SNAD: anomaly detection for large scale time-domain astronomy

Patrick Aleo (UIUC), Emille Ishida (CNRS-LPC), Matwey Kornilov (SAI MSU), Vladimir Korolev, Konstantin Malanchev (UIUC), Maria Pruzhinskaya (SAI MSU), Etienne Russeil (CNRS-LPC), Sreevarsha Sreejith (BNL), Alina Volnova (IKI RAS)

on the behalf of the SNAD team
+ all our side collaborators

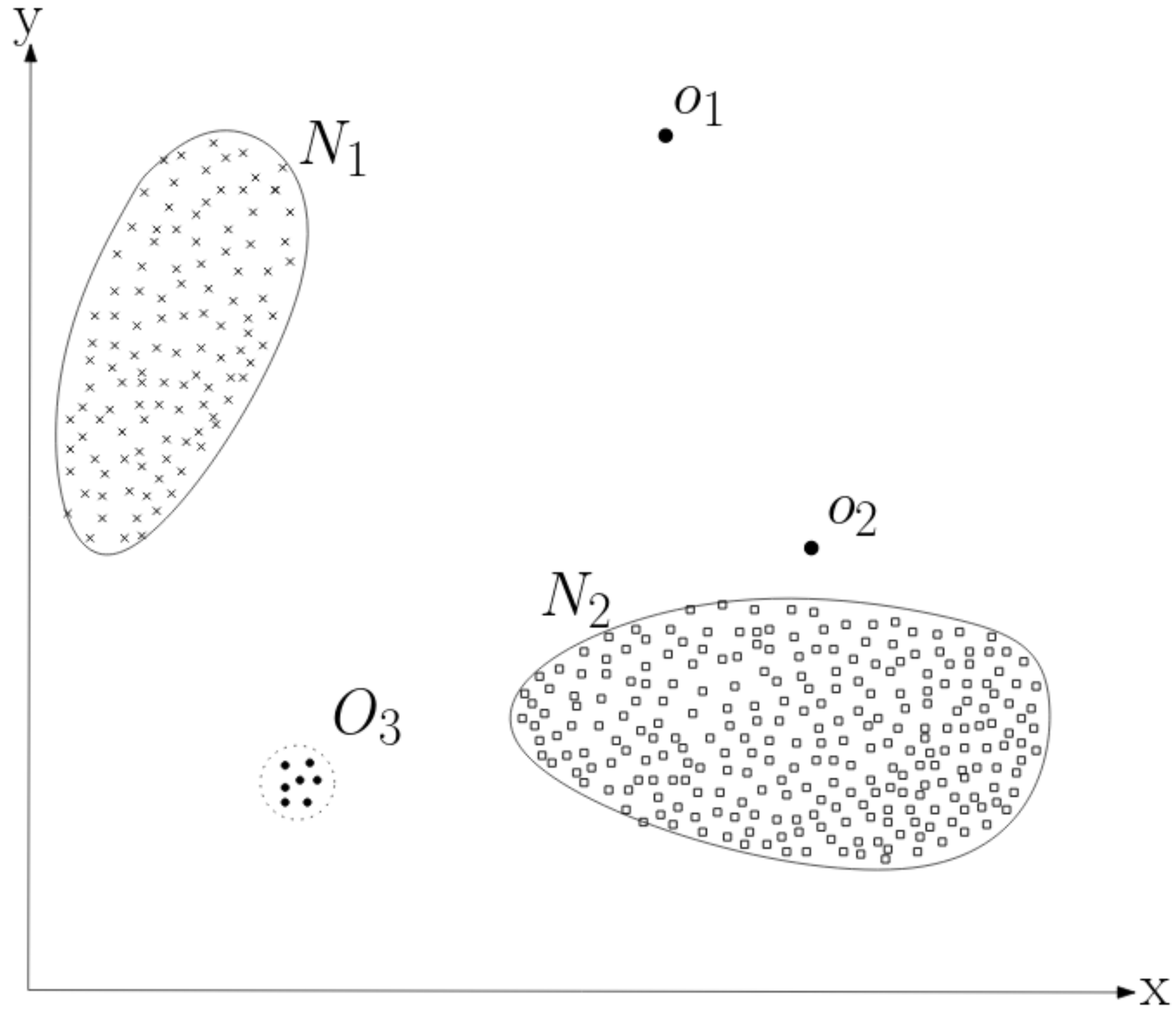
Paris, 2022.06.21

Anomaly detection

We look for anomalies

- Def. *Outlier* is an object located in a sparse region of the feature space

- Def. *Anomaly* is an astrophysical source having unusual properties for its class or a representative of some rare class



Chandola+ 2009

Discovery

ML only produces recommendations

Light curves

Preprocessing

Machine learning
Outlier detector

Potentially interesting anomalies:

- Candidate 1
- Candidate 2
- Candidate 3
- …
- …
- …
- …

metadata
images
simulations
catalogs

Not interesting

Interesting

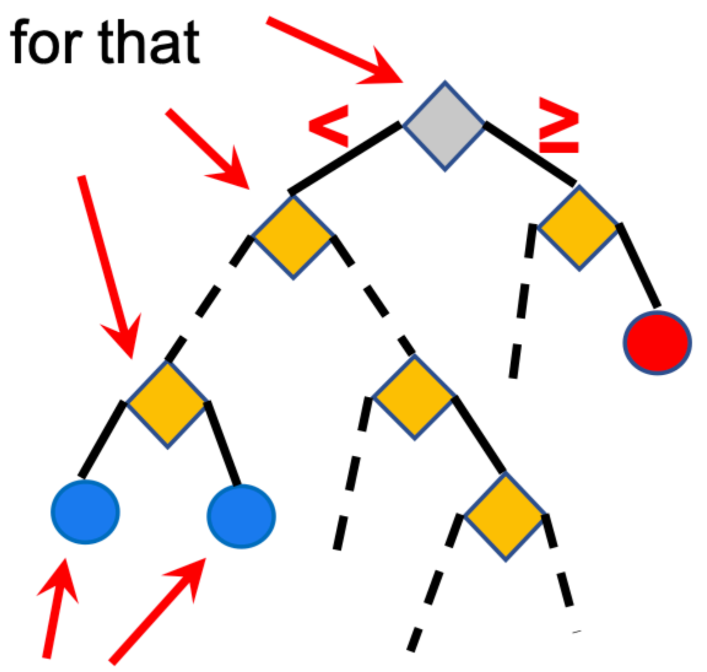
Very interesting!

Get more data and Publication

Outlier detection: Isolation Forest

iTree

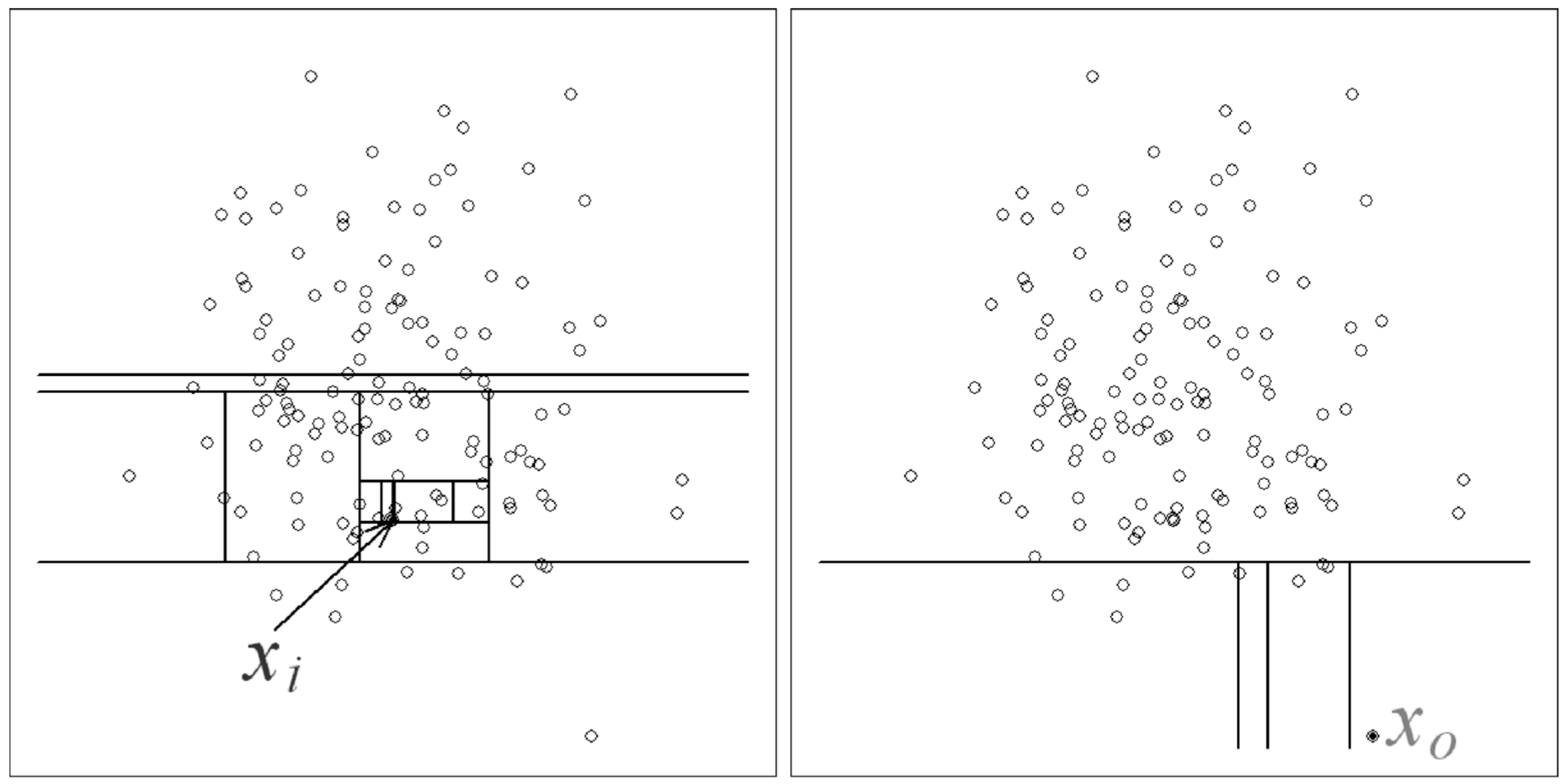
Select a random feature at each node, and a random split point for that feature



Shallower leaf nodes have higher anomaly scores, whereas, deeper leaf nodes have lower anomaly scores.

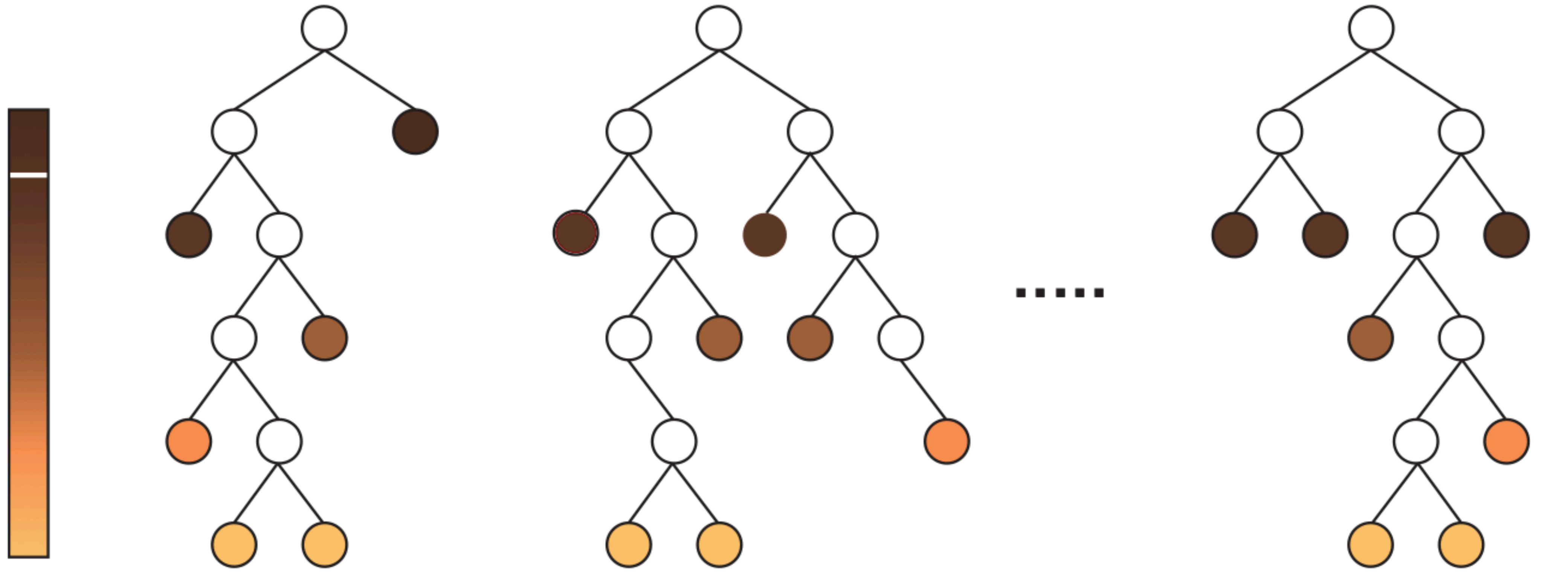
Leaf instance

arXiv:1708.0944

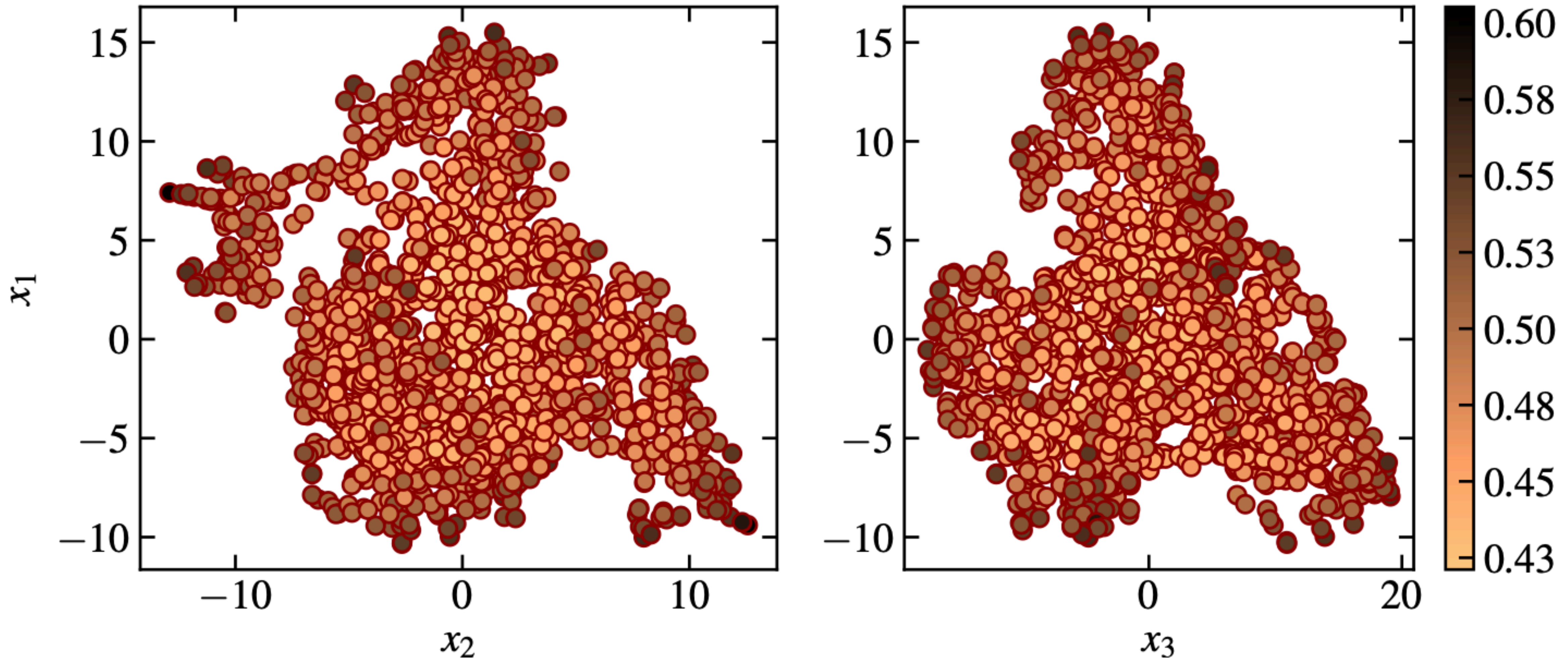


(a) Isolating x_i (b) Isolating x_o

Liu+ 2008, Liu+ 2012

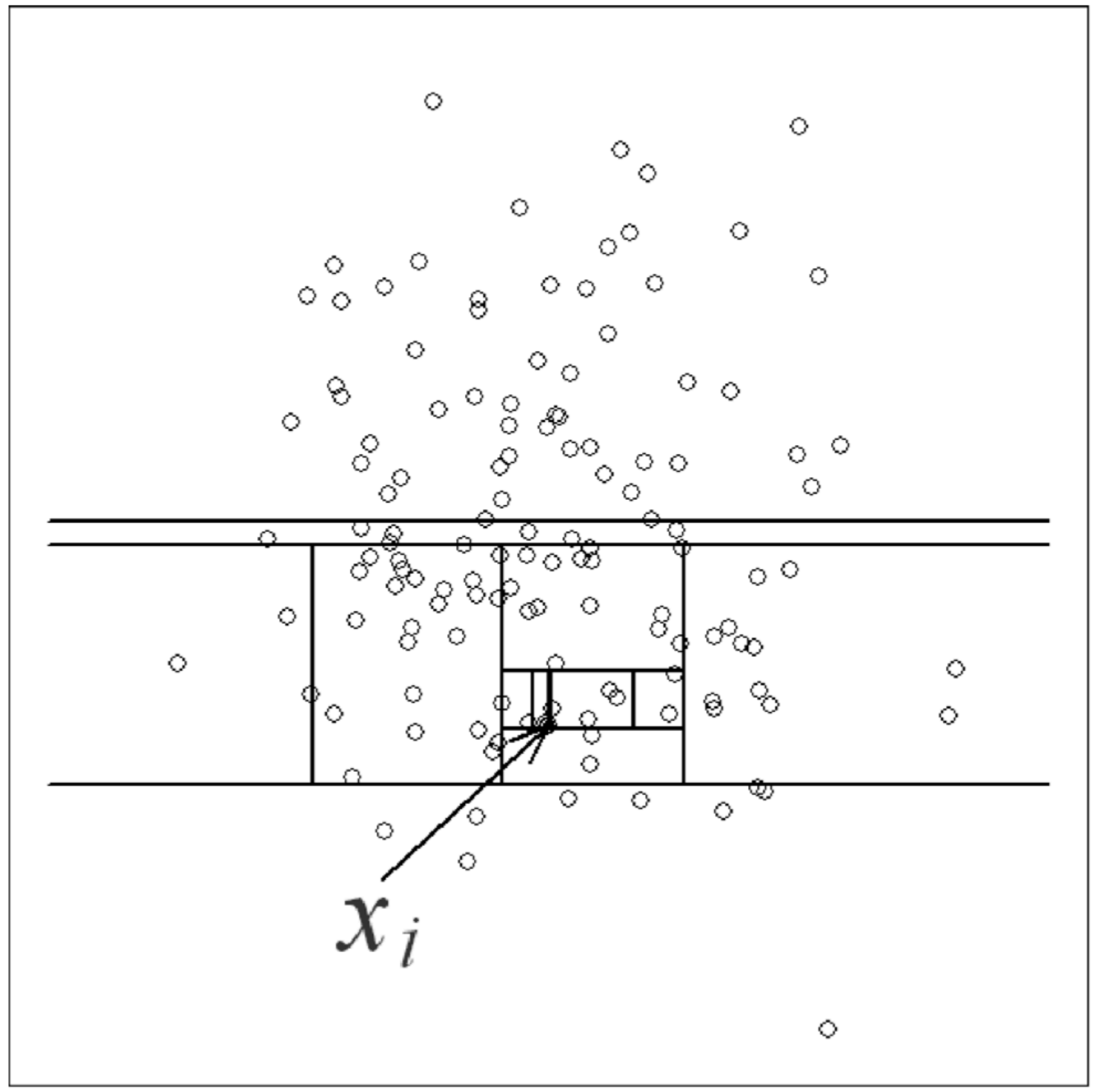


Darker is more anomalous

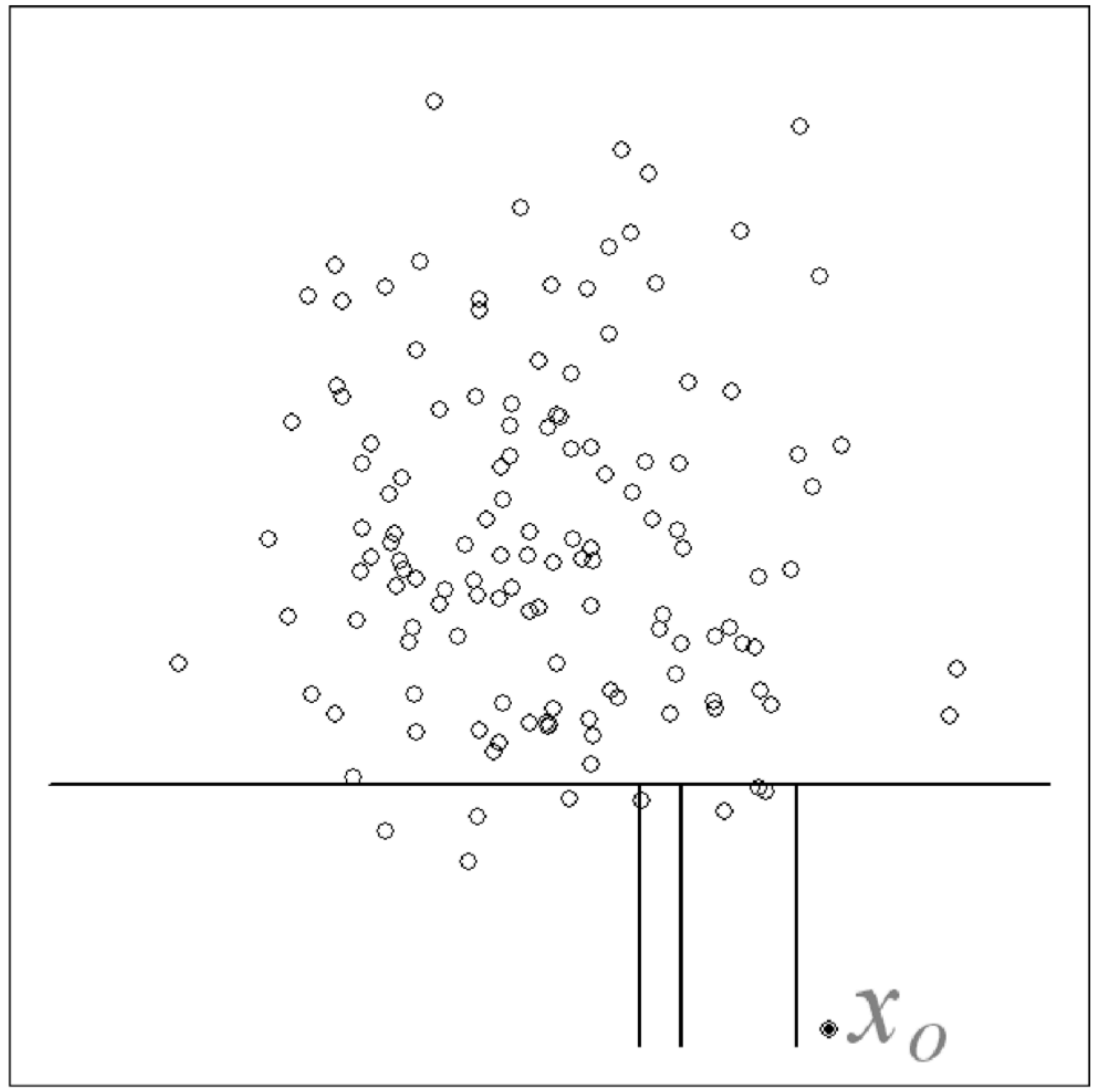


arXiv:1905.11516

Isolation Tree



(a) Isolating x_i



(b) Isolating x_o

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/\psi & \text{for } \psi > 2, \\ 1 & \text{for } \psi = 2, \\ 0 & \text{otherwise,} \end{cases}$$

