Quasar classification and redshift estimation

Machine Learning at the ICCUB

Ignasi Pérez-Ràfols
Universitat Politècnica de Catalunya
Departament de Física

Outline

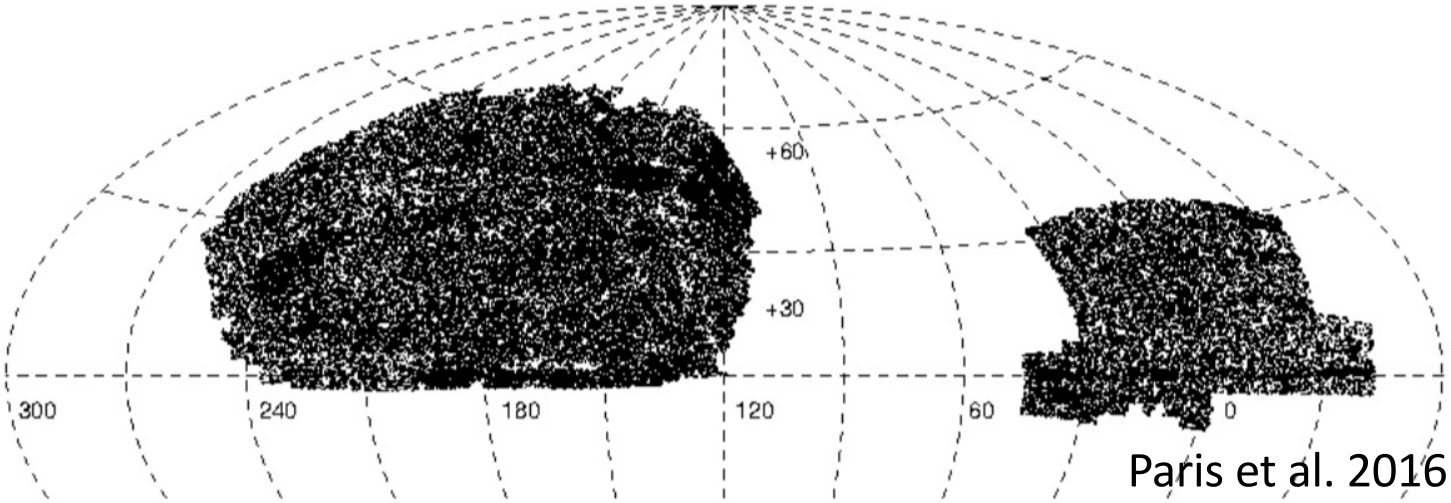
- The problem
- Machine learning strategy
- SQUEzE overview
- Results
  - Performance
  - Flexibility
  - Explainability
- Combining with other algorithms
- Conclusions

The problem

- Science goals: study the spectra of quasars
  - BAO to measure the structure of the Universe
  - Intergalactic and Circumgalactic Physics
  - Quasar Science

- Current surveys are producing hundreds of thousands of spectra. We need to confirm which spectra are those of quasars
- Future surveys will only increase these numbers.

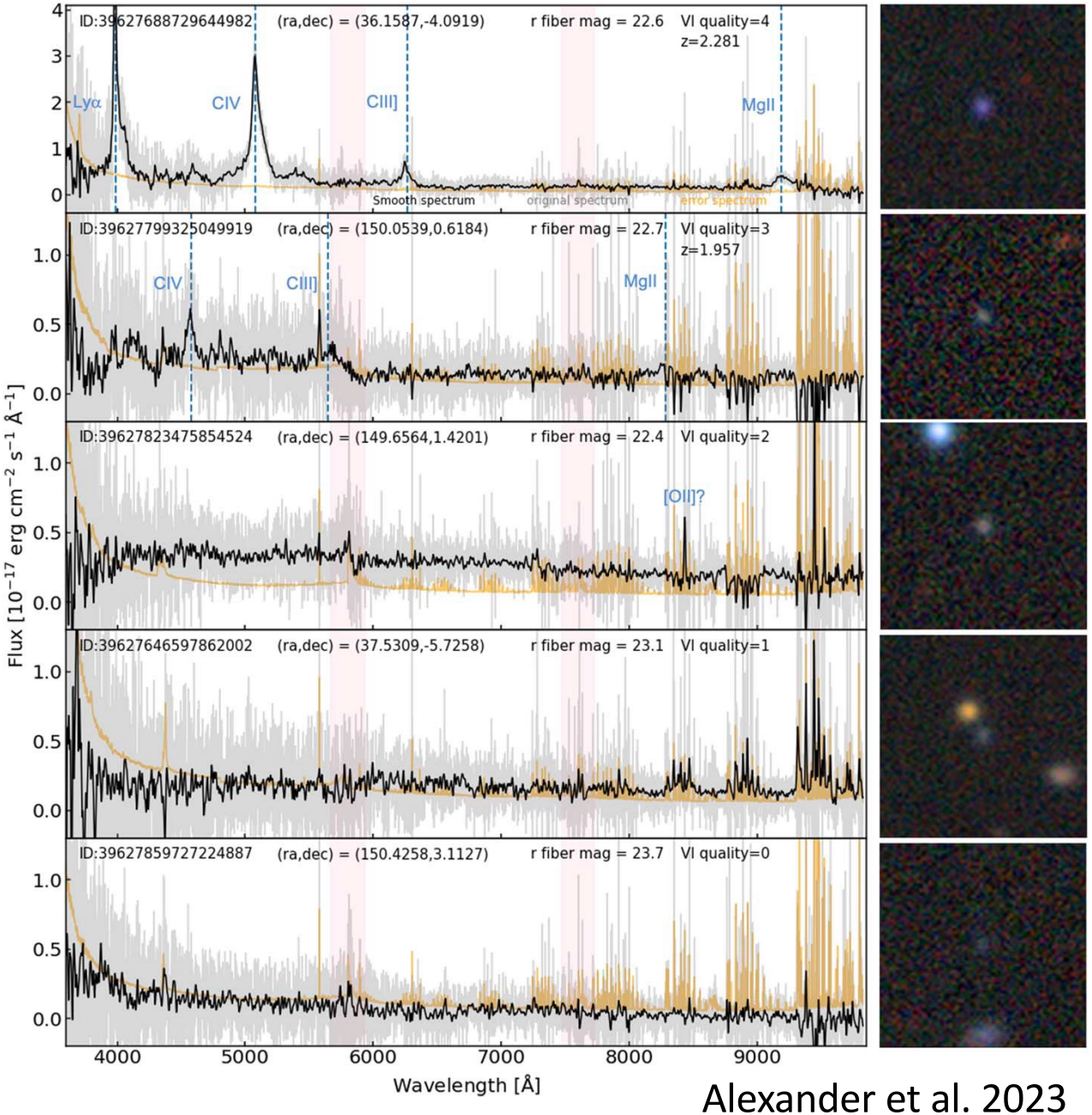
Example: BOSS DR12 → 546 856 objects visually inspected → 297 301 quasars



Paris et al. 2016

The problem

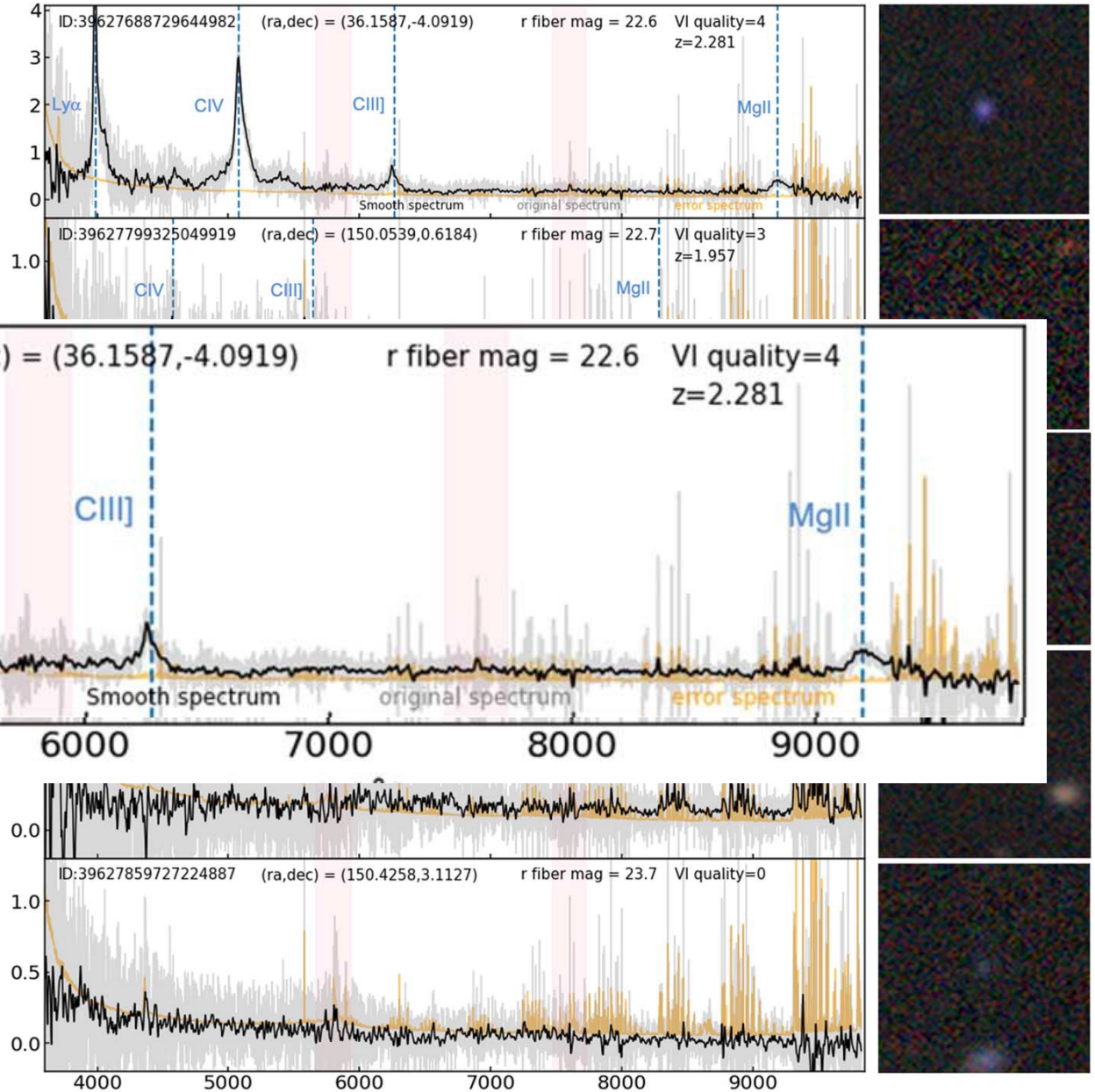
- How do we know if we have a quasar?



