



The Baryon Cycle Project: ByCycle

Detecting Cool CGM Quasar Absorbers via Contextual Anomaly Detection with Variational Autoencoders



nicolas.guerravaras@eso.org

Nicolás Guerra Varas (he/him), Céline Péroux, Paula Sánchez Sáez,
Andrea Merloni, Tamsyn O'Beirne, Rajesh Potlia, Ramona Augustin,
Hannah Bish, Emily Kerrison, Simon Weng & the ByCycle Team

ML4Astro3, September 2nd, 2026, Valletta, Malta

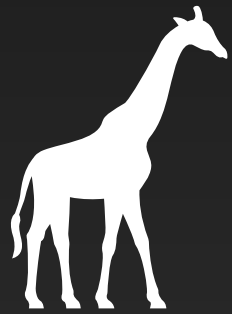




The Baryon Cycle Project: ByCycle

(Cool CGM Contextual Anomaly Detection)

The COOLCAD Pipeline for Automatic Absorber Identification



nicolas.guerravaras@eso.org

Nicolás Guerra Varas (he/him), Céline Péroux, Paula Sánchez Sáez,
Andrea Merloni, Tamsyn O'Beirne, Rajesh Potlia, Ramona Augustin,
Hannah Bish, Emily Kerrison, Simon Weng & the ByCycle Team

ML4Astro3, September 2nd, 2026, Valletta, Malta



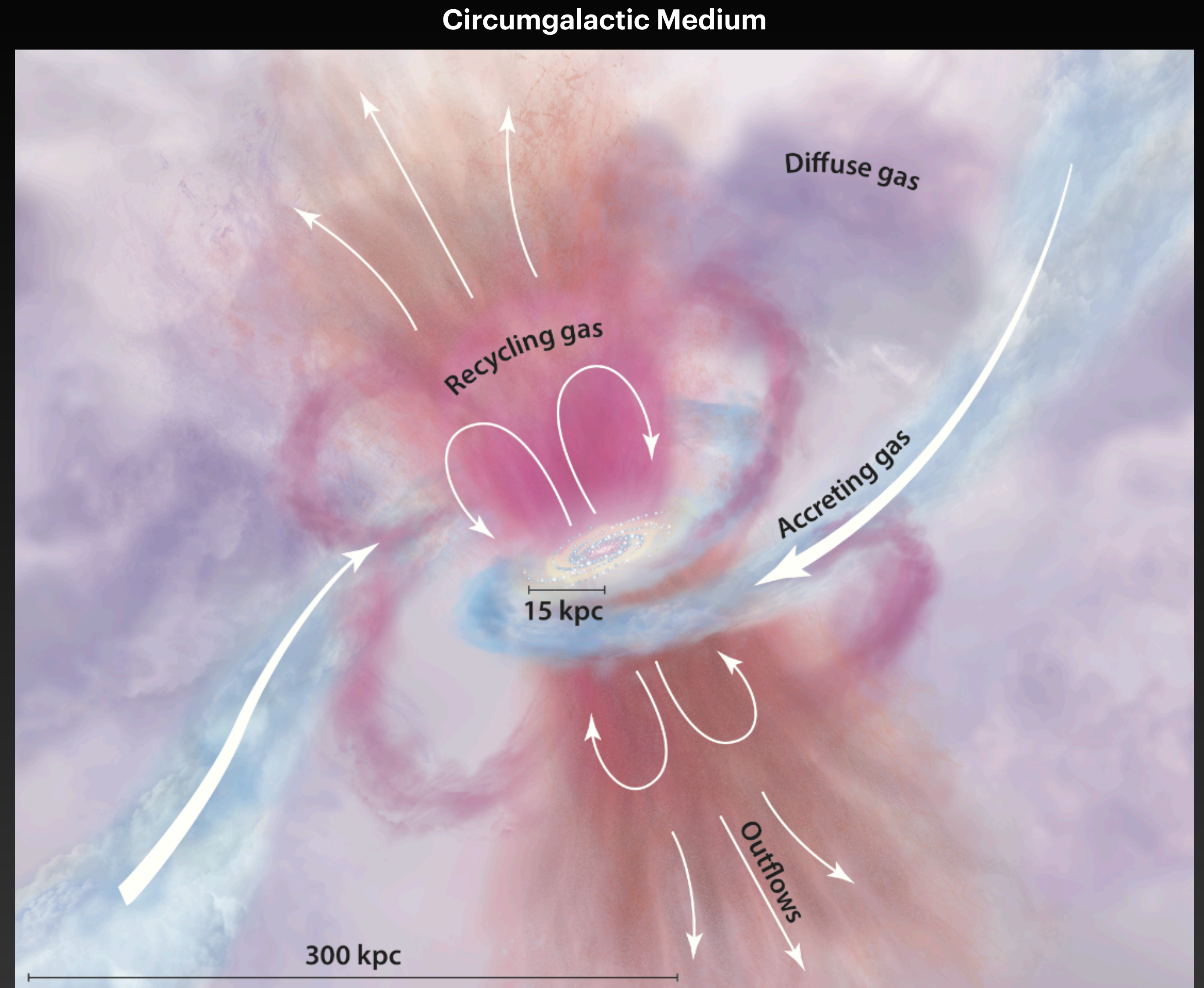
What are the baryons of the Universe up to?

The Baryon Cycle

how gas flows into, through
and out of galaxies

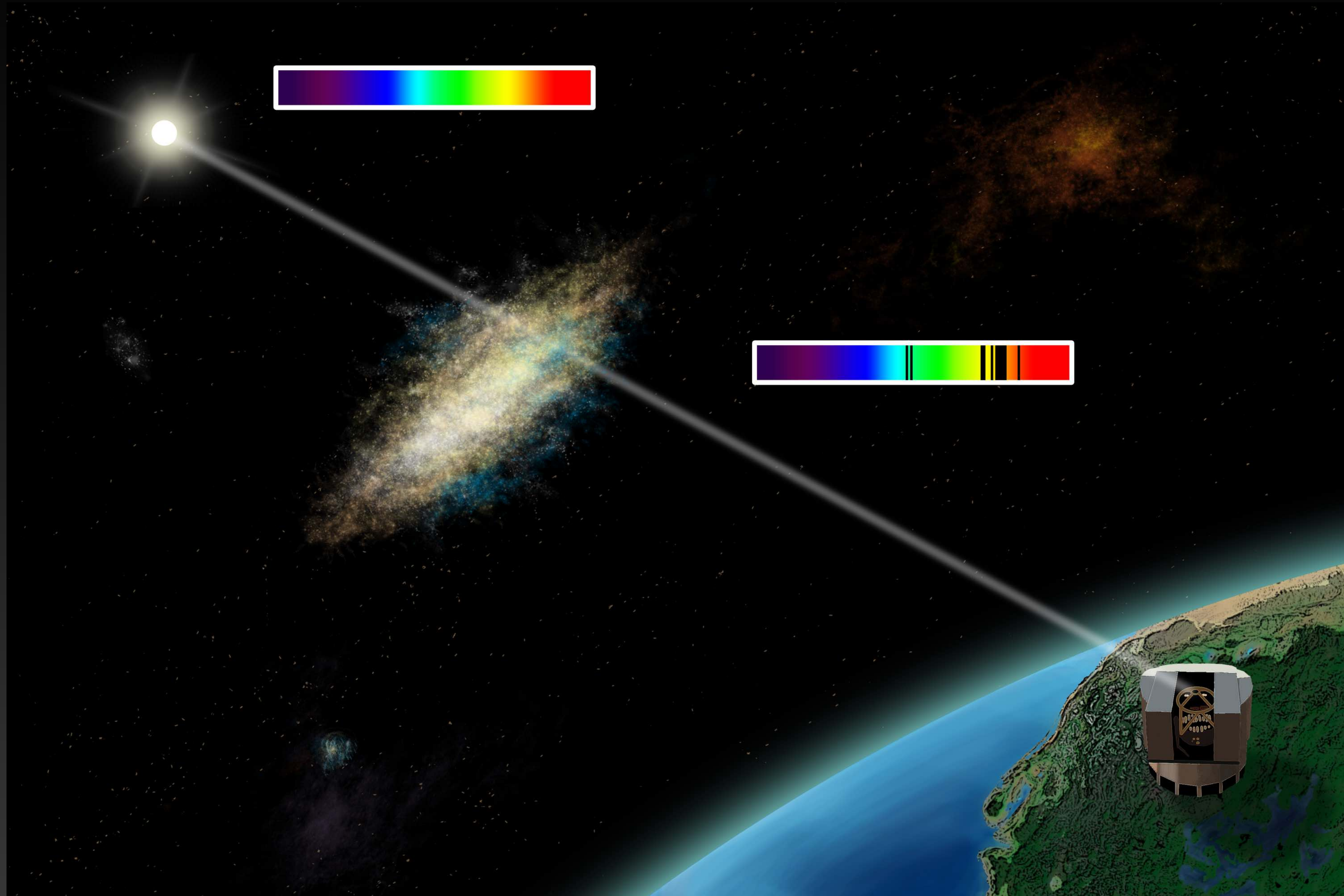
Key to understand:

- Galaxy formation and evolution
- Star formation
- Missing satellites problem
- Dark matter distr. & cosmological parameter estimation



