Co-funded by the European Union

I FEEL SLOVENIA

SMASH machine learning for science and humanities postdoctoral program

TVSSC

Early classification of Tidal Disruption Events with Machine Learning

Razim O., Umuzigazuba E., Contardo G., Bhattacharyya S., Concepcion E., Gomboc A., Petrushevska T.

University of Nova Gorica, Slovenia

2nd SMASHING workshop
Vipava, Slovenia, 19.10.2026

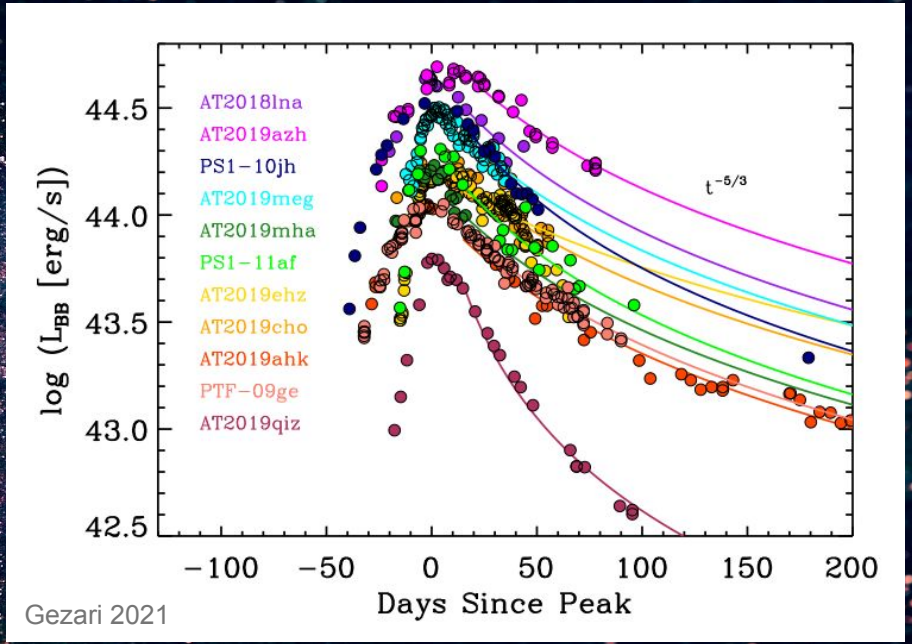
Tidal Disruption Events

- A radiation burst when a star is being disrupted by a black hole and an accretion disk is formed
- Mostly detected in X-ray, UV, optical; radio and IR observed (sometimes)with a delay
- About 200 known events, mostly discovered with ZTF and ASASN
- Intrinsically rare (~2 orders of magnitude less common than SN Ia)



Tidal Disruption Events

- Easy to confuse with SNe or AGN flare (and TDE can happen in an AGN too)
- Potentially useful for studying black hole population



