FLAMEING: are we able to find SC4K-like LAEs using AI?

Ana Paulino-Afonso

Institute of Astrophysics and Space Sciences (IA-U.Porto)

FCT official team members: Andrew Humphrey (DTx CoLab/IA-U.Porto), Jarle Brinchmann (IA-U.Porto), Israel Matute (IA-U.Lisboa), Rodrigo Pizarro (IA-U.Lisboa/FCUL), Thomas Scott (IA-U.Porto), Pedro Cunha (IA-U.Porto/FCUP), José Fonseca (IA-U.Porto/FCUP), **Afonso do Vale (IA-U.Porto/FCUP)**, and **Bruno Cerqueira (IA-U.Porto/FCUP)**

Unofficial team members, but people always available to discuss and help: Bruno Ribeiro (GBL/Celfocus), David Sobral (BNP Paribas), XGAL team, and CRISP team

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) through the research grants UIDB/04434/2020 and UIDP/04434/2020, the FCT Investigador FCT Contract No. 2020.03946.CEECIND, and in the form of an exploratory project with the EXPL/FIS-AST/1085/2021 reference (PI: Paulino-Afonso).

ia fct Fundação para a Ciência e a Tecnologia

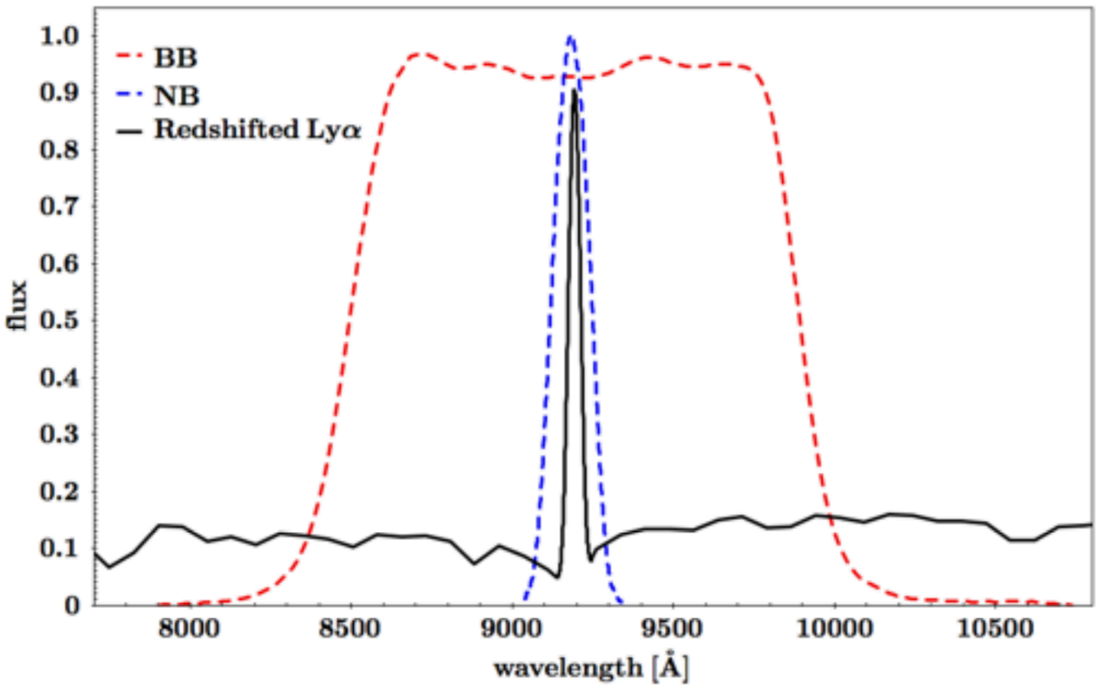
Escape of Lyman radiation from galactic labyrinths @ OAC, Kolymbari, Crete

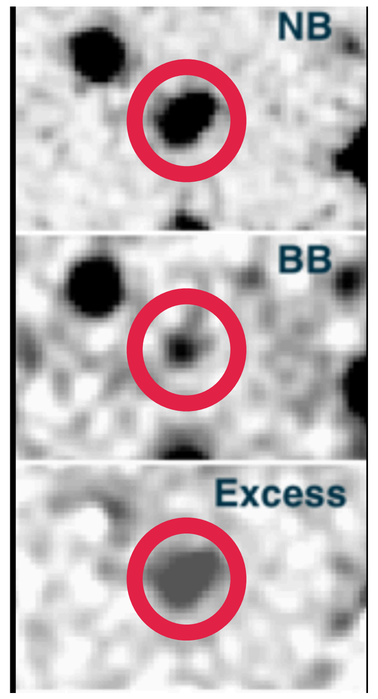
April 21, 2023

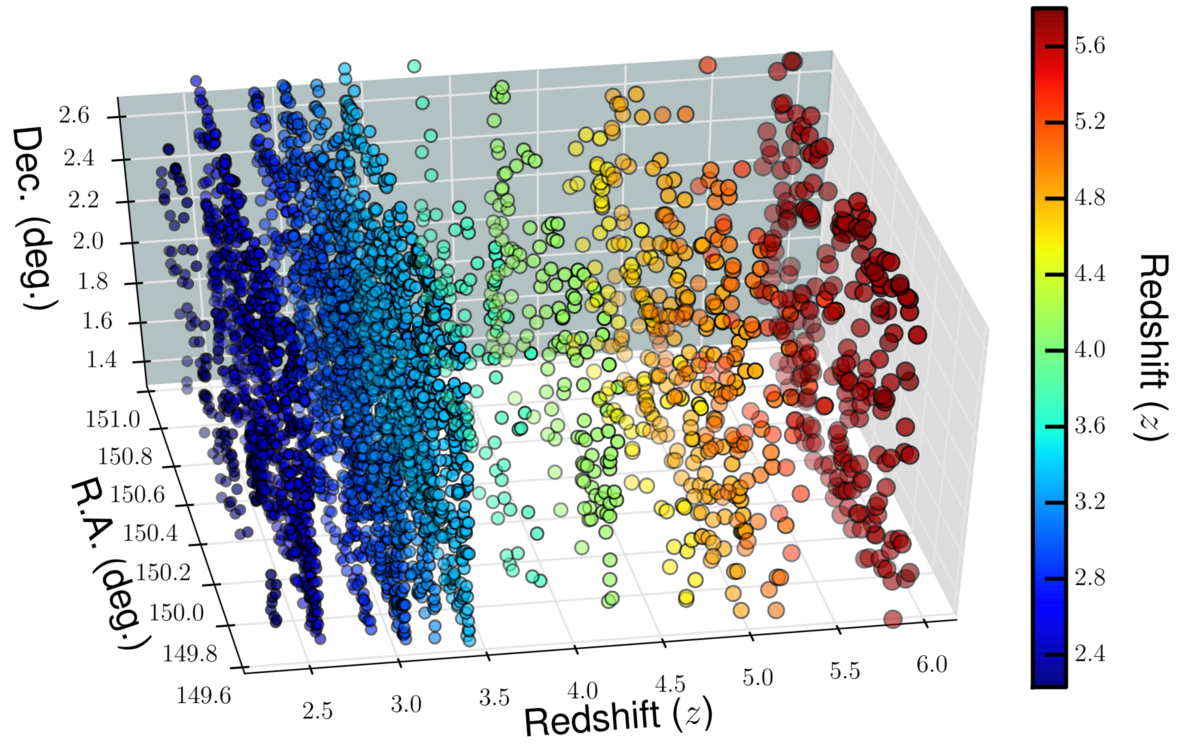
Introducing & Motivation: SC4K

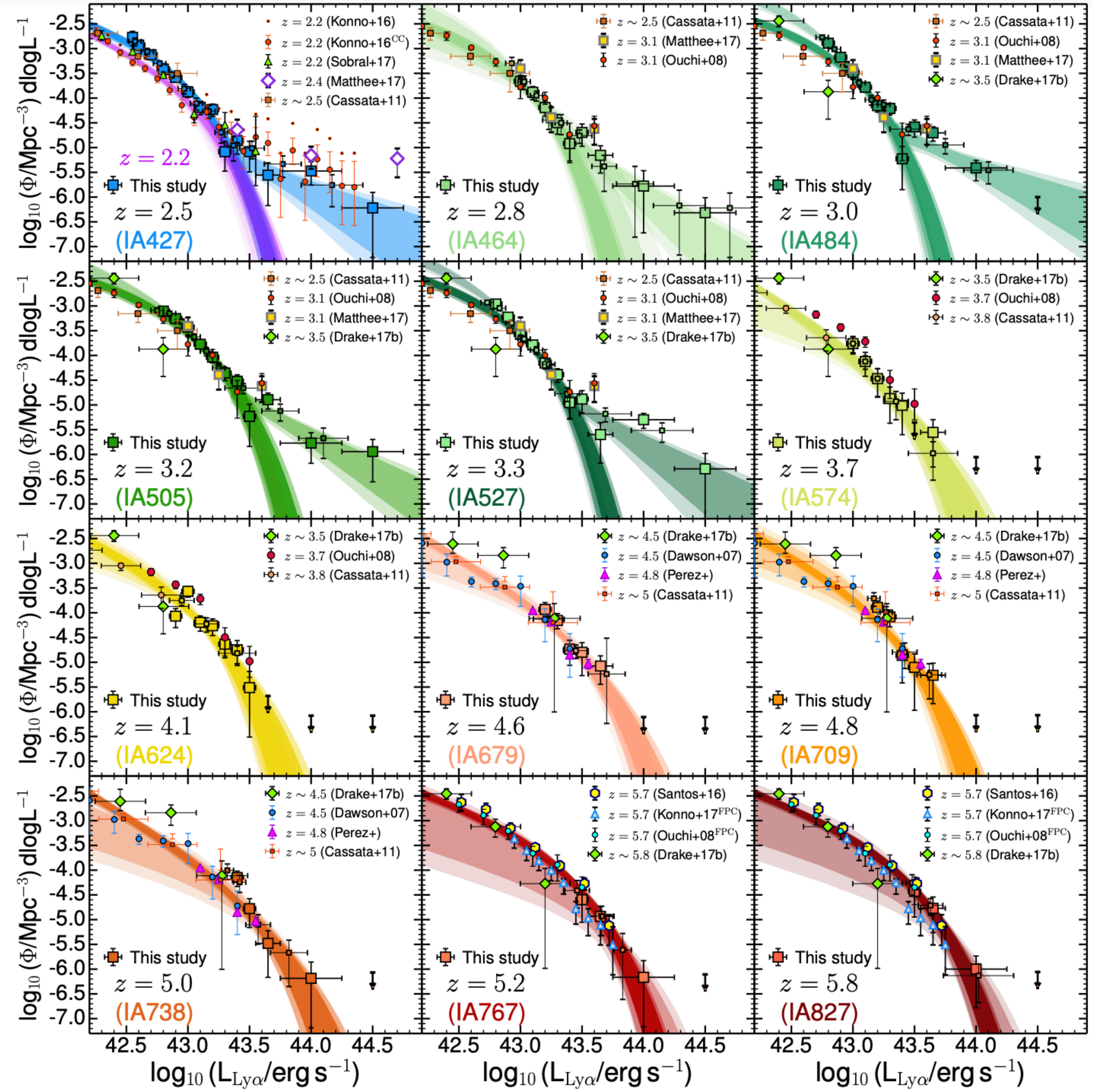
Sobral et al. 2018a

- A **good, well-understood selection** that can be applied with current instrumentation
- Well calibrated + sensitive + resulting in **a representative population of galaxies**
- Able to uniformly select **large samples**
- **Different epochs** + **large areas** + **best-studied fields**