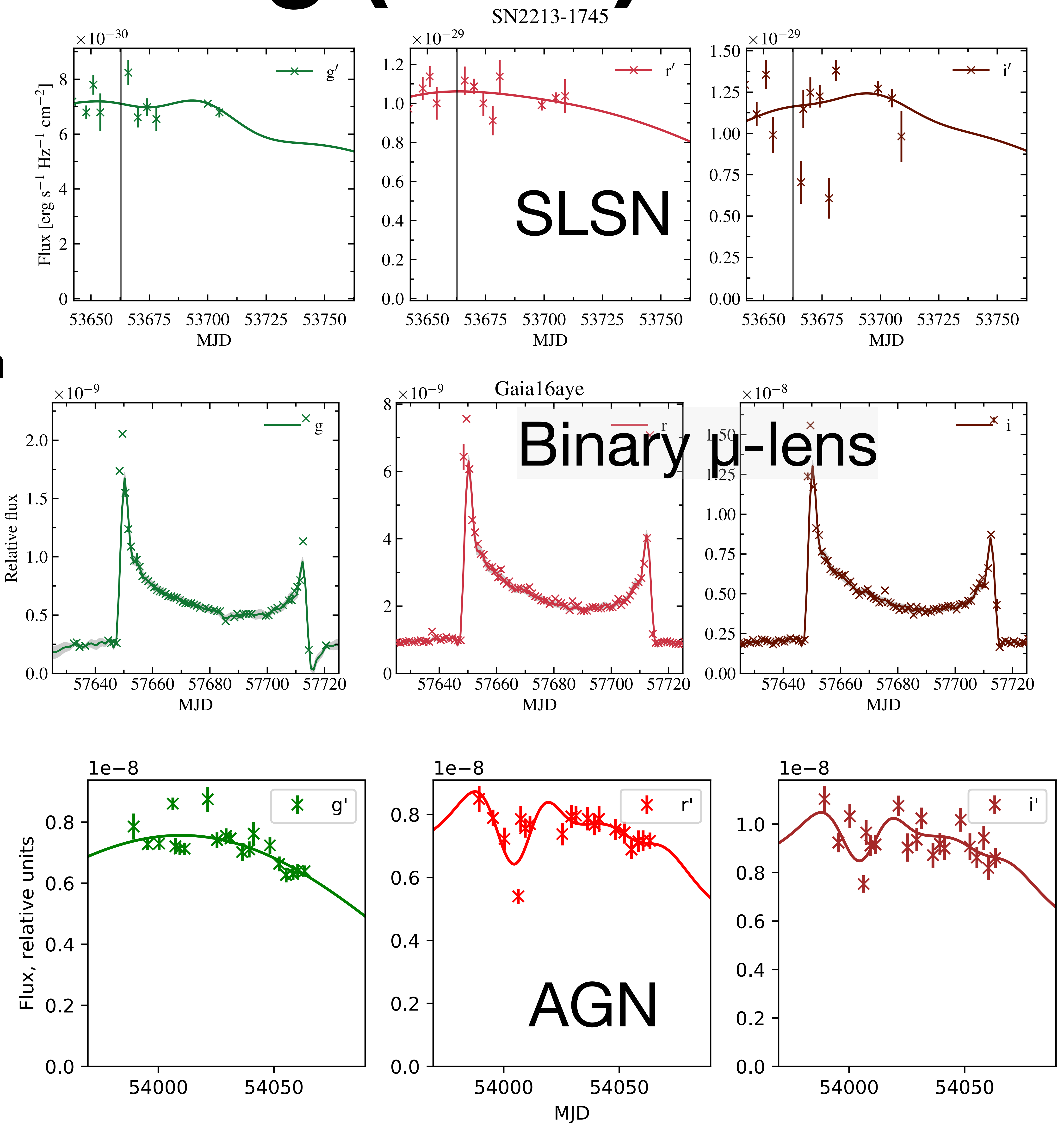
$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}},$$

Liu et al 2008, Liu et al 2012

Case: Open Supernova Catalog (OSC)

arXiv:1905.11516

- 1999 SNe in *gri*, *g'r'i'* & *BRI* taken from the OSC (Guillochon+ 2017)
- Multivariate Gaussian process approximation (Semenikhin+ in prep.) & t-SNE
- 30/100 anomaly candidates
  - Two known SLSNe
  - Several known peculiar SNe
  - Several known cases of misclassification, including binary μ-lens
  - **16 previously unknown cases of misclassification** (10 stars and 6 AGNs), including SN 2006kg suggested as a "template" SN II (Okumara+ 2014)



Multivariable Gaussian processes

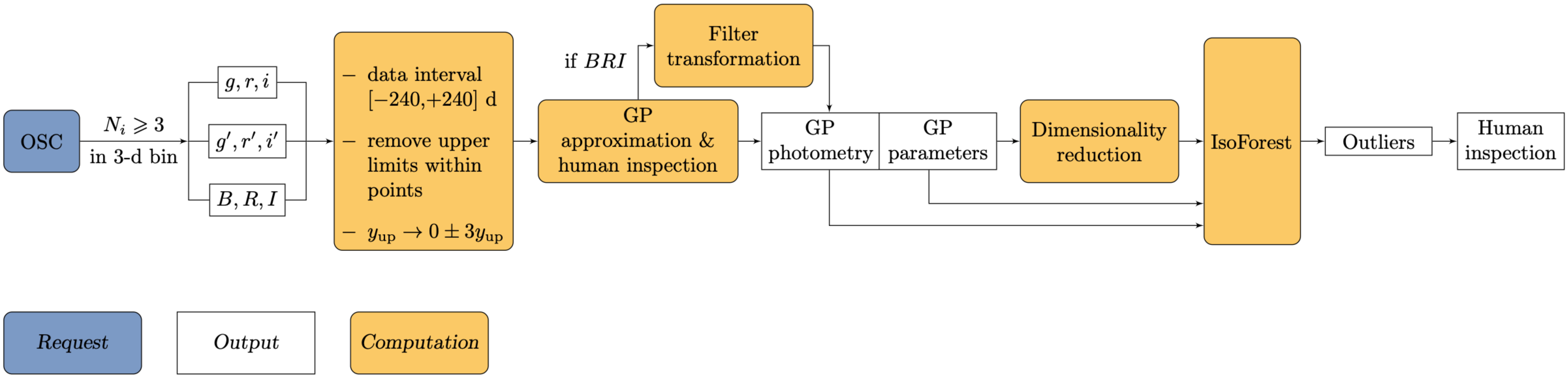
We implement it via correlation between components (passbands), not via 2D kernels

$$\boldsymbol{y}(t) = \mathrm{M}\left[\boldsymbol{\nu}(t)\right] \equiv \int \boldsymbol{\nu} p_{\boldsymbol{\nu}}(\boldsymbol{\nu}, t; \boldsymbol{\theta}) d\boldsymbol{\nu}$$

$$p_{\boldsymbol{\nu}}(\boldsymbol{\nu_1}, t_1, \boldsymbol{\nu_2}, t_2; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi \det |\Sigma|}^2} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}\boldsymbol{\nu_1}\\ \boldsymbol{\nu_2}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}, \Sigma^{-1}\left(\begin{bmatrix}\boldsymbol{\nu_1}\\ \boldsymbol{\nu_2}\end{bmatrix} - \begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}\right)\right)\right) \qquad \Sigma = \begin{pmatrix}\Sigma_d \Sigma_s\\ \Sigma_s \Sigma_d\end{pmatrix}$$

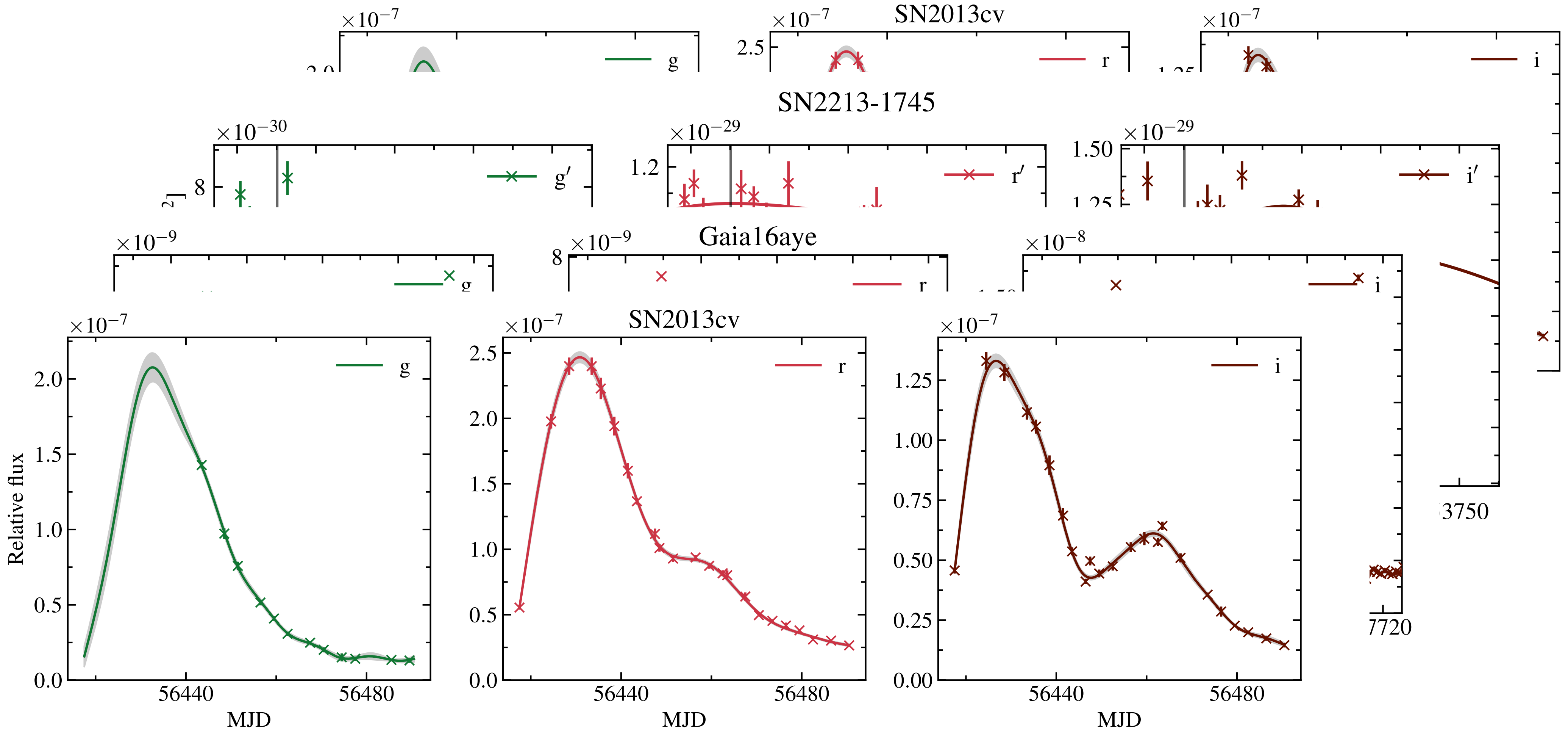
$$\Sigma = \begin{pmatrix}
\sigma_1^2 & 0 & \cdots & 0 & \sigma_1^2 K_1(t_1,t_2) & 0 & \cdots & 0\\
0 & \sigma_2^2 & \cdots & 0 & 0 & \sigma_2^2 K_2(t_1,t_2) & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & \cdots & \sigma_k^2 & 0 & 0 & \cdots & \sigma_k^2 K_k(t_1,t_2)\\
\sigma_1^2 K_1(t_1,t_2) & 0 & \cdots & 0 & \sigma_1^2 & 0 & \cdots & 0\\
0 & \sigma_2^2 K_2(t_1,t_2) & \cdots & 0 & 0 & \sigma_2^2 & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & \cdots & \sigma_k^2 K_k(t_1,t_2) & 0 & 0 & \cdots & \sigma_k^2
\end{pmatrix}$$

OCS anomaly detection pipeline



Three OCS feature sets

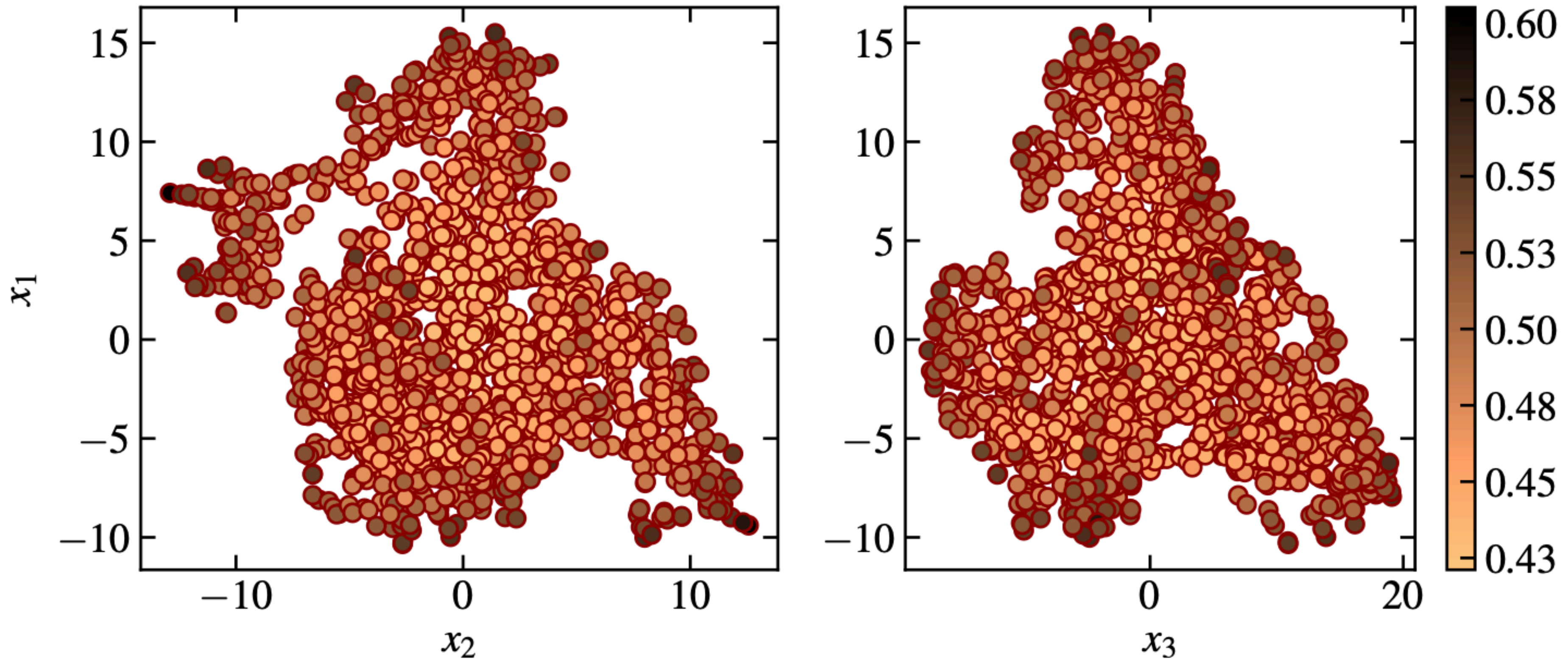
1. 364 Gaussian processes approximated points: 3 passbands ✖ 121 points $\in [-20; 100]$ days after peak in *r* normalized to peak, and peak flux itself



2. 10 parameters of Gaussian process fit: 6 values of correlation matrix, 3 lengths of kernels, likelihood

$$\{\log L, l_g, l_r, l_i, M_{gg}, M_{rr}, M_{ii}, M_{gr}, M_{gi}, M_{ri}\}$$

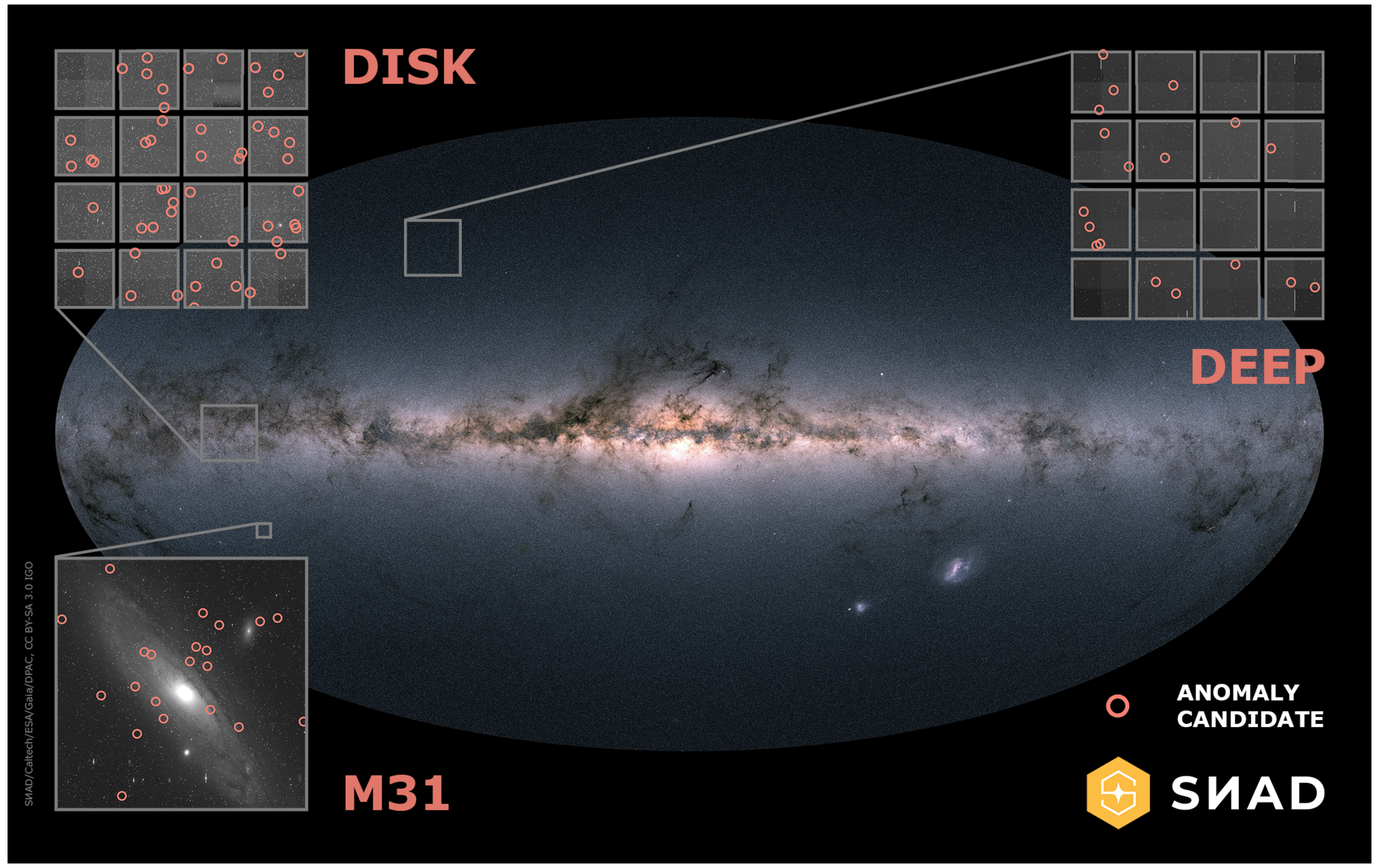
3. Eight datasets obtained by reducing 374 Gaussian process features to 2–9 t-SNE dimensions

