Transient classifiers for Fink: benchmarks for LSST

https://arxiv.org/pdf/2404.08798
Submitted to A&A

Bernardo Fraga (CBPF)

Clecio de Bom (CBPF), André Santos (CBPF) + Fink Team

Enabling Astronomical Transient discoveries in the Rubin era: the Fink-Brazil Workshop
May 2024

Summary

- LSST-like simulated data - ELAsTiCC

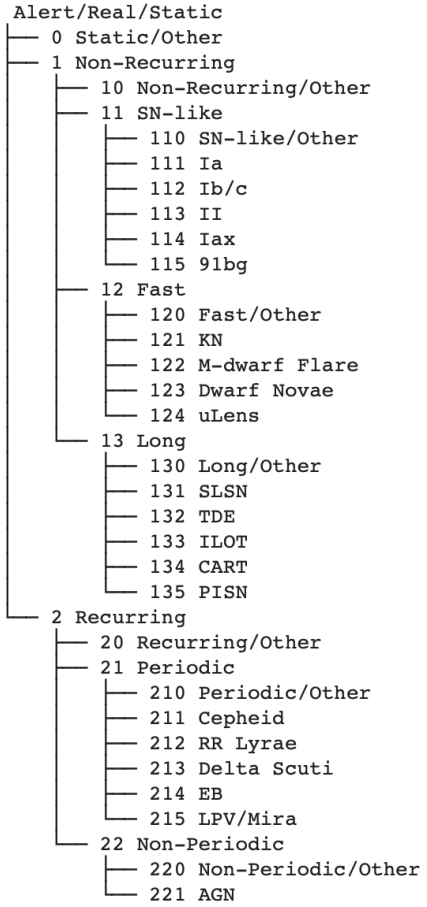
- Fink Classifiers
  - **CATS Broad Classifier**
  - SuperNNova
  - Early Supernova Ia
  - SLSN

- Fink infrastructure performance

CBPF

ELAsTiCC

- ELAsTiCC: The successor of PLAsTiCC (**P**hotometric **L**SST **As**tronomical **Ti**me-series **C**lassification **C**hallenge)
- **E**xtended **L**SST **As**tronomical **T**ime-seri**e**s **C**lassification **C**hallenge led by LSST DESC
- Simulated time-domain events and host galaxy associations
- Designed to test broker's infrastructure and evaluate ML algorithms
- Millions of synthetic light curves using SNANA
  - 6 filters (ugrizY)
  - Rubin depth and cadence
  - Extinction and atmospheric noise
  - Realistic host galaxy associations
  - Several Different models in a tree-based taxonomy
- Closest to LSST-type data so far
- ELAsTiCC V2 is under way



The streamed sample

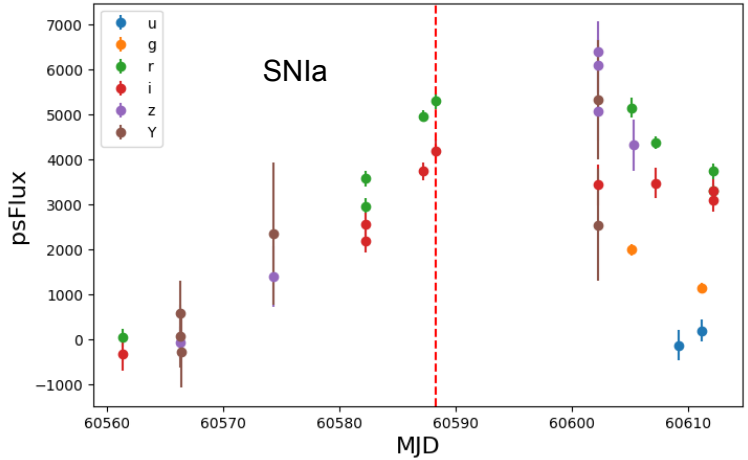
- ELASTiCC team streamed alerts daily through the ZTF Alert Distribution Server. 3 years of simulated light curves.

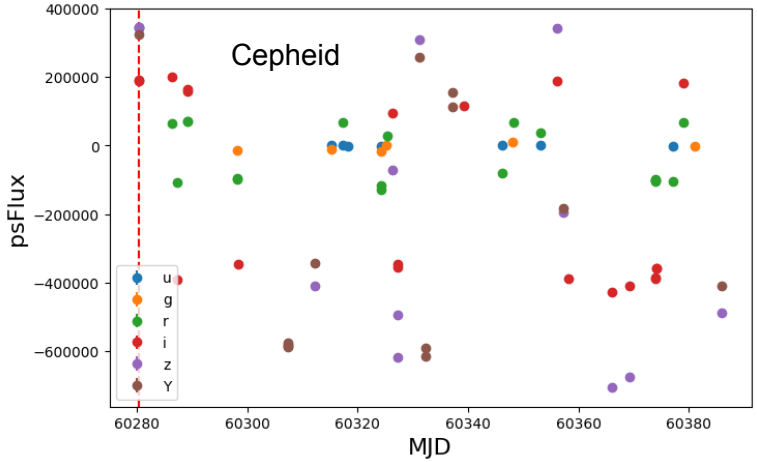
- Truth Table released after 3 years of data were streamed.

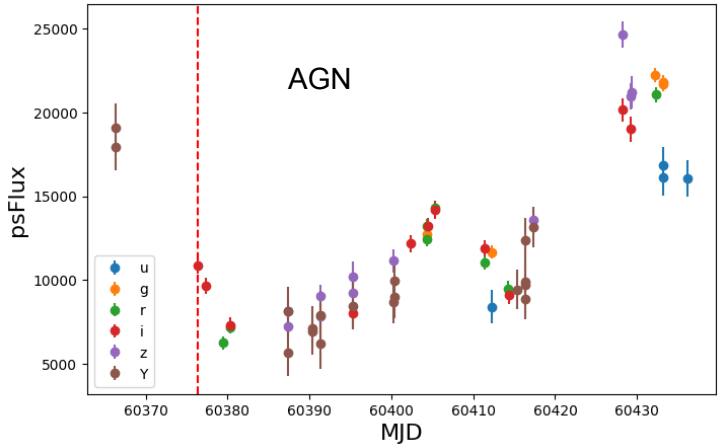
- After the unblinding, we used the first year of streamed alerts as training sample.
  - Enhanced version of the initial training sample

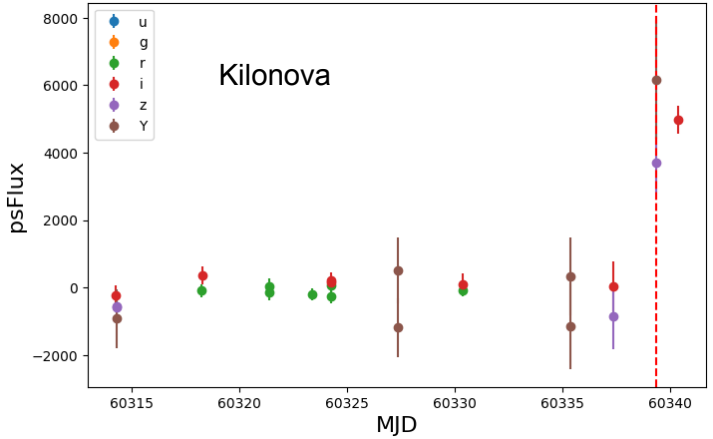
CBPF

Example Light Curves



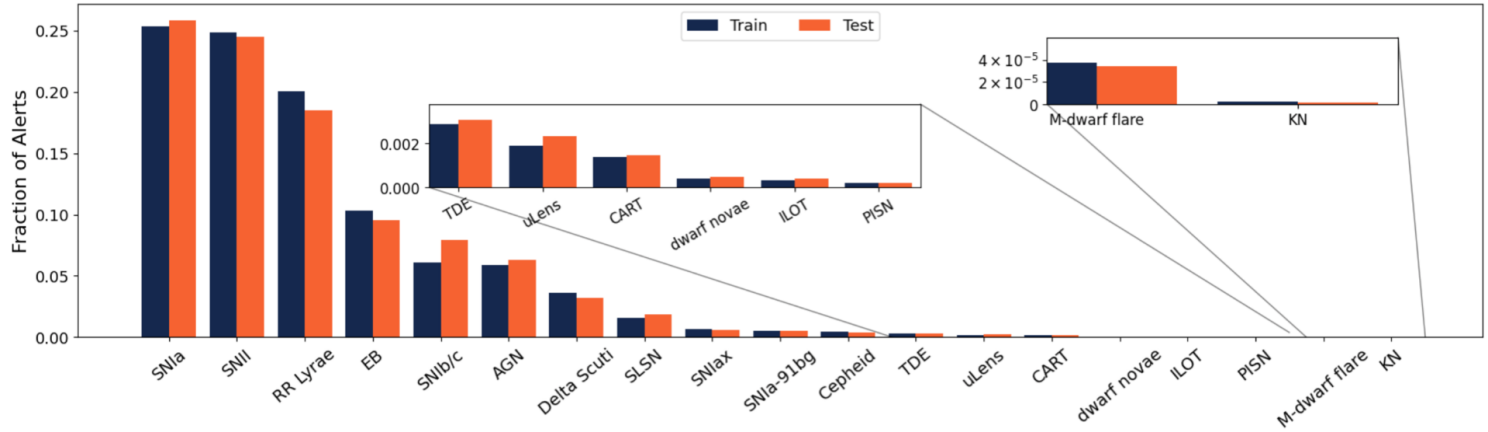


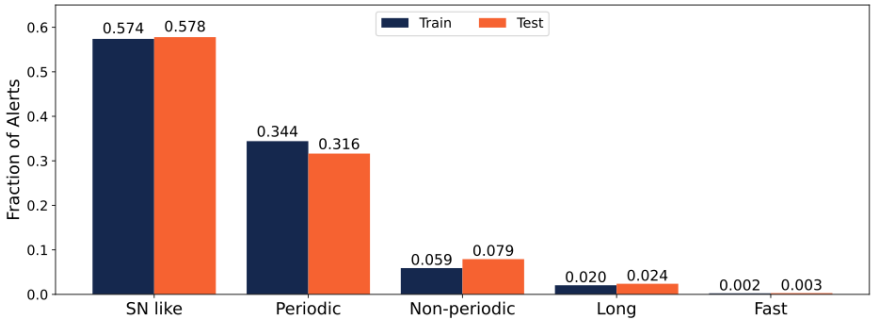


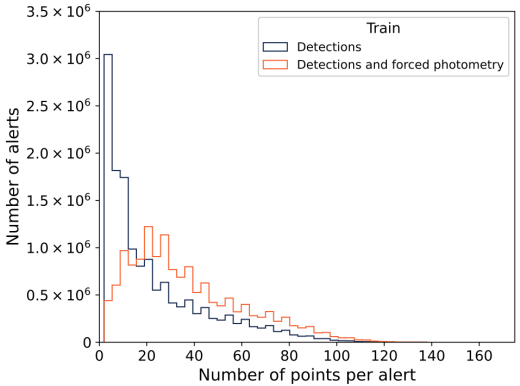


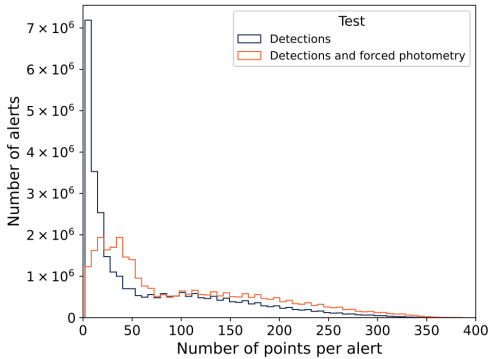
Statistics for ELASTiCC

17,233,868 alerts - 1,676,431 unique objects for training
34,872,745 alerts - 2,865,642 unique objects for testing