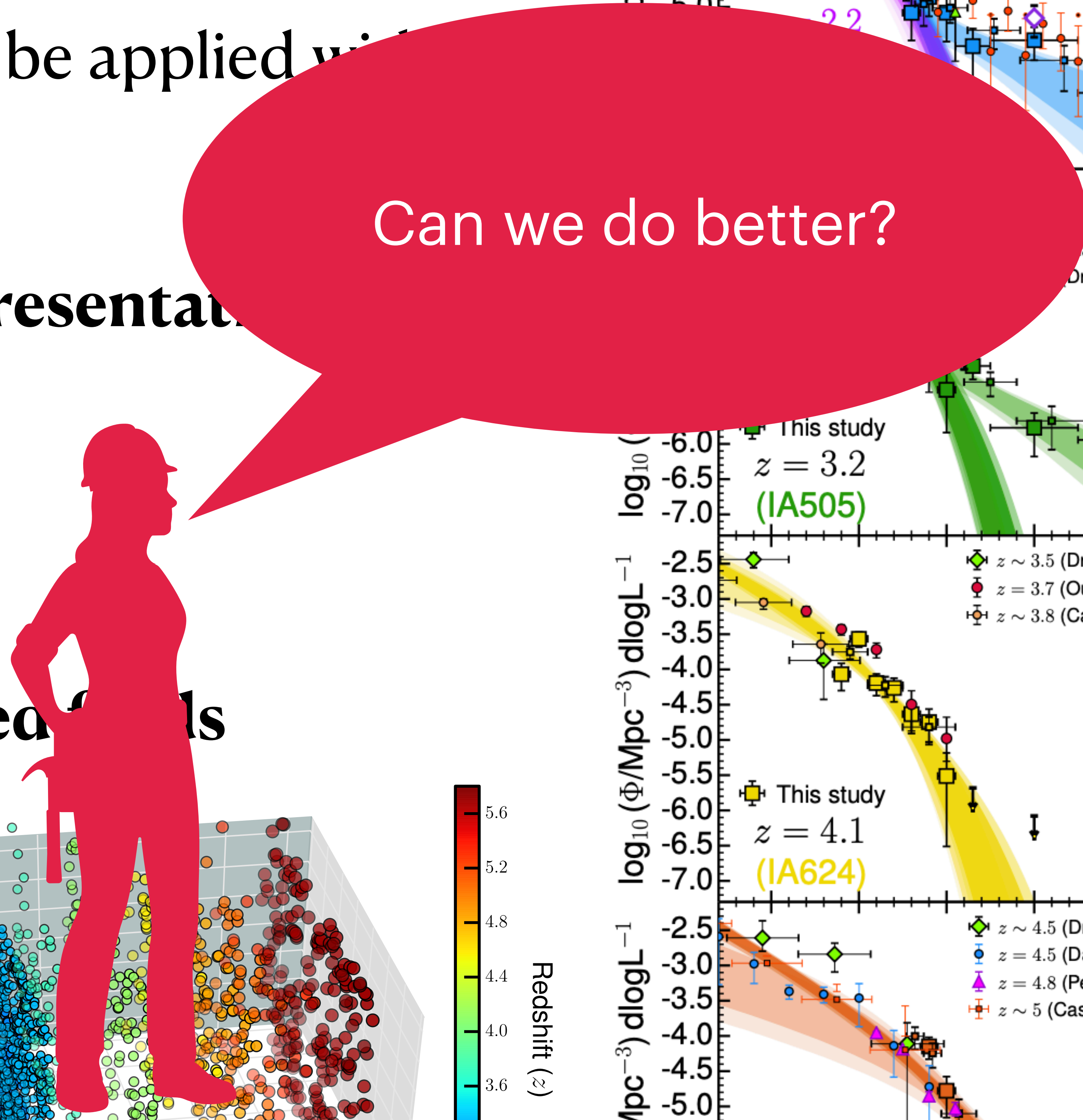


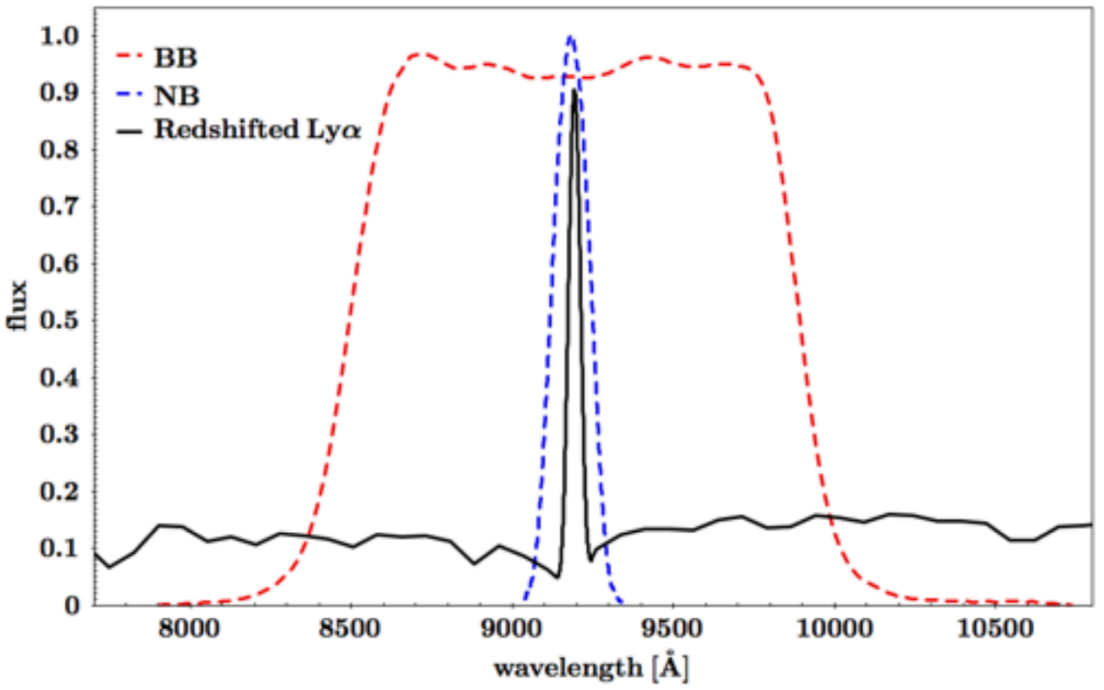


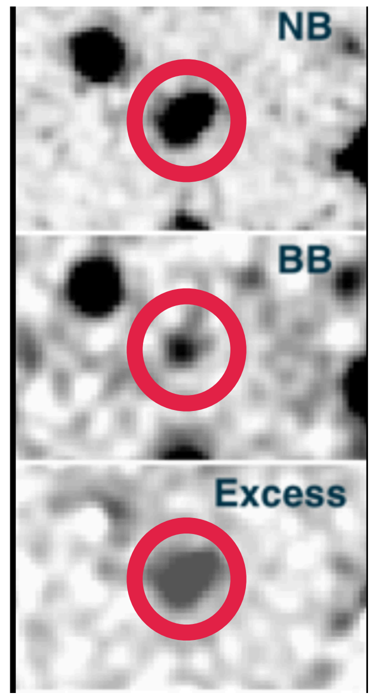
Introducing & Motivation: SC4K

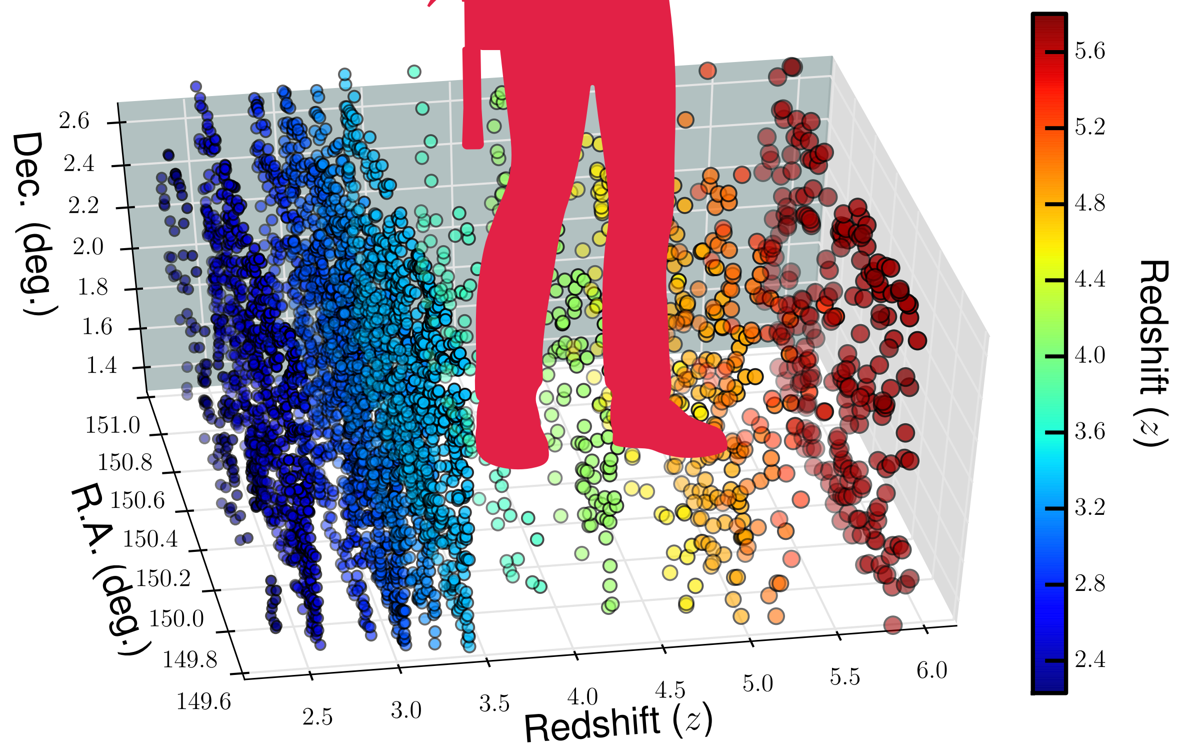
Sobral et al. 2018a

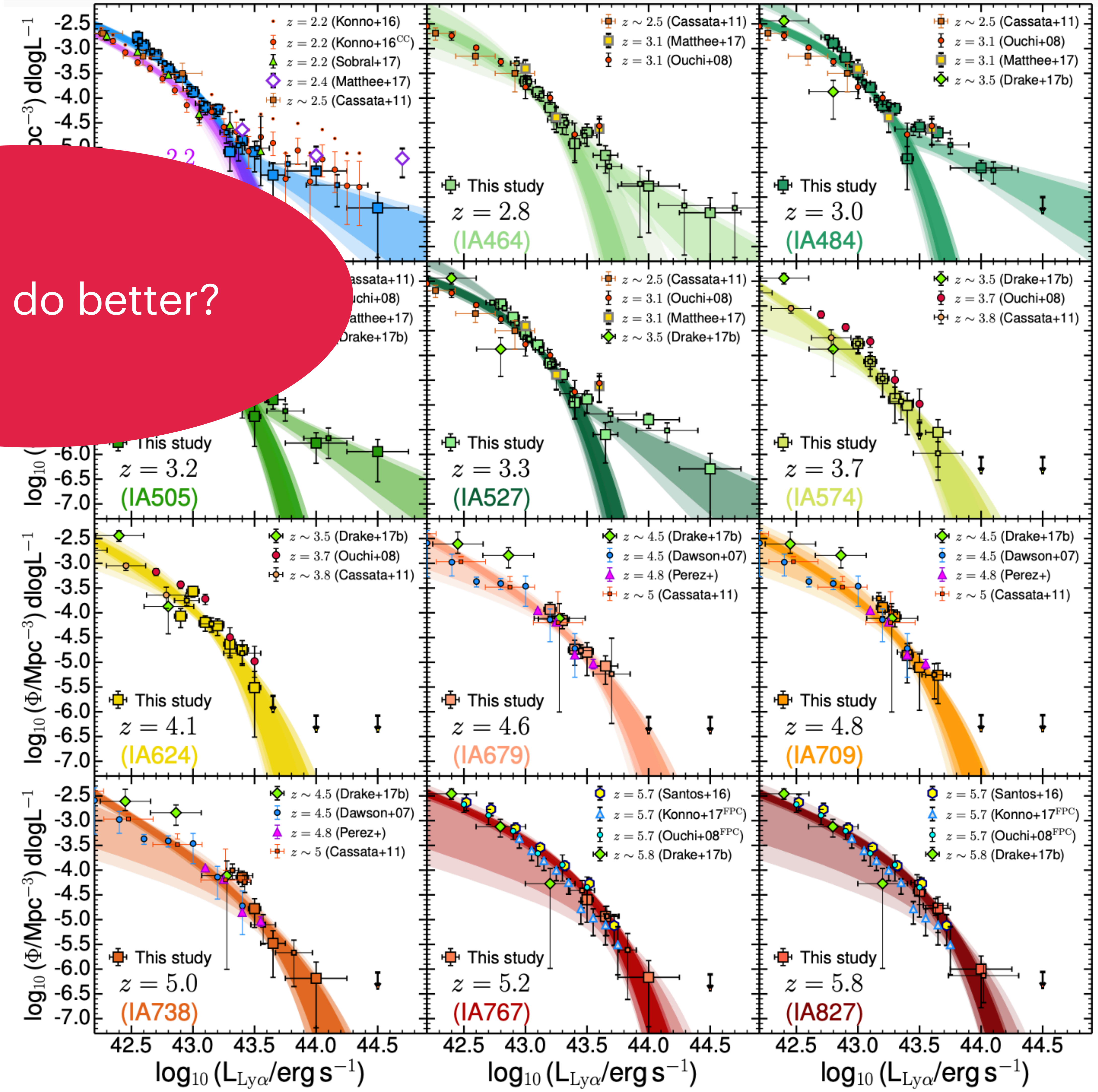
- A **good, well-understood selection** that can be applied wi
  current instrumentation
- Well calibrated + sensitive + resulting in **a representat**
  population of galaxies
- Able to uniformly select **large samples**
- **Different epochs** + **large areas** + **best-studied**











Introducing & Motivation: SILVERRUSH

Ono et al. 2022

ACCEPTED FOR PUBLICATION IN ApJ
Preprint typeset using LaTeX style emulateapj v. 12/16/11

9,318 new LAE candidates @ 2.2 < z < 7.0

177 of them with spectroscopic confirmation

SILVERRUSH X: MACHINE LEARNING-AIDED SELECTION OF 9,318 LAES AT $z = 2.2, 3.3, 4.9, 5.7, 6.6$, AND 7.0 FROM THE HSC SSP AND CHORUS SURVEY DATA

YOSHIAKI ONO[1], RYOHEI ITOH[1,2], TAKATOSHI SHIBUYA[3], MASAMI OUCHI[4,1,5], YUICHI HARIKANE[1,6], SATOSHI YAMANAKA[7,8], AKIO K. INOUE[9,7], TOSHIYUKI AMAGASA[10,11], DAICHI MIURA[10], MAIKI OKURA[10], KAZUHIRO SHIMASAKU[12,13], IKURU IWATA[4], YOSHIAKI TANIGUCHI[14], SEIJI FUJIMOTO[15], MASANORI IYE[4], ANTON T. JAELANI[16,17], NOBUNARI KASHIKAWA[12,13], SHOTARO KIKUCHIHARA[1,12], SATOSHI KIKUTA[11], MASAKAZU A.R. KOBAYASHI[18], HARUKA KUSAKABE[19], CHIEN-HSIU LEE[20], YONGMING LIANG[4], YOSHIKI MATSUOKA[8], RIEKO MOMOSE[12], TOHRU NAGAO[8], KIMIHIKO NAKAJIMA[4], AND KEN-ICHI TADAKI[4]

Accepted for publication in ApJ

ABSTRACT

We present a new catalog of 9,318 Lyα emitter (LAE) candidates at $z = 2.2$, 3.3, 4.9, 5.7, 6.6, and 7.0 that are photometrically selected by the SILVERRUSH program with a machine learning technique from large area (up to 25.0 deg^2) imaging data with six narrowband filters taken by the Subaru Strategic Program with Hyper Suprime-Cam (HSC SSP) and a Subaru intensive program, Cosmic HydrOgen Reionization Unveiled with Subaru (CHORUS). We construct a convolutional neural network that distinguishes between real LAEs and contaminants with a completeness of 94% and a contamination rate of 1%, enabling us to efficiently remove contaminants from the photometrically selected LAE candidates. We confirm that our LAE catalogs include 177 LAEs that have been spectroscopically identified in our SILVERRUSH programs and previous studies, ensuring the validity of our machine learning selection. In addition, we find that the object-matching rates between our LAE catalogs and our previous results are $\simeq$ 80–100% at bright NB magnitudes of $\lesssim 24$ mag. We also confirm that the surface number densities of our LAE candidates are consistent with previous results. Our LAE catalogs will be made public on our project webpage.

Keywords: galaxies: formation — galaxies: evolution — galaxies: high-redshift

Introducing & Motivation: SILVERRUSH

Ono et al. 2022

9,318 new LAE candidates @ 2.2 < z < 7.0

177 of them with spectroscopic confirmation

ACCEPTED FOR PUBLICATION IN ApJ
Preprint typeset using LaTeX style emulateapj v. 12/16/11

SILVERRUSH X: MACHINE LEARNING-AIDED SELECTION OF 9,318 LAES AT $z = 2.2, 3.3, 4.9, 5.7, 6.6$, AND 7.0 FROM THE HSC SSP AND CHORUS SURVEY DATA

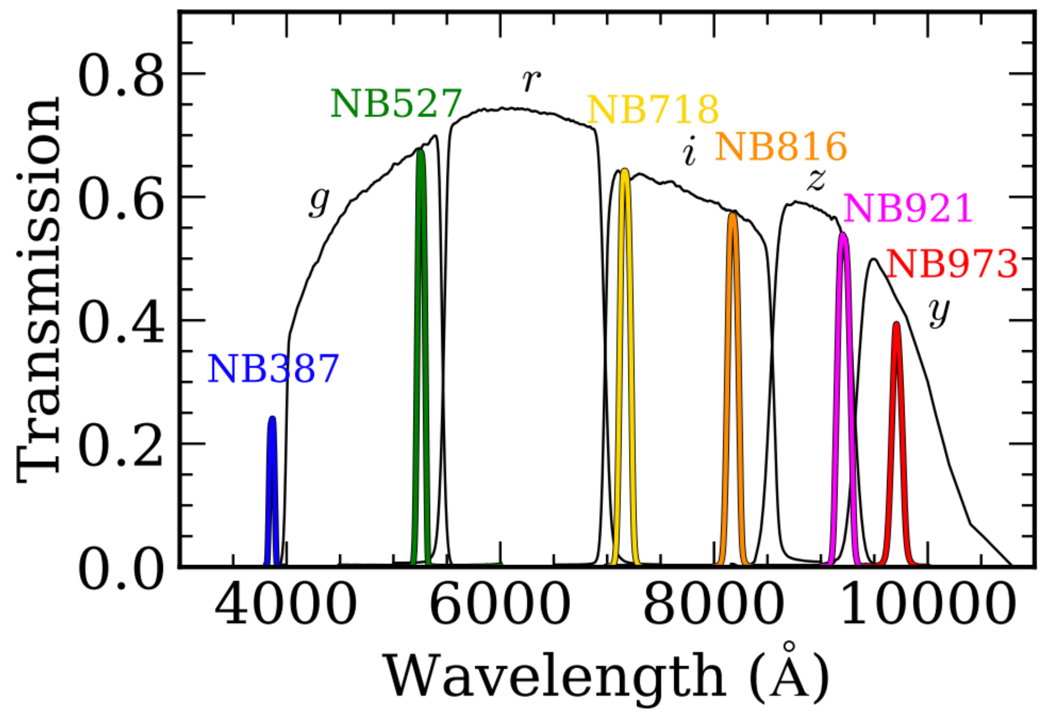
YOSHIAKI ONO[1], RYOHEI ITOH[1,2], TAKATOSHI SHIBUYA[3], MASAMI OUCHI[4,1,5], YUICHI HARIKANE[1,6], SATOSHI YAMANAKA[7,8], AKIO K. INOUE[9,7], TOSHIYUKI AMAGASA[10,11], DAICHI MIURA[10], MAIKI OKURA[10], KAZUHIRO SHIMASAKU[12,13], IKURU IWATA[4], YOSHIAKI TANIGUCHI[14], SEIJI FUJIMOTO[15], MASANORI IYE[4], ANTON T. JAELANI[16,17], NOBUNARI KASHIKAWA[12,13], SHOTARO KIKUCHIHARA[1,12], SATOSHI KIKUTA[11], MASAKAZU A.R. KOBAYASHI[18], HARUKA KUSAKABE[19], CHIEN-HSIU LEE[20], YONGMING LIANG[4], YOSHIKI MATSUOKA[8], RIEKO MOMOSE[12], TOHRU NAGAO[8], KIMIHIKO NAKAJIMA[4], AND KEN-ICHI TADAKI[4]