Alexander et al. 2023

The problem

- How do we know if we have a quasar?
- Visual inspection:

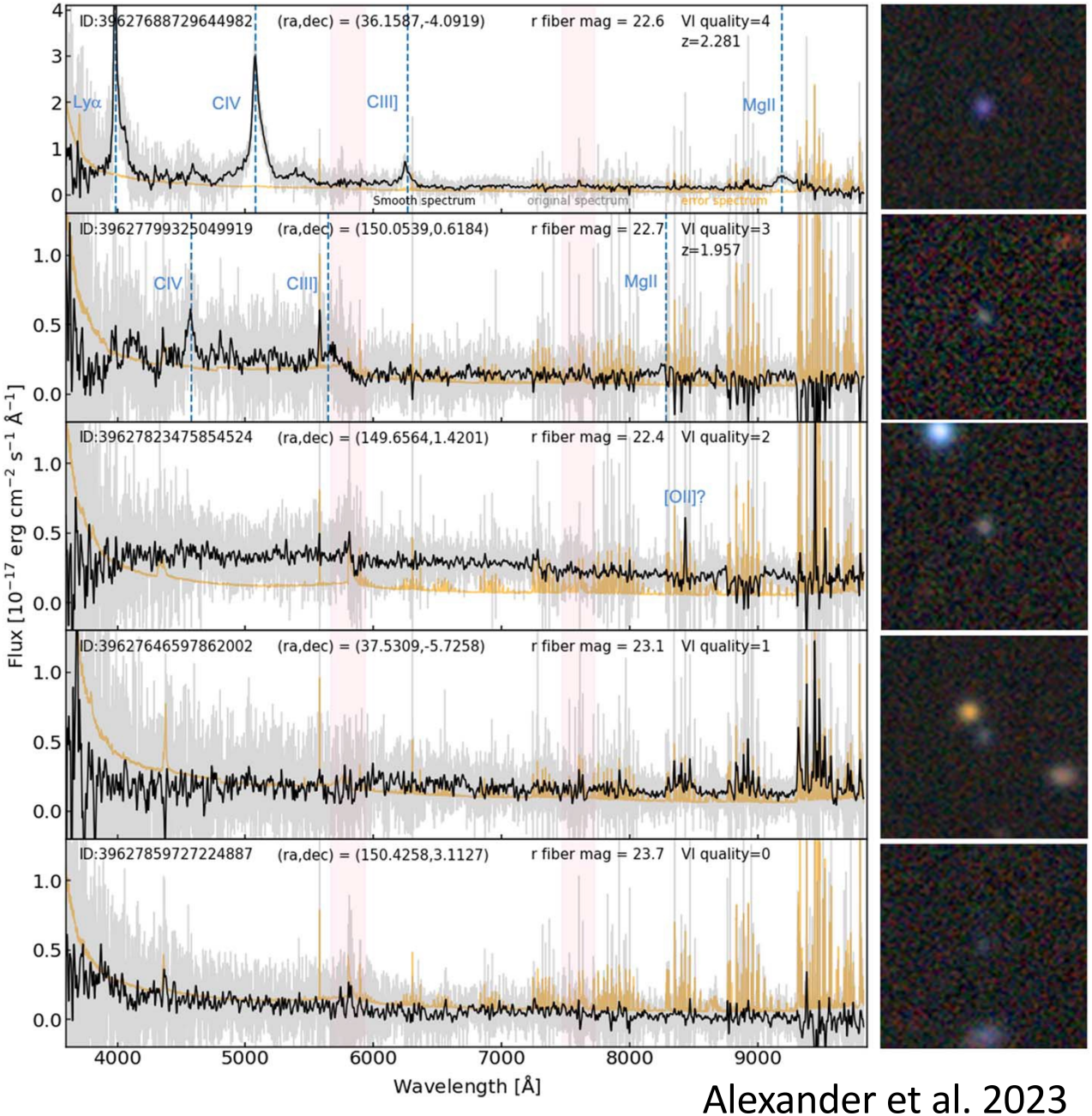


- Outputs of the visual inspection:
  - classification
  - redshift

Alexander et al. 2023

The problem

- How do we know if we have a quasar?
- Visual inspection:



- Outputs of the visual inspection:
  - classification
  - redshift

Alexander et al. 2023

The problem

- Approaches to the identification
  - ~~Human visual inspection~~
  - Template-based fitting
  - Machine learning

- Past and current surveys:
  - SDSS-III (BOSS) and SDSS-IV (eBOSS)
  - DESI
  - WEAVE
  - JPAS

Machine learning strategy

- Results wanted:
  - Classification
  - Redshifts

- Requirements:
  - Classification + regression
  - Different surveys → versatility
  - Explainability

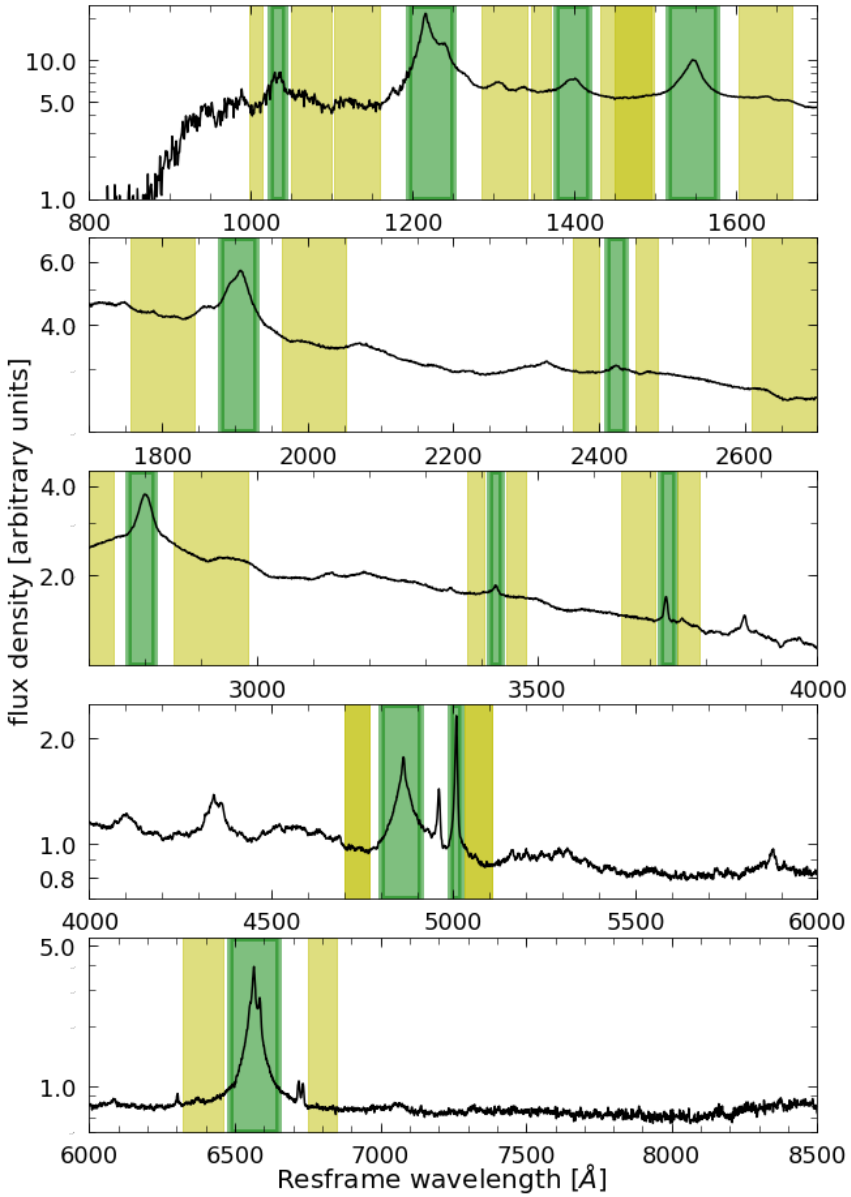
SQUEzE overview

- Mimic the visual inspection procedure
- 3 steps
  1. Peak identification
  2. Metric computation
  3. Classification