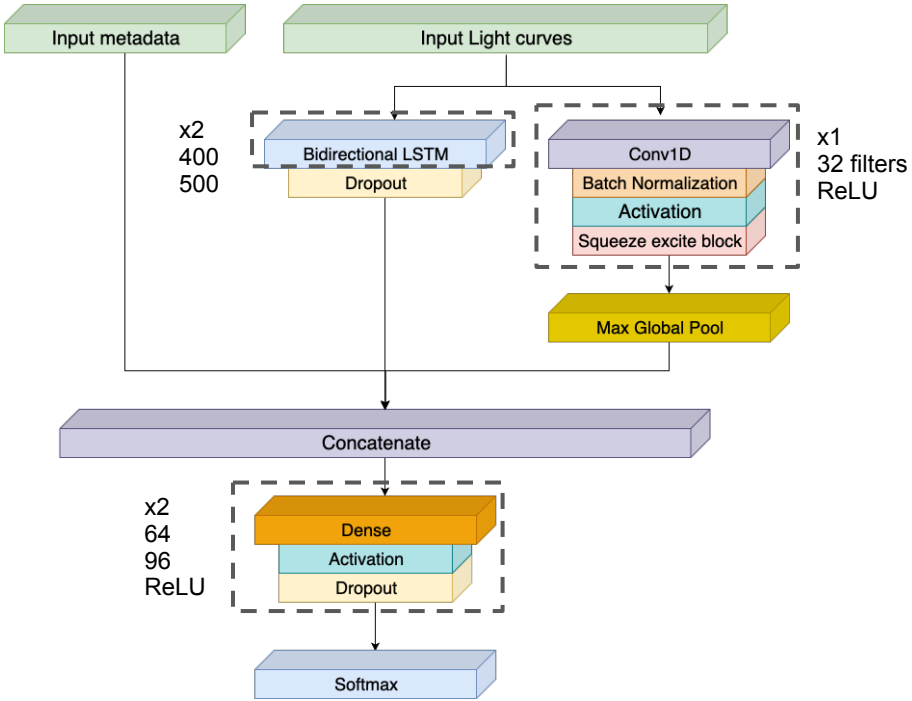
Fink Classifiers

- **CATS (CBPF Alert Transient Search) Broad Classifier**
- SuperNNova (Anais Möller) - binary and broad class
- SLSN (Etienne Russeil)
- Early SNIa (Marco Leoni, Emille Ishida)

CBPF

CATS

- Broad Classifier: Uses 5 superclasses from the taxonomy
- Multivariate LSTM Fully Convolutional Network adapted to two different inputs
- Hyperparameter and architecture optimization
- Light curve: gap in MJD to the first point, normalized flux, normalized flux error, filter
- Metadata Used:
  - Extinction
  - Host galaxy photoz + error
  - Transient z + error



CBPF

CATS - Training

- 5-fold cross-validation
- 80% of the unique objects
- Split the unique IDs: alerts of a given object are either in training or validation
- Alerts + Forced photometry (only LCs with more than 1 point)
- Model with the lowest validation loss across all folds chosen as the best

CBPF

SuperNNova

- SuperNNova:
  - Recurrent Neural Networks with Long-Short Term Memory
- Detection + forced photometry used to make light curve data (+preprocessing)
- Metadata used:
  - Host galaxy redshift
  - Milk way extinction
- Binary and Multi-class
- Balanced training set

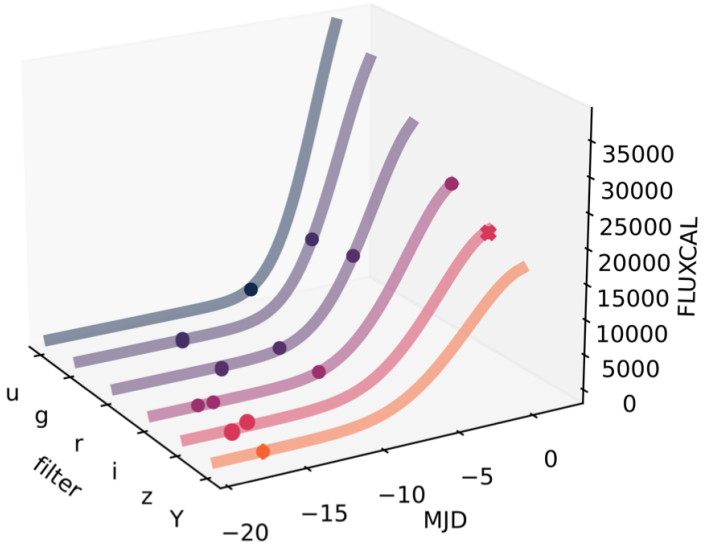
CBPF

Early SNIa

- Binary classifier using Random Forest algorithm.
- Catch SNIa during its rising phase (at least 2 bands)
- Use RAINBOW (Russeil et al. 2024) to fit all bands together

$$f(t) = \frac{\texttt{amplitude}}{1 + \exp(-\frac{t - t_0}{\texttt{rise\_time}})}$$

- Metadata used:
  - number of points
  - host galaxy redshift
  - separation between transient and host



CBPF

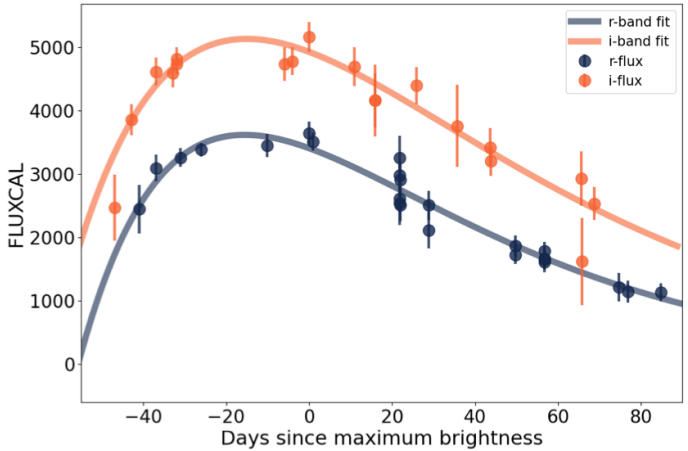
SLSN

- SLSN+PISN binary using Random Forest
- Multi-view symbolic regression (Russeil et al. 2024) on ZTF light curve of a SLSN candidate (*r* and *i* bands)

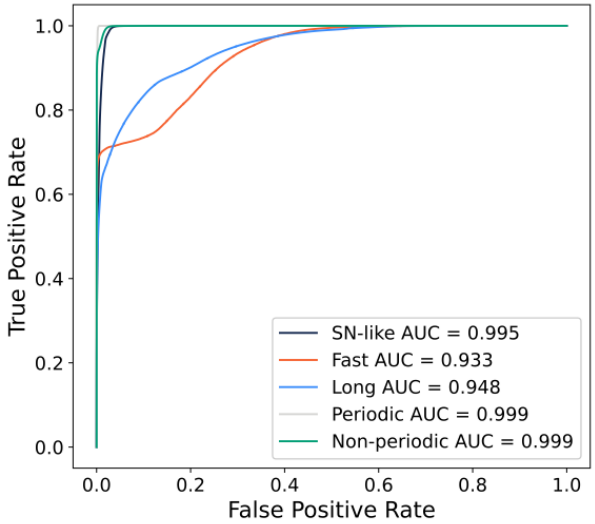
$$f(t) = A(t - t_0) \times e^{-\frac{t-t_0}{t_{fall}}}$$

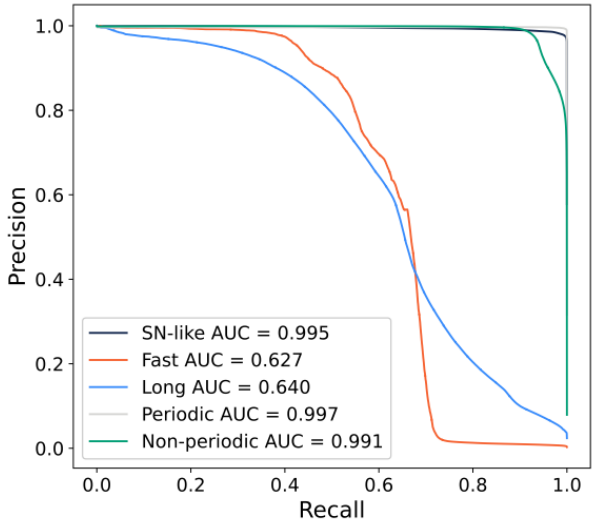
- Metadata used:
  - RA, DEC., host galaxy z (+ error), host galaxy distance

