Searching for Changing-state AGNs in massive data sets with anomaly detection

Paula Sánchez Sáez

H. Lira, L. Martí, N. Sánchez-Pi, J. Arredondo, F. E. Bauer, A. Bayo, G. Cabrera-Vives, C. Donoso-Oliva, P. A. Estévez, S. Eyheramendy, F. Förster, L. Hernández-García, A. M. Muñoz Arancibia, M. Pérez-Carrasco, M. Sepúlveda, and J. R. Vergara

MOTIVATION

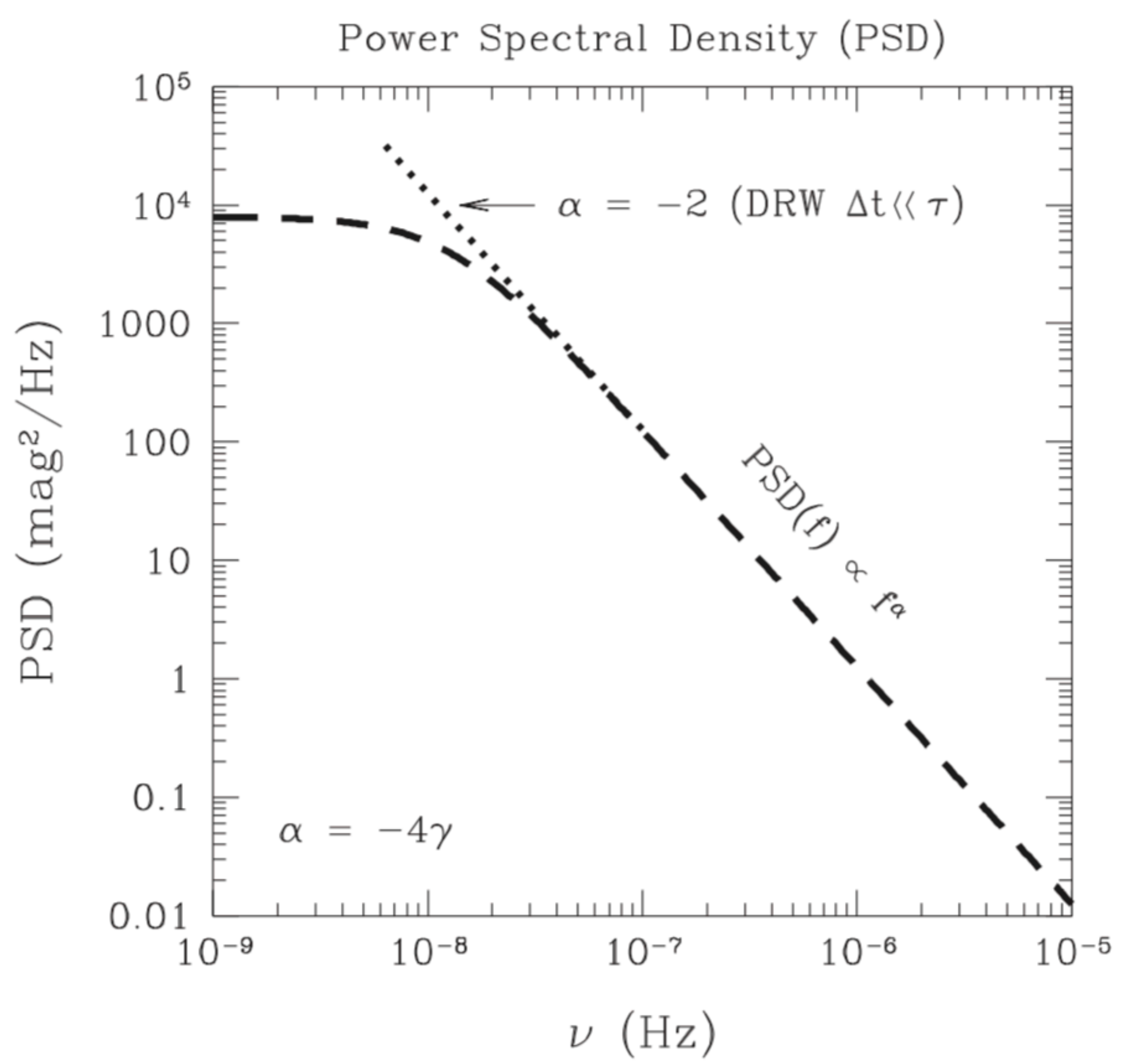
Active Galactic Nuclei (AGN)

AGNs are powered by the release of gravitational energy related with the accretion of material onto a supermassive black hole (SMBH), with masses larger than $10^6 M_\odot$

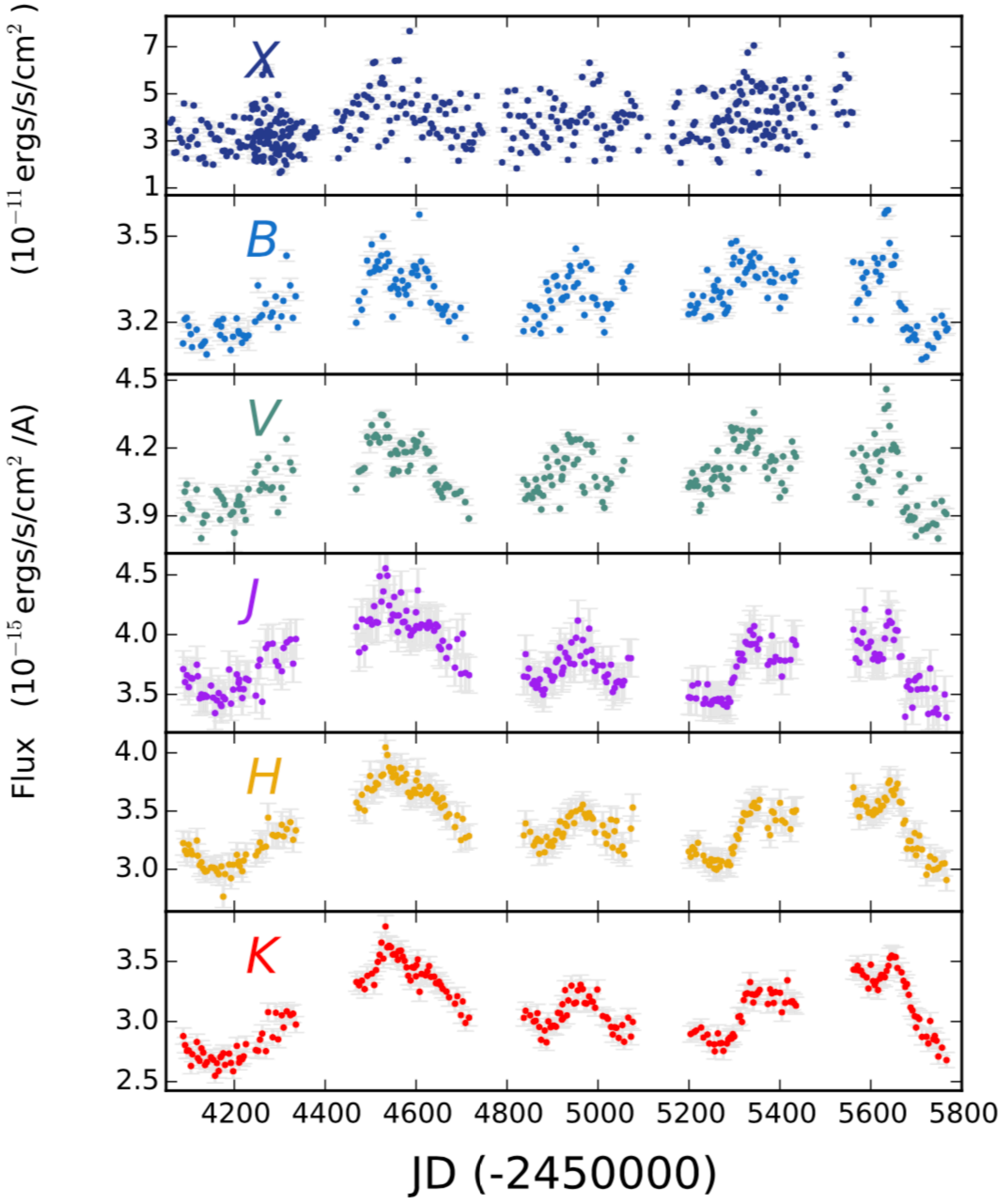
M87, Credits: ESO

AGN are variable

- AGN variability seems to be well described as a stochastic process.
- The characteristic time-scales of the variability range from hours to years, with the shortest time-scales being associated with shorter emission wavelengths.



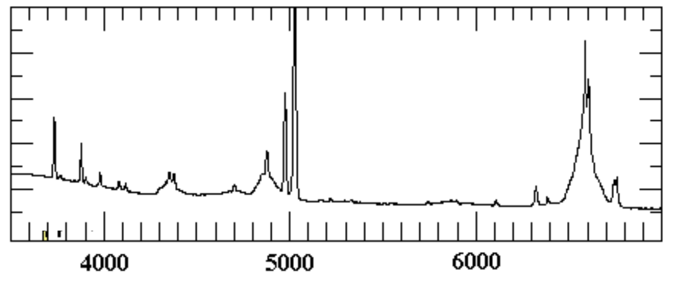
Kozlowski 2016



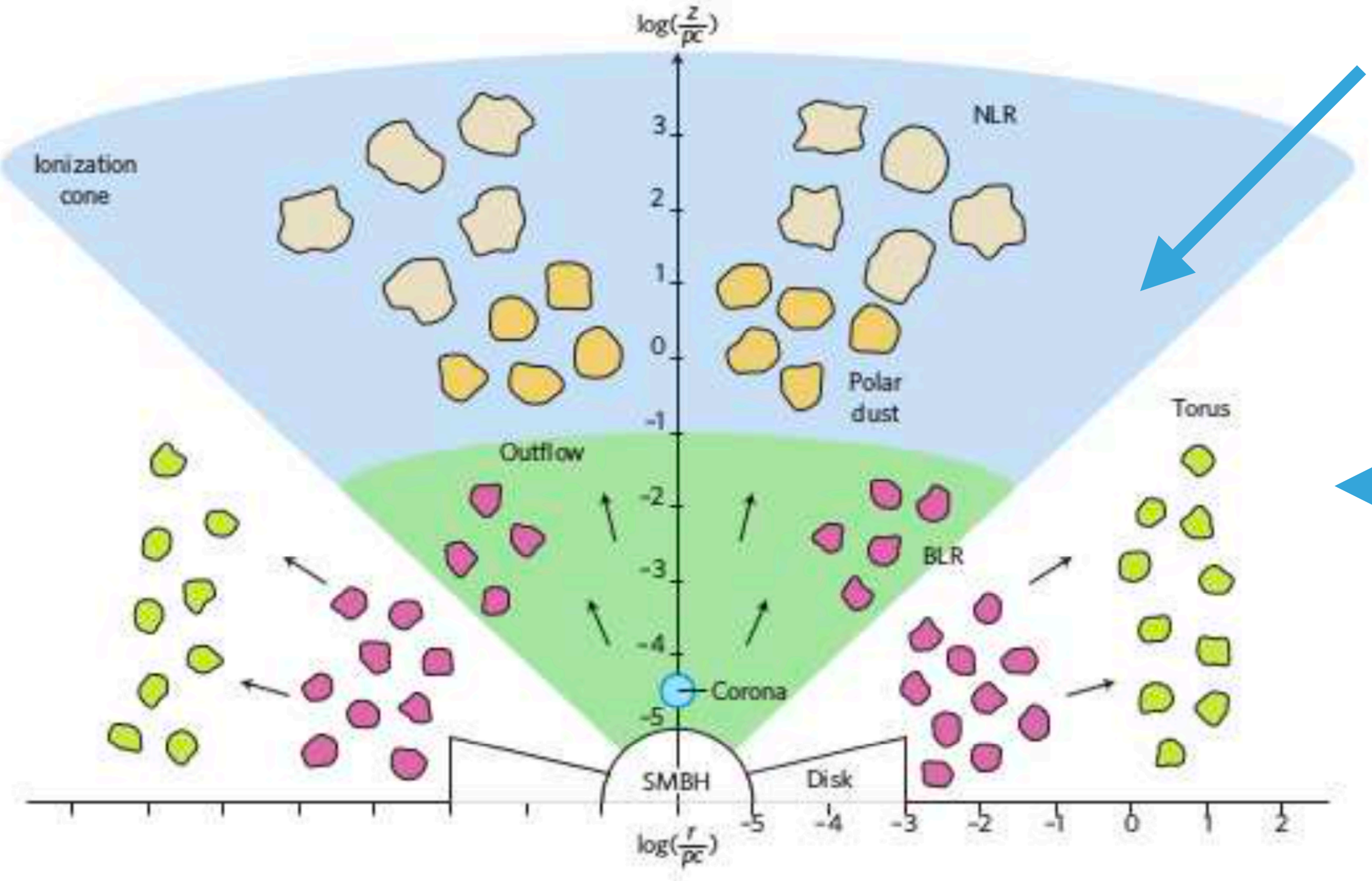
MCG-6-30-15 from Lira et al. 2015

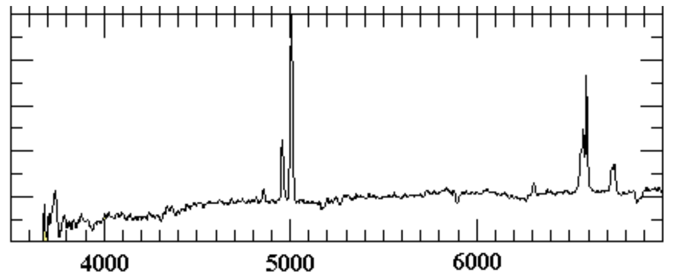
Changing-state AGNs

Changing-state AGNs (CSAGNs) in the optical range correspond to sources that change their classification as type 1 or type 2 AGN, as well as to sources that present large changes in the flux of their broad emission lines, within a timescale of months or years. This transition phase is accompanied by a drastic change in the AGN continuum flux.



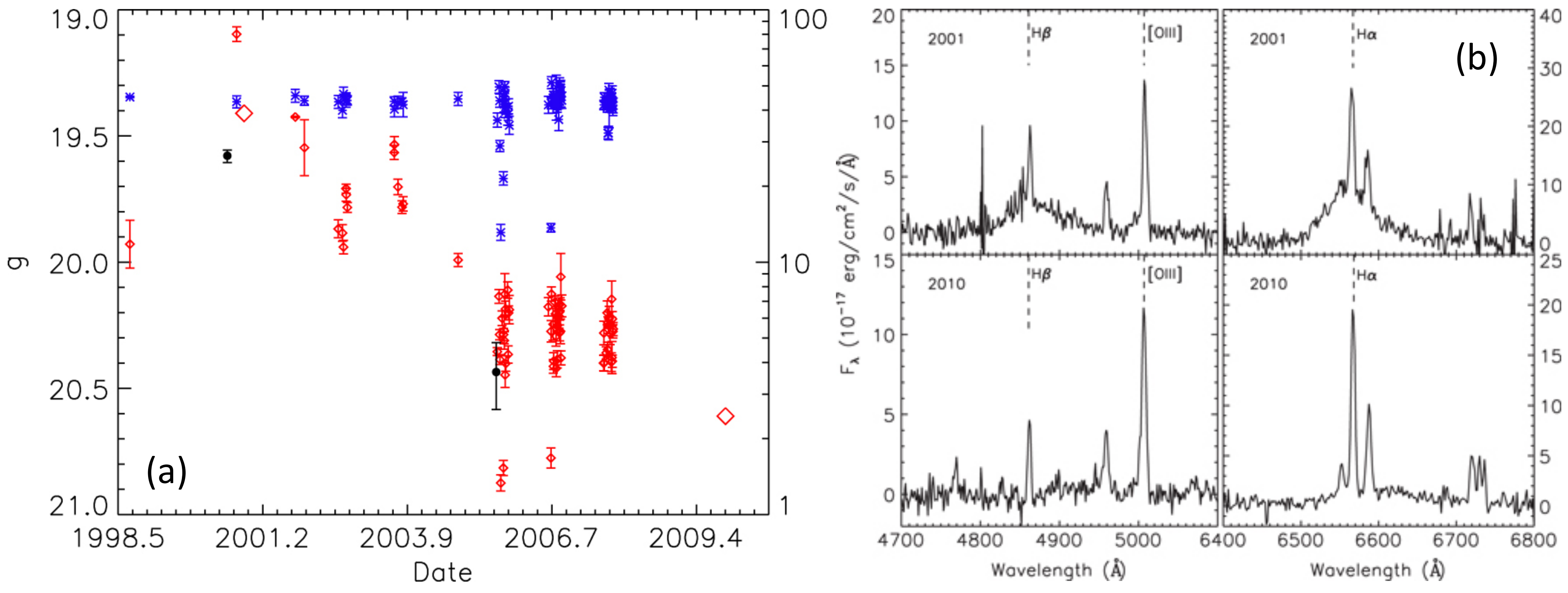
Type 1





Type 2

Ramos Almeida & Ricci (2017)

