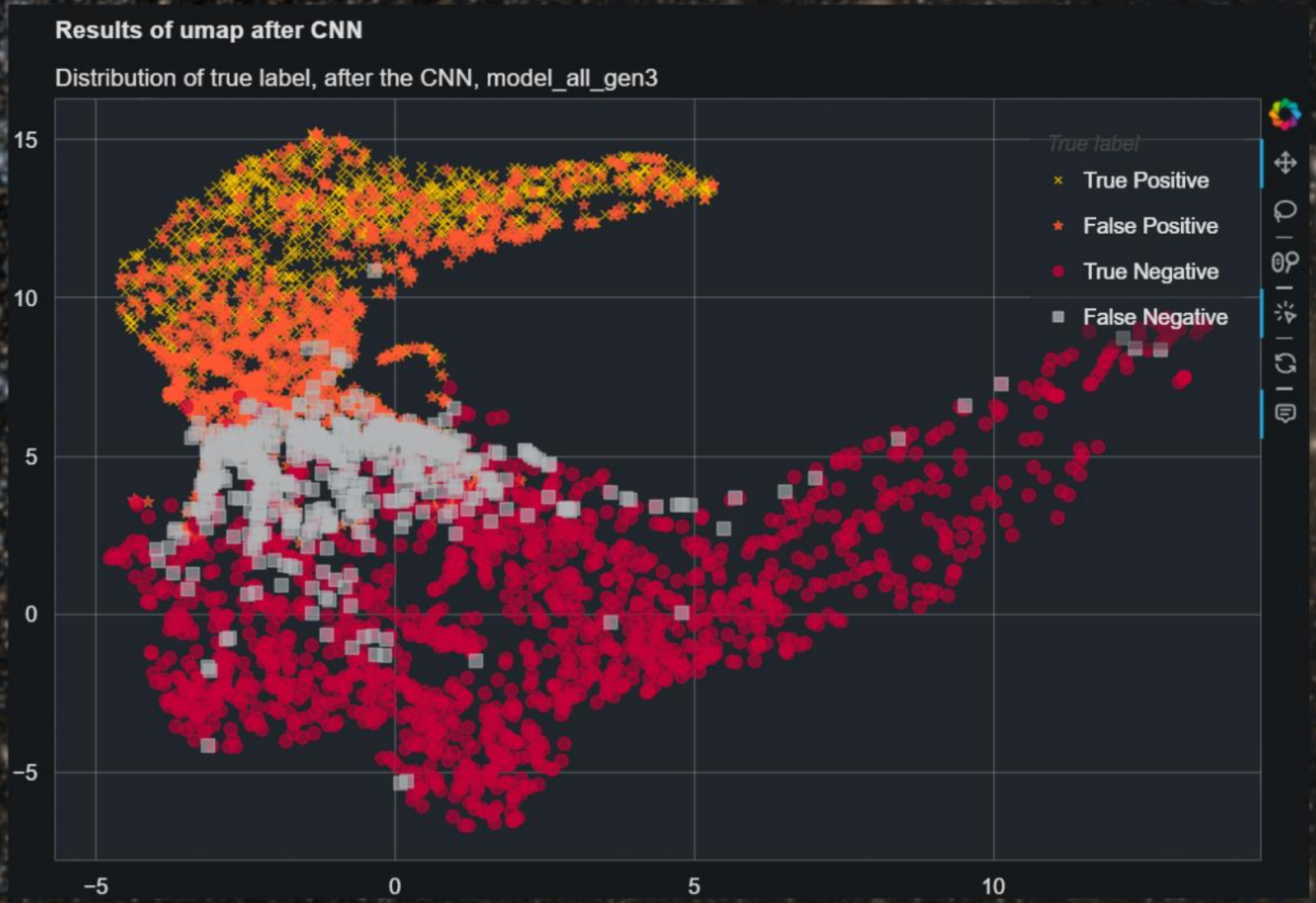


We are looking at the output probability comparison between the full data set and the injection.

- Injection should be predicted around 1.
- The additional data predicted around 1 are our potential transient.
- Clear split with nearly no 'in-between' predictions.
- We also want to reduce or explain the injection predicted around 0.

UMAP with data class predictions.

Trained on all different available visits :

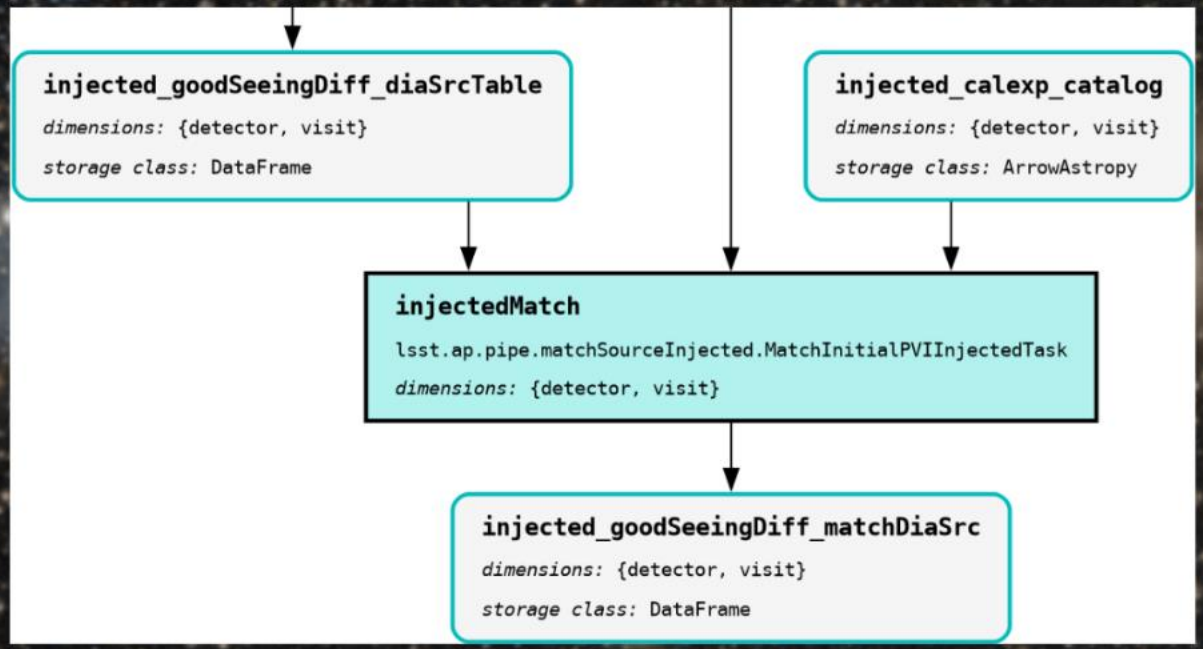


UMAP with data class predictions.

Trained on all different available visits, using co teaching method.

Light Curve Confirmation :

Adding a matching step to the pipeline on step 4, we build a DIA object table.



Light Curve Confirmation

We also developed a tool to inspect this LC on the UMAP.