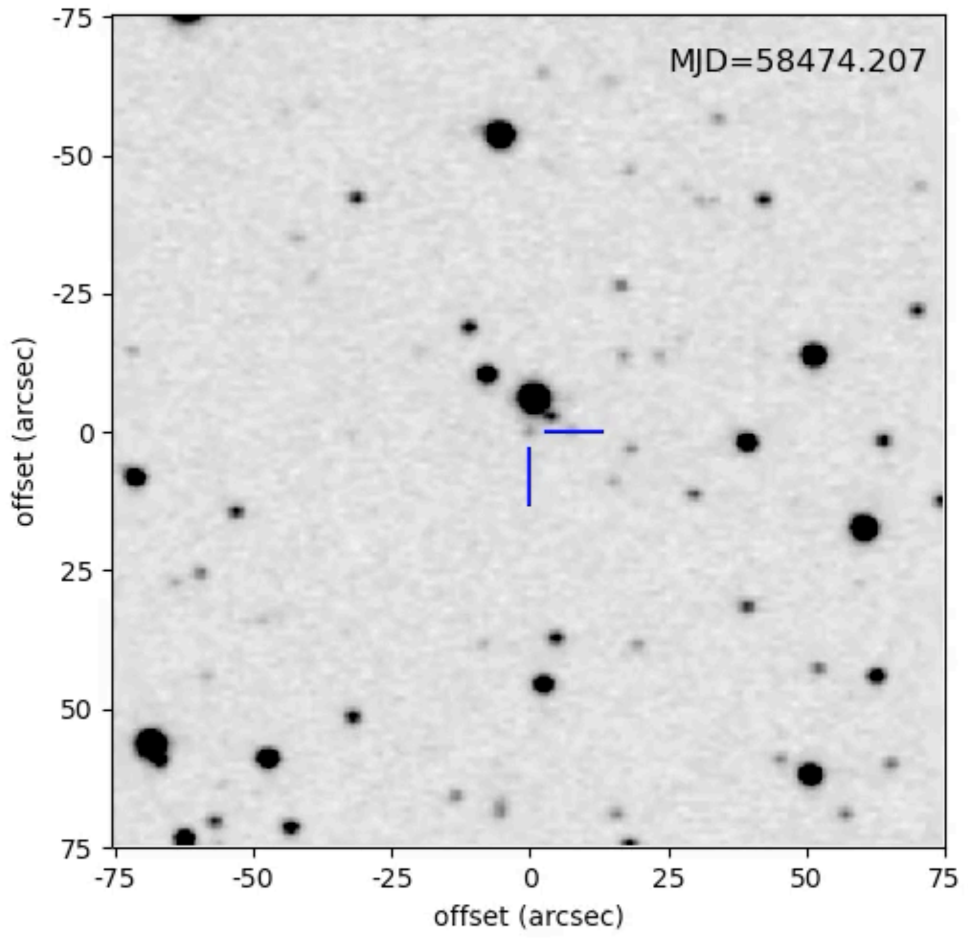


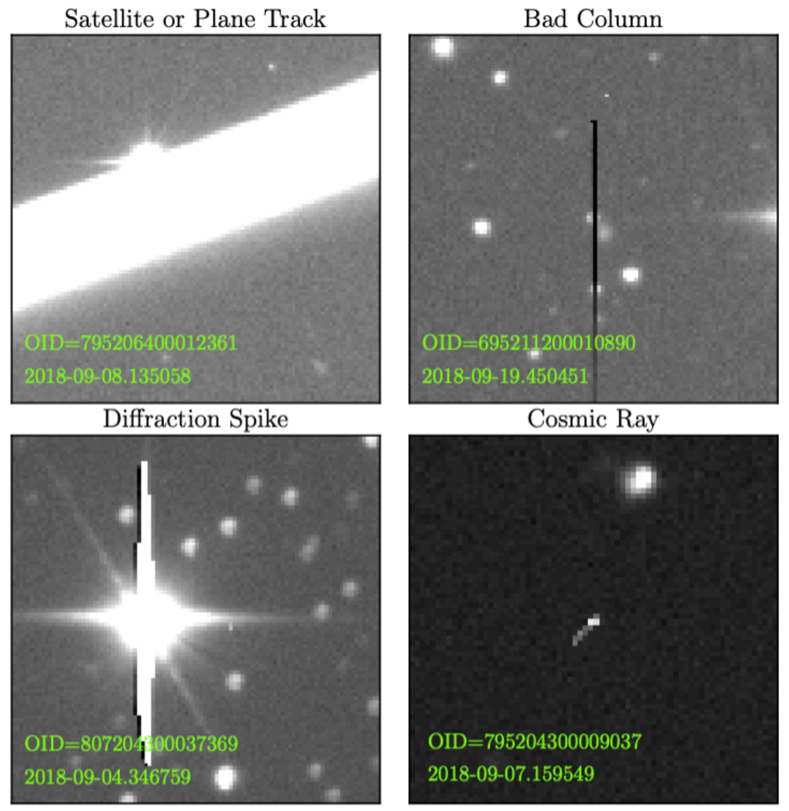
Case: Zwicky Transient Facility DR3

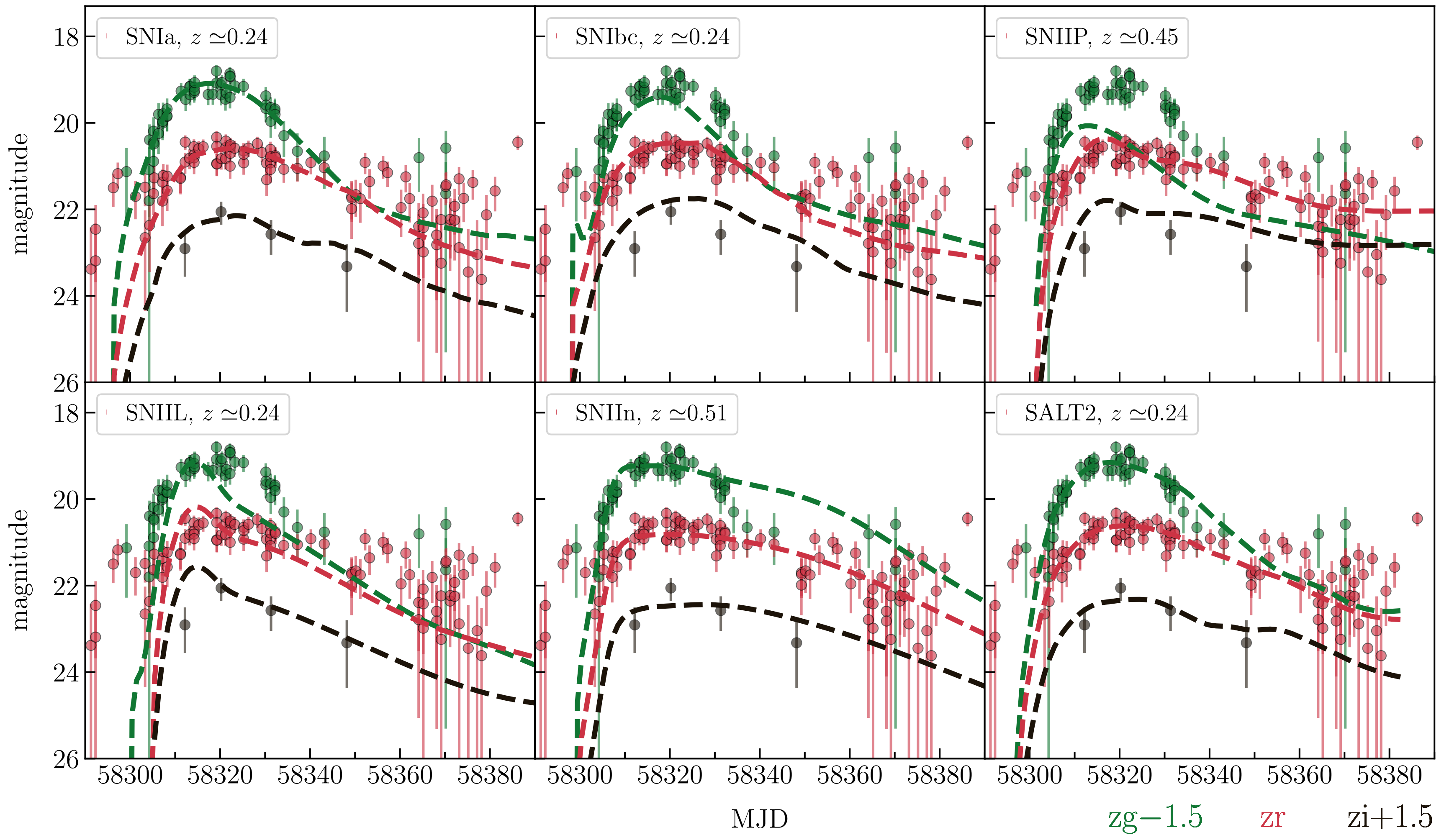
arXiv:2012.01419

- Three fields of ZTF DR3
- **~2×10^6 objects total**
- Four outlier detection algorithms
- 89/227 anomaly candidates
  - **Six (5/6 are new!) SN Ia candidates**
  - RS CVn (confirmed by our spectra)
  - Mira binary candidate
- 188/277 bogus light curves
  - Double star defocusing
  - Bright Mira "echos"
  - Asteroid overlap
  - Bad columns, satellites, spikes, ghosts, etc



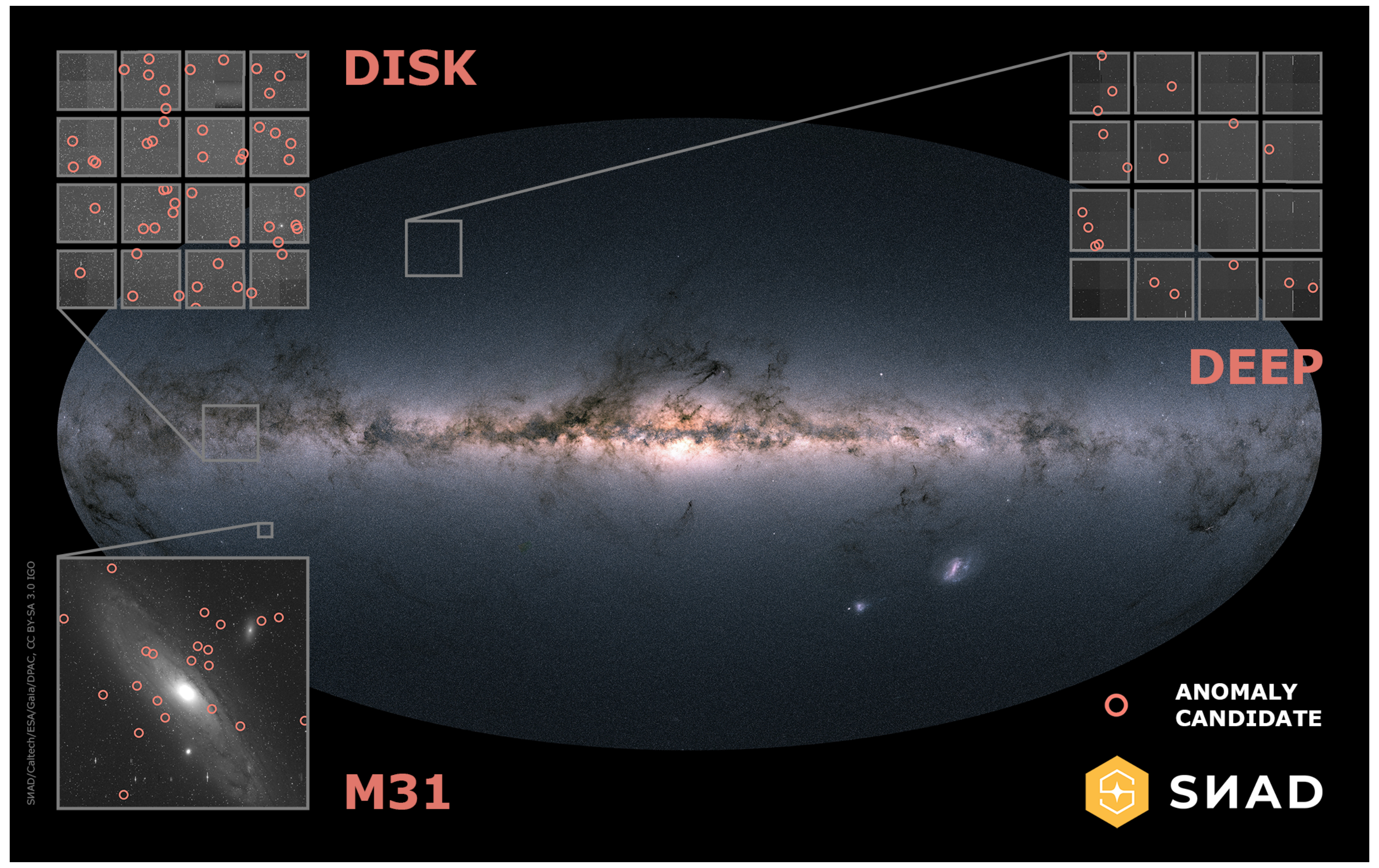






Anomaly Detection Pipeline

https://github.com/snad-space/zwad



$1.8 \cdot 10^6$

$4 \cdot 10^5$

$6 \cdot 10^4$

Outlier detection algorithms:

- Isolation Forest
- Gaussian Mixture Model
- Local Outlier Factor
- One-class Support Vector Machine

40
per field
per algo

277

Expert domain analysis

sncosmo
A Python Library for Supernova Cosmology

Bogus light curves

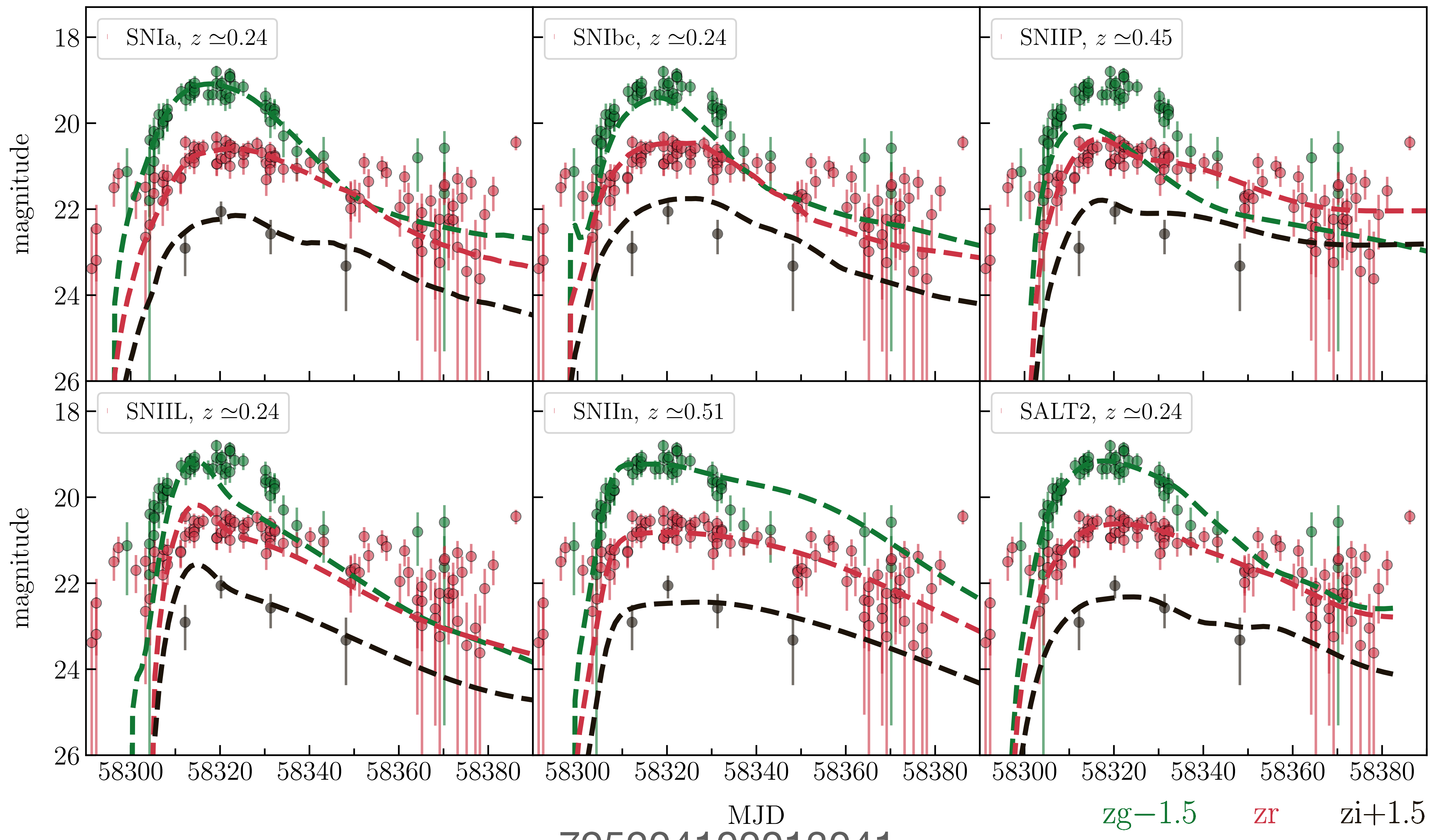
Anomaly candidates

SNAD ZTF DR3 Supernova Candidates

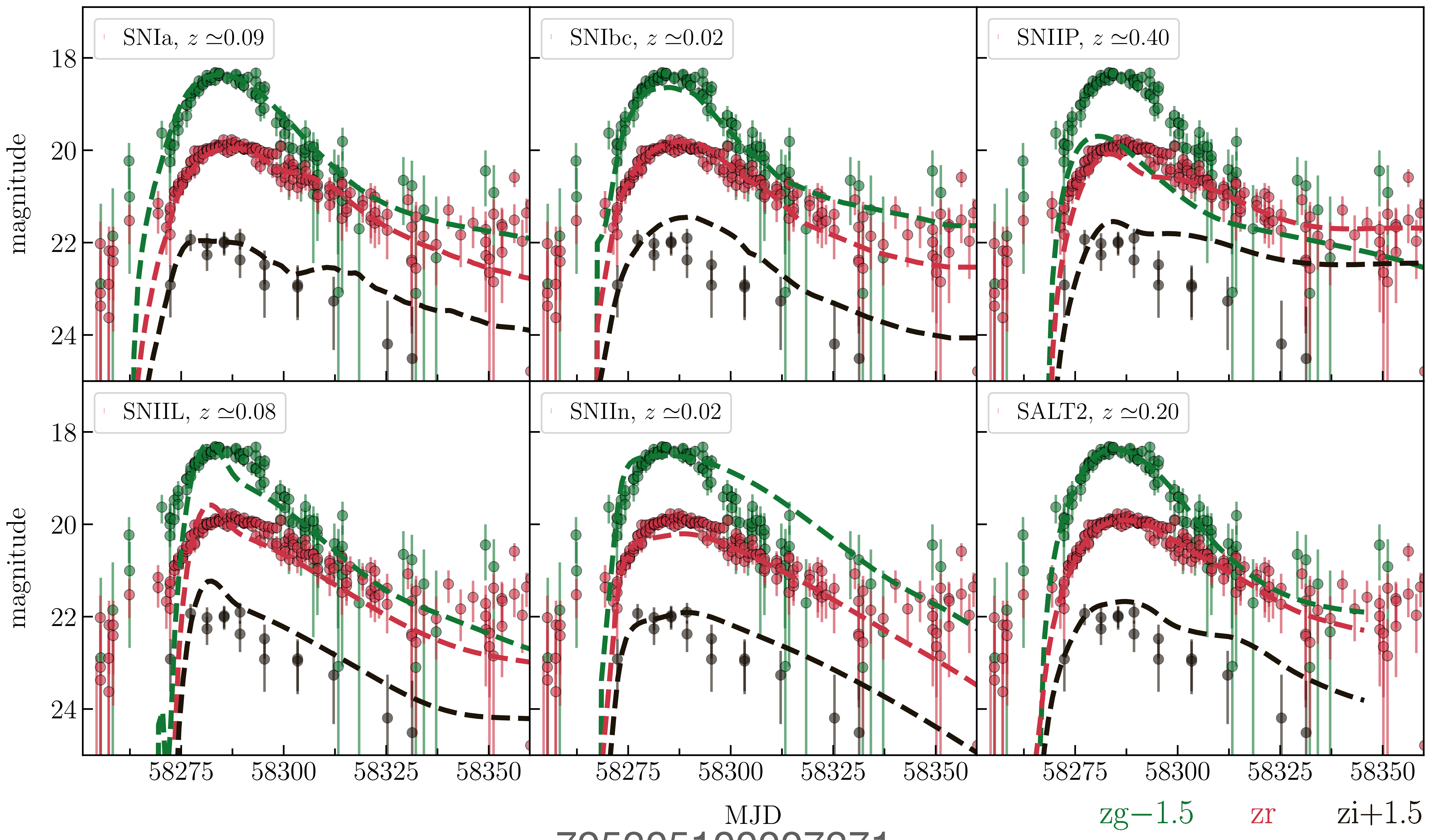
Six candidates from DEEP field (400 000 objects), only one is in TNS

Table 2. Results of the light curve fit with the SALT2 model for supernova candidates from the DEEP field.