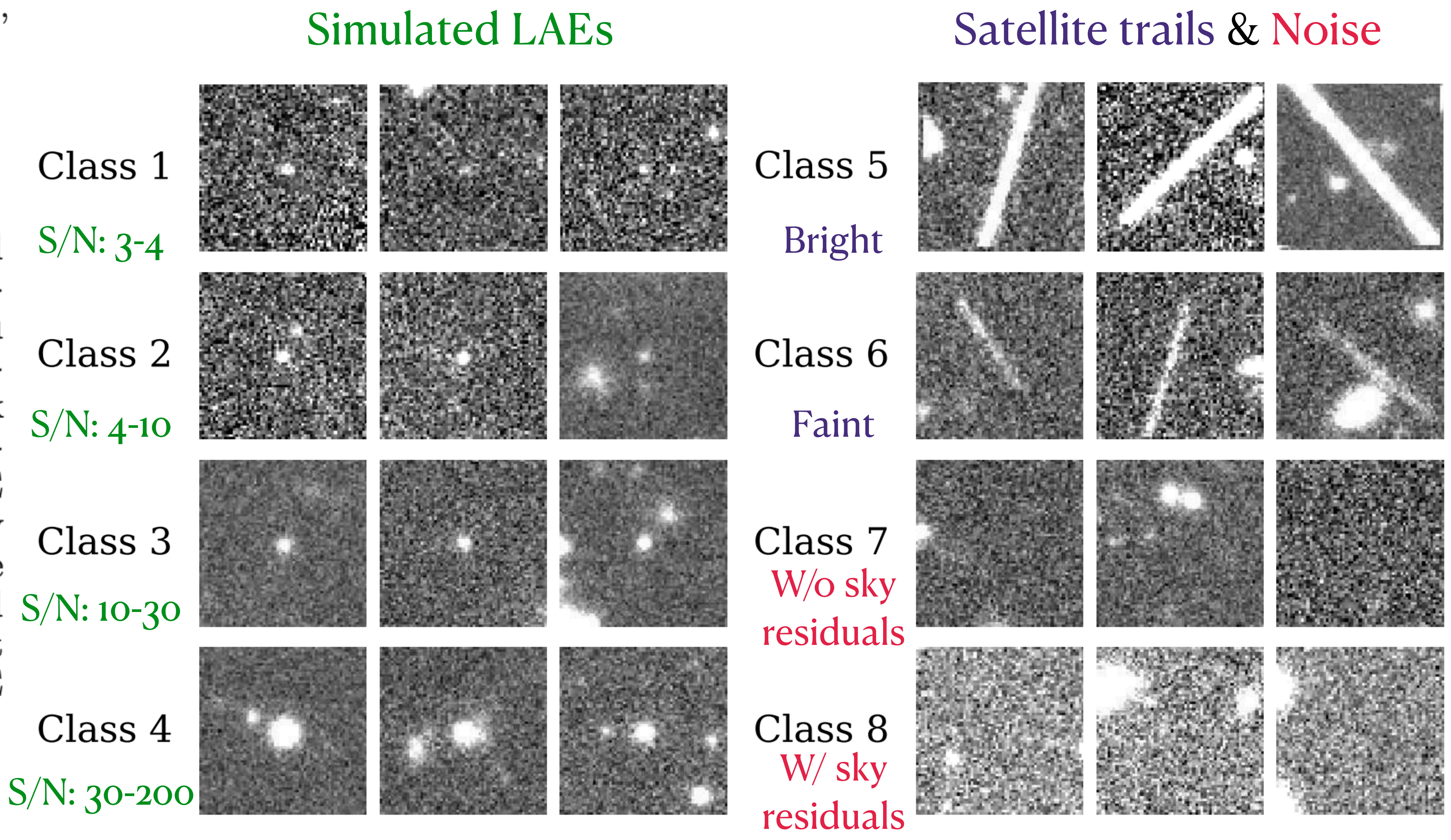
Accepted for publication in ApJ

ABSTRACT

We present a new catalog of 9,318 Lyα emitter (LAE) candidates at $z = 2.2, 3.3, 4.9, 5.7, 6.6$, and 7.0 that are photometrically selected by the SILVERRUSH program with a machine learning technique from large area (up to 25.0 deg^2) imaging data with six narrowband filters taken by the Subaru Strategic Program with Hyper Suprime-Cam (HSC SSP) and a Subaru intensive program, Cosmic HydrOgen Reionization Unveiled with Subaru (CHORUS). We construct a convolutional neural network that distinguishes between real LAEs and contaminants with a completeness of 94% and a contamination rate of 1%, enabling us to efficiently remove contaminants from the photometrically selected LAE candidates. We confirm that our LAE catalogs include 177 LAEs that have been spectroscopically identified in our SILVERRUSH programs and previous studies, ensuring the validity of our machine learning selection. In addition, we find that the object-matching rates between our LAE catalogs and our previous results are $\simeq$ 80–100% at bright NB magnitudes of $\lesssim 24$ mag. We also confirm that the surface number densities of our LAE candidates are consistent with previous results. Our LAE catalogs will be made public on our project webpage.

Keywords: galaxies: formation — galaxies: evolution — galaxies: high-redshift





Introducing & Motivation: SILVERRUSH

Ono et al. 2022

9,318 new LAE candidates @ 2.2 < z < 7.0

177 of them with spectroscopic confirmation

ACCEPTED FOR PUBLICATION IN ApJ
Preprint typeset using LaTeX style emulateapj v. 12/16/11

SILVERRUSH X: MACHINE LEARNING-AIDED SELECTION OF 9,318 LAES AT $z = 2.2, 3.3, 4.9, 5.7, 6.6$, AND 7.0 FROM THE HSC SSP AND CHORUS SURVEY DATA

Yoshiaki Ono[1], Ryohei Itoh[1,2], Takatoshi Shibuya[3], Masami Ouchi[4,1,5], Yuichi Harikane[1,6], Satoshi Yamanaka[7,8], Akio K. Inoue[9,7], Toshiyuki Amagasa[10,11], Daichi Miura[10], Maiki Okura[10], Kazuhiro Shimasaku[12,13], Ikuru Iwata[4], Yoshiaki Taniguchi[14], Seiji Fujimoto[15], Masanori Iye[4], Anton T. Jaelani[16,17], Nobunari Kashikawa[12,13], Shotaro Kikuchihara[1,12], Satoshi Kikuta[11], Masakazu A.R. Kobayashi[18], Haruka Kusakabe[19], Chien-Hsiu Lee[20], Yongming Liang[4], Yoshiki Matsuoka[8], Rieko Momose[12], Tohru Nagao[8], Kimihiko Nakajima[4], and Ken-ichi Tadaki[4]

Accepted for publication in ApJ

ABSTRACT

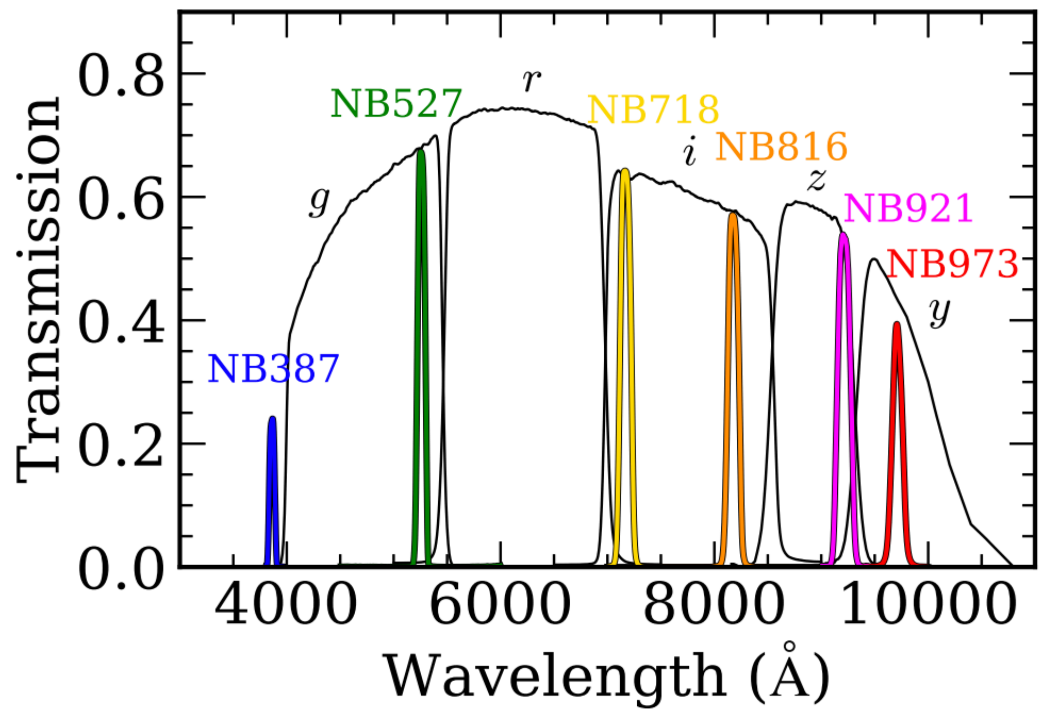
We present a new catalog of 9,318 Lyα emitter (LAE) candidates at $z = 2.2, 3.3, 4.9, 5.7, 6.6$, and 7.0 that are photometrically selected by the SILVERRUSH program with a machine learning technique from large area (up to 25.0 deg^2) imaging data with six narrowband filters taken by the Subaru Strategic Program with Hyper Suprime-Cam (HSC SSP) and a Subaru intensive program, Cosmic HydrOgen Reionization Unveiled with Subaru (CHORUS). We construct a convolutional neural network that distinguishes between real LAEs and contaminants with a completeness of 94% and a contamination rate of 1%, enabling us to efficiently remove contaminants from the photometrically selected LAE candidates. We confirm that our LAE catalogs include 177 LAEs that have been spectroscopically identified in our SILVERRUSH programs and previous studies, ensuring the validity of our machine learning selection. In addition, we find that the object-matching rates between our LAE catalogs and our previous results are $\simeq$ 80–100% at bright NB magnitudes of $\lesssim$ 24 mag. We also confirm that the surface number densities of our LAE candidates are consistent with previous results. Our LAE catalogs will be made public on our project webpage.

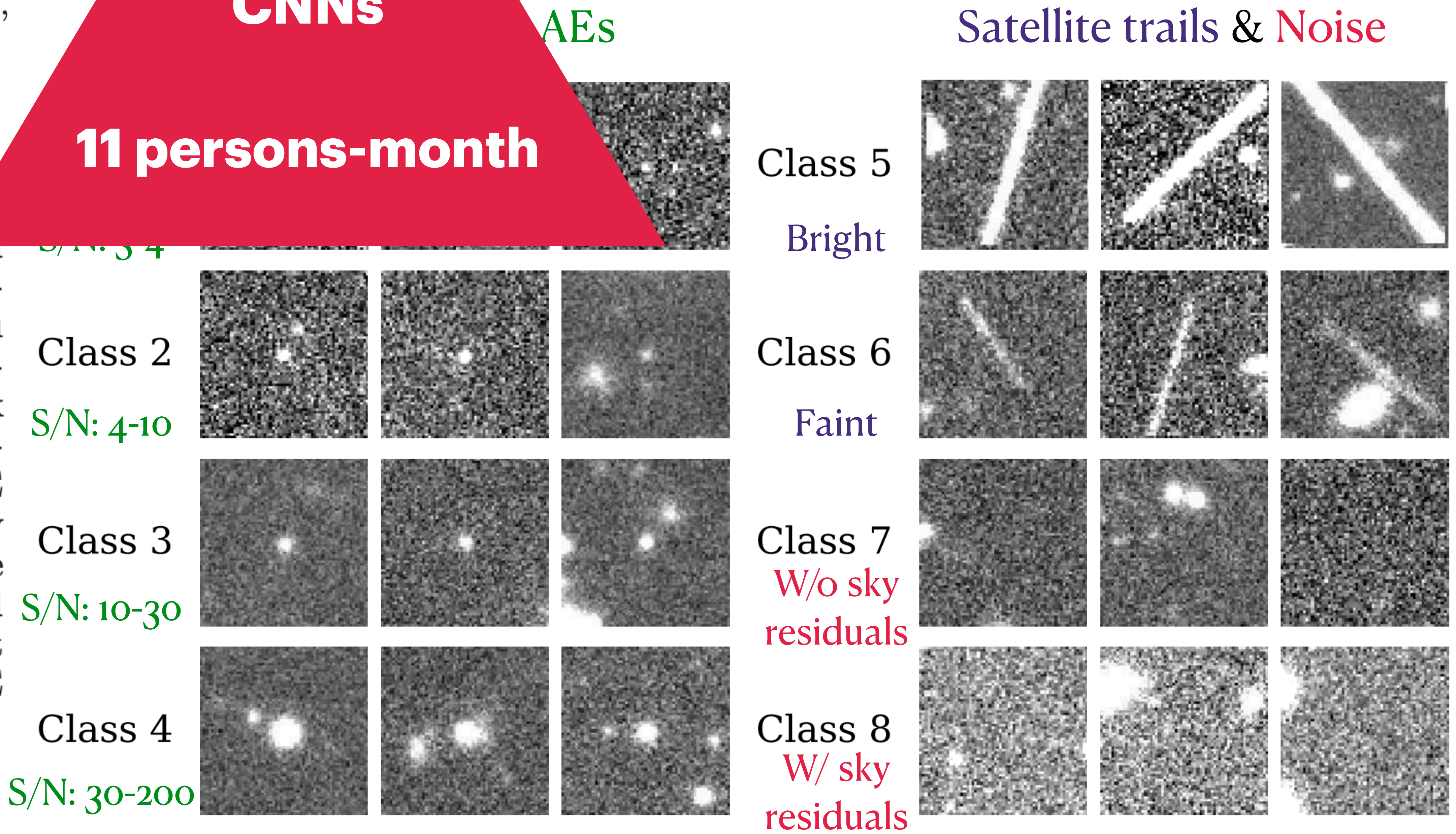
Keywords: galaxies: formation — galaxies: evolution — galaxies: high-redshift

!!!