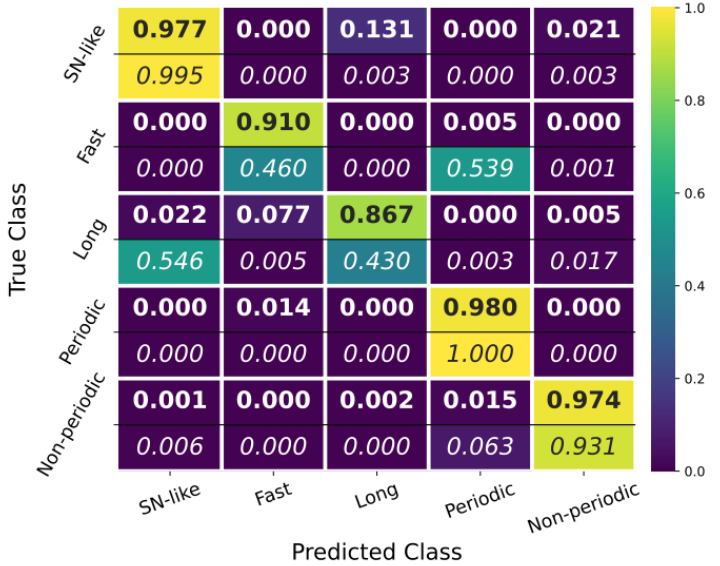
- Active learning loop



Results on Y2 and Y3

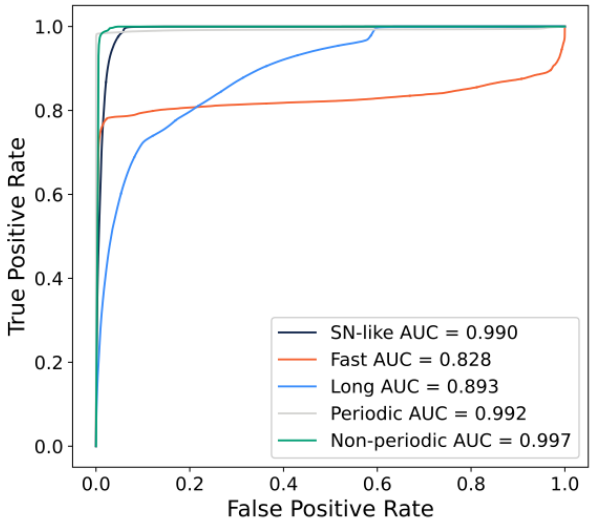


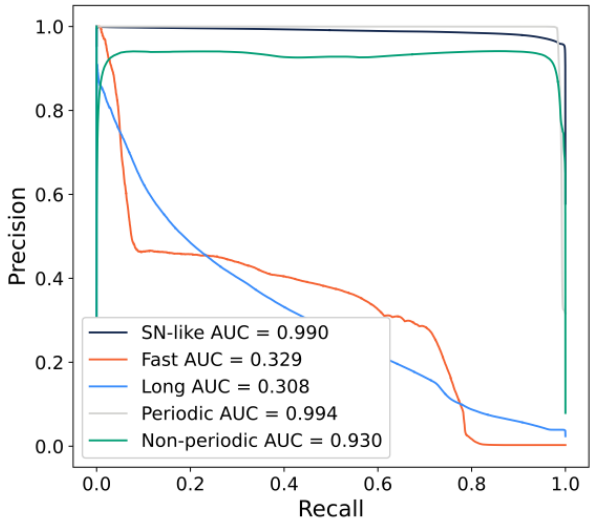


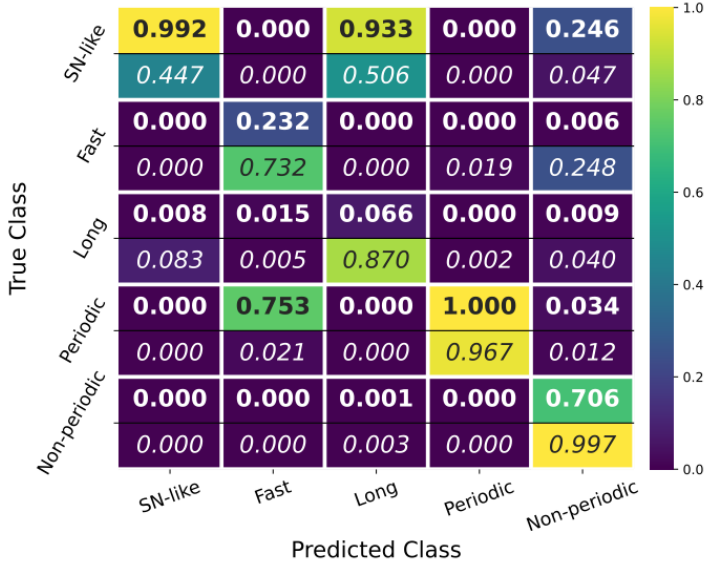


(a) CATS

Results on Y2 and Y3

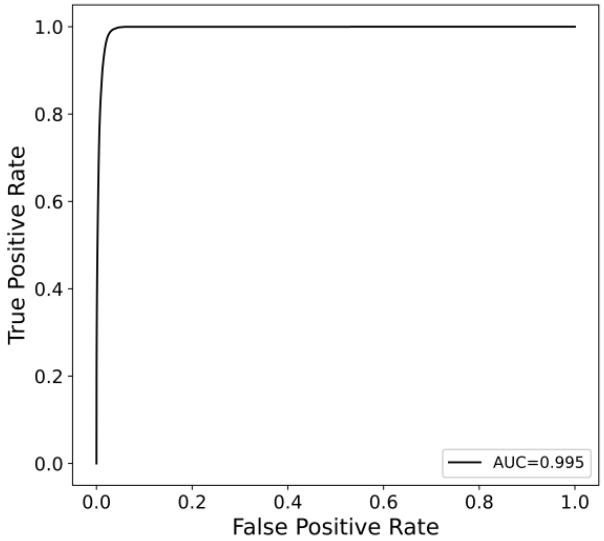


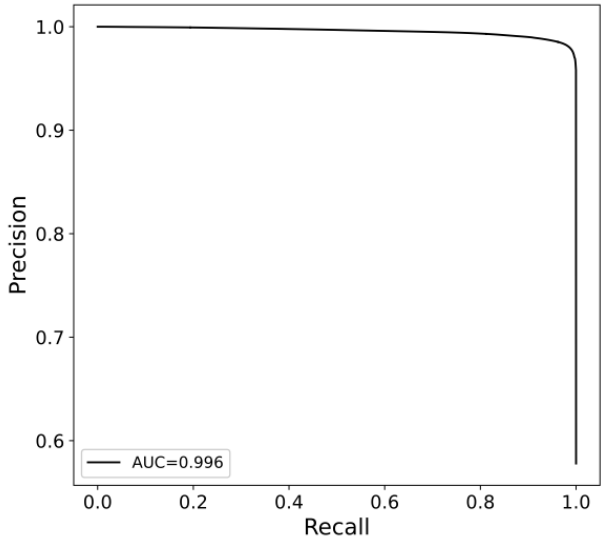


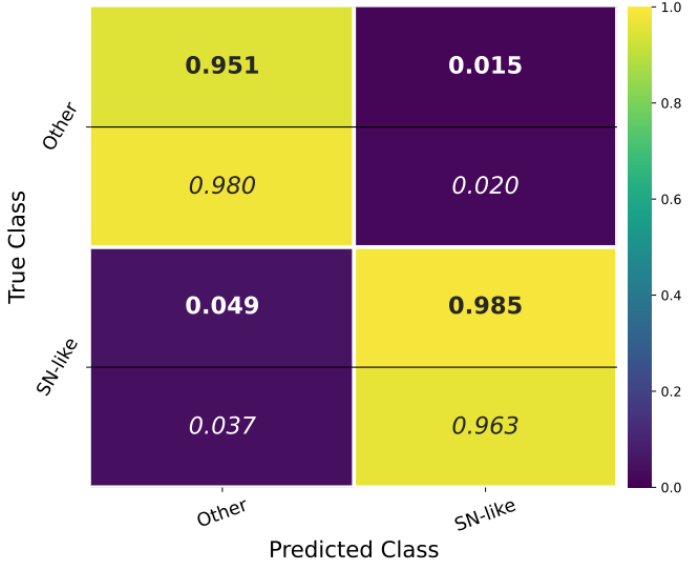


(b) SuperNNova broad

Results on Y2 and Y3

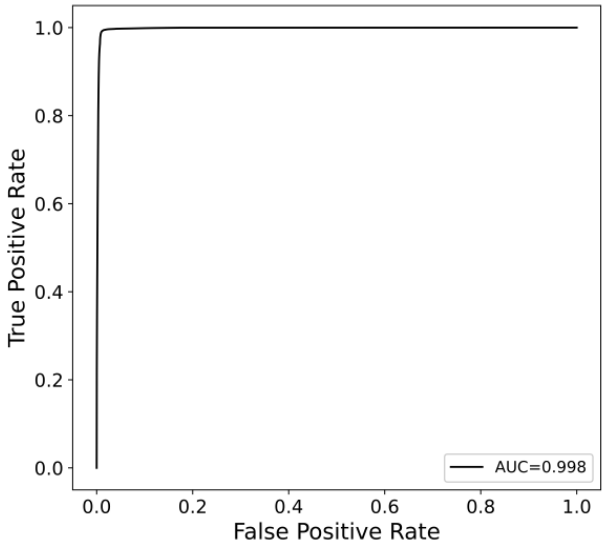


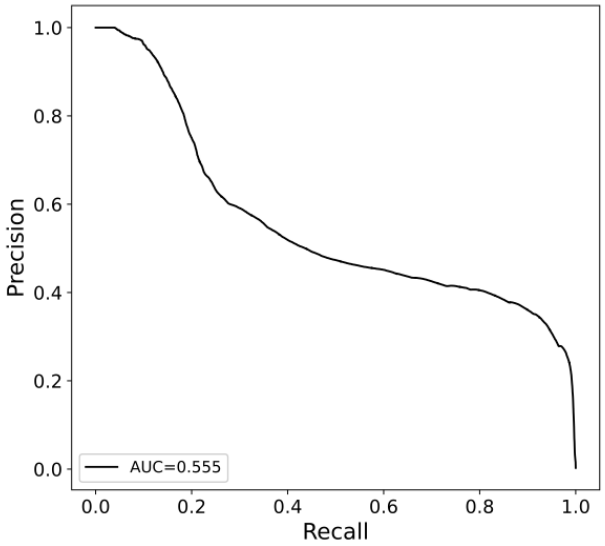


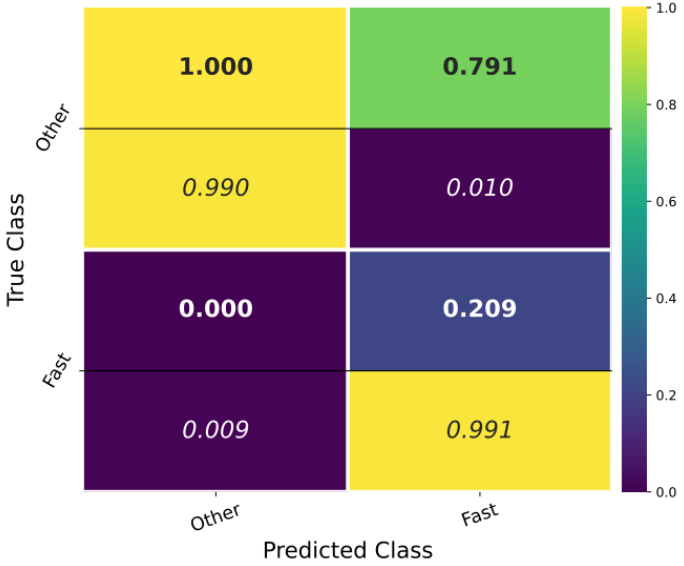


(c) SuperNNova binary for SN-like

Results on Y2 and Y3

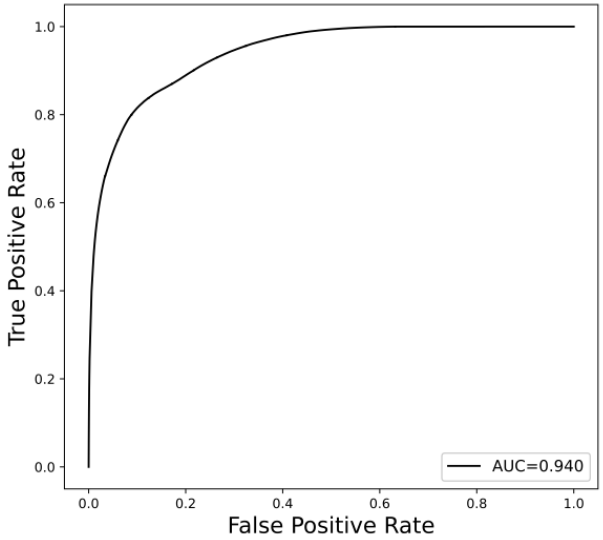


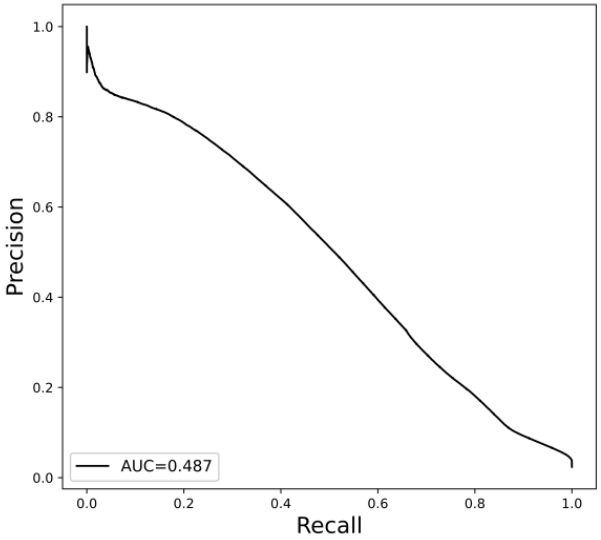


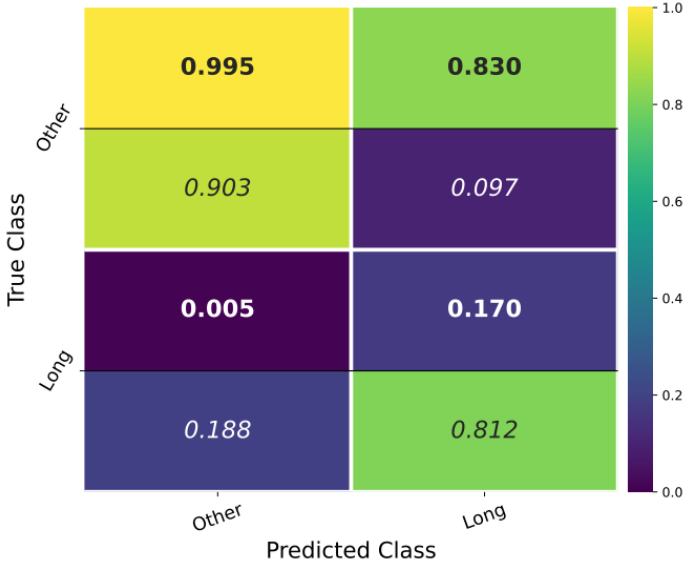


(d) SuperNNova binary for Fast

Results on Y2 and Y3

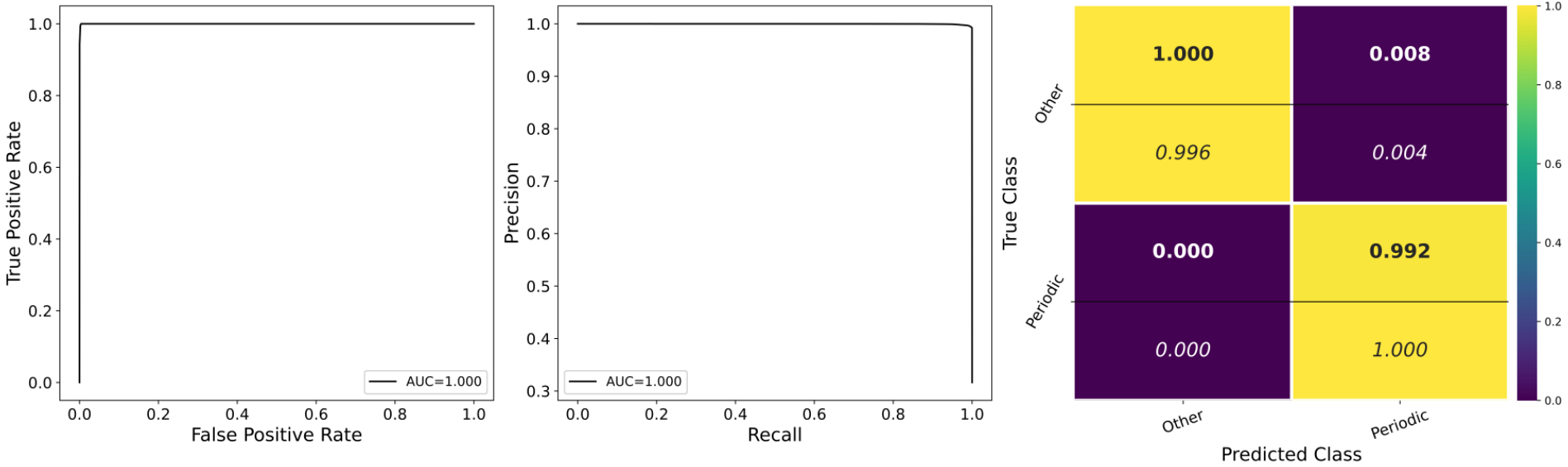






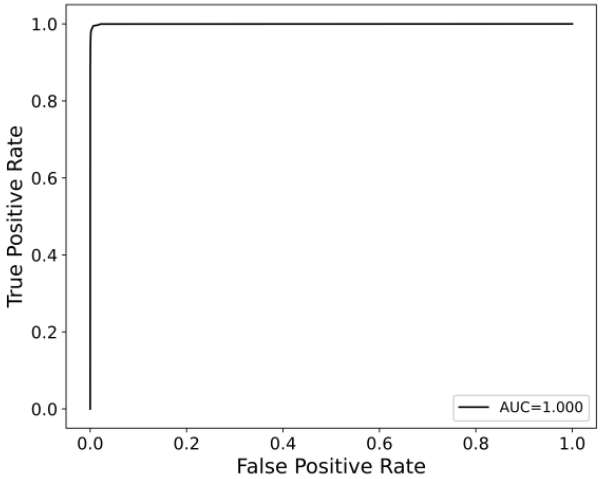
(e) SuperNNova binary for Long

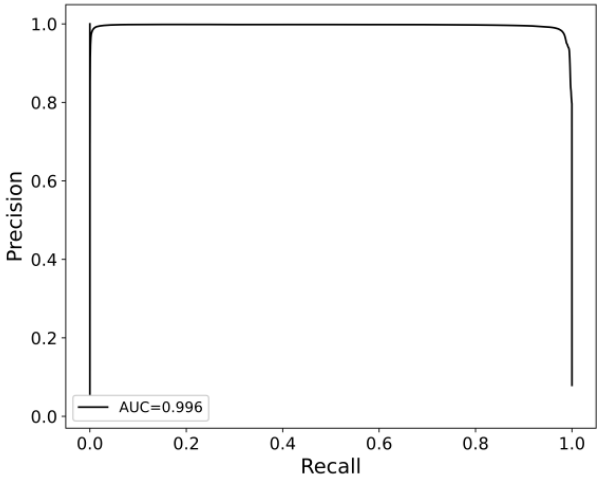
Results on Y2 and Y3

