ESO

ALeRCE
Automatic Learning for the
Rapid Classification of Events

Searching for different AGN populations in massive datasets with Machine Learning

Paula Sánchez Sáez, ESO Garching Fellow

Napoli, June 30th 2023

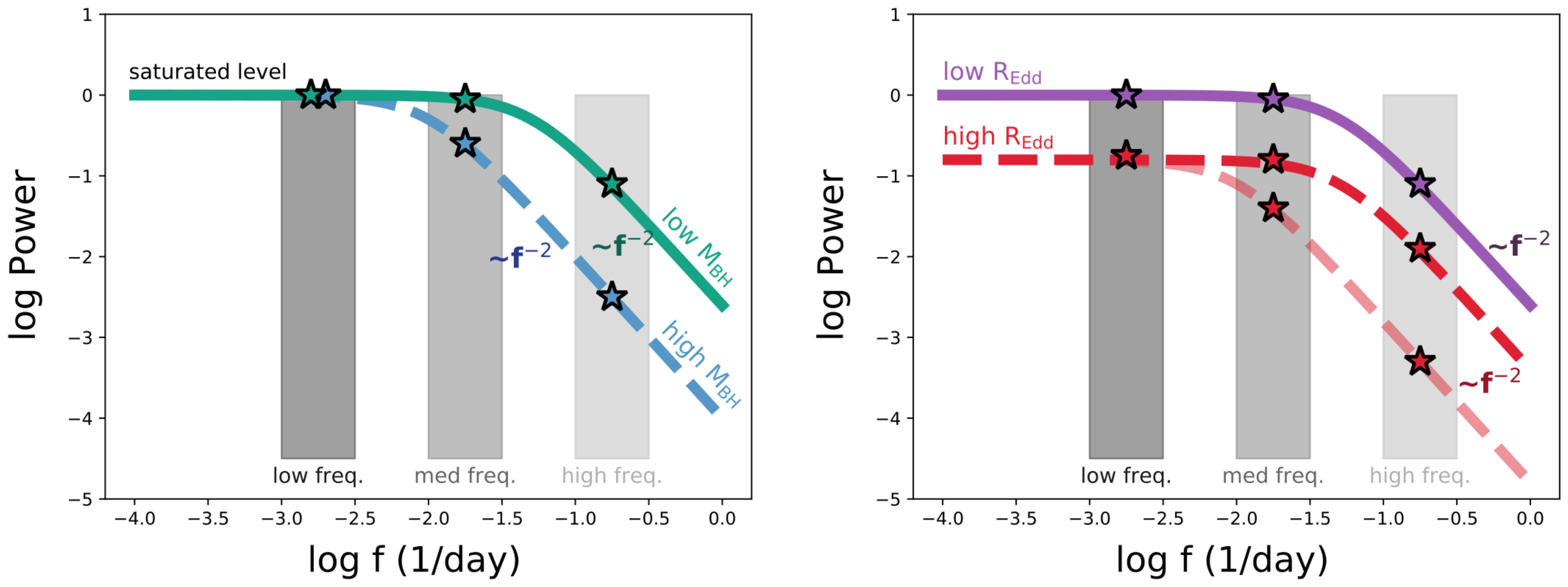
1. VARIABILITY AND COLOR–BASED AGN CLASSIFIER

2. SEARCHING FOR CSAGNS WITH ANOMALY DETECTION

1. VARIABILITY AND COLOR-BASED AGN CLASSIFIER

Variability-based selection of AGN candidates

The variability properties (i.e., PSD normalization and breaking time scale) are correlated with the AGN physical properties



Arévalo et al. 2023a, 2023b, submitted (**5400 sources**): correlation between timescale of the variations and the black hole mass and accretion rate, and negative correlation between accretion rate and variability amplitude.

Variability selection is particularly efficient in finding low-mass and low-accretion rate candidates!

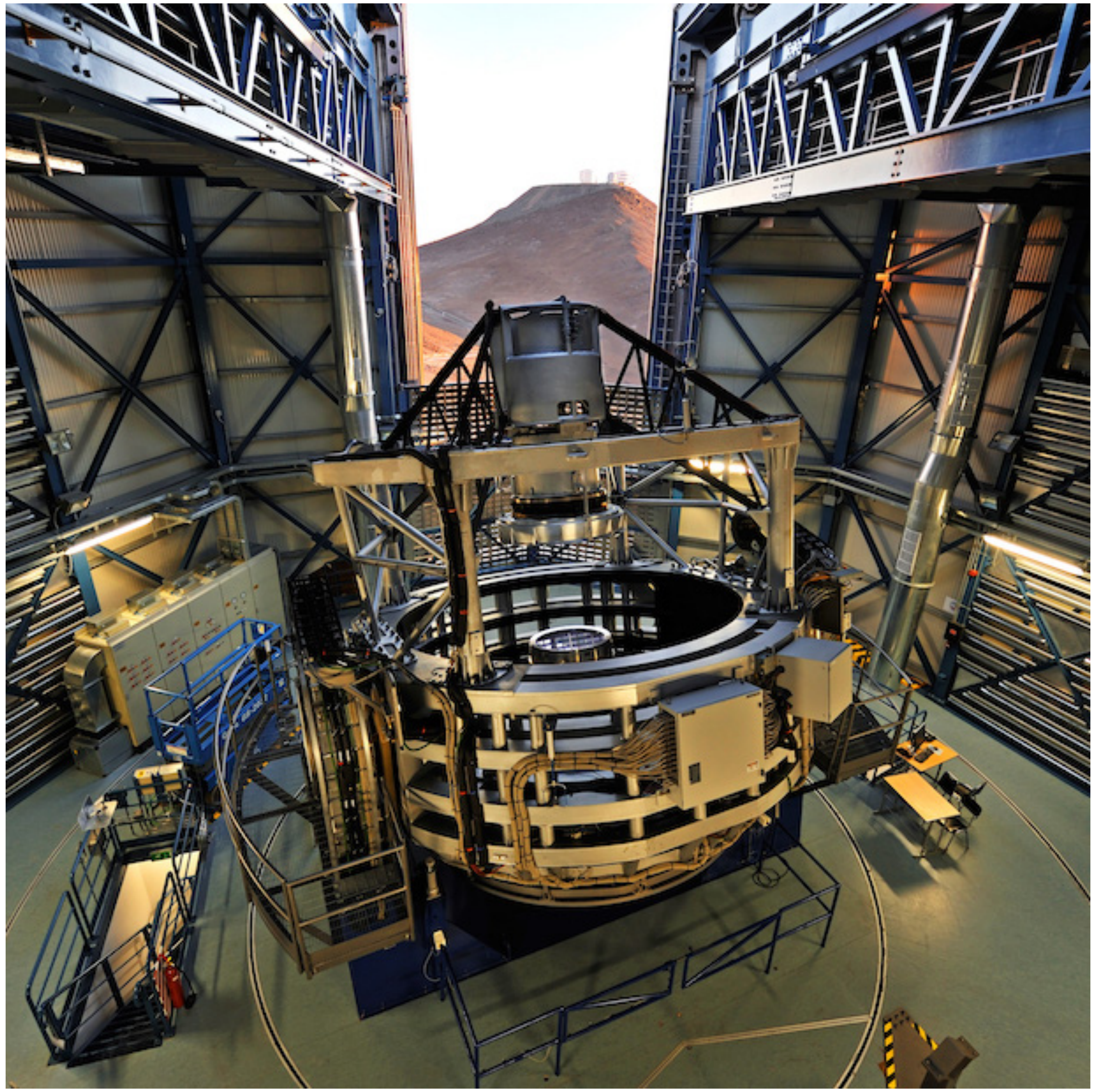
The 4MOST Chilean AGN/Galaxy Evolution Survey (ChANGES)

4MOST-ChANGES (PIs: Franz Bauer, Paulina Lira) will target a legacy sample of AGNs, **based on optical continuum variability** and SED selection from several existing surveys, and ultimately complemented by Rubin LSST to:

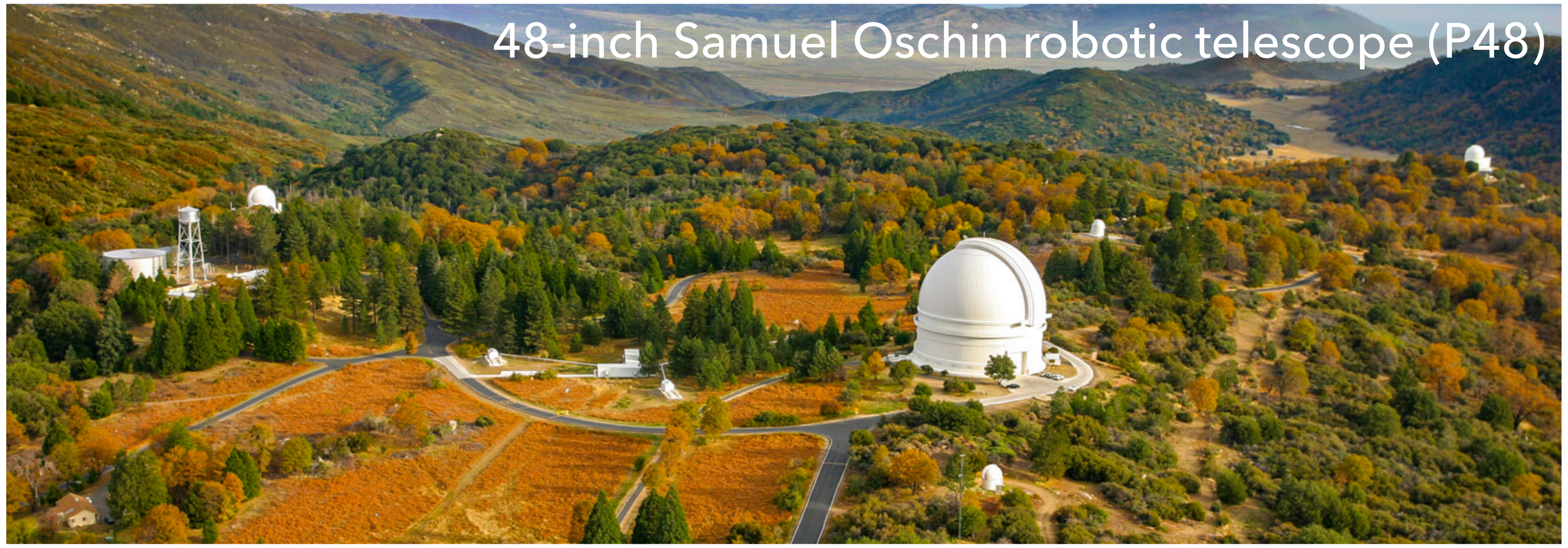
1) constrain the low-MBH, and low-L/LEdd end

2) investigate correlations among AGN (MBH, L/LEdd, UV slope, outflows, variability) and host properties (stellar age, metallicity, kinematics)

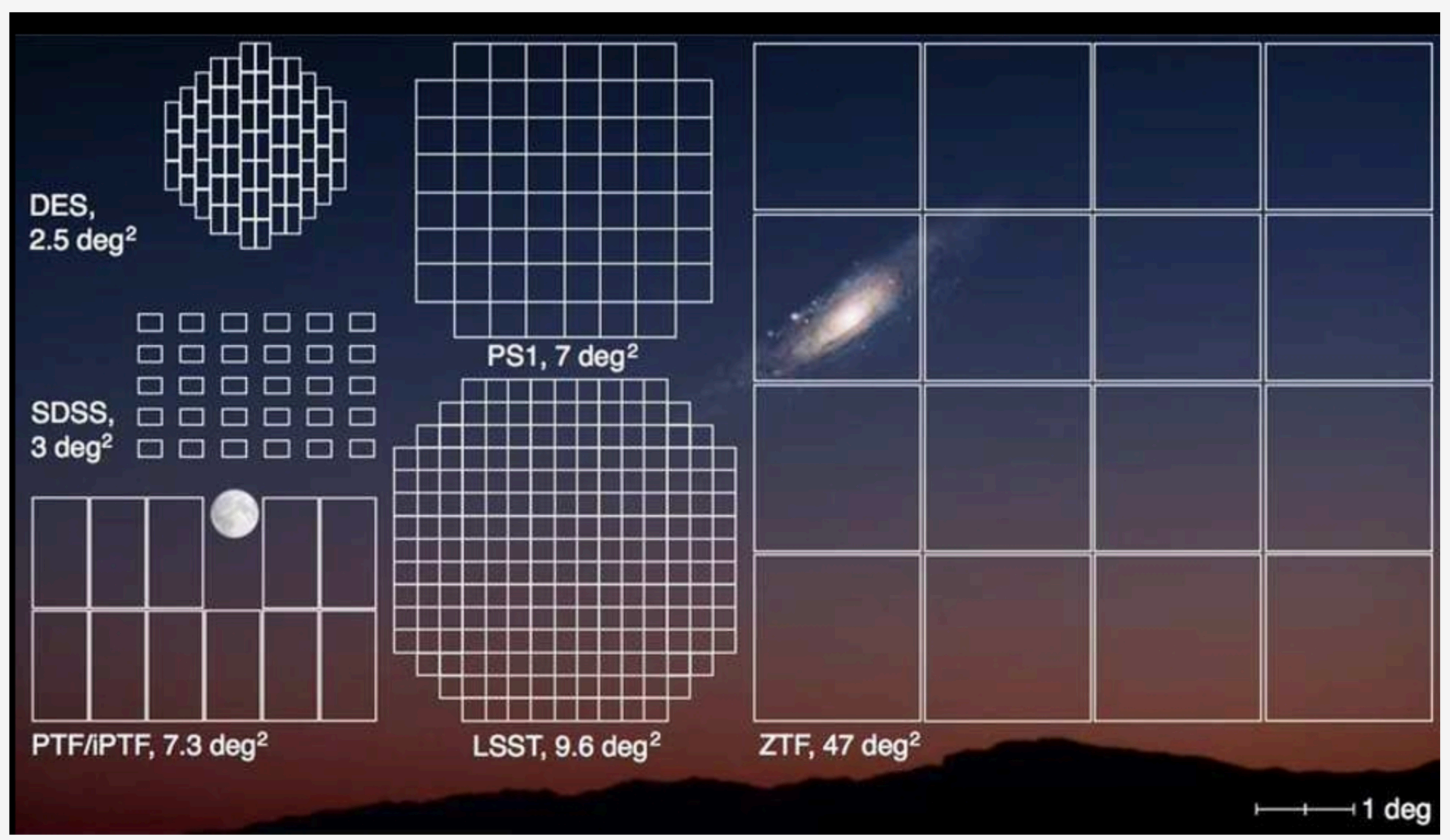
3) target/confirm rare BH subsamples (extreme variability, tidal disruption events, lensed, intervening absorption line systems)

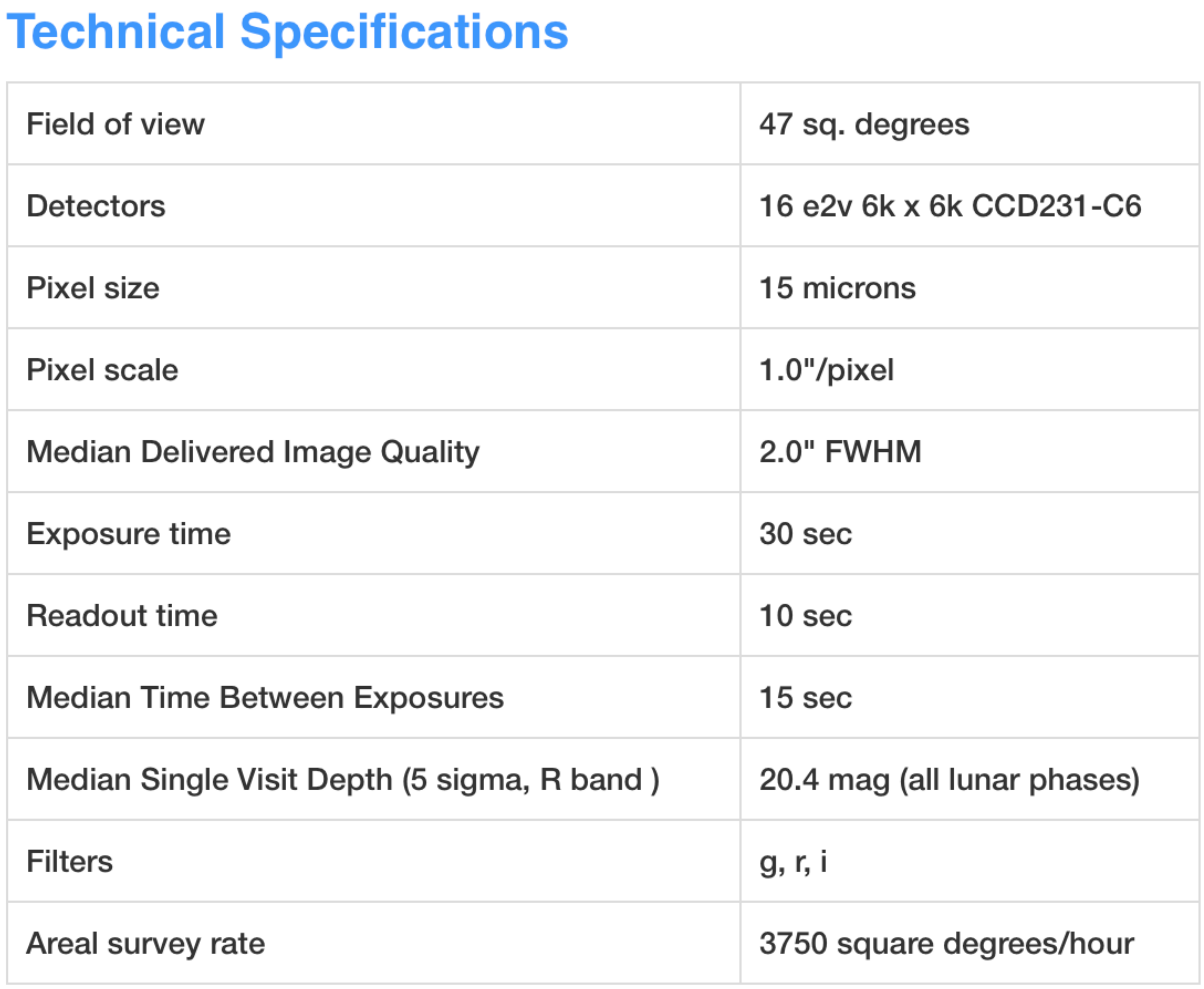


The Zwicky Transient Facility (ZTF)