Tumlinson, Peebles & Werk 2017

Observing Cool Gas as Quasar Absorbers



Credit: JK Krogager

4MOST

4MOST: 4-metre Multi-Object Spectroscopic Telescope

VISTA Telescope in Paranal, Chile



Credit: ESO

Two LRS at
R~6500

&

One HRS at
R~20,000

Instrument Layout

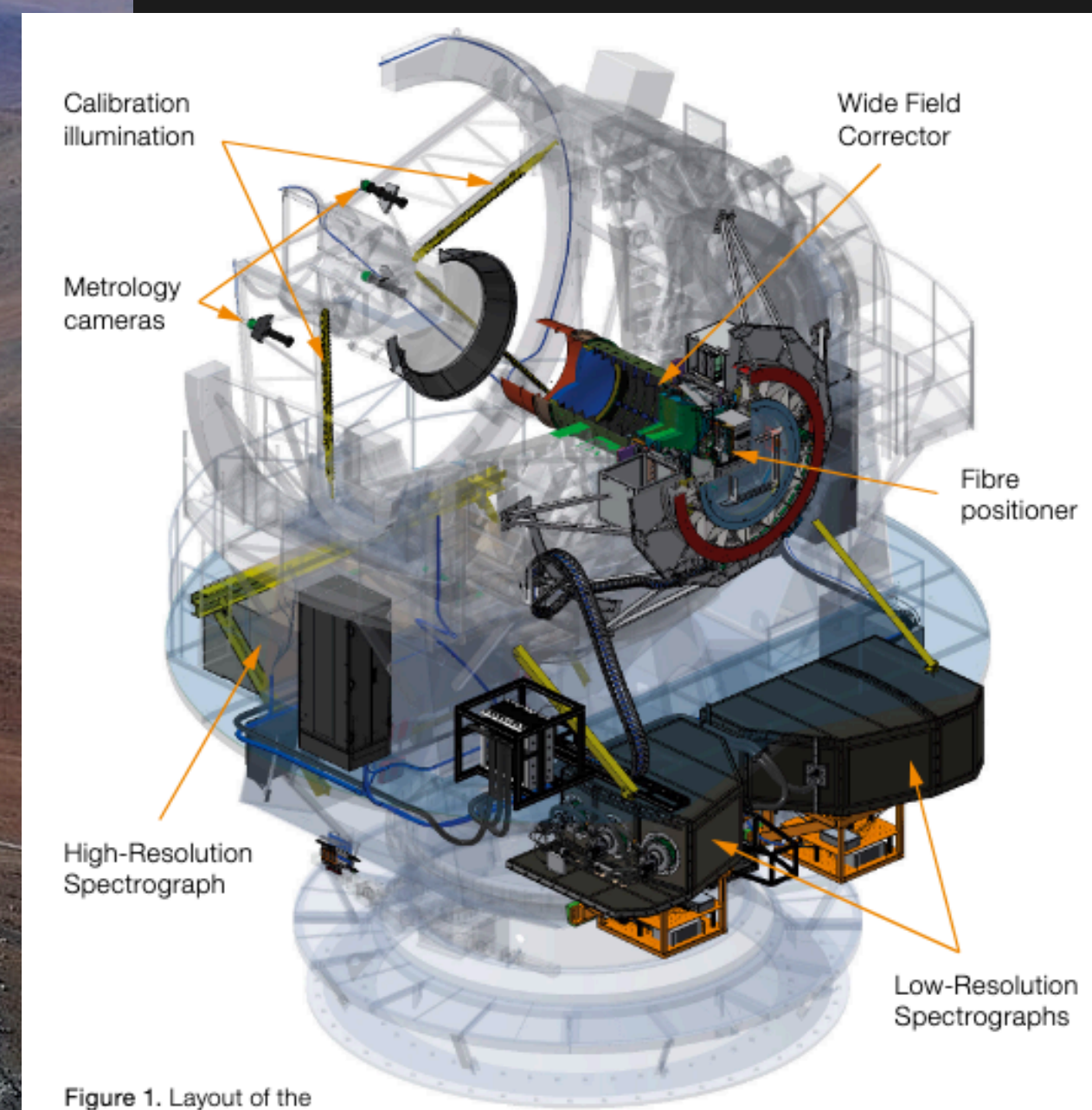
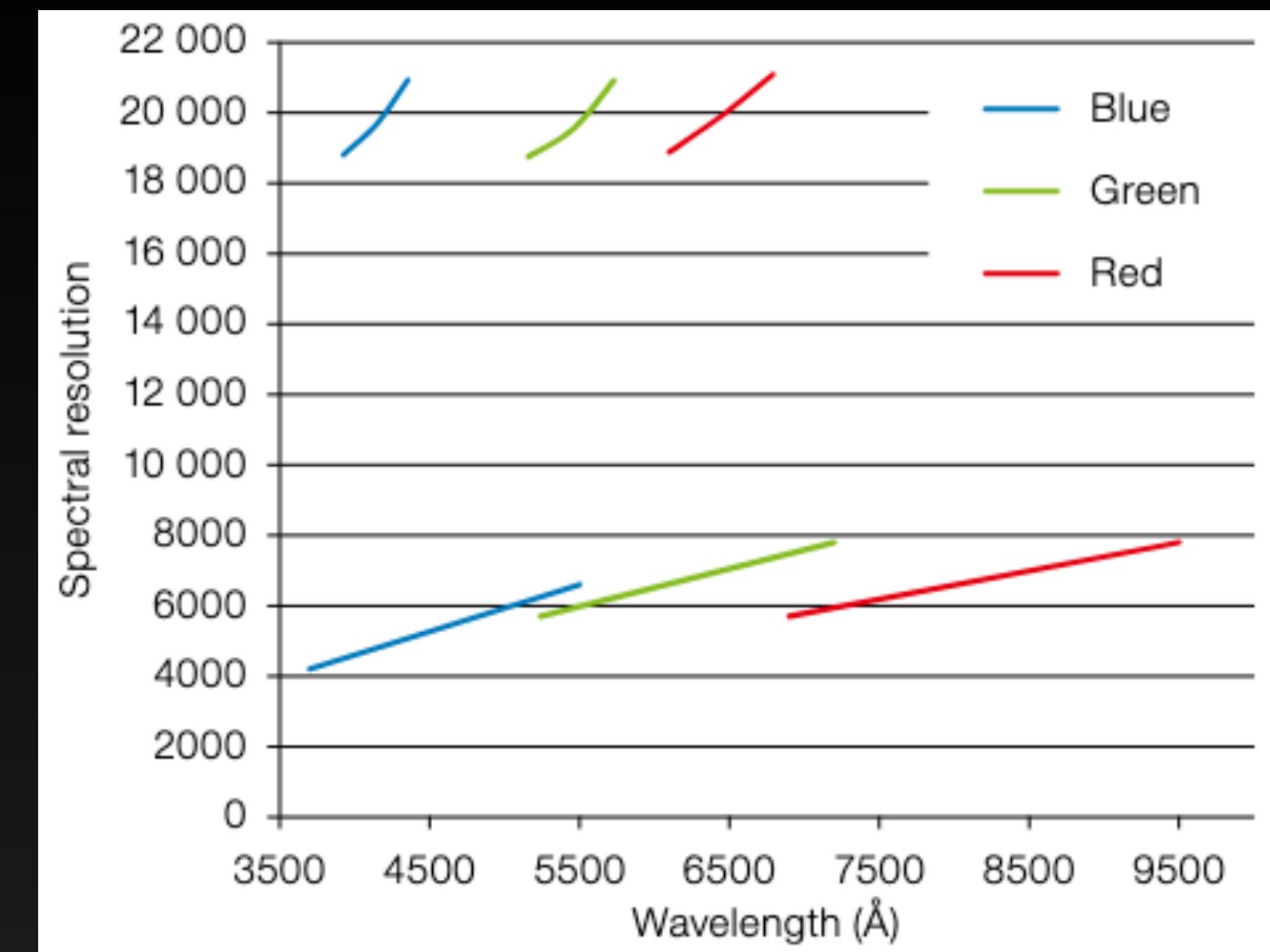


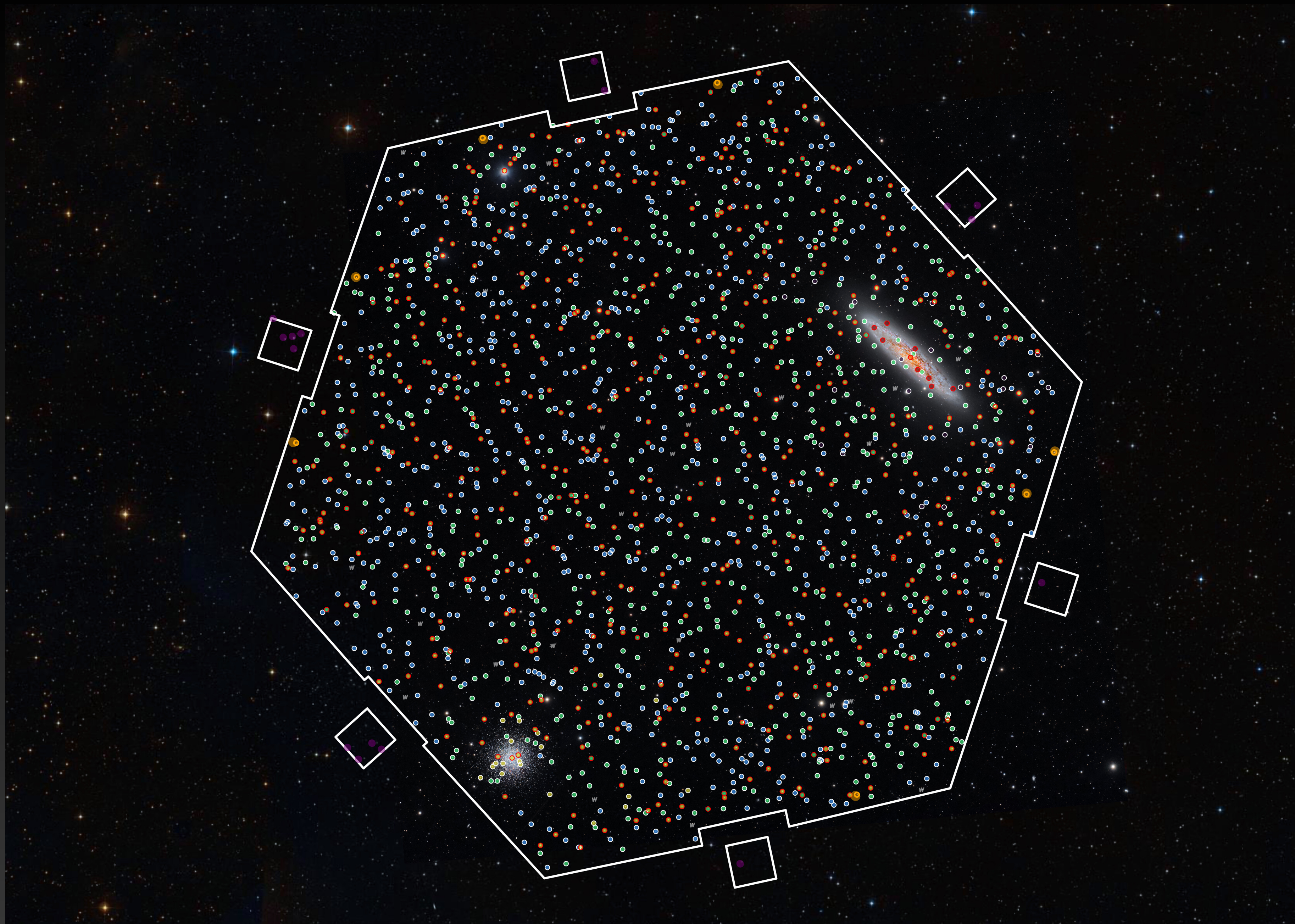
Figure 1. Layout of the

de Jong et al. 2019



de Jong et al. 2019

- FoV of ~4.2 deg²
 - 812 fibers per spectrograph
 - 18 surveys simultaneously
- Will cover >17,000 deg²
at least twice in 5 years