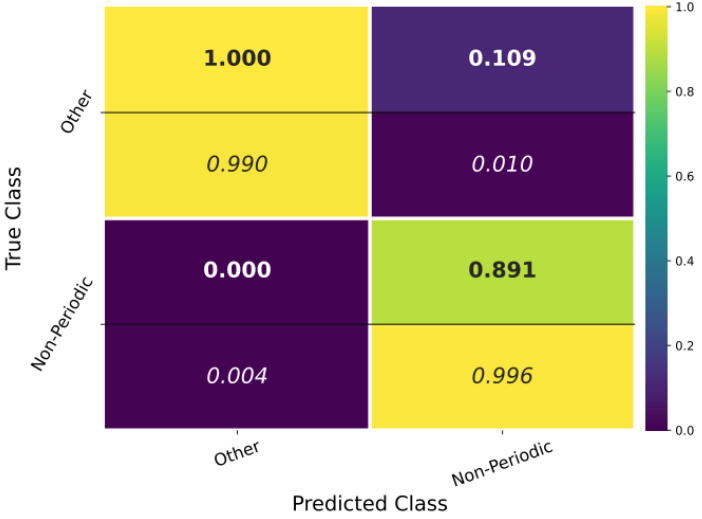


(f) SuperNNova binary for Periodic

Results on Y2 and Y3

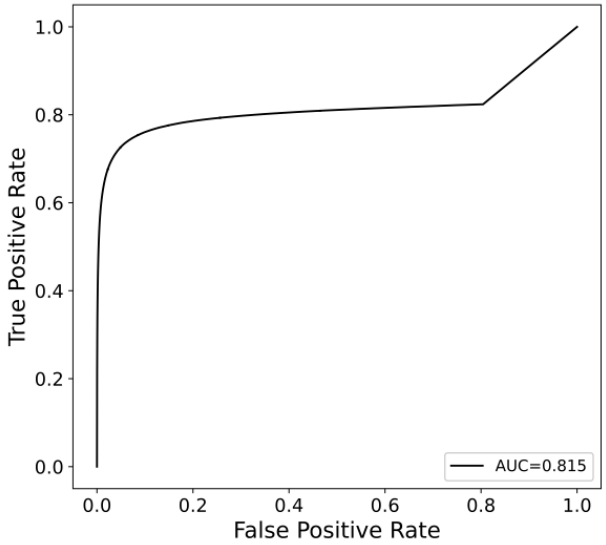


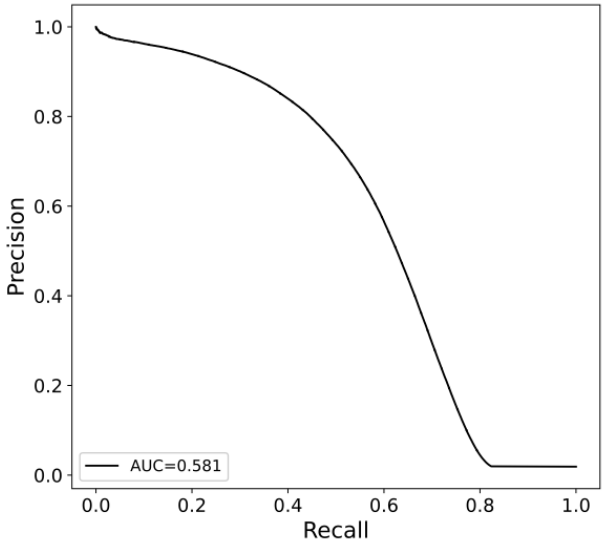


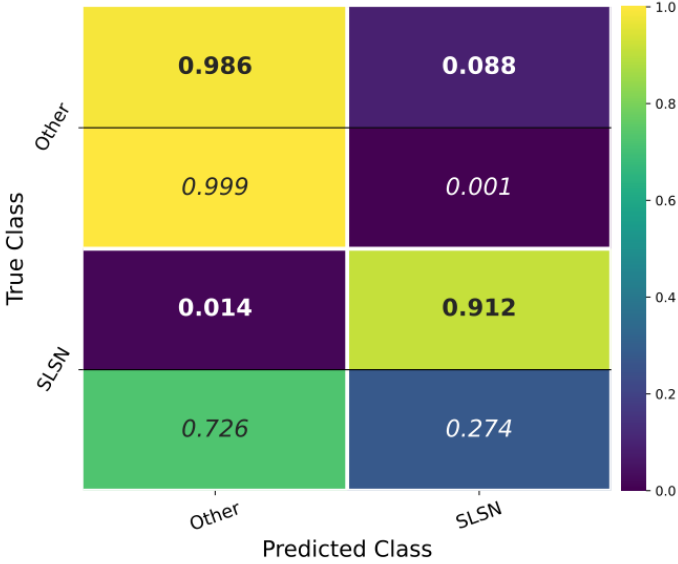


(g) SuperNNova binary for non-Periodic

Results on Y2 and Y3

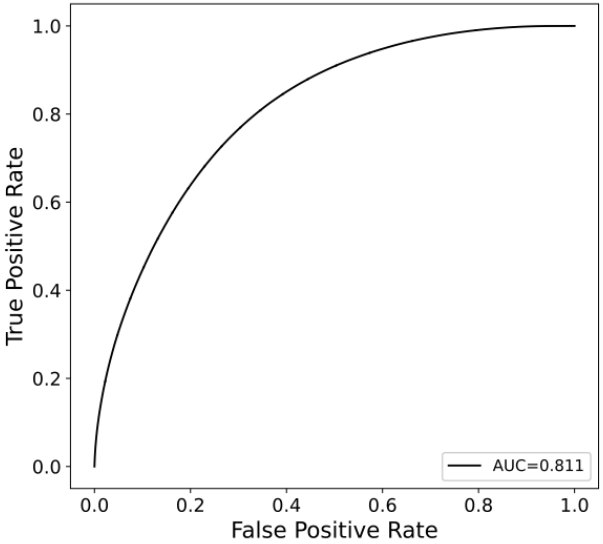


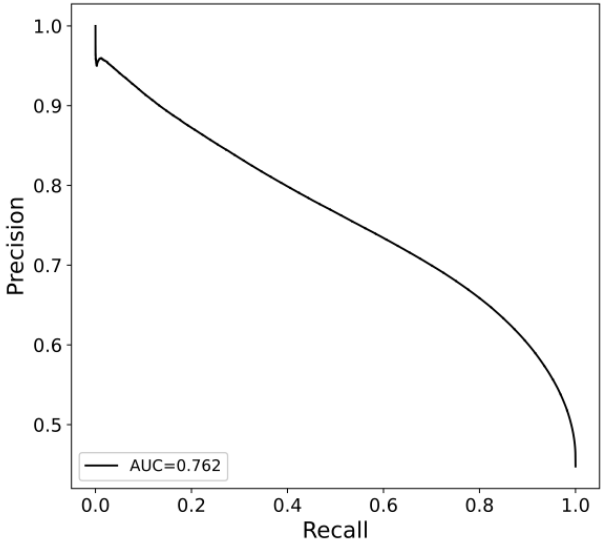


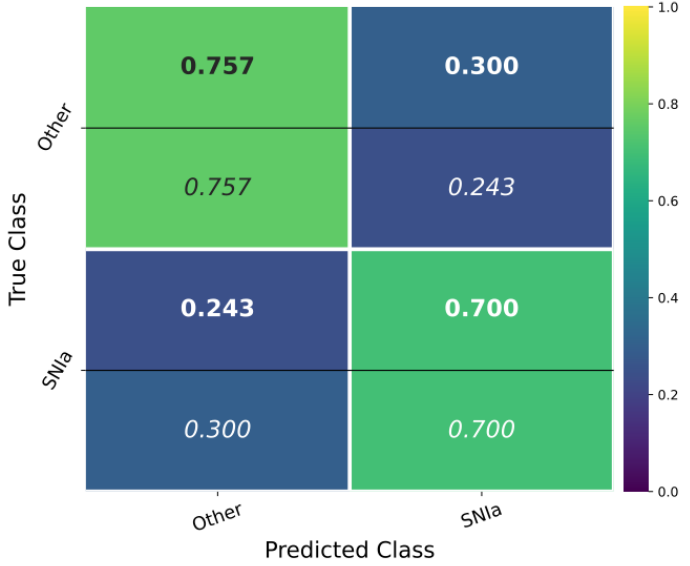


(h) SLSN

Results on Y2 and Y3

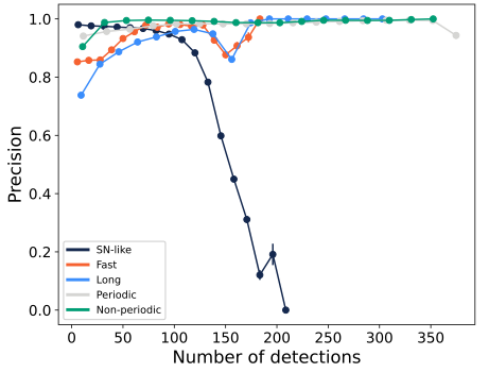


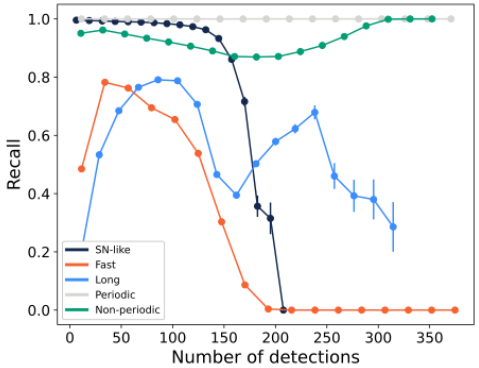


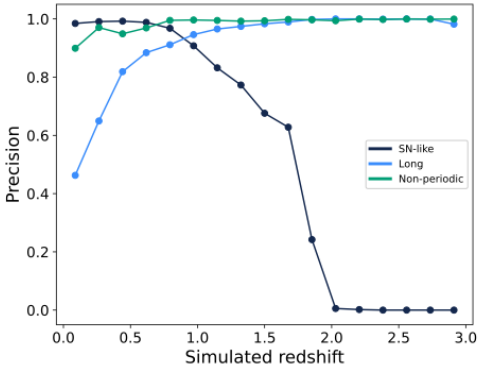


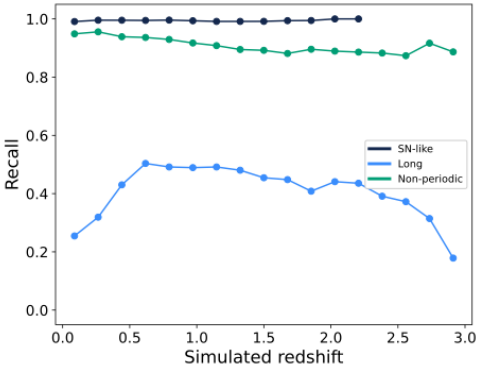
(i) EarlySNIa

Evolution of metrics





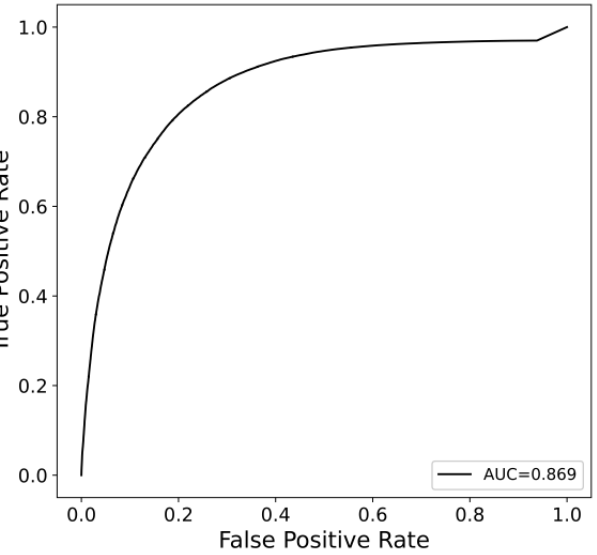


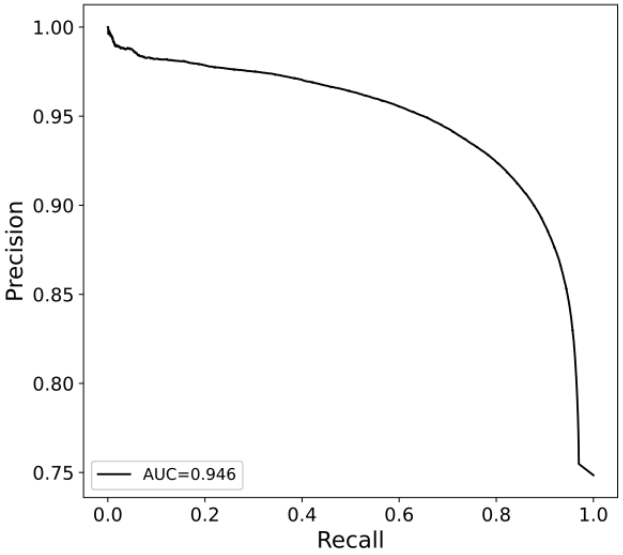


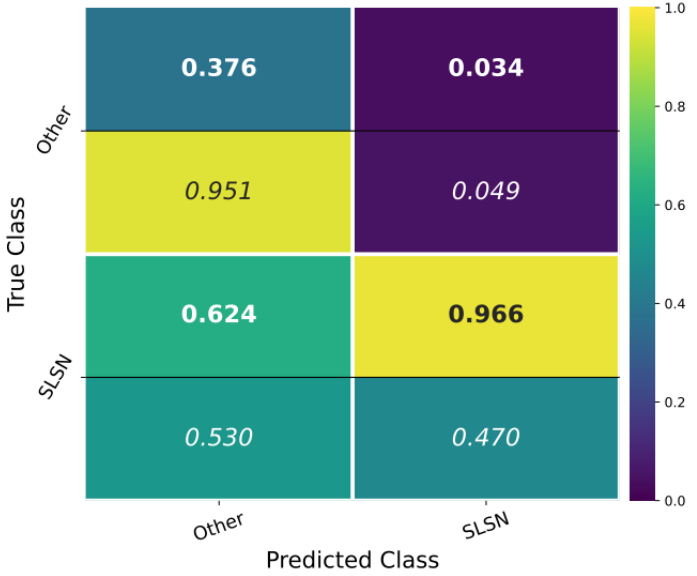
(a) CATS

Combining classifiers - CATS as a first step

- SLSN only sees alerts classified as Long by CATS

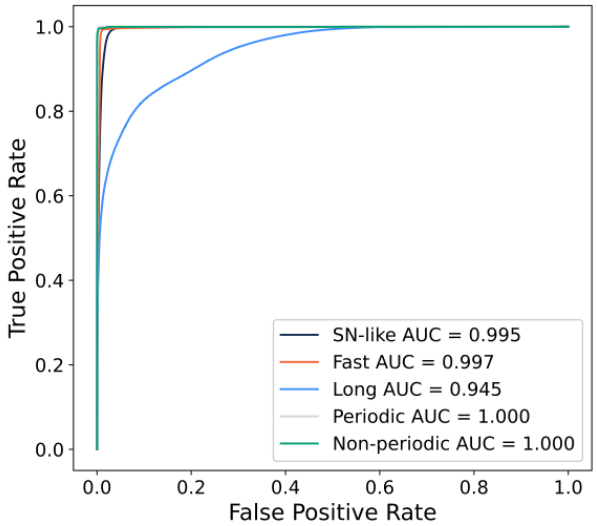


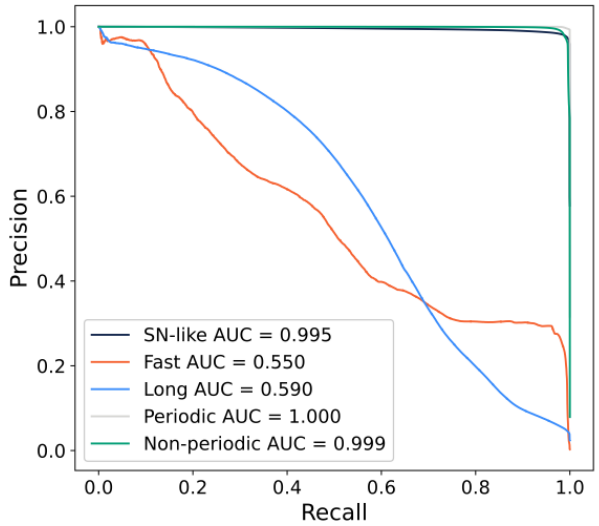