Gezari 2021

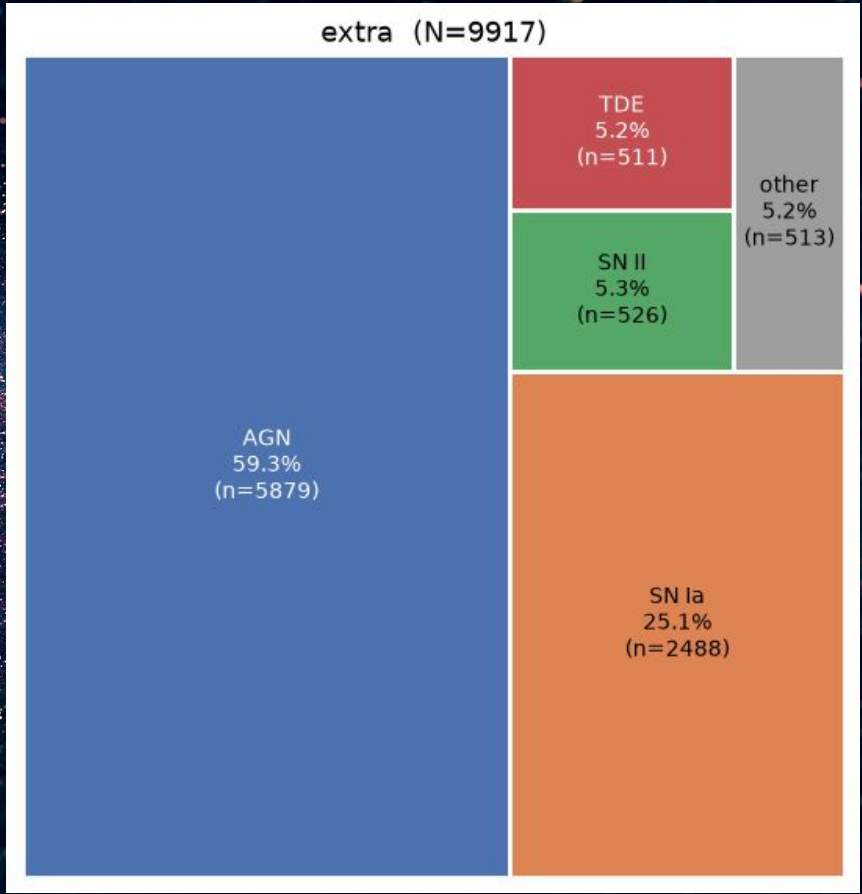
Tidal Disruption Events

- LSST *was* expected to see ~10^3 TDEs/year (re-evaluation needed for newer cadence strategies)...
- ...but class imbalance is brutal.



MALLORN Challenge

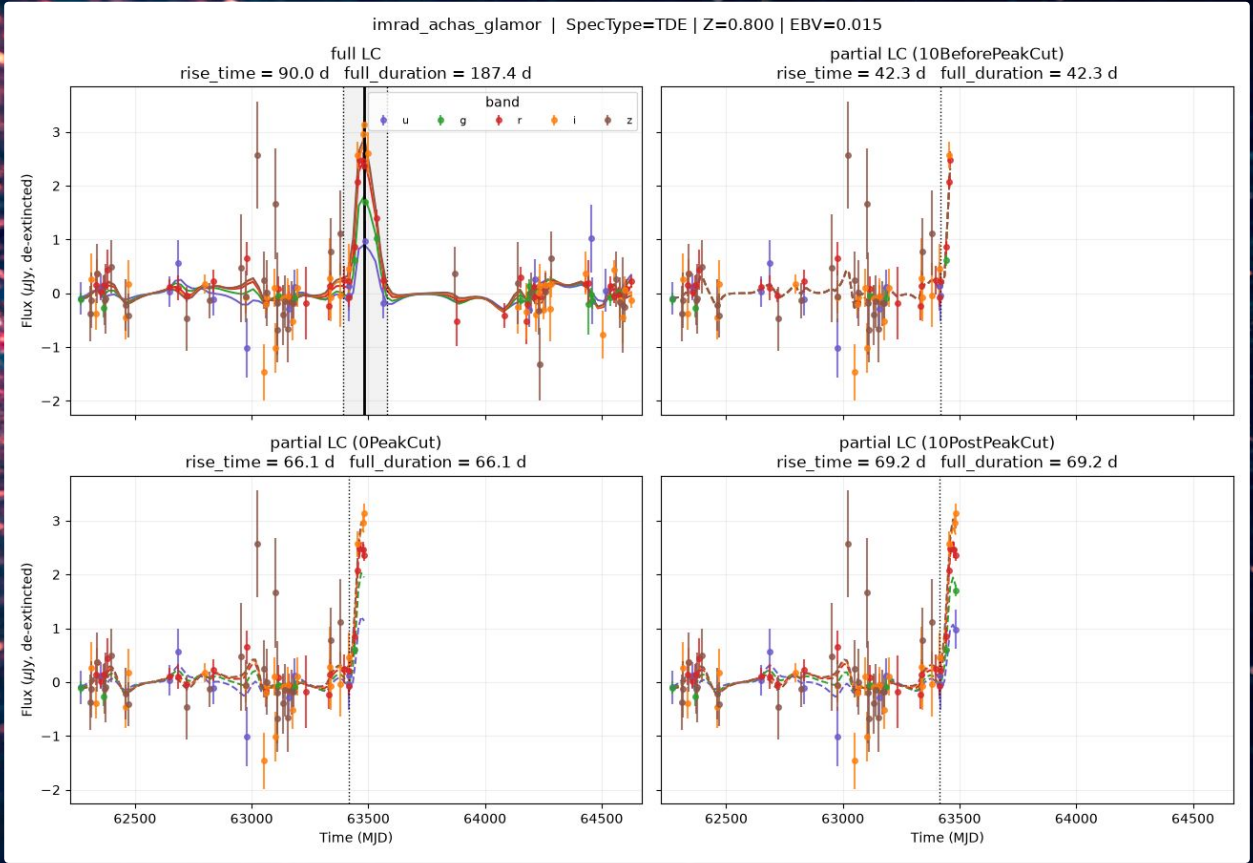
- Kaggle challenge for nuclear transient classification that took place in late 2025;
- 3043 (Kaggle sample) + 9917 (Extra sample, private communication with the MALLORN team) LSST-like ugrizy synthetic lightcurves of nuclear transients made from real ZTF events with spectroscopic labels, including AGNs, SNe Ia, SNe II, SLSNe, peculiar SNe, and TDEs