OID	Host galaxy*	z_{ph}	z	t_0	x_1	c	Comments†
795202100005941/ZTF18aanbnjh	SDSS J163437.92+521642.2	0.424 ± 0.103	—	—	—	—	Blazar
795204100013041/ZTF18abgvctp	SDSS J160913.83+521251.3	0.375 ± 0.138	~0.24	58320.9336±0.4389	1.71±0.51	−0.044±0.035	—
795205100007271/ZTF18aayatjf	—	—	~0.20	58285.8334±0.1810	−0.54±0.18	−0.075±0.021	SN Ia
795209200003484/ZTF18abbpebf	—	—	~0.11	58299.7269±0.0008	0.60±0.12	−0.013±0.012	SN Ia
795212100007964/ZTF18aanbksg	SDSS J161144.90+555740.7	0.288 ± 0.122	~0.18	58214.4470±0.0002	0.40±0.20	−0.282±0.020	Blazar
795213200000671/ZTF18aaincjv	—	—	—	—	—	—	AGN-I



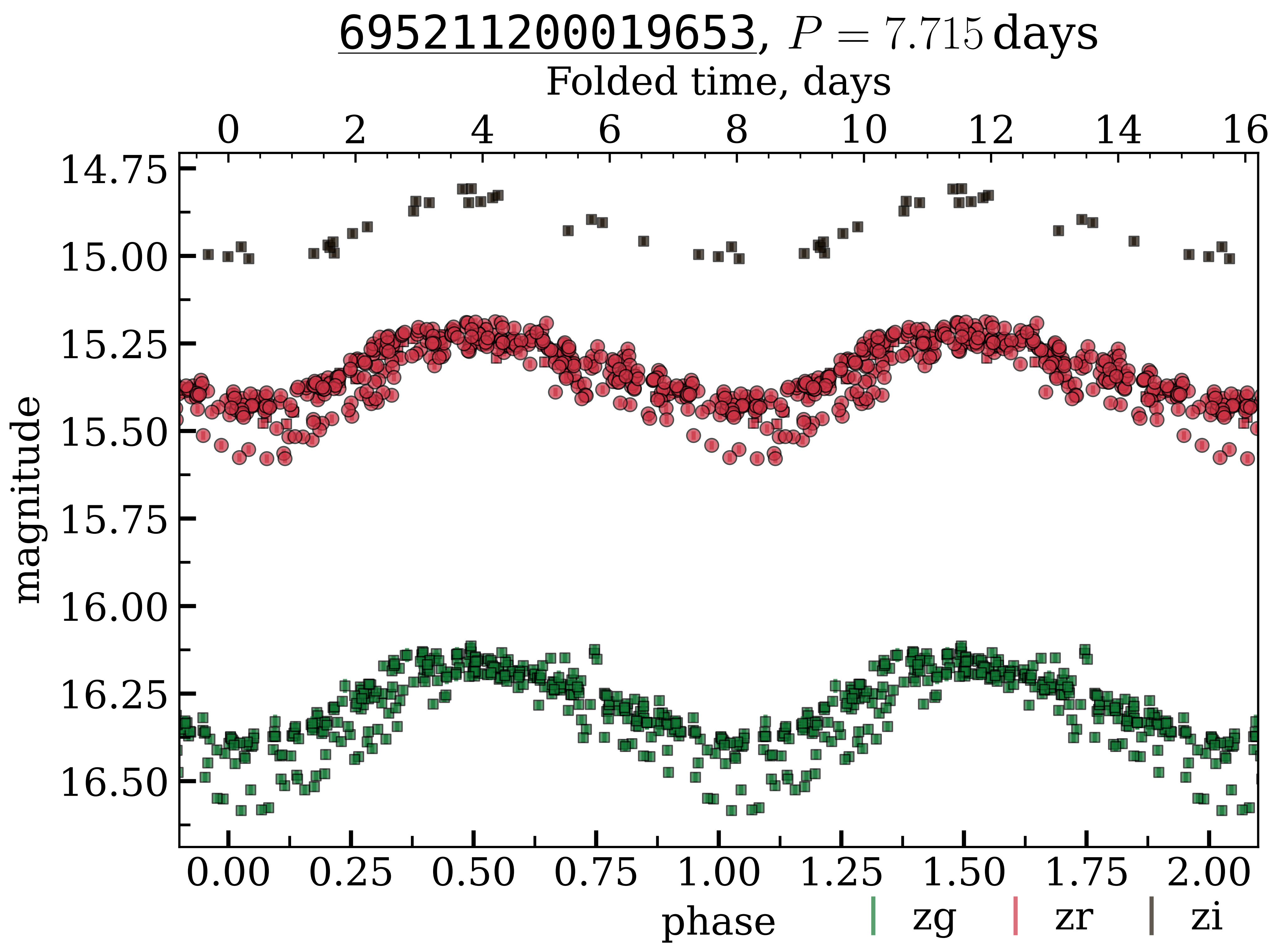
795204100013041

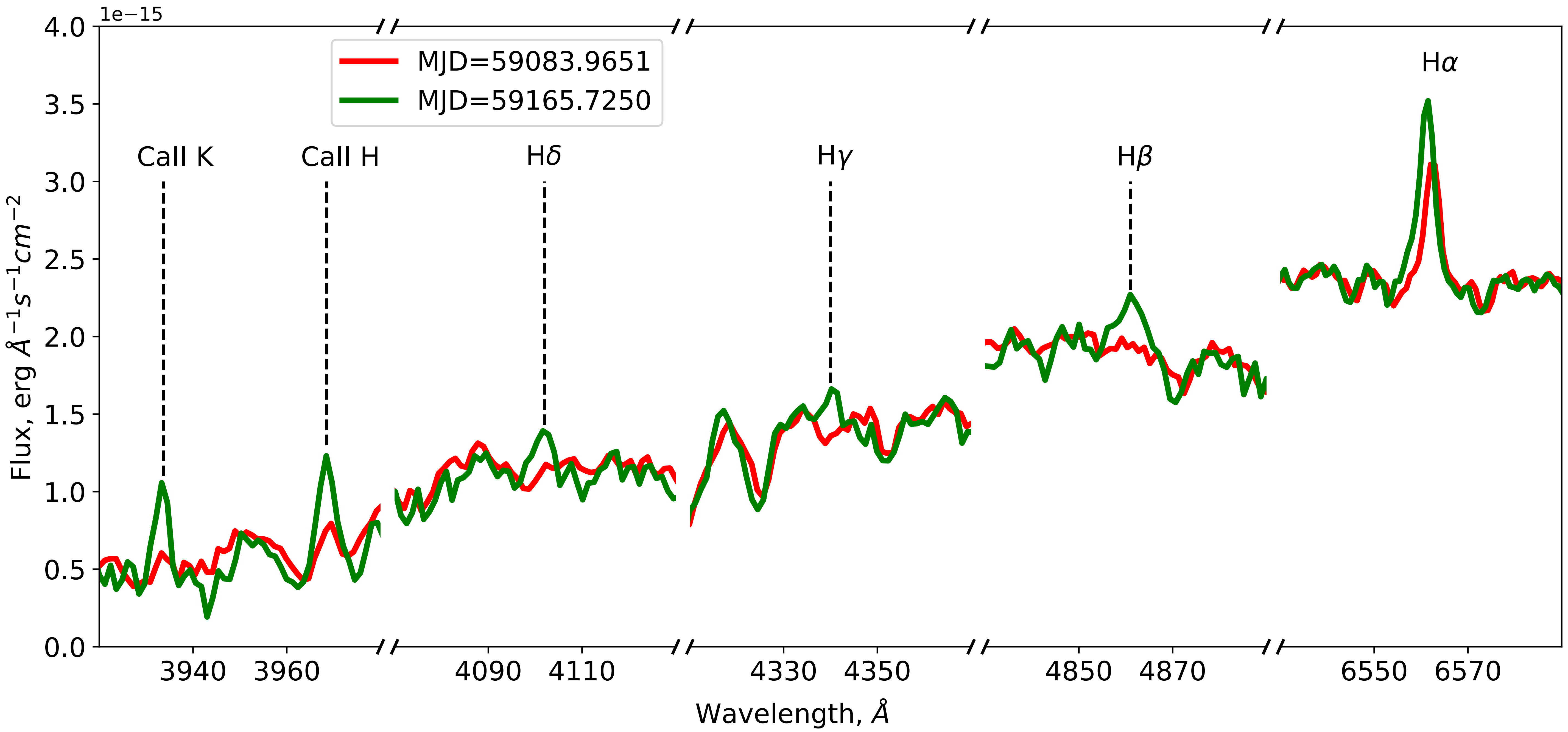


795205100007271

Classification of RS CVn (binary w/ spots)

Our spectra + period change + flare activity

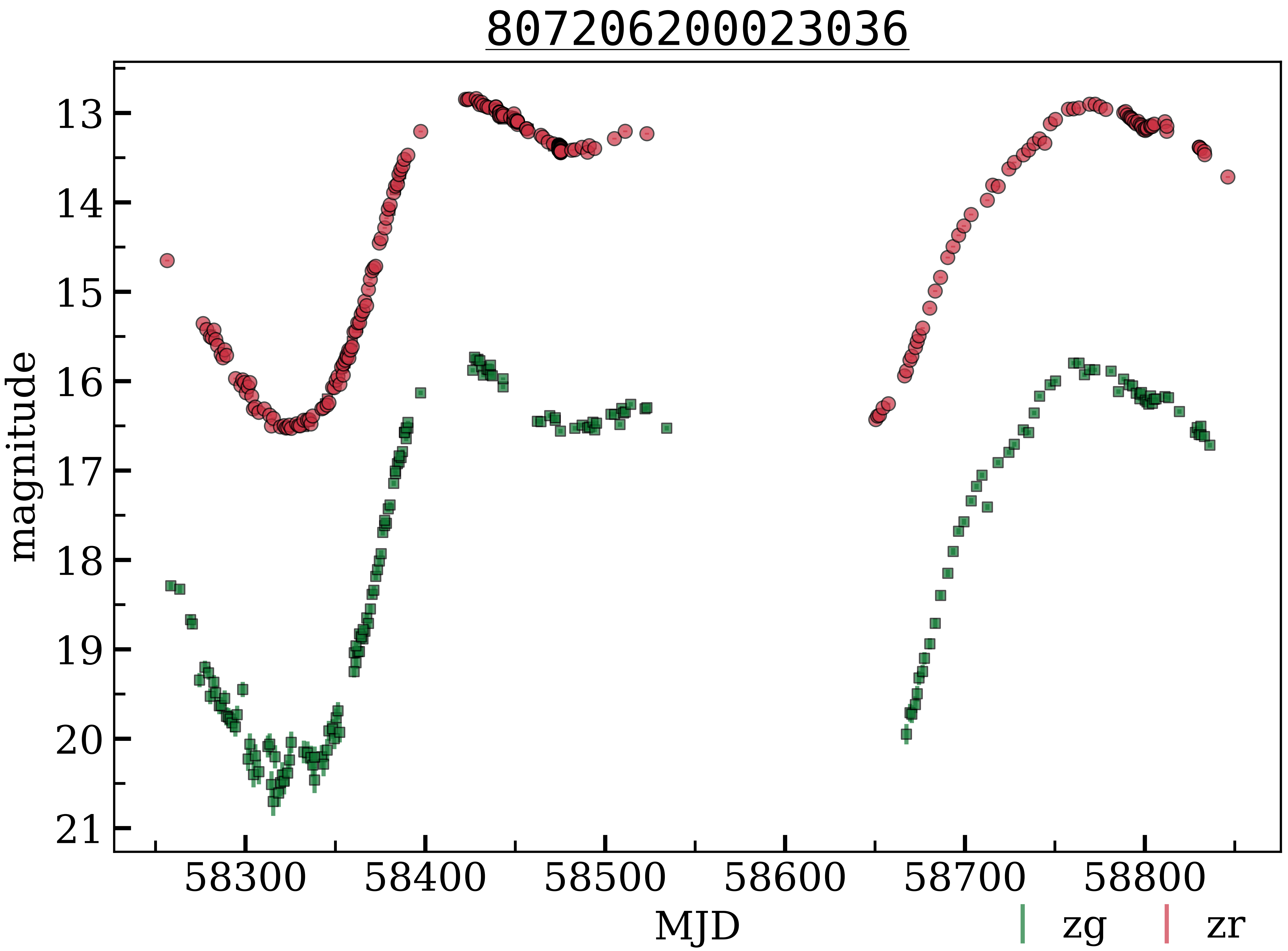




2.5m CMO SAI MSU spectra of the object at different phases of the orbital cycle

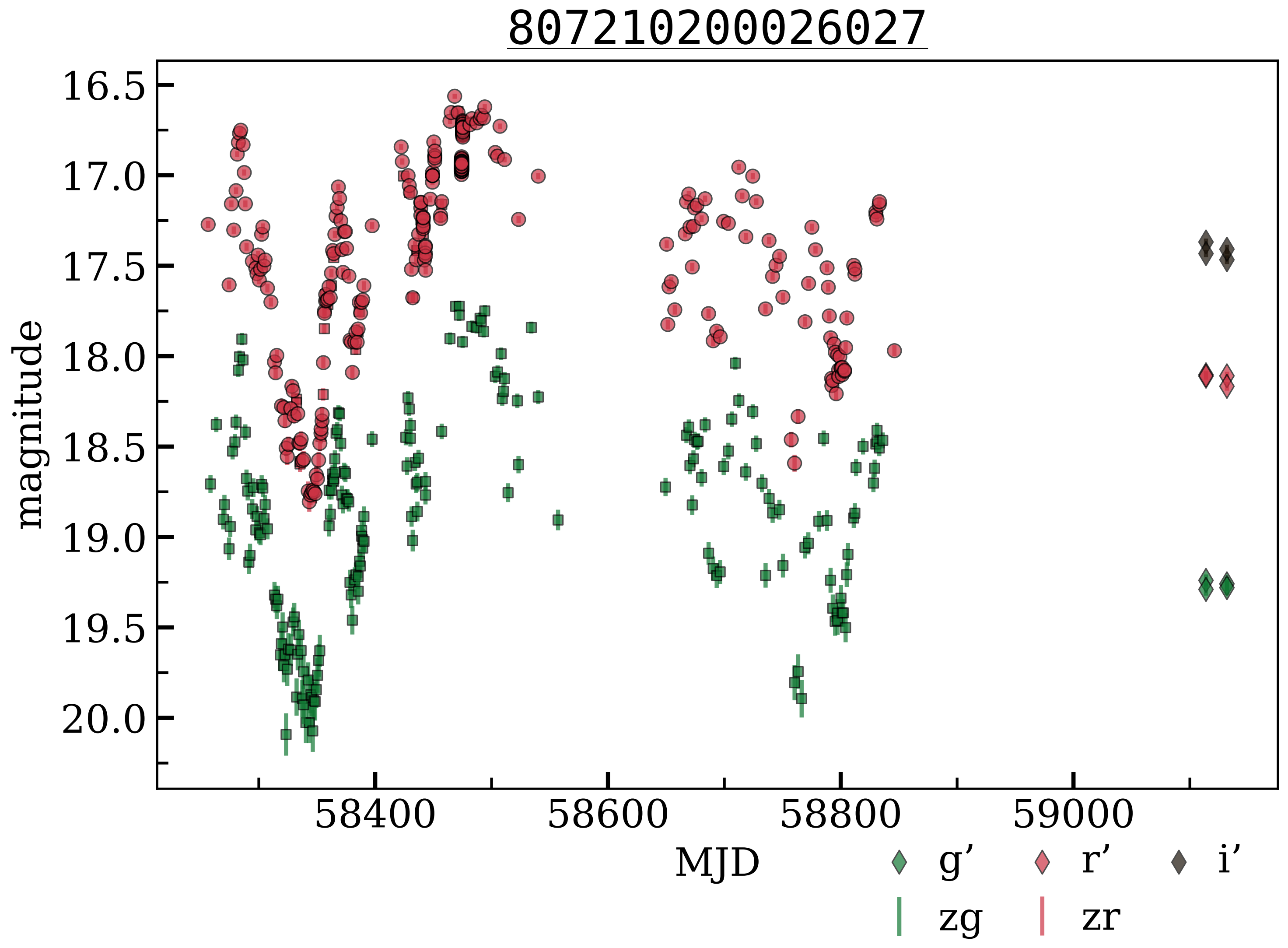
Mira Binary Candidate

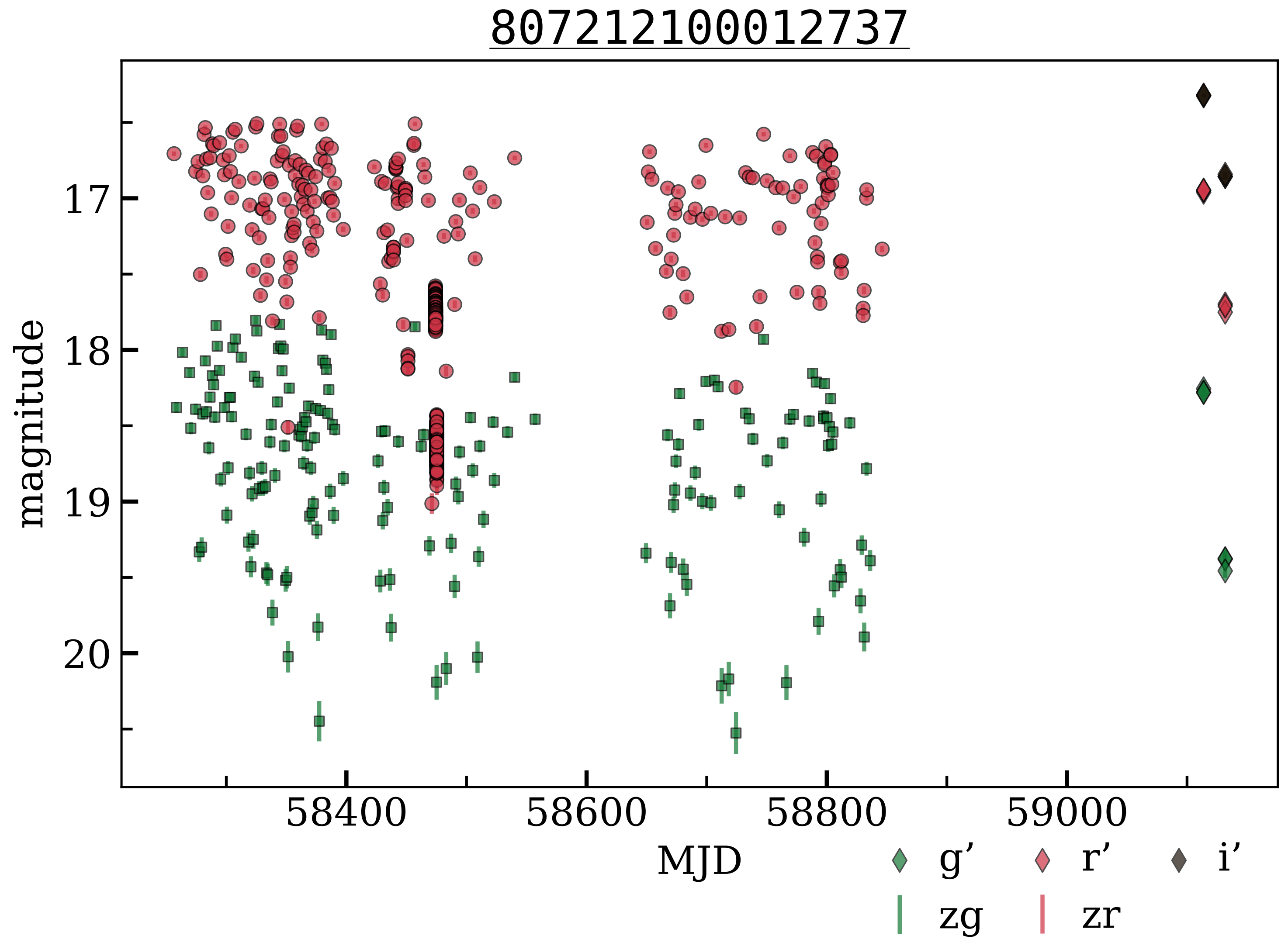
Light curve may indicate the presence of a companion

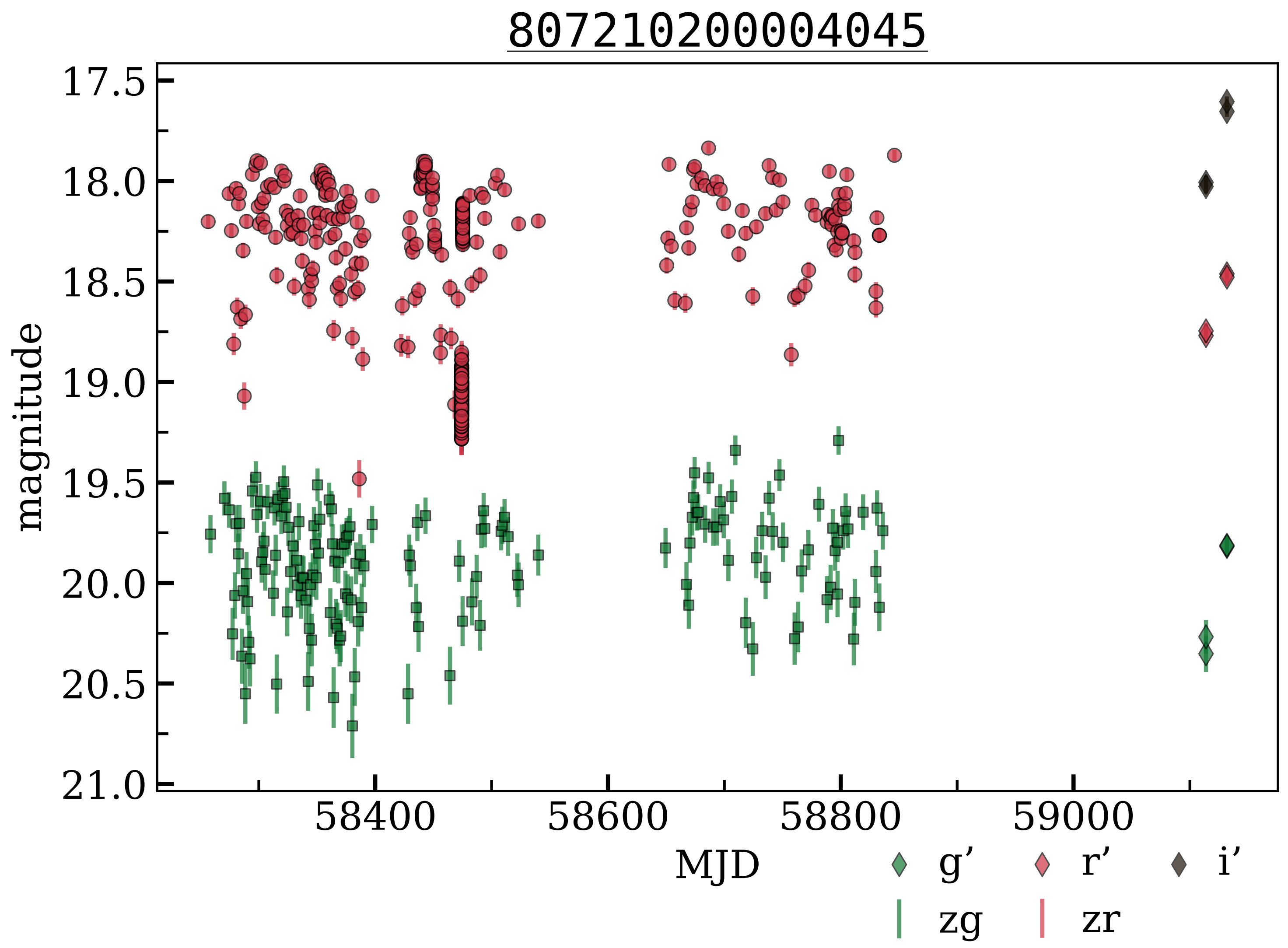


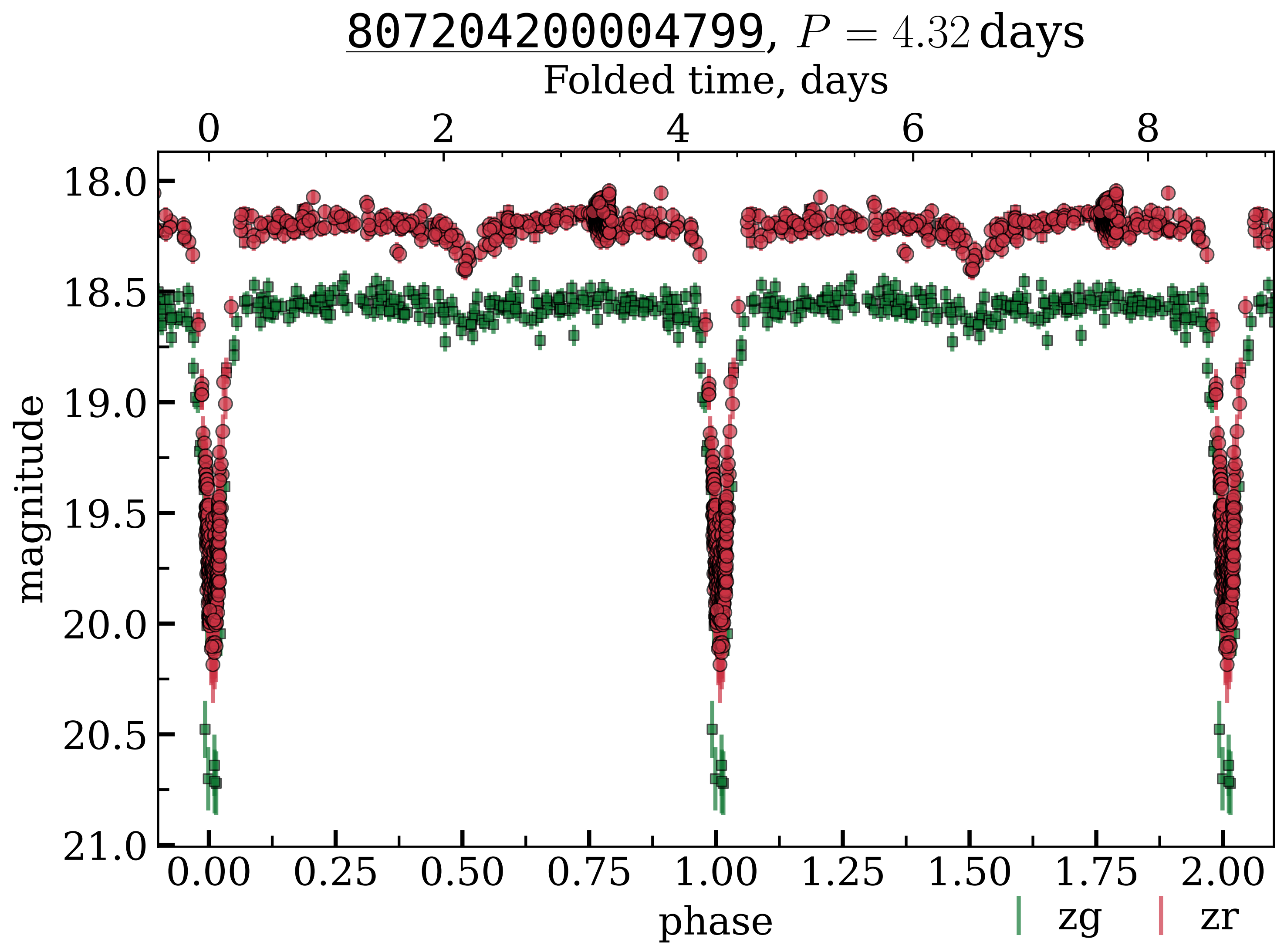
Non-catalogued Sources — 23 / 277

- Diamonds are our observations (the right group of observations)
- Red circles were considered by the outlier detection algorithms
- Squares are other light curves of the same source