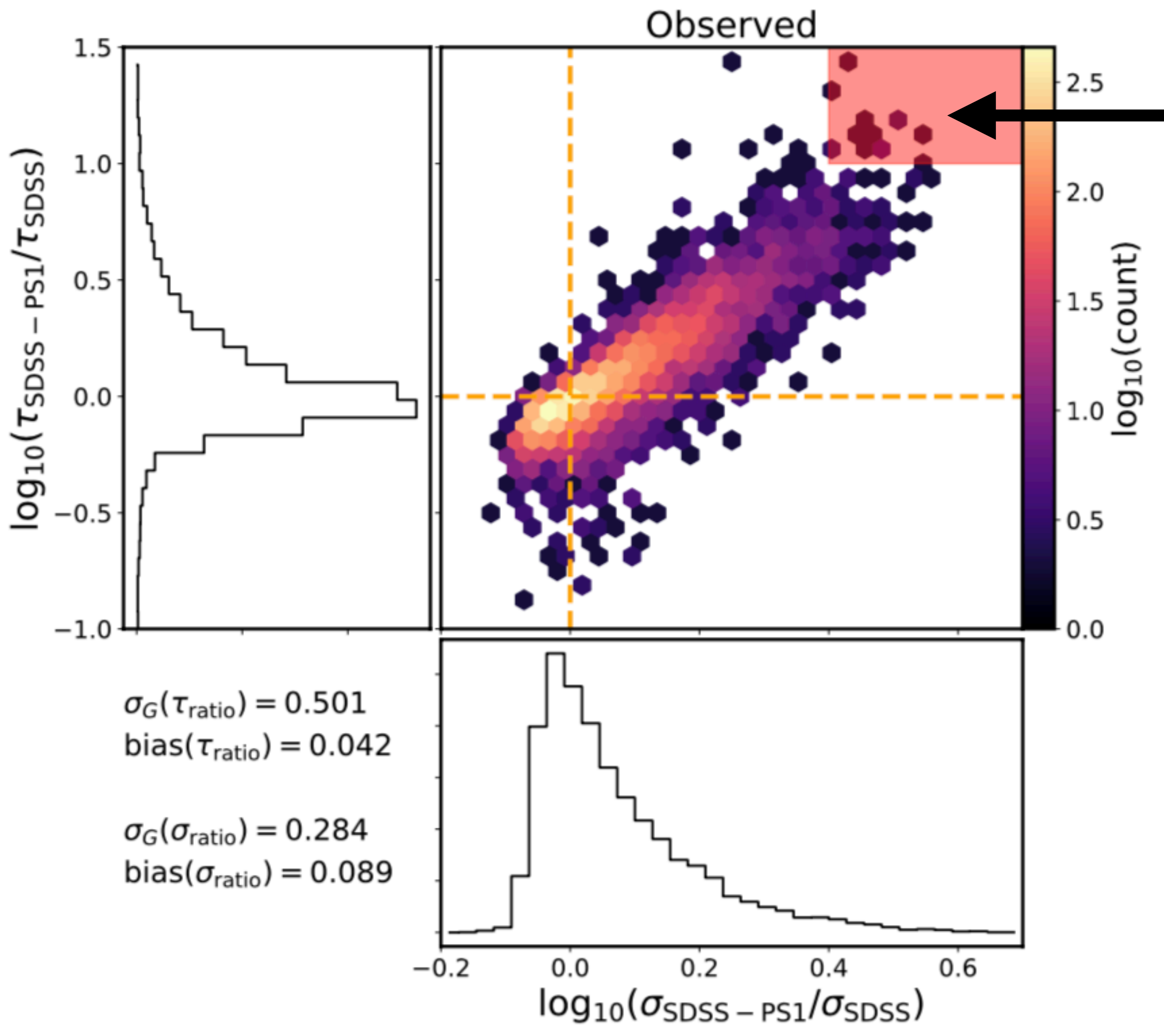


LaMassa et al. 2015

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

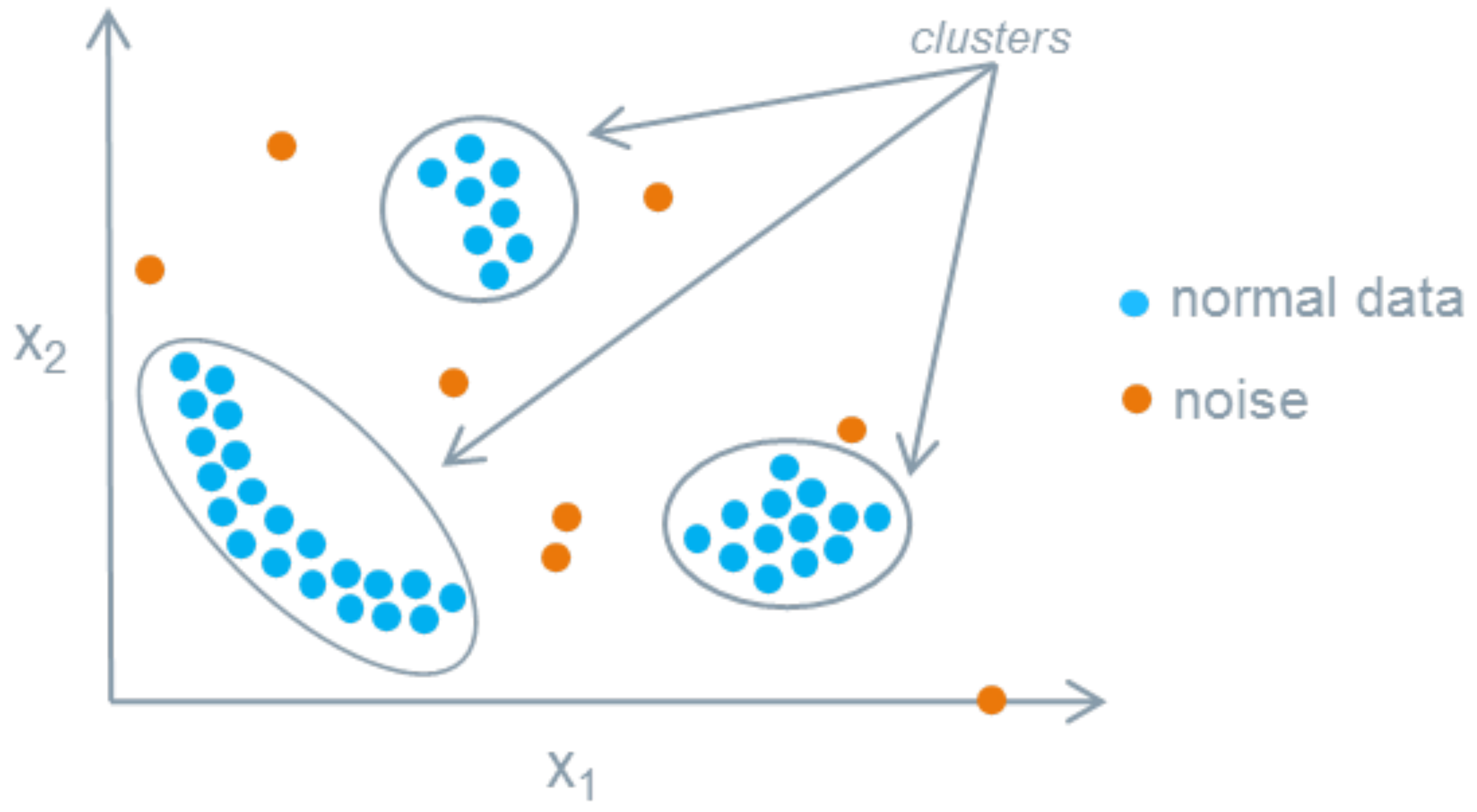
ZTF → VERA C. RUBIN OBSERVATORY

ANOMALY DETECTION TECHNIQUES FOR TIME SERIES

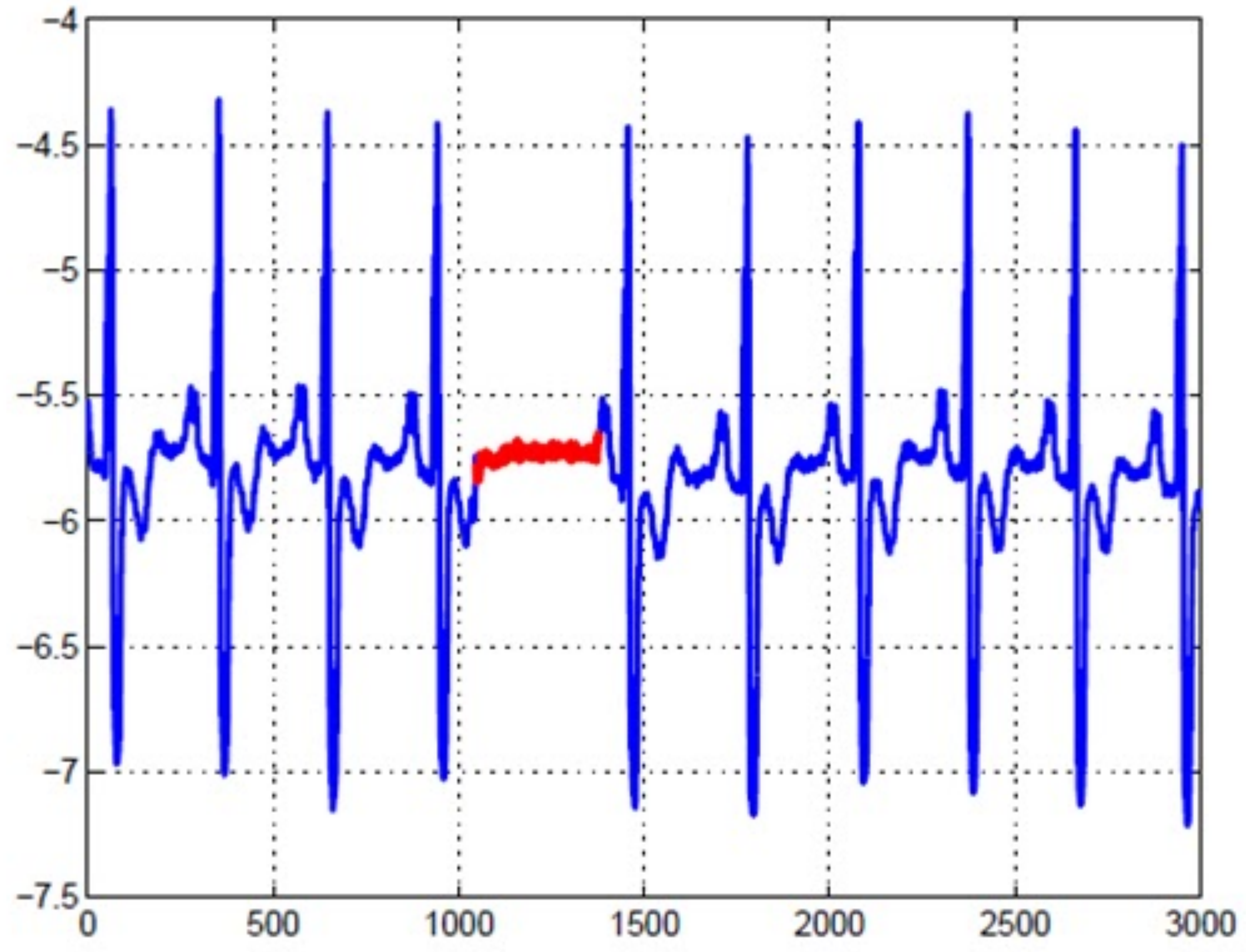
Anomaly detection (AD)

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.

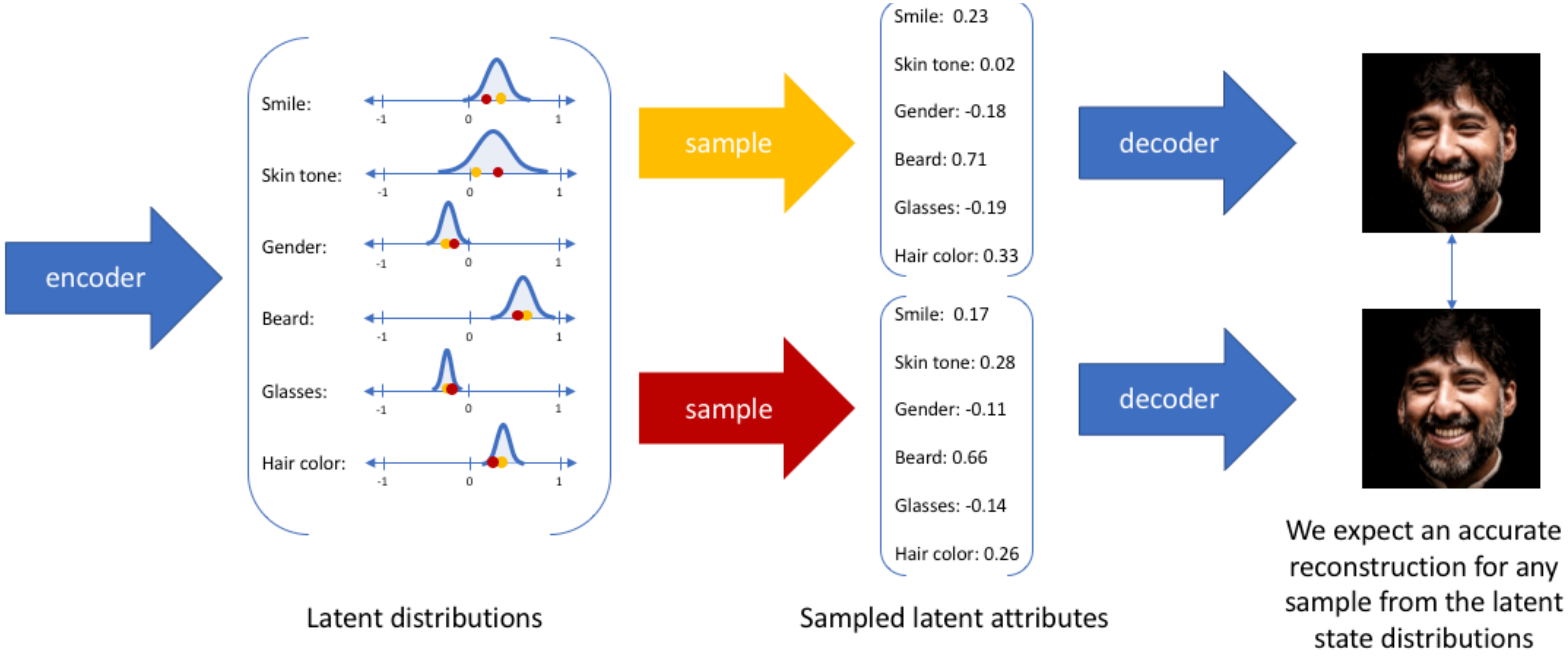


Contextual anomaly: searching for objects that suddenly start presenting unusual behaviors.