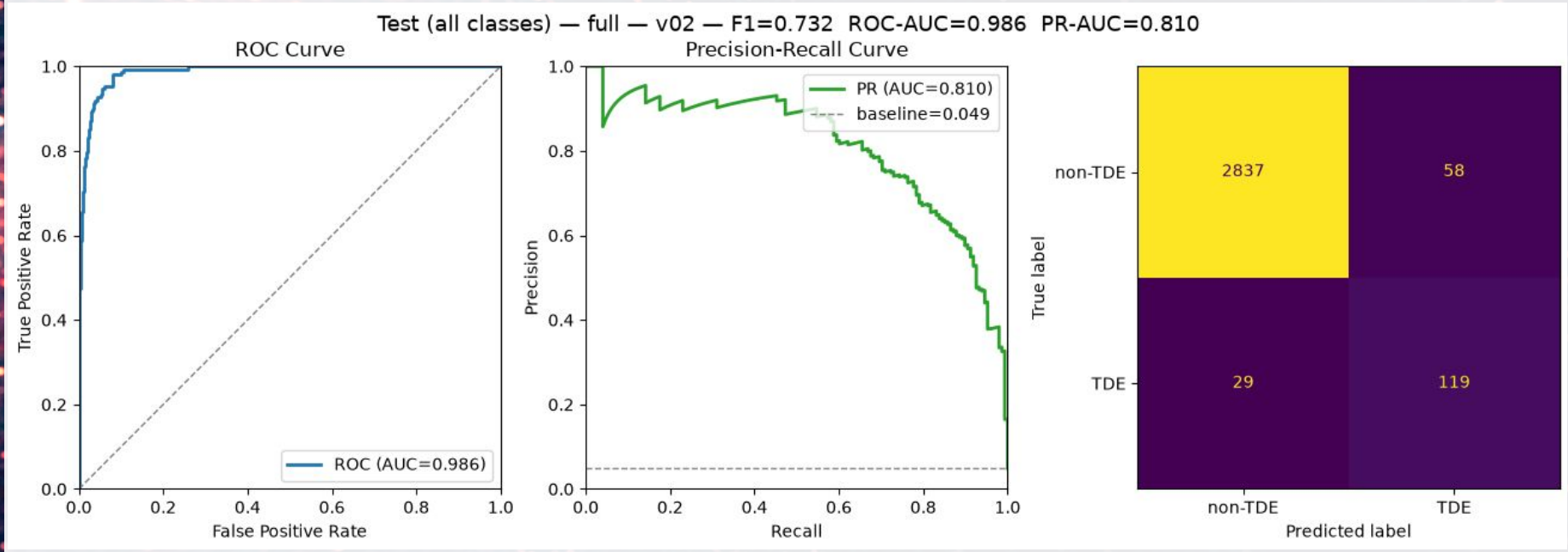
ML setup

- LightGBM classifier, trained with stratified K-fold cross-validation, with conditional permutation importance feature selection;
- over 500 features originally from several families: basic statistics, Bazin function fit parameters, Potica model fit parameters (in-house model), Gaussian Process fit parameters (done with Redback), Mexican-hat power-spectrum features (MHPS), statistics in- and out of event window;
- ~60 features after the feature selection;
- F1~0.66-0.7 on full LCs (top 20 of Kaggle leaderboard).