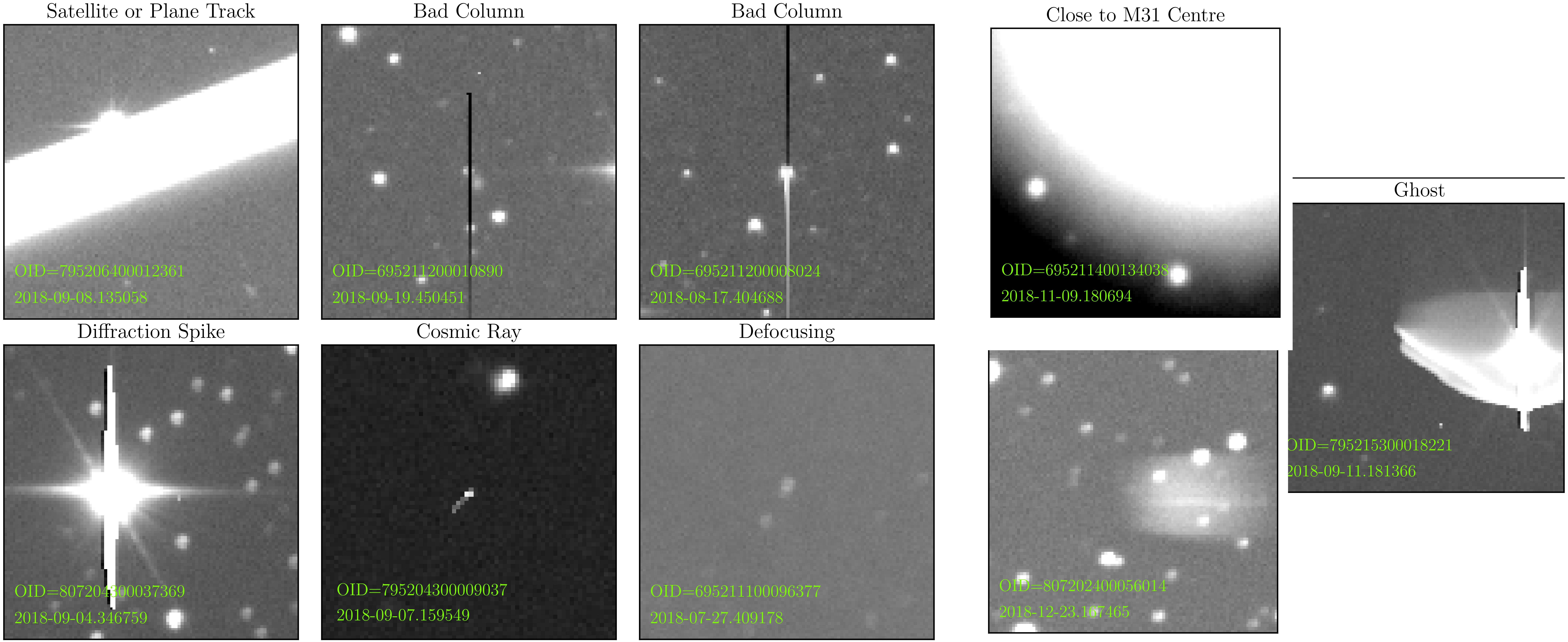






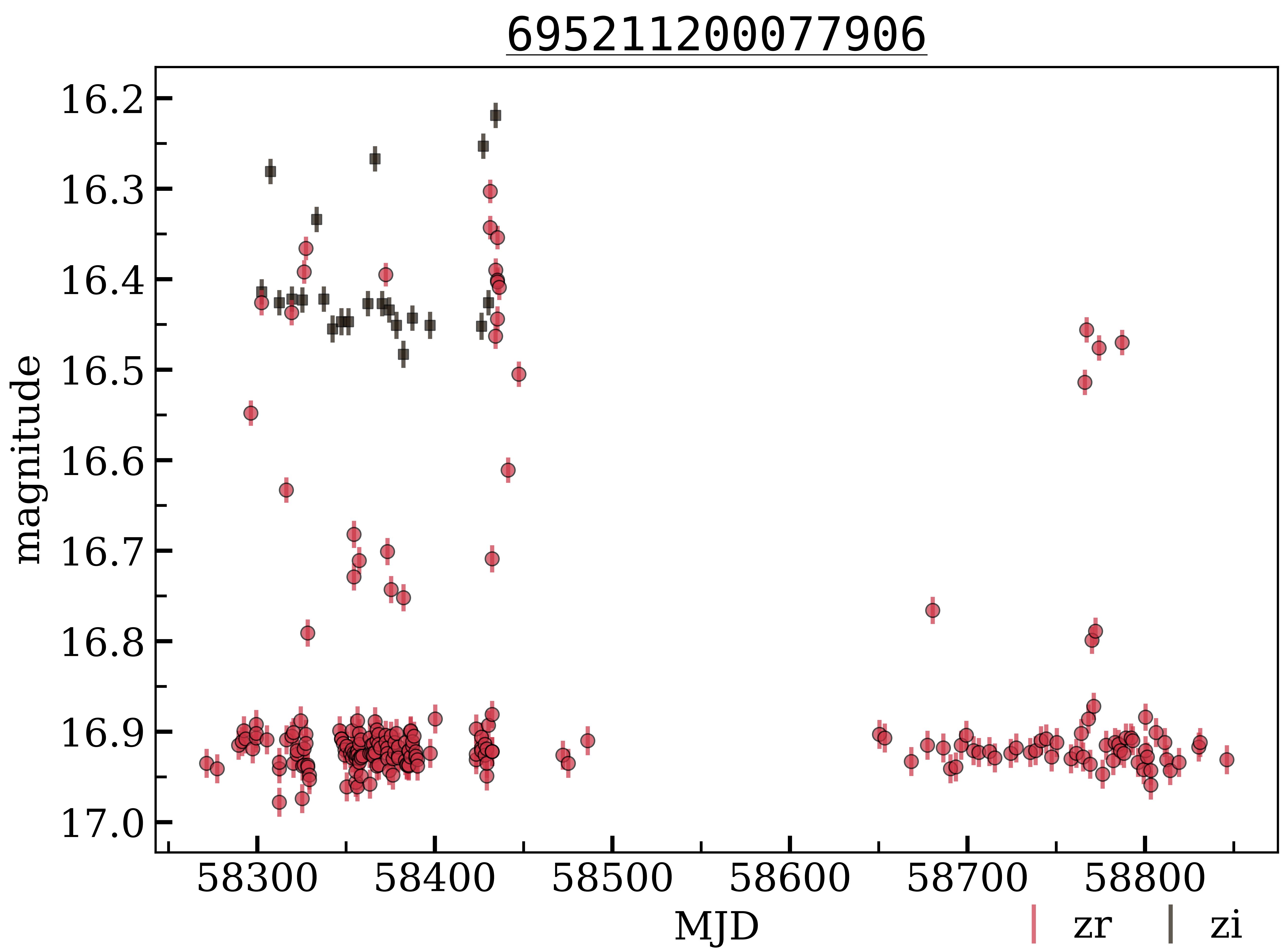
Bogus Detections

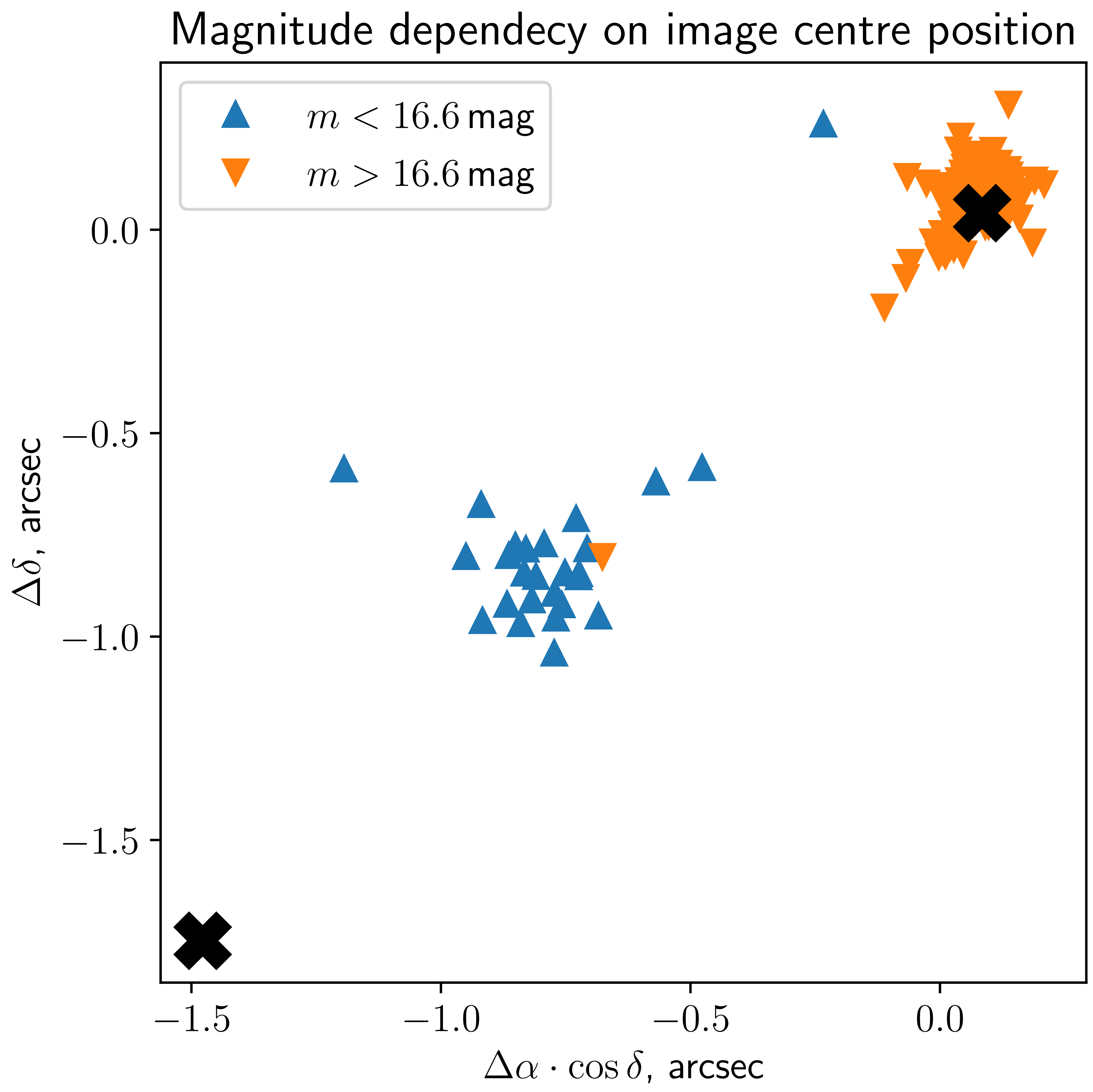
The objects are in the frame centers



Double Stars Defocusing

If a separation is about 2” (typical FWHM) then defocusing can cause the false variability

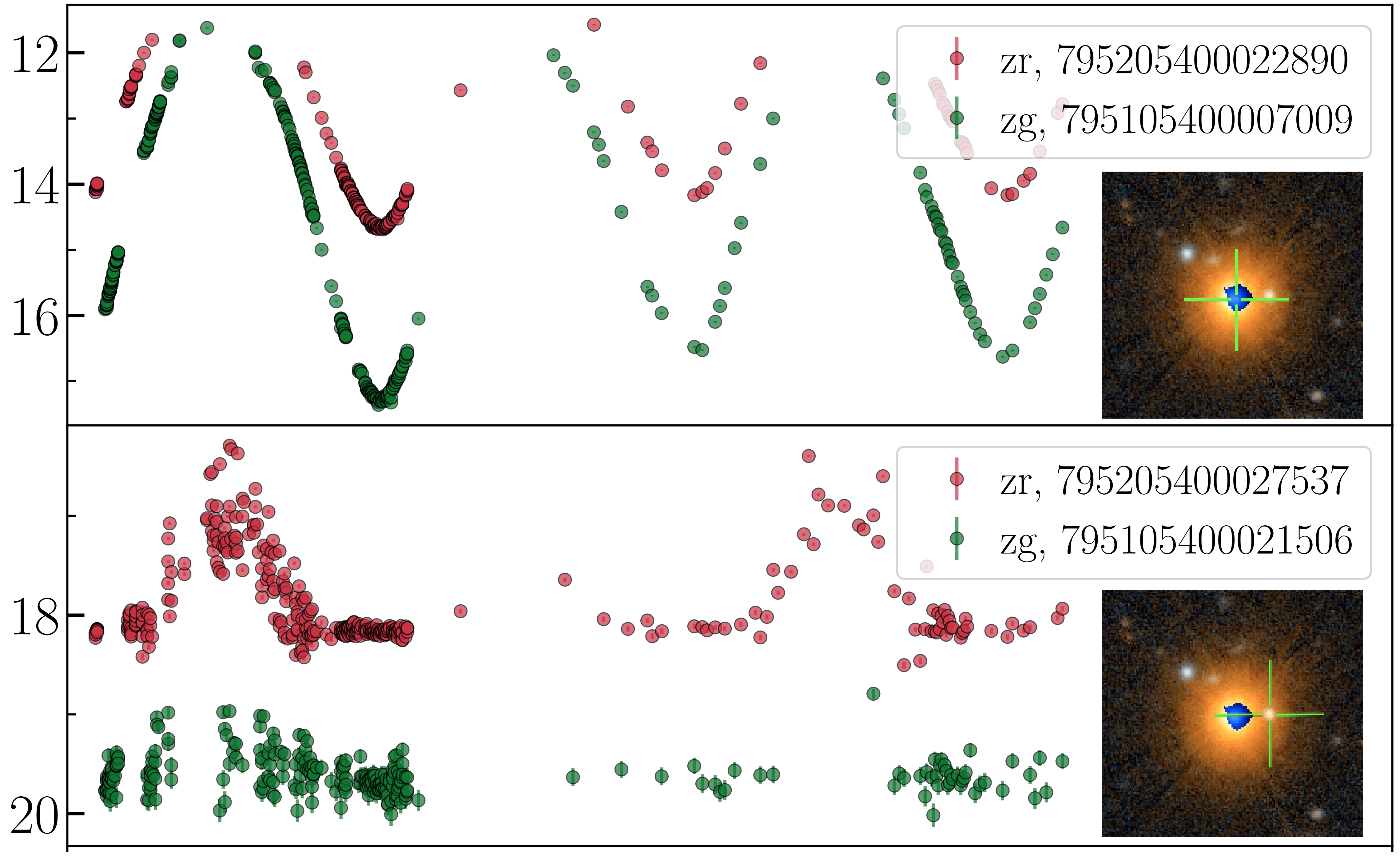




IW Dra "Echos"

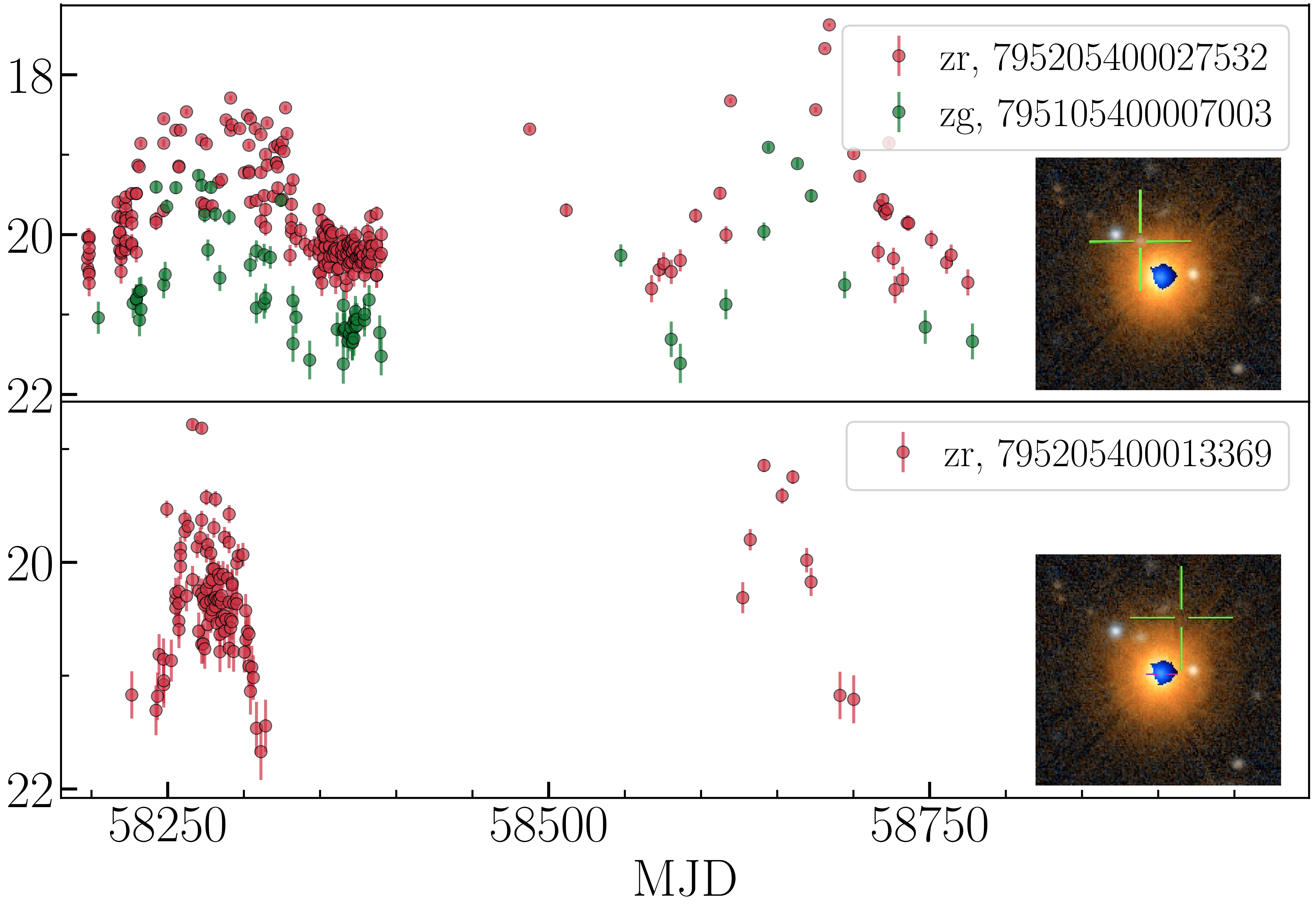
All four objects are found by the outlier detection pipeline

Mira-like IW Dra



Echo 1

Echo 2



Echo 3

Applying expert bias to anomaly detection

From outlier to anomaly detection algorithm

- How to discriminate annoying non-anomalies sources and bogus light curves?
  - We can ask an expert interactively about each new outlier
  - If it is not an anomaly, set lower probability to objects like this
  - Retrain, ask the expert again
- We can do the opposite: highlight interesting class of objects for classification of rare objects. Listen Emille Ishida's talk about this