CNNs

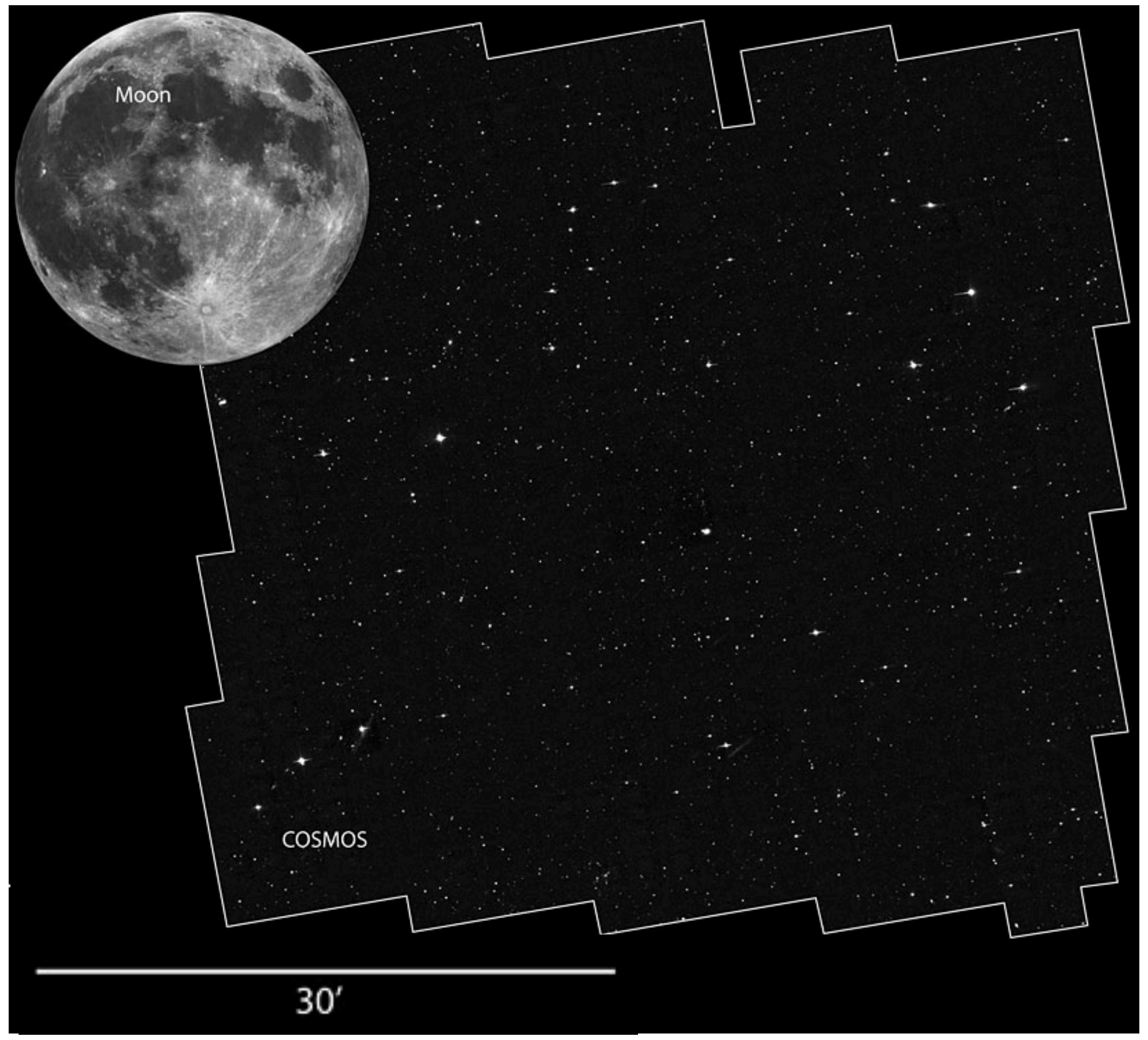
11 persons-month





Data & Feature Engineering

- COSMOS 2020 (Weaver et al. 2022a)
- SC4K 2018 (Sobral et al. 2018a)

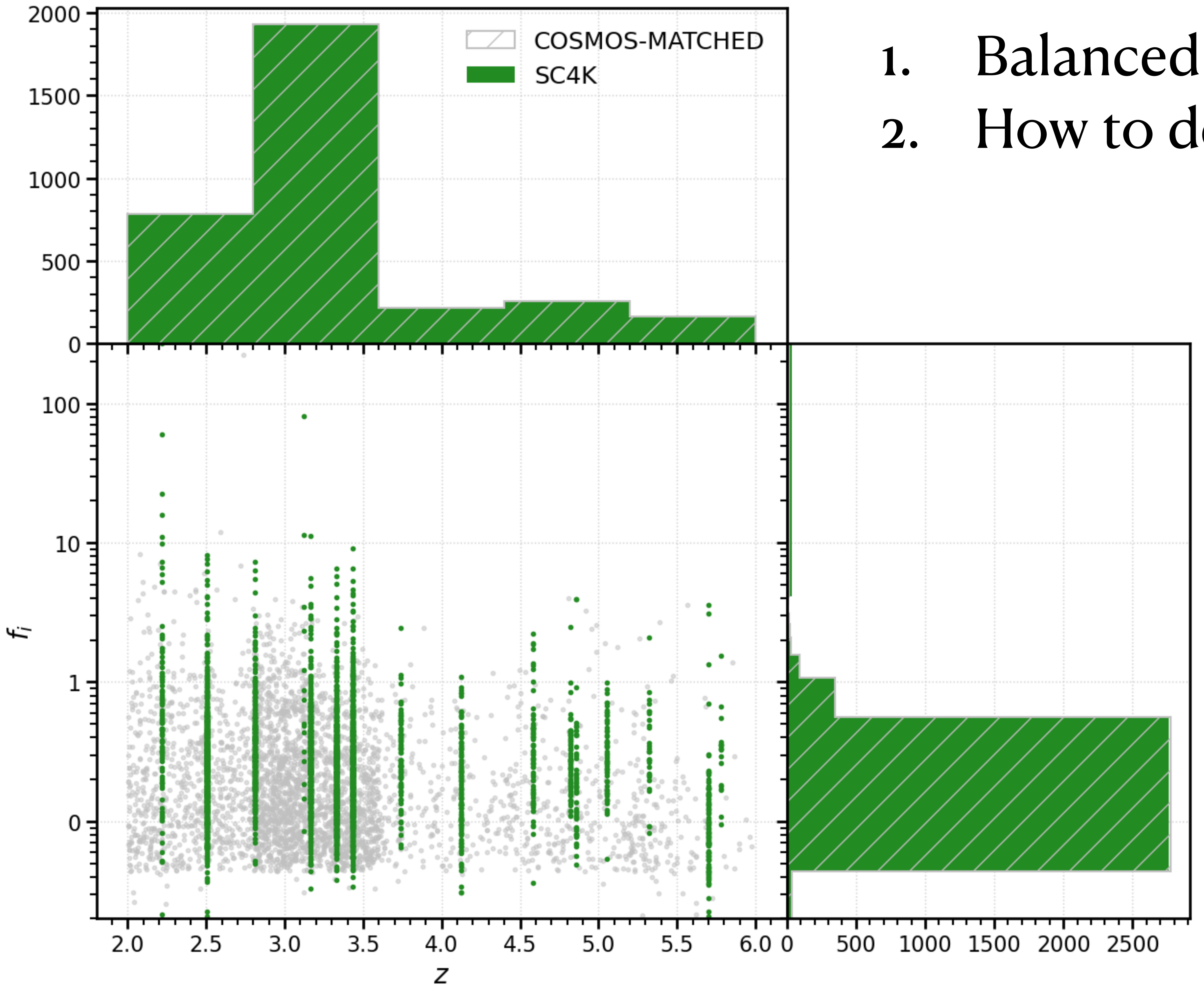


COSMOS field
Credit: NASA, ESA, and Z. Levay (STScI)

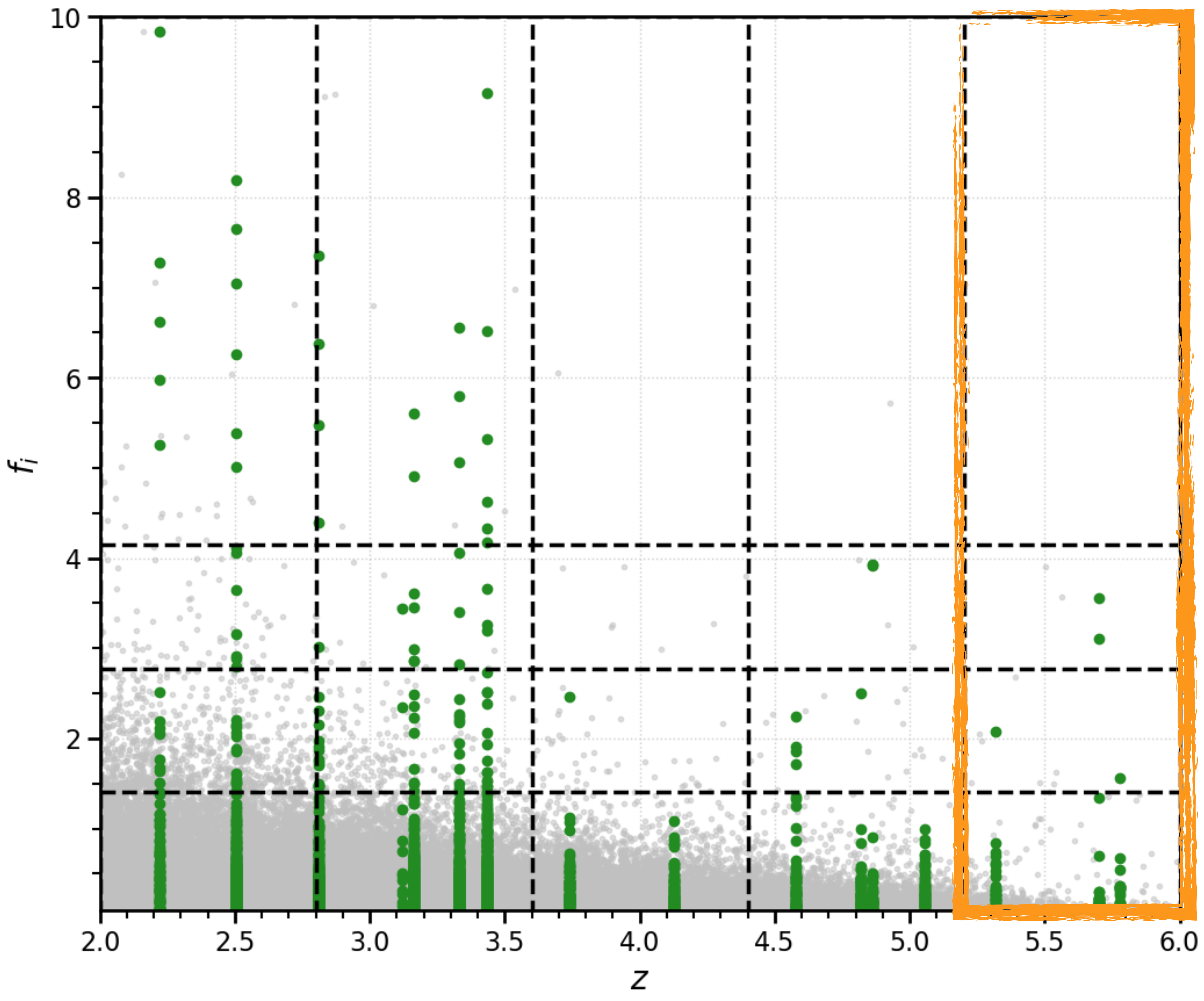
Optical to Near-Infrared

Fluxes + Uncertainties & Normalisation (unit-norm)

Data & Feature Engineering

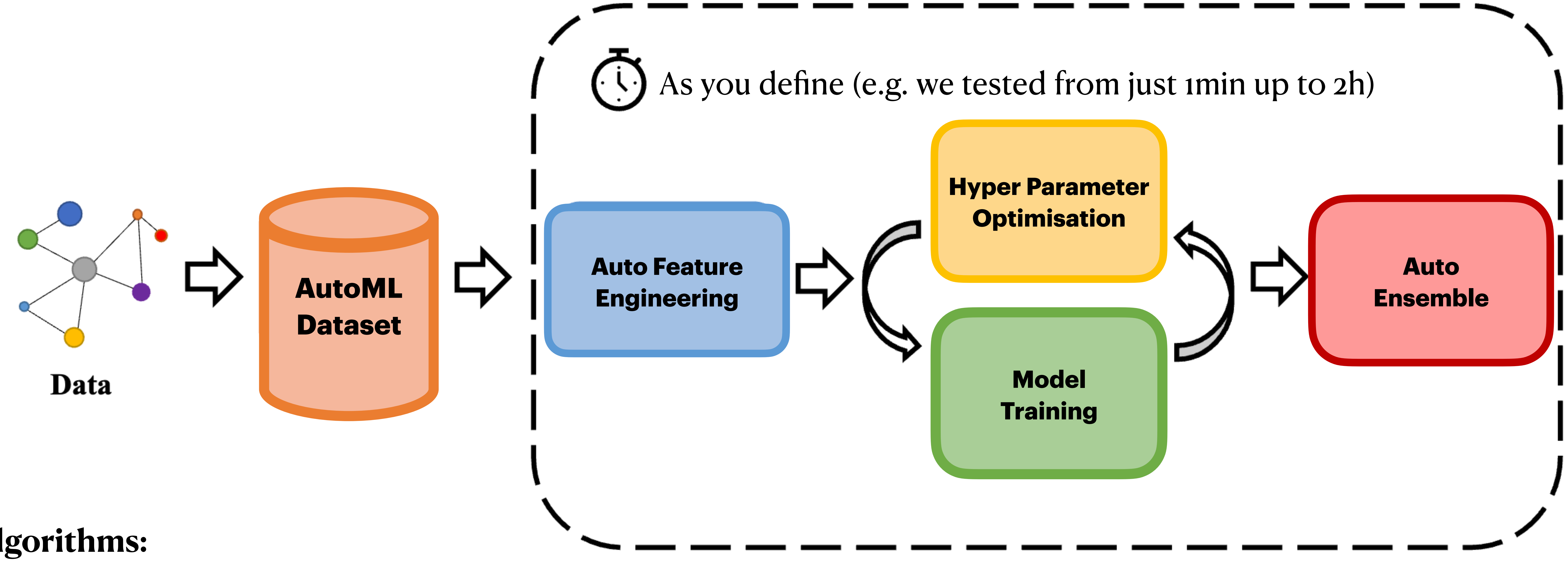


1. Balanced or unbalanced data?
2. How to define a proper sample of non-LAEs?



Methods: H20 AutoML

LeDell & Poirier 2020



Algorithms:

- Distributed Random Forest (DRF, including both Random Forest and Extremely Randomised Trees (XRT) models)
- Generalised Linear Model (GLM)
- XGBoost
- ***Gradient Boosting Machines (GBM)***
- DeepLearning
- ~~Stacked Ensemble~~