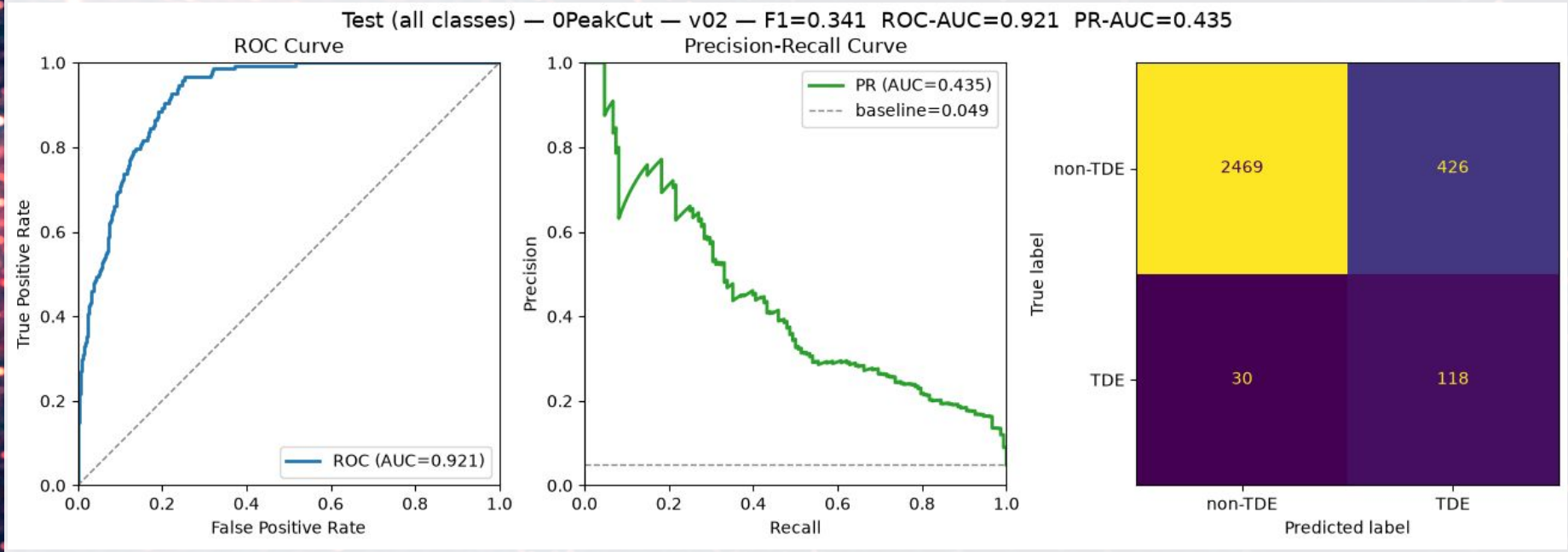
Real-world scenario: partial LCs



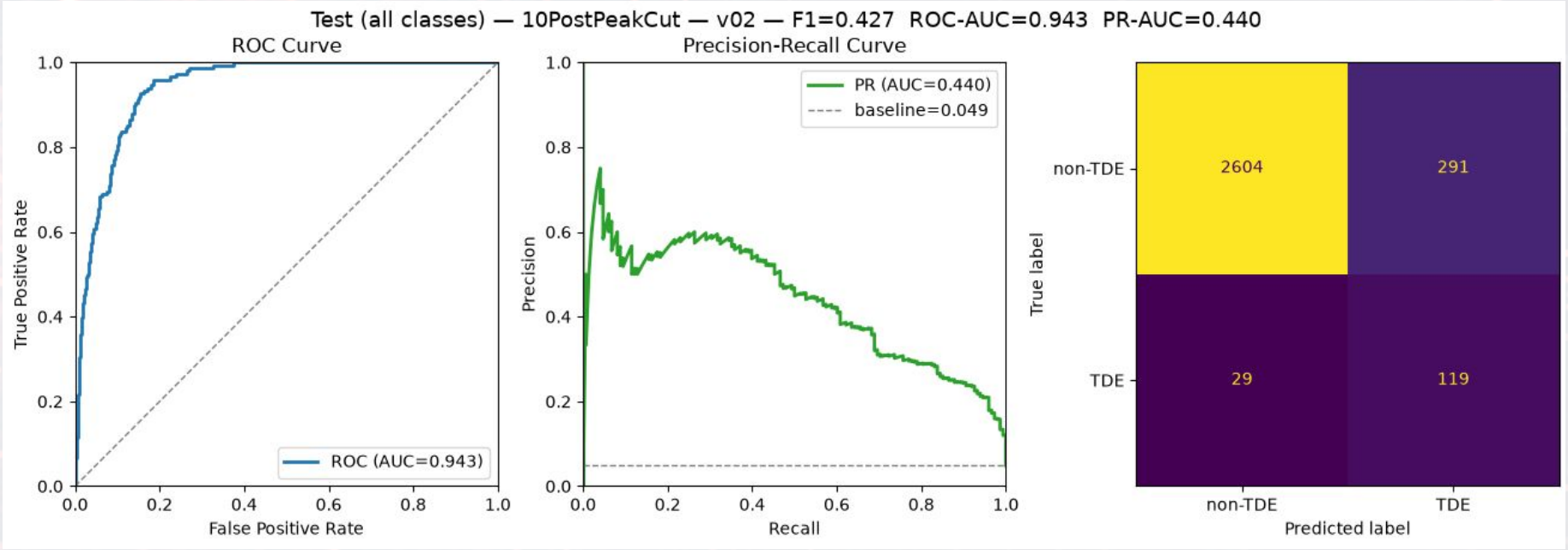
ML classification performance on full