Data

↓ **Train initial model**

Machine

↓ **The best outlier up date**

↑ **Update model with the outlier label**

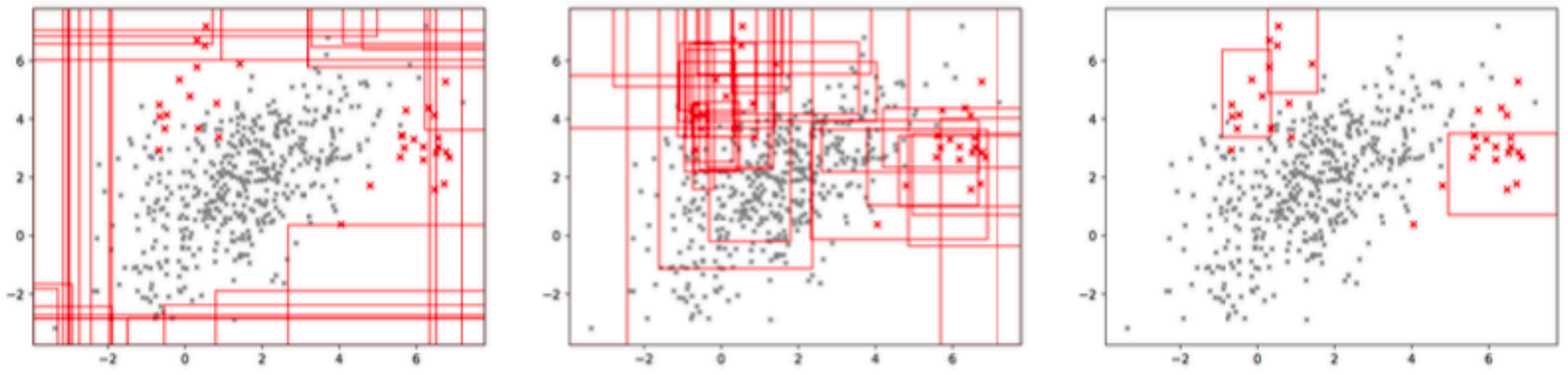
Inspect ouliers using external data

Active anomaly detection (AAD)

Implementation of the machine—expert loop, Das+2018

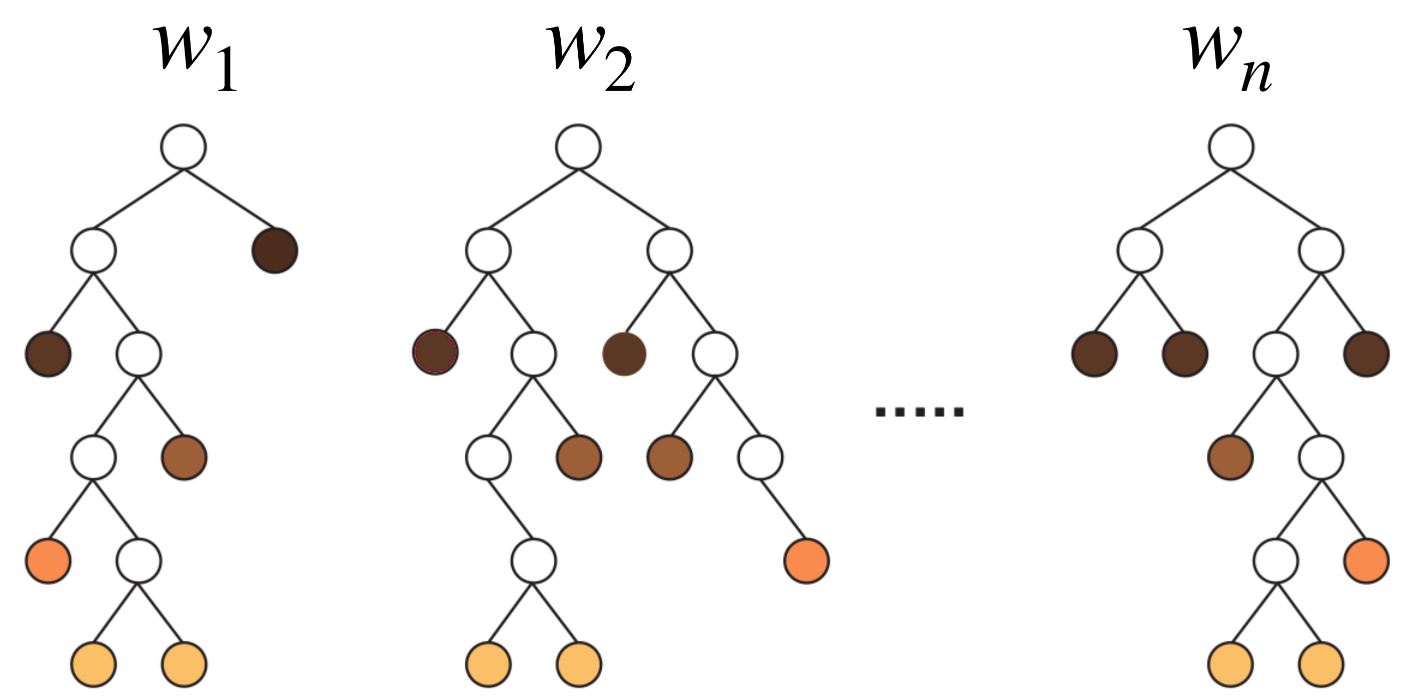
Algorithm:

1. Initialize isolation forest, set equal w_i to each iTree
2. Ask the forest for the outlier with the largest score
3. Ask an expert to classify the object as normal or anomaly
4. If anomaly, go to step 2 and ask next outlier
5. If normal, update $\{w_i\}$ to give lower influence to wrong detectors, go to step 2



(a) Baseline (b) AAD (c) Descriptions

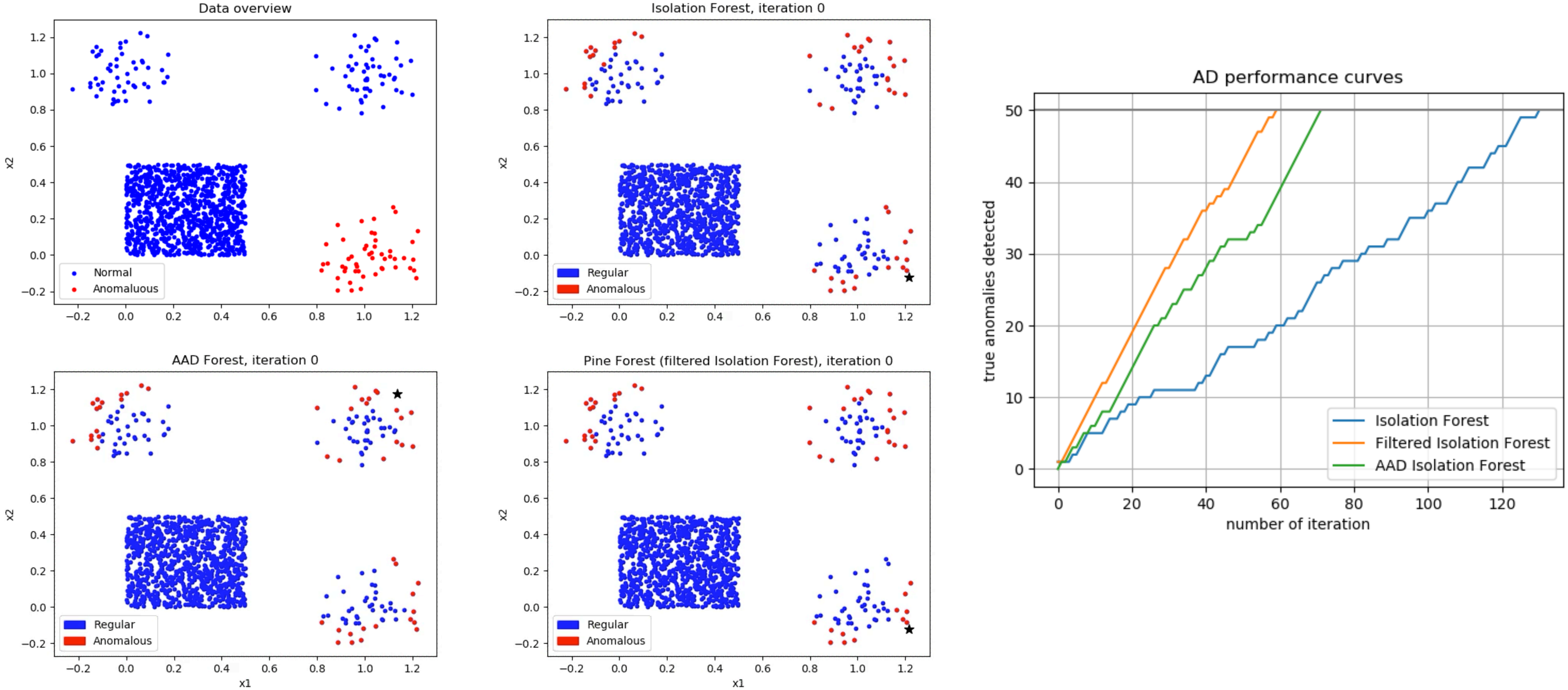
arXiv:1809.06477



There are other algorithms to solve this problem, we are developing a (better) alternative (Korolev+, in prep)

Pine Forest

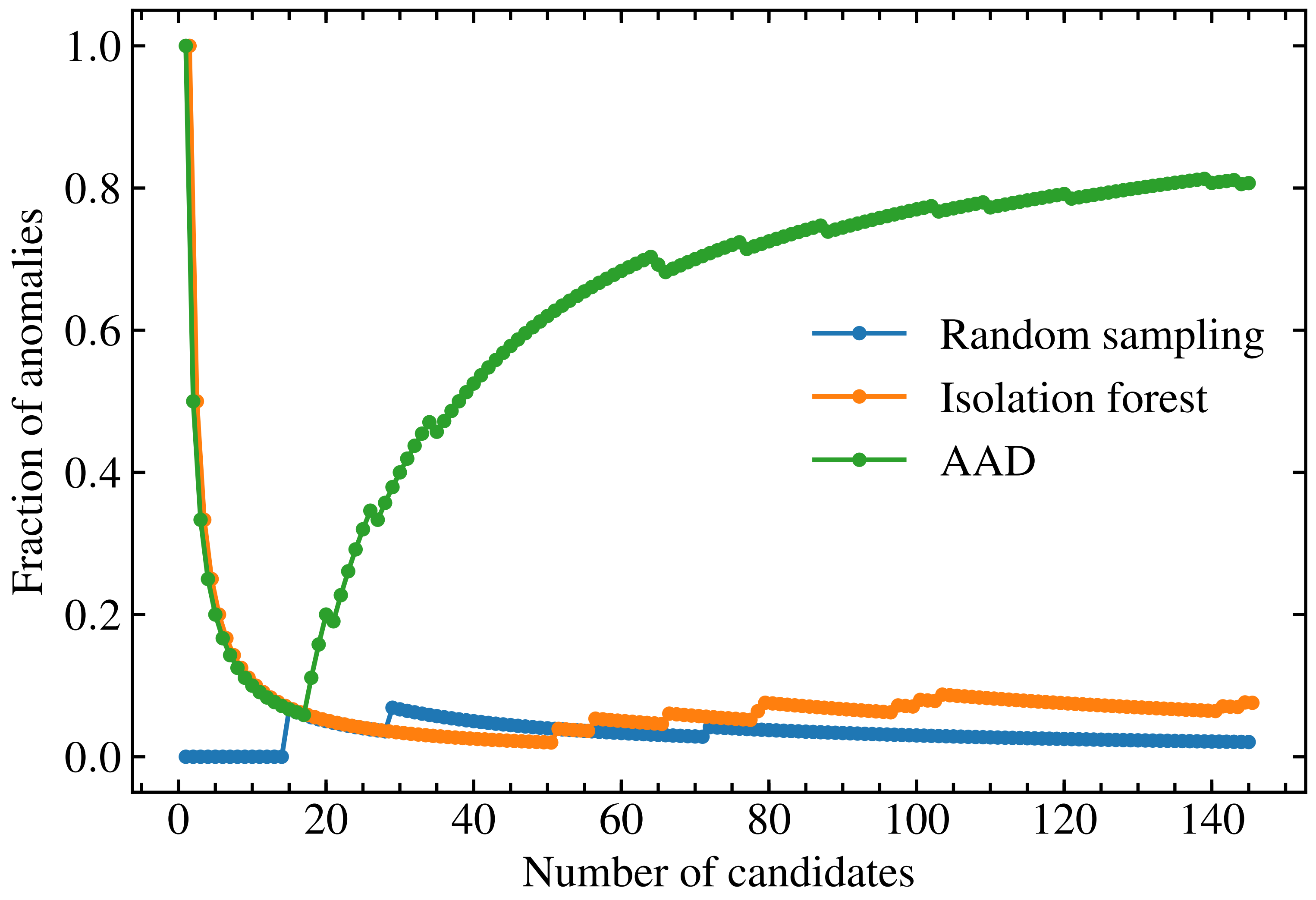
Cutting bad trees and growing good trees, Korolev+ in prep.



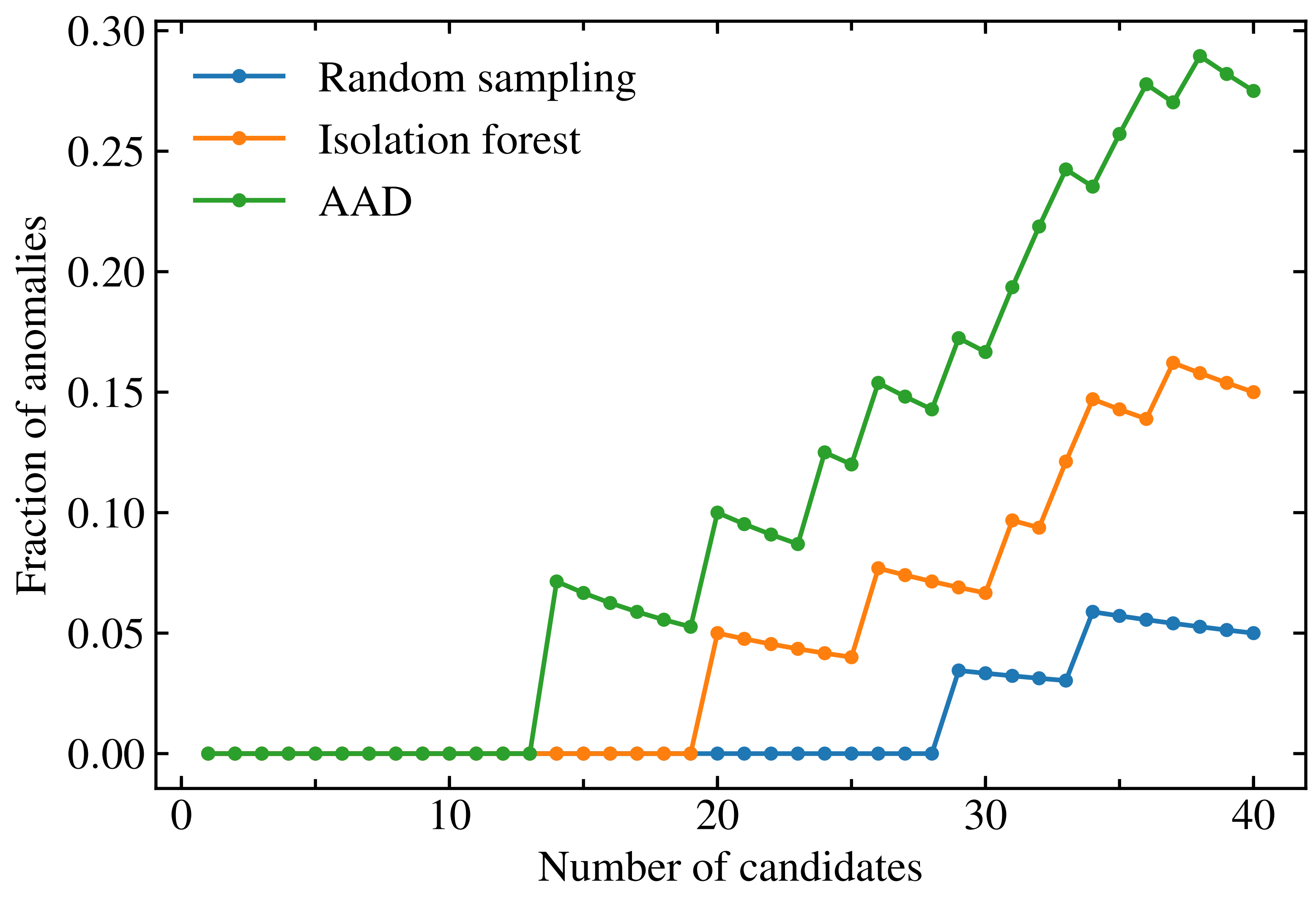
Case: PLAsTiCC & OSC

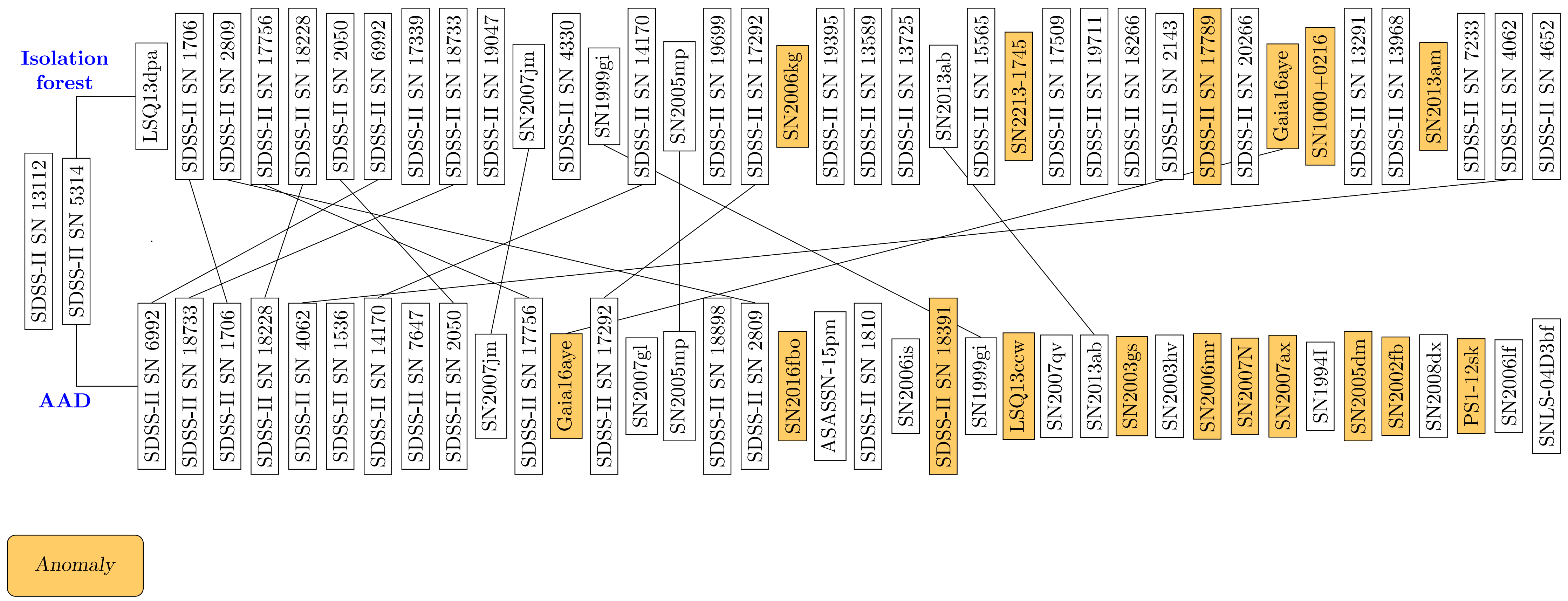
arXiv:1909.13260

PLAsTiCC (LSST sims)



OSC

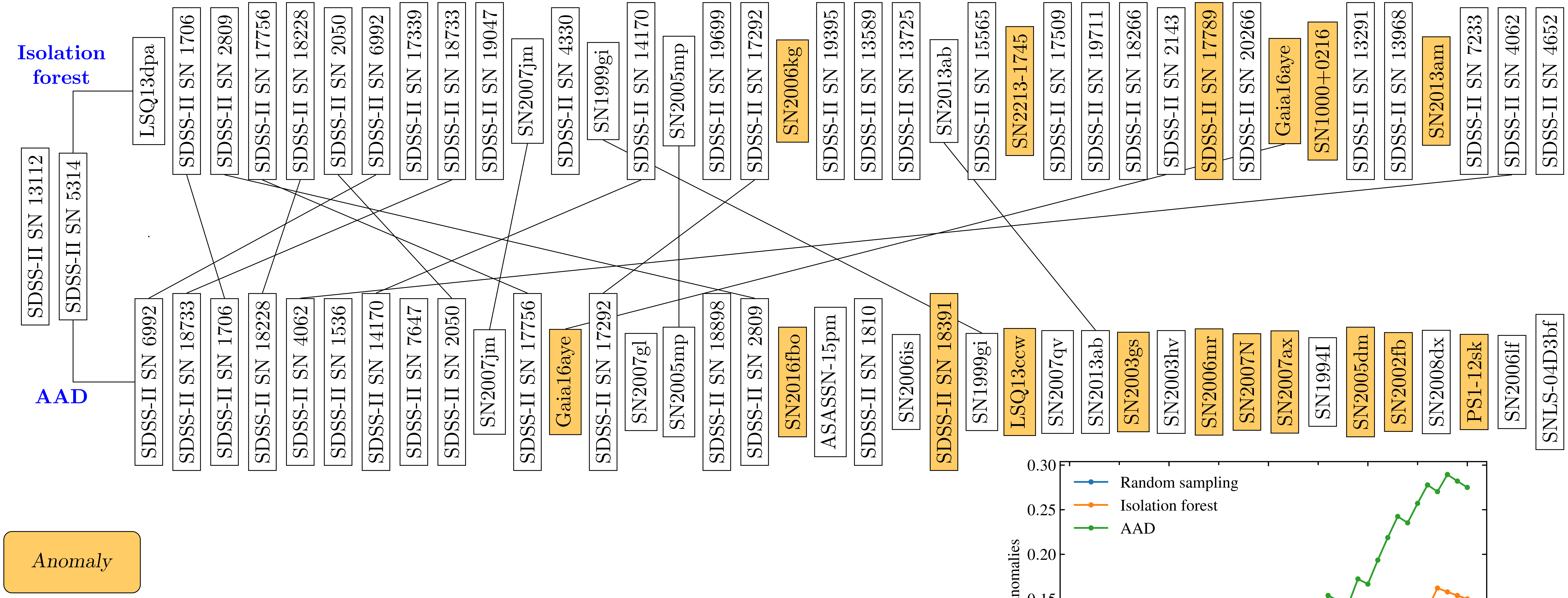


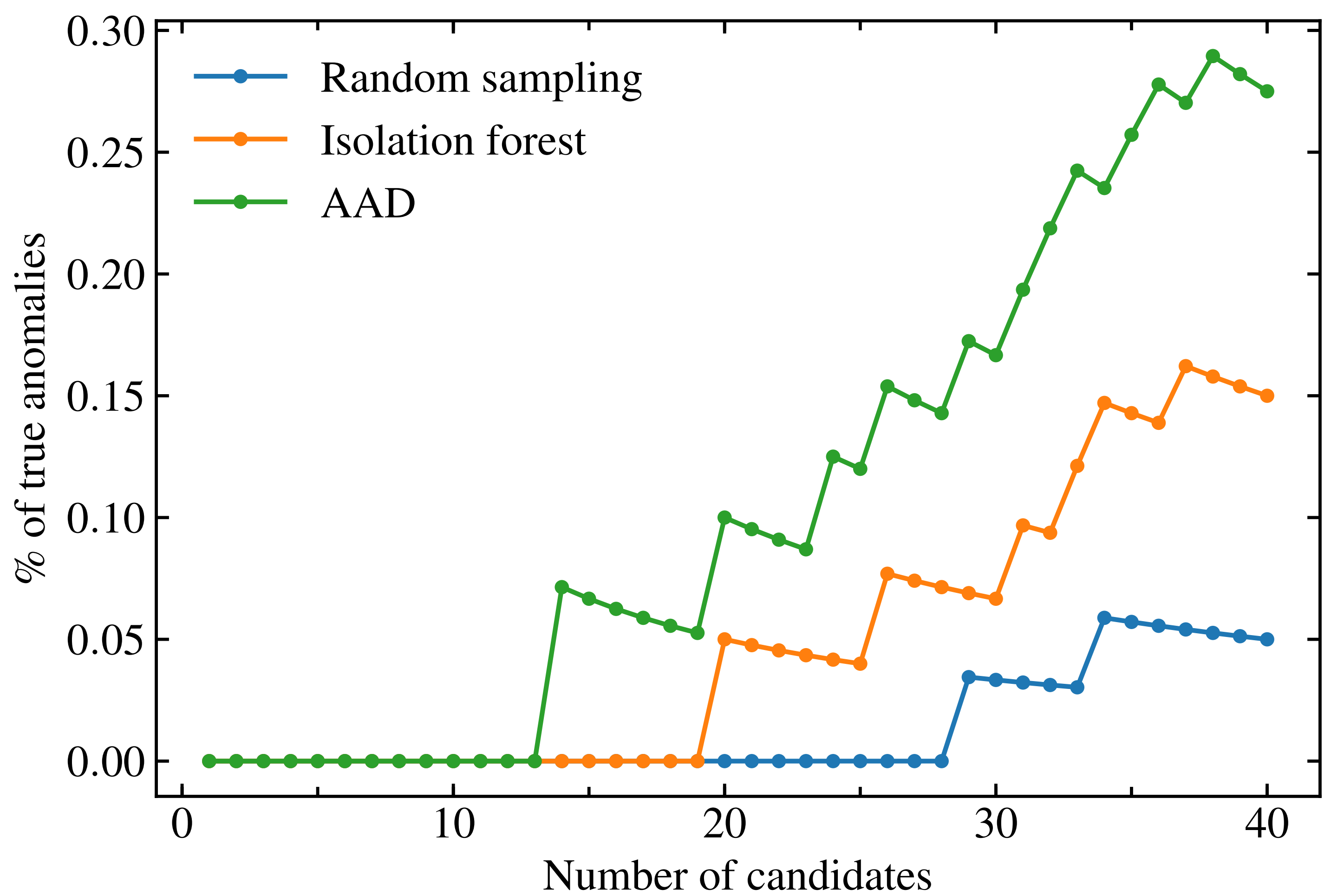


Lessons learnt:

- AAD works
- Anomaly definition (expert bias) matters
- Real data is much harder!

AAD: OSC Case

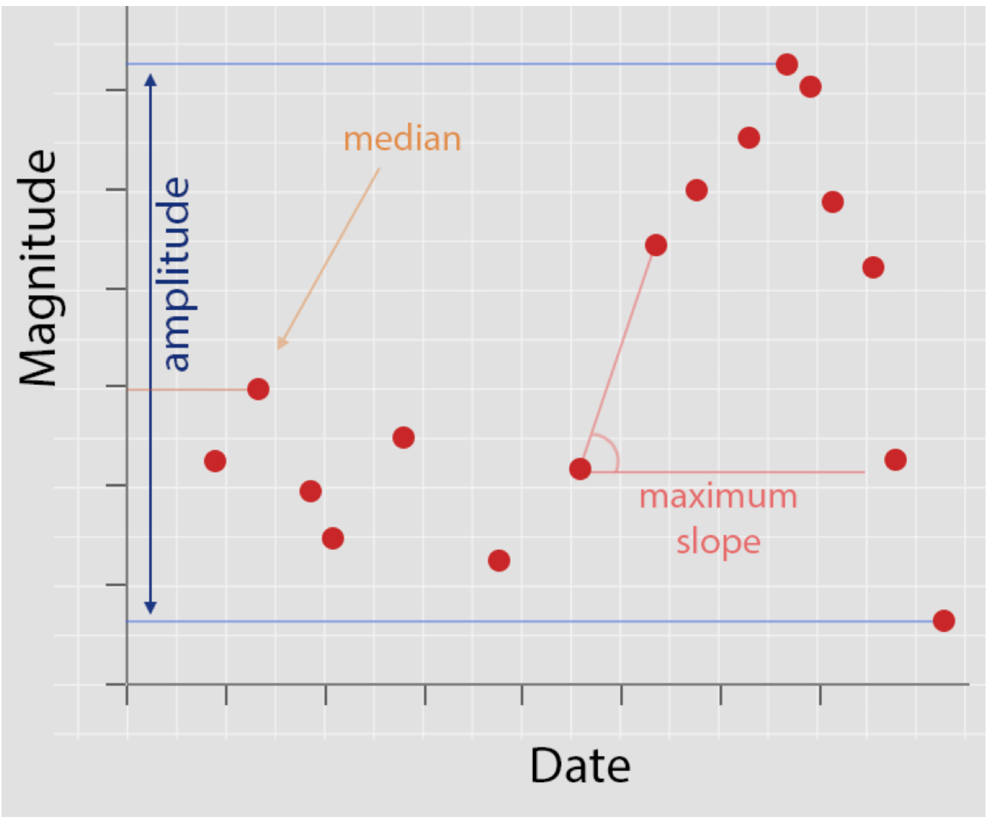




arXiv:1909.13260

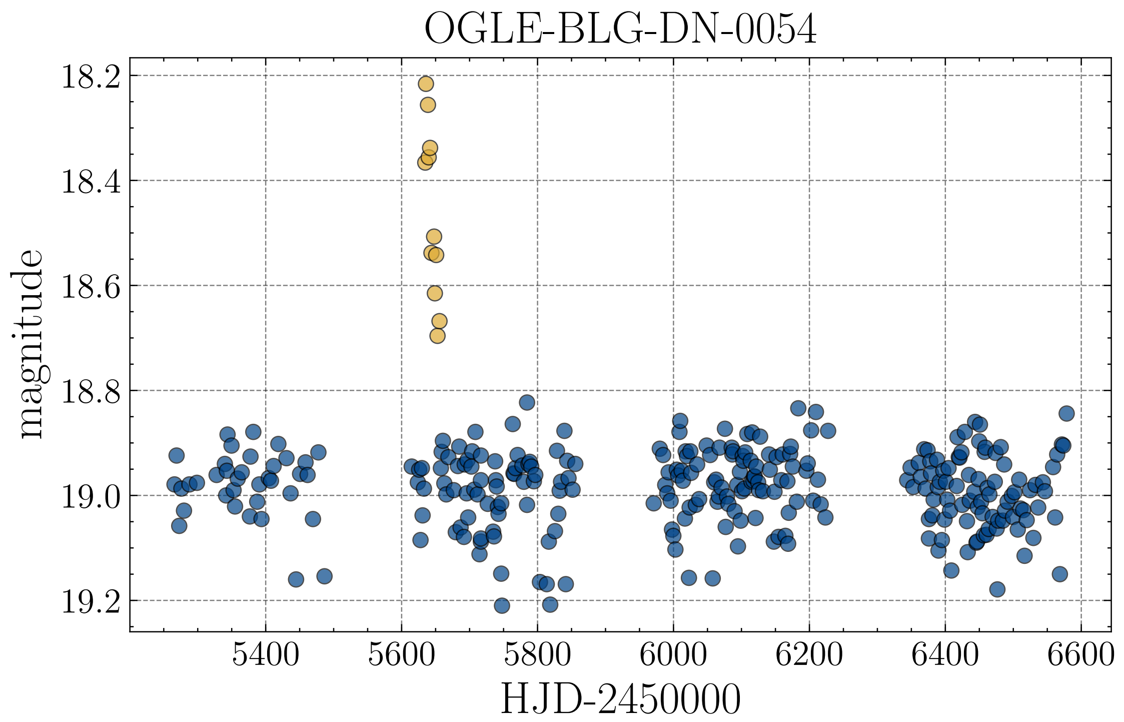
By-product: `light-curve` feature extractor

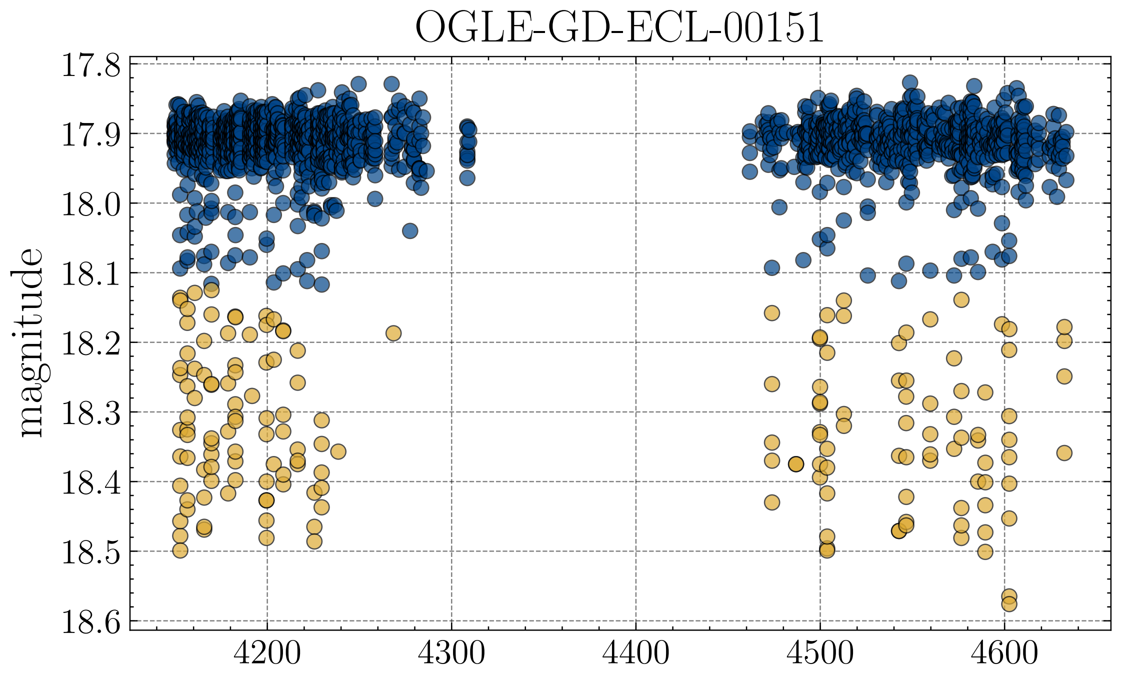
https://github.com/light-curve



Anastasia Lavrukhina

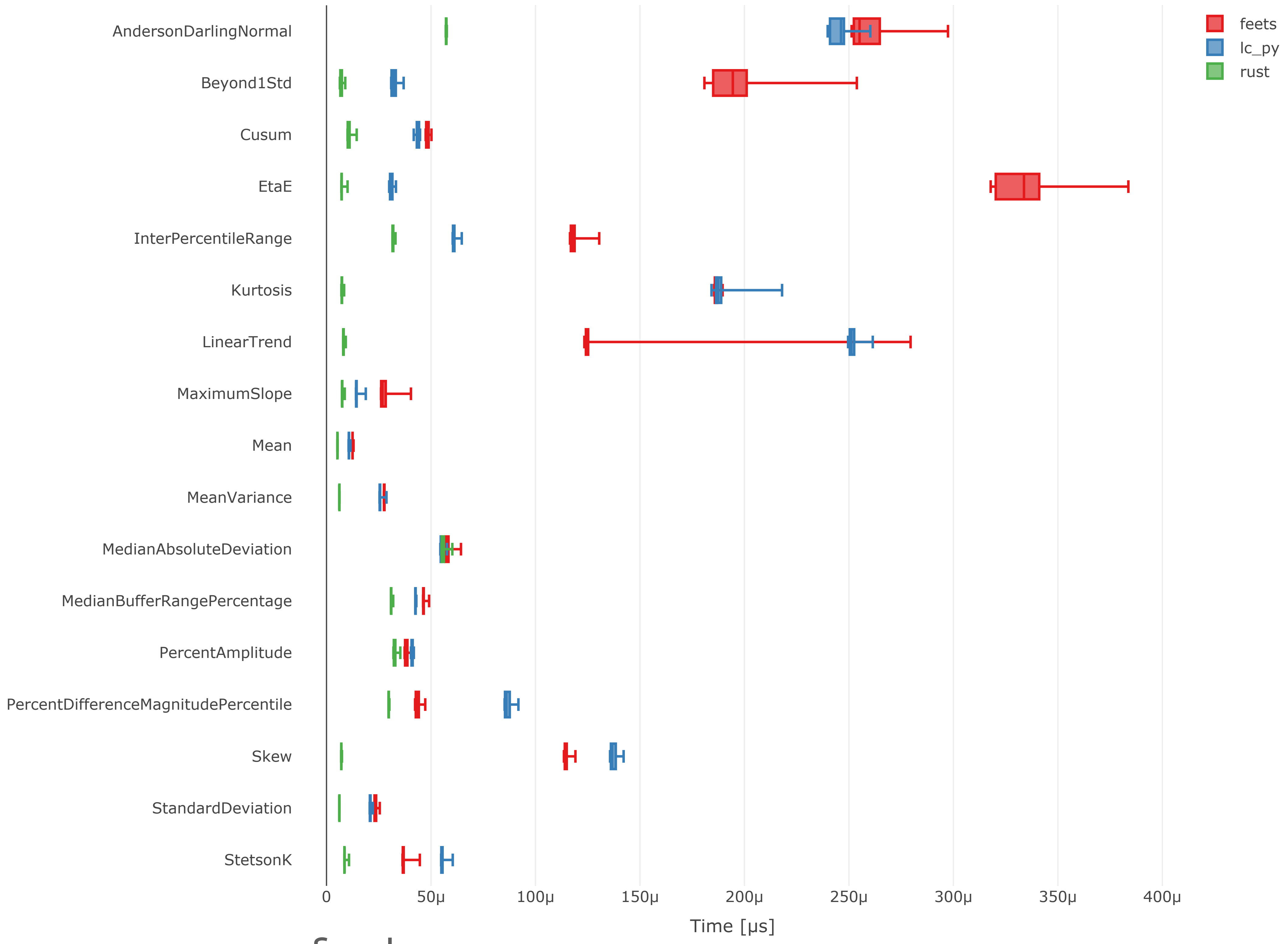
- **Performant** Rust/Python code: processing of ~10^6 light-curves, Nobs ≳ 100, takes few CPU hours
- Rich feature set
  - Magnitude statistics: mean-, median-, momentum- quartile-based
  - Shape-based: Stetson (1996) K, η^e (Kim+ 2014)
  - "Fast" Lomb–Scargle periodogram peaks and other derivatives
  - Parametric fits: linear, SN-like functions (Bazin+ 2009, Villar+ 2019)
  - **New Otsu-split extractor**: powerful features to classify recurrent outbursts, eclipsing binaries, etc (Lavrukhina & Malanchev in prep.)
- Hundreds of unit tests, packages for Linux and Intel Macs
- Serves **three ZTF/LSST brokers**: Ampel, Antares, Fink
- `python3 -m pip install light-curve`





`light-curve` **benchmarks**

- 1.5—140 times faster than feets
- Periodogram is few times faster than "fast" implementation in `astropy` and `gatspy`.
- Large set of "cheap" features (w/o periodogram and parametric fits) can be done in few ms * CPU for Nobs=1000
- Realistic feature set including periodogram, Bazin and Villar fits is ~25 ms * CPU for six *ugrizy* LSST 3-year light curves (tested on the ELAsTiCC training set).
- Single CPU is (almost?) enough to process all LSST alerts in real time!



`feets`
vs Python implementation of `light-curve` (lc_py)
vs Rust implementation of `light-curve` (rust).
Smaller is better

By-product: SNAD ZTF viewer

https://ztf.snad.space

ztf.snad.space/dr8/view/633207400004730

login

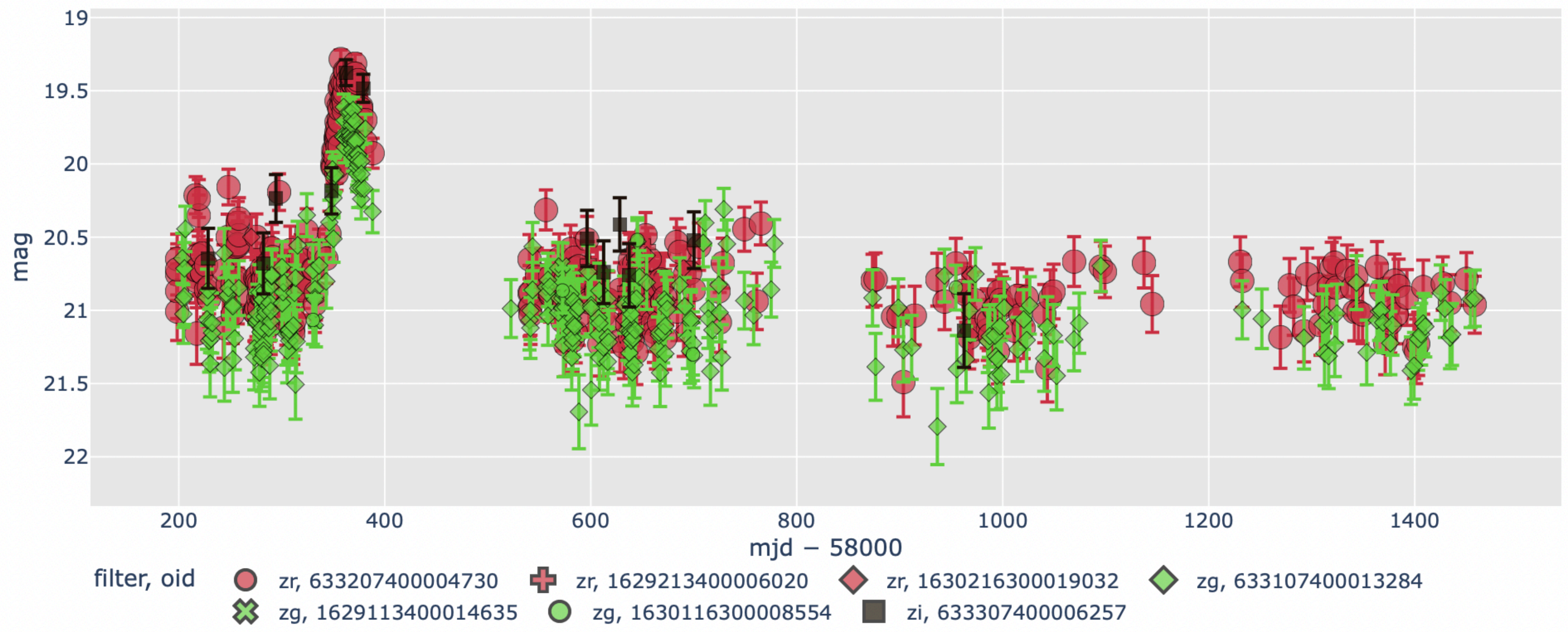
SNAD ZTF DR8 object viewer

OID E.g. 633207400004730 GO

Coordinates 00h00m00s +00d00m00s radius (arcsec) 1 GO

SNAD101 — 633207400004730

Self-matched ZTF light-curve



Download PNG, PDF, CSV

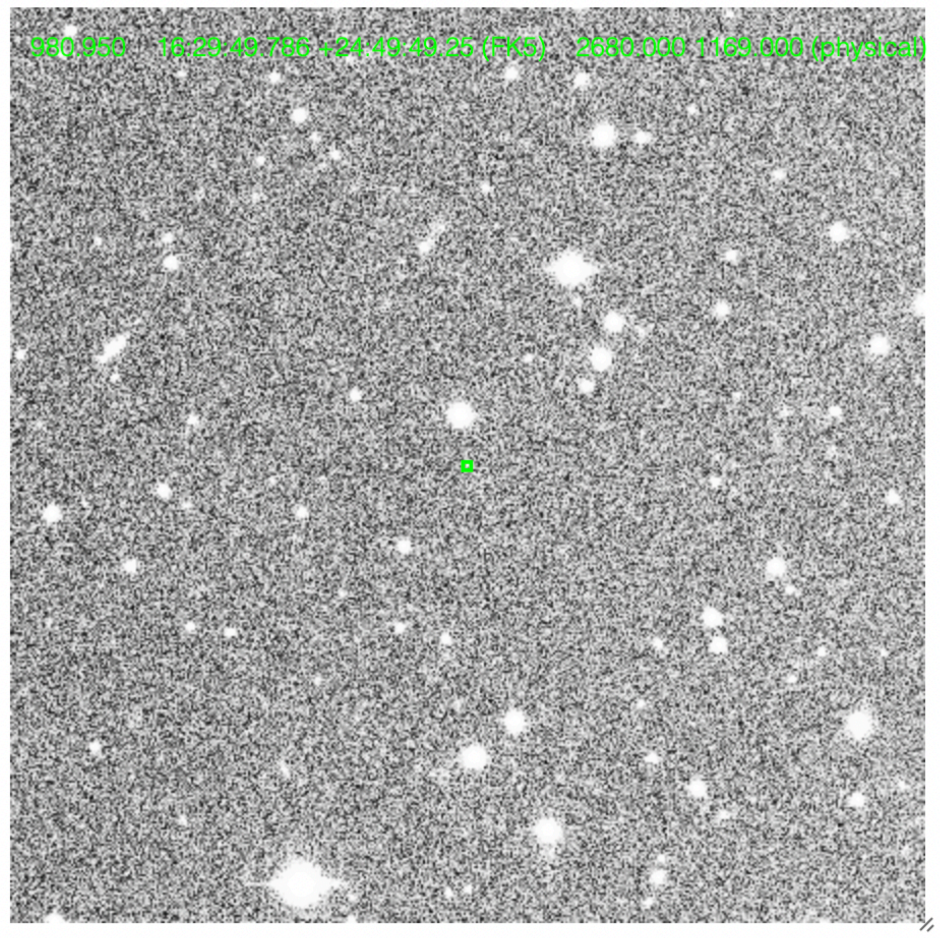
"Short" light curve: 58194.0 ≤ MJD ≤ 58972.0