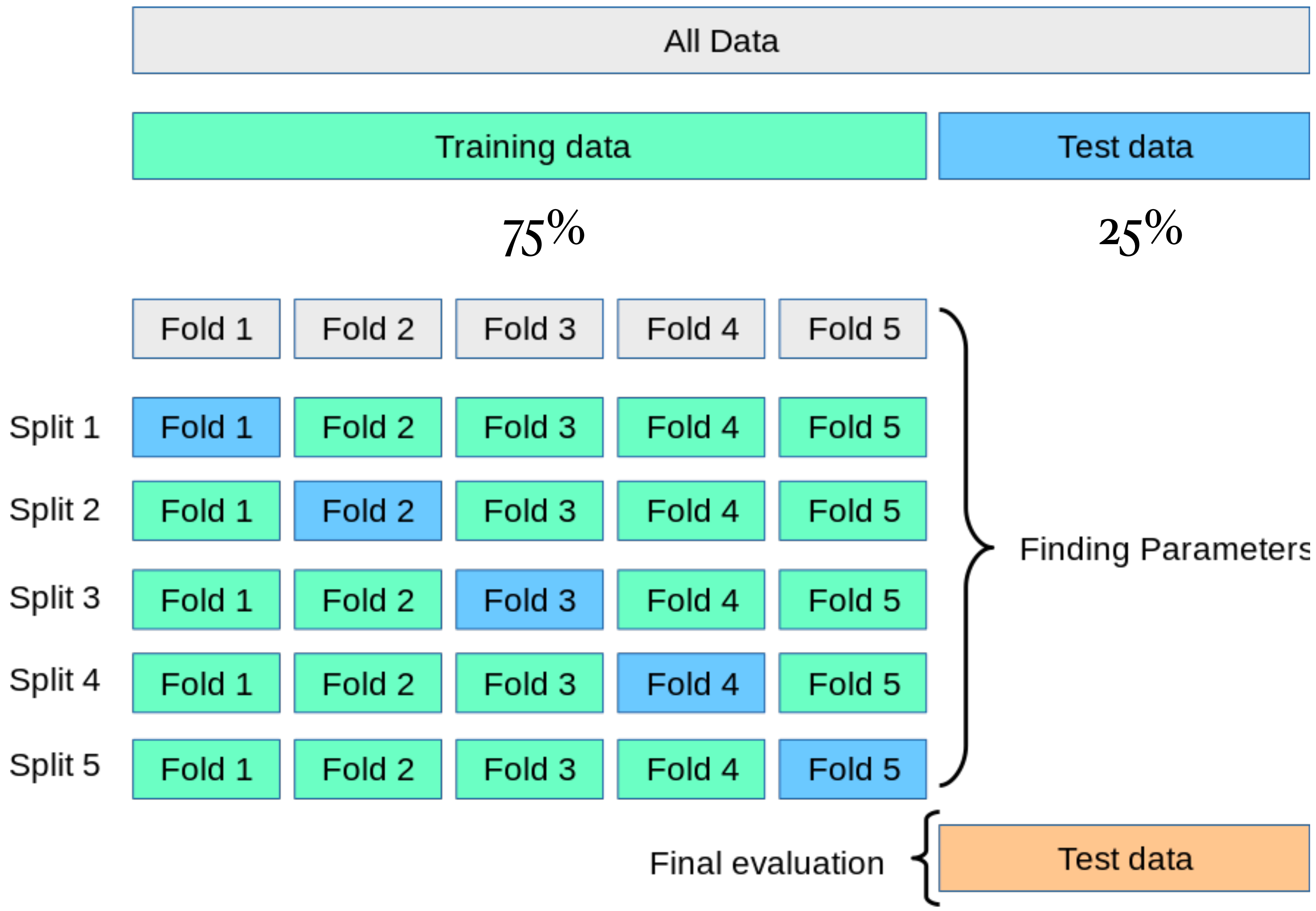
Classification
Regression

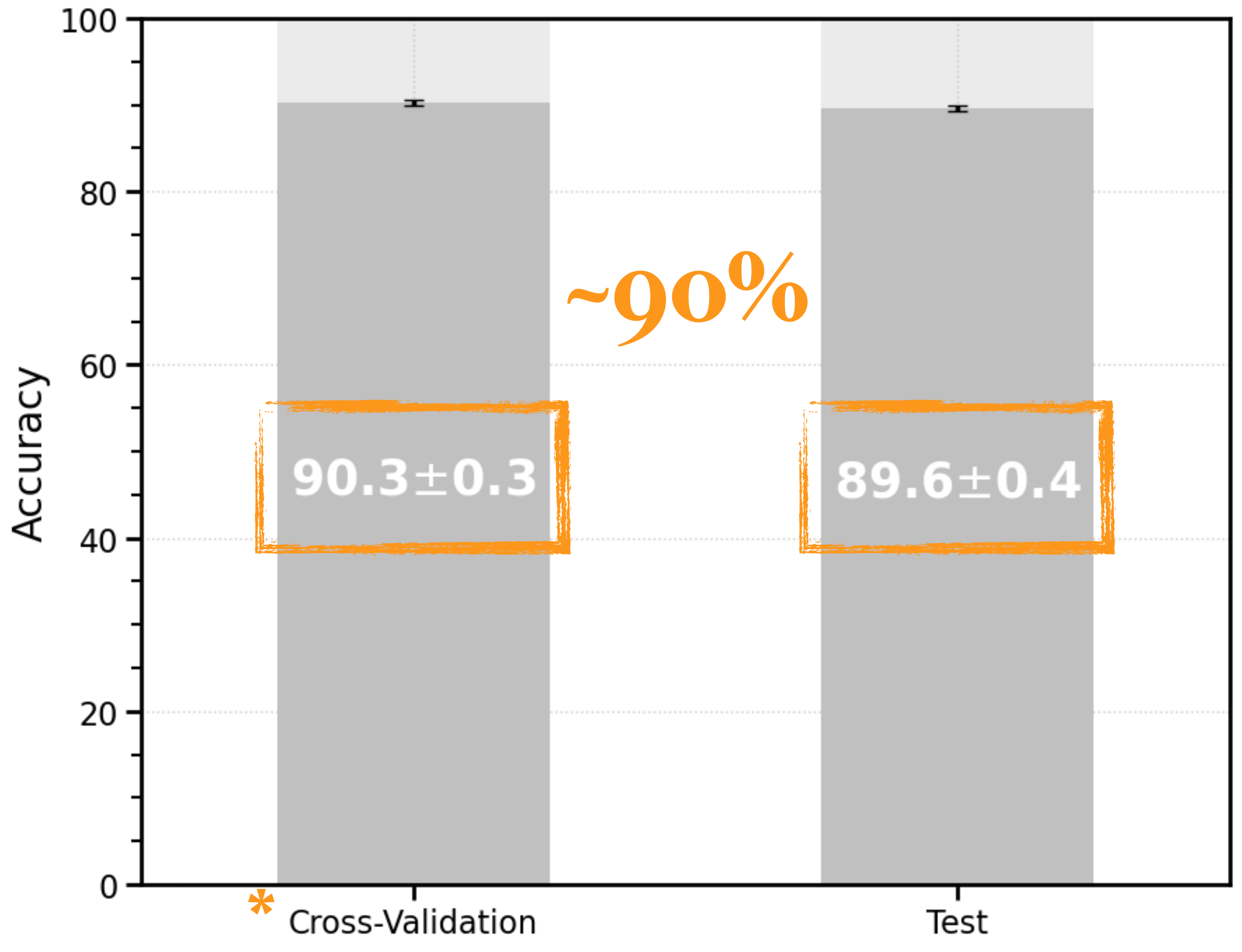
Supervised
Unsupervised

Tabular (**see Afonso's talk!**)
Image (**see Bruno's talk!**)
...

AutoML: evaluating the classification model

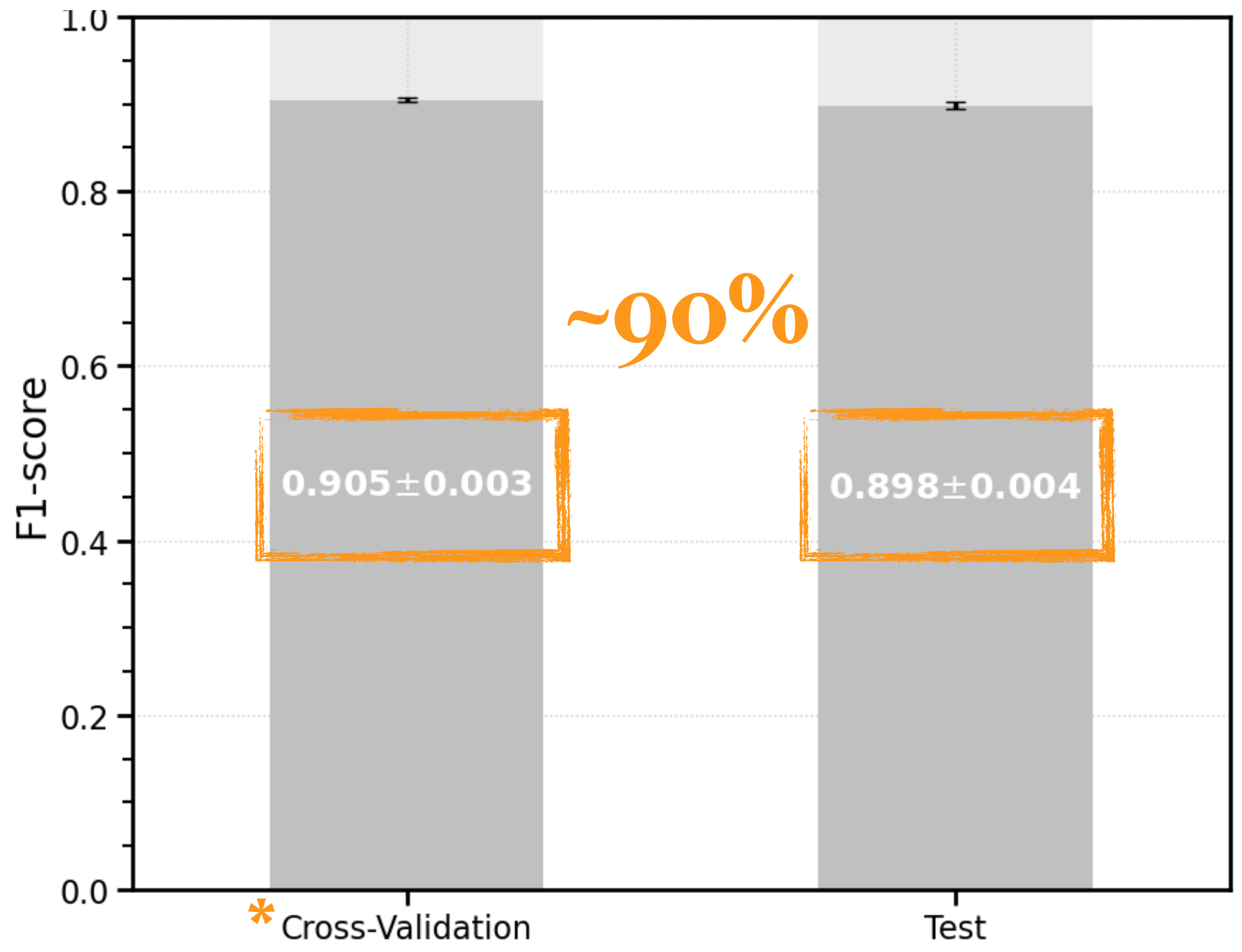
* ***Cross-validation***: a resampling method that uses different portions of the data to test and train a model on different iterations. **Important** to protect against overfitting, particularly when the amount of data is limited.





Accuracy is the ratio of predictions that exactly math the true class labels.

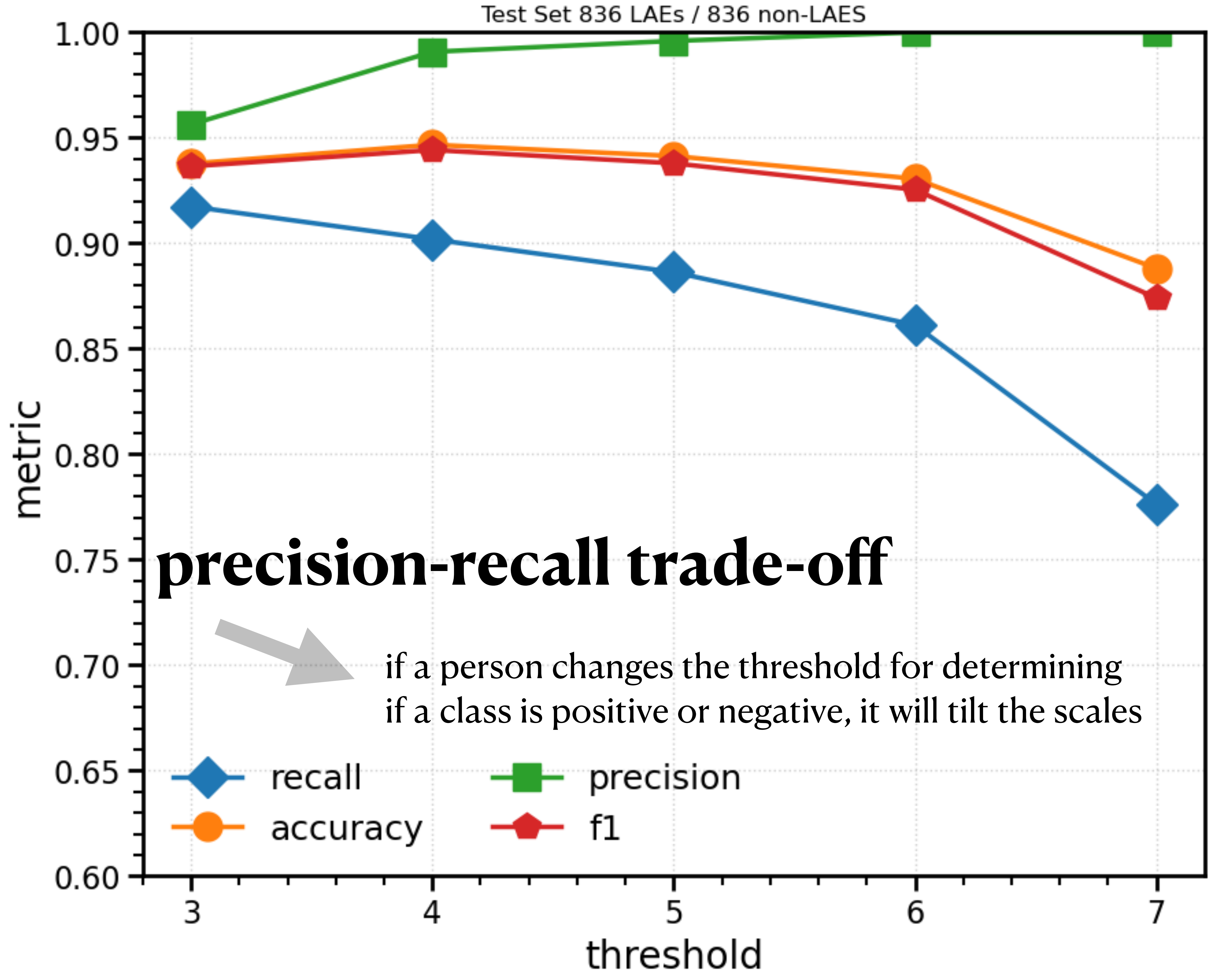
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

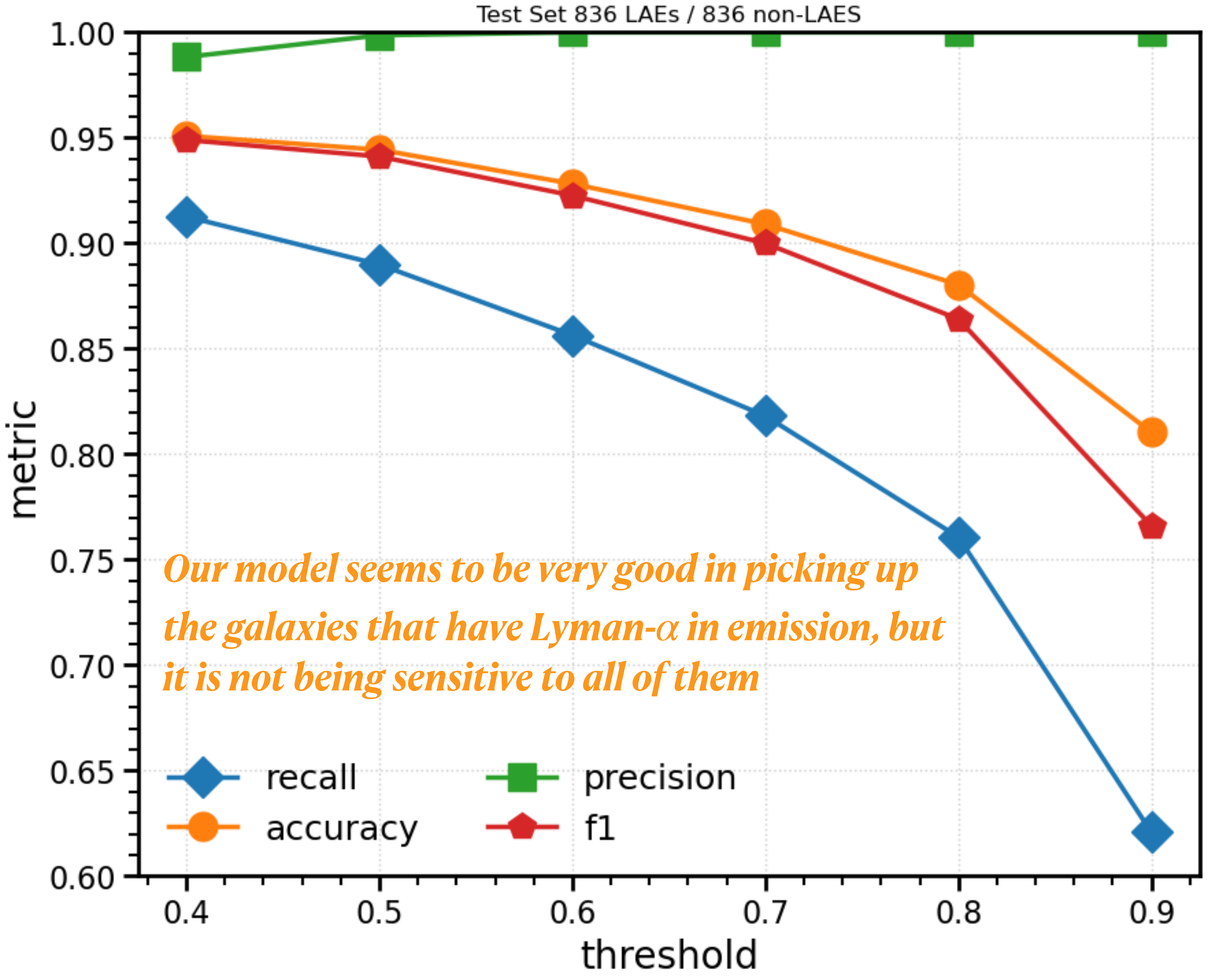


F1-score is a good balanced measure of both false positives and false negatives.