Variational Autoencoders (VAEs)

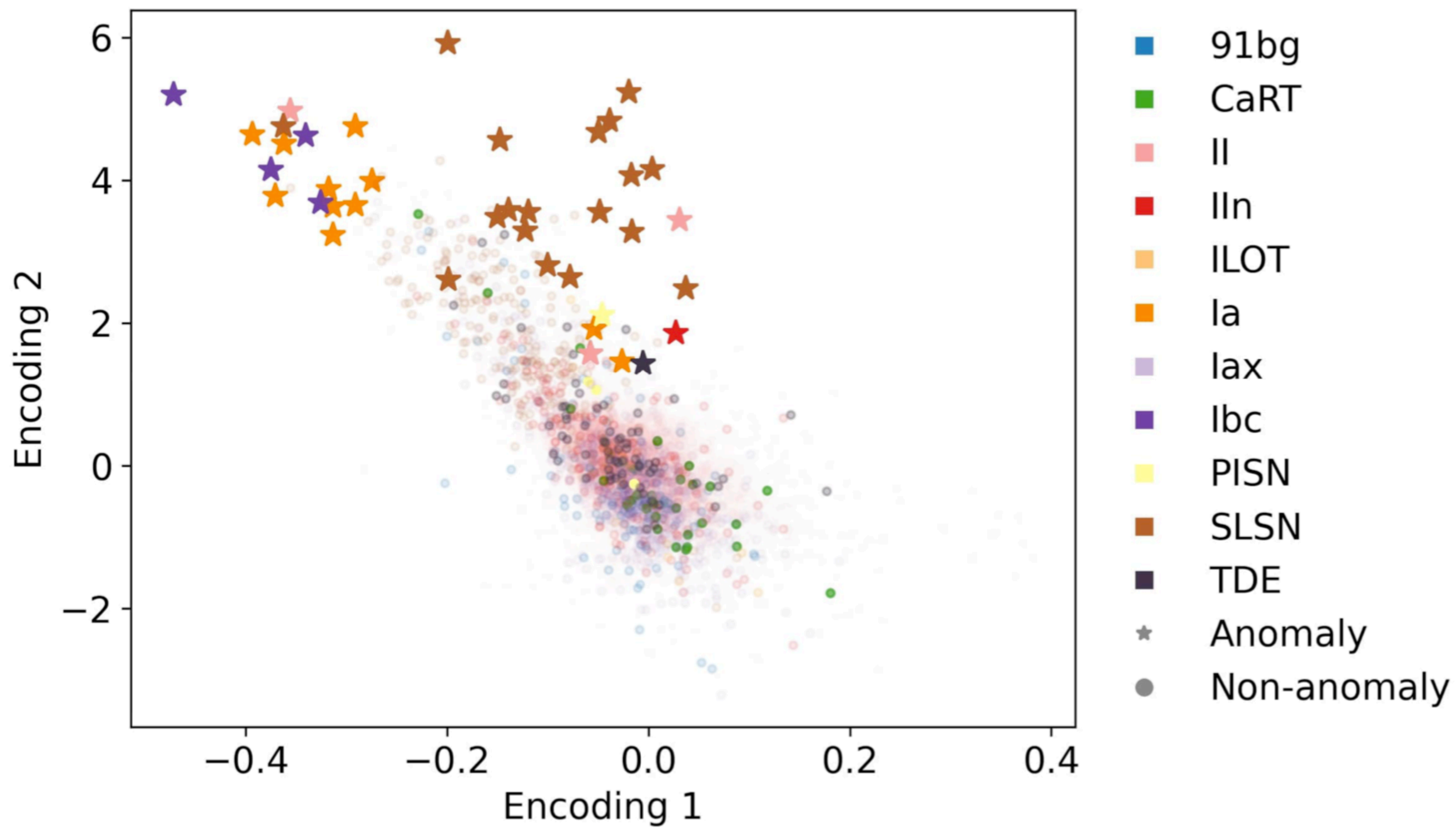
VAEs correspond to a modification of the more classical Autoencoder (AE) architectures. In this case, the latent representations are described by multivariate normal distributions, where each attribute or feature in the latent space is described by a latent mean (μ) and a latent variance (σ^2), which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



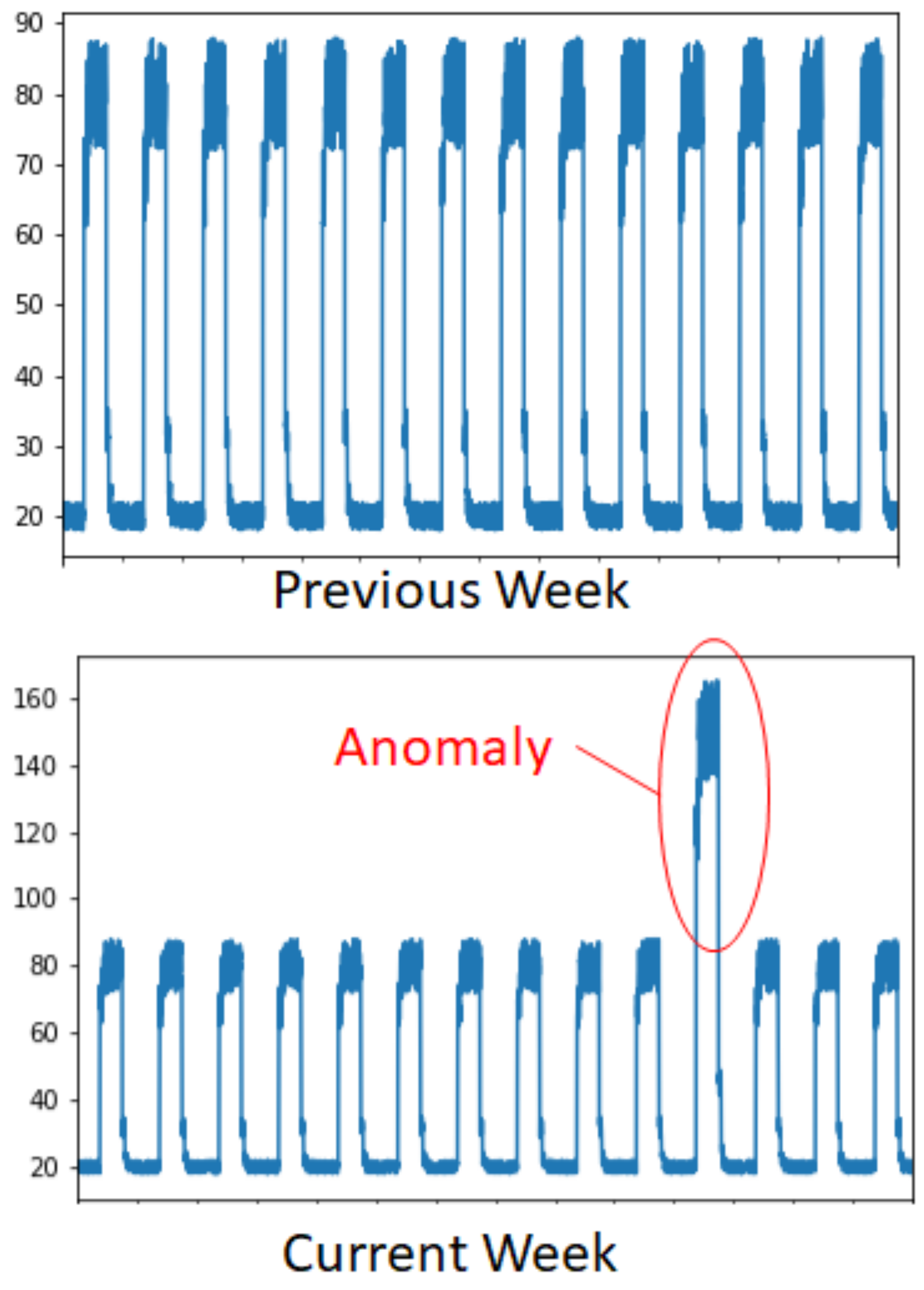
Credits: https://www.jeremyjordan.me/variational-autoencoders/

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



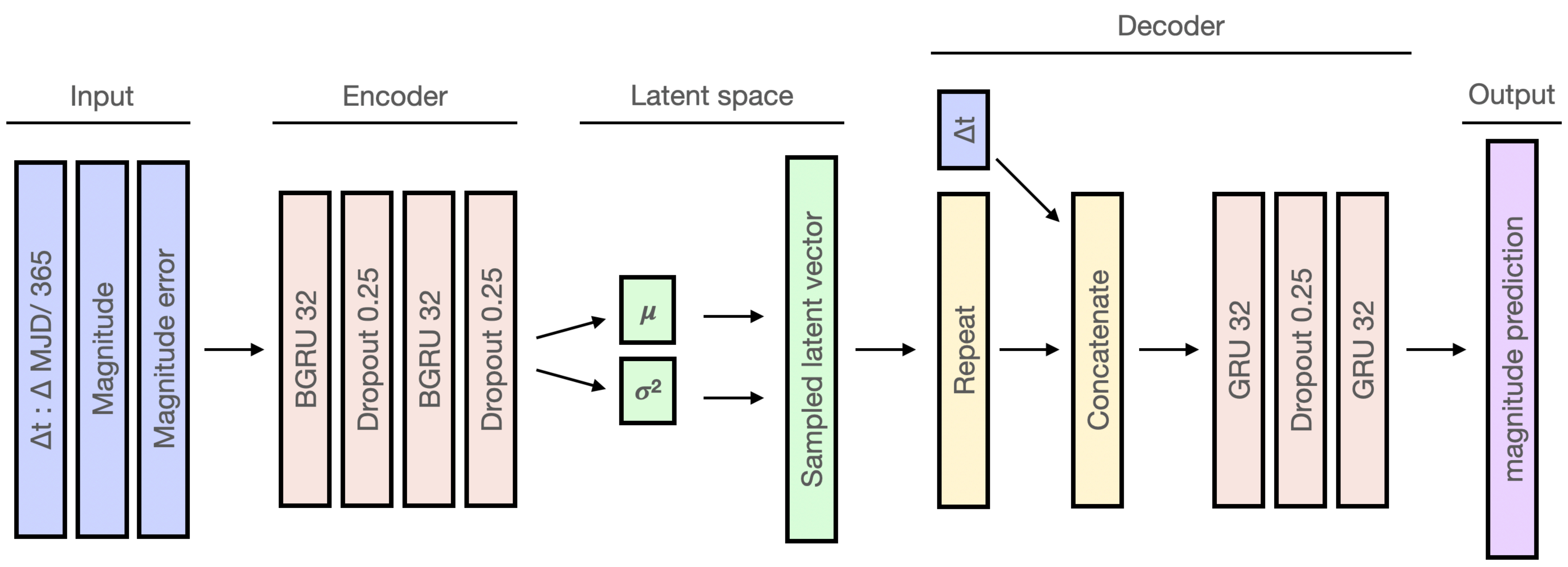
SEARCHING FOR ANOMALOUS AGN VARIABILITY WITH ANOMALY DETECTION

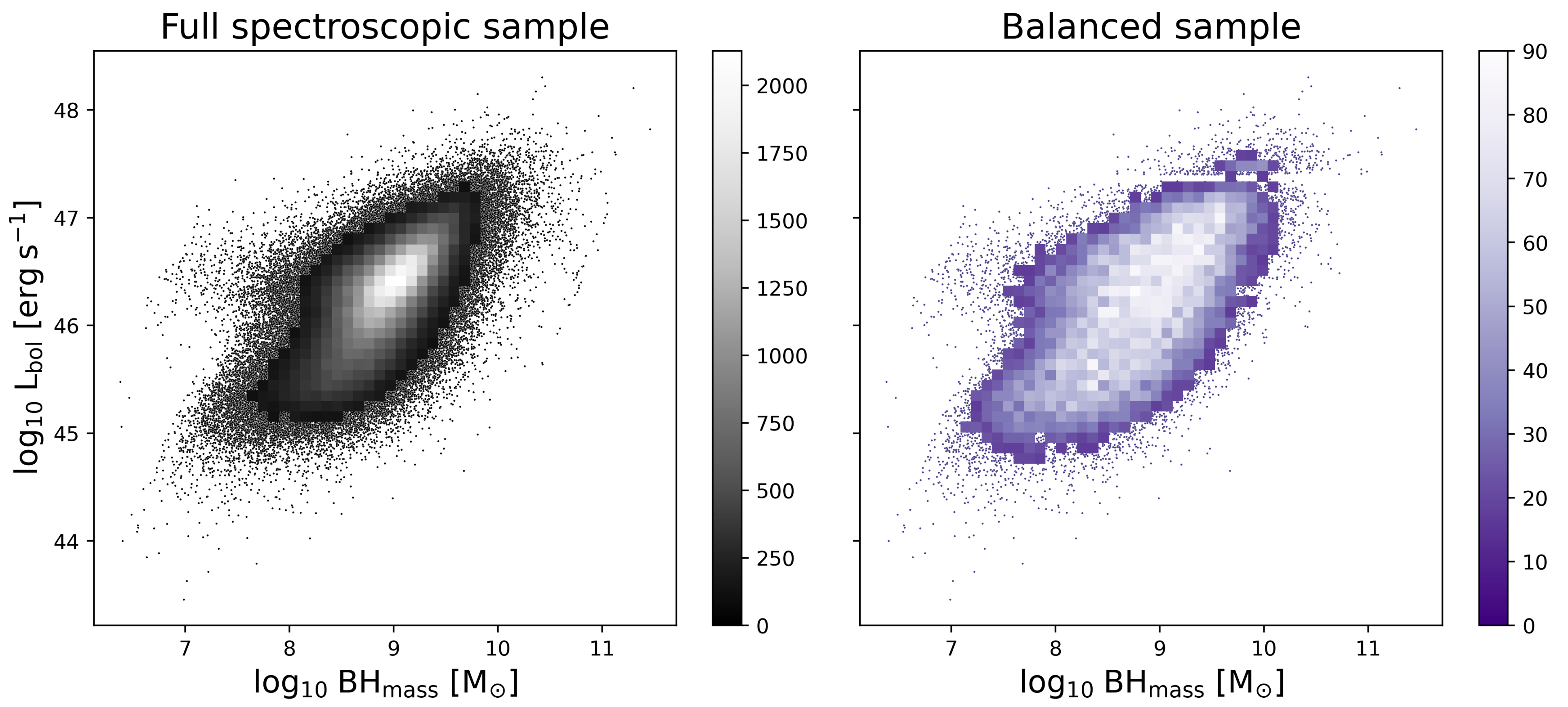
VRAEs to model AGN variability

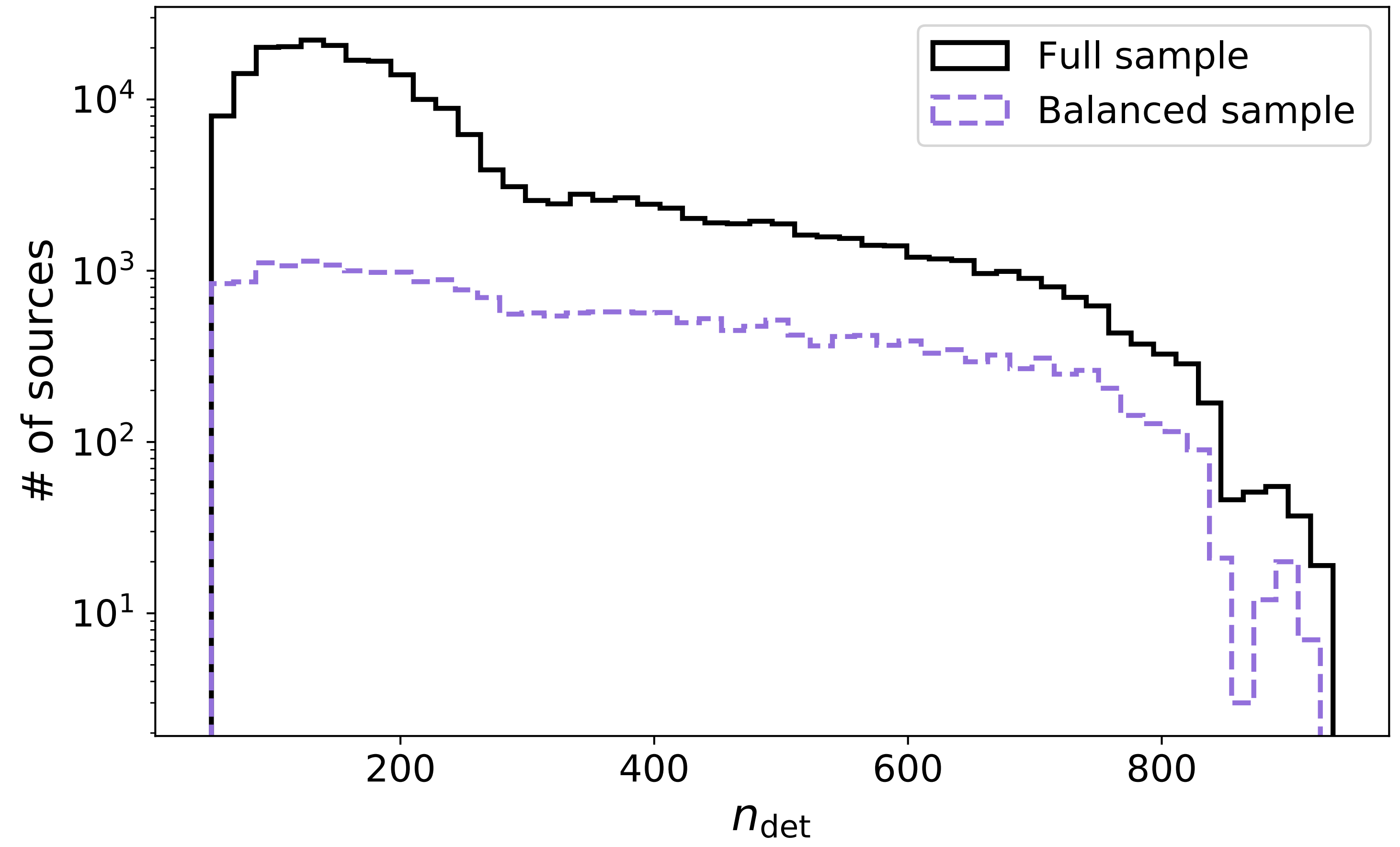
Sánchez-Sáez et al. 2021, AJ, 162, 206

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.