Aerial shot of the Palomar Observatory in Southern California, USA Image credit: Palomar Observatory/Caltech



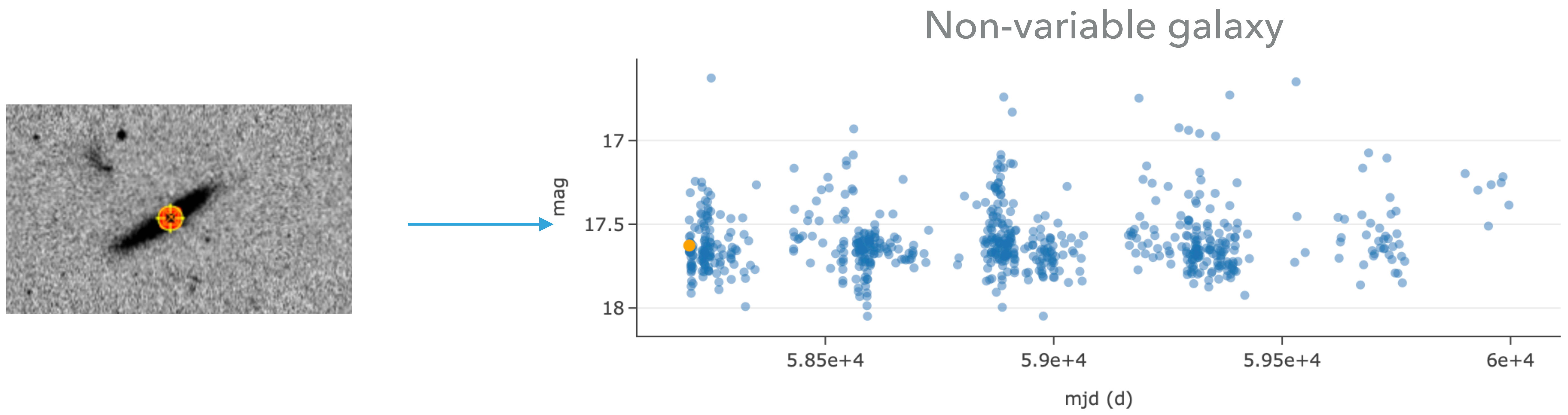


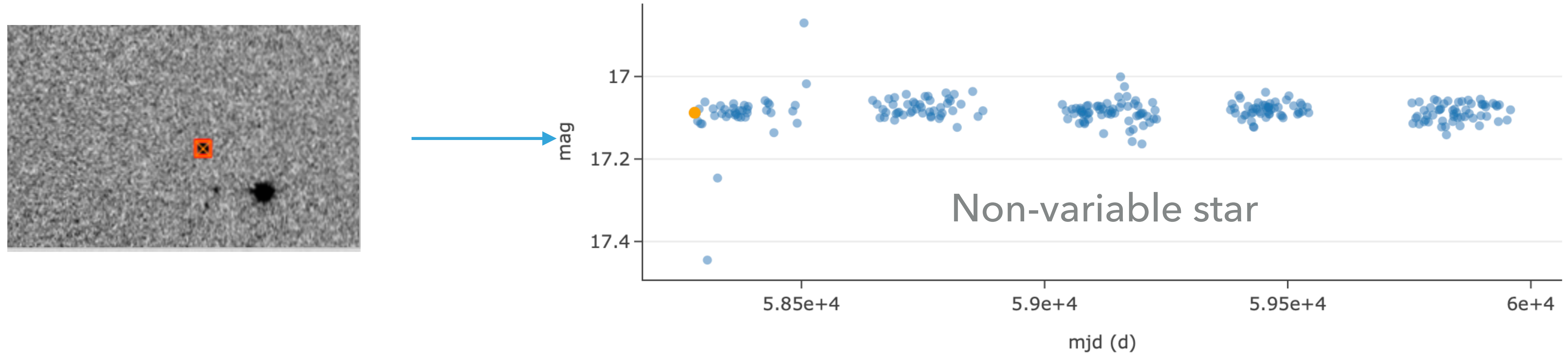
Technical Specifications

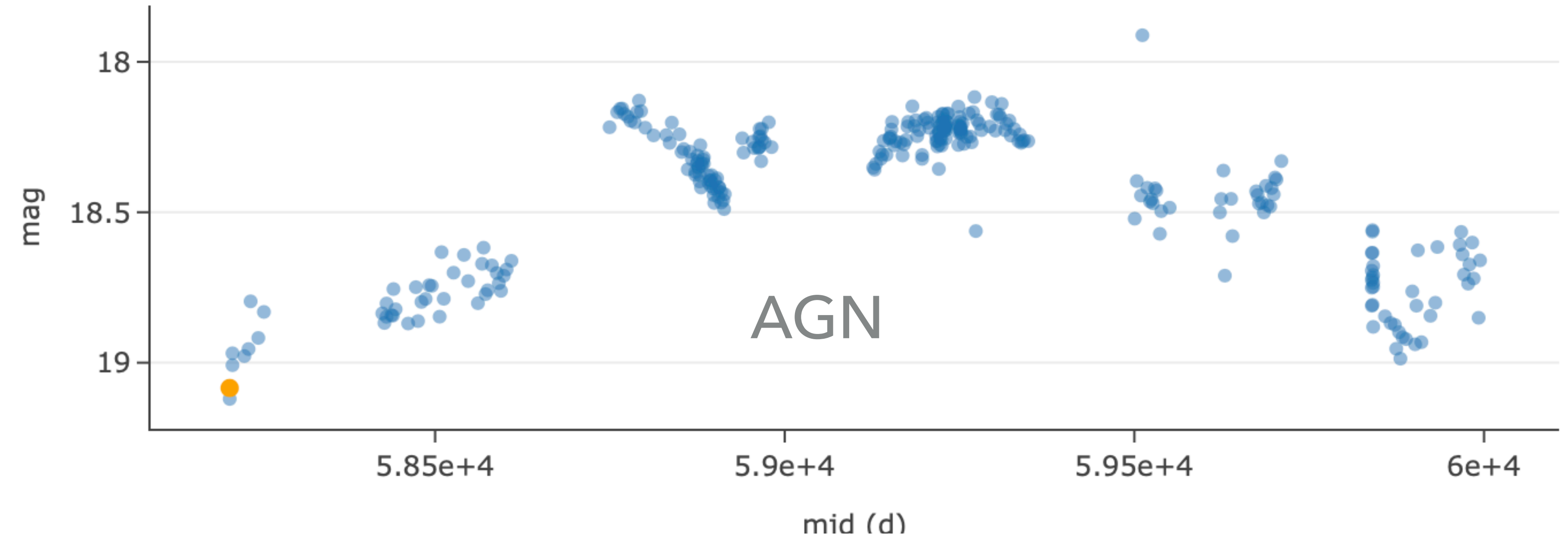
Field of view	47 sq. degrees
Detectors	16 e2v 6k x 6k CCD231-C6
Pixel size	15 microns
Pixel scale	1.0"/pixel
Median Delivered Image Quality	2.0" FWHM
Exposure time	30 sec
Readout time	10 sec
Median Time Between Exposures	15 sec
Median Single Visit Depth (5 sigma, R band)	20.4 mag (all lunar phases)
Filters	g, r, i
Areal survey rate	3750 square degrees/hour

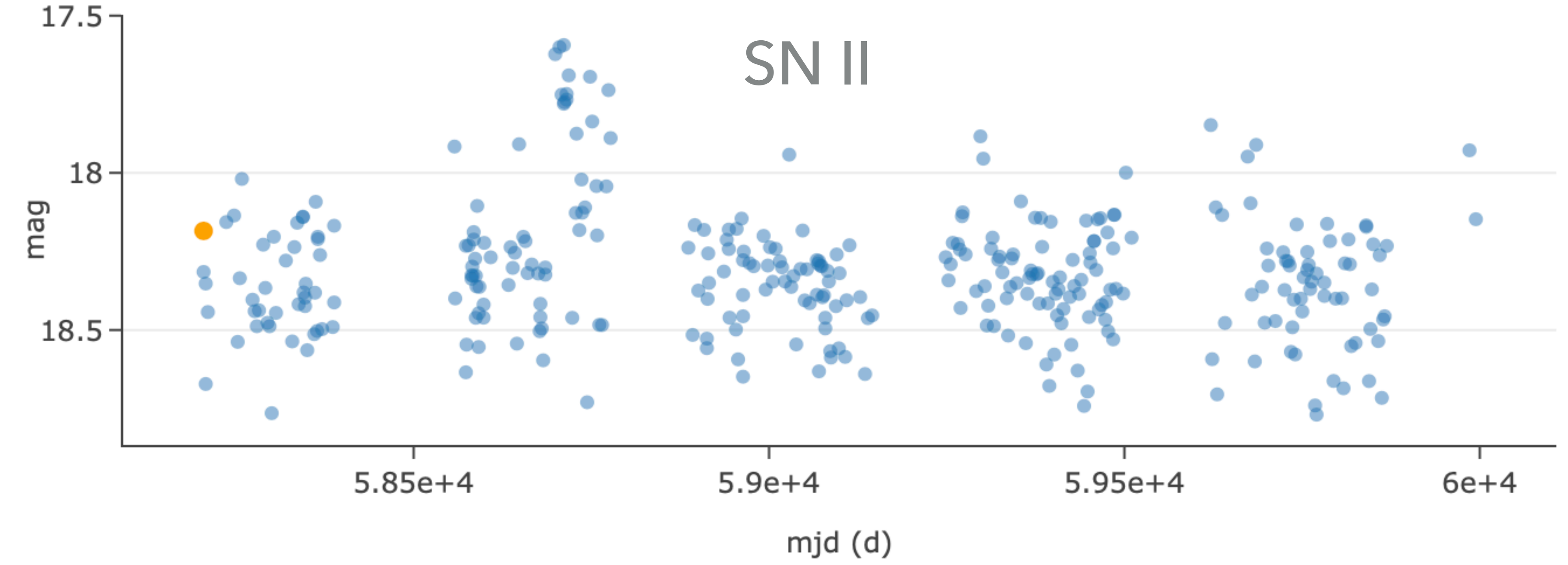
The Zwicky Transient Facility (ZTF) data releases (DR)

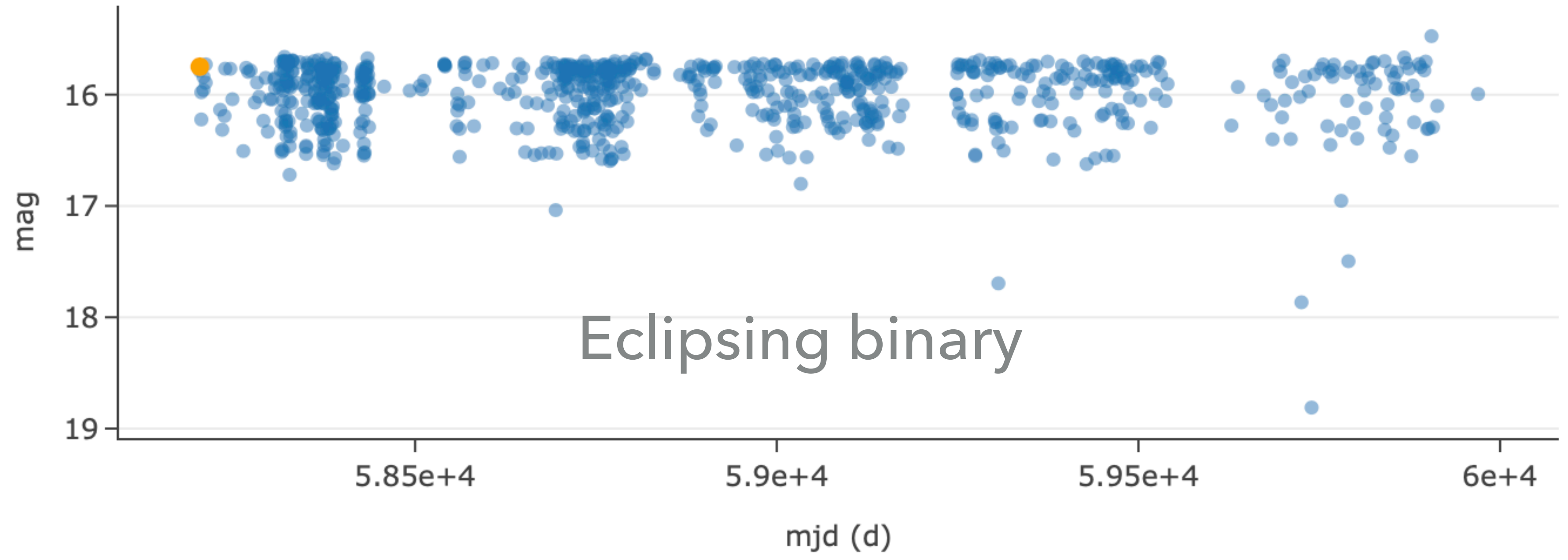
Data releases (DRs): PSF photometry over the all the ZTF images.