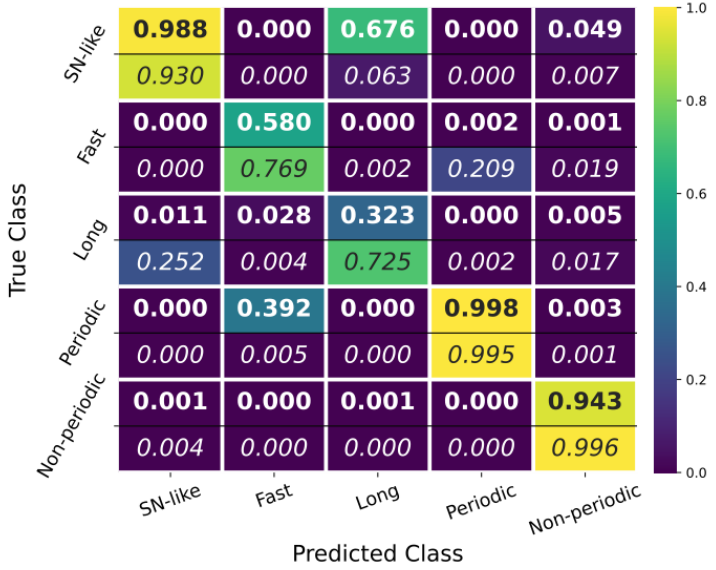


Combining classifiers - SNN binaries

- Combine all SNN binary classifiers into one multi-class

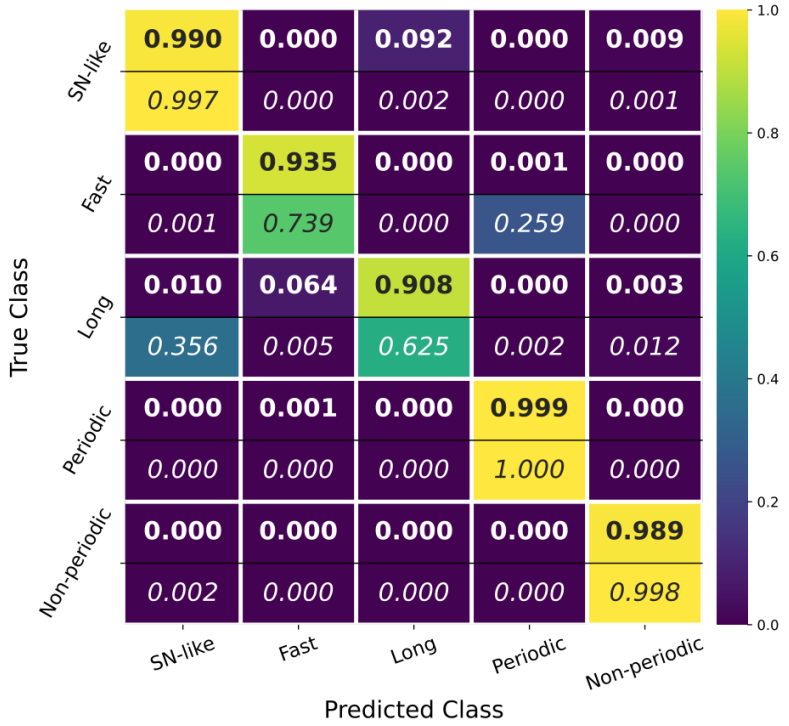






Combining classifiers - increasing purity

- Alerts where CATS and the combined SNN agree
- 94% of the test sample
- Significant improvement for the Long and Fast classes



Takeaways

- Data is heterogeneous, unbalanced classes - results are satisfactory
- Fast and Long classes are the hardest to classify. Choice of split?
- Confusion between classes is the same - similarities intrinsic to the dataset
- CATS is able to classify objects with high purity with less than 10 detections, and for all redshifts
- Combining specialist binary classifiers could yield better results than multi-class
- Hierarchical classification possible with a broad classifier as a first step
- Results improve by combining broad classifiers

CBPF

Fink infrastructure performance