$$F1\ \ Score = \frac{2TP}{2TP + FP + FN}$$

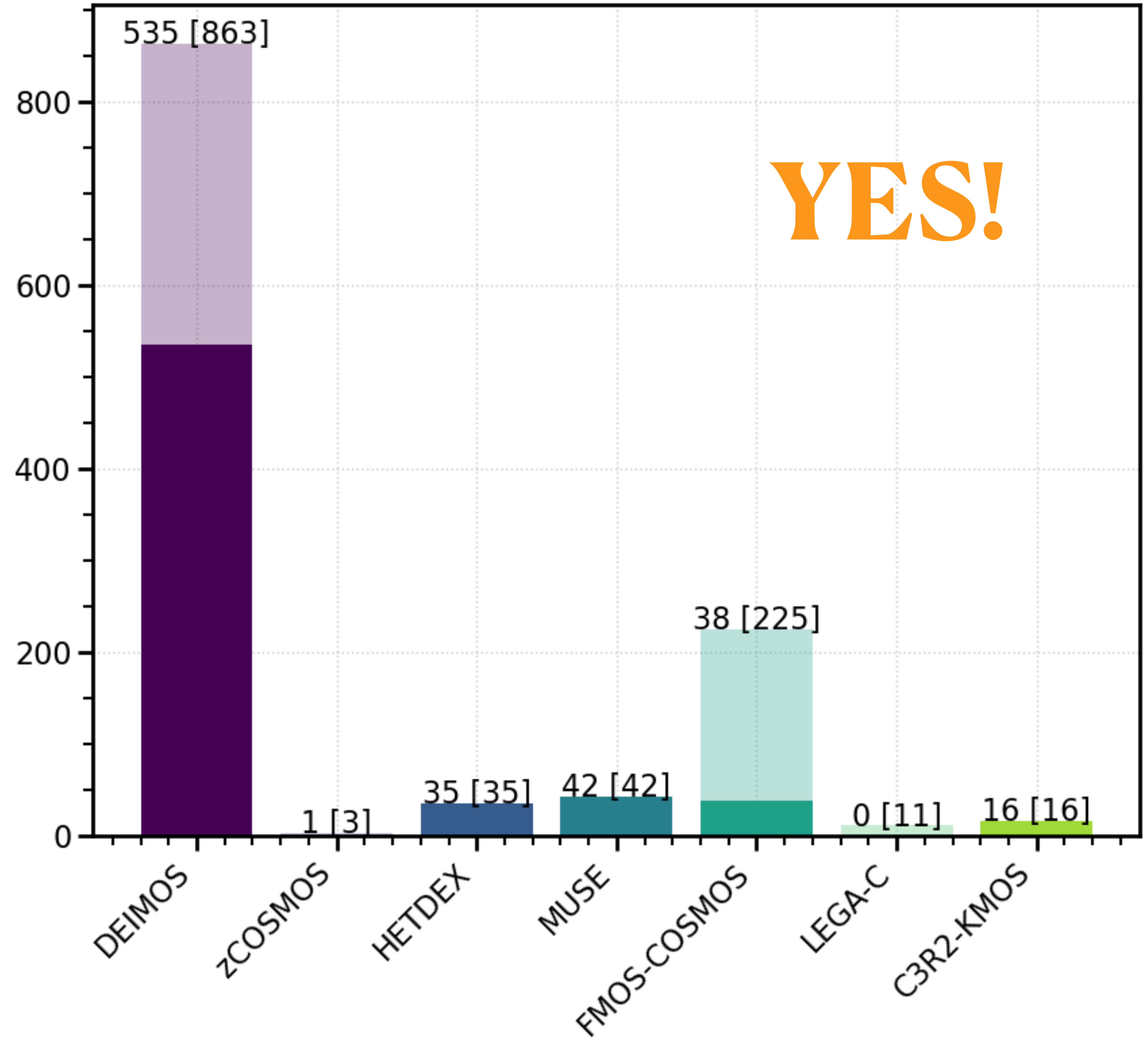
AutoML: evaluating the classification model



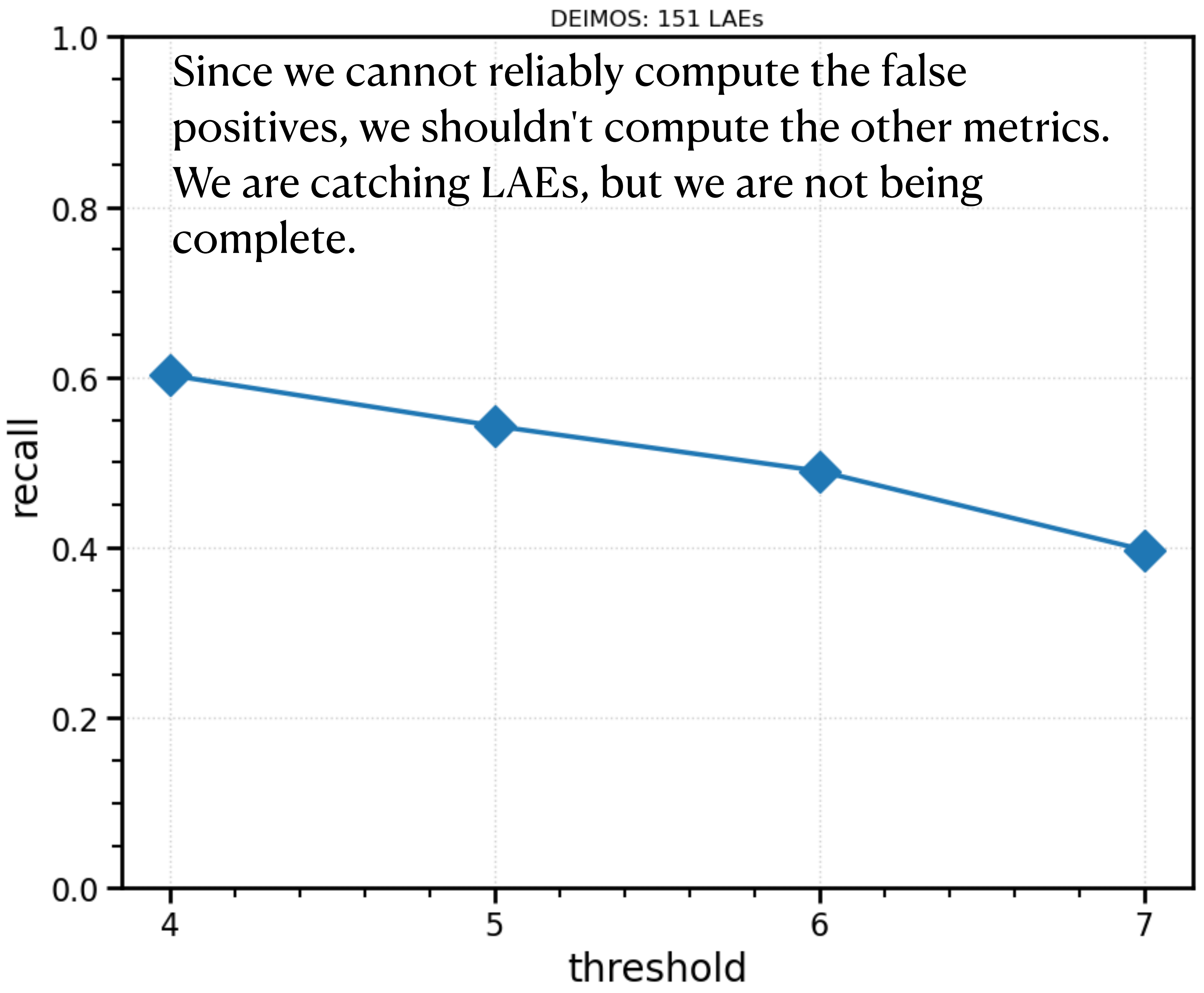


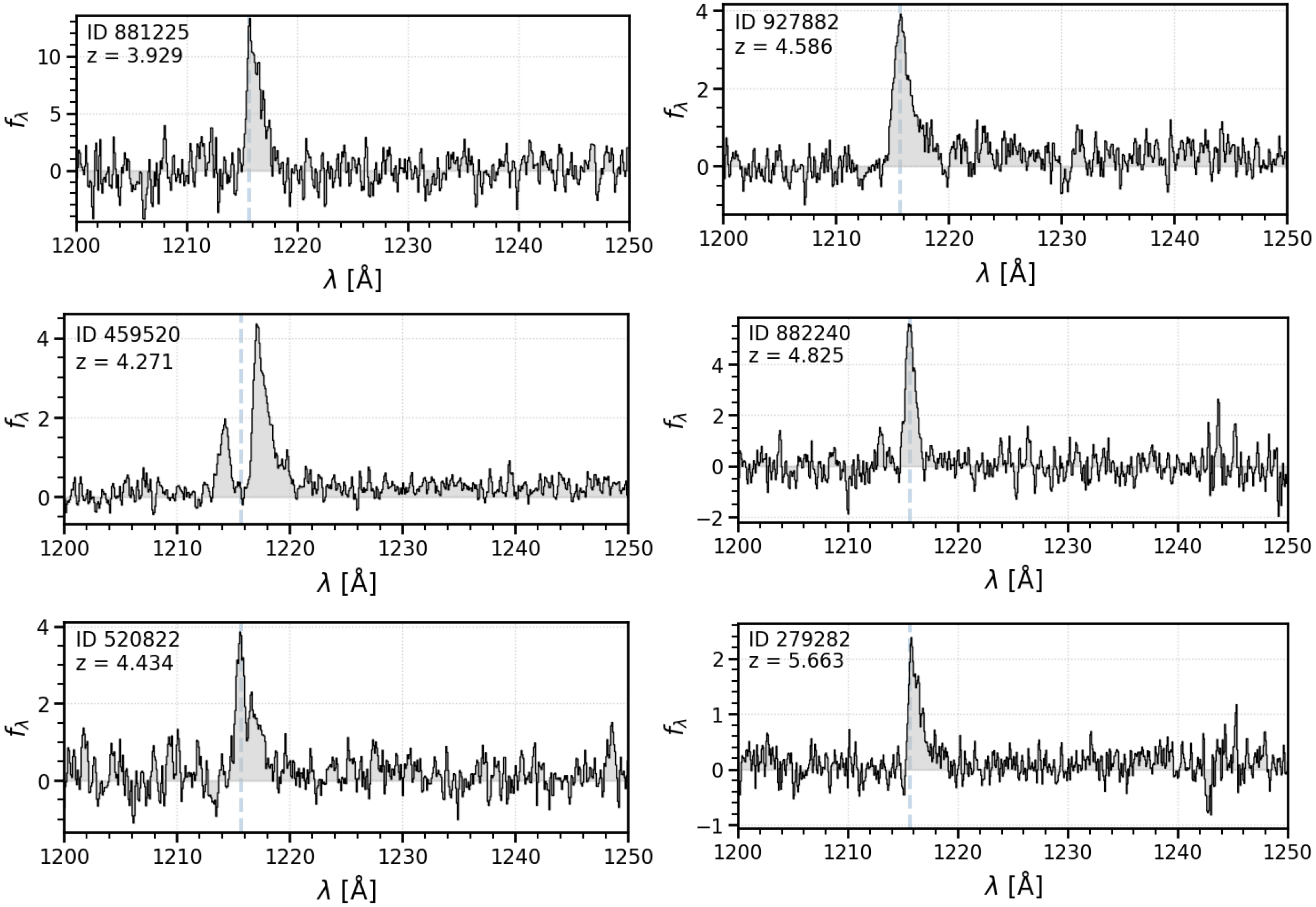
Results: how many LAEs are probably escaping us?

Results: can we spectroscopically confirm some of them?



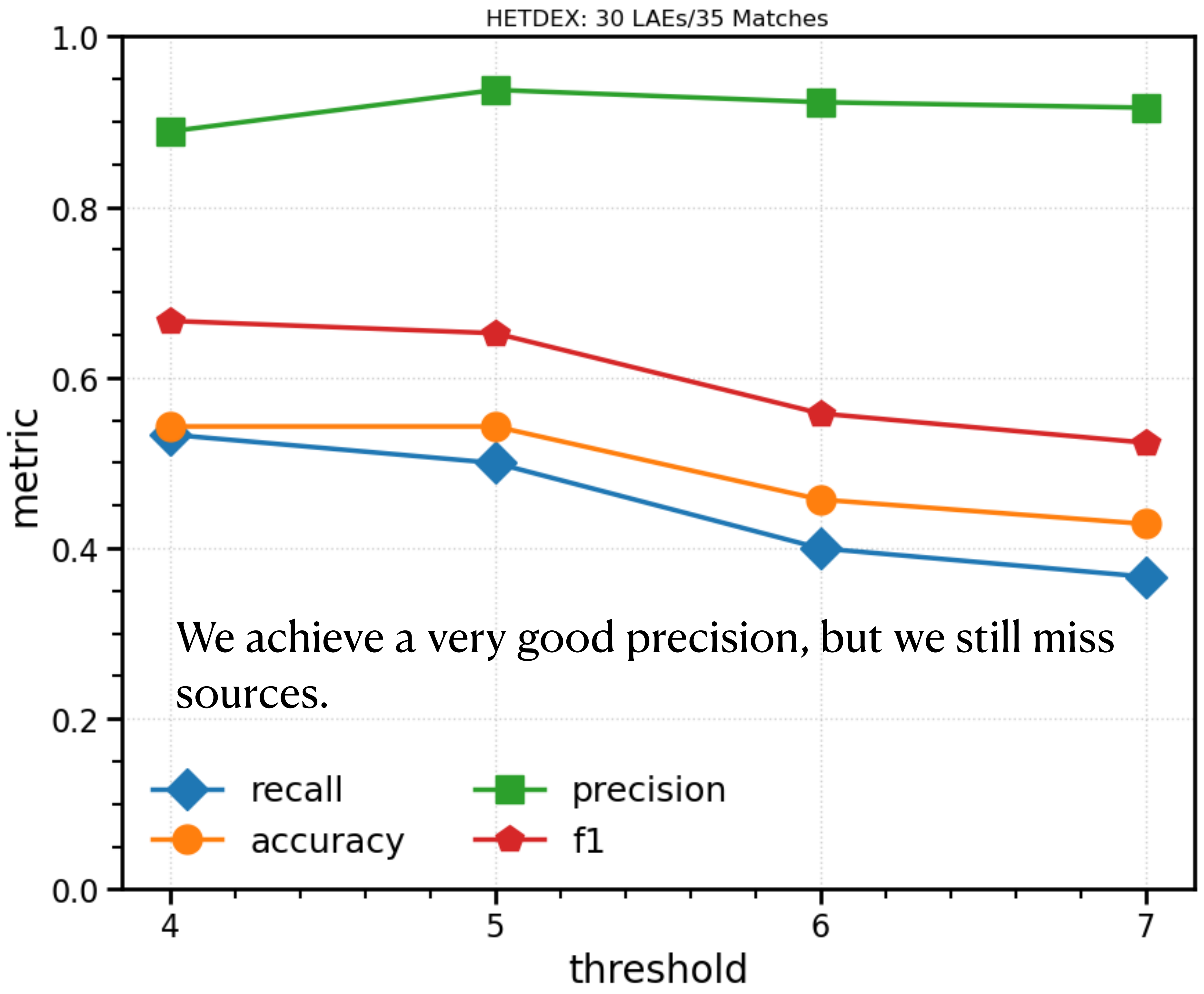
Results: DEIMOS spectroscopic confirmations