Non-variable galaxy


Non-variable star


AGN

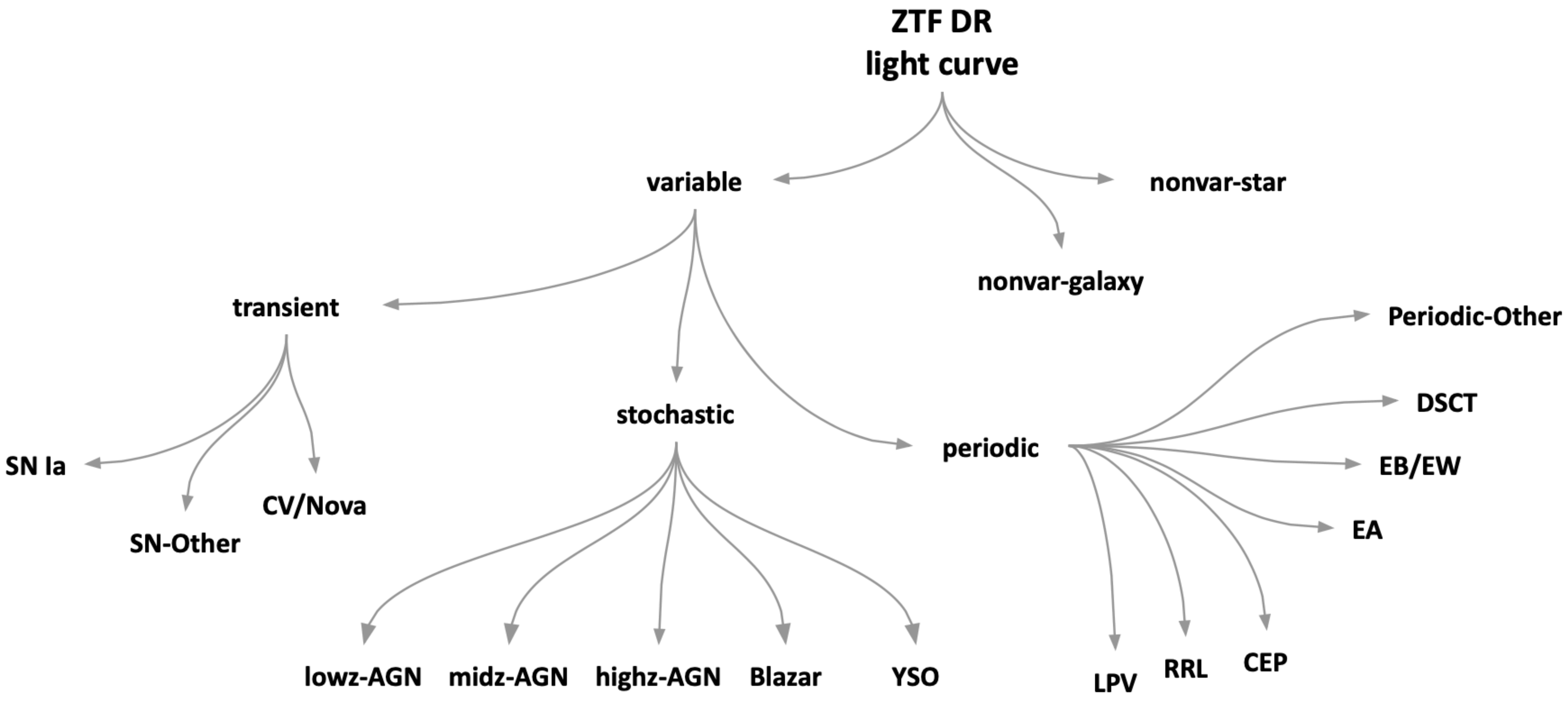

SN II


Eclipsing binary

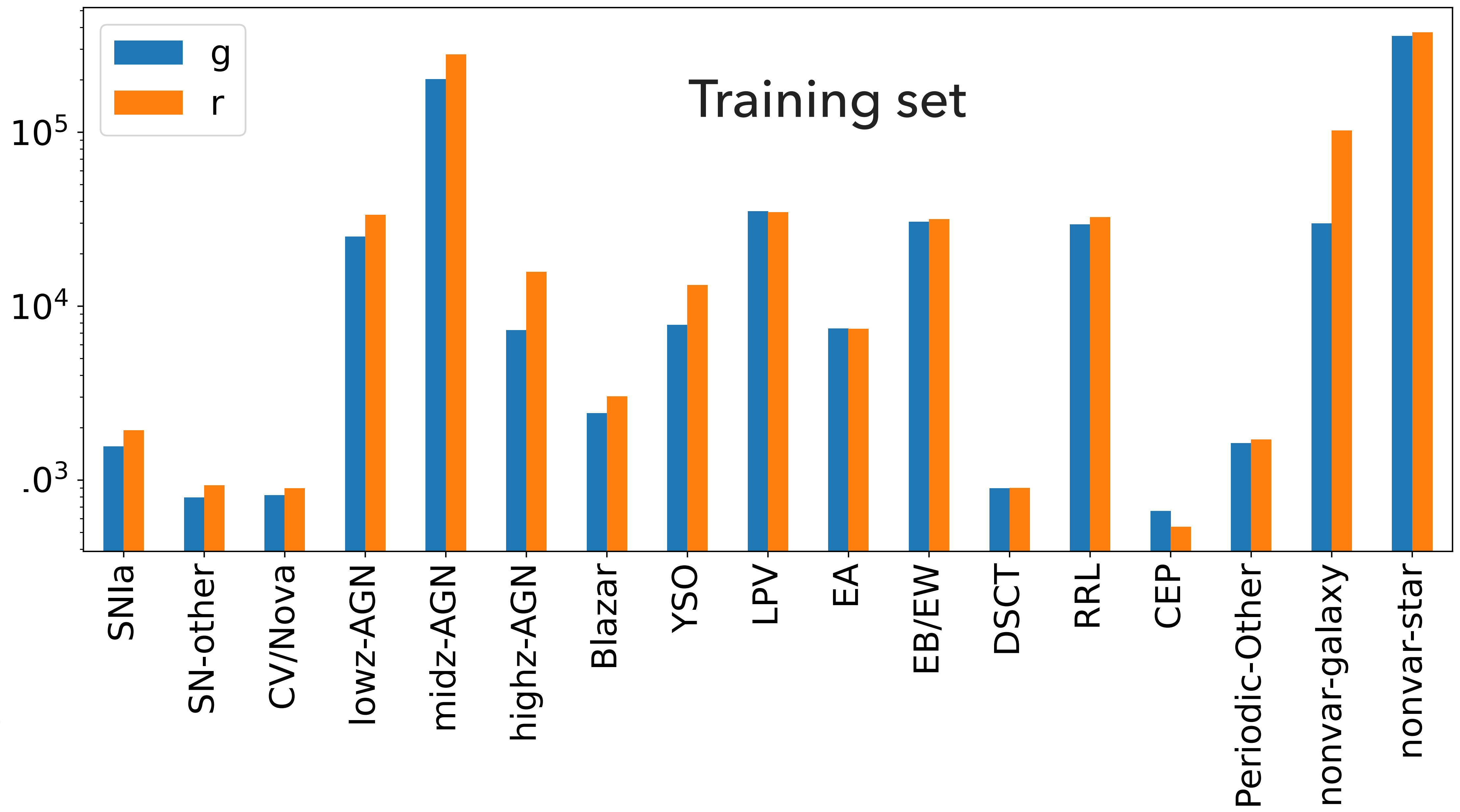
The ZTF DR light curve classifier

Sánchez-Sáez et al. 2023, A&A, in press.

- Using a Balanced Random Forest
- Hierarchical strategy
- Single ZTF band model

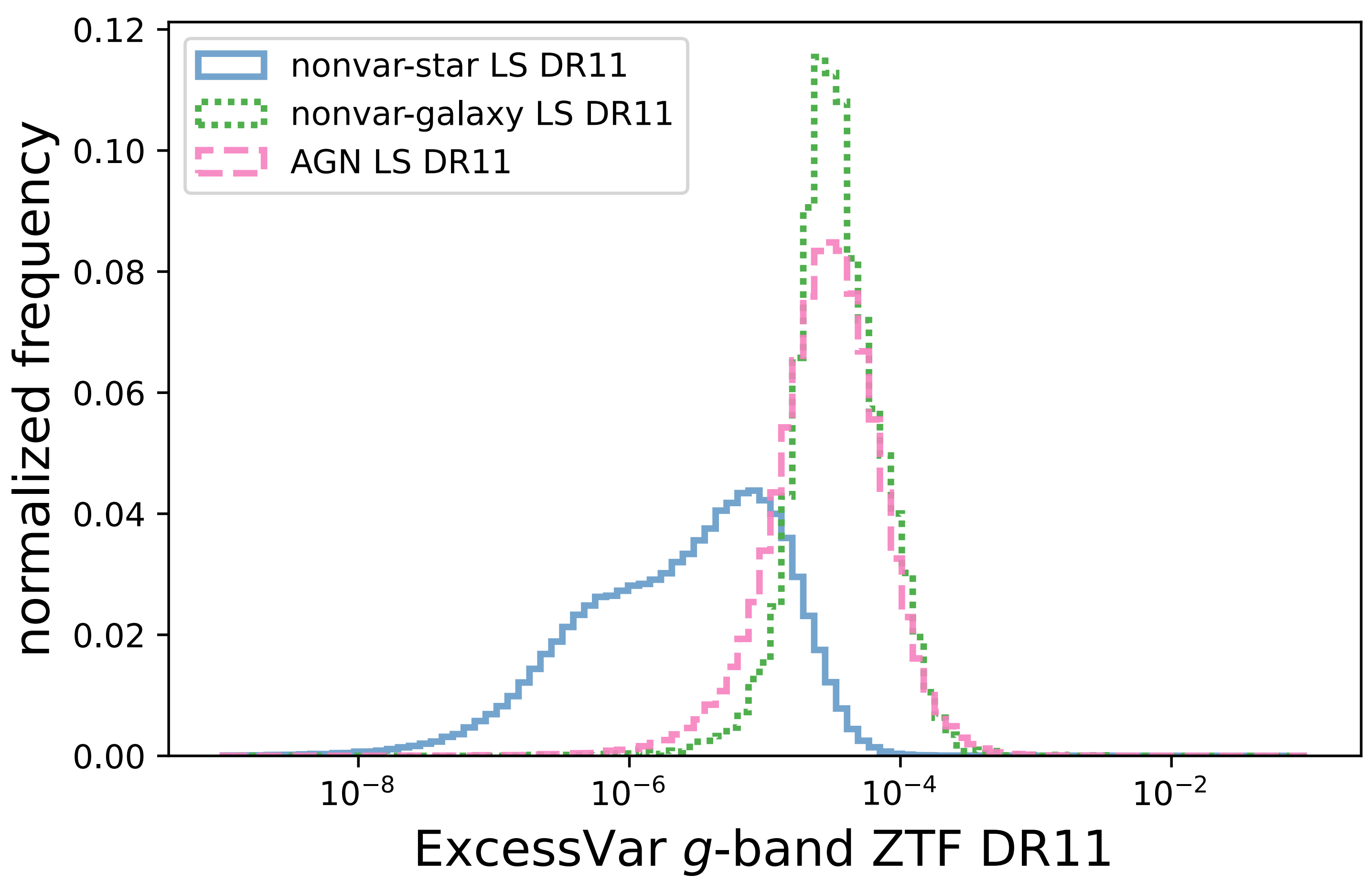


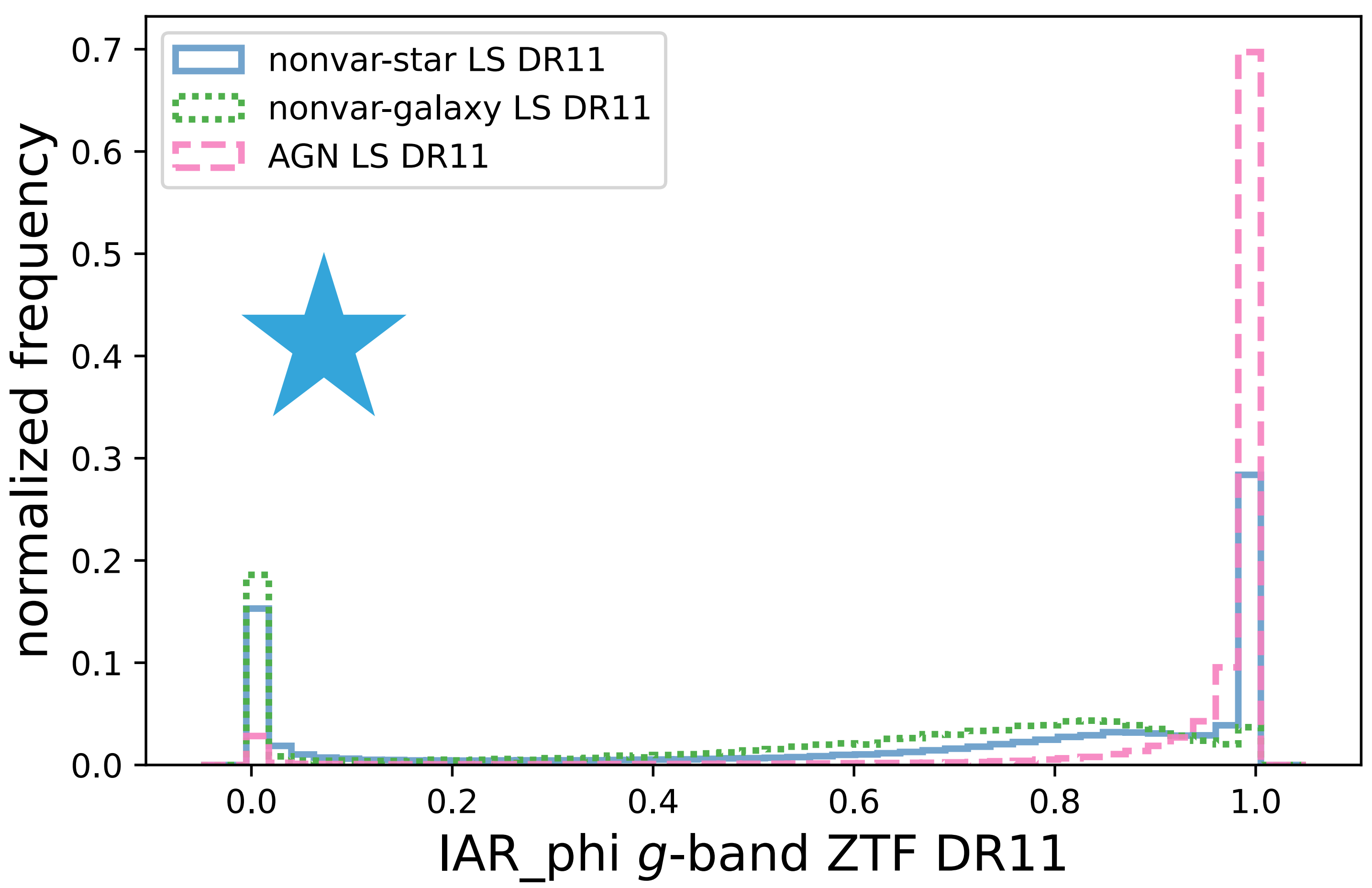
z<0.5 0.5<z<3 z>3

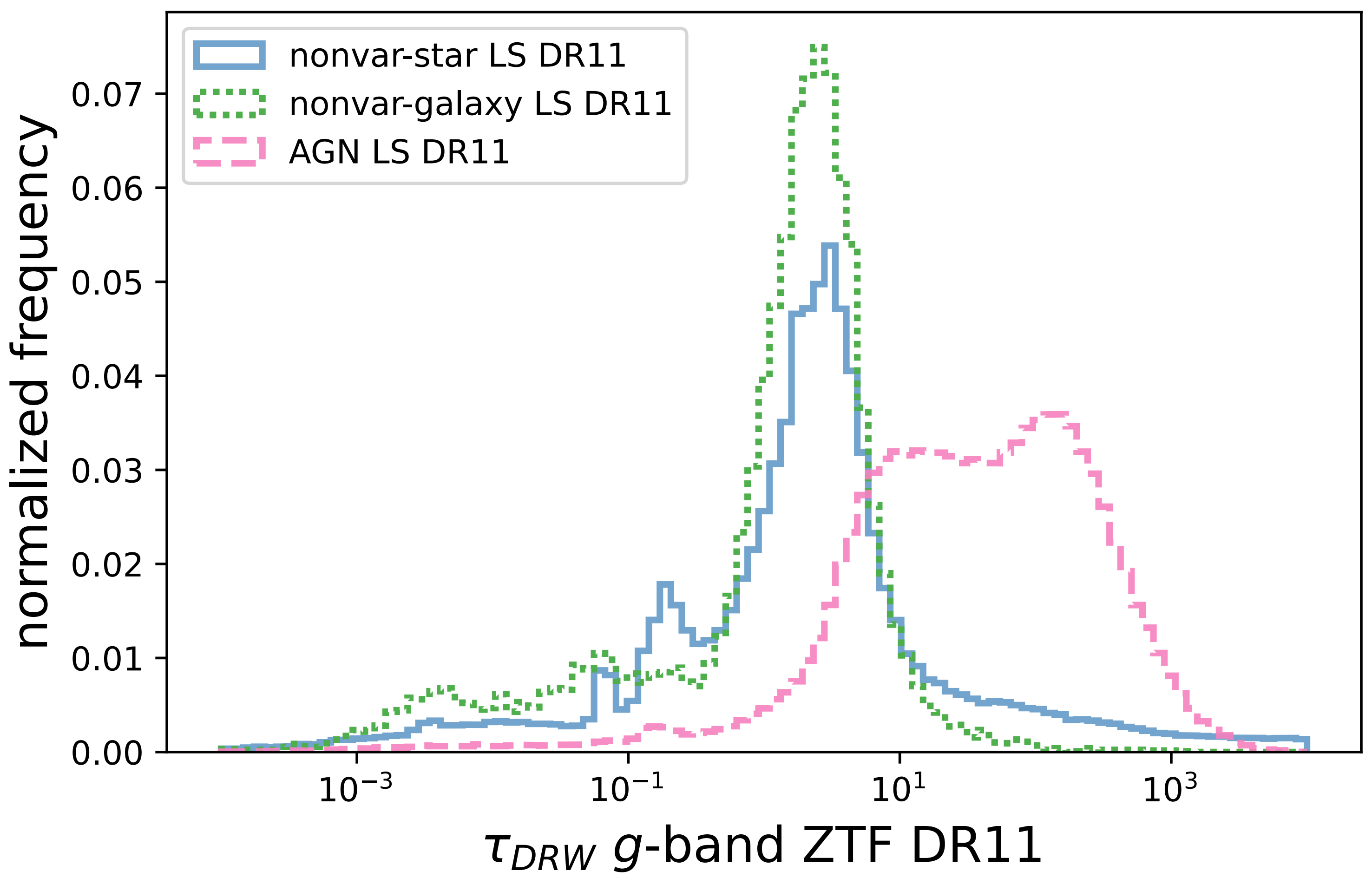


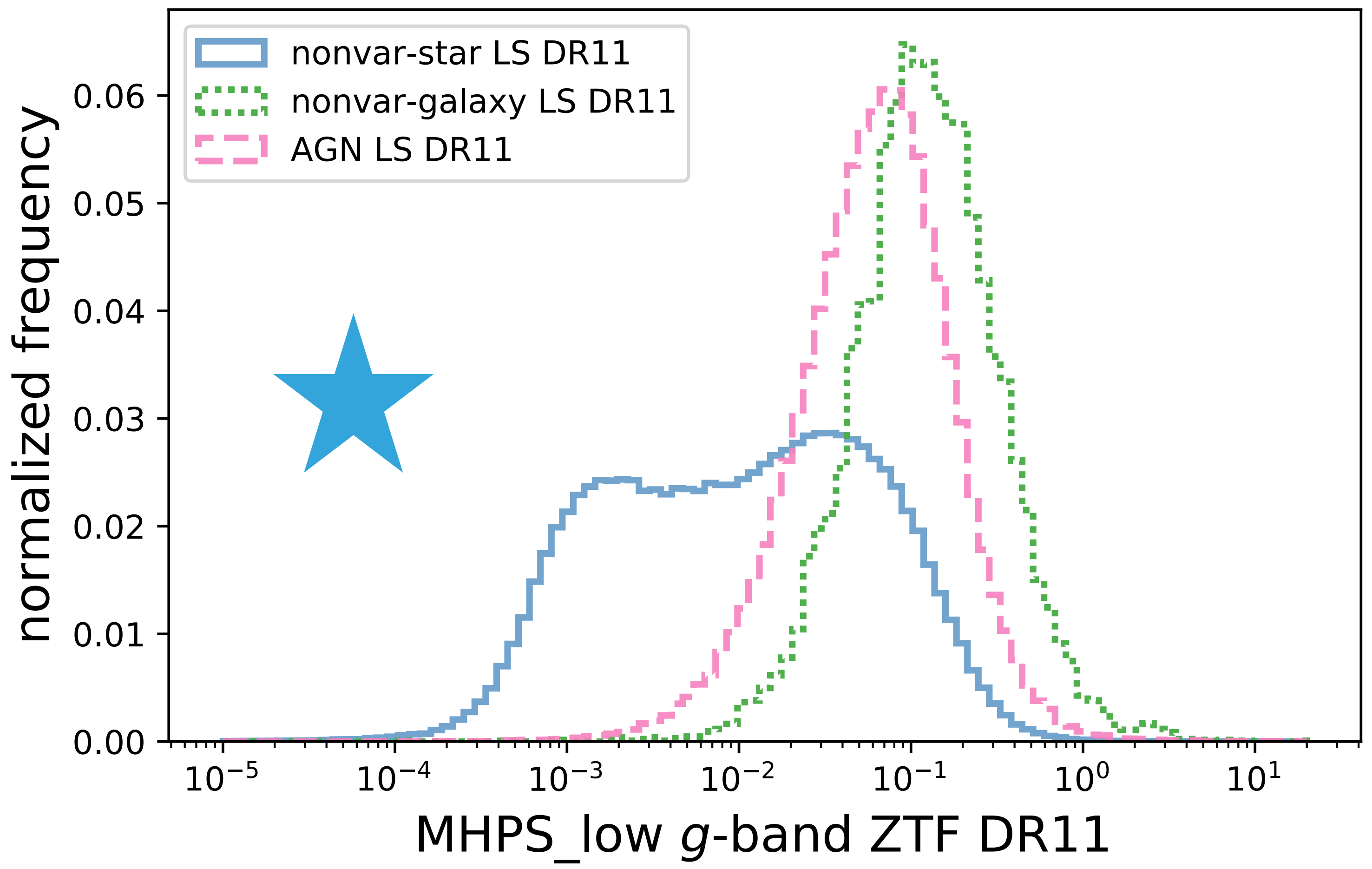
The ZTF DR light curve classifier: features