- Engineering team monitored the classifiers' performance - throughput and memory use
- LSST will stream thousands of alerts every 30 seconds - results must come fast
- Fink computing infrastructure impose limits
- Real-time processing - 82% of the alerts classified by all 9 classifiers in less than 30 seconds, 90% in less than a minute. No optimization attempted in this work!
- New service available to access ELAsTiCC data now through Fink!

CBPF

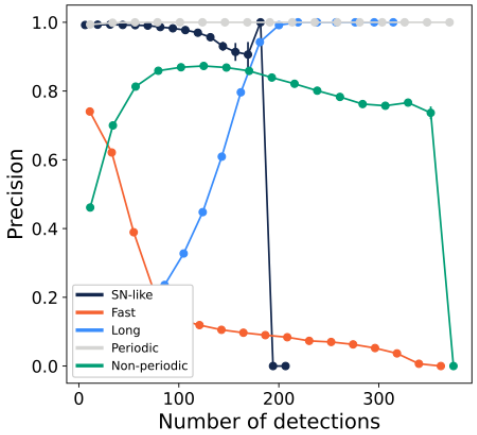
Conclusion

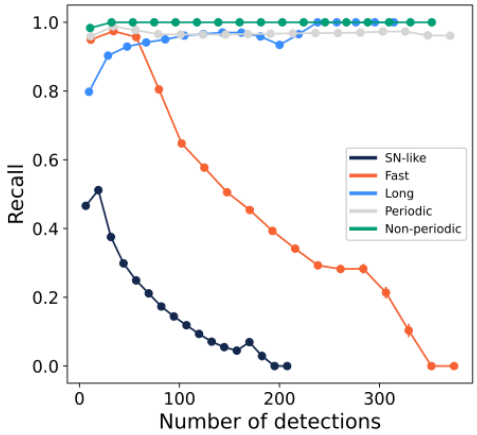
- Fink has several science products available for whoever's interested in ZTF data.
- LSST will be a different beast - hardware and ML models will need to be adapted
- Fink models and infrastructure are on the right path
- Easy access to data was paramount in this work - new service developed
- Fink is working to provide hardware for model training at scale
- 9 different classifiers can be used in different science cases - informed decision by the user