and partial LCs: fixed FN fraction!



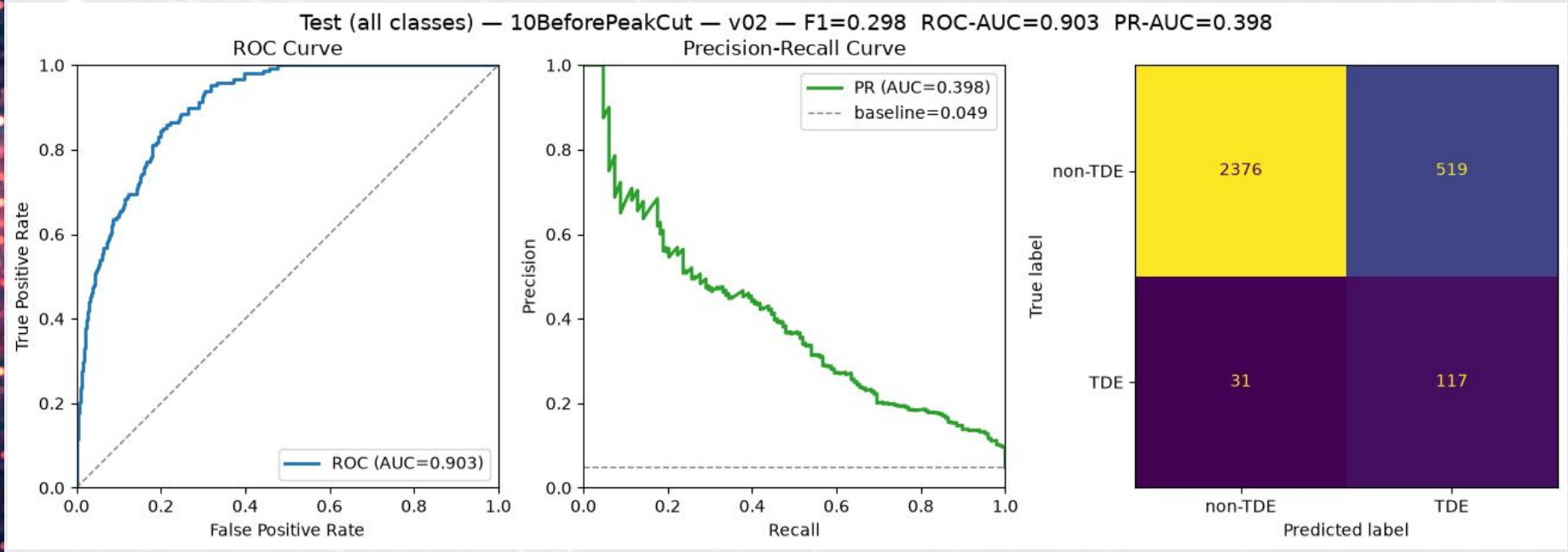
ML classification performance on full and partial LCs: fixed FN fraction!