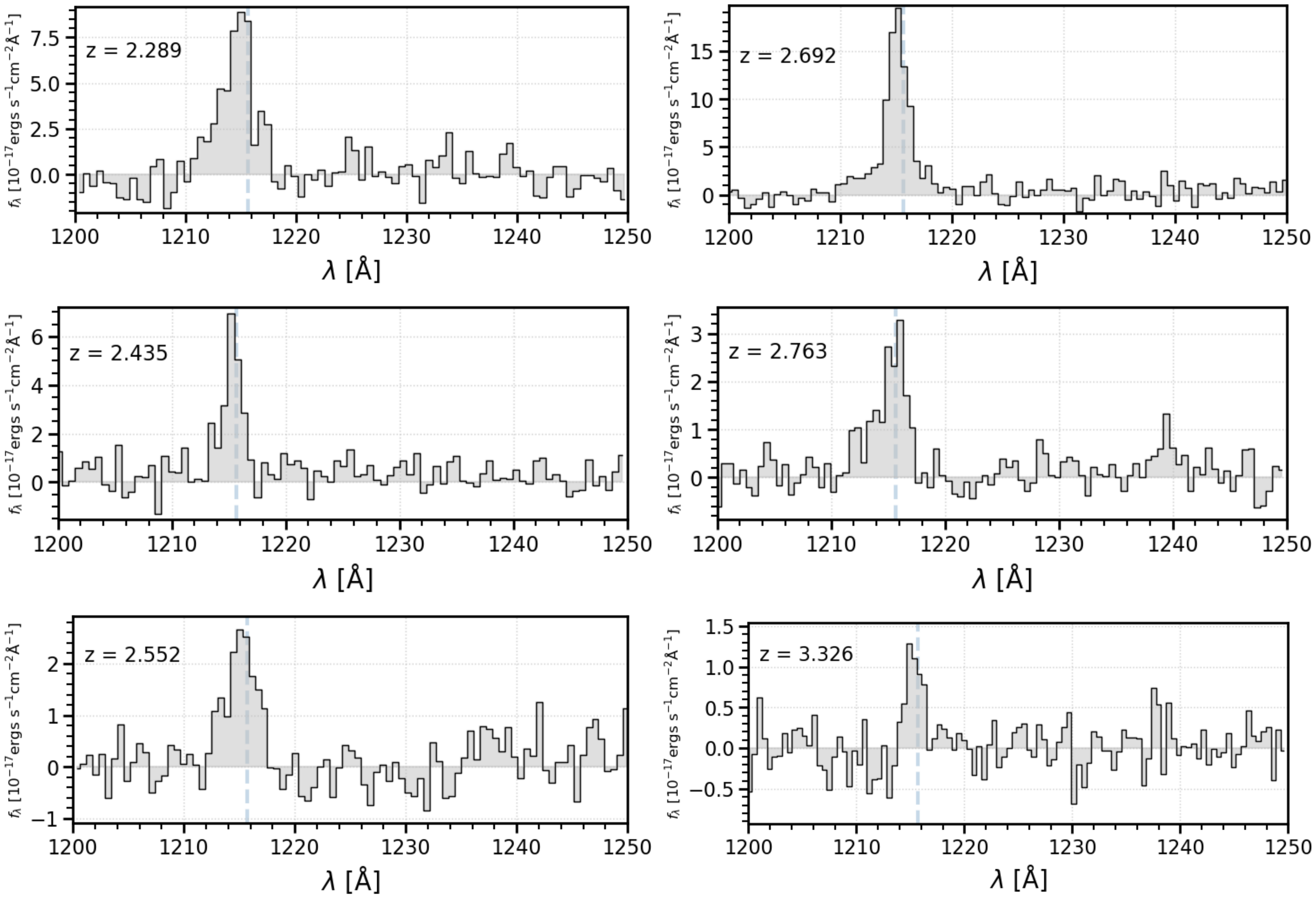


Hasinger et al. 2018

Results: HETDEX spectroscopic confirmations

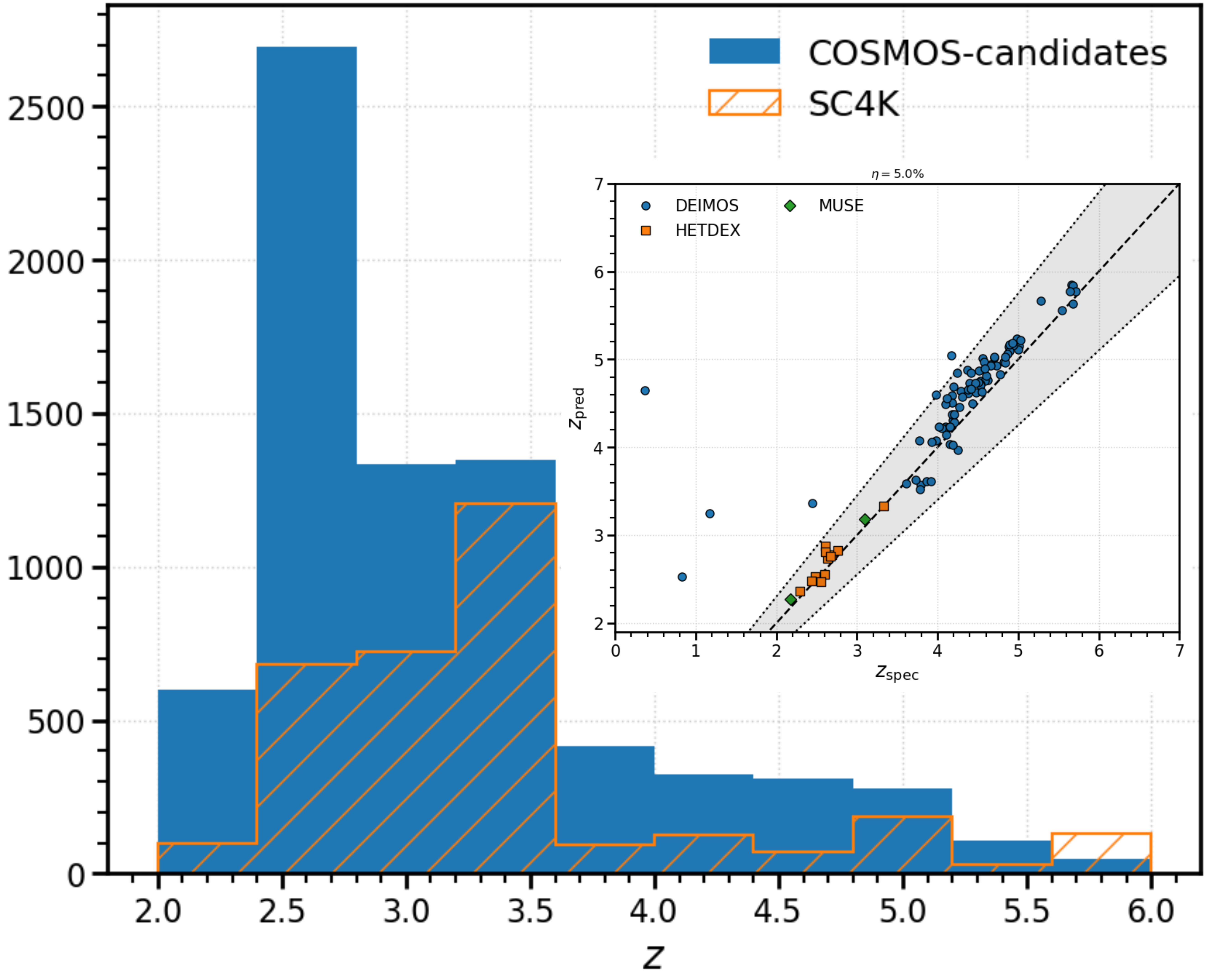


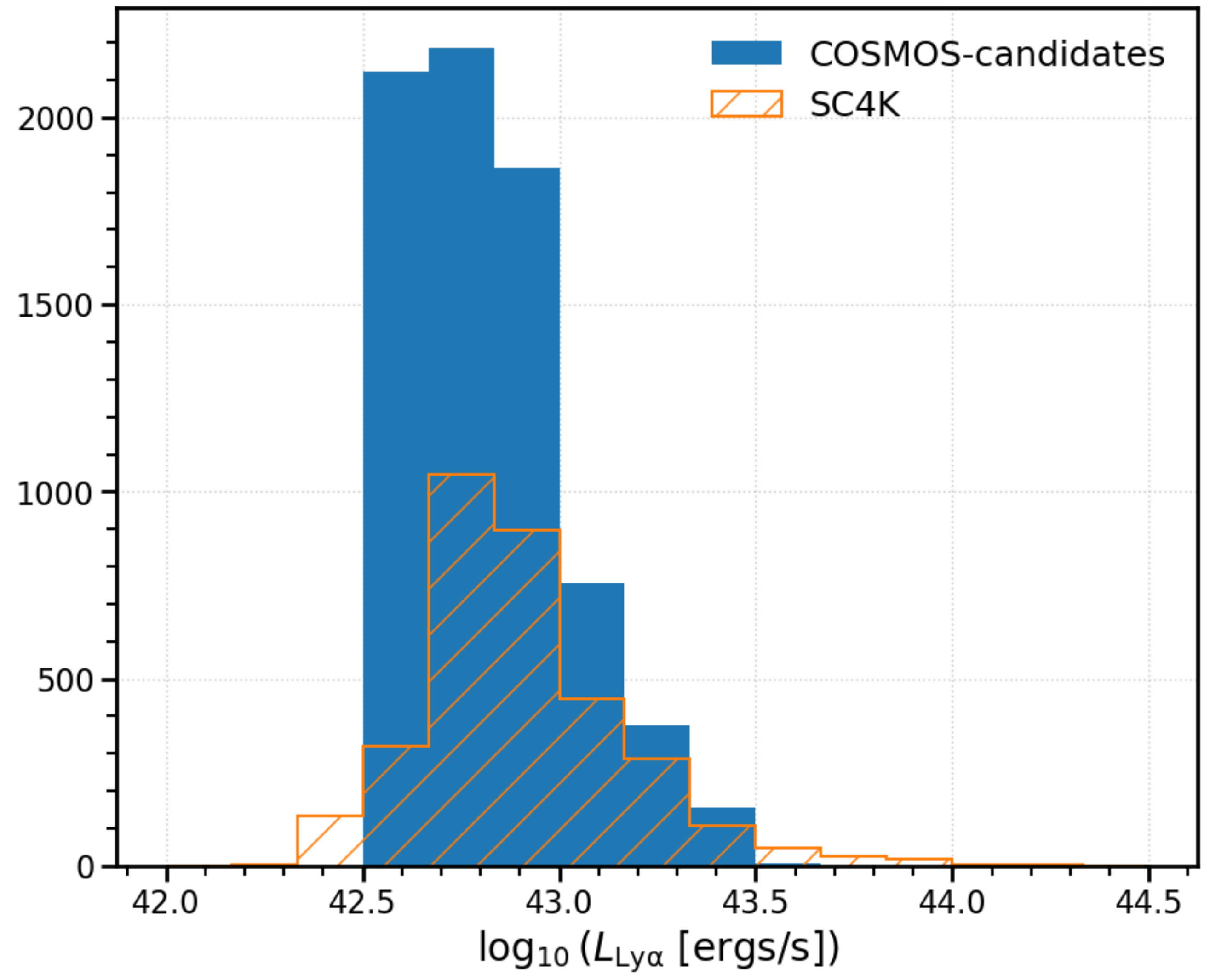
We achieve a very good precision, but we still miss sources.

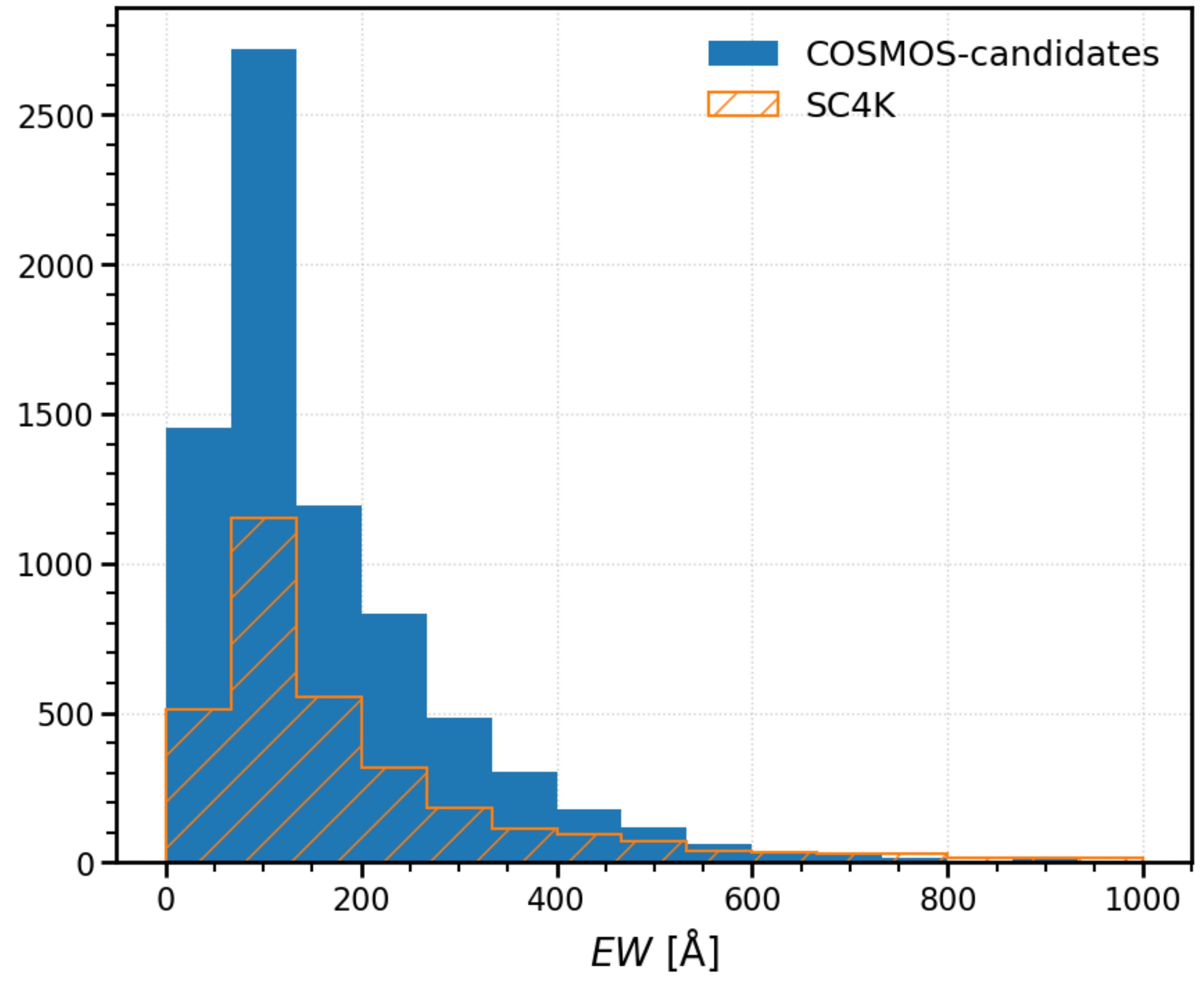


Cooper et al. 2023

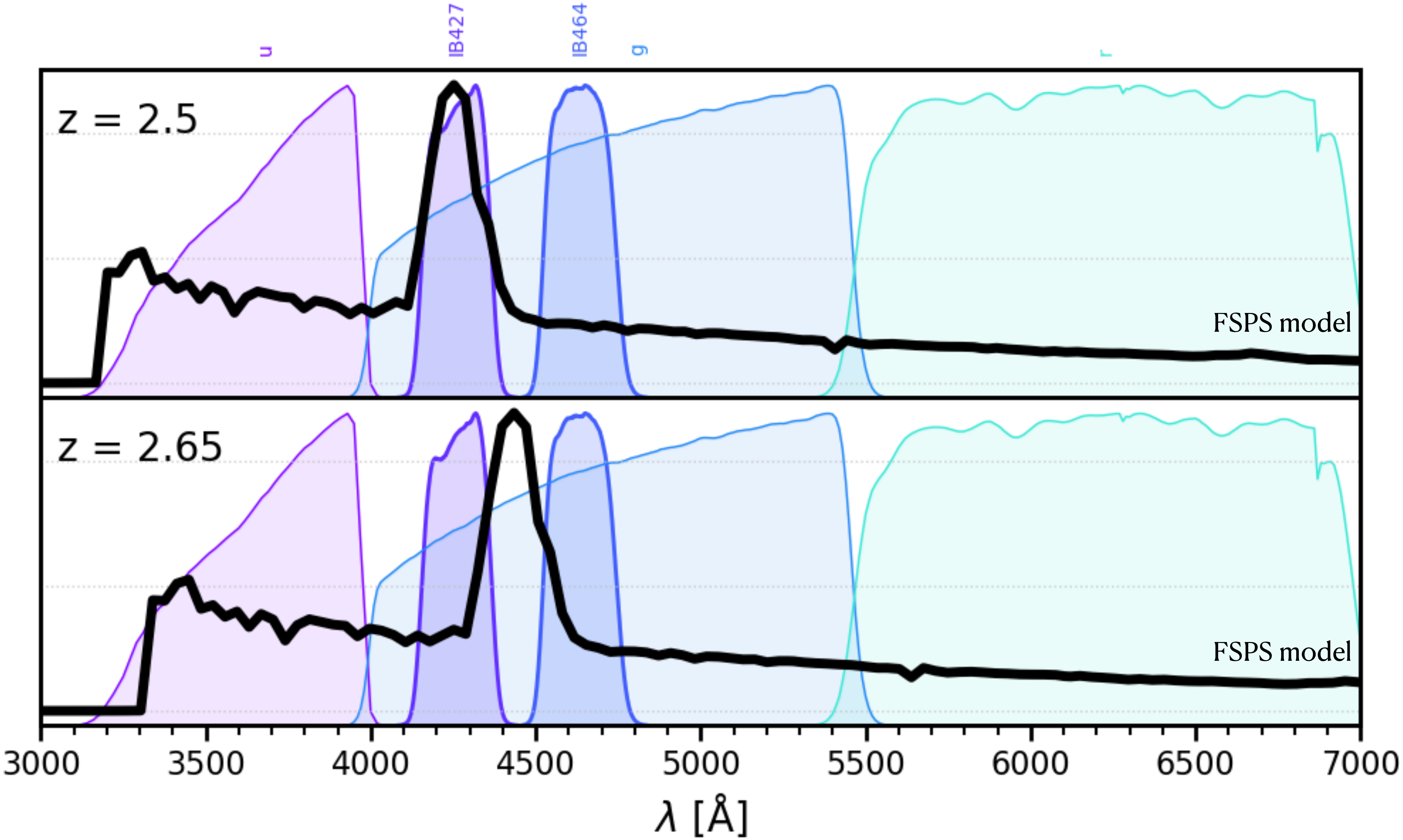
Results: predicting their zs, $L_{Ly\alpha}$, and EWs