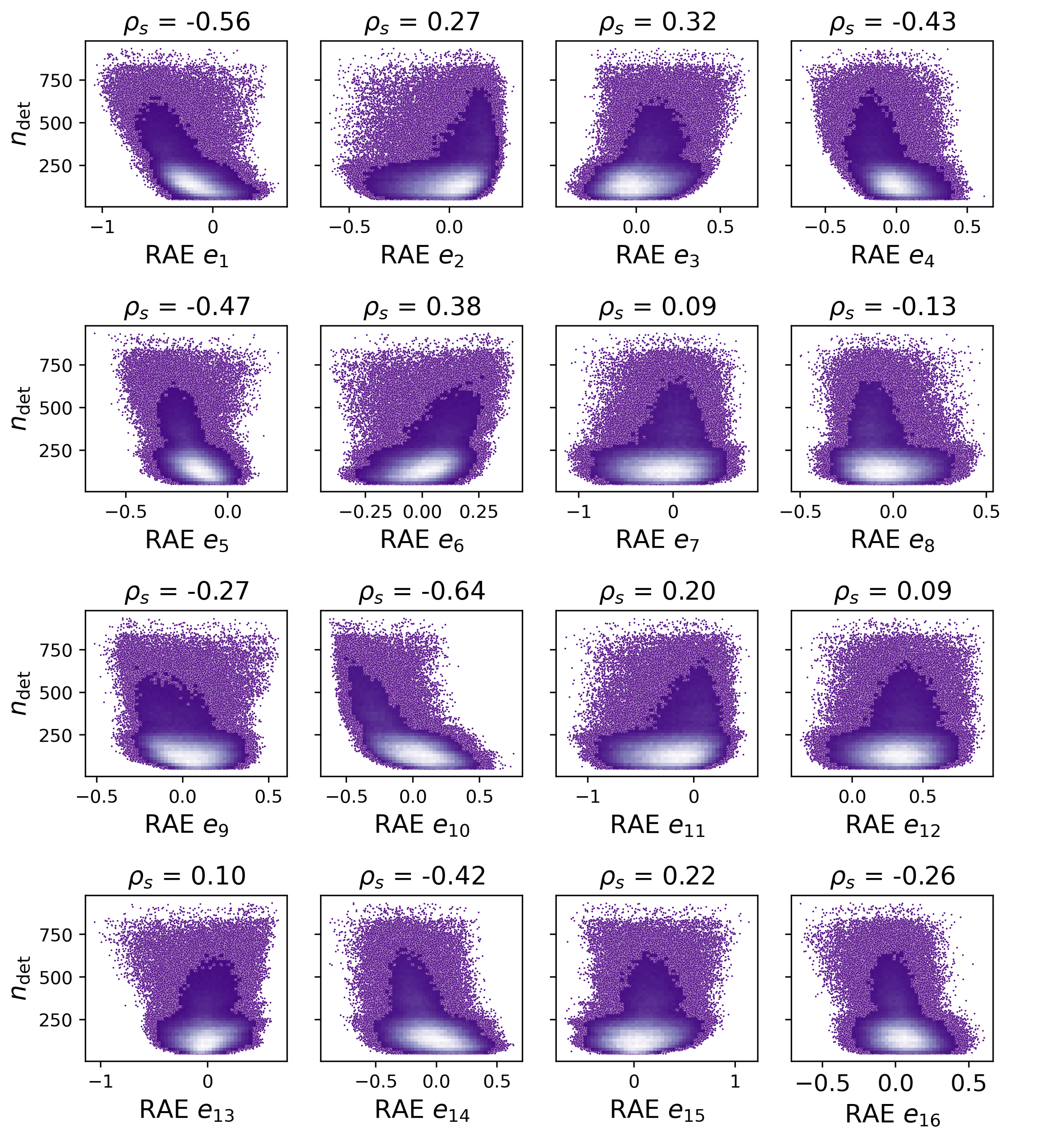




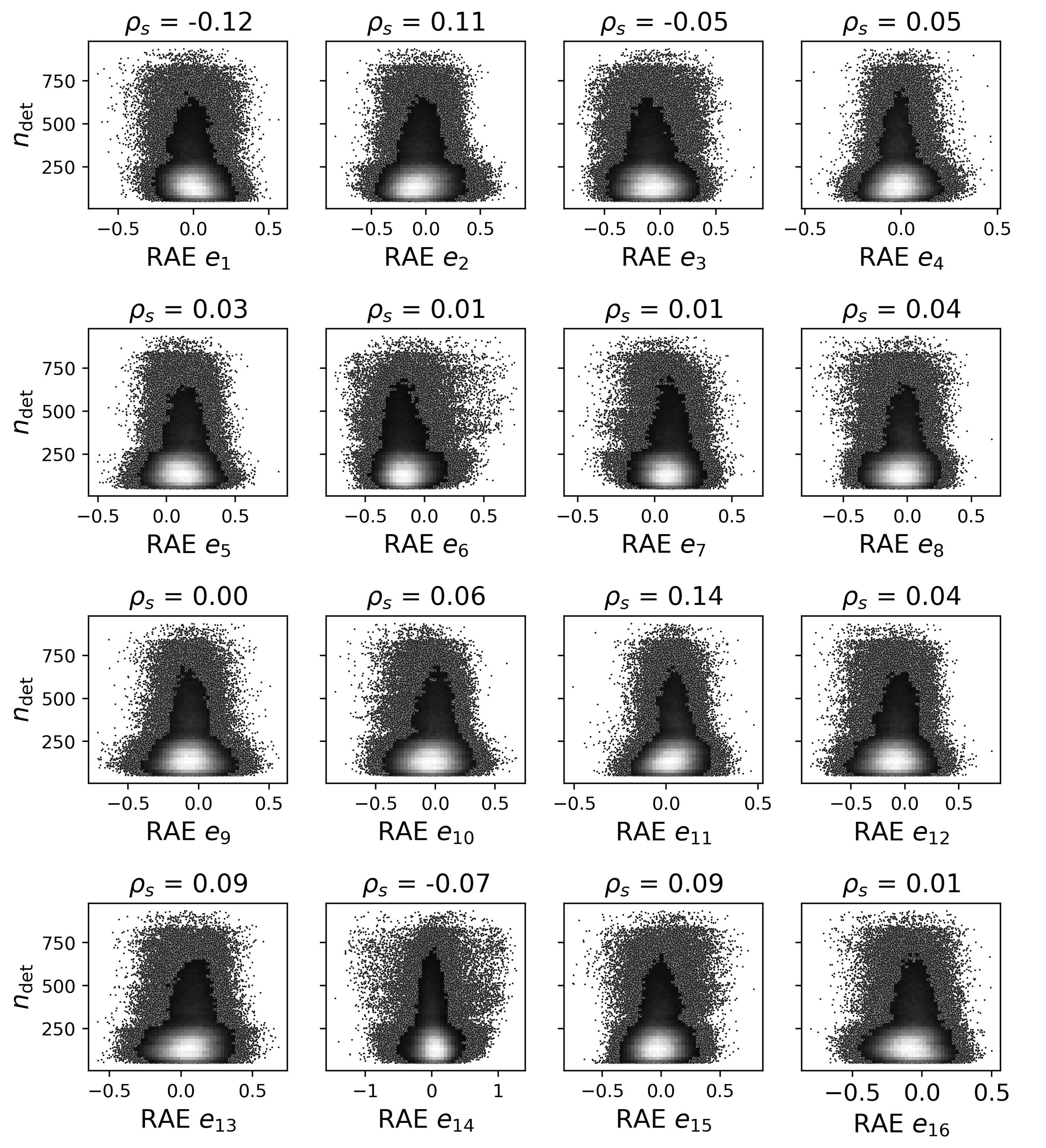
VRAEs to model AGN variability

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

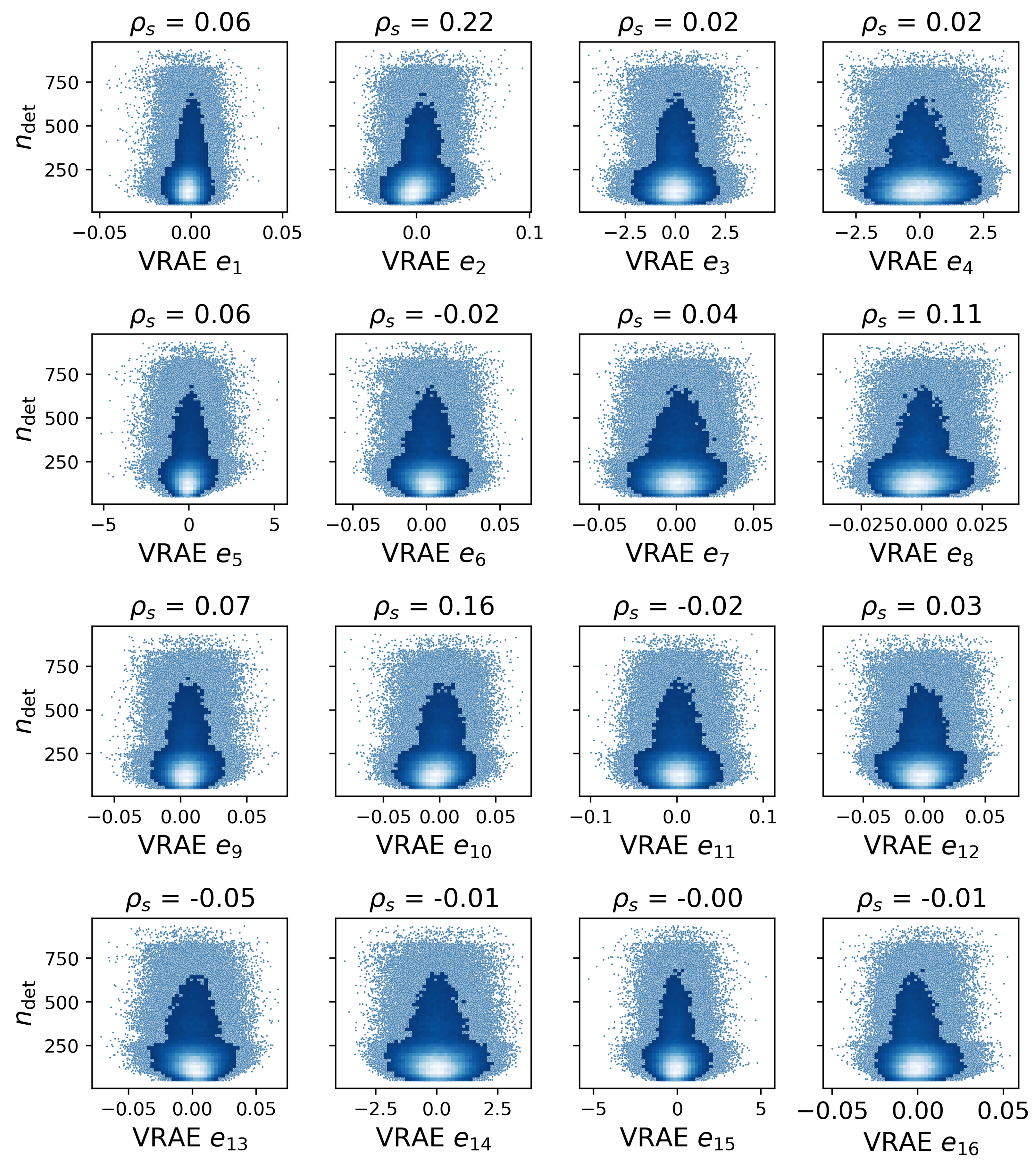
We first used the Tachibana+2020 architecture to model the AGN light curves from ZTF DR5. We found high correlations between the latent space attributes with the number of epochs per light curve, although with low reconstruction errors.



Original RAE architecture. From Tachibana et al. 2020



RAE architecture, TS balanced by number of epochs



VRAE architecture, TS balanced by number of epochs

VRAEs for AGN variability anomaly detection

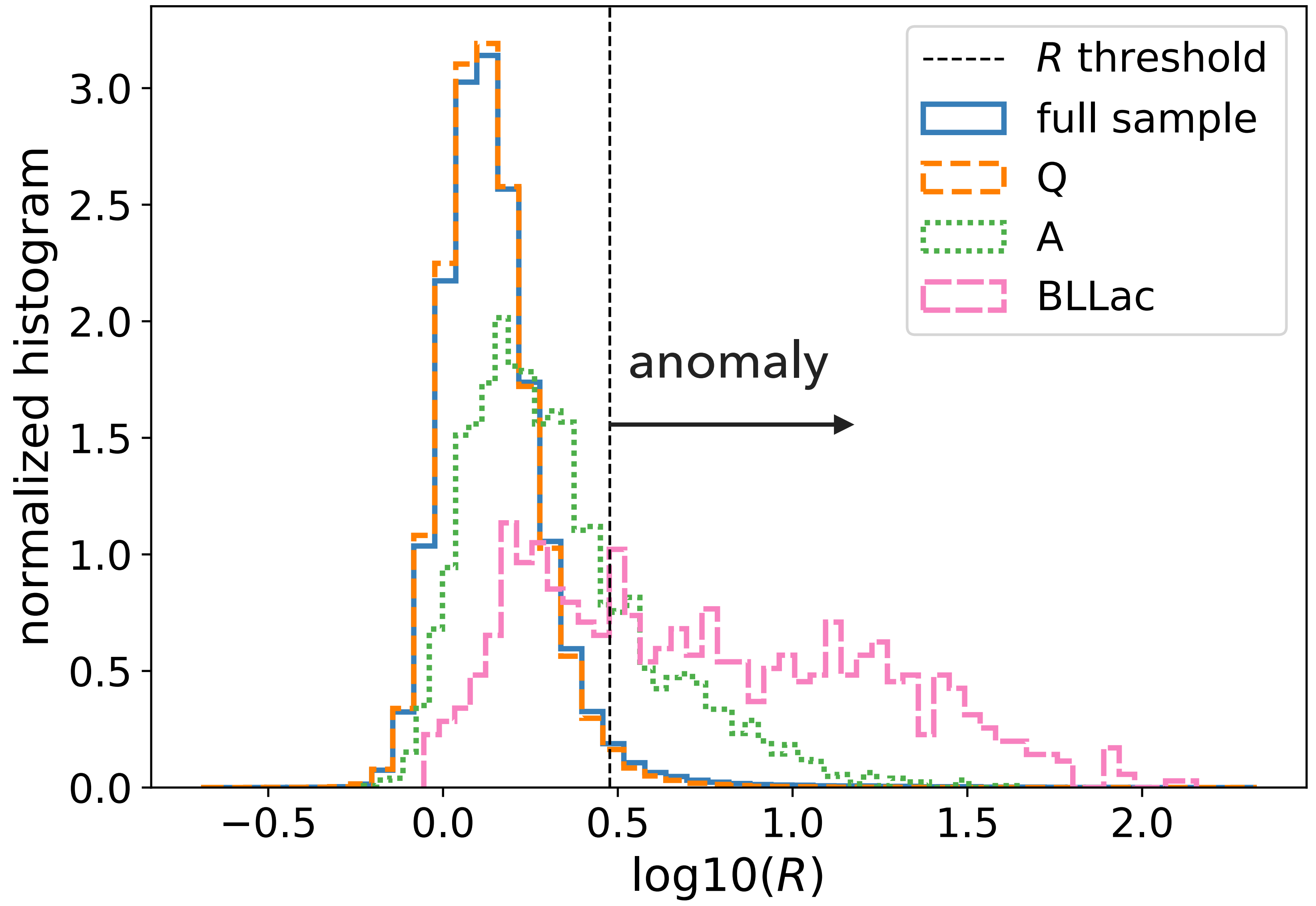
Sánchez-Sáez et al. 2021, AJ, 162, 206

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:

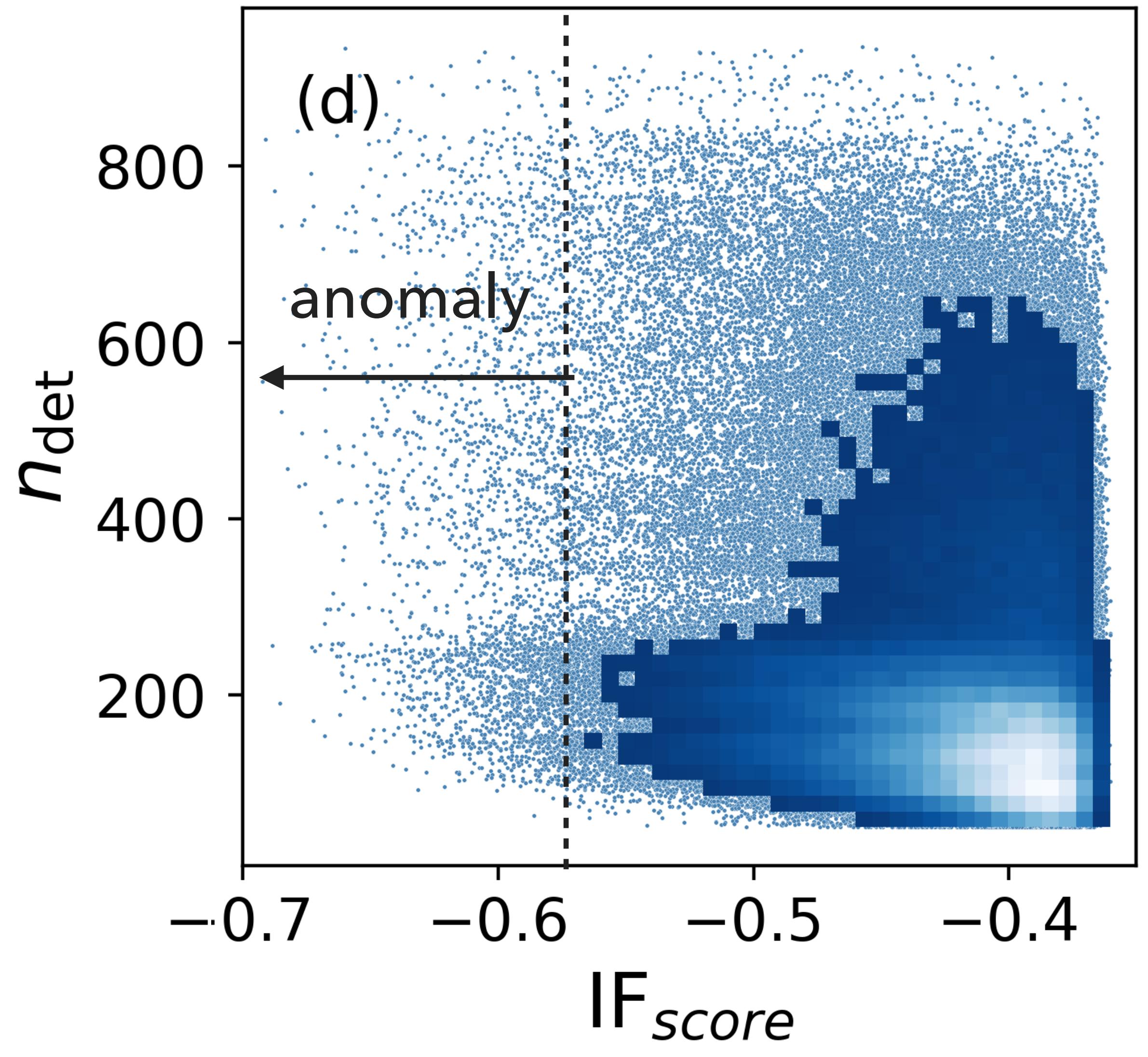
$R > 3$

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



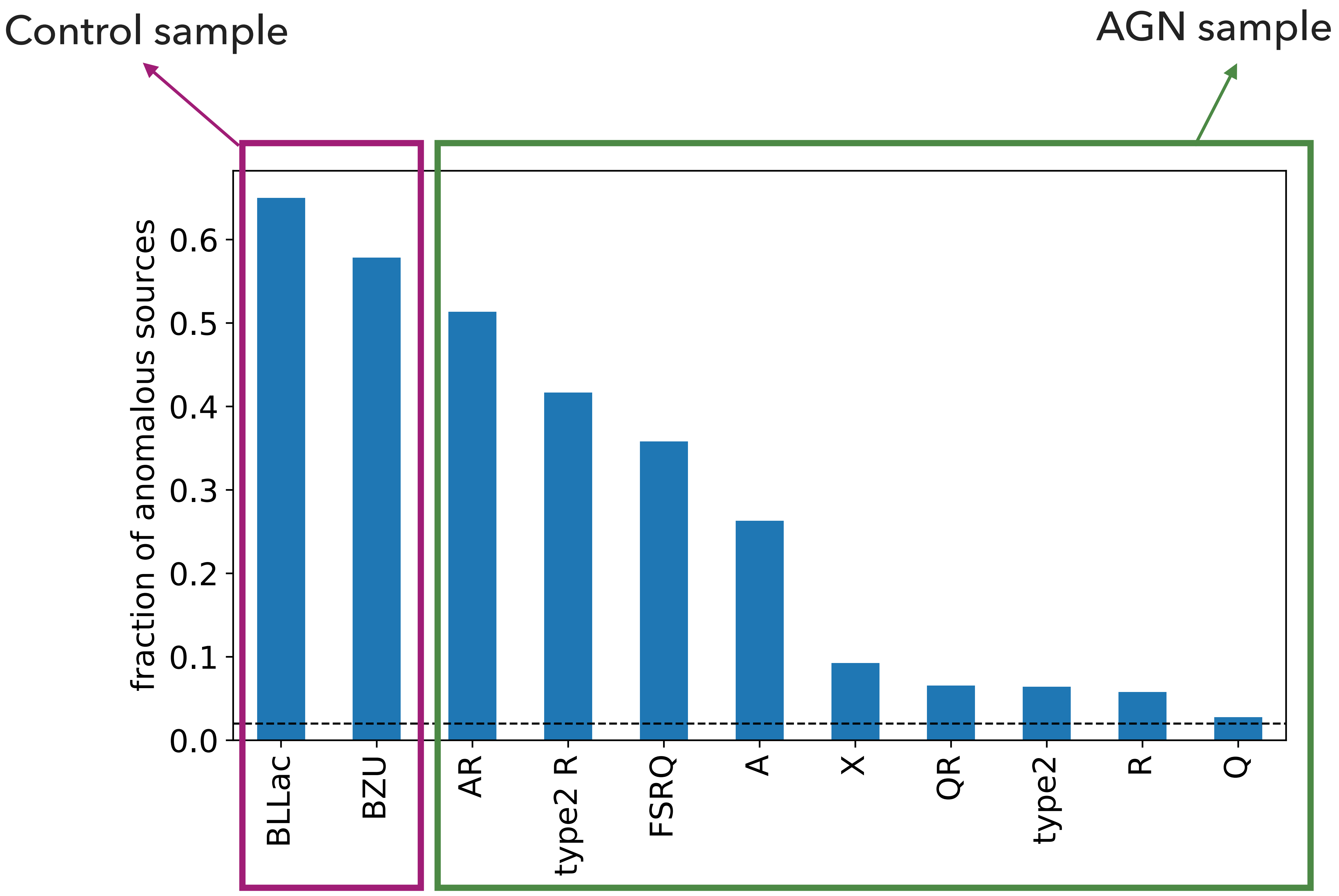
Using the latent space attributes with an Isolation Forest algorithm (IF):

IF_score < IF threshold 2% contaminants (-0.57633)

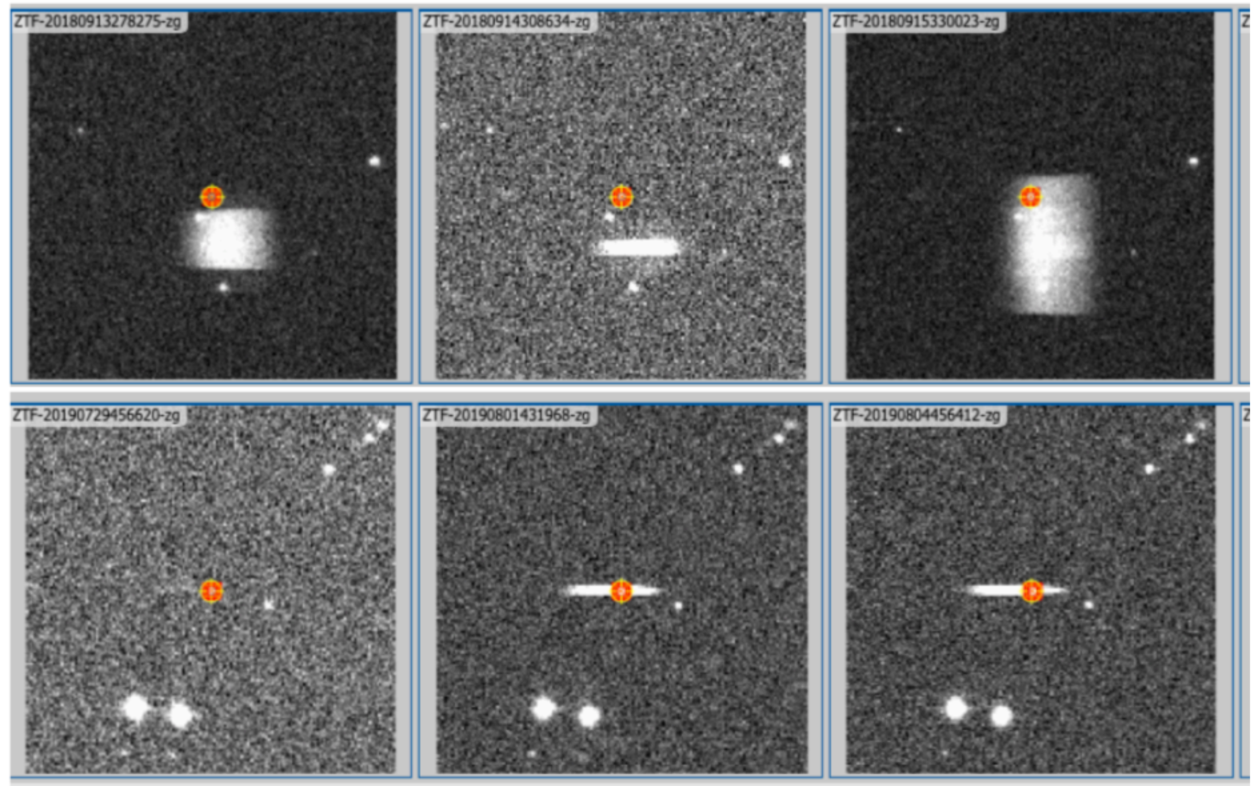


VRAEs for AGN variability anomaly detection: results

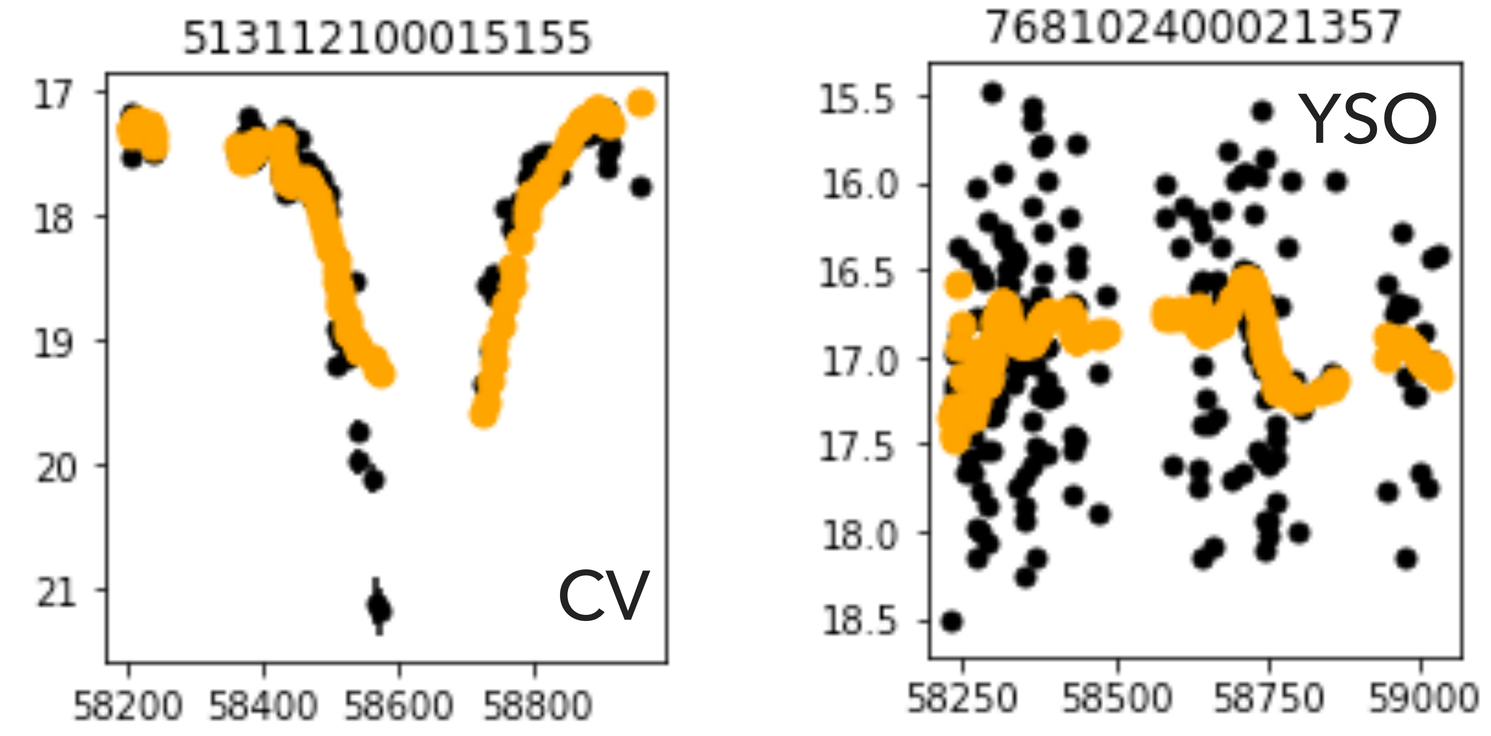
We selected 8,809 anomalies.



Dominated by photometric issues



And miss-classified sources



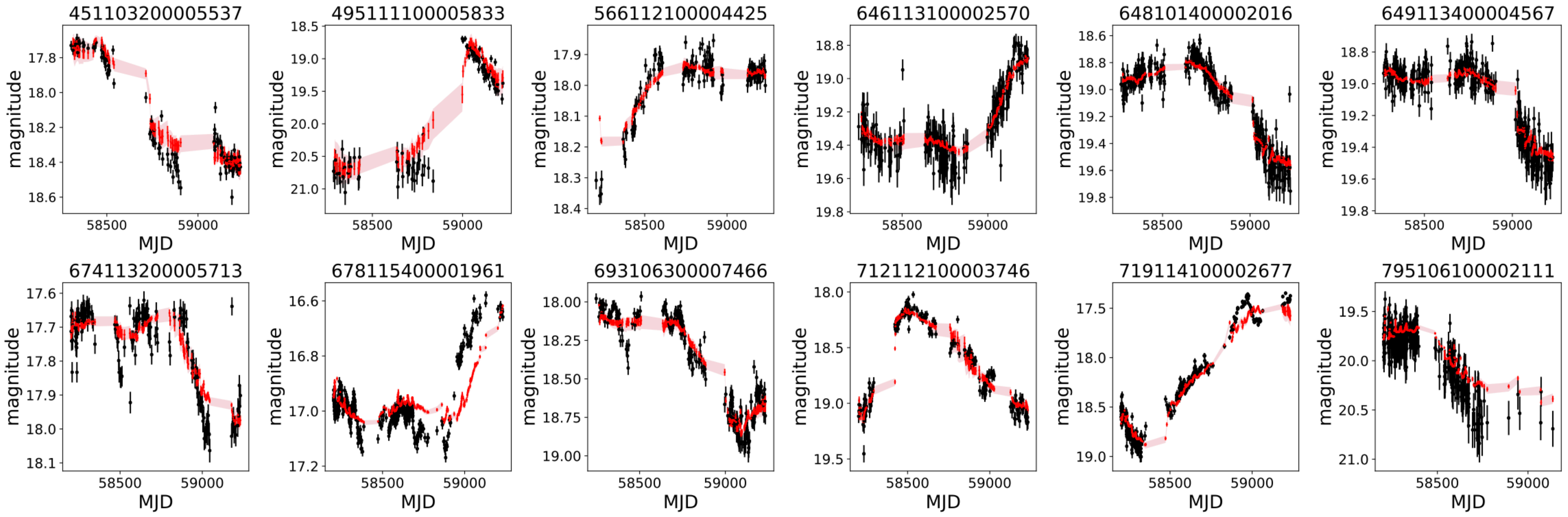
Sánchez-Sáez et al. 2021, AJ, 162, 206

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

- Detection of CSAGN events in massive data sets is crucial to understand these events and to improve our knowledge of the physical mechanisms behind AGN variability. Anomaly detection is a promising technique to search for CSAGN candidates.

- When using the original RAE architecture from Tachibana et al. 2020 we obtained a latent space that was highly correlated with the number of epochs per light curve.

- We used a Variational Recurrent Autoencoder (VRAE) architecture to model 230,451 AGN light curves from the ZTF DR5. 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.