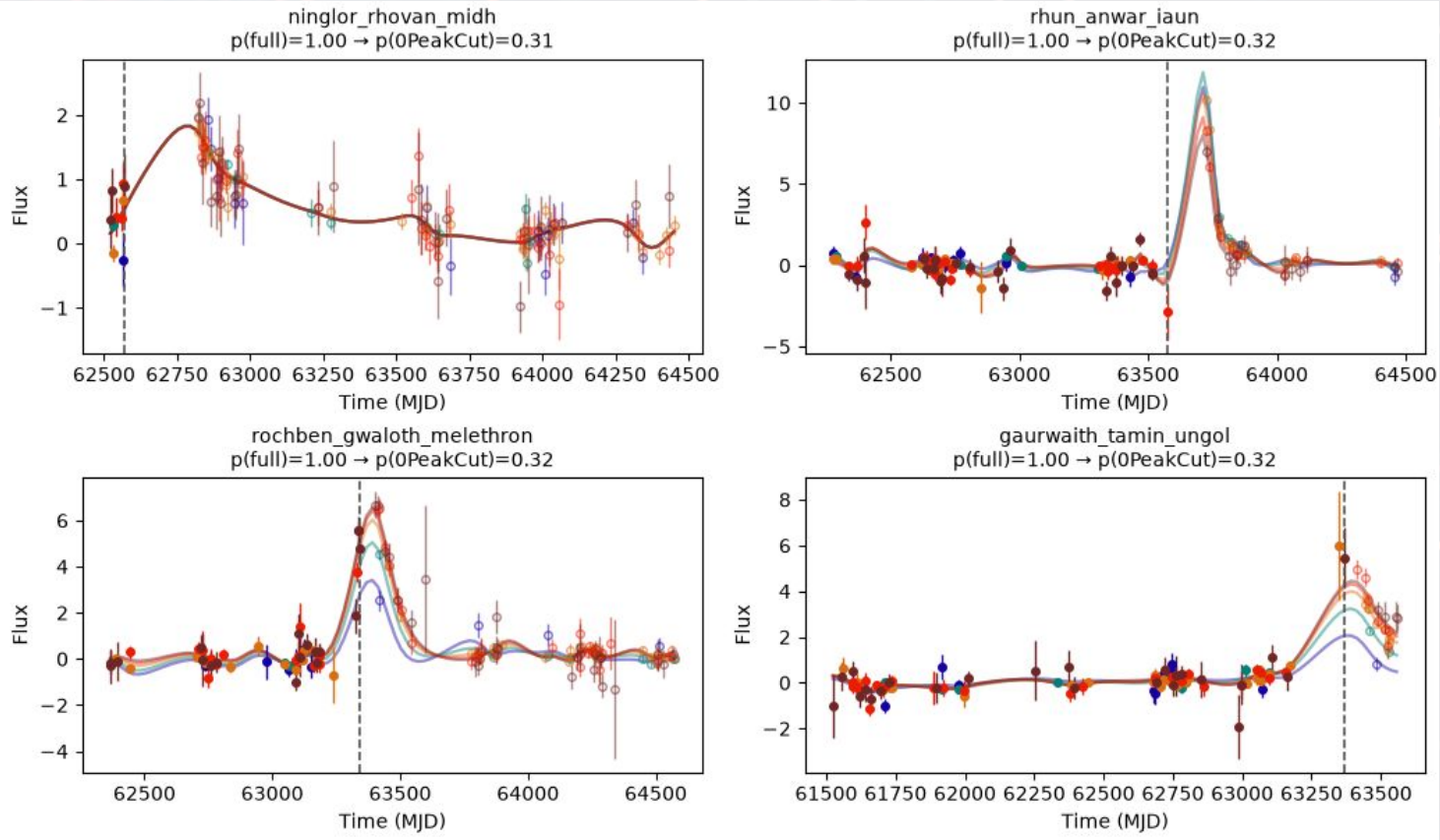
ML classification performance on full and partial LCs: fixed FN fraction!