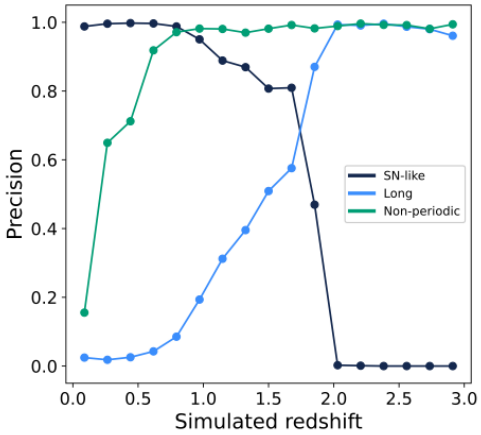
CBPF

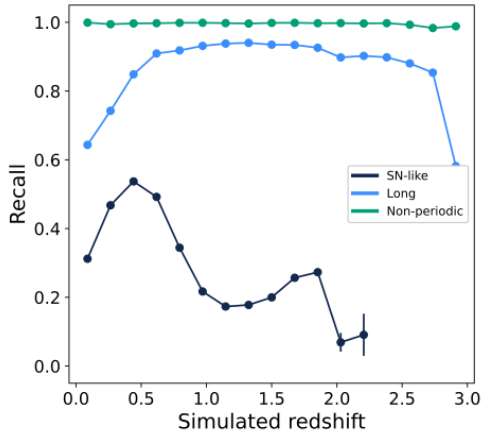
Thank you

Visit www.fink-portal.org

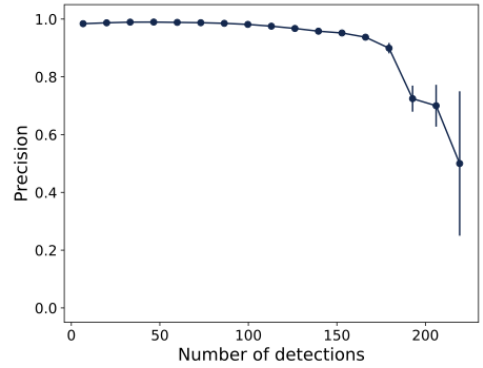


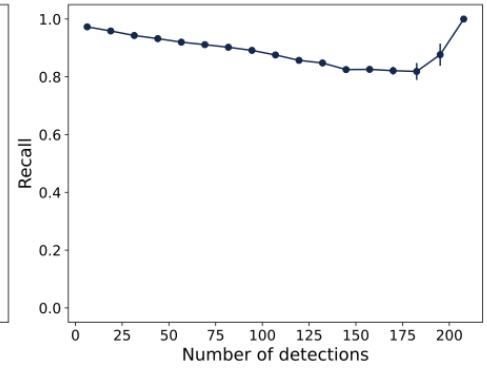


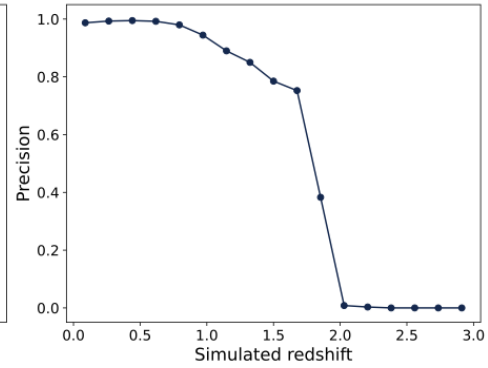


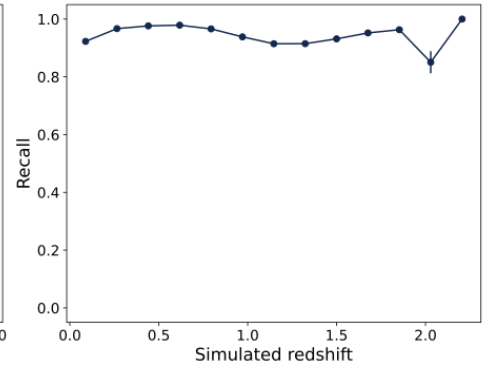


(b) SuperNNova broad

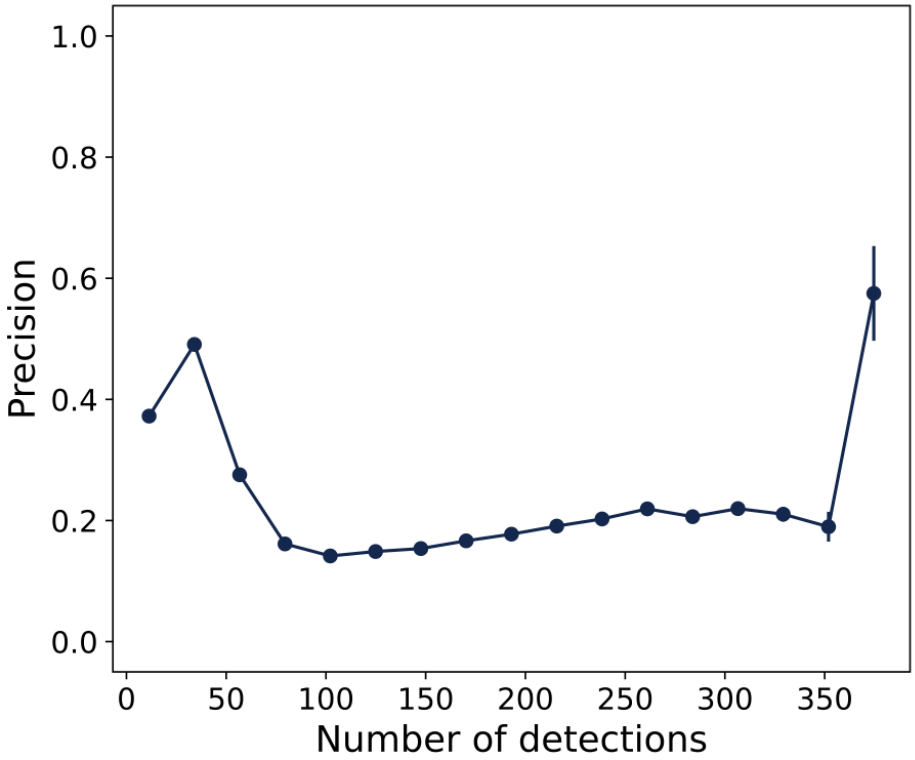


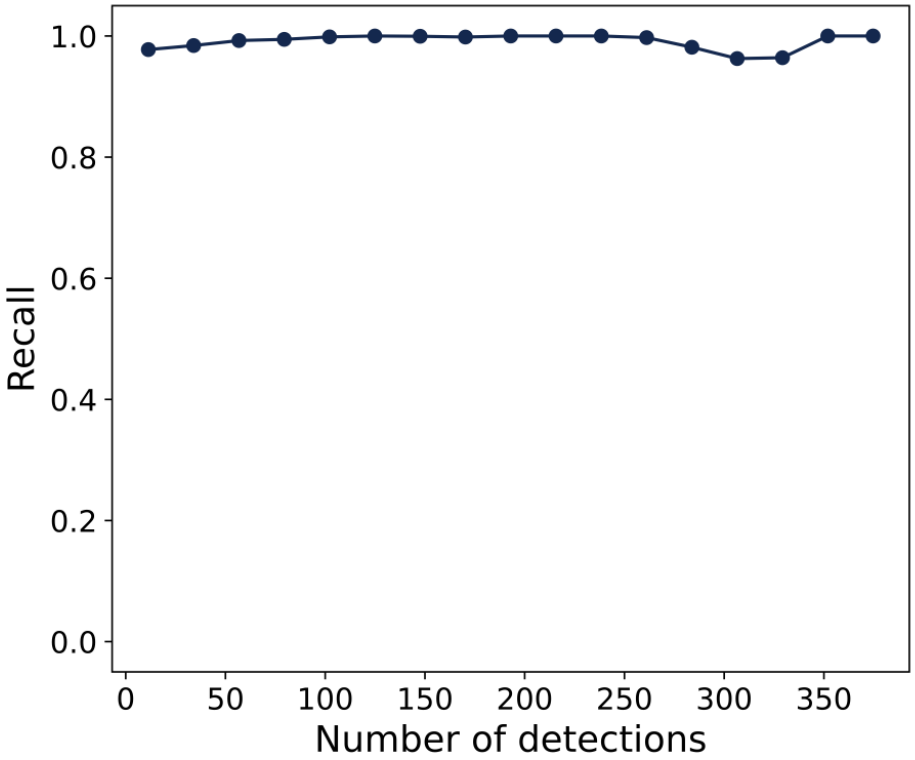




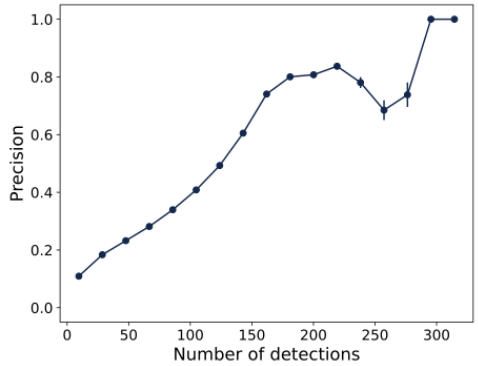


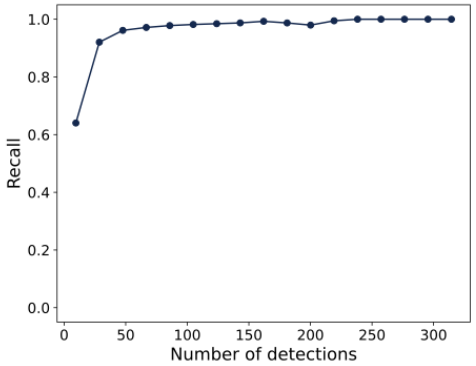
(c) SuperNNova binary for SN-like

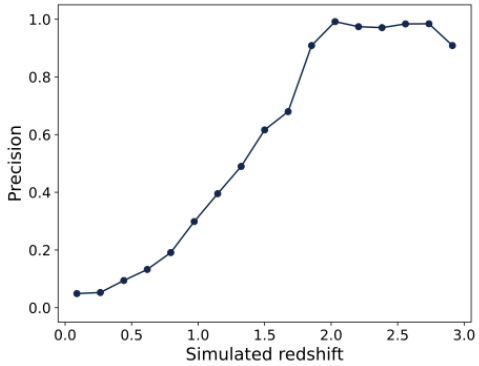