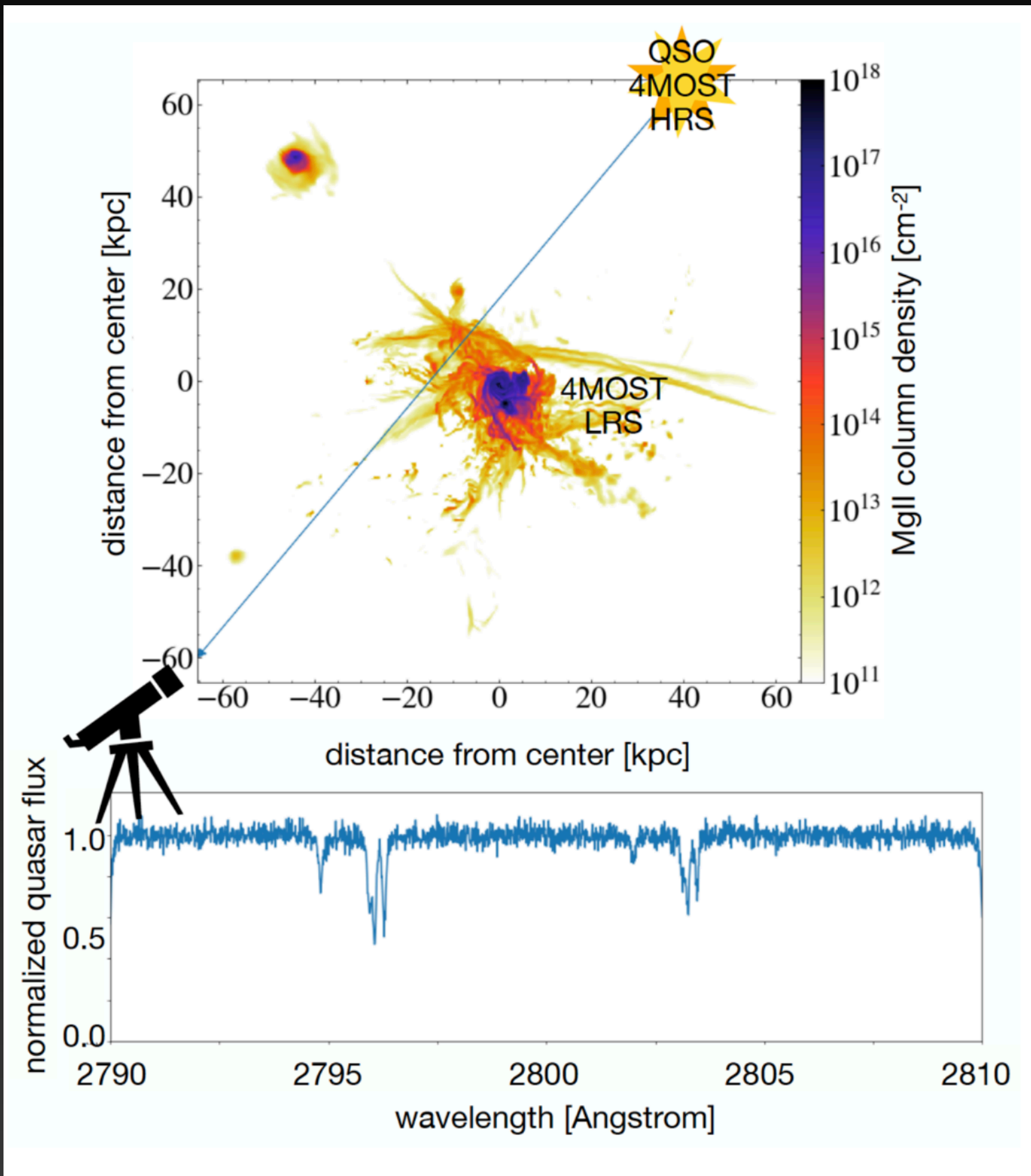


4MOST First Light (Credit: AIP/Background: Harshwardhan Pathak/Telescope Live)

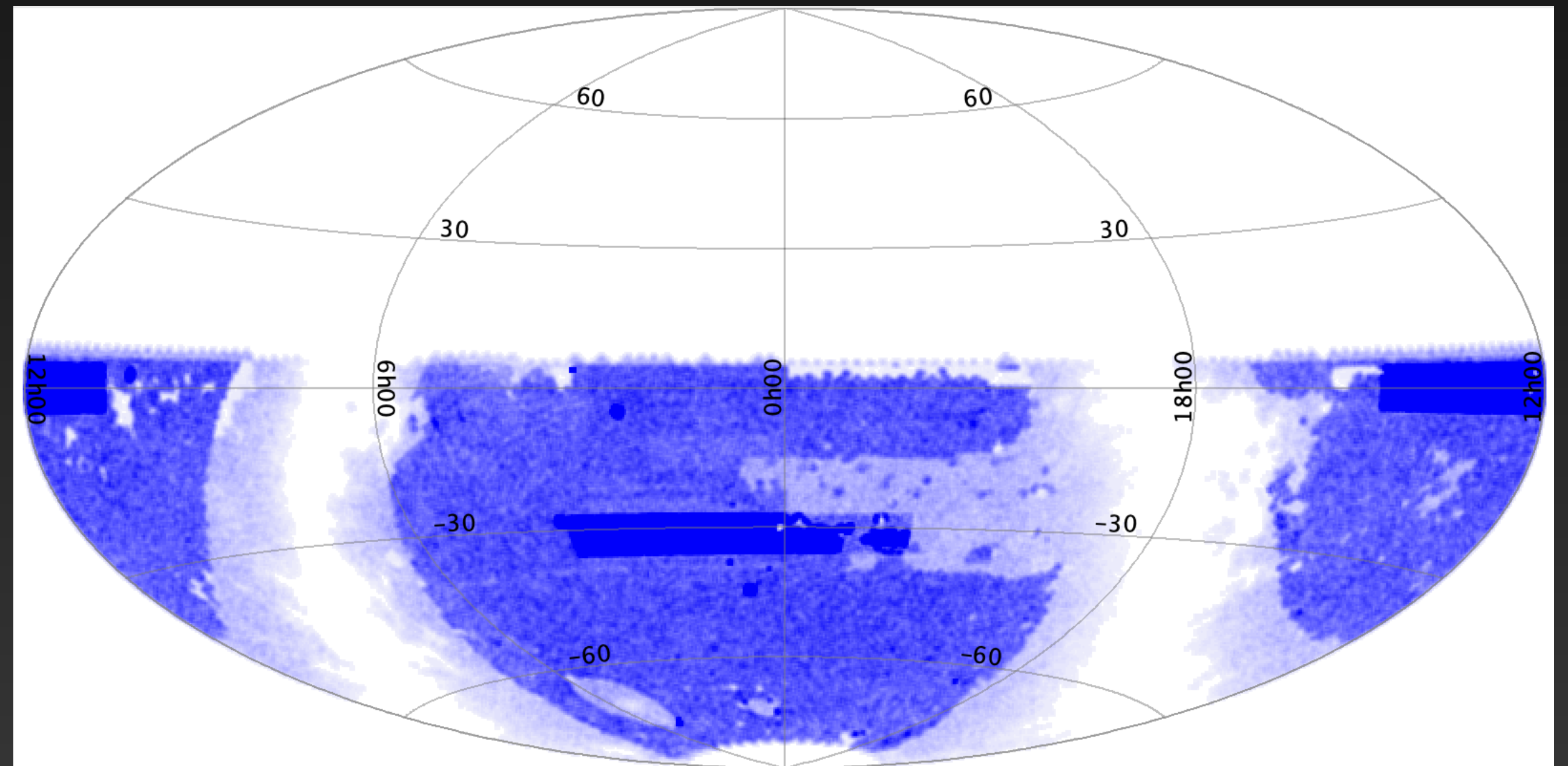
The Baryon Cycle Project

A High-Resolution Extragalactic Community 4MOST Survey

Aim: characterise *how* the CGM gas changes phase and moves into, through and out of galaxies



Credit: Péroux et al.



Hundreds of thousands of QSO spectra

Science made possible by ByCycle

CGM Metal Enrichment

- * **Extent** of the CGM around galaxies?
- * Scale of **metal mixing**?
- * Comparing to **simulations**?

Baryon Budget

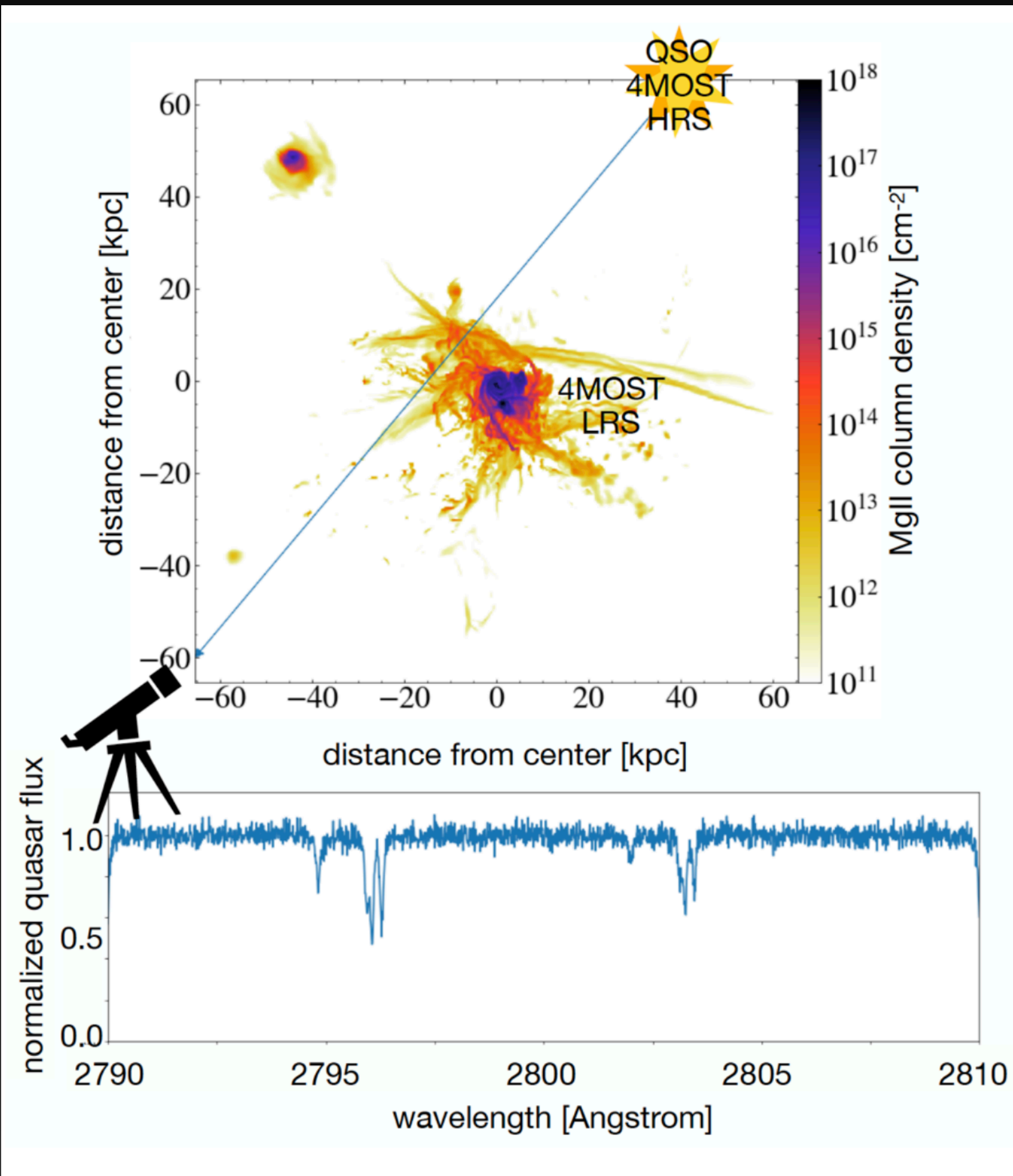
- * Is the gas **photoionised** or **collisionally ionised**?
- * **Dynamical state** of the CGM?
- * Total **baryonic content** of haloes?