$$\text{line ratio: } l_i = \frac{2p_i}{b_i + r_i},$$

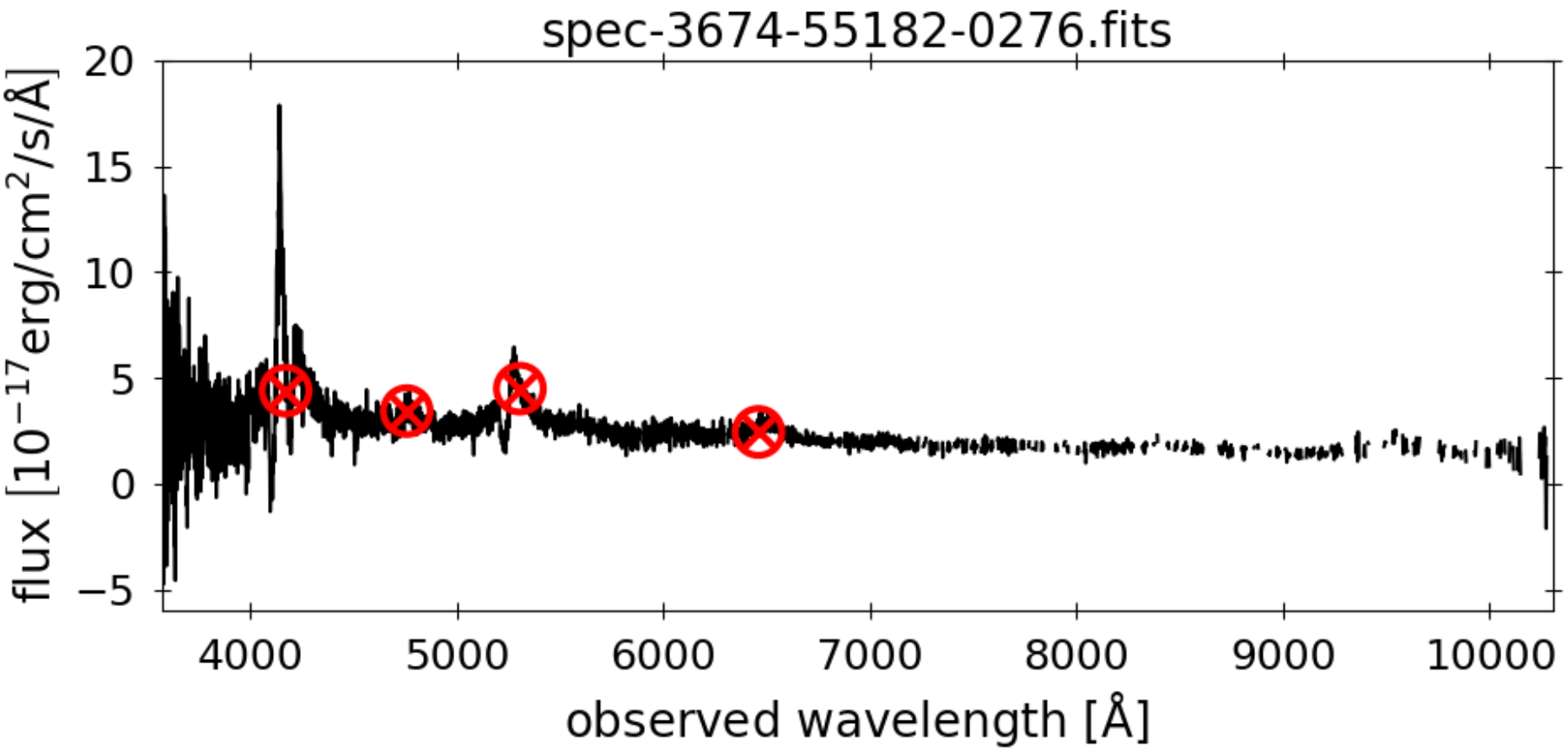
$$\text{line contrast ratio: } c_i = (l_i - 1)/e_i,$$

$$\text{line continuum slope: } s_i = \left| \frac{r_i - b_i}{r_i + b_i} \right|.$$



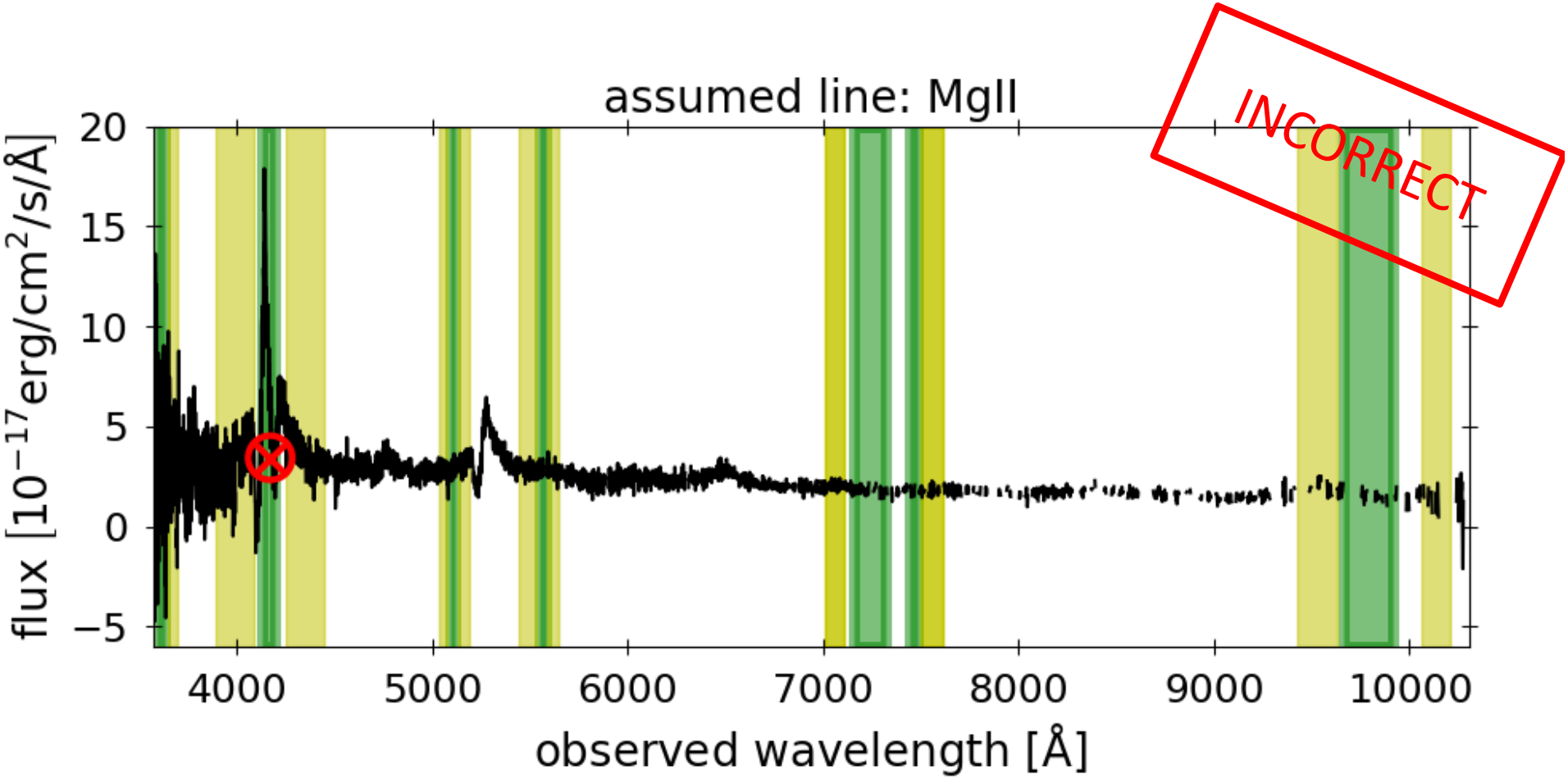
Pérez-Ràfols et al. 2020b

SQUEzE overview



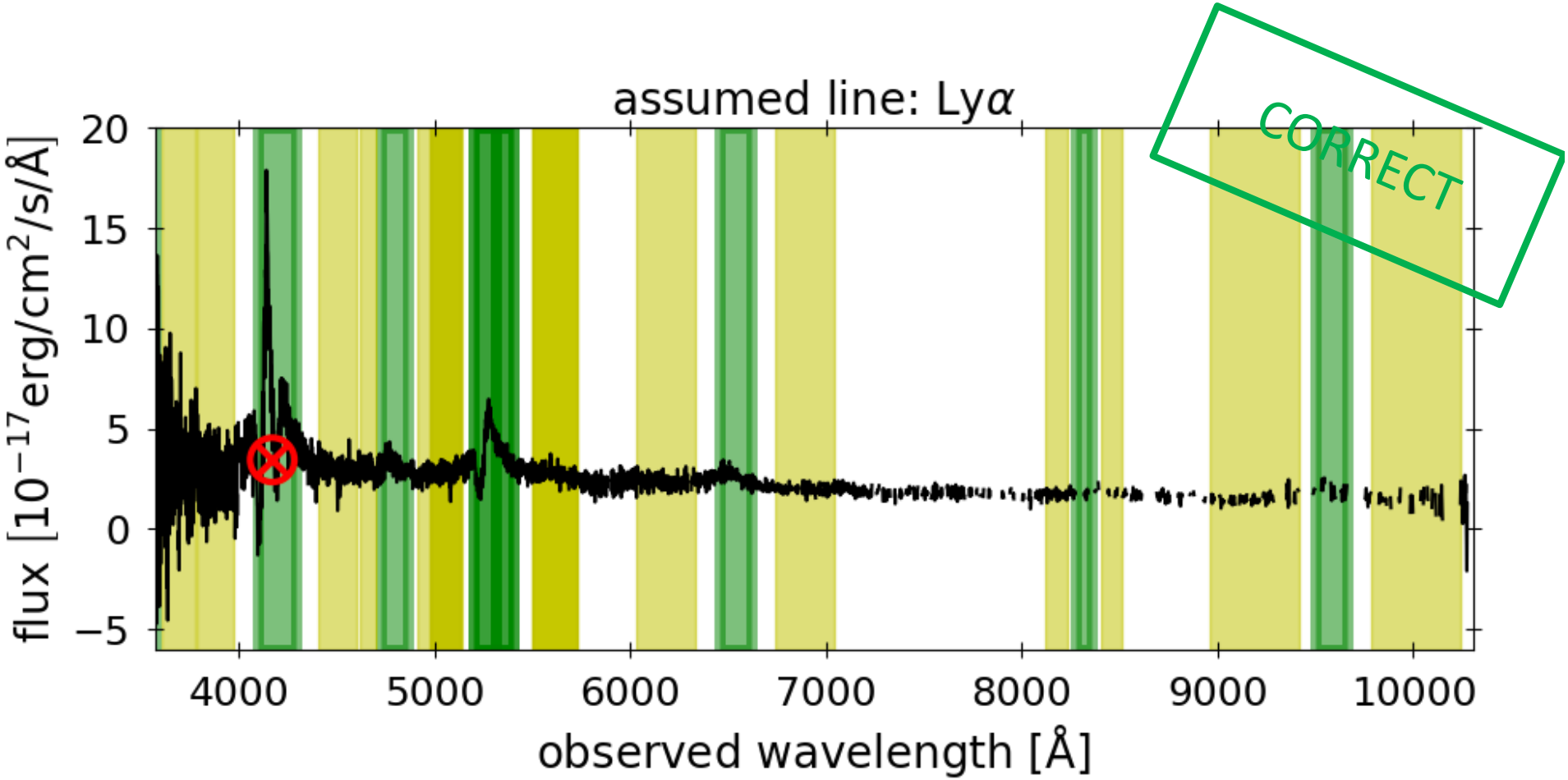
Pérez-Ràfols et al. 2020b

SQUEzE overview



Pérez-Ràfols et al. 2020b

SQUEzE overview



Pérez-Ràfols et al. 2020b

SQUEzE performance

- Definition of correct:
  - Spectra is indeed a quasar
  - The estimated redshift is correct