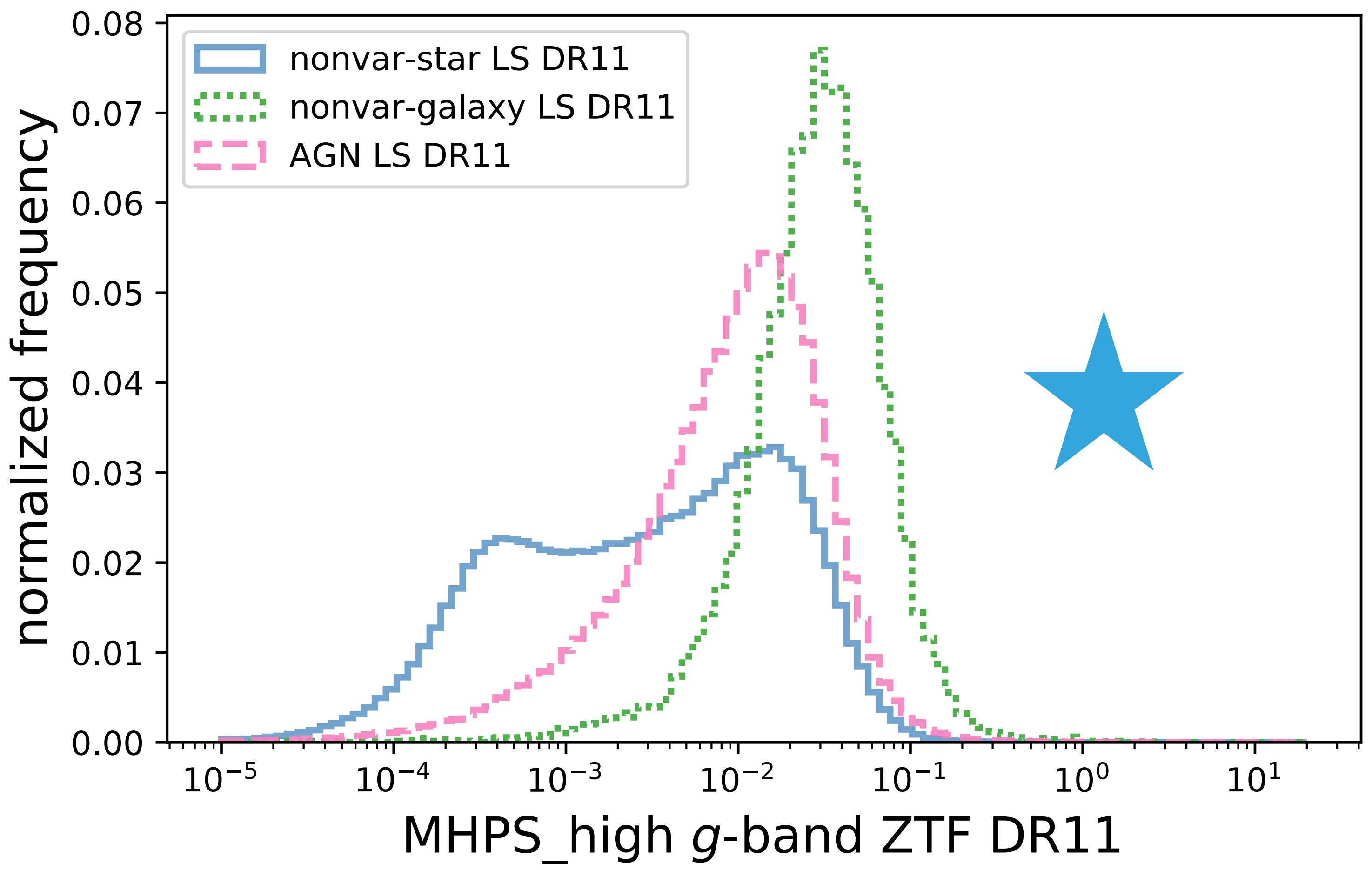
Sánchez-Sáez et al. 2023, A&A, in press.

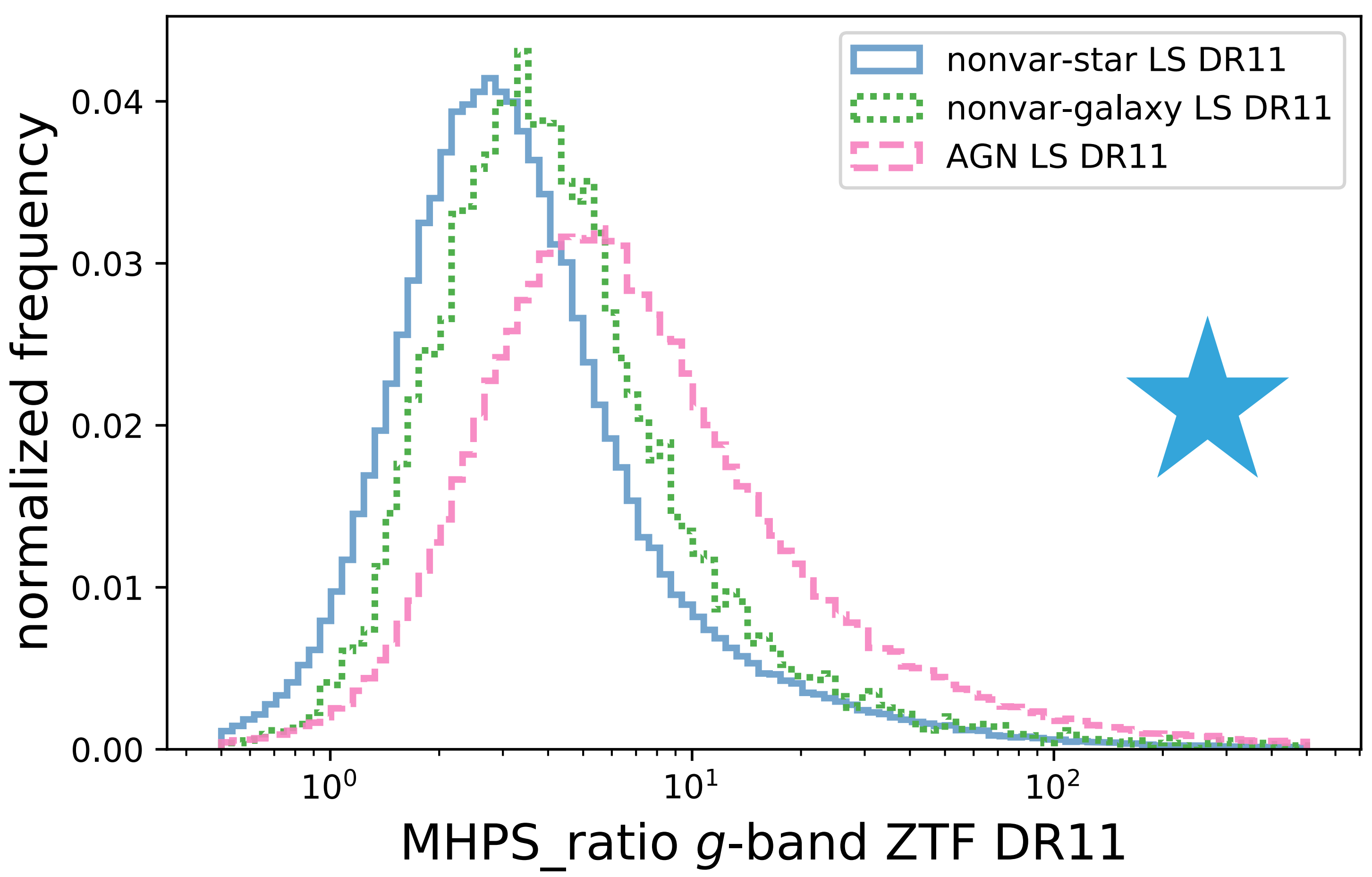








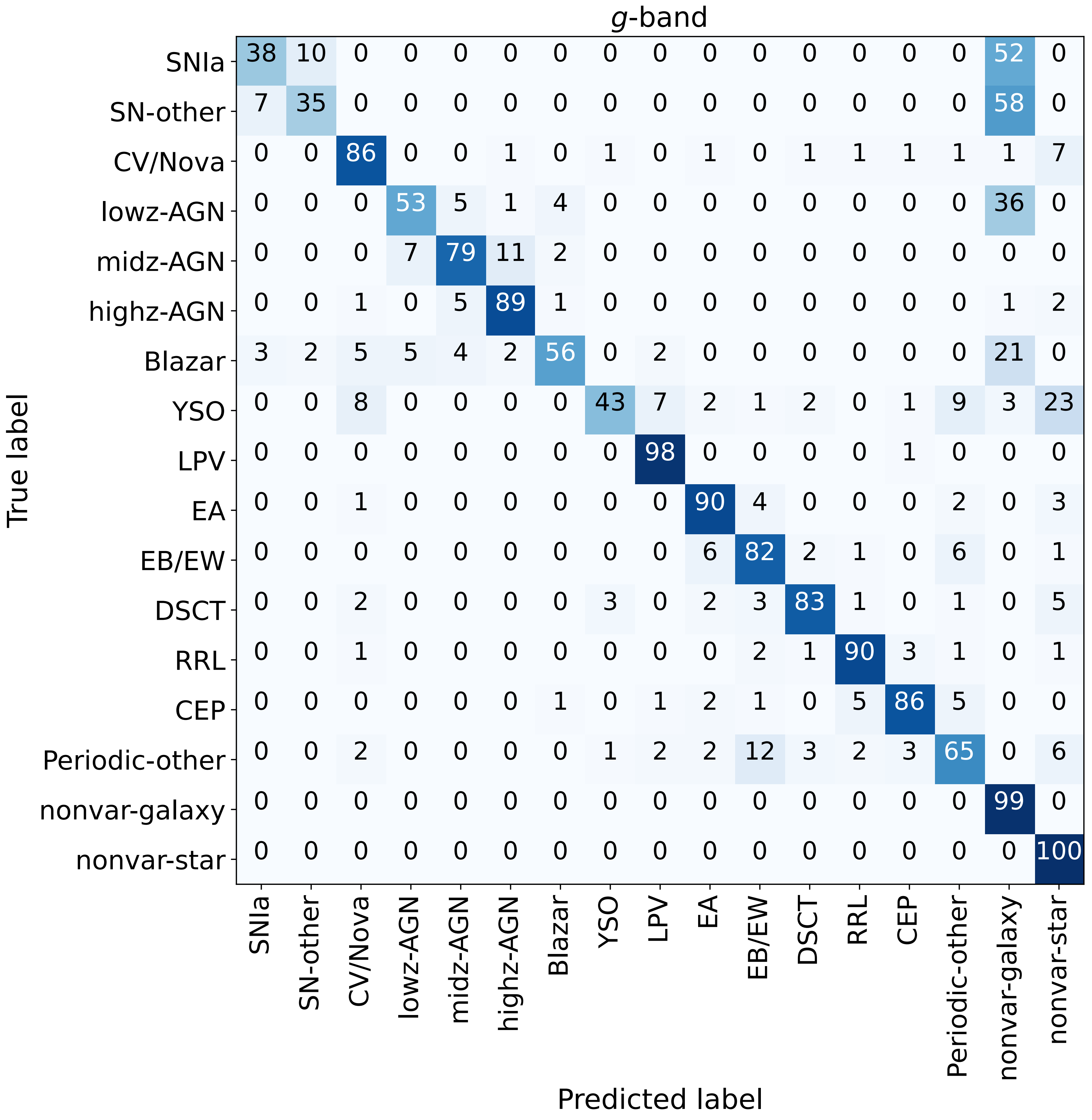




★ New variability features used by ALeRCE

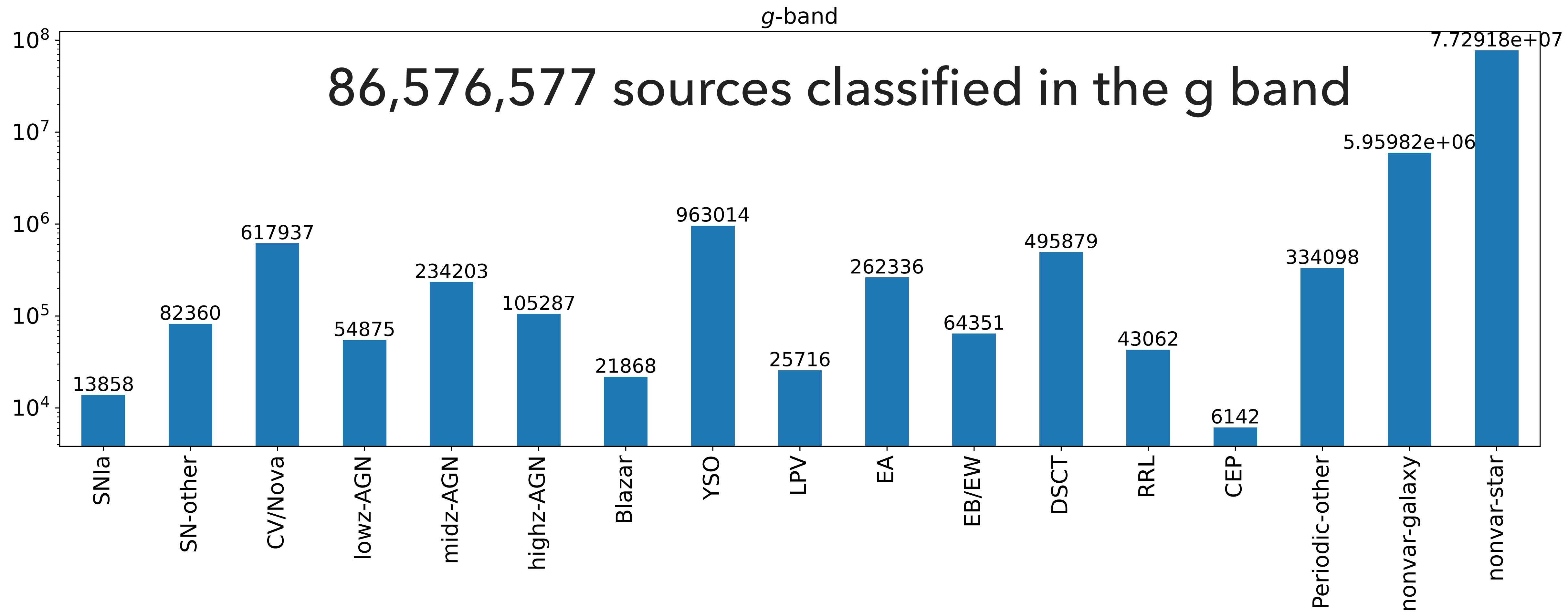
The ZTF DR light curve classifier: results

Sánchez-Sáez et al. 2023, A&A, in press.