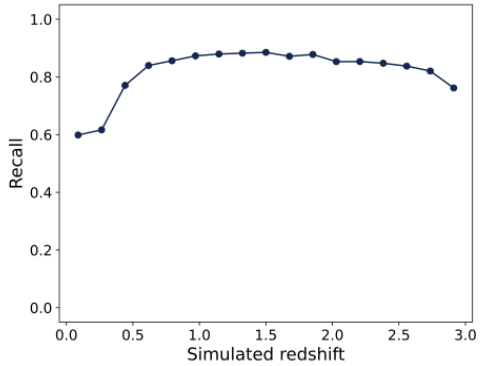


(d) SuperNNova binary for Fast

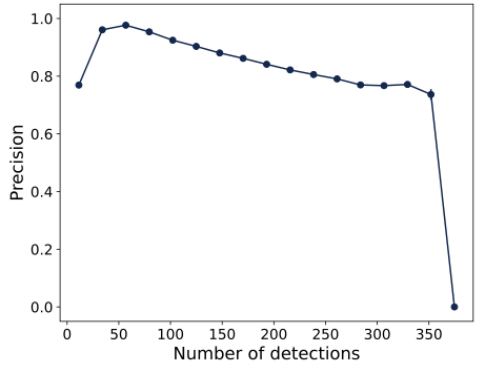


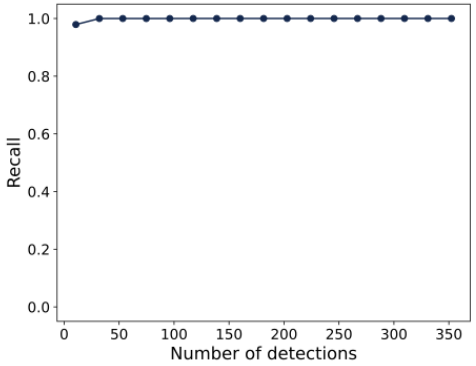


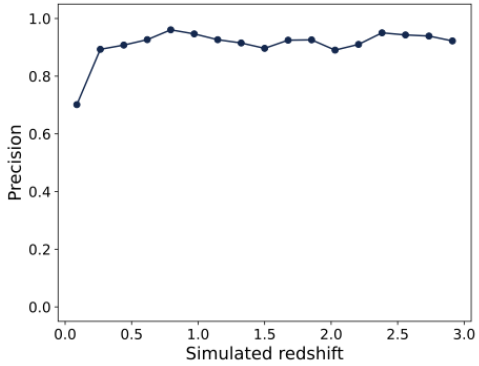


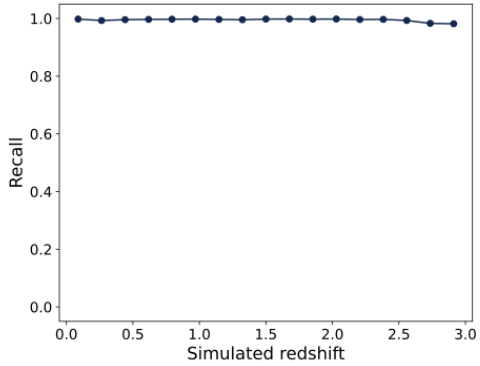


(e) SuperNNova binary for Long

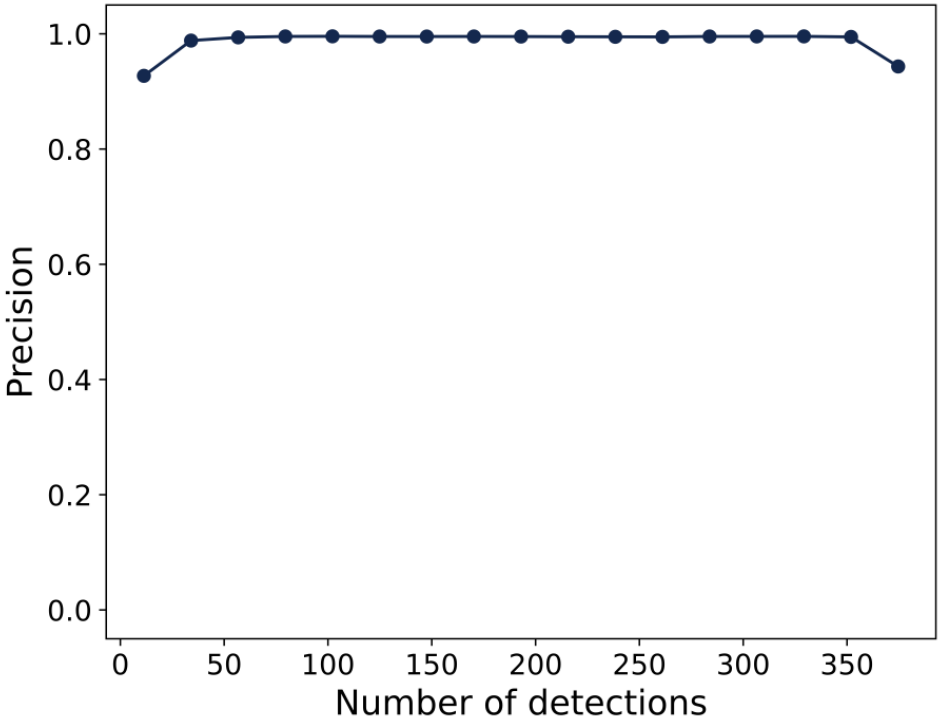


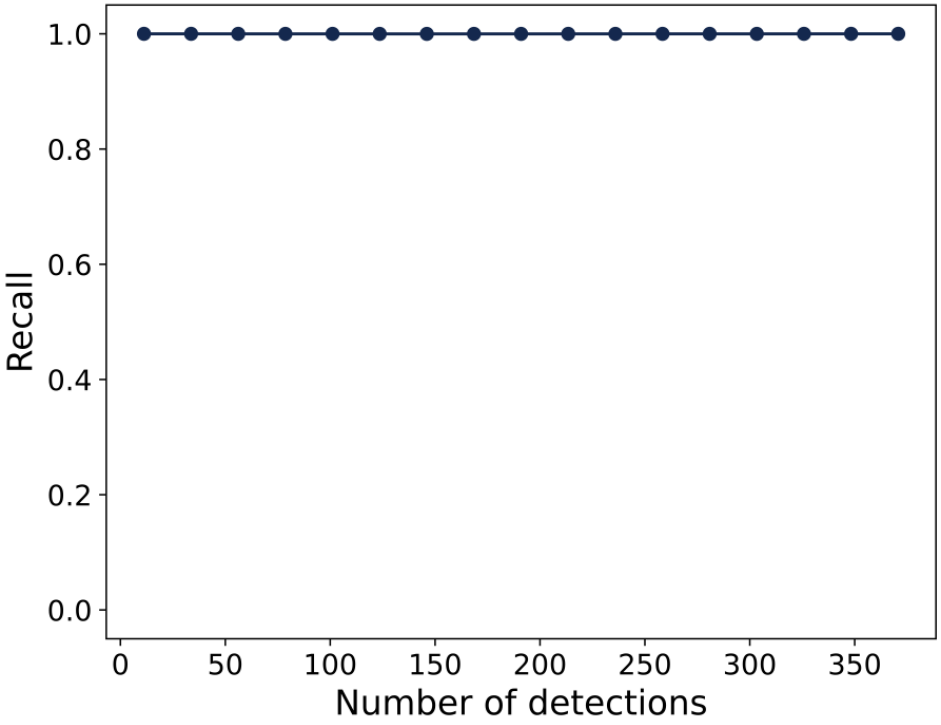




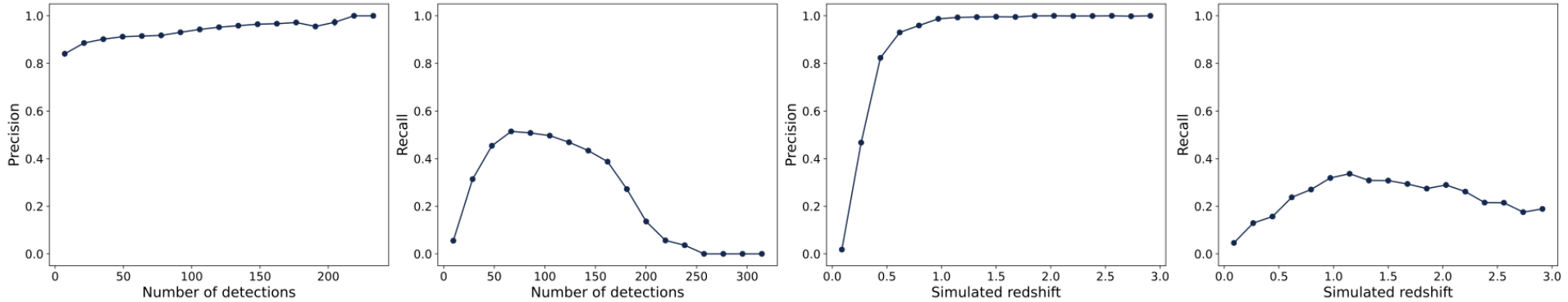


(f) SuperNNova binary for non-Periodic

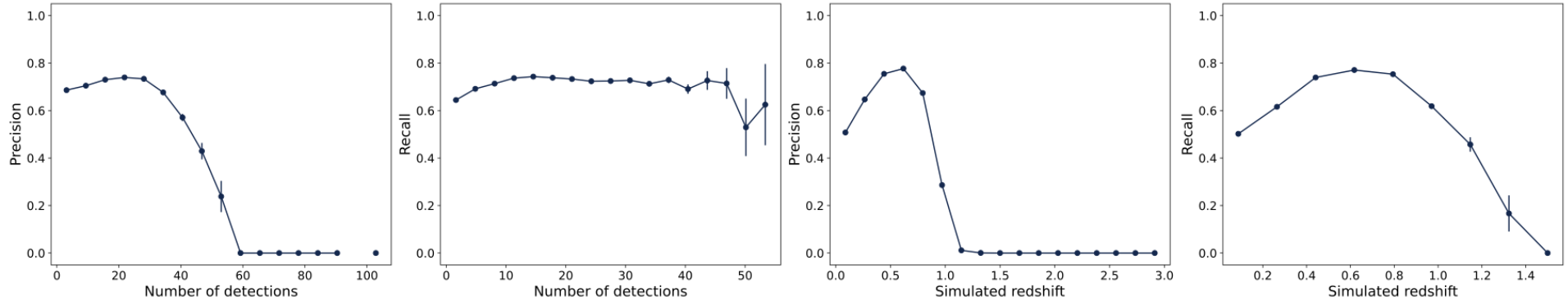




(g) SuperNNova binary for Periodic



(h) SLSN classifier



(i) EarlySNIa