- Anomaly detection cannot be blindly used to search for CSAGN candidates. Interaction between AD algorithms with domain experts is needed in order to find reliable samples of candidates

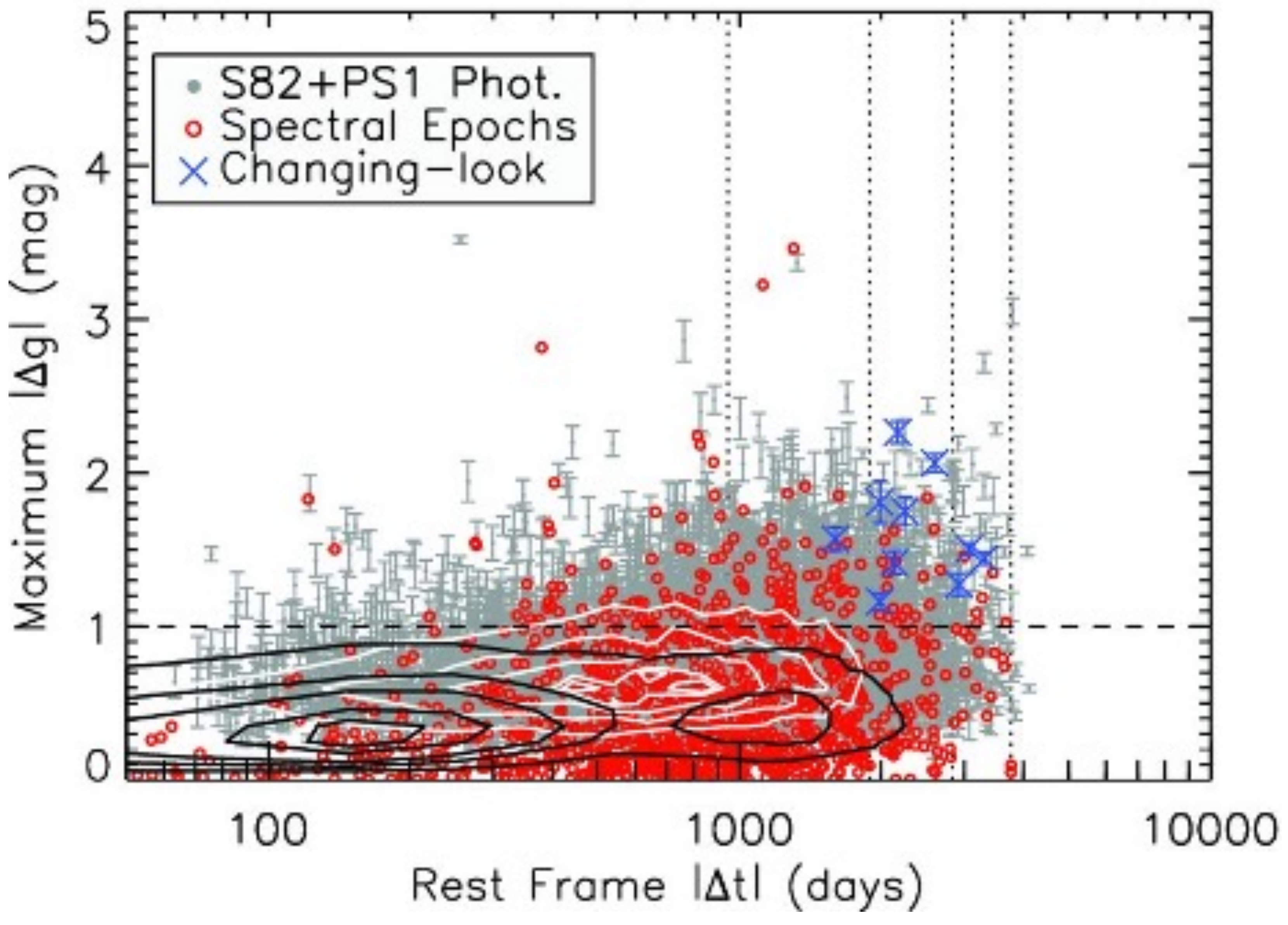
pasanchezsaez@gmail.com

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

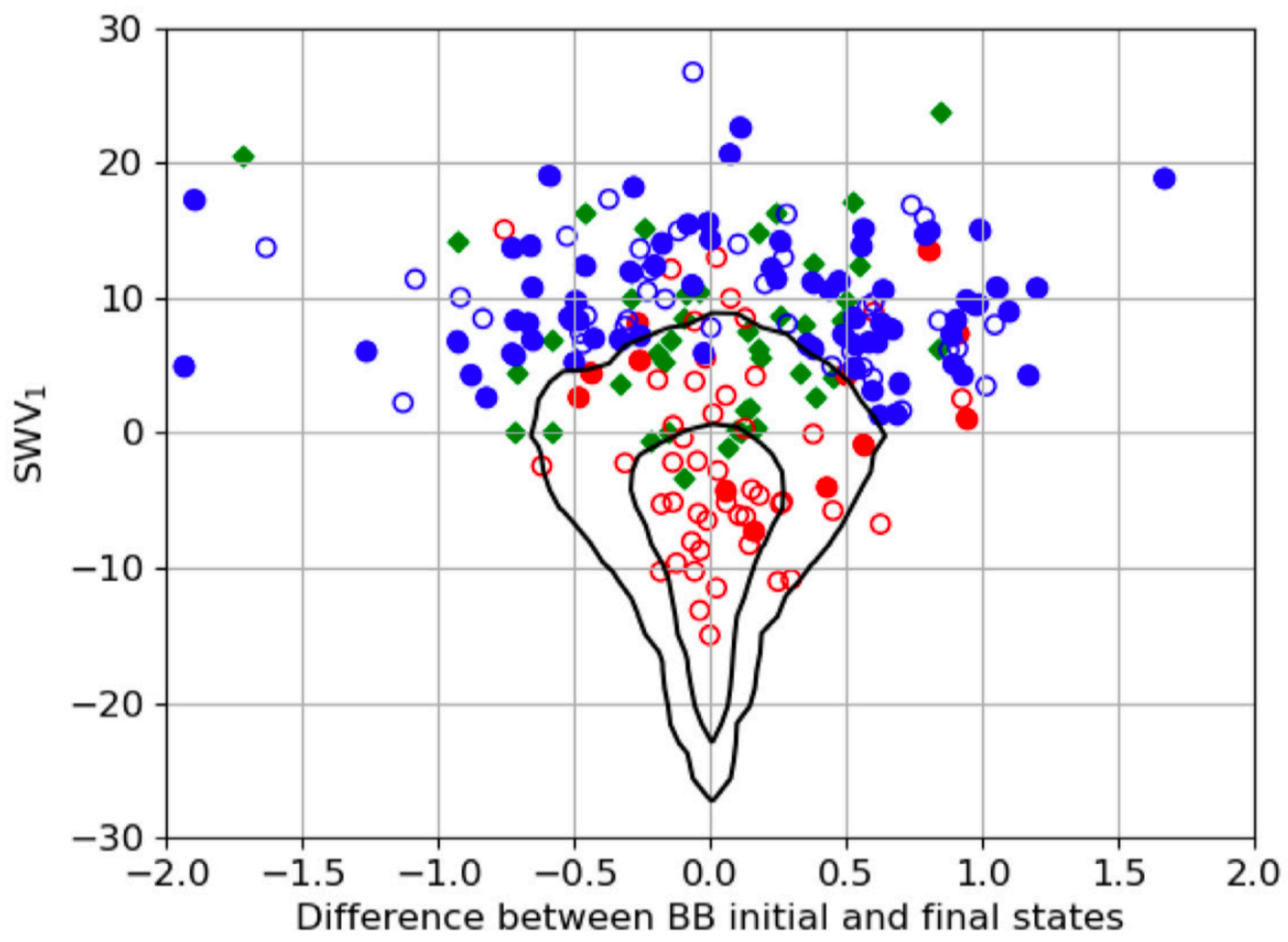
Buck up

Previous detections of CSAGN candidates

MacLeod+2016: candidates selected from SDSS and PS1 photometry. Criteria: $|\Delta g|>1$

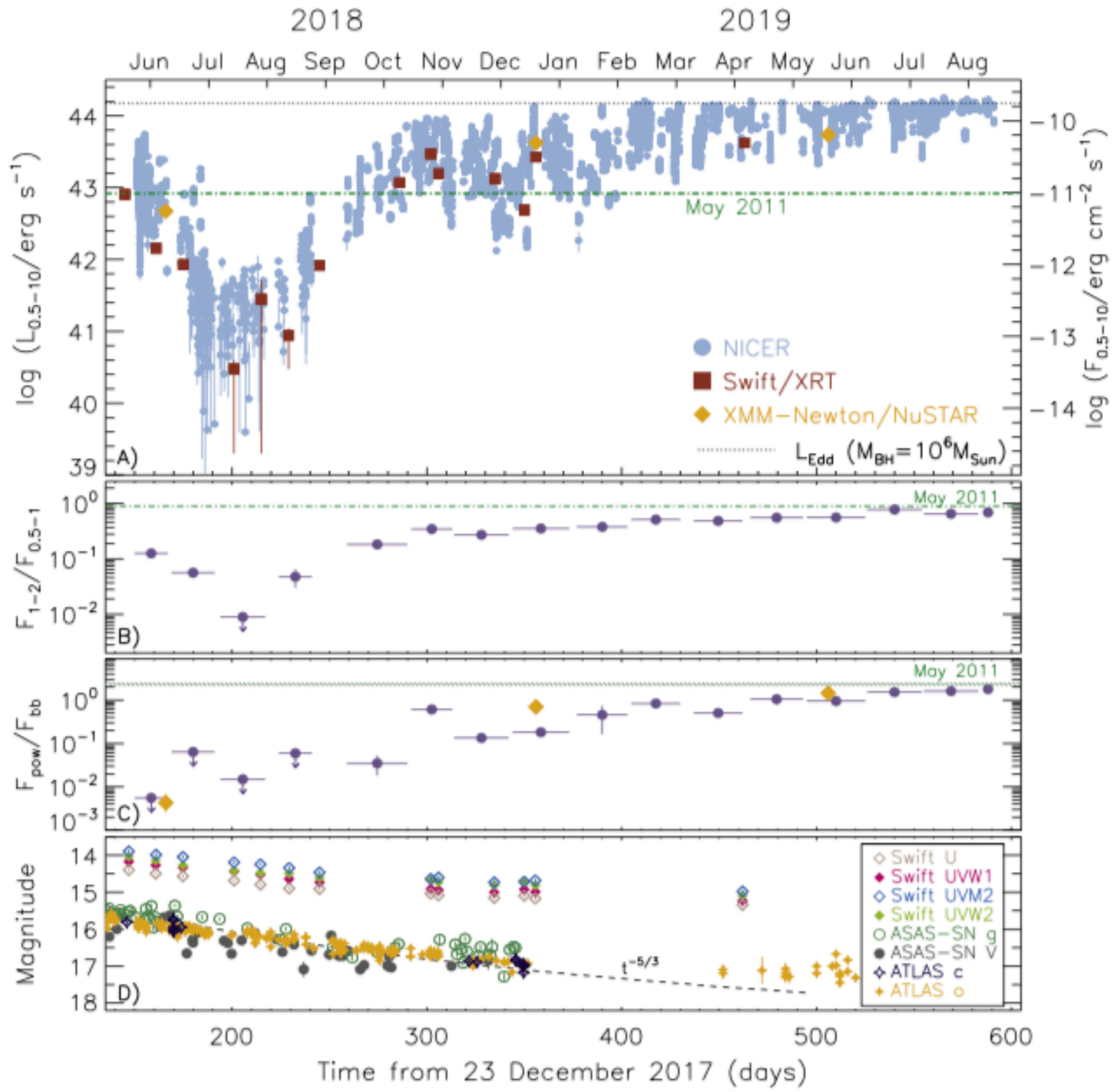


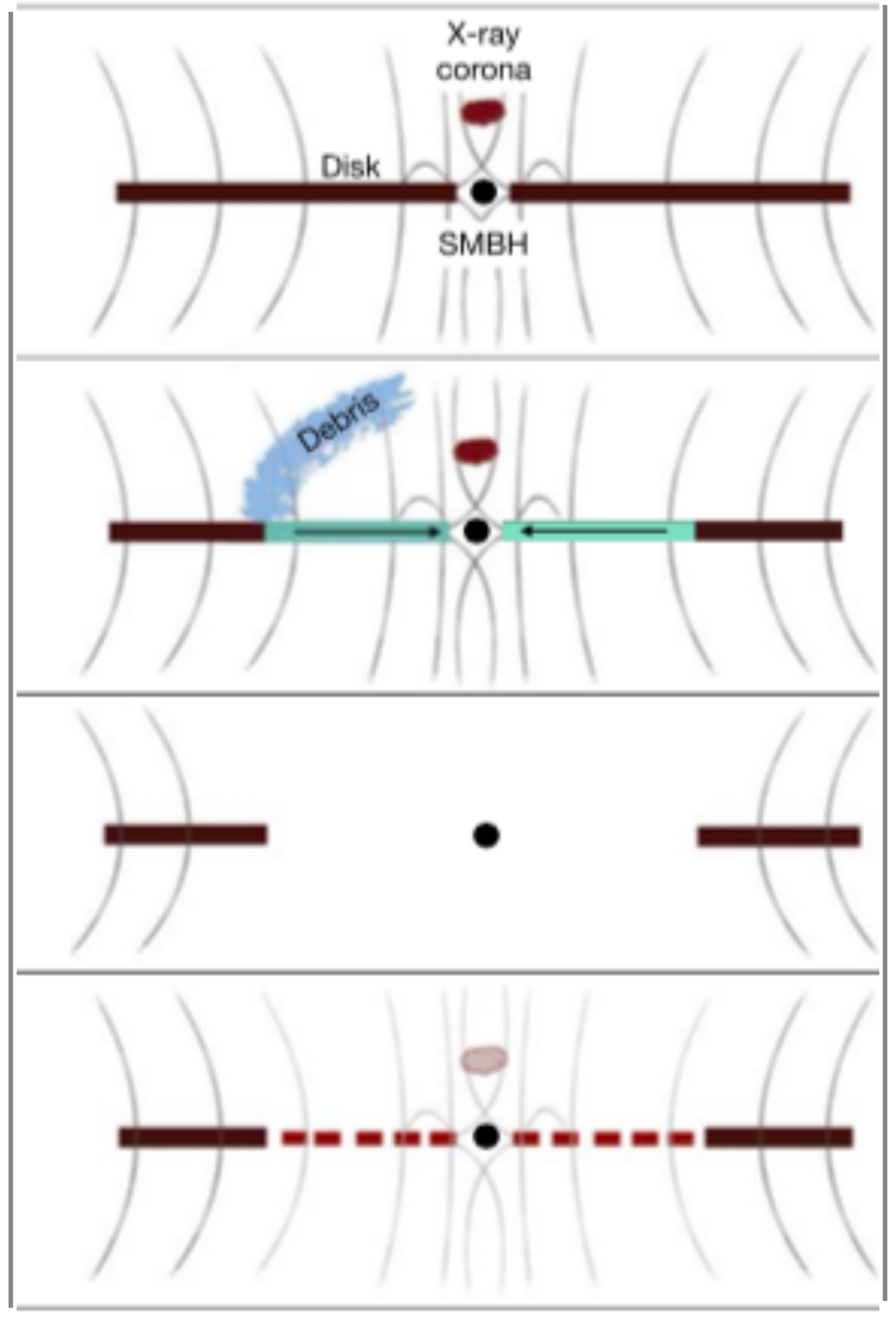
Graham+2020: candidates selected from CRTS photometry. Using Bayesian blocks (BB) representation and Slepian wavelet variance.



Understanding the physics behind the CSAGN phenomenon

Real-time detection of CSAGNs is crucial!