Impact of Galaxy Properties on CGM

- * Is the CGM of **AGN** or galaxies in **denser environments** more highly enriched?

The Baryon Cycle Project

A High-Resolution Extragalactic Community 4MOST Survey



MgII absorbers trace CGM gas at $T \sim 10^4 \text{ K}$

Goal:

Find MgII absorbers in
the spectra

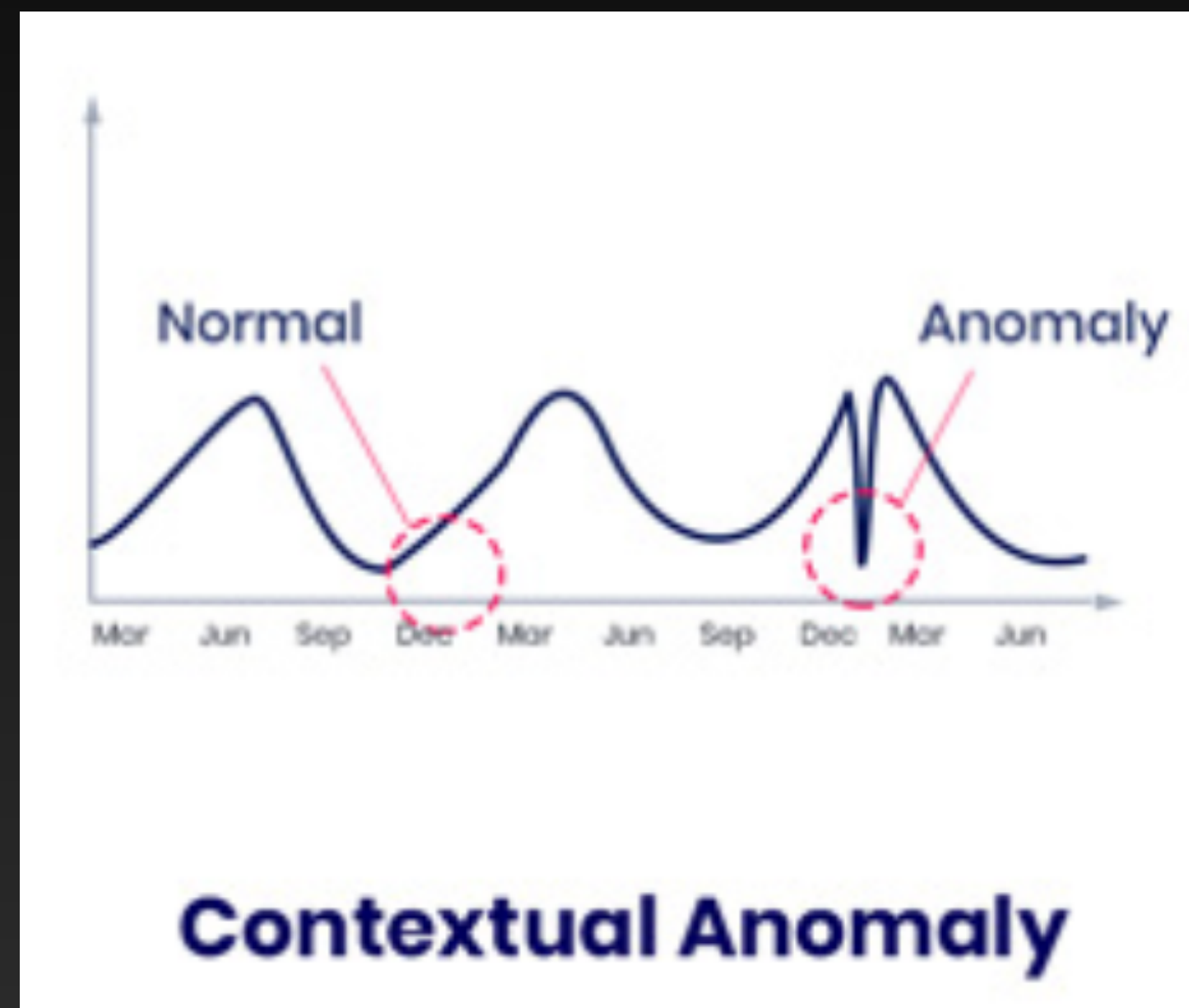
Real spec example with MgIIs

The Baryon Cycle Project

Contextual Anomaly Detection



vs.



What does our model learn?

What a non-absorbed
("normal") spectrum looks

Then what?

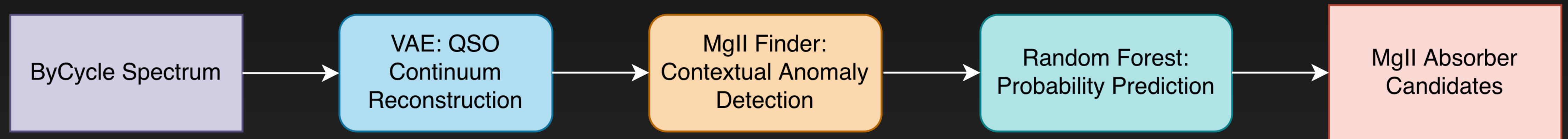
It won't be able to reproduce
the absorbers =>
Reconstruction error will *peak*

Credit: Einer Partners

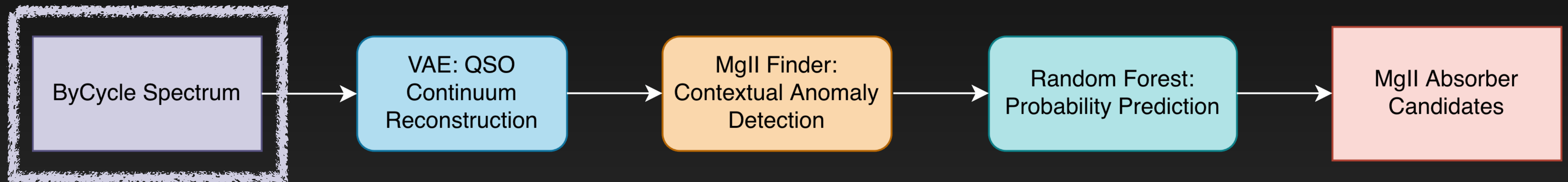
Extra advantages:

- No need for explicit quasar fitting
- No quasar z prior needed

The COOLCAD Pipeline



The COOLCAD Pipeline

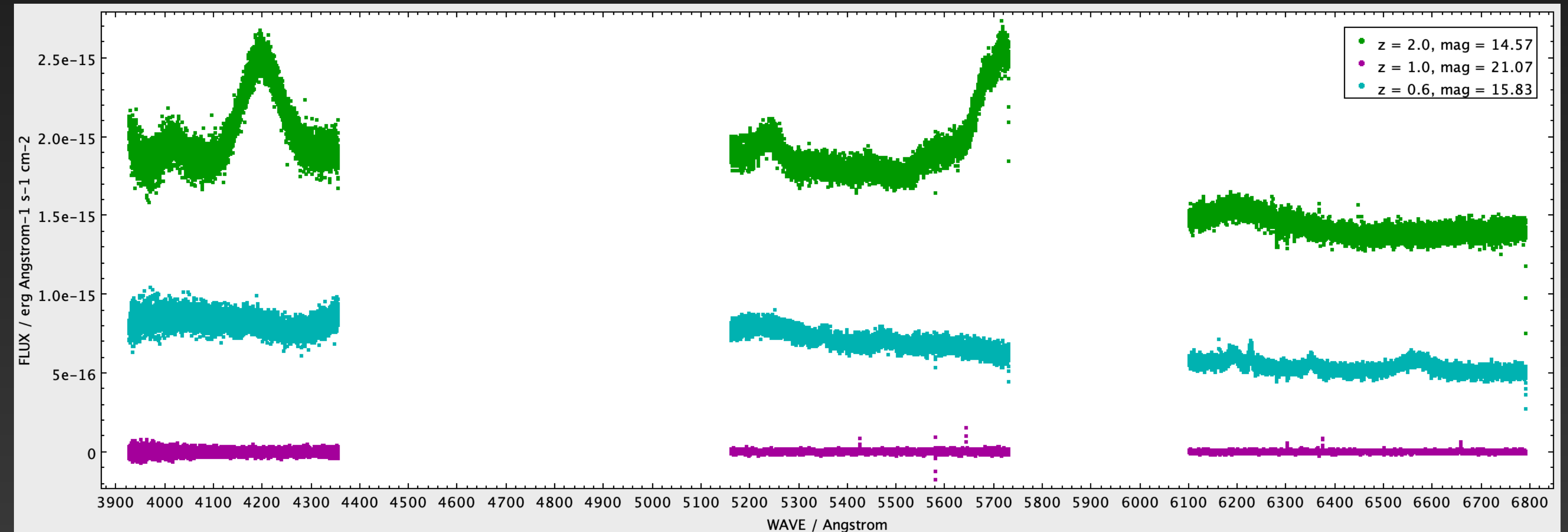
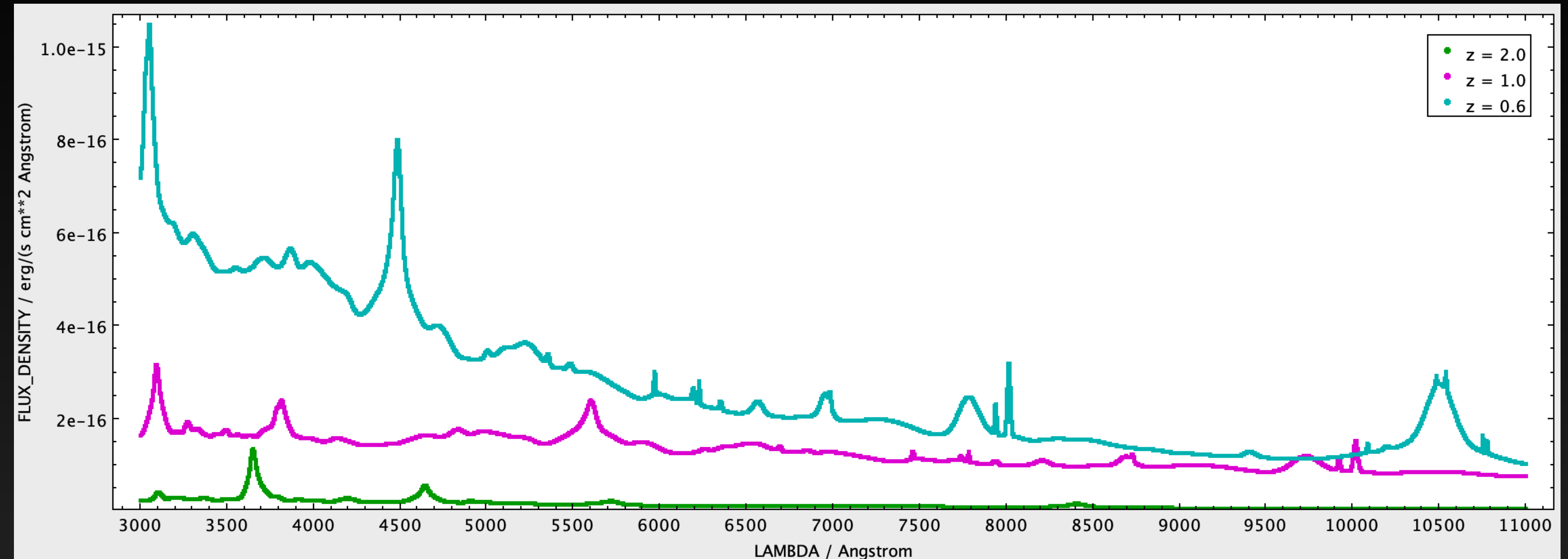