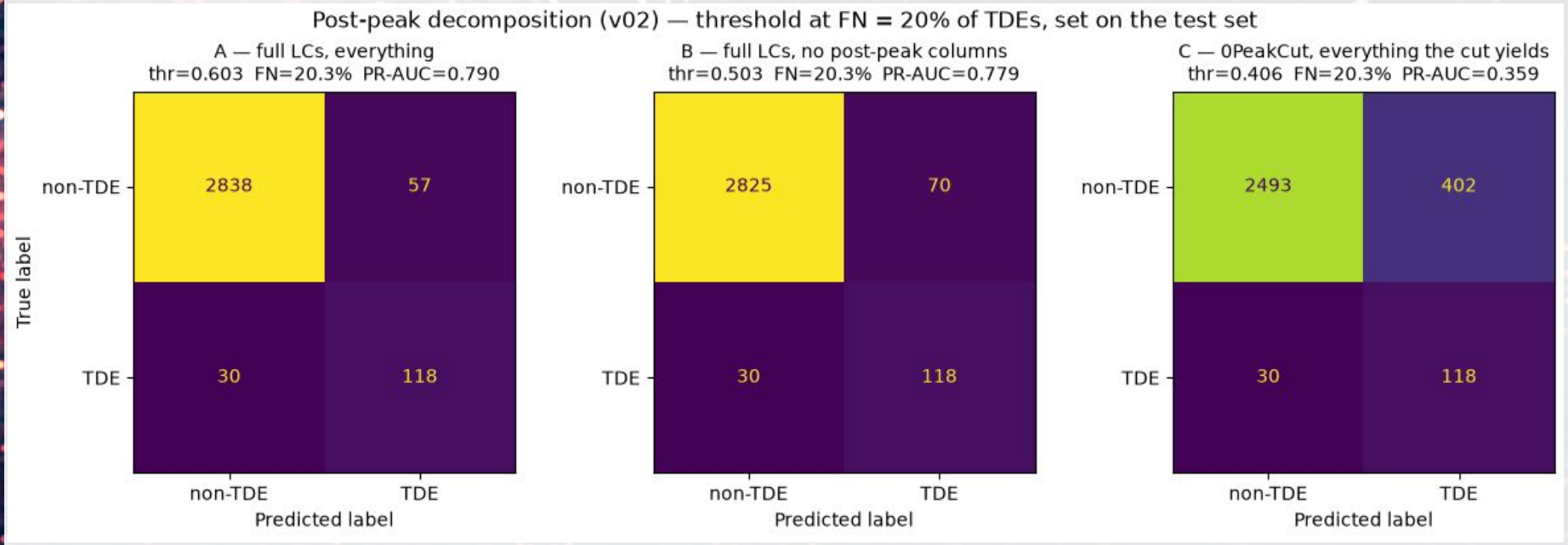
ML classification performance on full and partial LCs: fixed FN fraction!