- Purity (fraction of correctly classified quasars)

$$p = \frac{TP}{TP + FP}$$

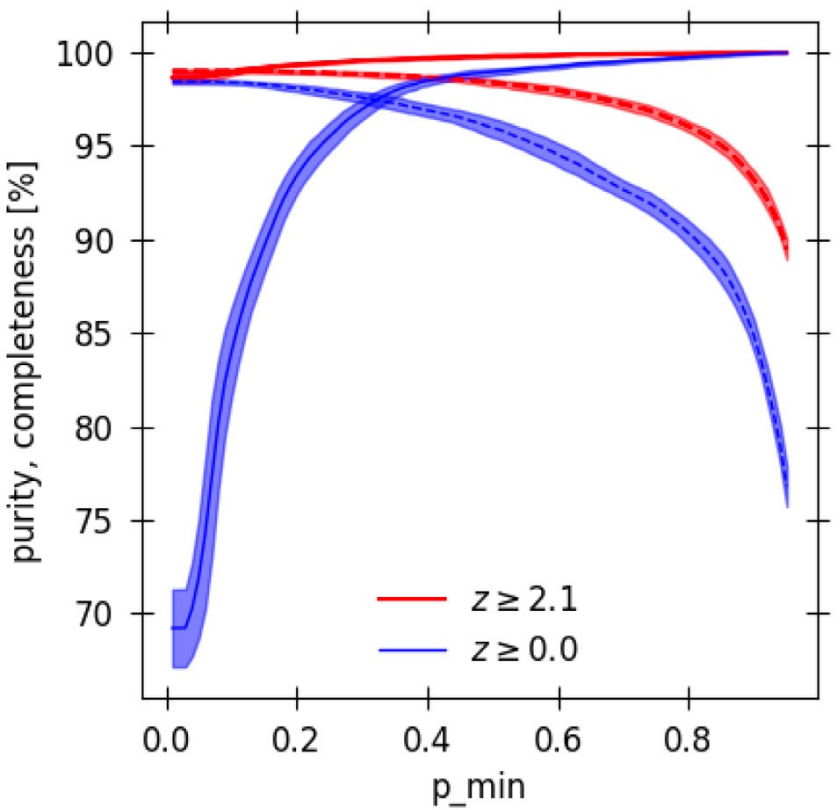
- Completeness (fraction of found quasars)

$$c = \frac{TP}{TP + FN}$$

- f1 score (sweet spot)