Mock ByCycle Spectra

Quasar Continua
with **no absorbers**
from simqso



Quasar Spectra
with **realistic**
4MOST noise



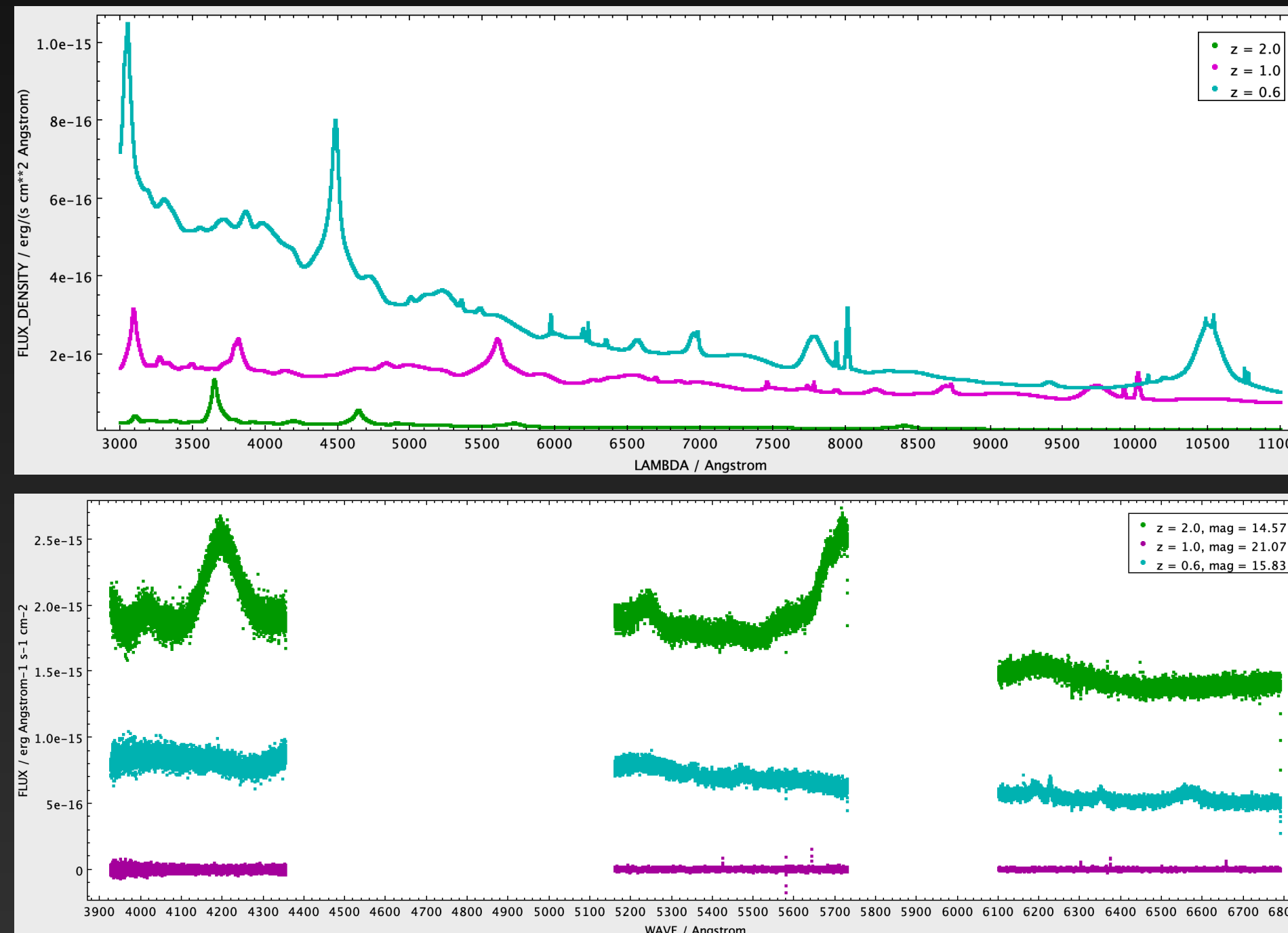
Mock ByCycle Spectra

Quasar Spectra

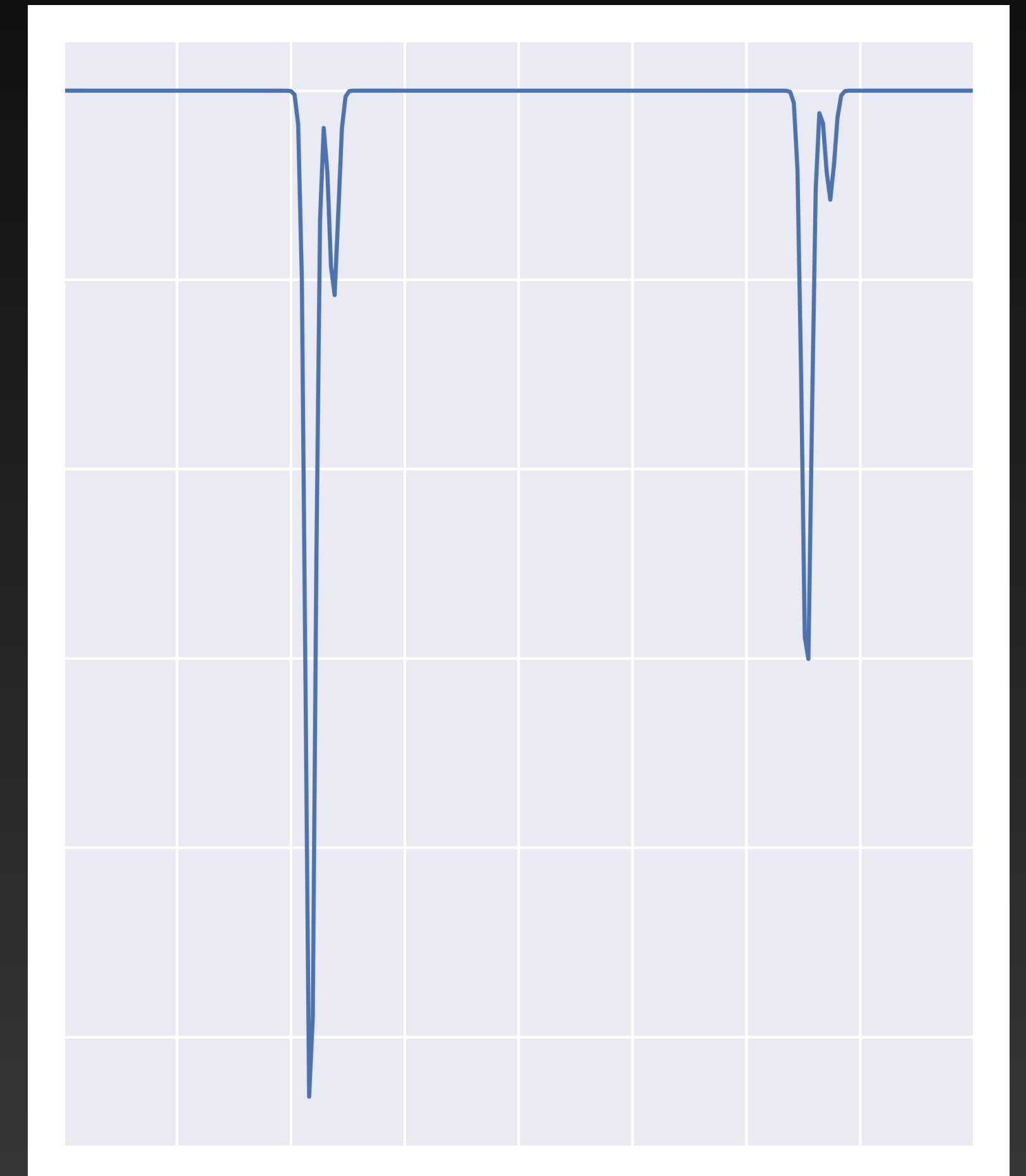
Quasar Continua
with **no absorbers**
from simqso



Quasar Spectra
with **realistic**
4MOST noise

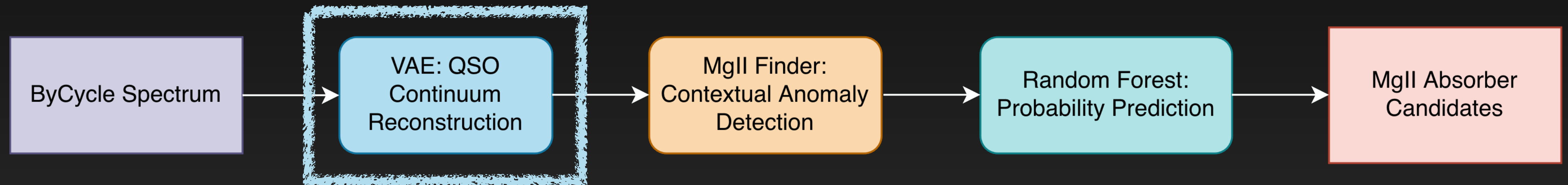


Metal Absorbers from TNG50



SALSA (Nelson et al. 2025)