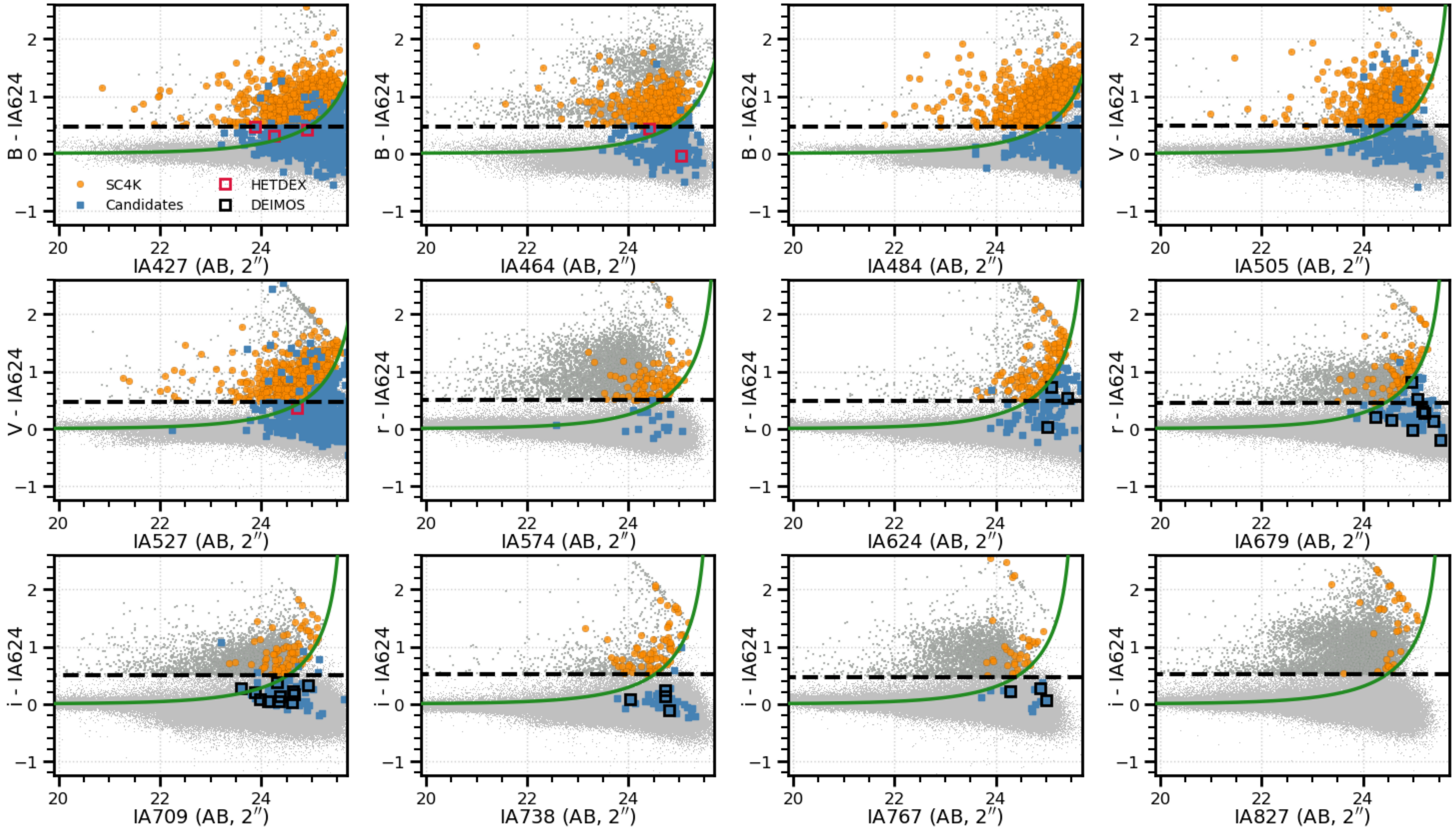




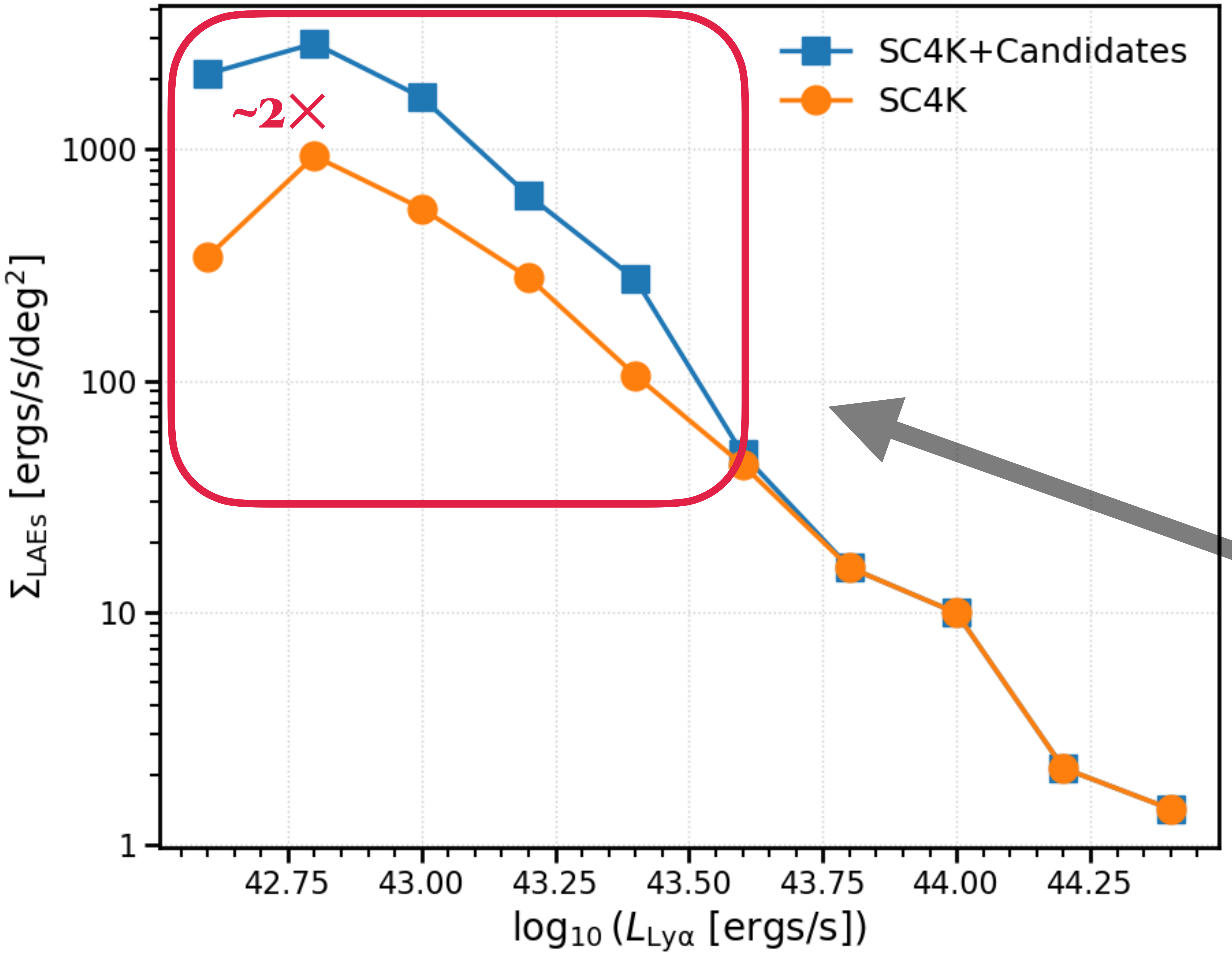
Discussion: why are we missing them?

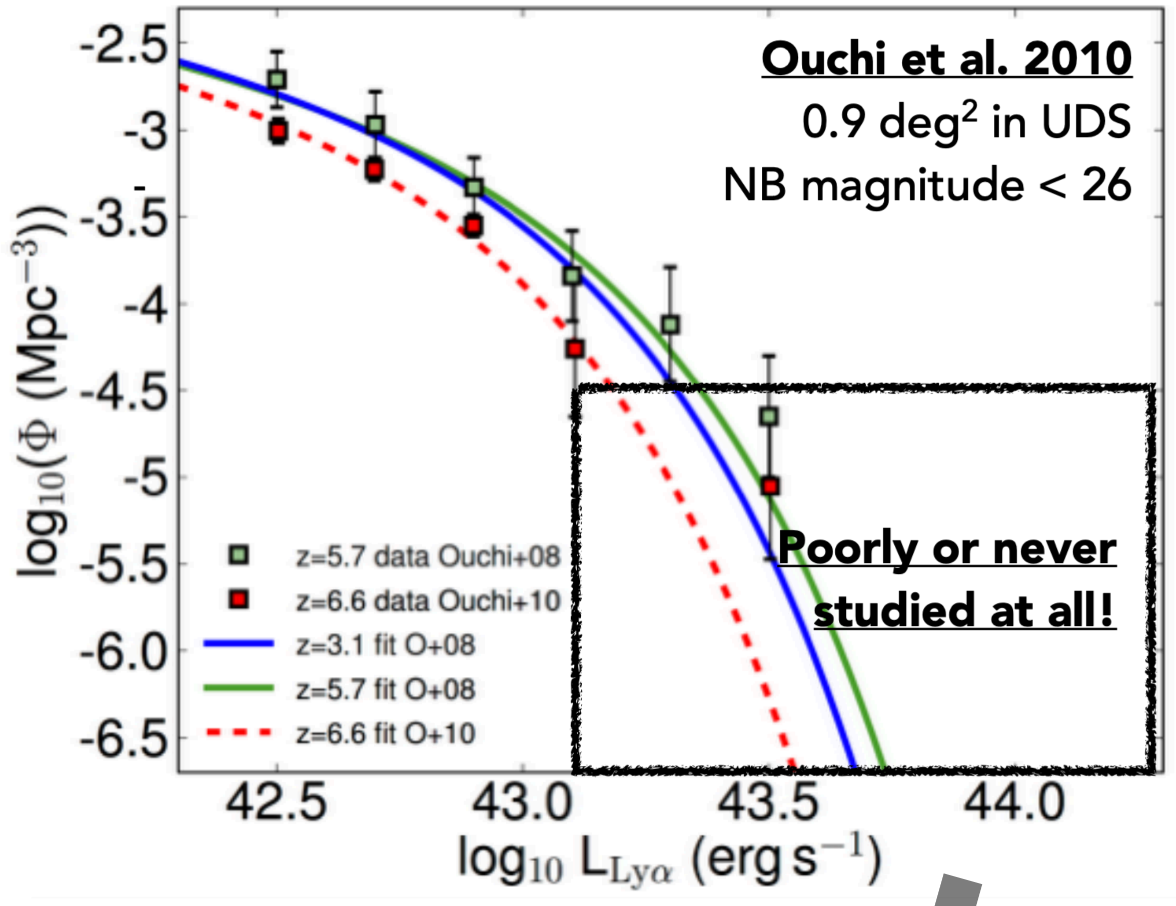


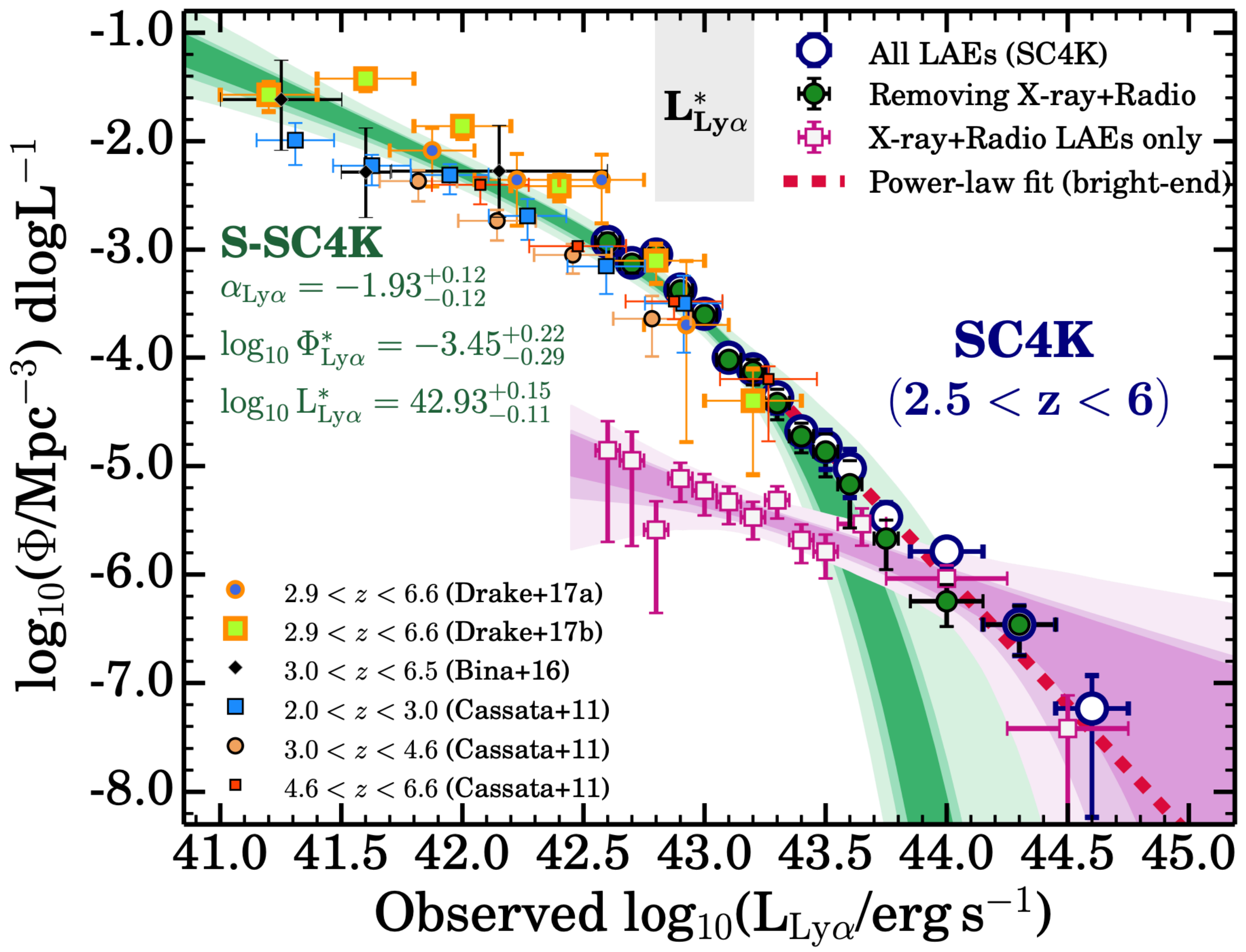
Discussion: why are we missing them?



Discussion: what are the implications?







Take-home messages

- AI can be used to increase samples of LAEs.
- We are getting pure sample, but we are still far from having complete samples.
- We can also predict basic properties of LAEs which is useful for effective target selection.
- We still need to prove generalisation capabilities and discuss uncertainties.

Future directions

FLAEMING: Finding Lyman-α emitters through machine learning

T-FLAEMING: Transcoding the way of finding LAEs in the era of large surveys and AI