Full light curve Folded light curve

Closest Antares object, diff-photometry Closest Pan-STARRS object, apparent

Magnitude Flux diff Magnitude diff Flux

Summary



Open in JS9 Download FITS Product directory

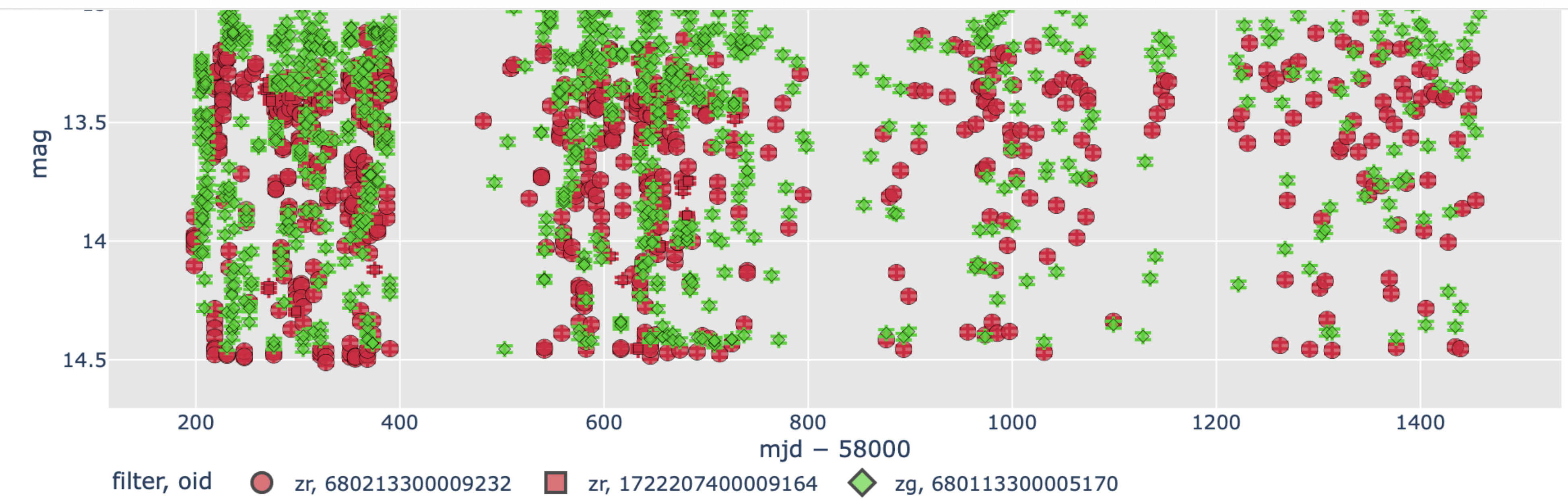
ZTF science image for any detection

Aladin

By-product: SNAD ZTF viewer

https://ztf.snad.space

ztf.snad.space/dr8/view/680213300009232



Download PNG, PDF, CSV

"Short" light curve: 58194.0 ≤ MJD ≤ 58972.0

Full light curve   Folded light curve

Closest Antares object, diff-photometry   Closest Pan-STARRS object, apparent

Magnitude   Flux   diff Magnitude   diff Flux

Summary

Name: ZTF18aabpzic (0.266″ Alerce), ZTF18aabpzic (0.353″ Fink), J254.4575+35.3423 (0.124″ ATLAS), 1338822021487330304 (0.115″ Gaia EDR3 Distances), HZ Her (0.711″ GCVS), PSO J254.4575+35.3423 (0.109″ Pan-STARRS DR2 Stacked), V* HZ Her (0.081″ Simbad), 15037 (0.720″ VSX), ZTFJ165749.81+352032.4 (0.124″ ZTF Periodic)

Type: LMXB (0.353″ Fink), IRR (0.124″ ATLAS), XPR+E (0.711″ GCVS), LowMassXBin (0.081″ Simbad), LMXB:/XPR+E (0.720″ VSX), EW (0.124″ ZTF Periodic)

Period, days: 1.700 (periodogram S/N=78.620), 1.700 (0.124″ ATLAS), 1.700 (0.711″ GCVS), 1.700 (0.081″ Simbad), 34.875 (0.720″ VSX), 3.400 (0.124″ ZTF Periodic)

Distance: 7.00 kpc (0.115″ Gaia EDR3 Distances), 6.60 kpc (0.081″ Simbad)

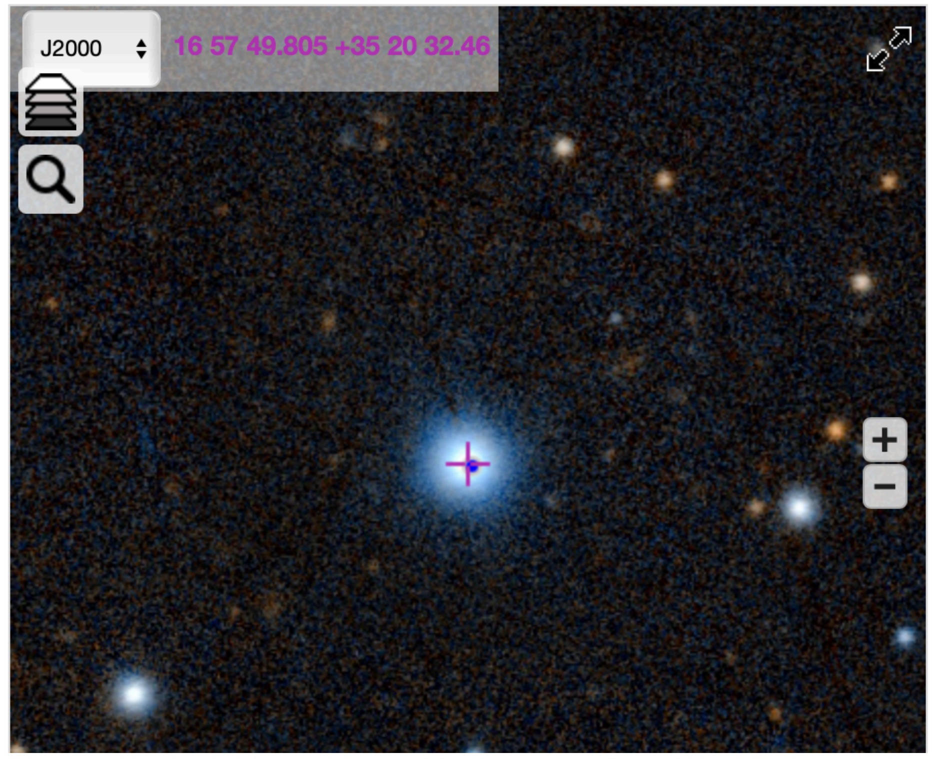
Average mag (including neighbourhood): zg 13.55, zr 13.68, (zg–zr) -0.13

Extinction: SFD E(B-V) = 0.01, Bayestar & Gaia EDR distance Ag = 0.07 Ar = 0.05 Ai = 0.03

Search in brokers: ALeRCE, Antares, Fink, MARS

Coordinates: Eq 254.45752 35.34235, Gal 58.149 37.5231

Aladin



Name, type, period, distance & extension from other catalogs and our periodogram

By-product: SNAD ZTF viewer

https://ztf.snad.space

Home page

ZTF object ID / SNAD ID

Eq coordinates / common name

By-product: SNAD ZTF viewer

https://ztf.snad.space

ztf.snad.space/dr8/search/hz%20her/1

Konstantin Malanchev

SNAD ZTF DR8 object viewer

OID E.g. 633207400004730 GO

Coordinates hz her radius (arcsec) 1 GO

Objects inside cone (254.45755 deg, 35.34236 deg), r = 1.0″

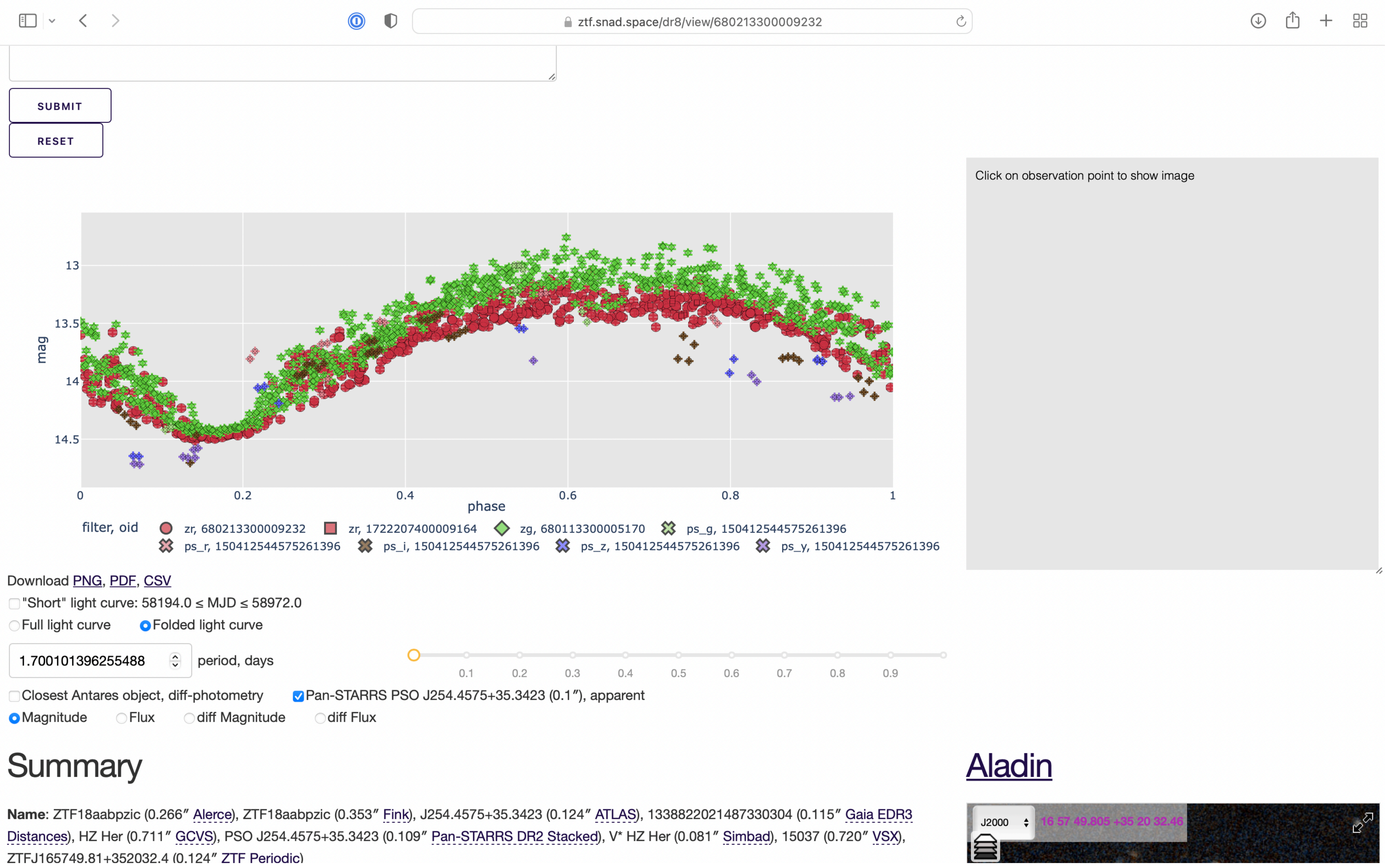
Same source, different OIDs

OID	separation, arcsec	filter	Number of "good" observations	Duration, days
1722207400009164	0.069	zr	43	459.824
680113300005170	0.078	zg	734	1253.823
680213300009232	0.082	zr	686	1255.902

By-product: SNAD ZTF viewer

https://ztf.snad.space

Period folding and third-party photometry (Pan-STARRS)

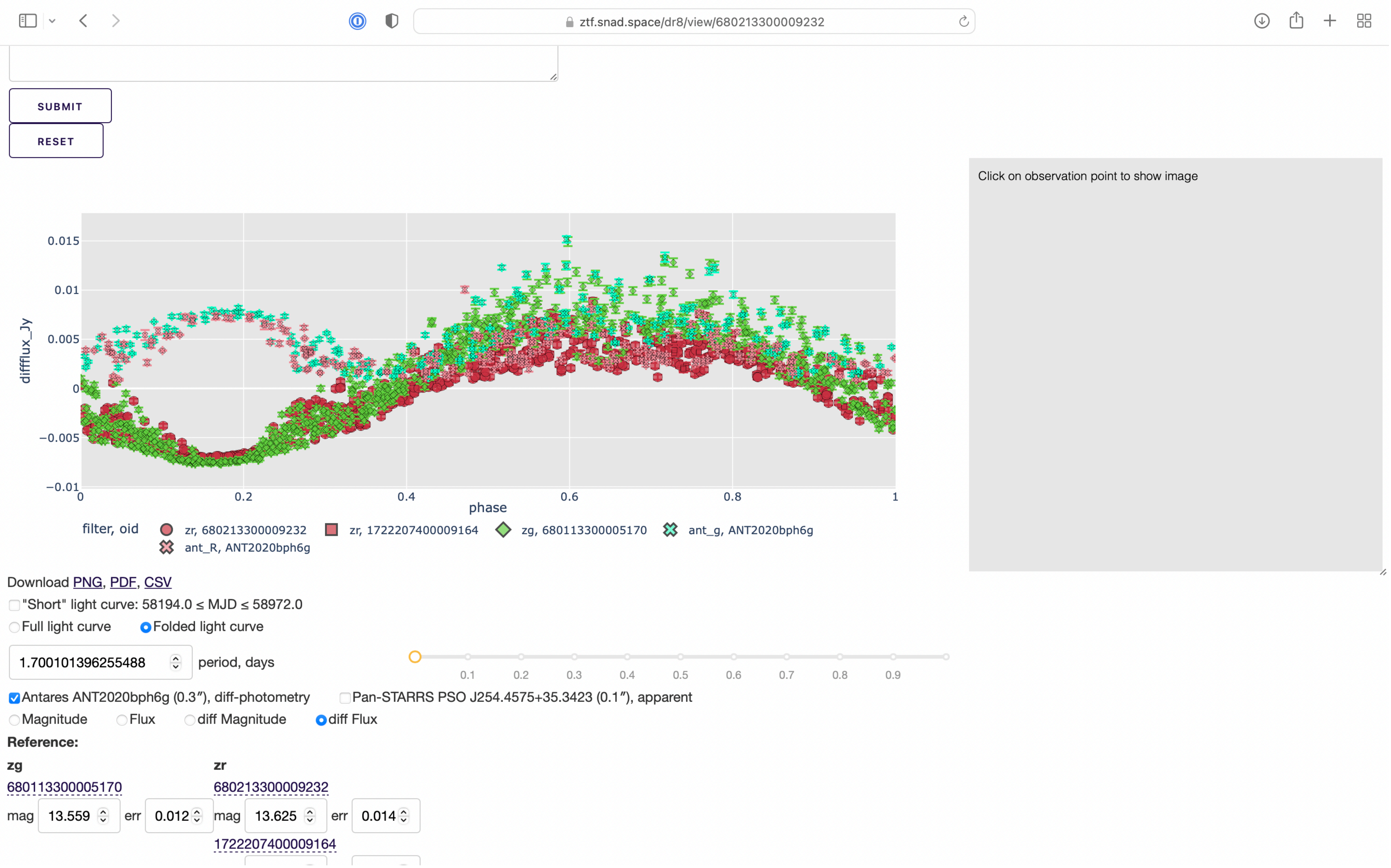


X

By-product: SNAD ZTF viewer

https://ztf.snad.space

Period folding and third-party diff photometry (ZTF alert stream from Antares broker)



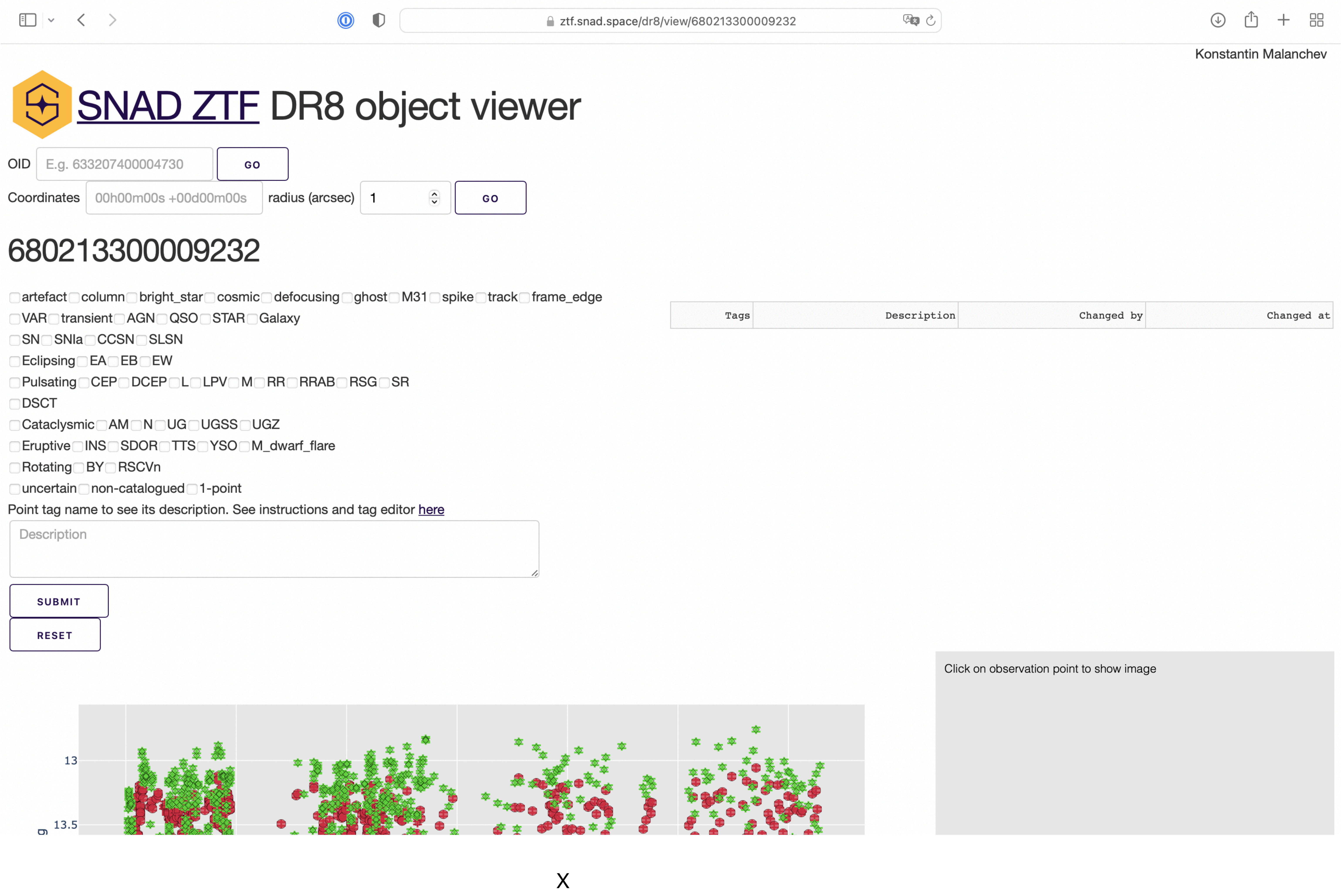
X

By-product: SNAD ZTF viewer

https://ztf.snad.space

SNAD

Tags and description DB frontend



By-product: SNAD ZTF viewer

https://ztf.snad.space

SNAD experts tagged >2000 objects, >70 are submitted to the TNS!

ztf.snad.space/akb

Konstantin Malanchev

SNAD ZTF DR8 object viewer

OID E.g. 633207400004730 GO

Coordinates 00h00m00s +00d00m00s radius (arcsec) 1 GO

Anomaly knowledge base

OID	Tags	Description	Changed by	Changed at
filter data...		SNAD		
633207400004730	SN, uncertain	SNAD101	maria	2021-08-02T07:46:53.429000+00:00
633216300024691	SN, uncertain	SNAD102	maria	2021-08-02T07:47:54.227000+00:00
634108100006647	AGN, SN, uncertain	SNAD158	maria	2021-10-21T22:22:11.362000+00:00
643105300009229	AGN, SN, uncertain	SNAD153	maria	2021-10-21T21:39:27.291000+00:00
676212400013135	SN, uncertain	SNAD122	maria	2021-08-02T07:52:47.557000+00:00
679108100003227	SNIa, uncertain, non-catalogued	photo-z of host: 0.303 +/- 0.116... Possible absolute mag between -20.6 and -22.6. SLSN? Too bright for SN Ia... SNAD150	patrick	2021-10-21T14:19:32.575000+00:00
680109100003419	SN	SNAD168, PCA+ k-D tree	maria	2021-11-12T20:32:32.823000+00:00
682102200004200	SN, uncertain	SNAD176	maria	2022-03-03T14:49:49.398000+00:00
682209200018910	SN, uncertain	SNAD143	maria	2021-08-02T09:48:05.990000+00:00
684215200016923	SN, uncertain, non-catalogued	SNAD157	maria	2021-10-21T22:17:58.390000+00:00
692106300027877	SN, uncertain	SNAD174	novinskaya	2022-02-28T15:02:40.405000+00:00
718205300006523	AGN, uncertain, non-catalogued	SNAD155	maria	2021-10-22T09:30:51.131000+00:00
719202100004008	AGN, SN, uncertain	SNAD154	maria	2021-10-21T22:06:23.389000+00:00
720209400014960	SN, uncertain	SNAD123	maria	2021-08-02T10:06:24.720000+00:00
721210100012349	SN, uncertain	SNAD129	maria	2021-08-02T10:47:48.325000+00:00

Conclusion

SИAD

- Using real data from the very beginning
- Astronomical experts are queens: their opinion matters from the start of the algorithm construction to the last stage
- Developing new tools — and sharing them with the community
- Recent and ongoing projects:
  - Developing new active anomaly detection algorithm for new features, better computation and detection performance (Korolev+ in prep., ask me about it!)
  - Using AAD for classification, listen **talk by Emille Ishida** about SNe
  - Mining transients with k-D tree, see **poster by Patrick Aleo** (presented by me)