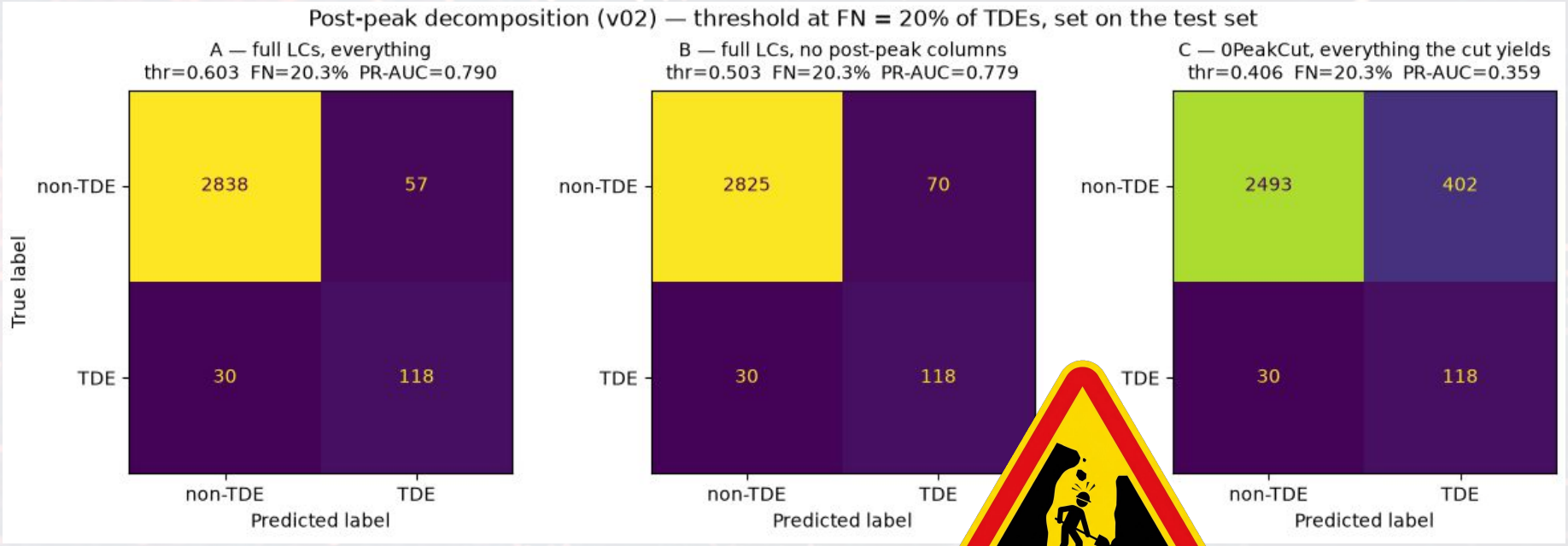
What ‘cut at peak’ actually means (spoiler: not what you would expect)



What if we get rid of post-peak features in the full LCs dataset?..



What if we get rid of post-peak features in the full LCs dataset?..



Takeaway

- For partial LCs, the performance goes from F1~0.67 to F1~0.35-0.4, but…
- …F1 isn't a good metric for us. Instead, we fix the fraction of missed TDEs (e.g. =0.2), and compare **the fraction of contaminants** at this threshold.
- The deterioration comes from both degenerate features (such as post-peak colors) and low quality of other features (such as rise/full duration of event);
- Main contaminating class switches from AGN to SN Ia when we move from full to partial LCs. **FP/TDE_count rises from ~0.4 to 3.7** (fractions units, not percents!)
- For full LCs, main contaminants are AGNs. For partial LCs, SNe dominate FPs;
- Without additional data, we don't know if the performance deterioration is inevitable and comes from genuine absence of discerning features that can be obtained from pre-peak optical LCs, or it is the result of insufficient representativeness of the training sample and/or insufficient cadence…
- But it seems like high-quality pre-peak features might be enough for a good classification (preliminary, WIP) ->
- Not only spectroscopic, but also **photometric follow-up observations of rising TDE candidates are needed**.