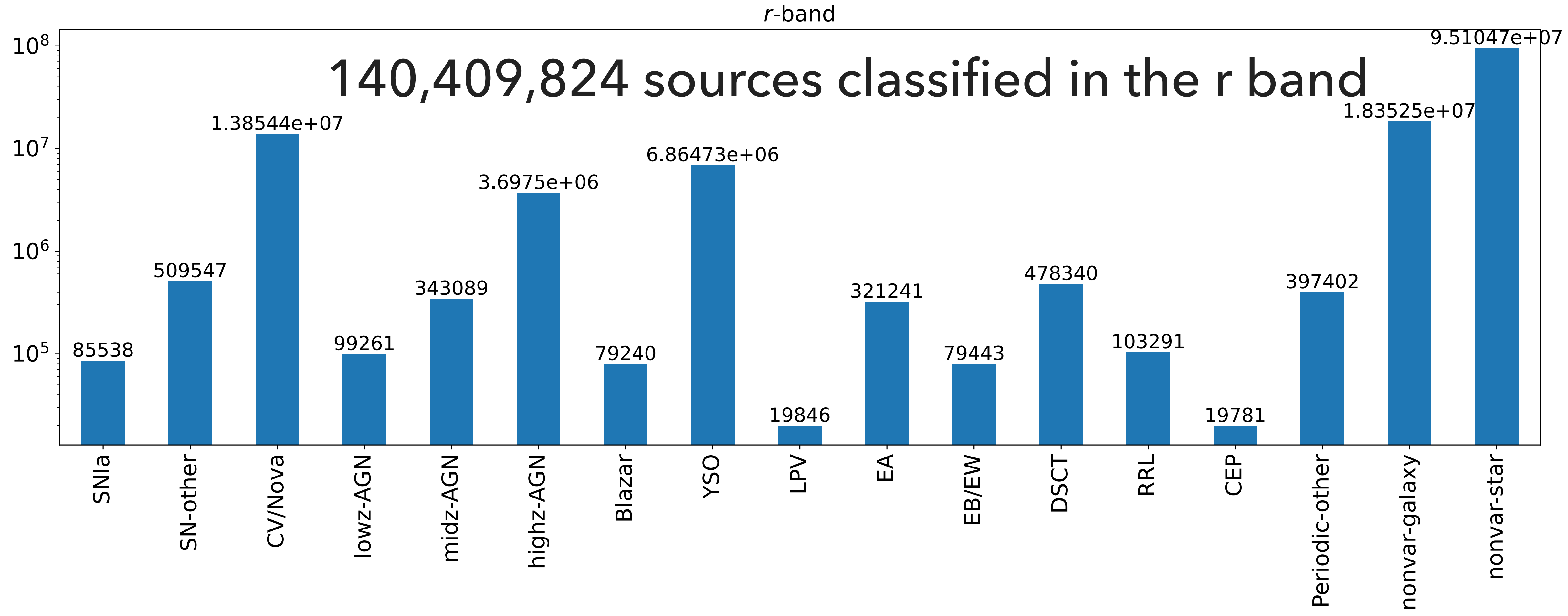


True label \ Predicted label	SNIa	SN-other	CV/Nova	lowz-AGN	midz-AGN	highz-AGN	Blazar	YSO	LPV	EA	EB/EW	DSCT	RRL	CEP	Periodic-other	nonvar-galaxy	nonvar-star
SNIa	38	10	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0
SN-other	7	35	0	0	0	0	0	0	0	0	0	0	0	0	0	58	0
CV/Nova	0	0	86	0	0	1	0	1	0	1	0	1	1	1	1	1	7
lowz-AGN	0	0	0	53	5	1	4	0	0	0	0	0	0	0	0	36	0
midz-AGN	0	0	0	7	79	11	2	0	0	0	0	0	0	0	0	0	0
highz-AGN	0	0	1	0	5	89	1	0	0	0	0	0	0	0	0	1	2
Blazar	3	2	5	5	4	2	56	0	2	0	0	0	0	0	0	21	0
YSO	0	0	8	0	0	0	0	43	7	2	1	2	0	1	9	3	23
LPV	0	0	0	0	0	0	0	0	98	0	0	0	0	1	0	0	0
EA	0	0	1	0	0	0	0	0	0	90	4	0	0	0	2	0	3
EB/EW	0	0	0	0	0	0	0	0	0	6	82	2	1	0	6	0	1
DSCT	0	0	2	0	0	0	0	3	0	2	3	83	1	0	1	0	5
RRL	0	0	1	0	0	0	0	0	0	0	2	1	90	3	1	0	1
CEP	0	0	0	0	0	0	1	0	1	2	1	0	5	86	5	0	0
Periodic-other	0	0	2	0	0	0	0	1	2	2	12	3	2	3	65	0	6
nonvar-galaxy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0
nonvar-star	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Similar matrix for the r band.



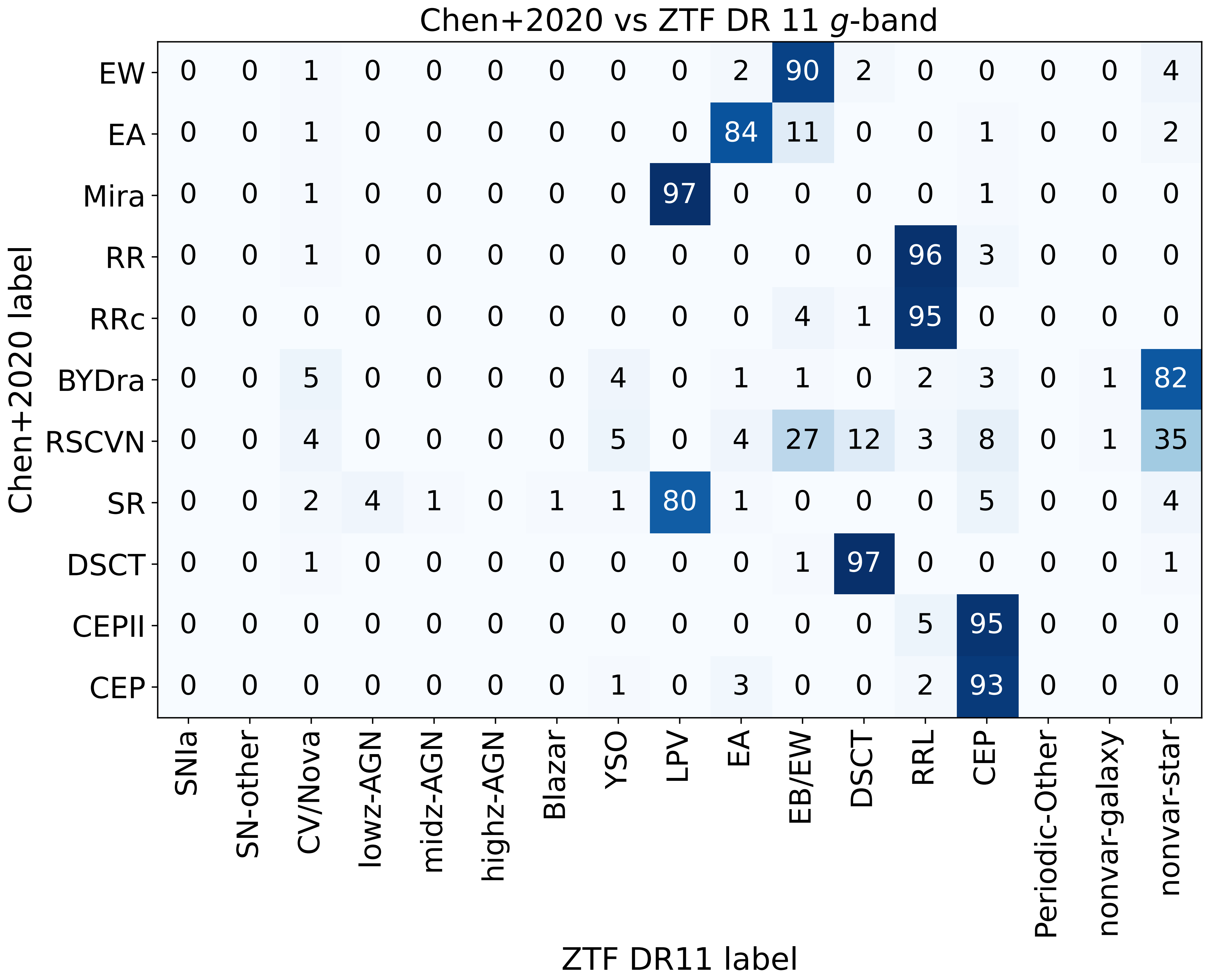
86,576,577 sources classified in the g band



140,409,824 sources classified in the r band

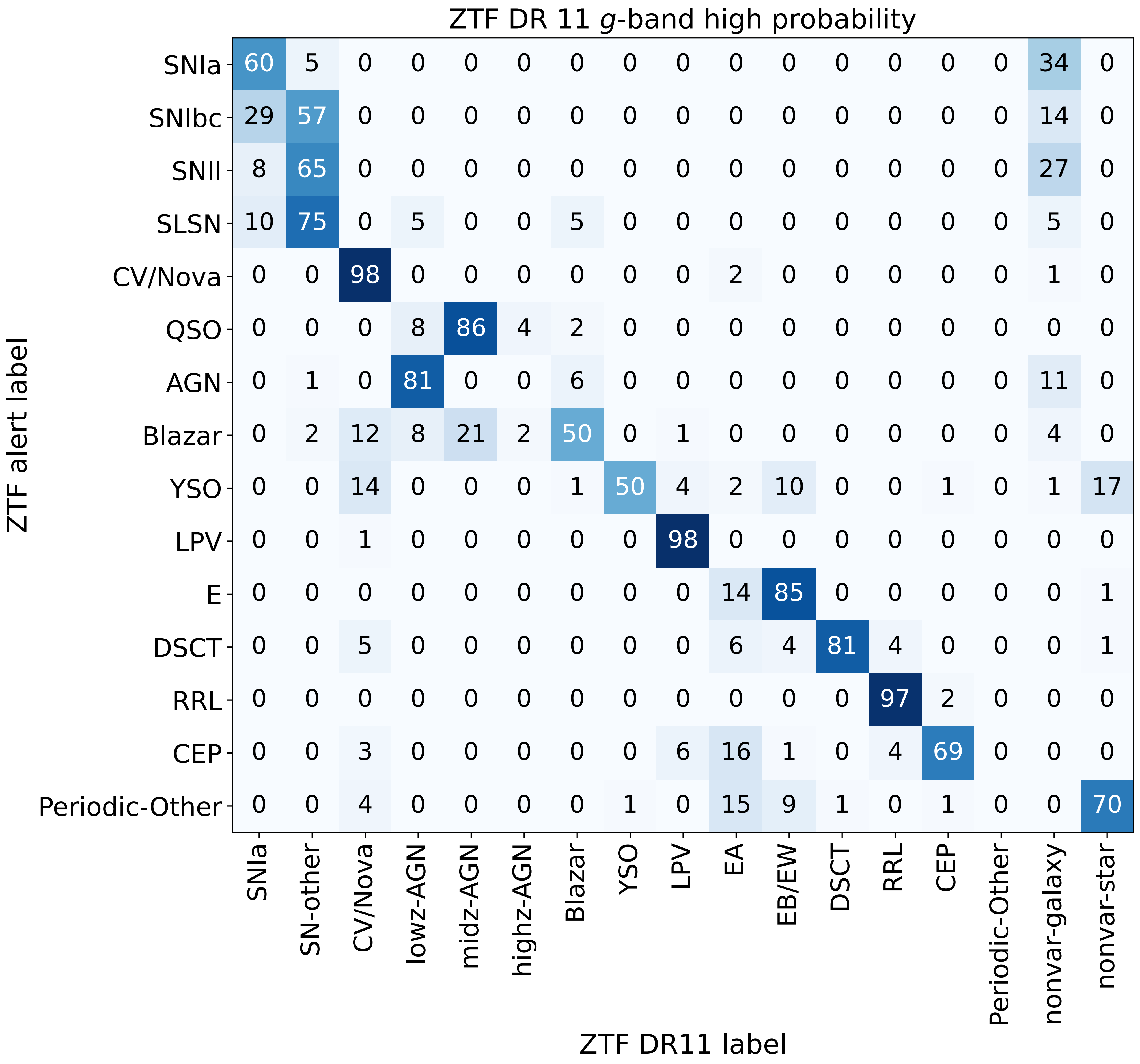
The ZTF DR light curve classifier: results

Comparison with Chen+2020



Chen+2020 label	SNIa	SN-other	CV/Nova	lowz-AGN	midz-AGN	highz-AGN	Blazar	YSO	LPV	EA	EB/EW	DSCT	RRL	CEP	Periodic-Other	nonvar-galaxy	nonvar-star
EW	0	0	1	0	0	0	0	0	0	2	90	2	0	0	0	0	4
EA	0	0	1	0	0	0	0	0	0	84	11	0	0	1	0	0	2
Mira	0	0	1	0	0	0	0	0	97	0	0	0	0	1	0	0	0
RR	0	0	1	0	0	0	0	0	0	0	0	0	96	3	0	0	0
RRc	0	0	0	0	0	0	0	0	0	0	4	1	95	0	0	0	0
BYDra	0	0	5	0	0	0	0	4	0	1	1	0	2	3	0	1	82
RSCVN	0	0	4	0	0	0	0	5	0	4	27	12	3	8	0	1	35
SR	0	0	2	4	1	0	1	1	80	1	0	0	0	5	0	0	4
DSCT	0	0	1	0	0	0	0	0	0	0	1	97	0	0	0	0	1
CEPII	0	0	0	0	0	0	0	0	0	0	0	0	5	95	0	0	0
CEP	0	0	0	0	0	0	0	1	0	3	0	0	2	93	0	0	0

Comparison with the ALeRCE alerts classifier



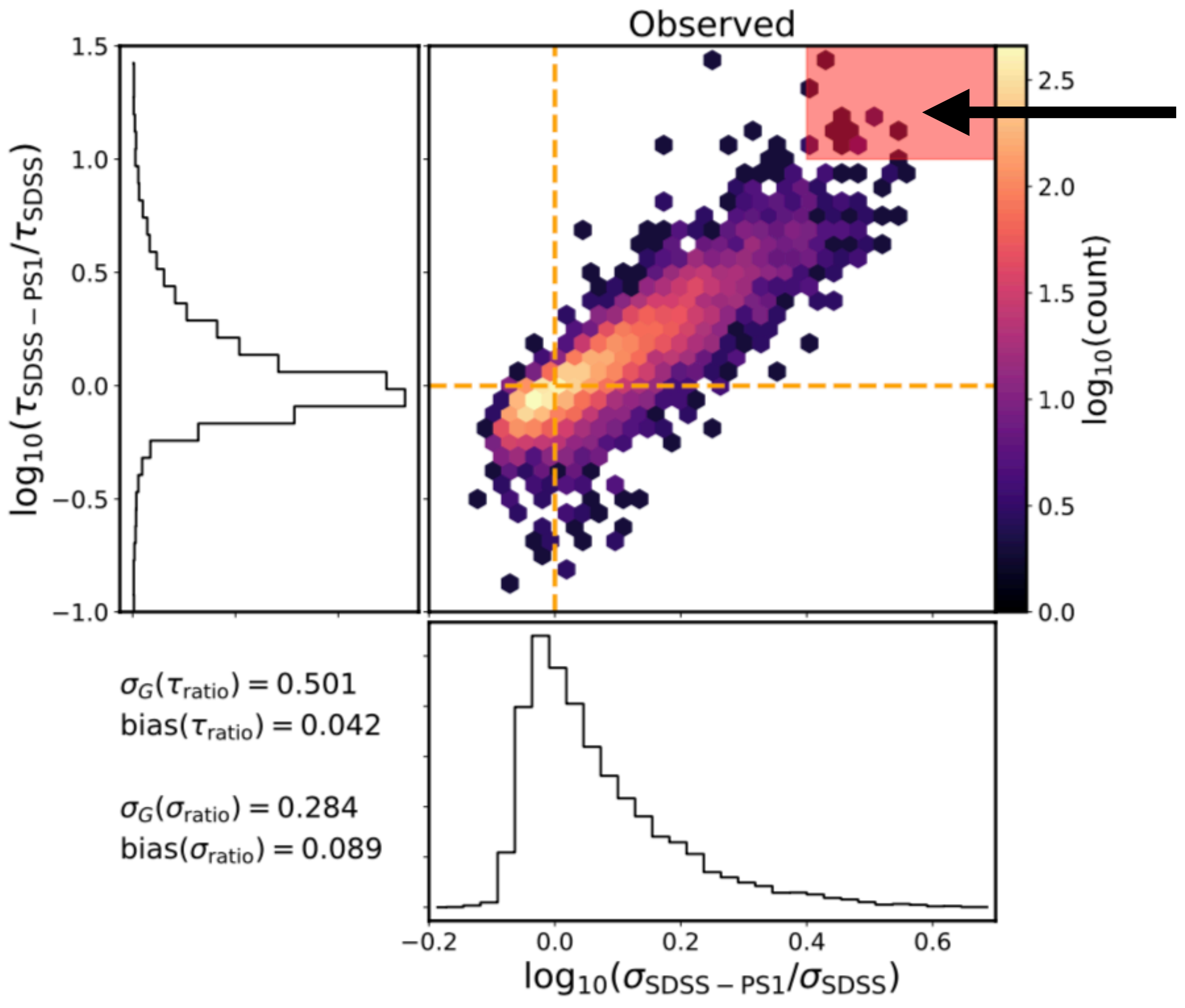
ZTF alert label	SNIa	SN-other	CV/Nova	lowz-AGN	midz-AGN	highz-AGN	Blazar	YSO	LPV	EA	EB/EW	DSCT	RRL	CEP	Periodic-Other	nonvar-galaxy	nonvar-star
SNIa	60	5	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0
SNIbc	29	57	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0
SNII	8	65	0	0	0	0	0	0	0	0	0	0	0	0	0	27	0
SLSN	10	75	0	5	0	0	5	0	0	0	0	0	0	0	0	5	0
CV/Nova	0	0	98	0	0	0	0	0	0	2	0	0	0	0	0	1	0
QSO	0	0	0	8	86	4	2	0	0	0	0	0	0	0	0	0	0
AGN	0	1	0	81	0	0	6	0	0	0	0	0	0	0	0	11	0
Blazar	0	2	12	8	21	2	50	0	1	0	0	0	0	0	0	4	0
YSO	0	0	14	0	0	0	1	50	4	2	10	0	0	1	0	1	17
LPV	0	0	1	0	0	0	0	0	98	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	14	85	0	0	0	0	0	1
DSCT	0	0	5	0	0	0	0	0	0	6	4	81	4	0	0	0	1
RRL	0	0	0	0	0	0	0	0	0	0	0	0	97	2	0	0	0
CEP	0	0	3	0	0	0	0	0	6	16	1	0	4	69	0	0	0
Periodic-Other	0	0	4	0	0	0	0	1	0	15	9	1	0	1	0	0	70

2. SEARCHING FOR CSAGNS WITH ANOMALY DETECTION

Detecting CSAGN events in massive datasets

The goal of this work is to create a method to search for CSAGN candidates in massive data sets, using anomaly detection techniques.