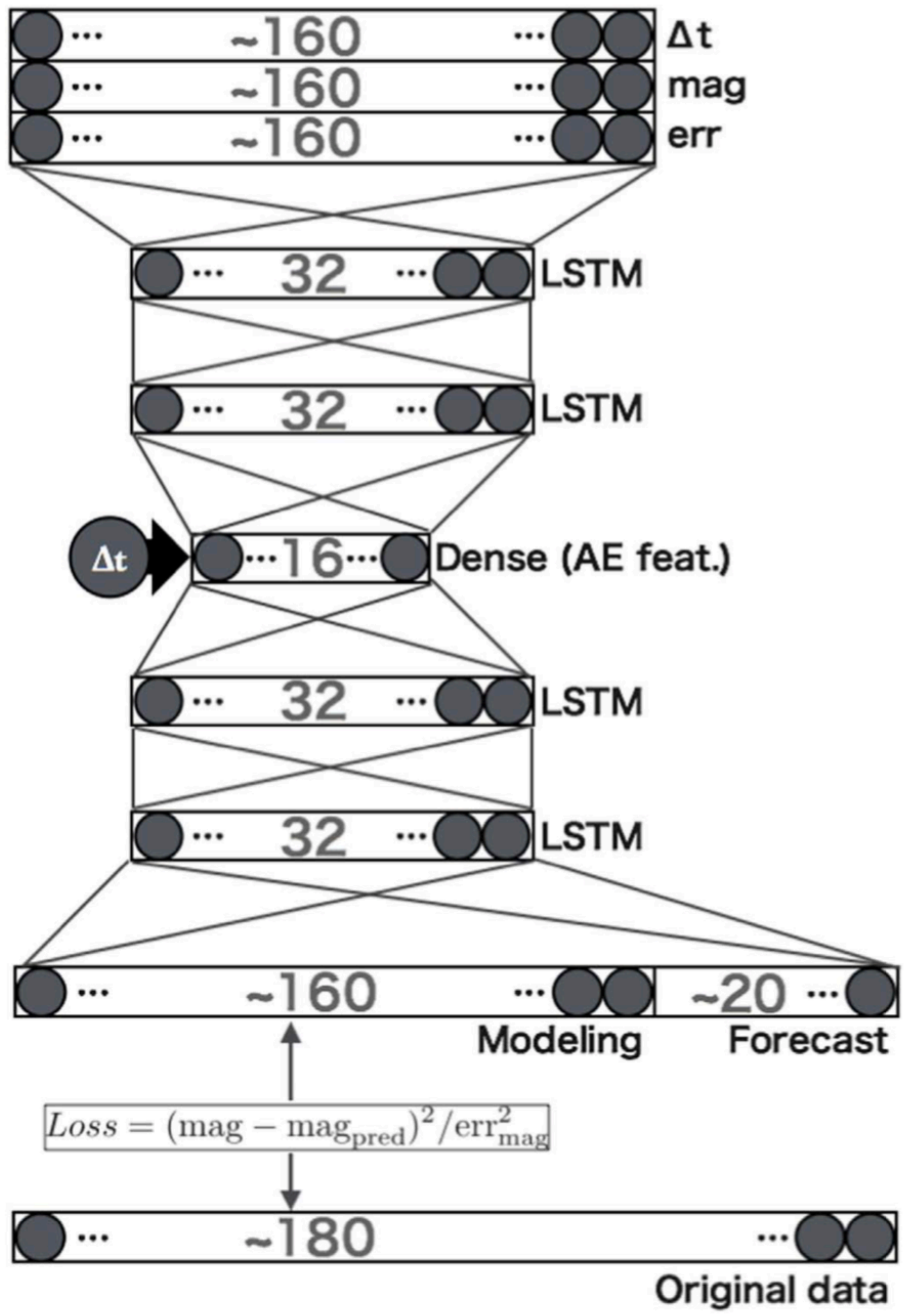


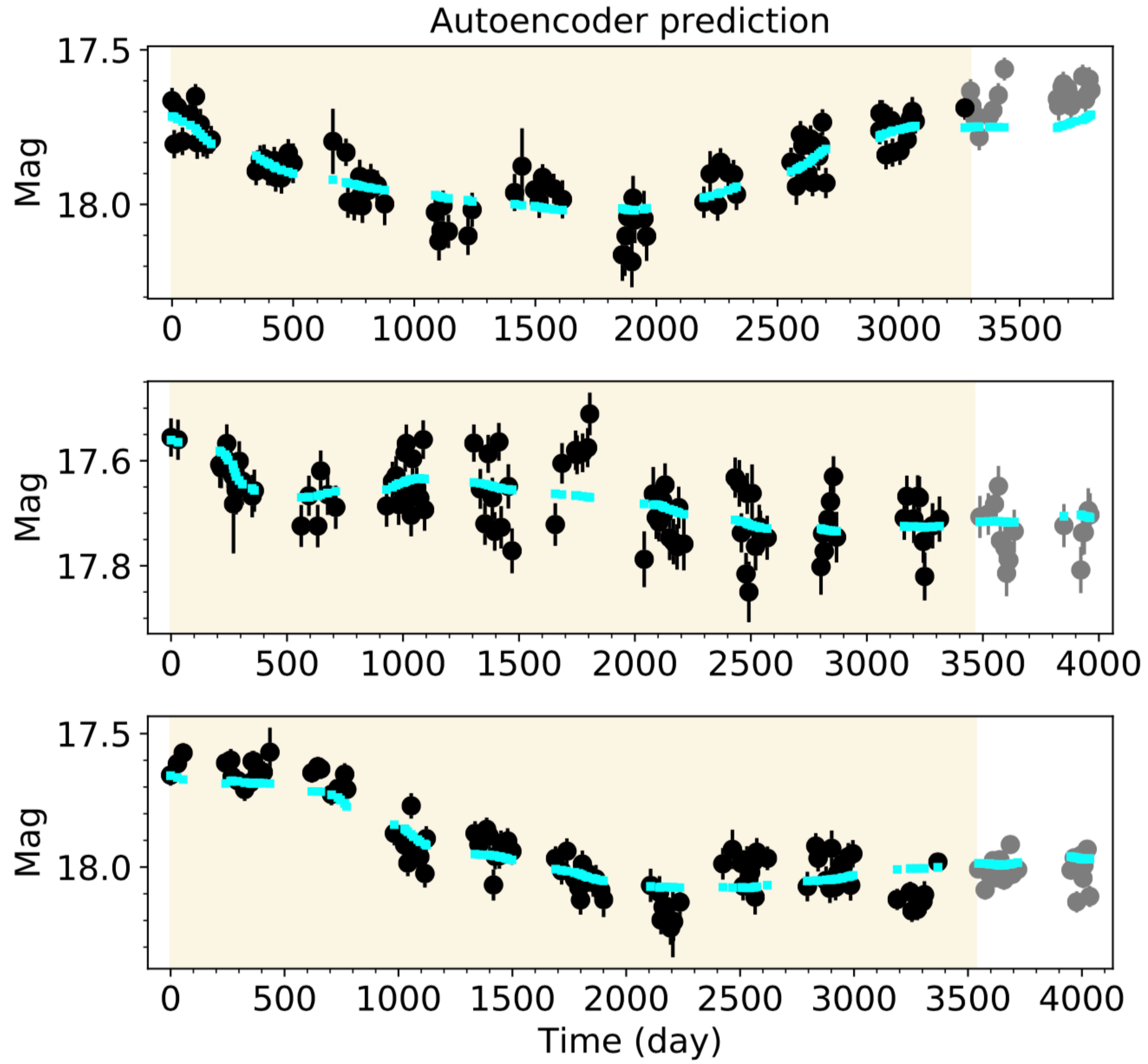


First CSAGN detected in real time. Probably related with a TDE. Ricci et al. 2020. See also Trakhtenbrot et al. 2019

Recurrent AEs to model AGN variability

Tachibana+2020 presented a novel algorithm to empirically model AGN light curves from CRTS using a Recurrent Autoencoder (RAE).

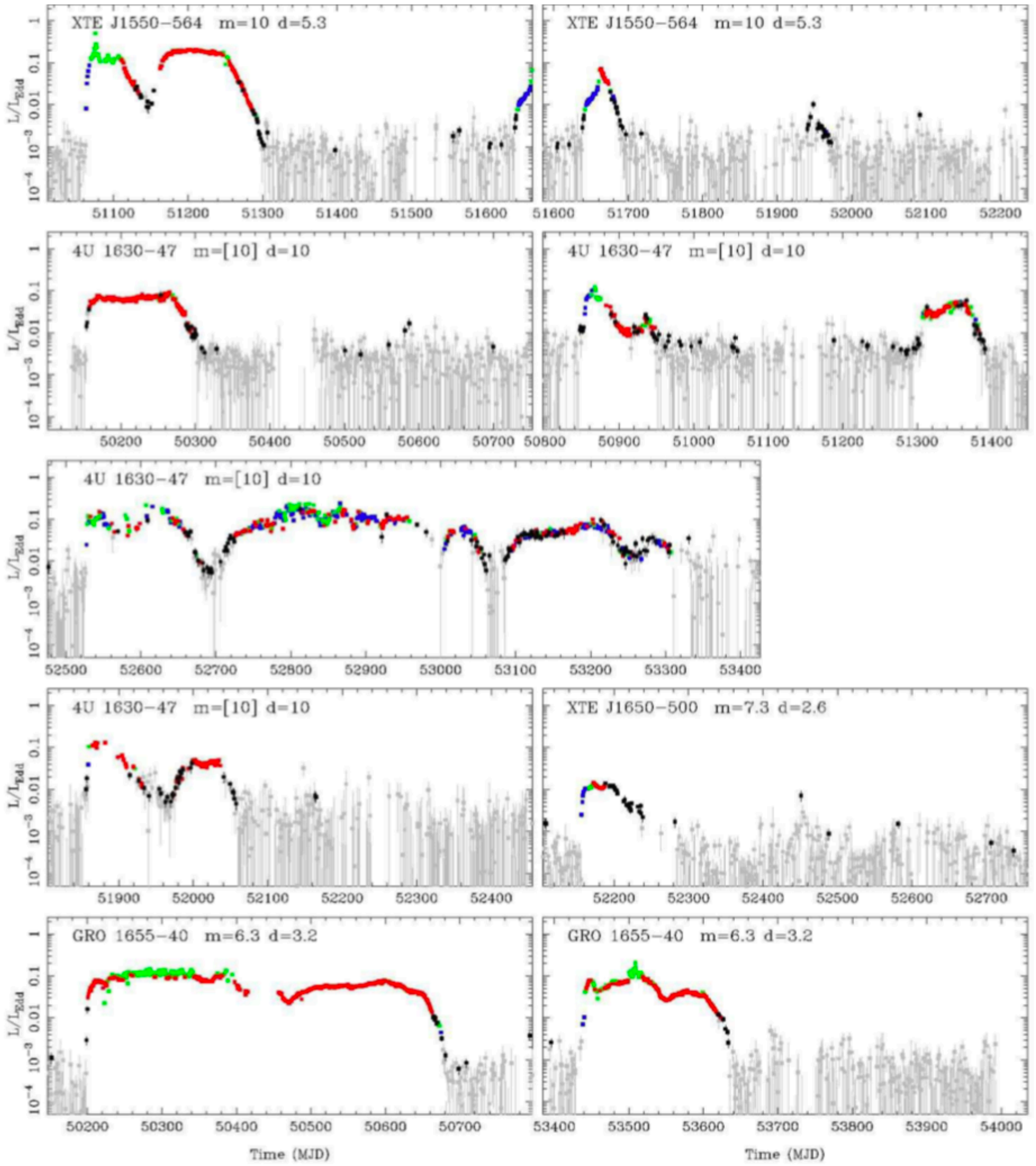




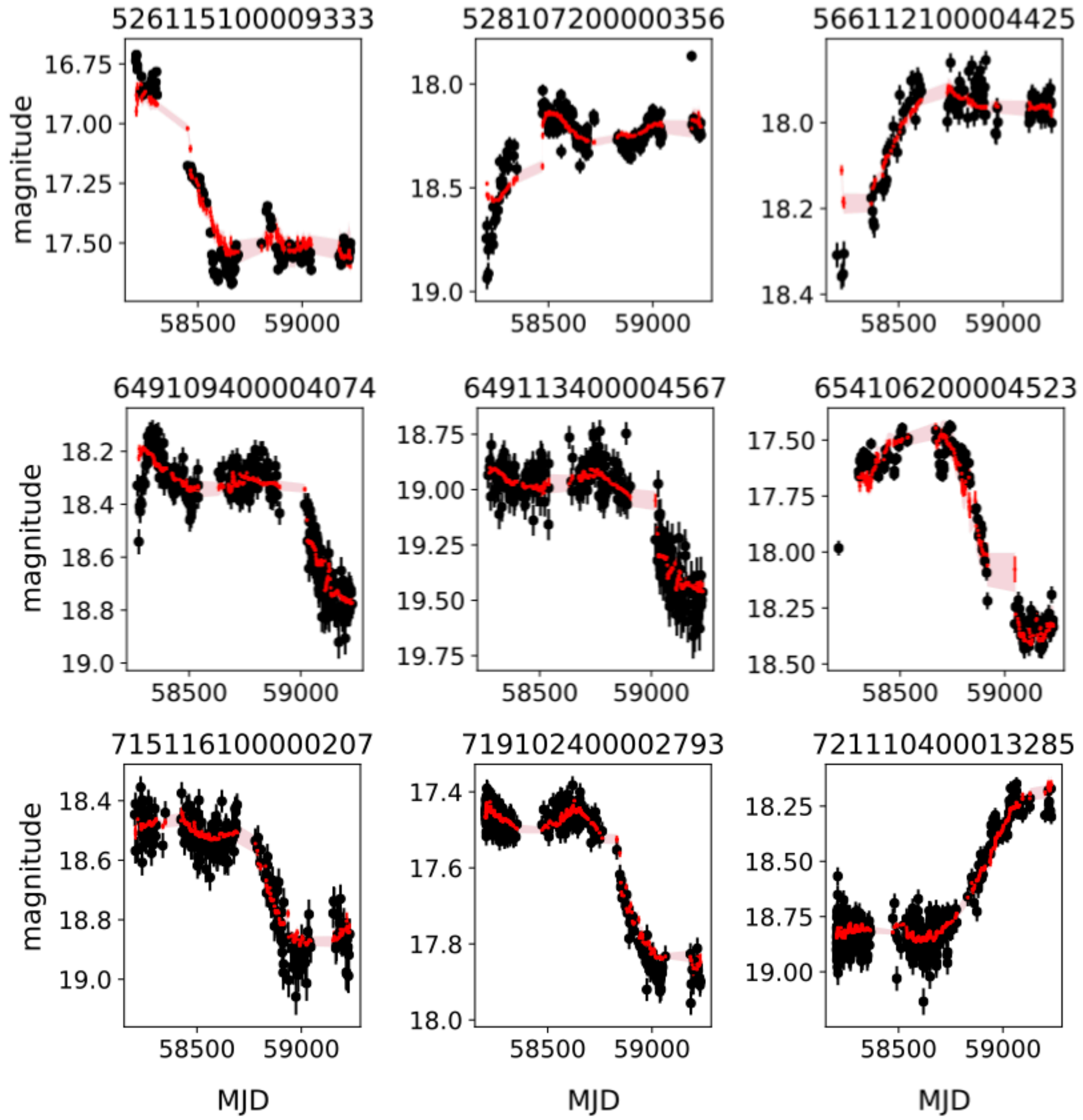
CSAGN candidates: the plateau feature

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

72% of the CSAGN candidates show a fast rise or decay and a "plateau" that spans hundreds of days. These light curve shapes are reminiscent of those previously observed in X-ray binaries at X-ray energies, although with much shorter timescales than expected if we extrapolate X-ray behaviors to AGNs (~ thousands of years).