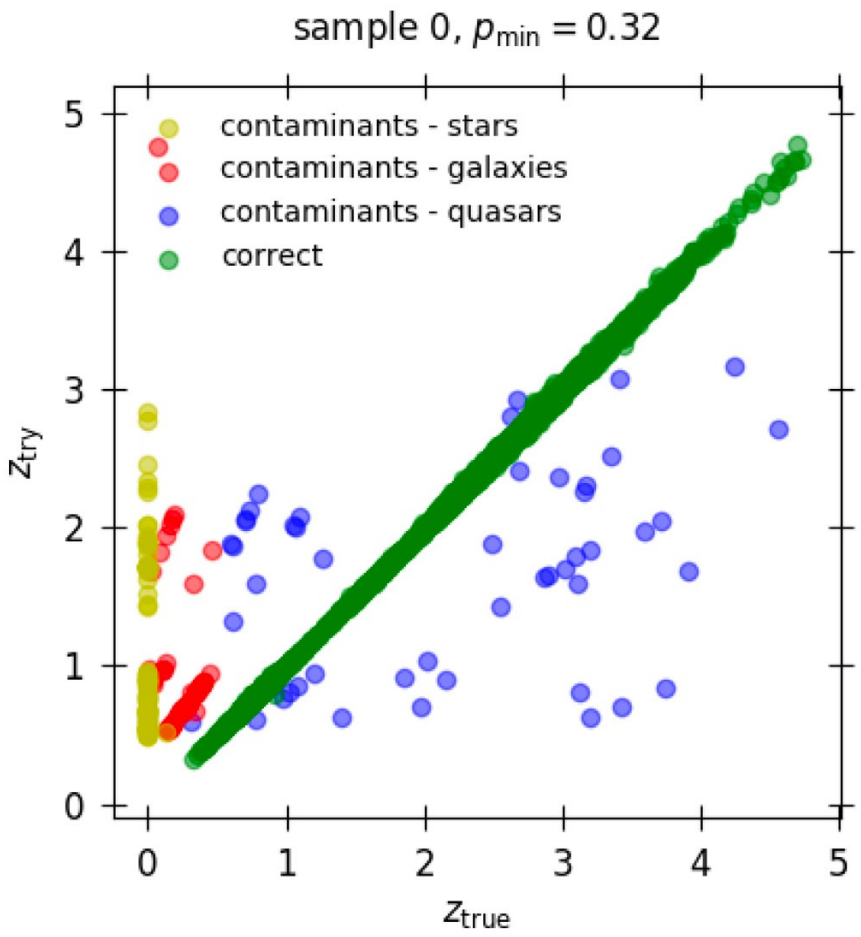
$$f_1 = \frac{2p\,c}{p + c}$$

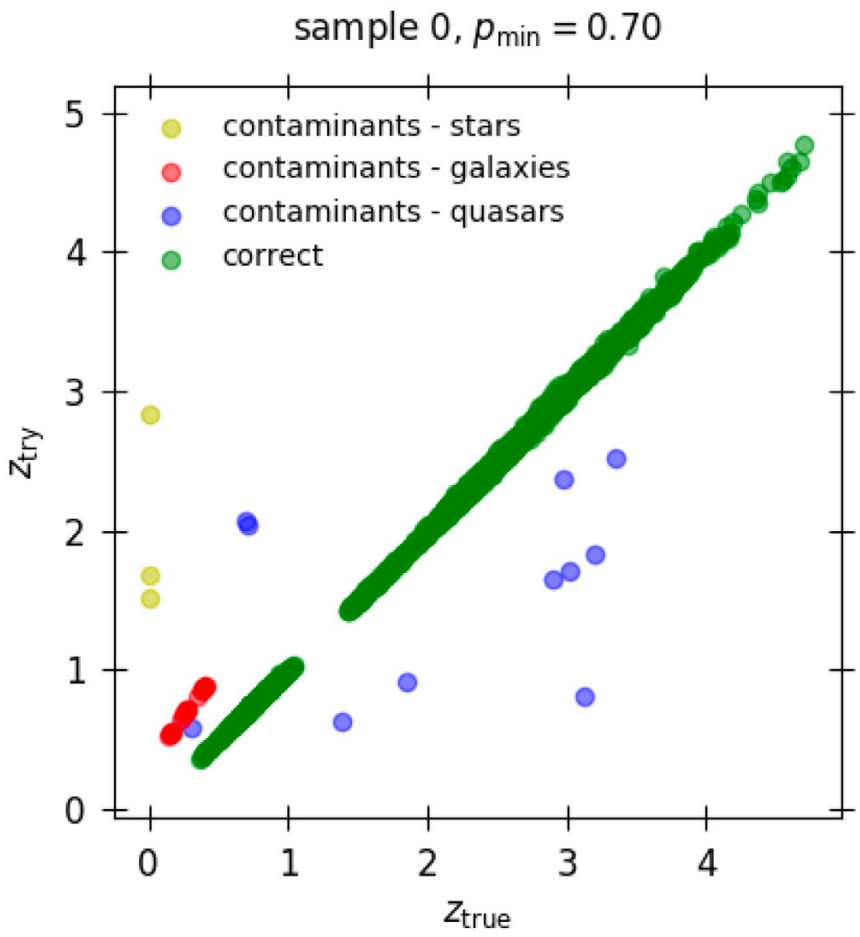
Performance on SDSS data



Pérez-Ràfols et al. 2020b

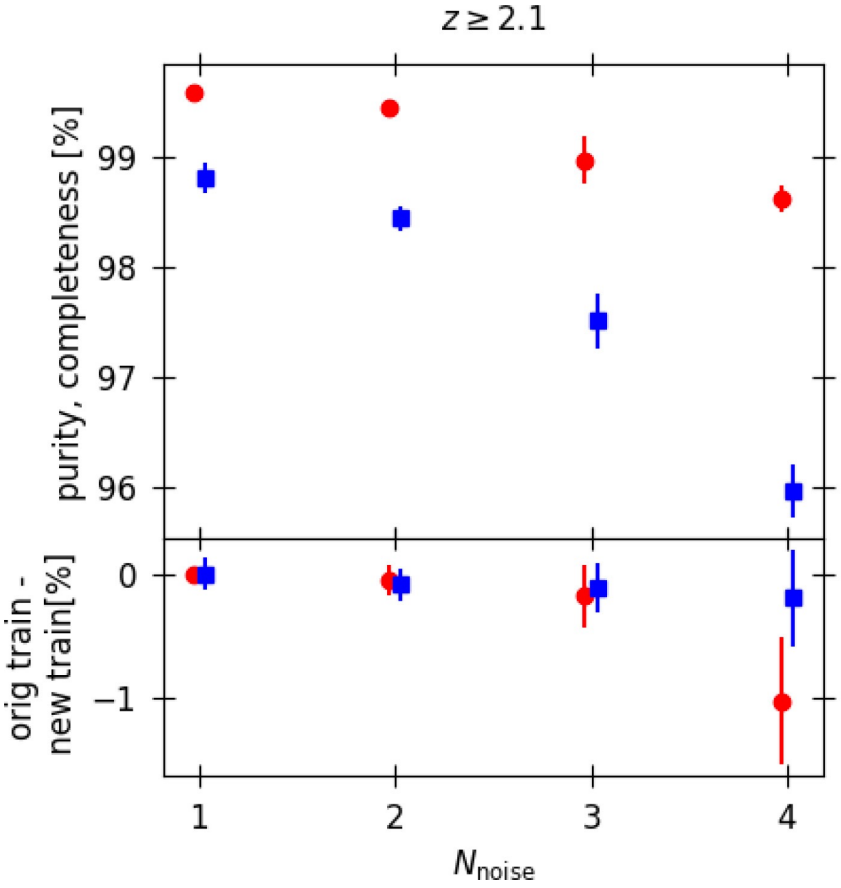
SQUEzE performance

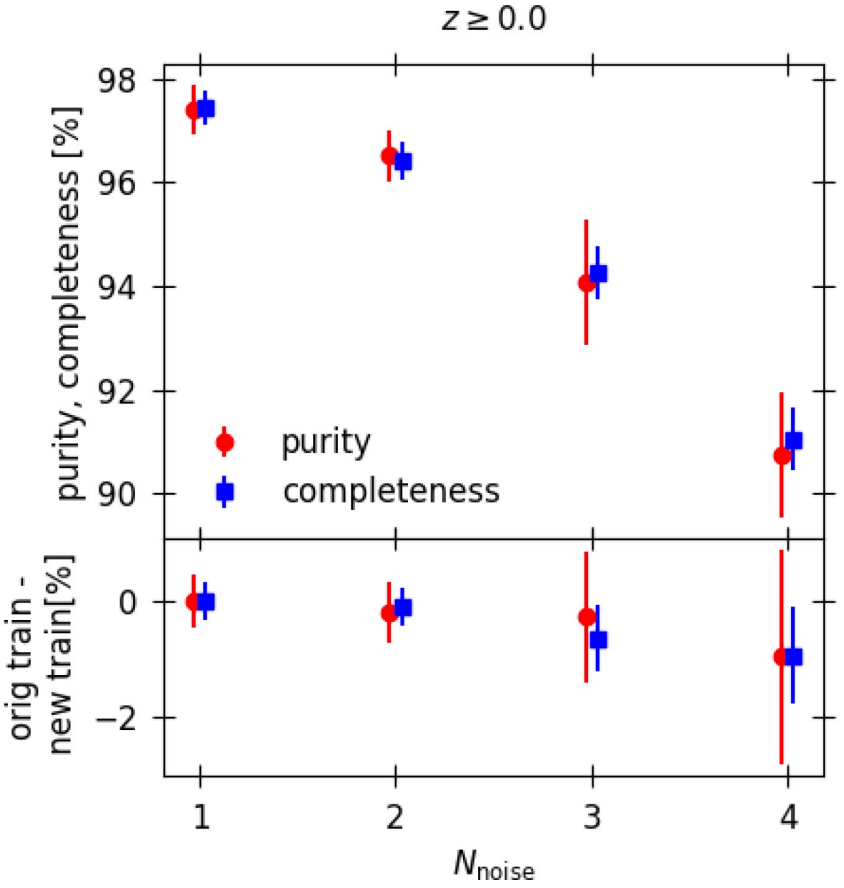




Pérez-Ràfols et al. 2020b

SQUEzE flexibility (noise)