Currently, we use data from the Zwicky Transient Facility data releases, and in the future we will apply this to Vera Rubin / LSST data.



CSAGN candidates

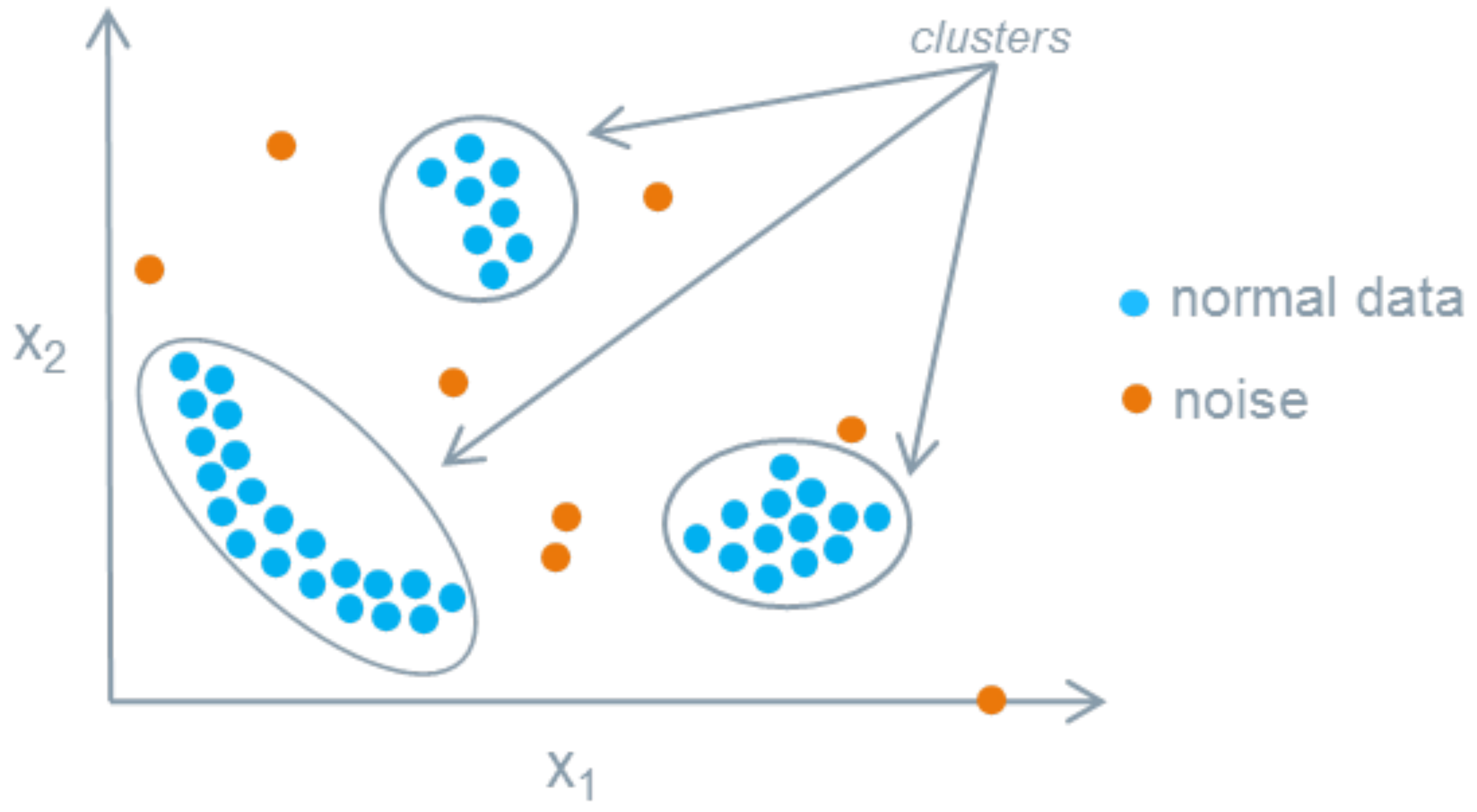
Suberlak et al. 2021

ZTF → VERA C. RUBIN OBSERVATORY

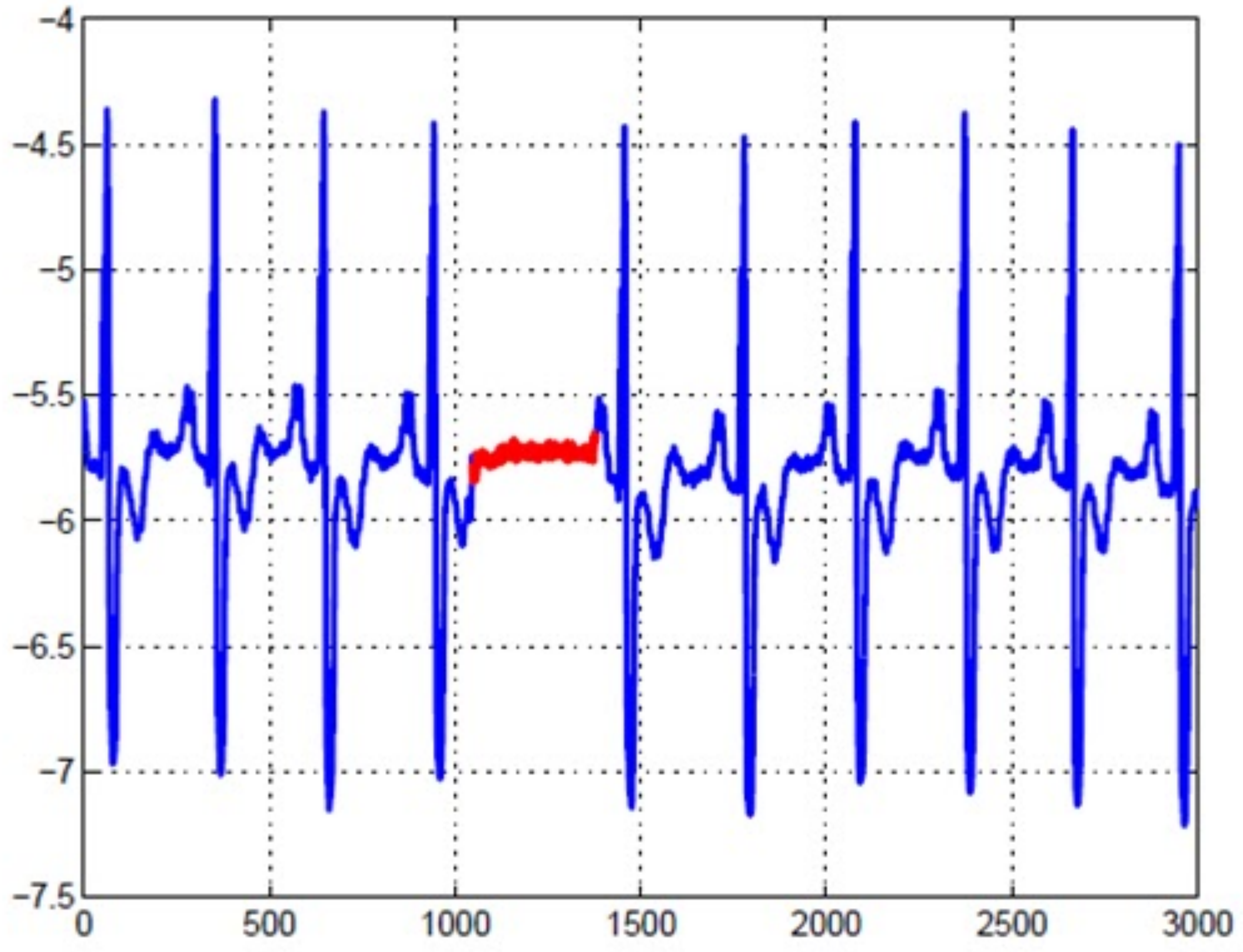
Anomaly detection (AD)

AD correspond to the identification of rare events or observations that differ significantly from the majority of the data.

Out of distribution anomaly: searching for unusual objects within datasets.

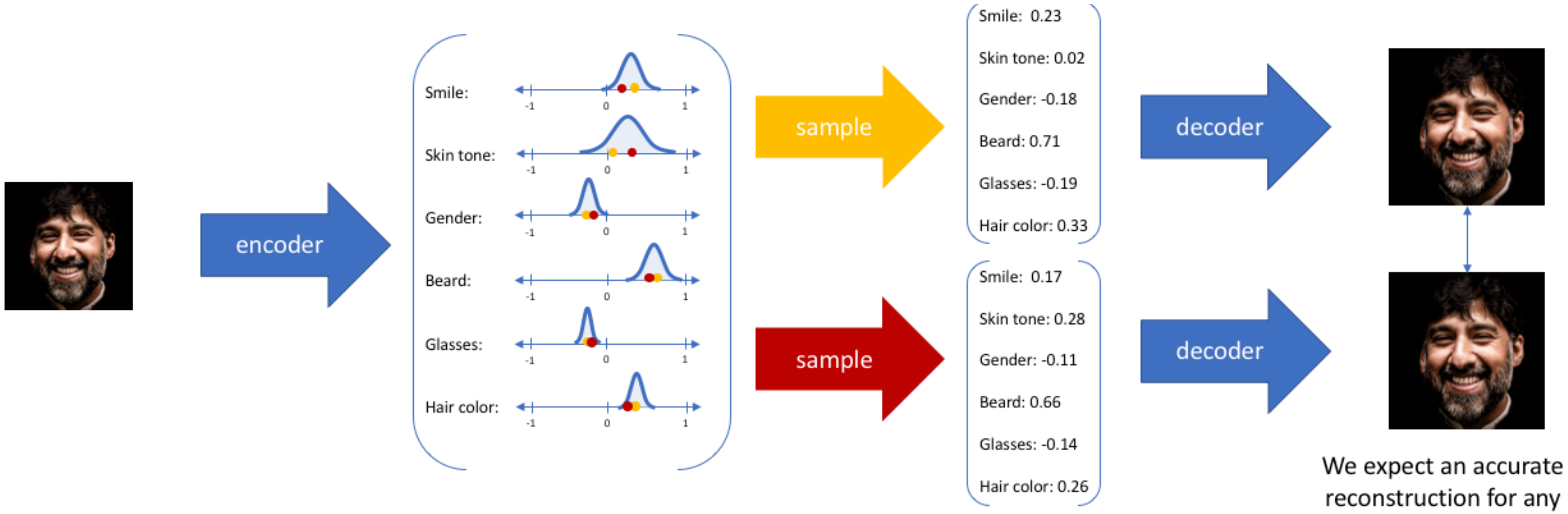


Contextual anomaly: searching for objects that suddenly start presenting unusual behaviors.



Variational Autoencoders (VAEs)

VAEs correspond to a modification of the more classical Autoencoder (AE) architectures. In this case, the latent representations are described by multivariate normal distributions, where each attribute or feature in the latent space is described by a latent mean (μ) and a latent variance ($σ^2$), which can be used to randomly sample a set of attributes.



Latent distributions

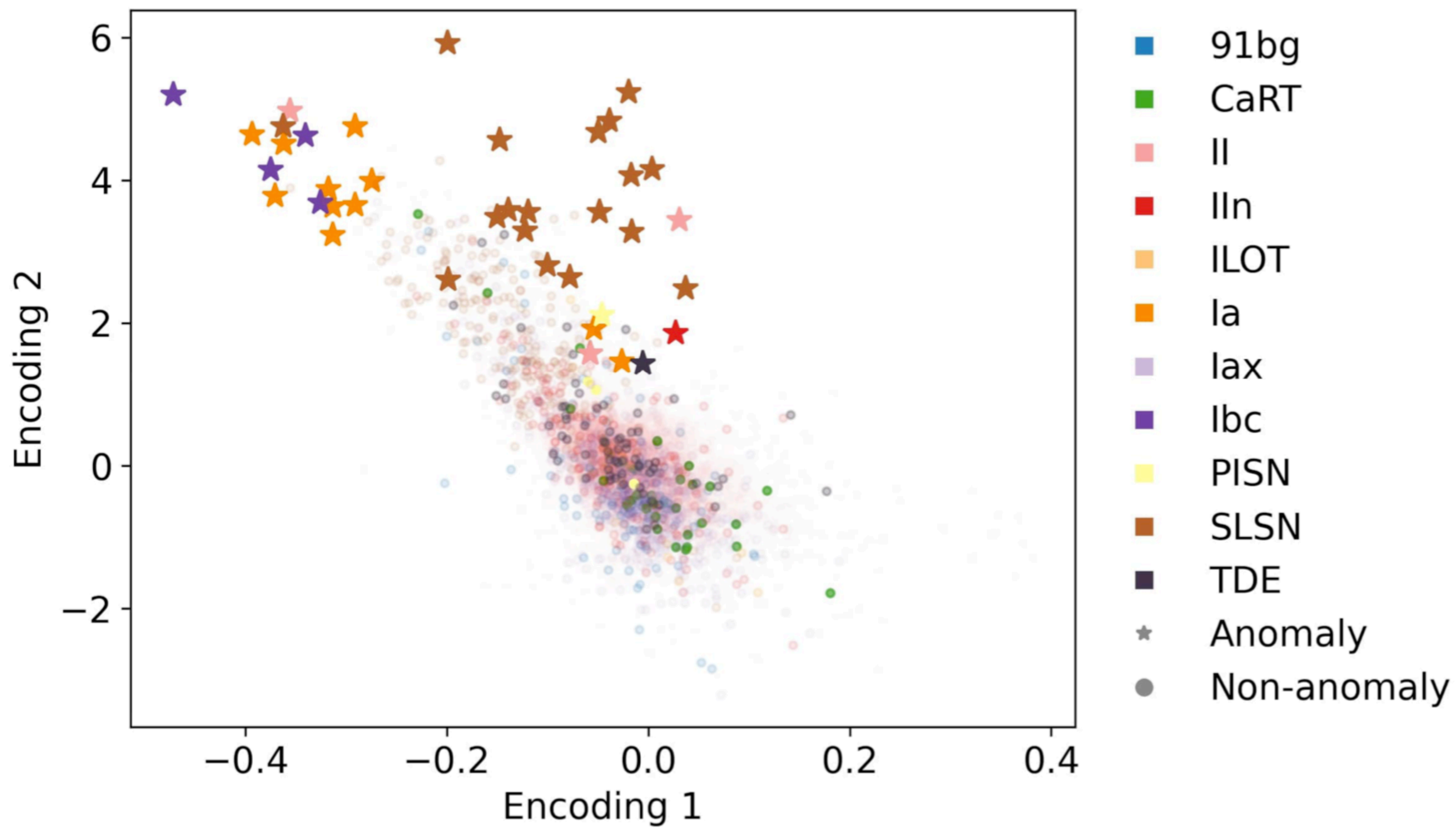
Sampled latent attributes

We expect an accurate reconstruction for any sample from the latent state distributions

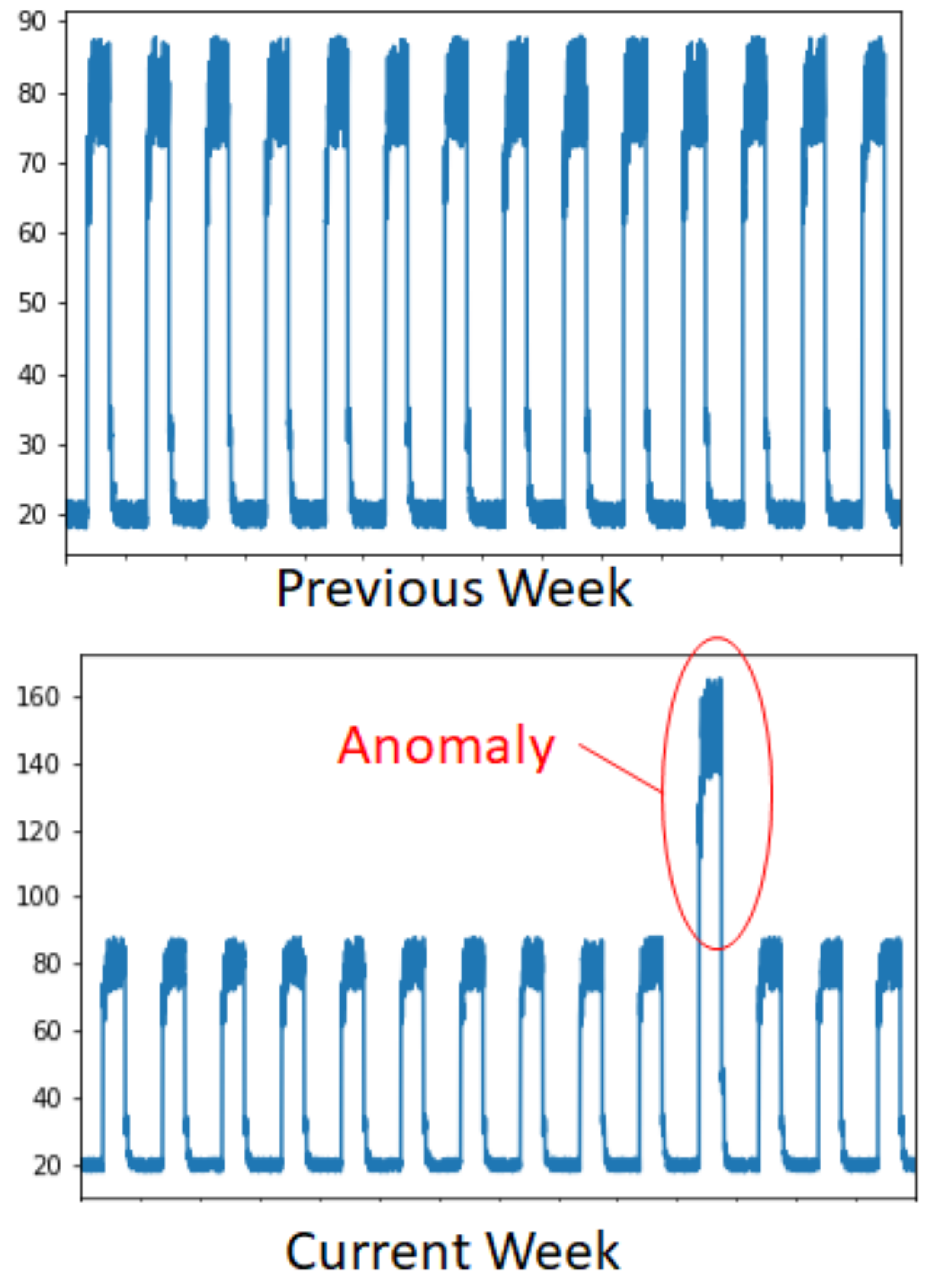
Credits: https://www.jeremyjordan.me/variational-autoencoders/