The COOLCAD Pipeline



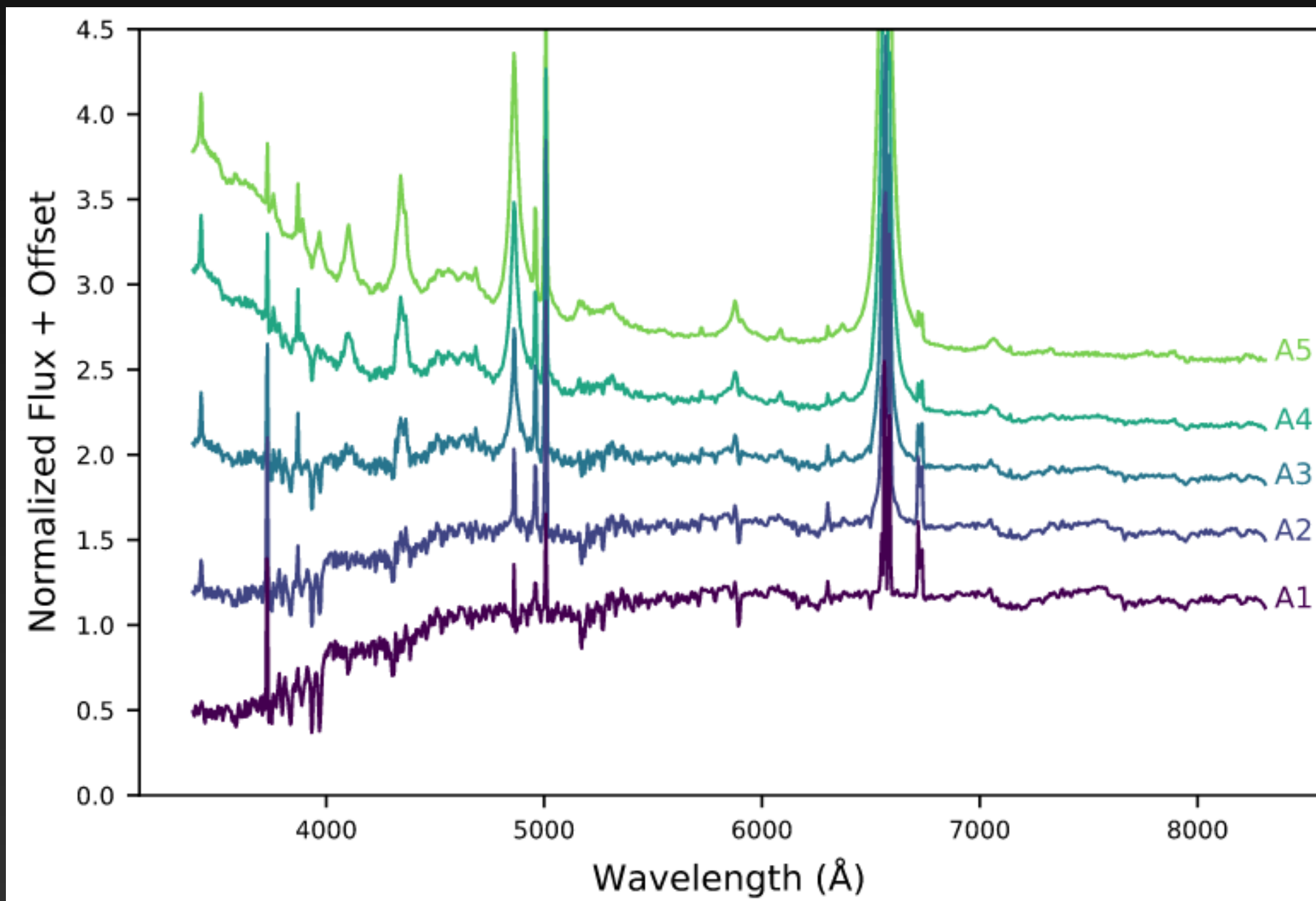
The ByCycle VAEs

Convolutional VAE

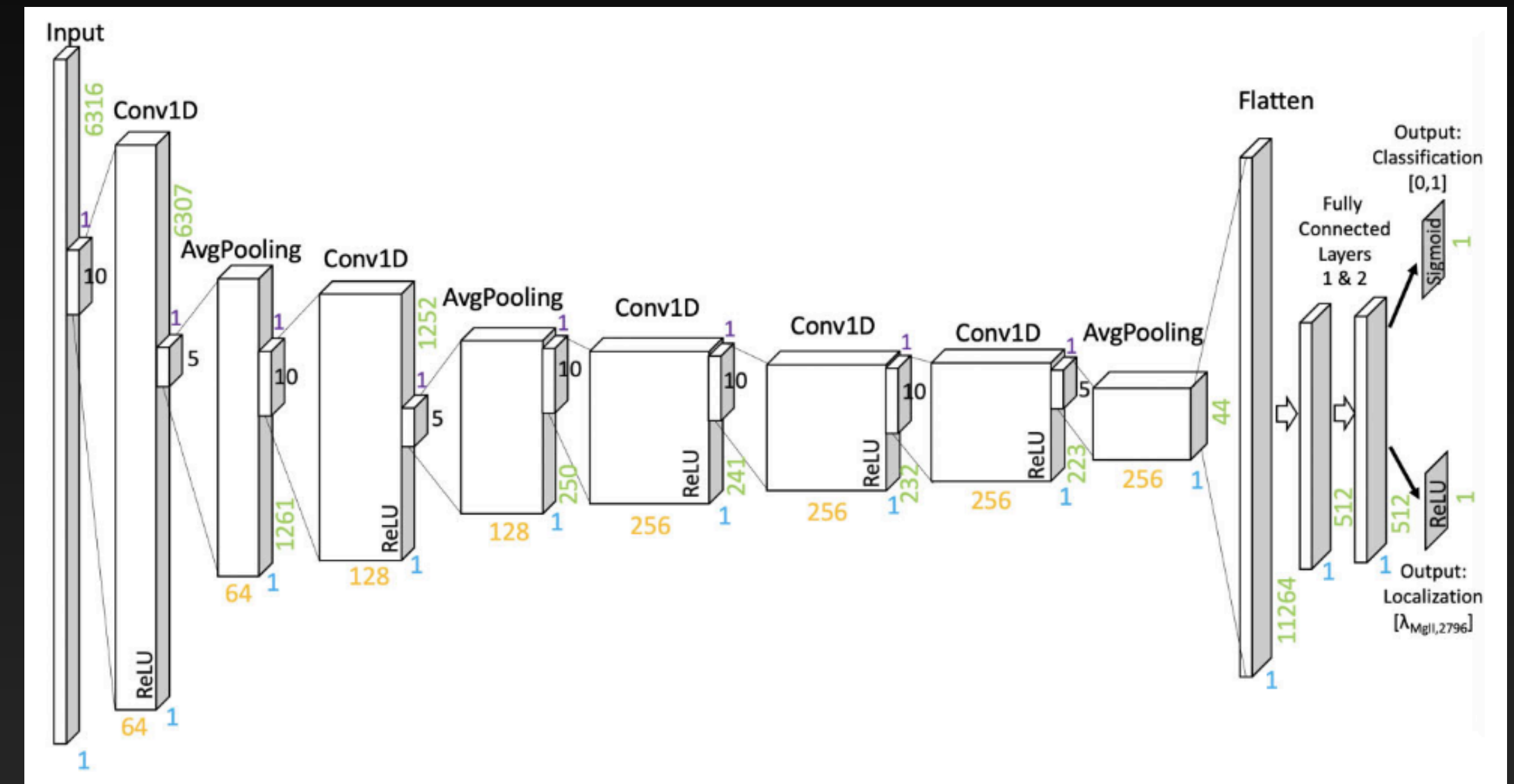
Trained on similar data: z and wavelength estimation