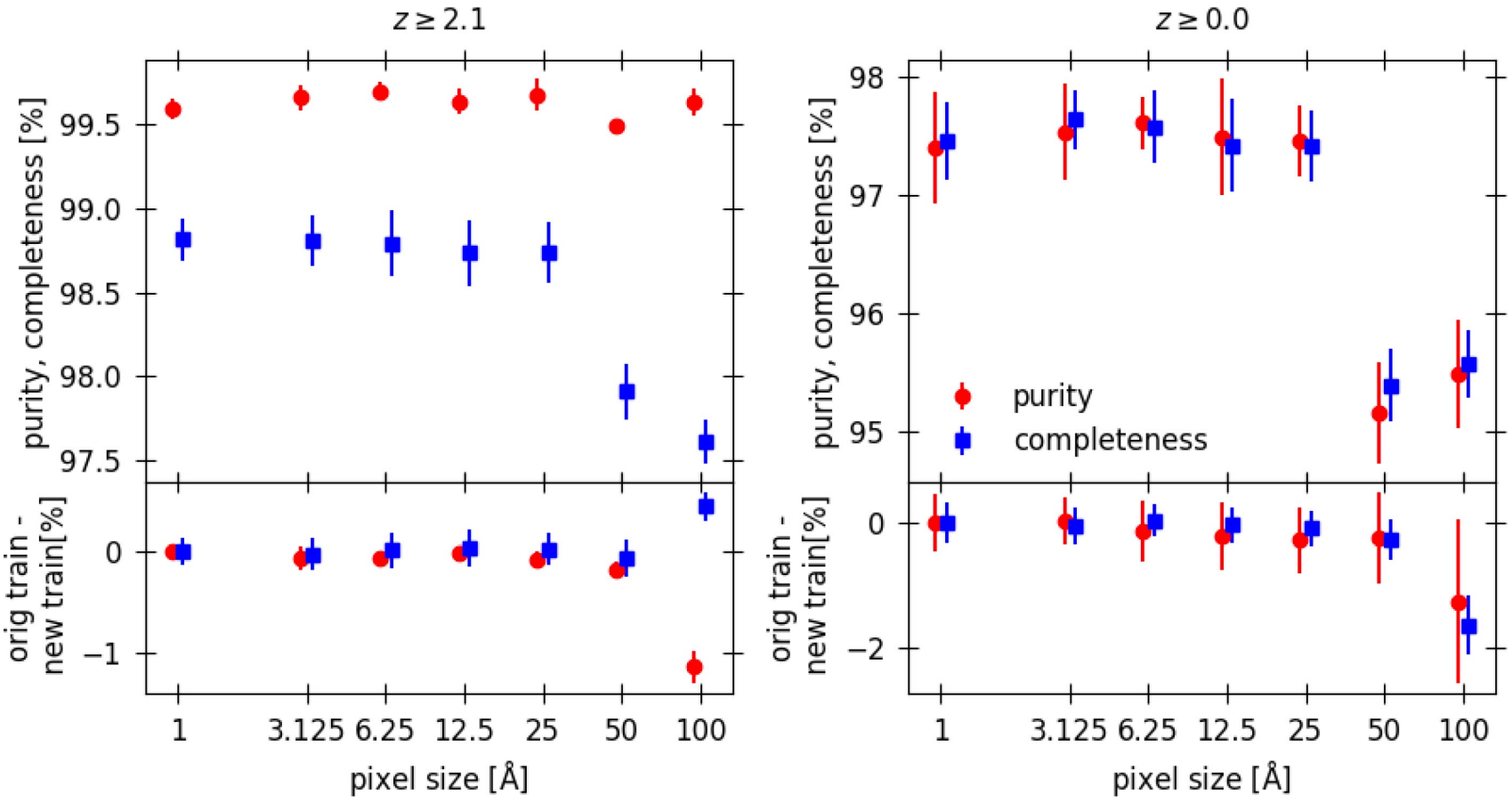


$$f_i' = f_i + (N_{\mathrm{noise}} - 1)\,\sigma_i\,G(0,1),$$

$$\sigma_i' = \sigma_i\sqrt{N_{\mathrm{noise}}},$$

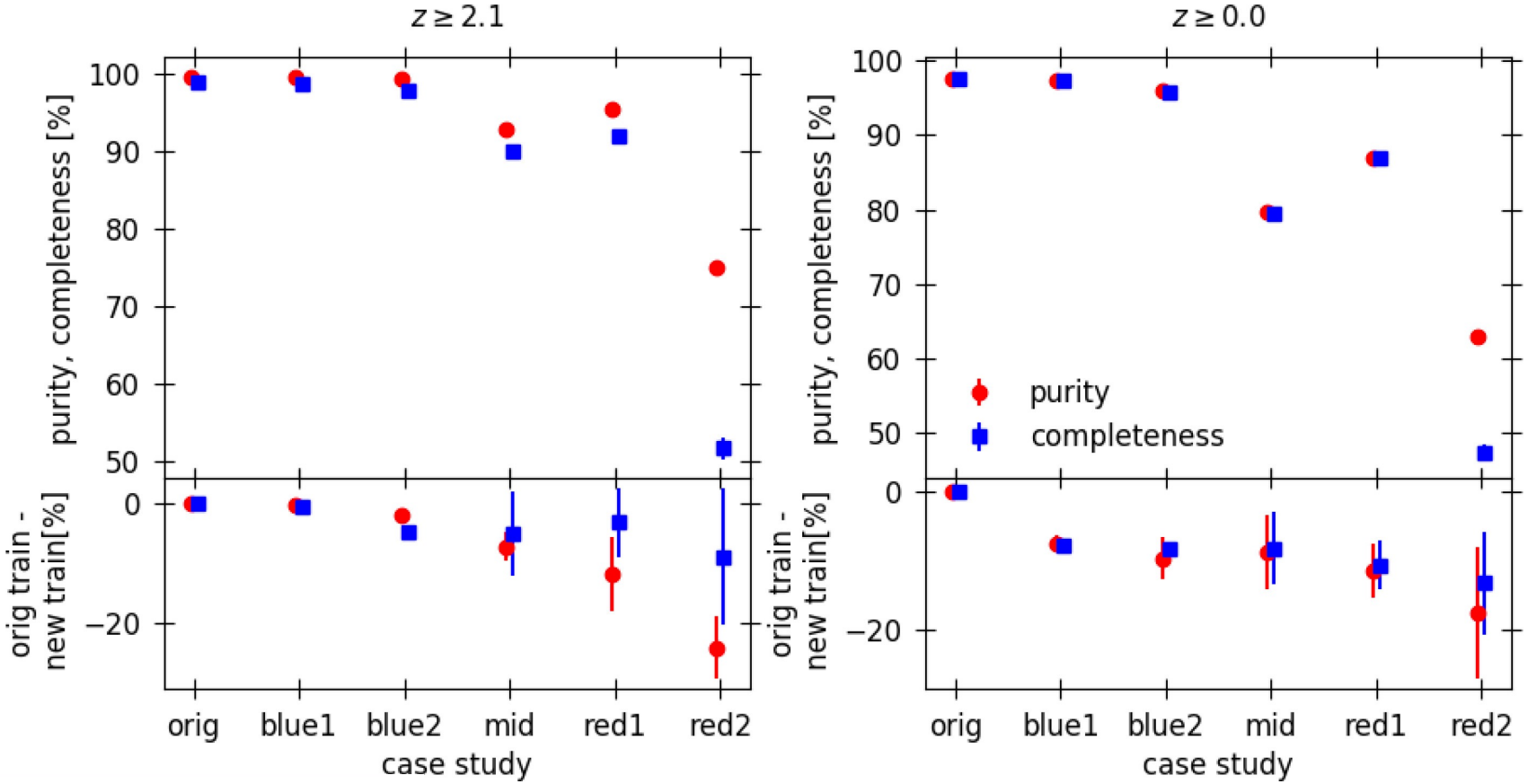
Pérez-Ràfols et al. 2020a

SQUEzE flexibility (rebinning)



Pérez-Ràfols et al. 2020a

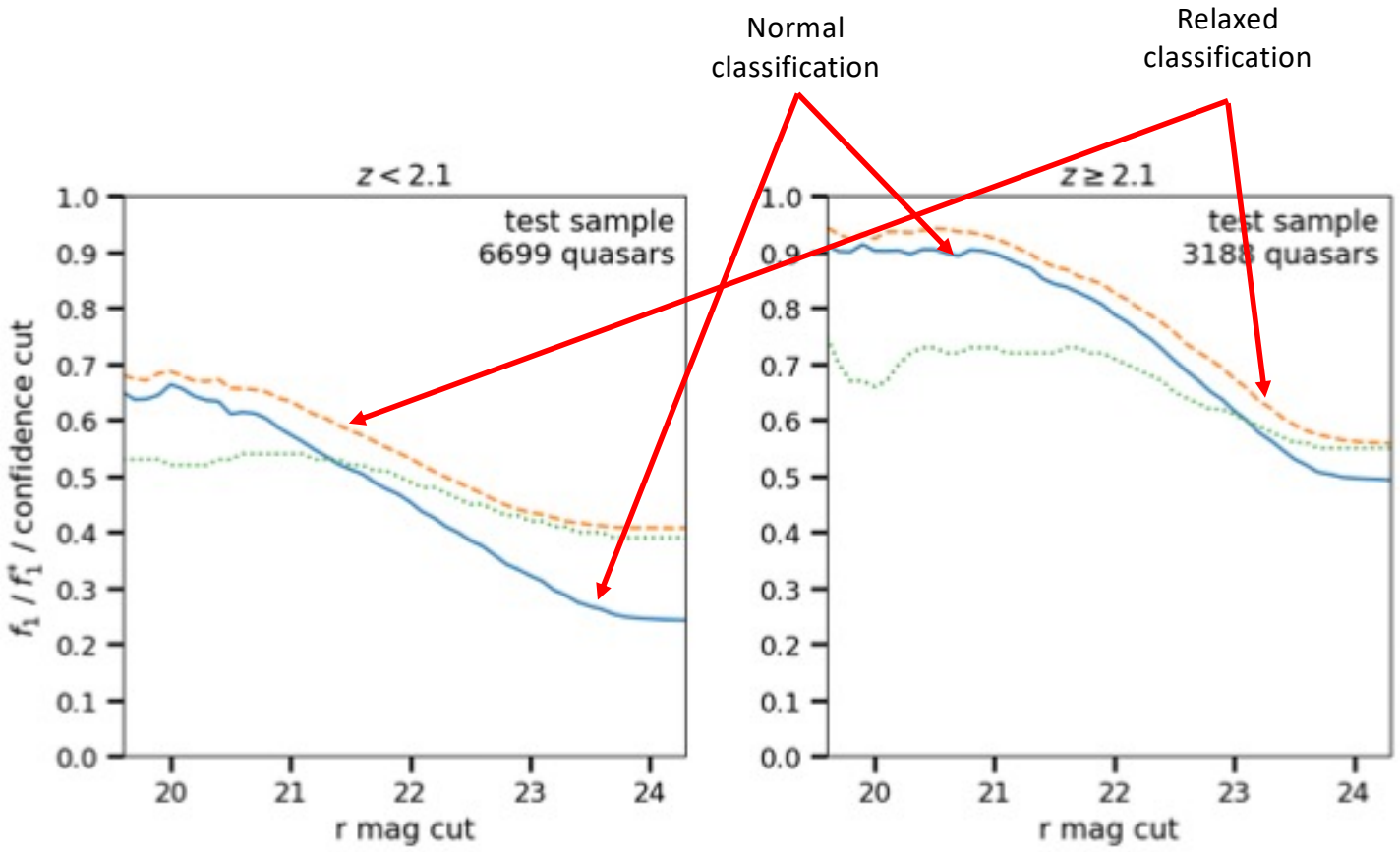
SQUEzE flexibility (resolution)



Sample	Wavelength coverage (Å)
orig	3600–10 400
red1	5300–10 400
red2	7000–10 400
blue1	3600–8700
blue2	3600–7000
mid	5300–8700

Pérez-Ràfols et al. 2020a

SQUEzE flexibility (photometric data)



More relaxed classification

- Object being a quasar
- Correct high-z (z>=2.1) and low-z classification

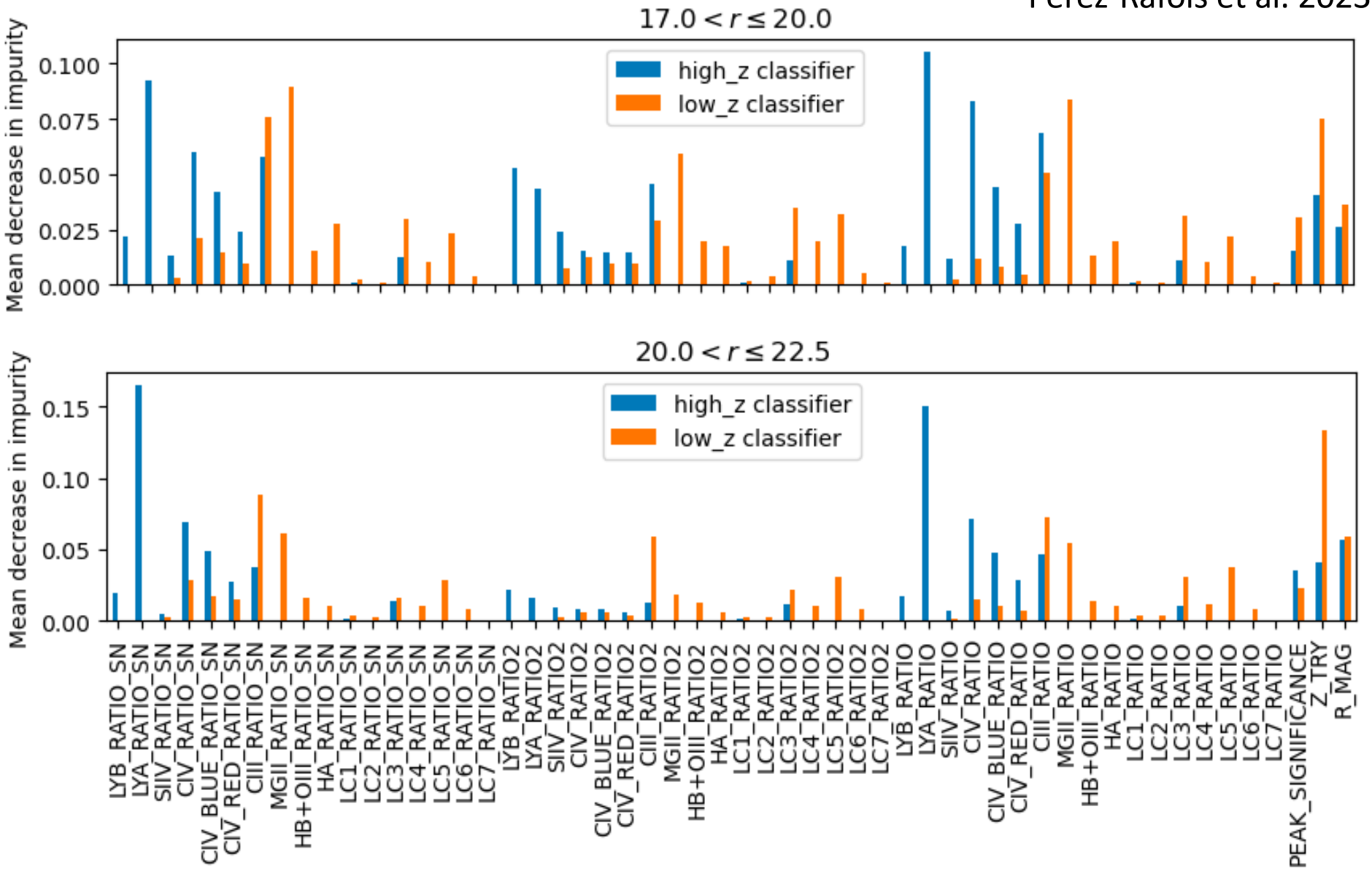
Pérez-Ràfols et al. 2023

SQUEzE explanation

- We can only correctly classify an object if a peak is identified
- We defined a set of metrics for the classification/regression
- Feature extraction allow us to understand the behaviour of the random forest
  This is performed by computing the mean (across the different tres in the forest) decrease in impurities when a particular feature is included or not. Higher values for the mean decrease indicate higher importance ofthe feature
- Doing splits in the sample help in understanding the behaviour