Variational Recurrent Autoencoders (VRAEs) for time series anomaly detection

Out of distribution AD: using the latent space to define outliers that are in atypical locations of latent space (e.g., Villar+2021)



Contextual AD: using the reconstruction error of the VRAE as an anomaly score

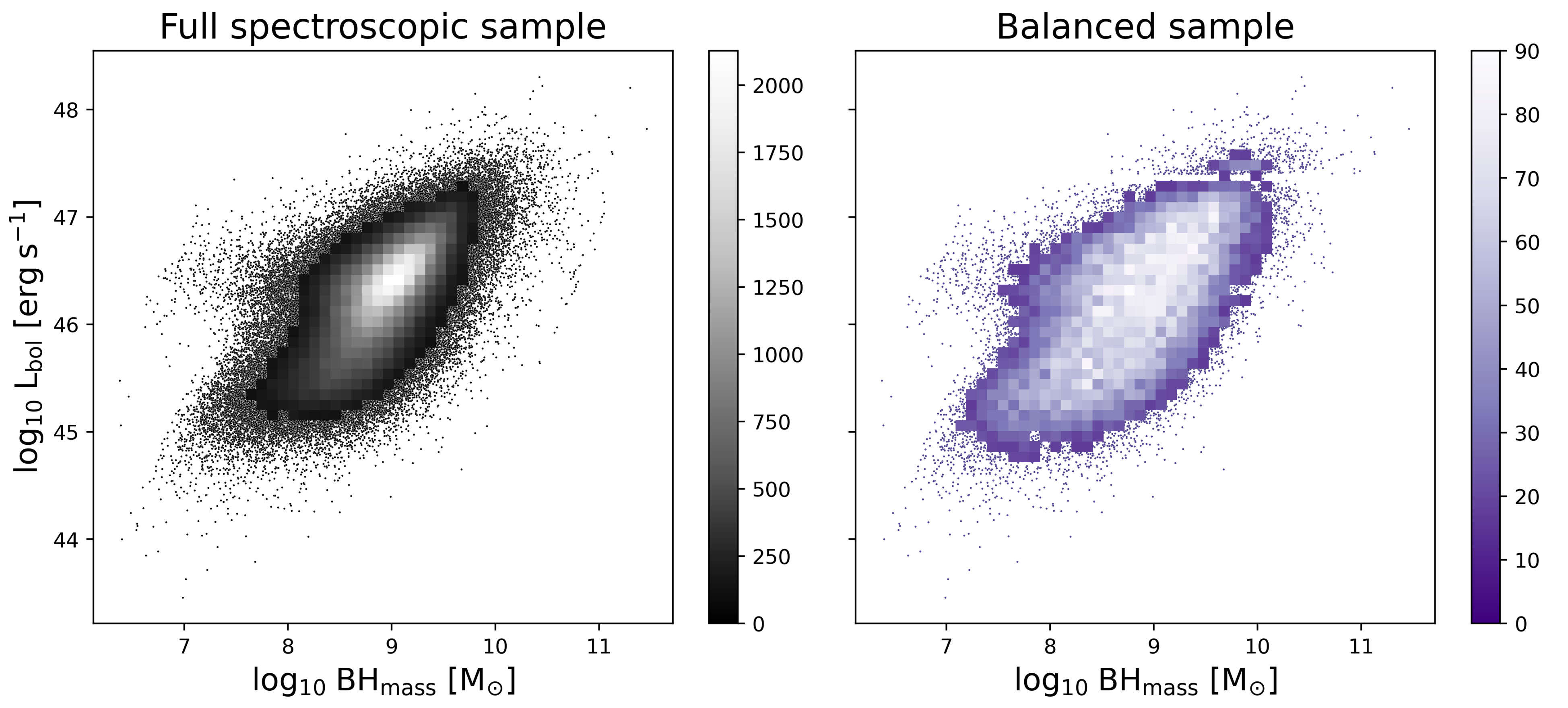


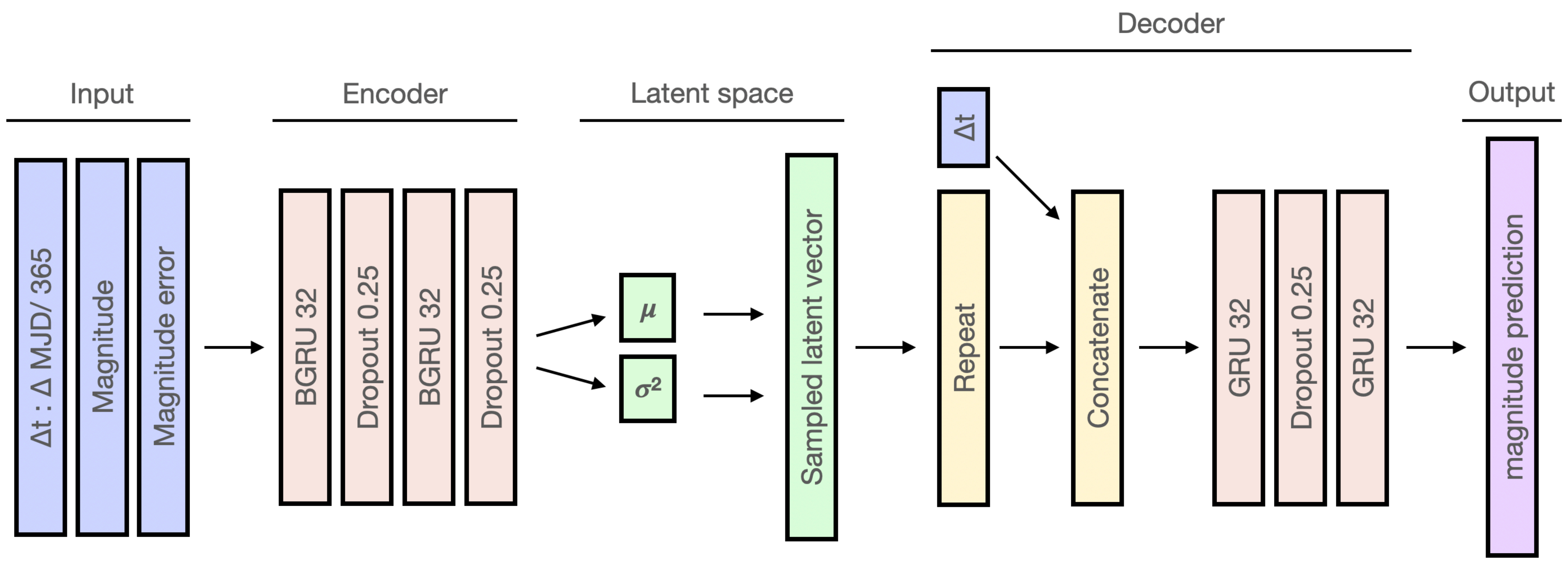
VRAEs to model AGN variability

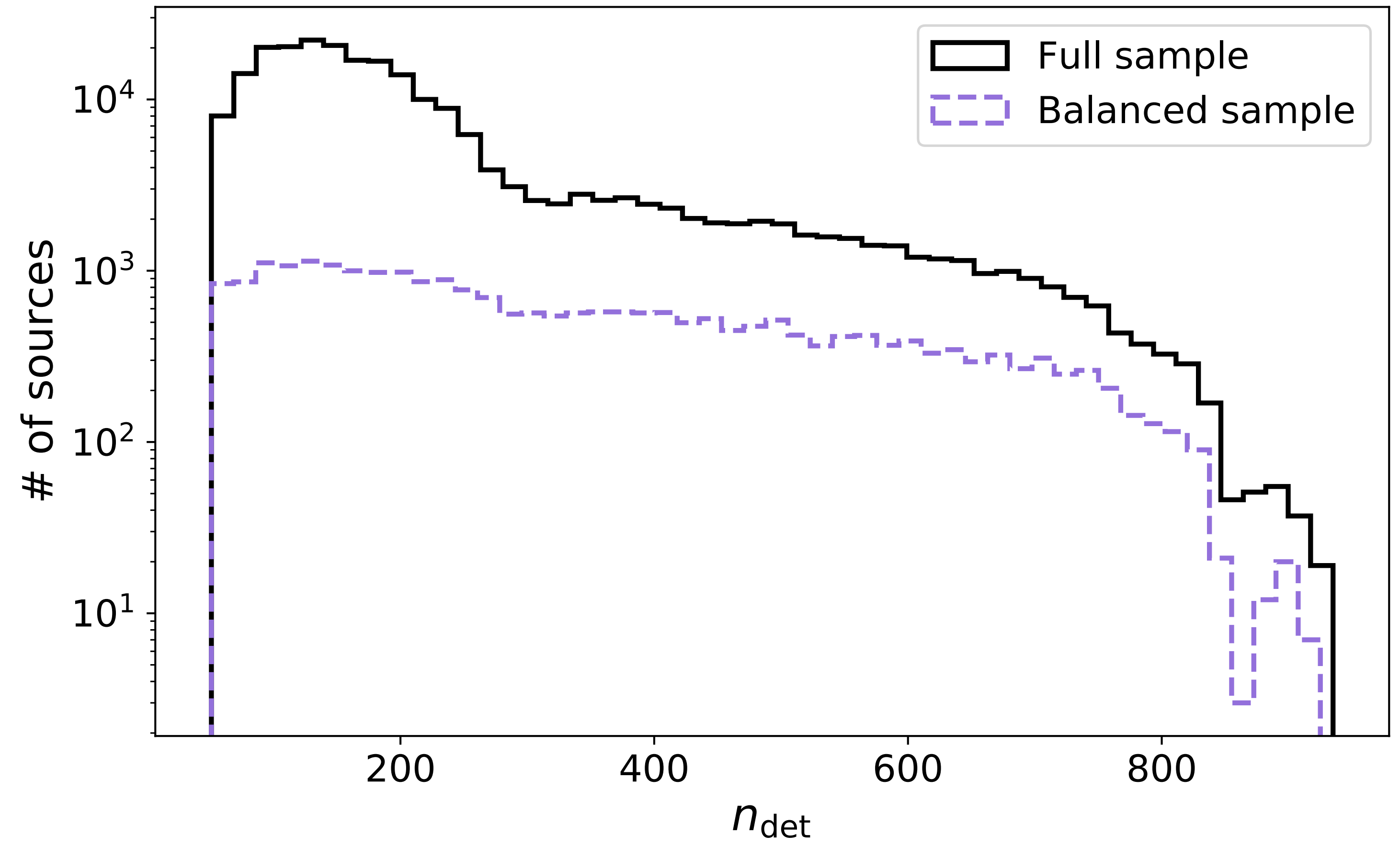
Sánchez-Sáez et al. 2021, AJ, 162, 206

230,451 AGN light curves from ZTF DR5 (including different classes from the MILLIQUAS and ROMABZCAT catalogs)

- VRAE architecture (inspired by Tachibana+2020's model)
- Trained with a dataset balanced by means of their physical properties and number of epochs per light curve.







VRAEs for AGN variability anomaly detection

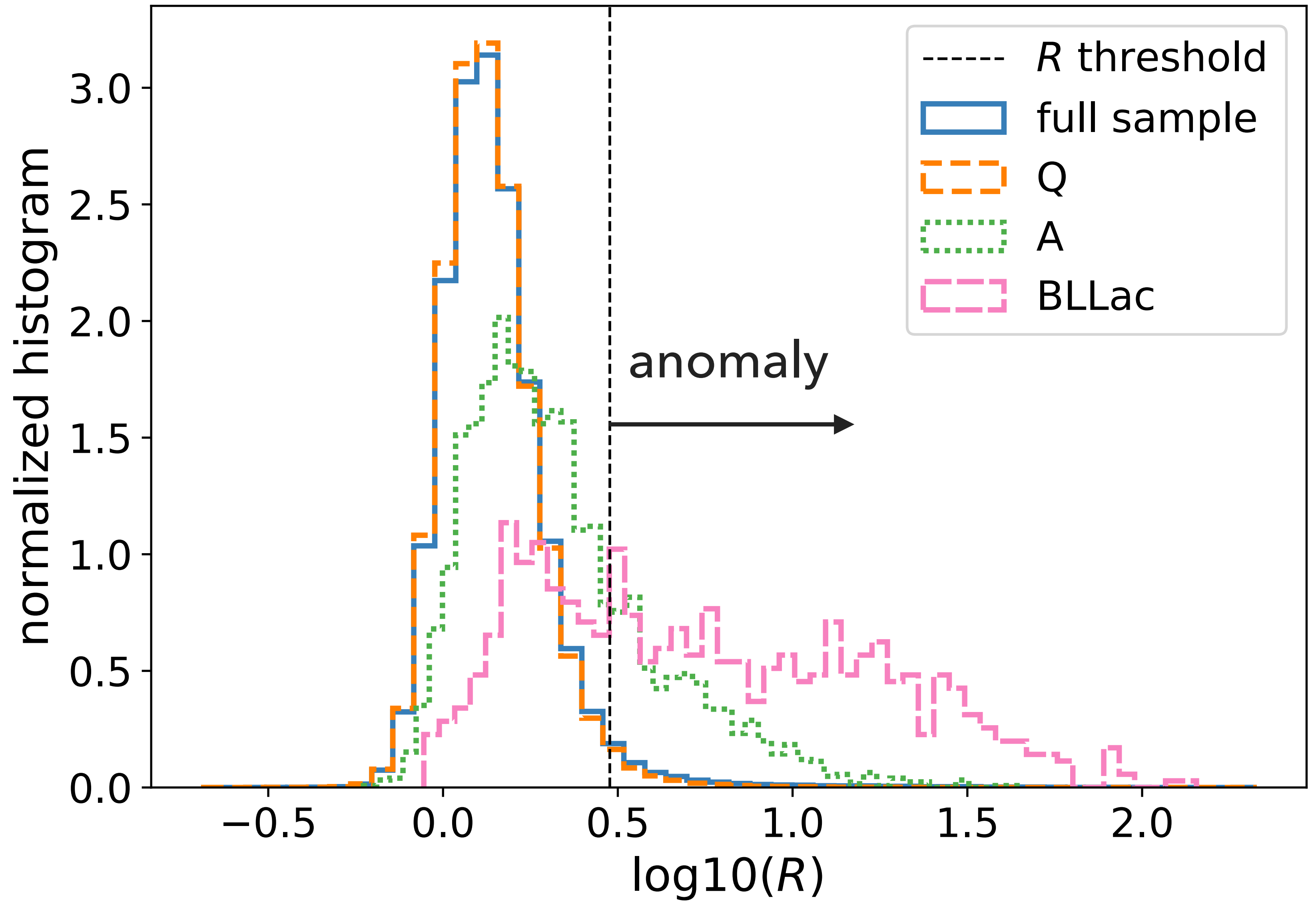
Sánchez-Sáez et al. 2021, AJ, 162, 206

We trained the VRAE architecture with a balanced sample, and then applied it to the full set of 230,451 light curves. We selected anomalies by:

Using the reconstruction error of the VRAE as an anomaly score:

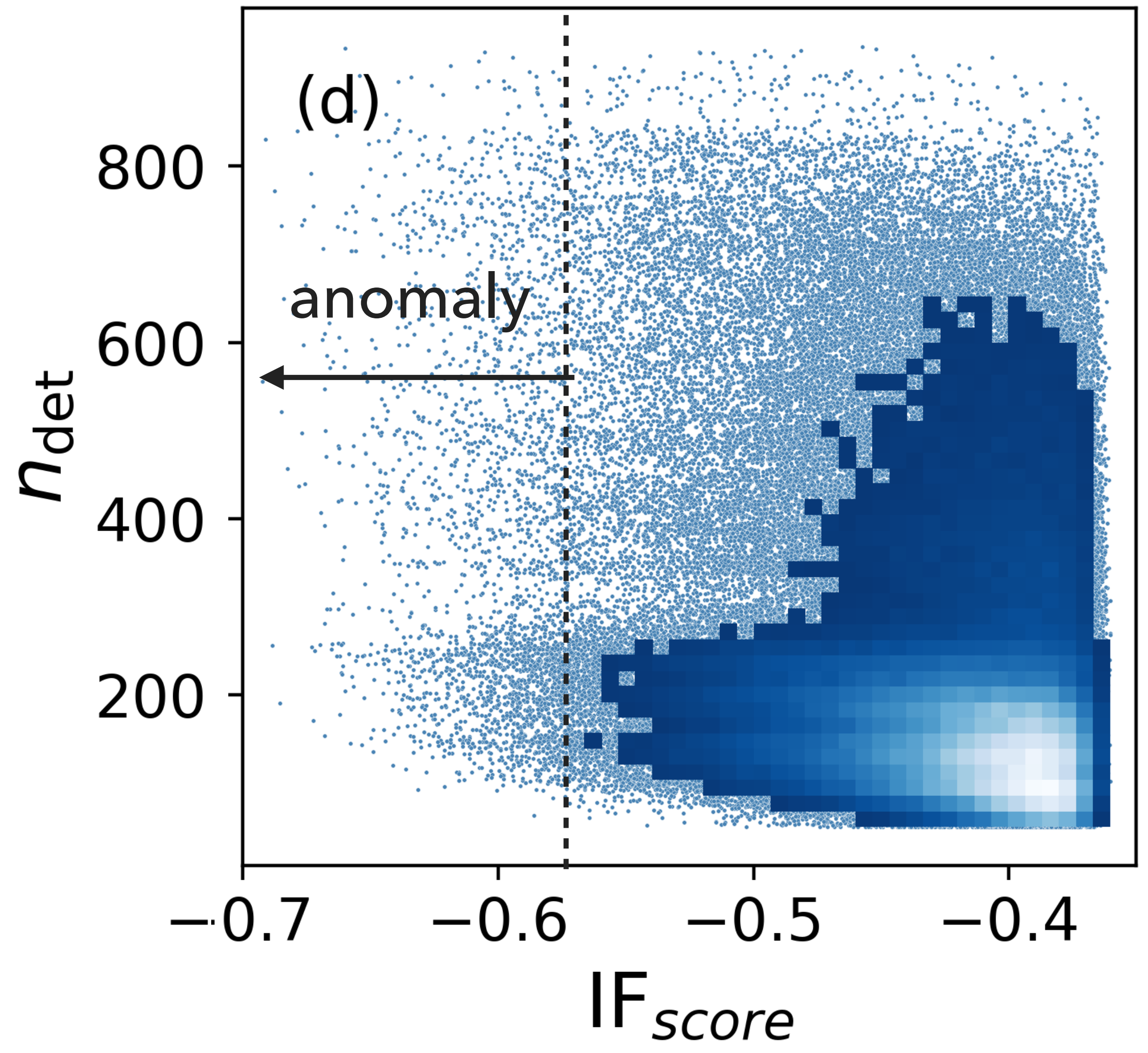
$R > 3$

$$R = \frac{1}{N_T}\sum_{i}^{N_T}\frac{(m_i - m_{pred_i})^2}{\mathrm{err}_i^2 + \mathrm{err}_{pred_i}{}^2}\,.$$



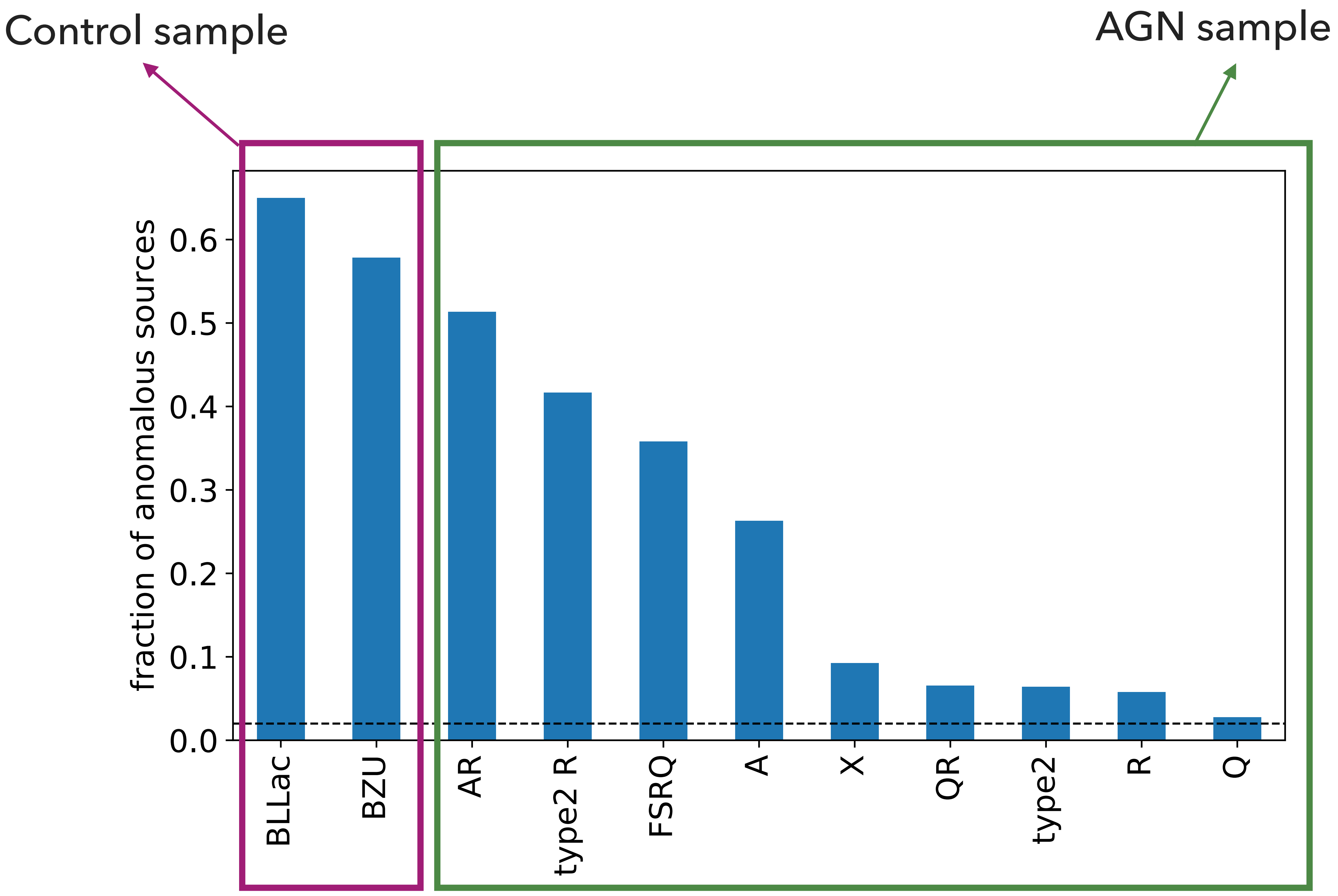
Using the latent space attributes with an Isolation Forest algorithm (IF):

IF_score < IF threshold 2% contaminants (-0.57633)



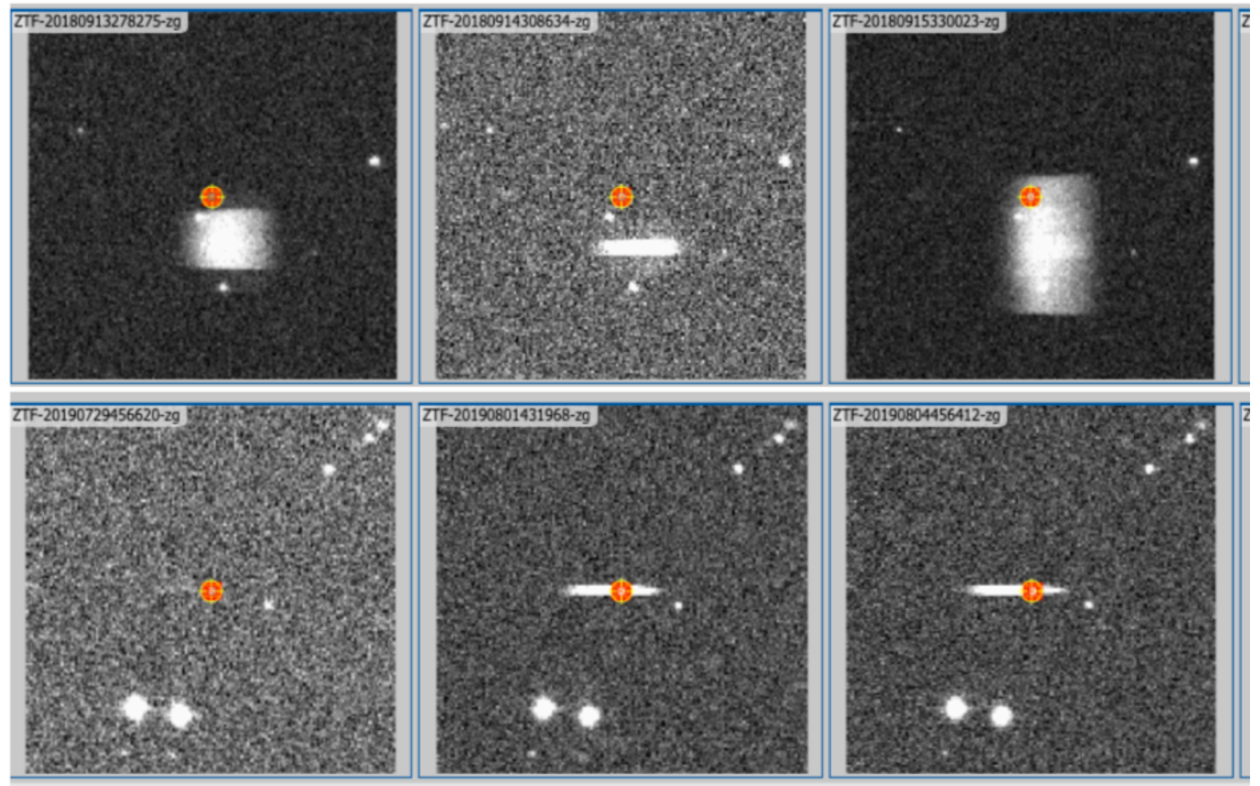
VRAEs for AGN variability anomaly detection: results

We selected 8,809 anomalies.

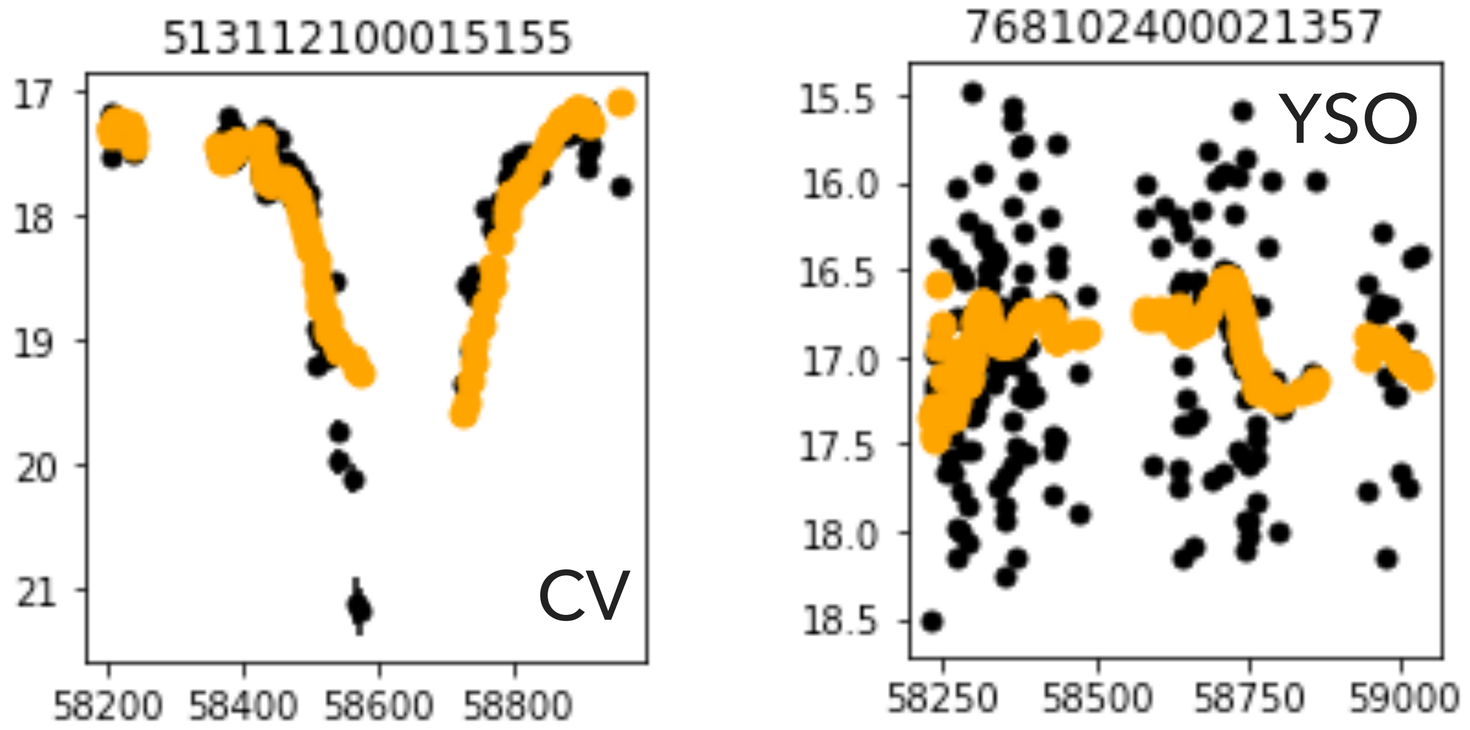


Sánchez-Sáez et al. 2021, AJ, 162, 206

Dominated by photometric issues



And miss-classified sources

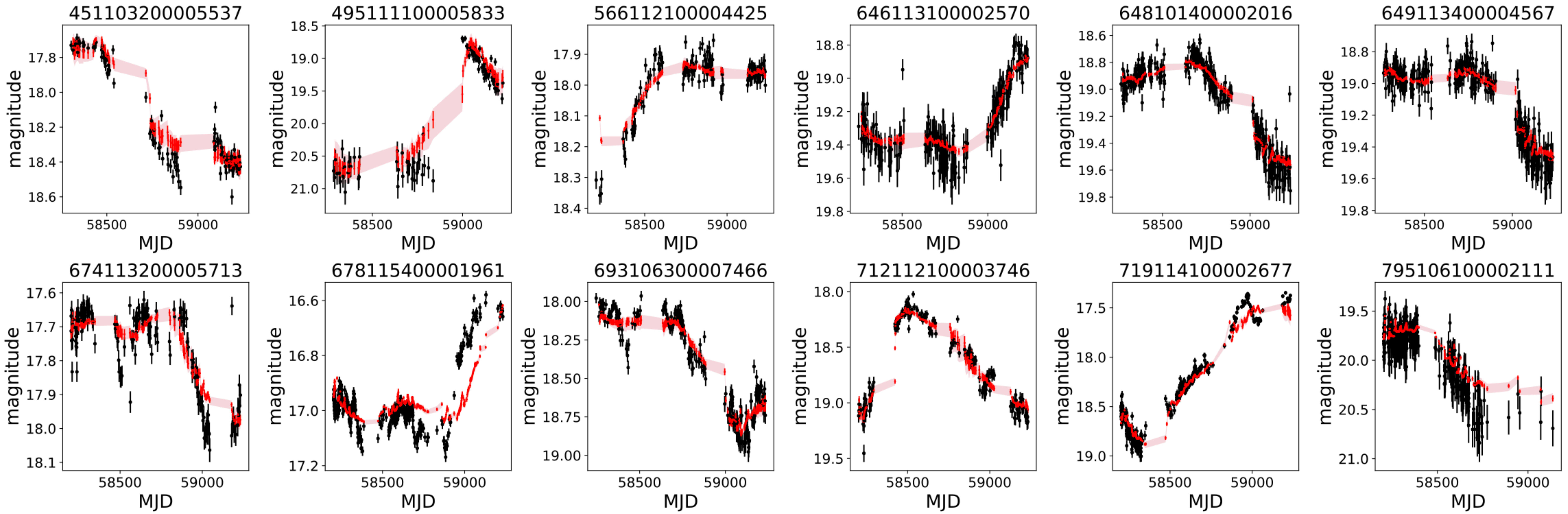


CSAGN candidates

Sánchez-Sáez et al. 2021, AJ, 162, 206

We visually inspected the list of candidates and selected as promising CSAGN candidates those anomalies that present evidence of flares, and/or abrupt increment or decrement in the luminosity. **We identified 75 CSAGN candidates** (65% are regular QSOs).

Further spectroscopic follow-up is required to confirm the nature of our candidates. Although 4 are known CSAGN candidates (Graham+2020), 2 have been spectroscopically confirmed (M. Graham, private communication), and 28 are candidates using other techniques (Graham+ in prep).



Summary

- Variability-ML-based classifiers can help us to select AGN populations that can me missed by more traditional selection techniques, particularly low-mass and low-Eddington rate sources.

- The ZTF DR light curve classifier corresponds to the first attempt to identify multiple classes of transients and persistently variable and non-variable sources from ZTF DF light curves of extended and point sources. The main motivation of this model was to identify AGN candidates, but it can be used for more general time-domain astronomy studies. We used a hierarchical local classifier per parent node approach, to classify a total of 17 classes, including non-variable objects, transients, and stochastic and periodic variables.

- Real-time detection of CSAGN events is crucial to understand these events and to improve our knowledge of the physical mechanisms behind AGN variability. We used a Variational Recurrent Autoencoder (VRAE) architecture to model AGN light curves from the ZTF DRs. We used reconstruction error and the latent space attributes to search for anomalous AGN light curves. We found 8,809 anomalies. These anomalies are dominated by bogus candidates (photometric issues, miss-classified sources in the original catalogs), but we were able to identify 75 promising CSAGN candidates.

Paula.SanchezSaez@eso.org

THANK YOU!