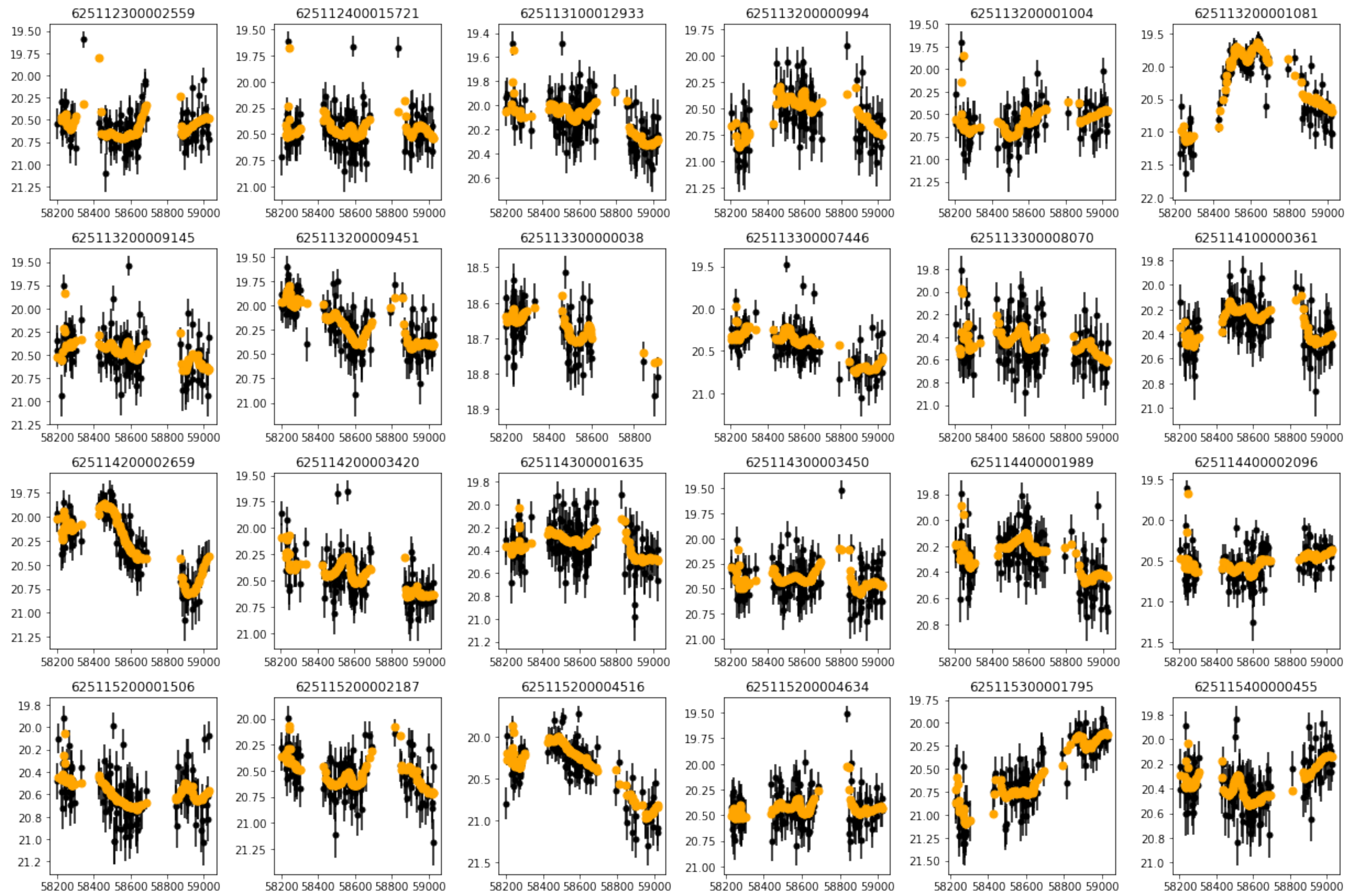
X-ray binaries light curves: Done+2007

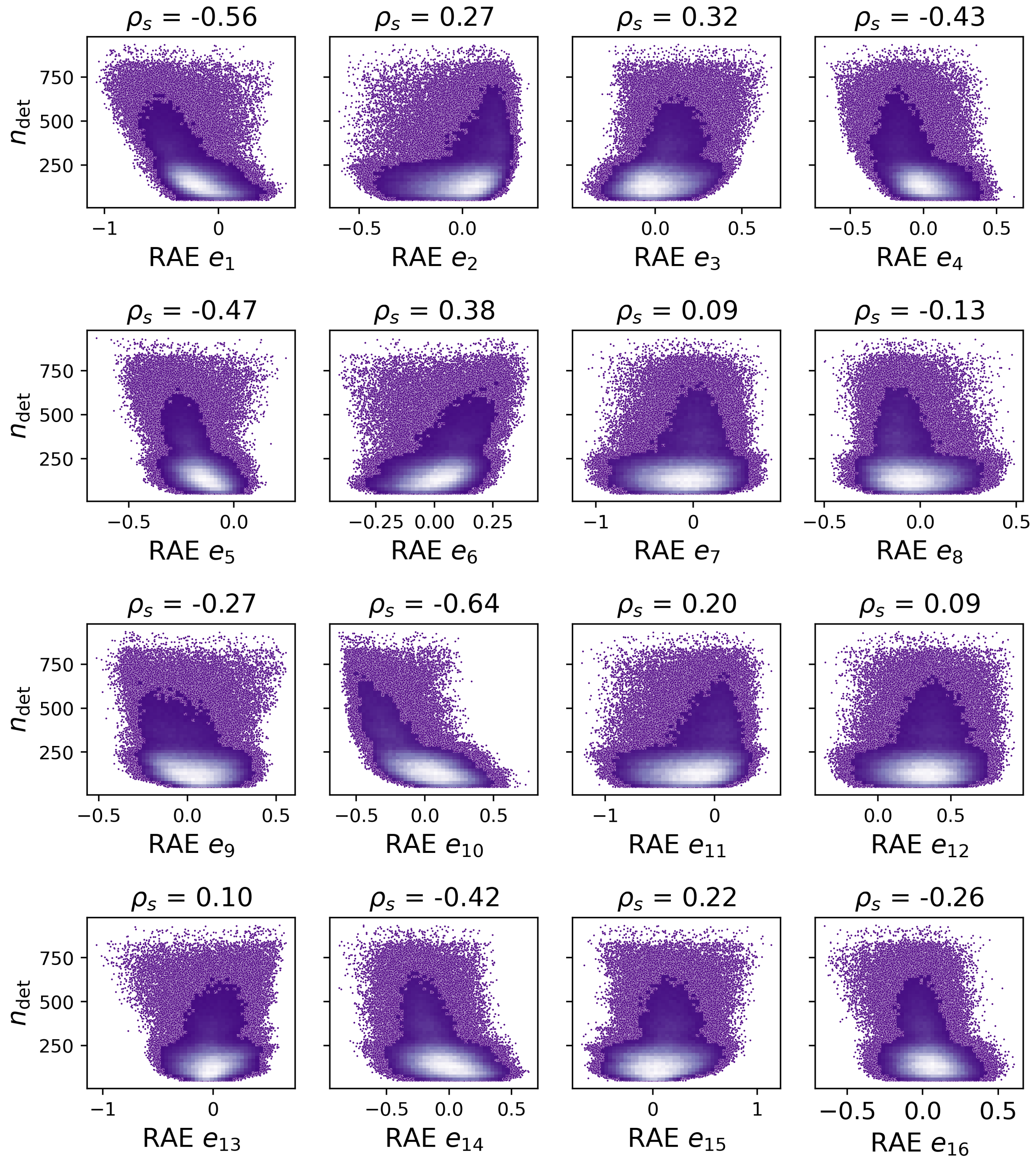


AEs for AGN variability anomaly detection

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

We used the Tachibana+2020 architecture to model 230,451 AGN light curves from ZTF DR5. We found high correlations between the latent space attributes with the number of epochs per light curve, although with low reconstruction errors.

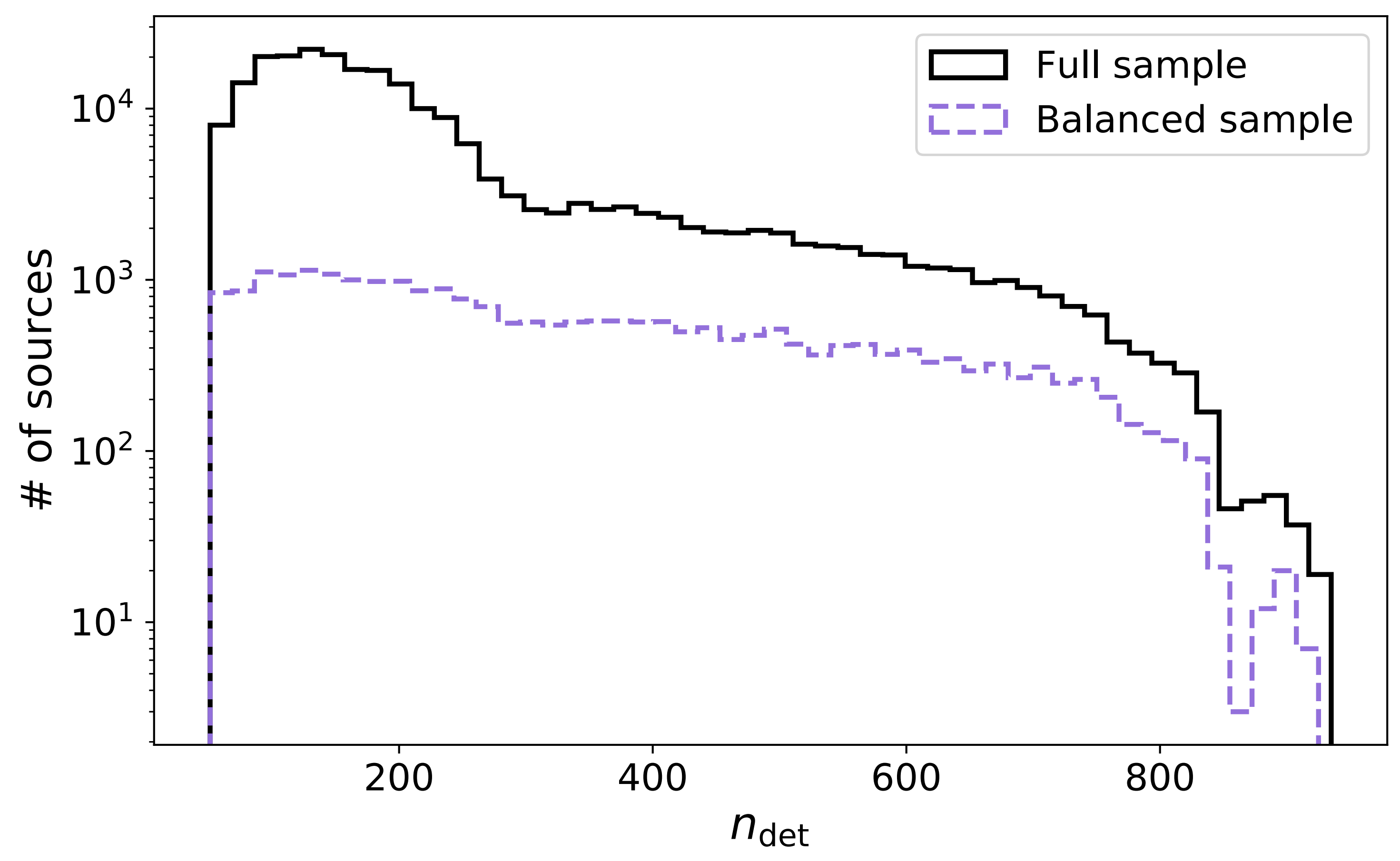


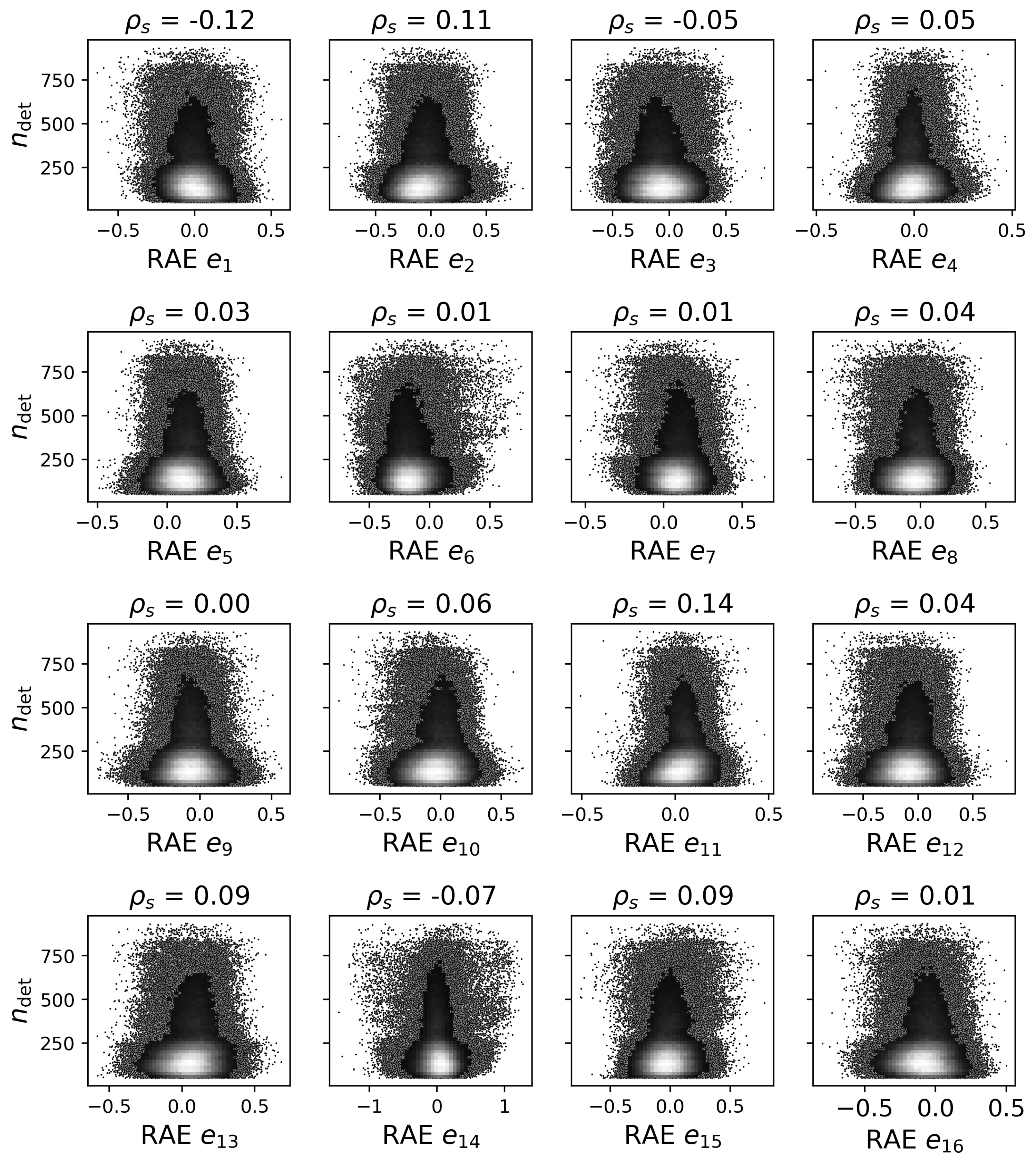


VAEs to model AGN light curves

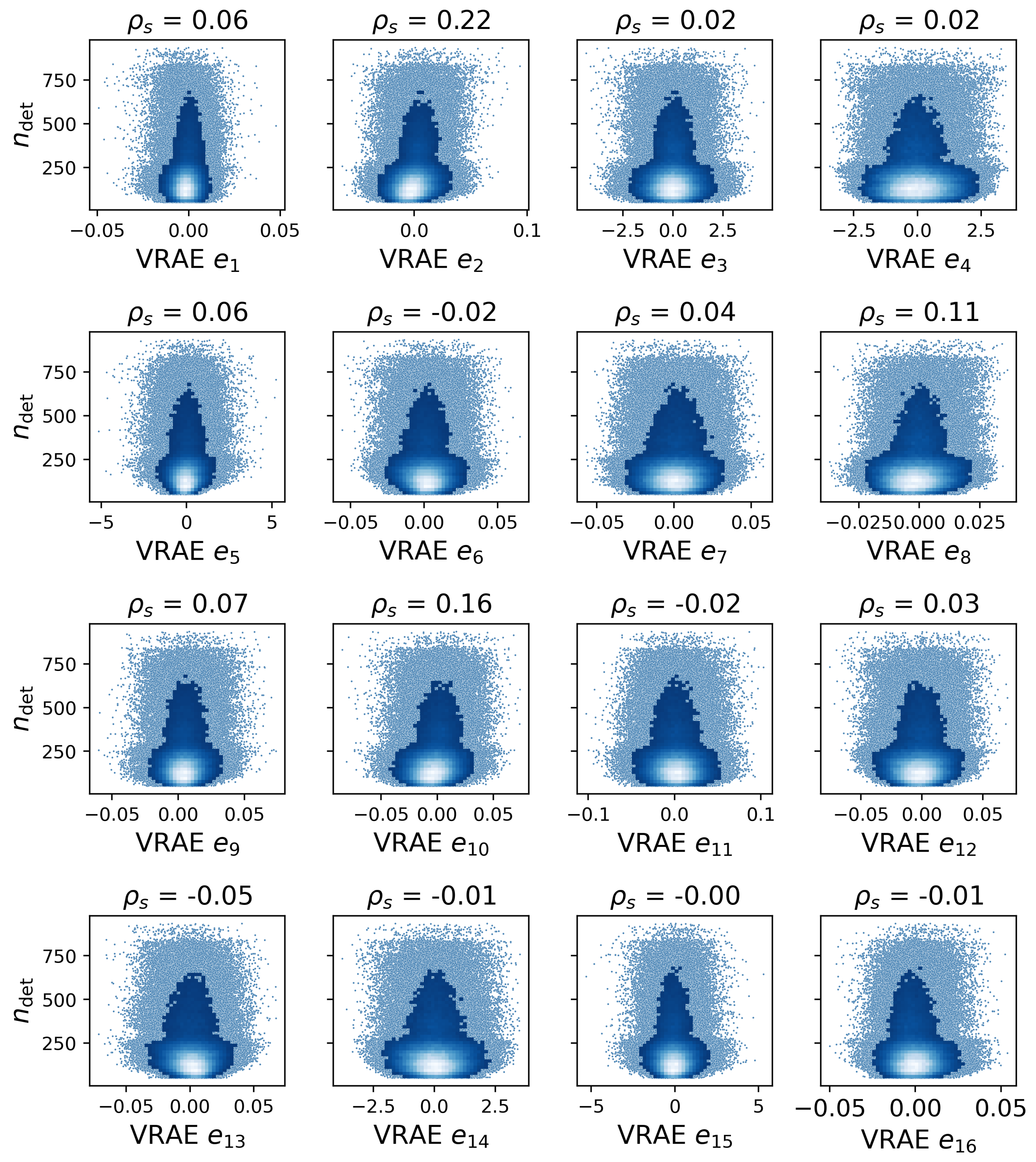
Sánchez-Sáez et al. 2021, AJ, 162, 206

To reduce even more the correlations we tested training with a dataset balanced also by means of the number of epochs per light curve. This improved the results of both VRAE and RAE architectures.





RAE architecture



VRAE architecture