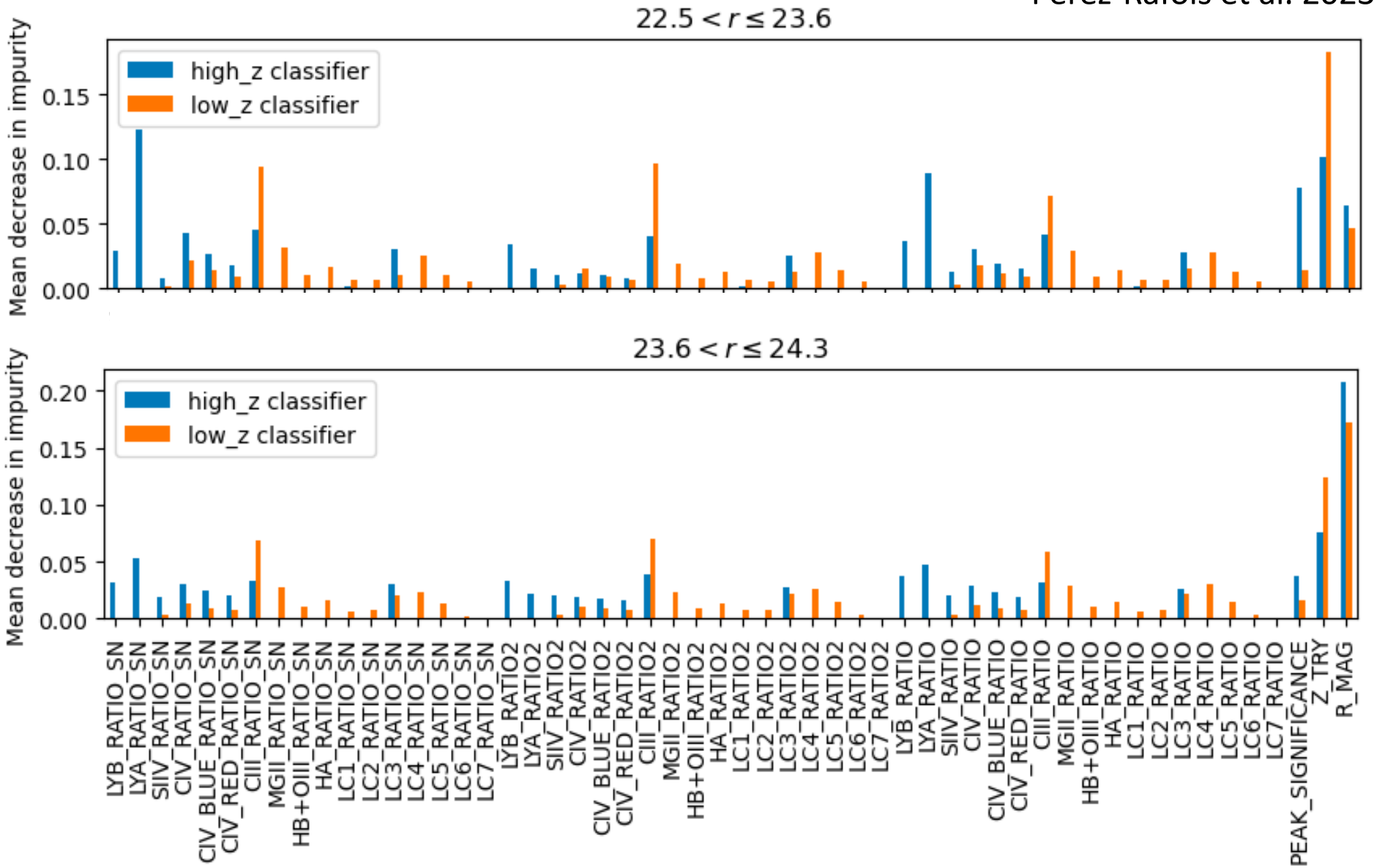
SQUEzE explanation

Pérez-Ràfols et al. 2023



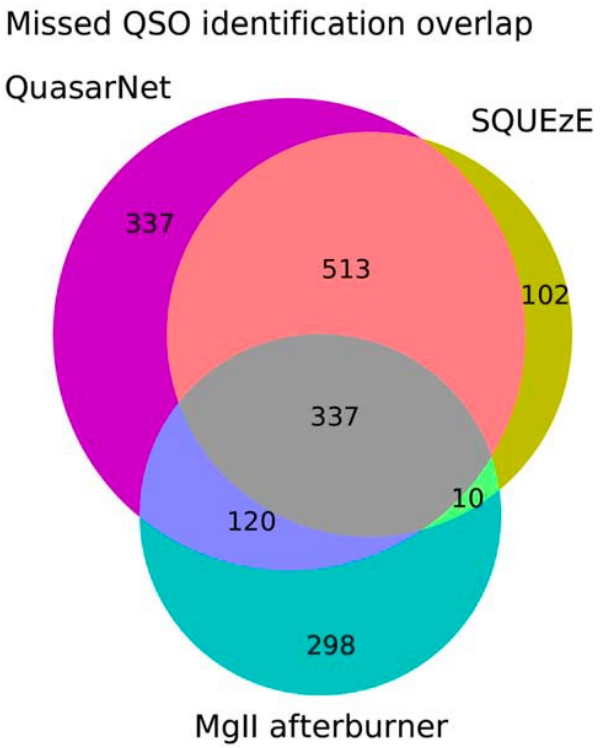
SQUEzE explanation

Pérez-Ràfols et al. 2023



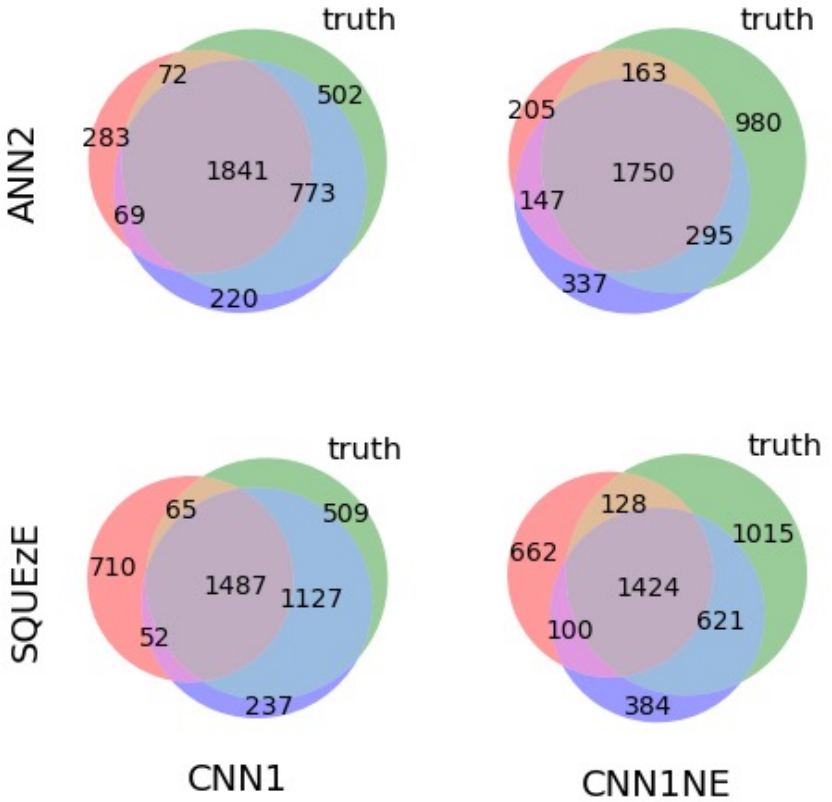
Combination with other algorithms

DESI



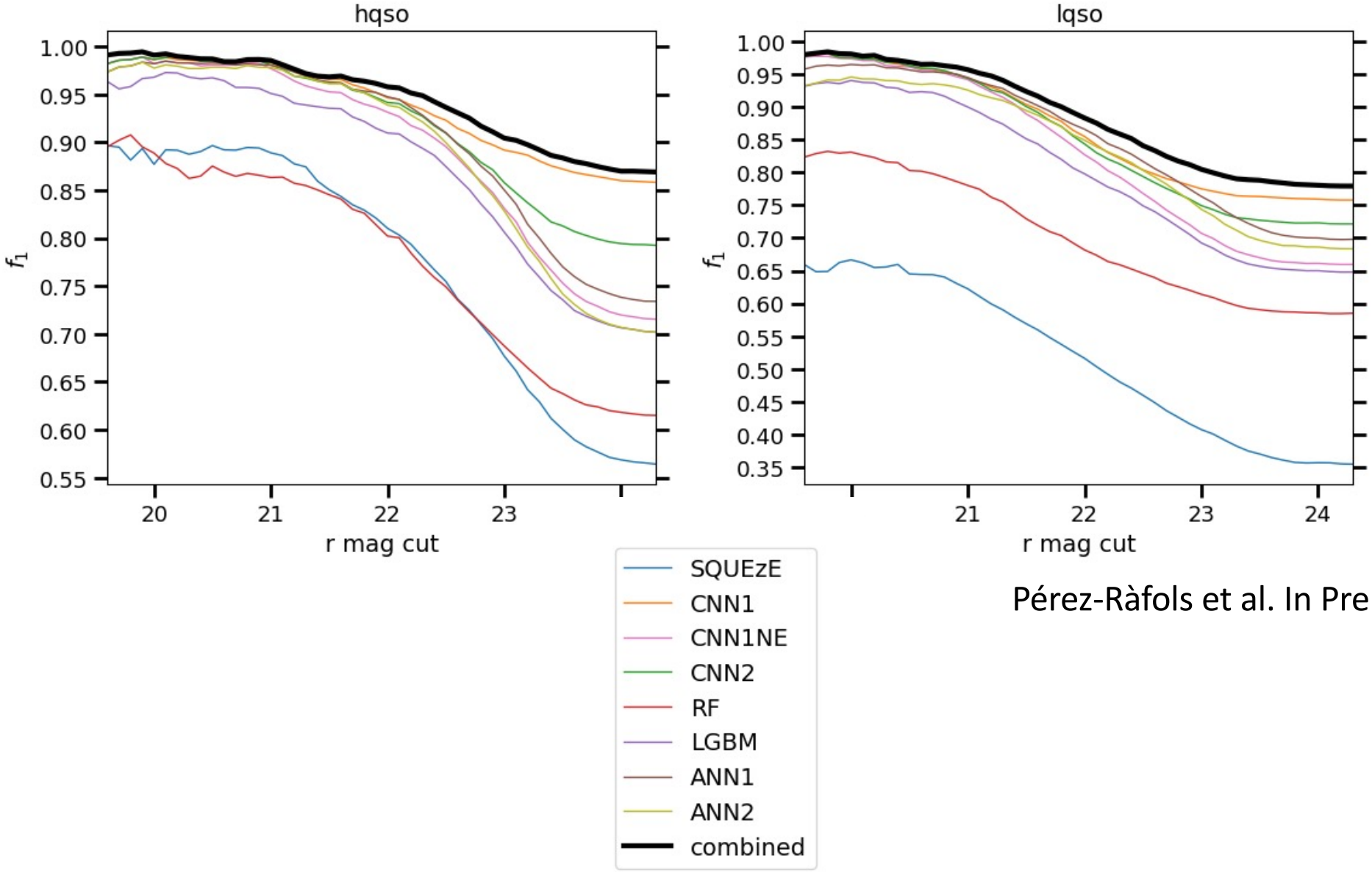
Alexander et al. 2023

JPAS



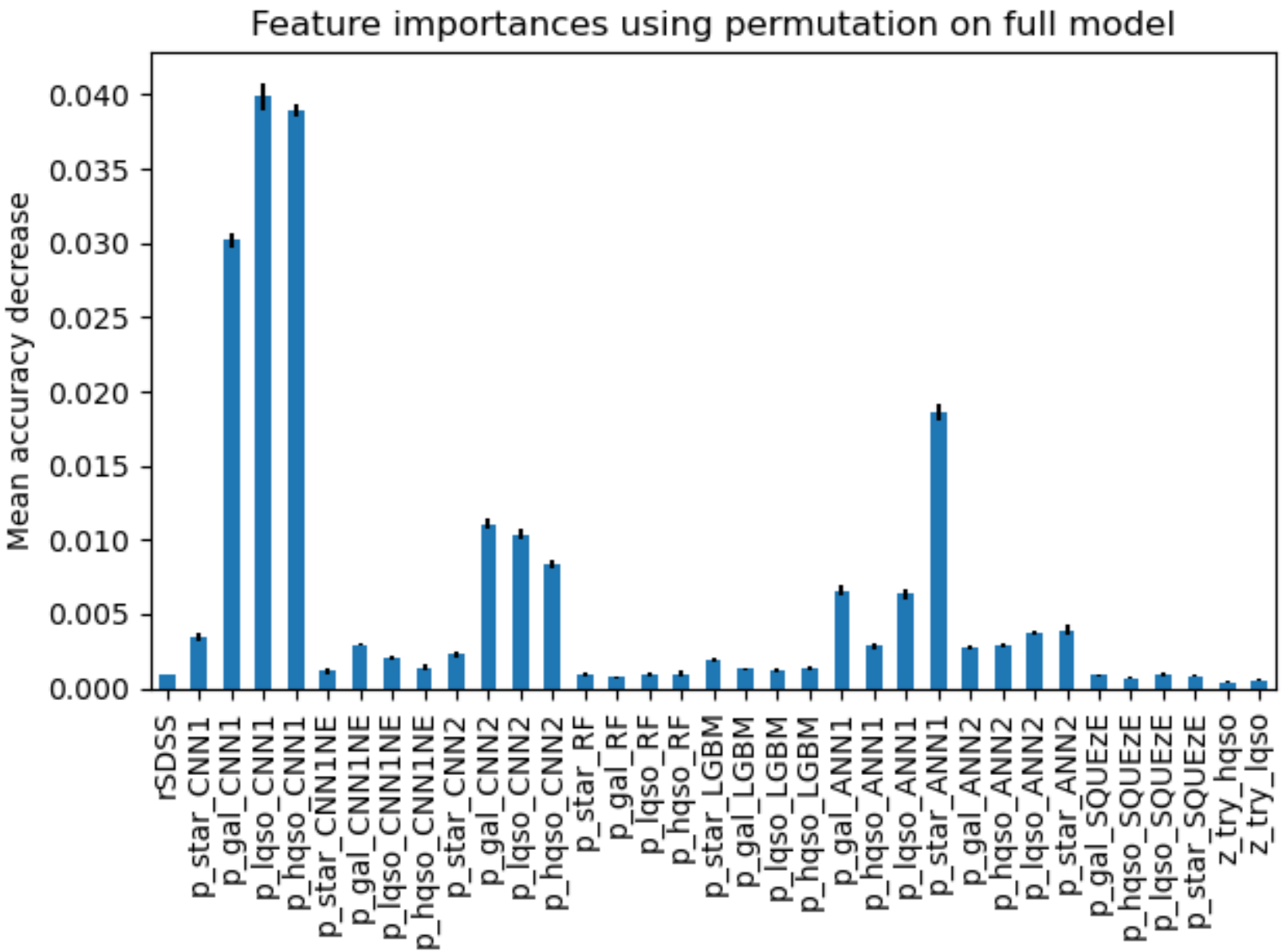
Pérez-Ràfols et al. In Prep

Combined algorithm



Pérez-Ràfols et al. In Prep

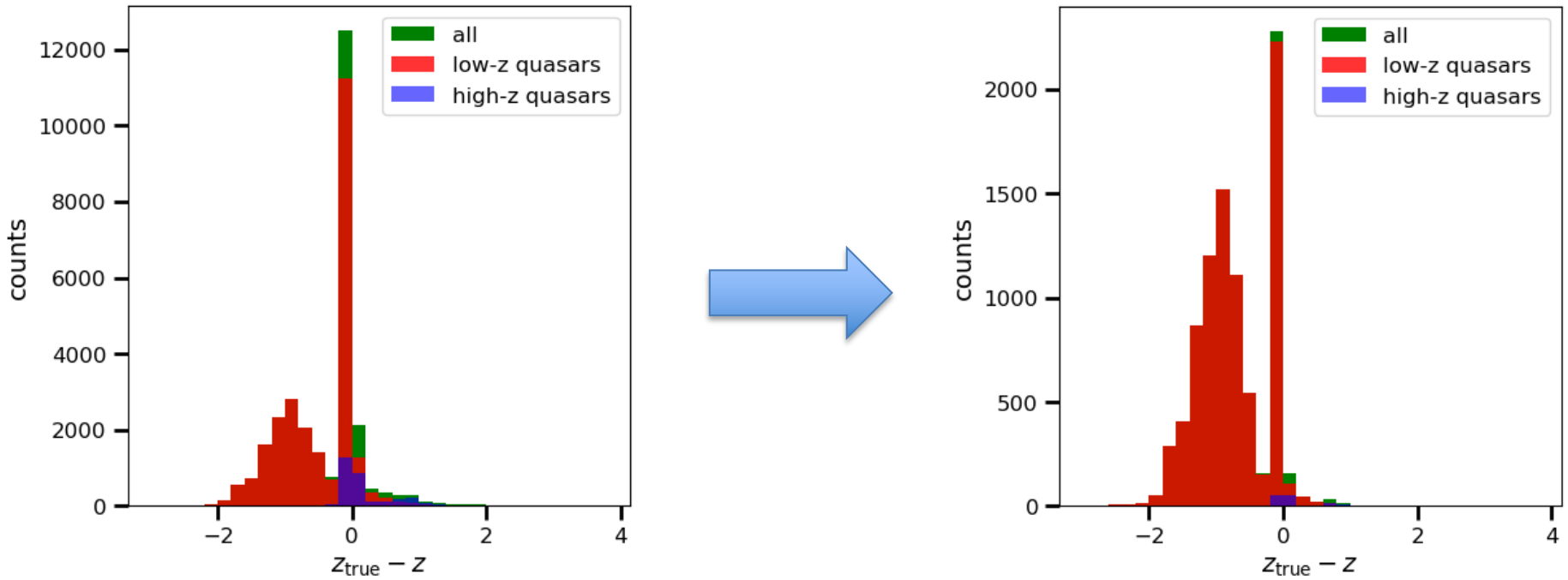
Combined algorithm



Combined algorithm

Final redshift is assigned as

- if class = star → z = 0
- if class = galaxy → z=z_try_lqso
- if class = lqso → z=z_try_lqso
- if class = hqso → z=z_try_hqso



Conclusions

- An intelligent choice of features can allow models do more than they are supposed to be → Don't jump directly to deep learning!

- Simpler models (potentially) offer
  - Higher degree of flexibility → predictability
  - Higher degree of explainability

- However, sometimes performance is not as high