Dense VAE

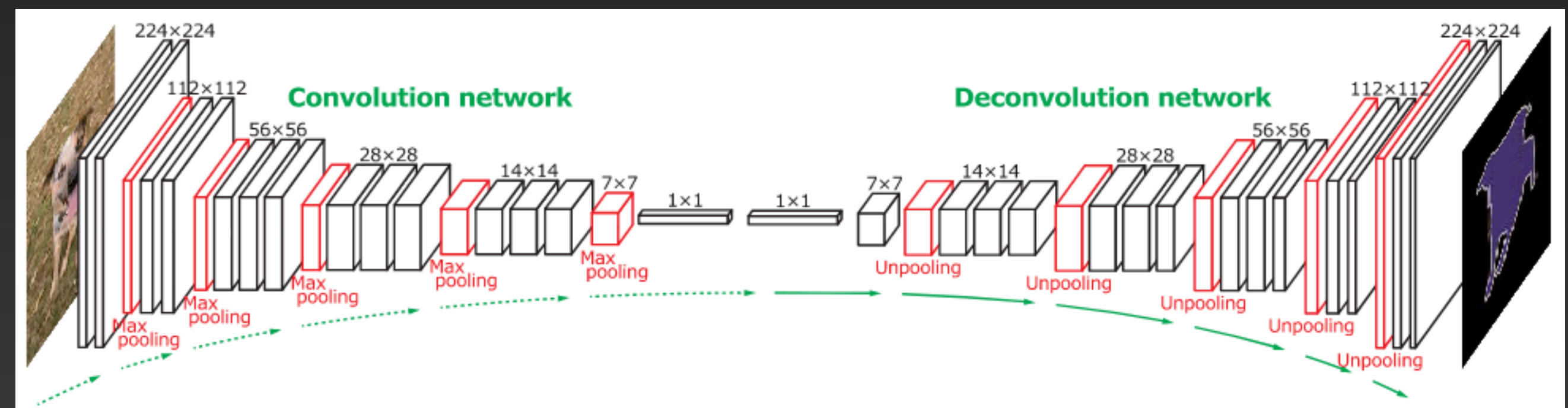
Trained on SDSS spectra



Portillo et al. 2020



Szakacs et al. 2023



Noh et al. 2015

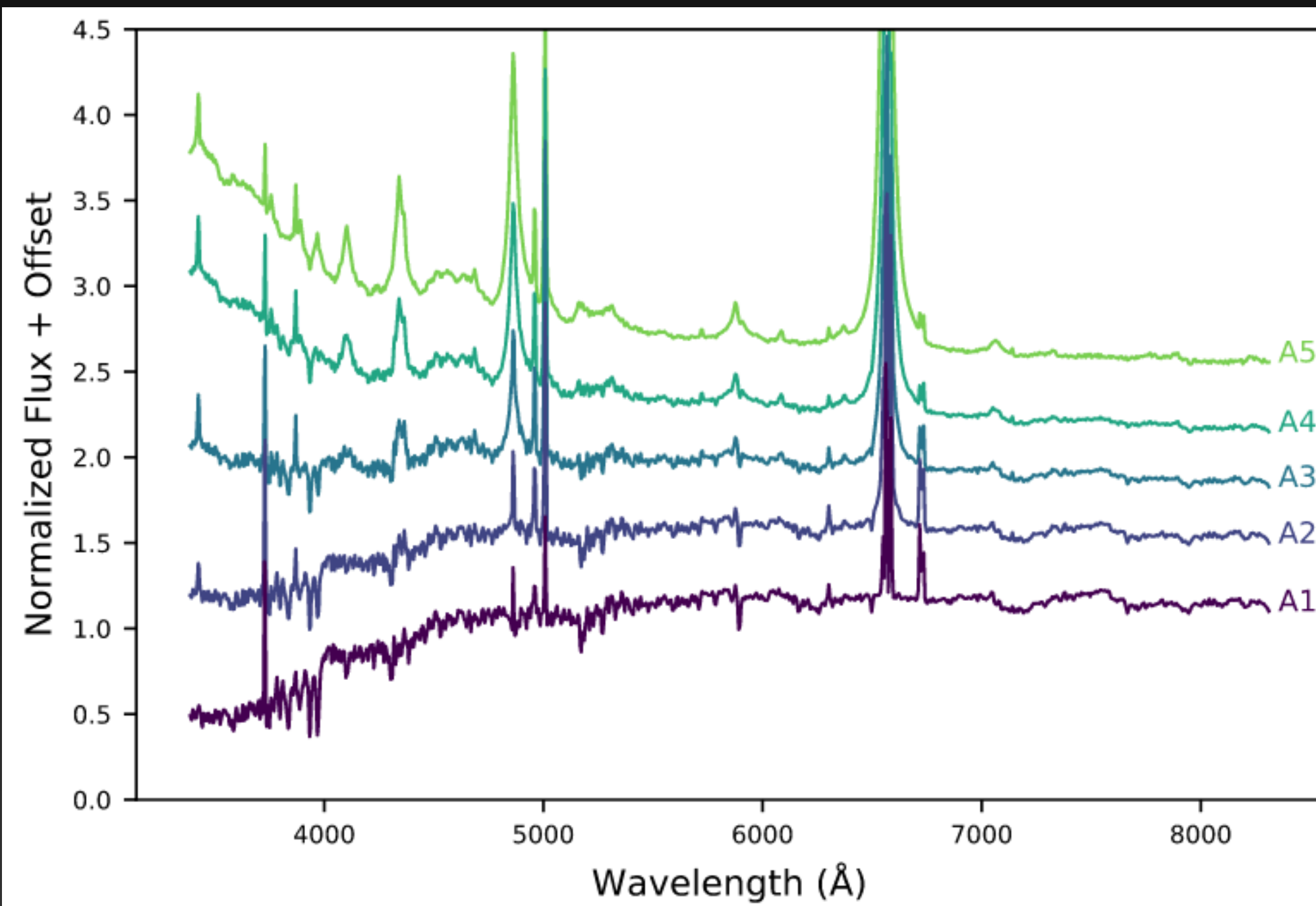
The ByCycle VAEs



Convolutional VAE

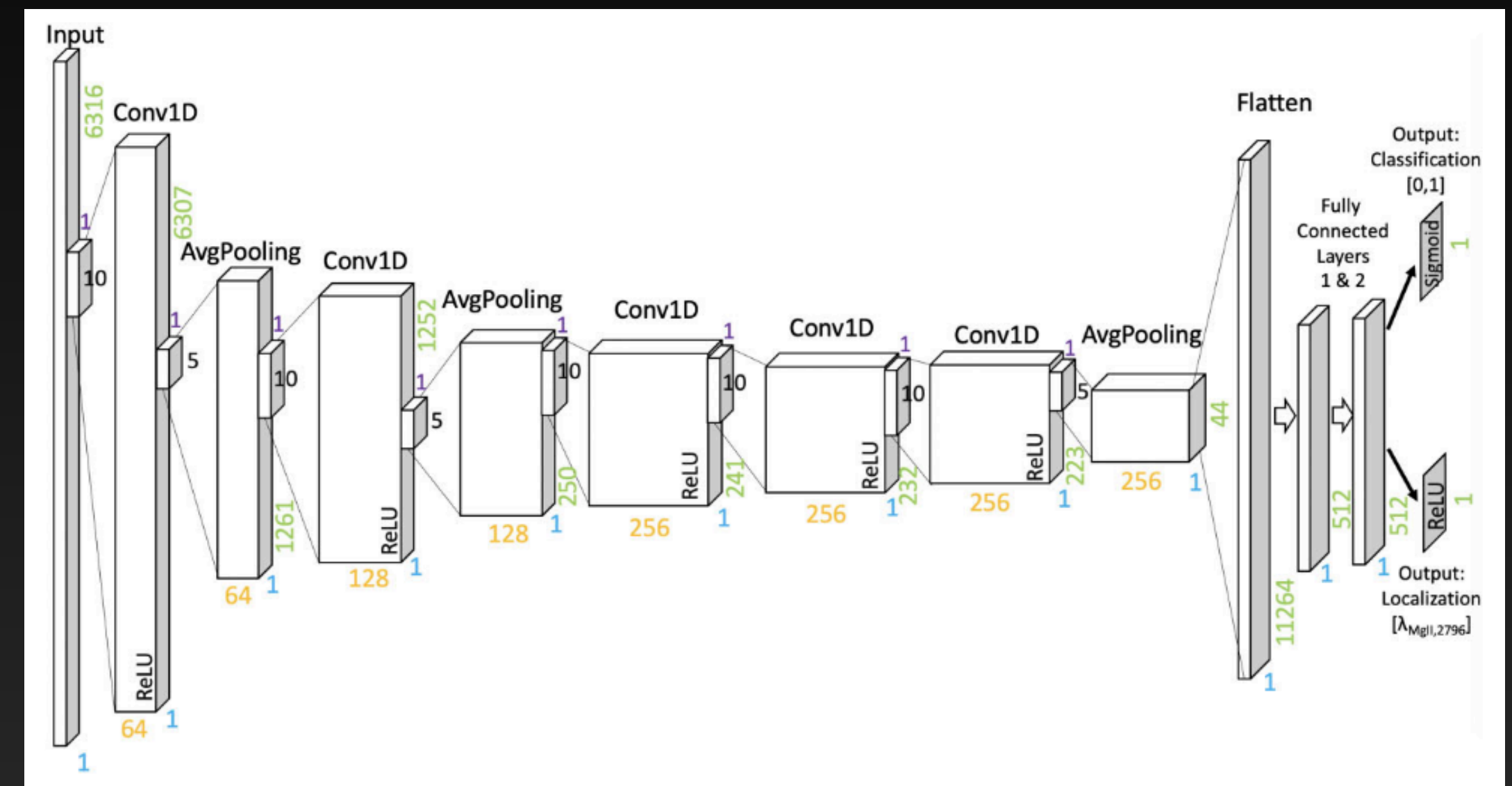
Dense VAE

Trained on SDSS spectra

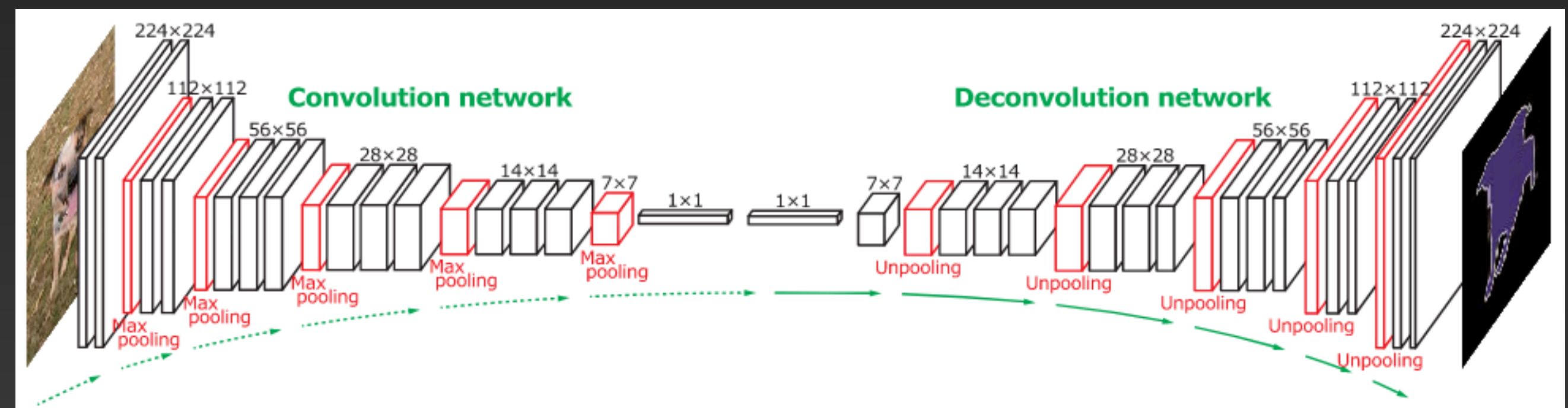


Portillo et al. 2020

Trained on similar data: z and wavelength estimation



Szakacs et al. 2023

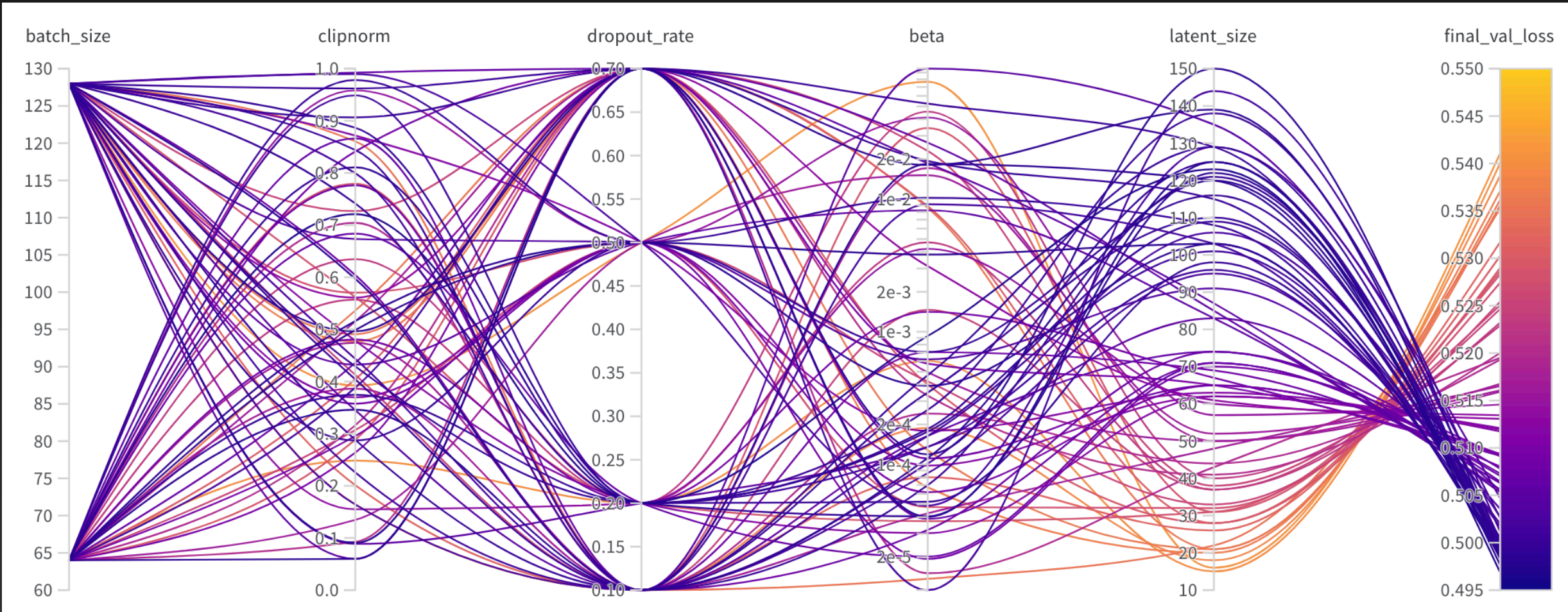


Noh et al. 2015

ByCycle VAEs Architecture

Hyper-parameter
optimisation with
weights & biases:

	Layer	In → Out	Notes
Encoder	Input	l	
	FC	$5971 \rightarrow 1453$	ReLU
	Dropout	$1453 \rightarrow 1453$	$d = 0.5$
	FC	$1453 \rightarrow 194$	ReLU
	Dropout	$194 \rightarrow 194$	$d = 0.5$
Latent	FC μ	$194 \rightarrow 15$	
	FC $\log \sigma^2$	$194 \rightarrow 15$	
	Latent Space	$15 \rightarrow 15$	$z = \mu + \sigma \odot \varepsilon$
Decoder	FC	$15 \rightarrow 194$	ReLU
	Dropout	$194 \rightarrow 194$	$d = 0.5$
	FC	$194 \rightarrow 1453$	ReLU
	Dropout	$1453 \rightarrow 1453$	$d = 0.5$
	FC (output)	$1453 \rightarrow l$	Linear

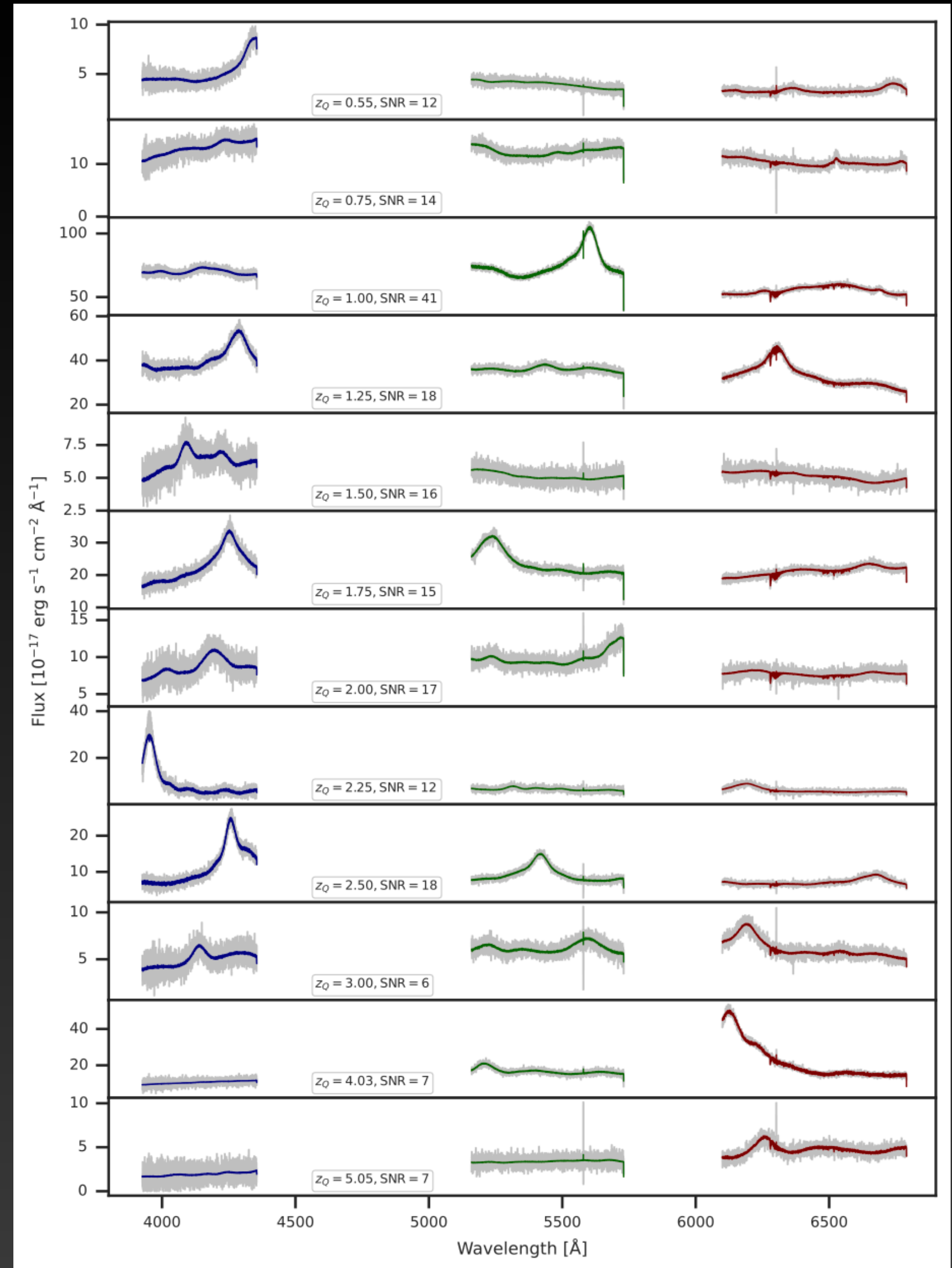


The ByCycle VAEs

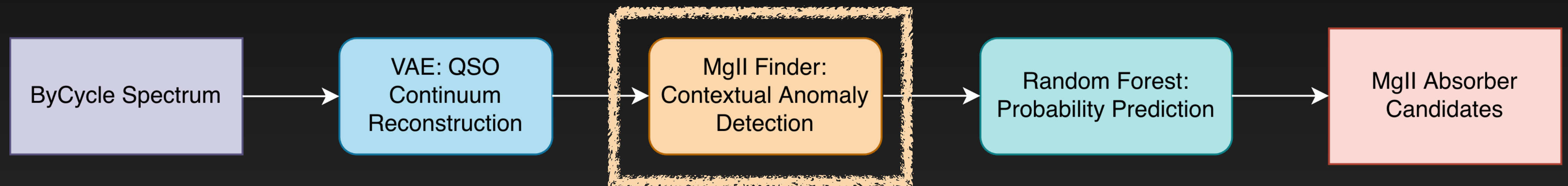
One fully-connected VAE per spectral arm

Reconstructions

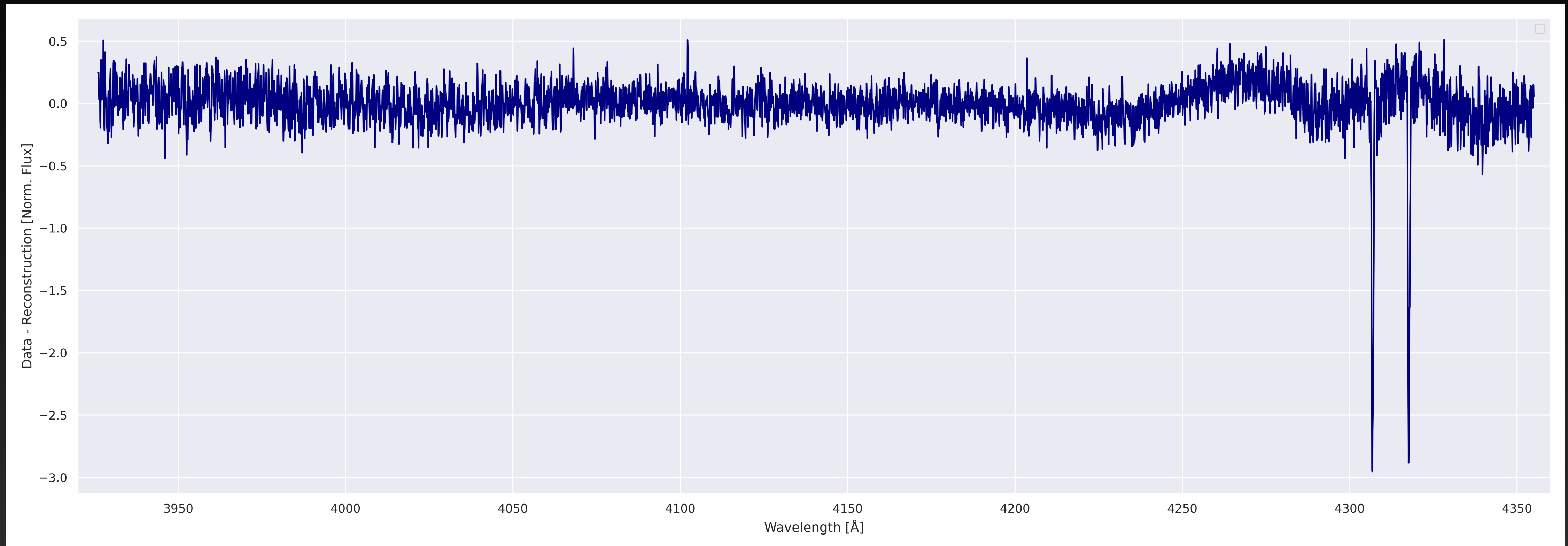
- are $\sim 2\times$ less noisy
- closely follow the true continua



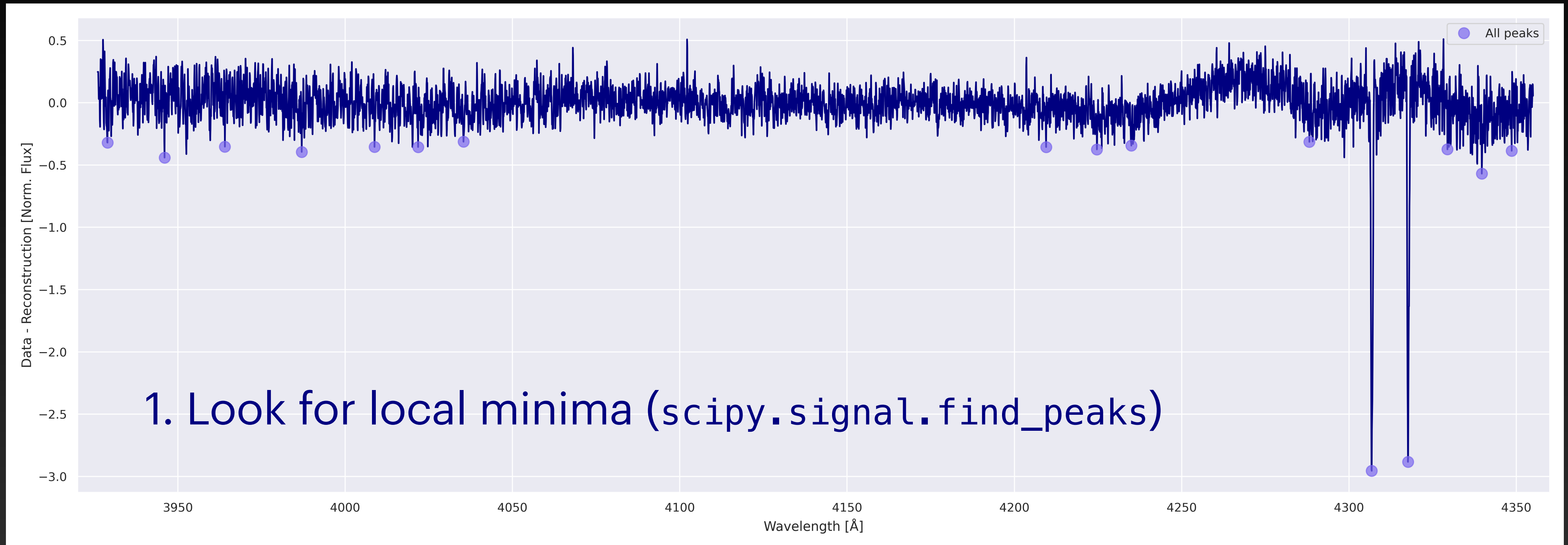
The COOLCAD Pipeline



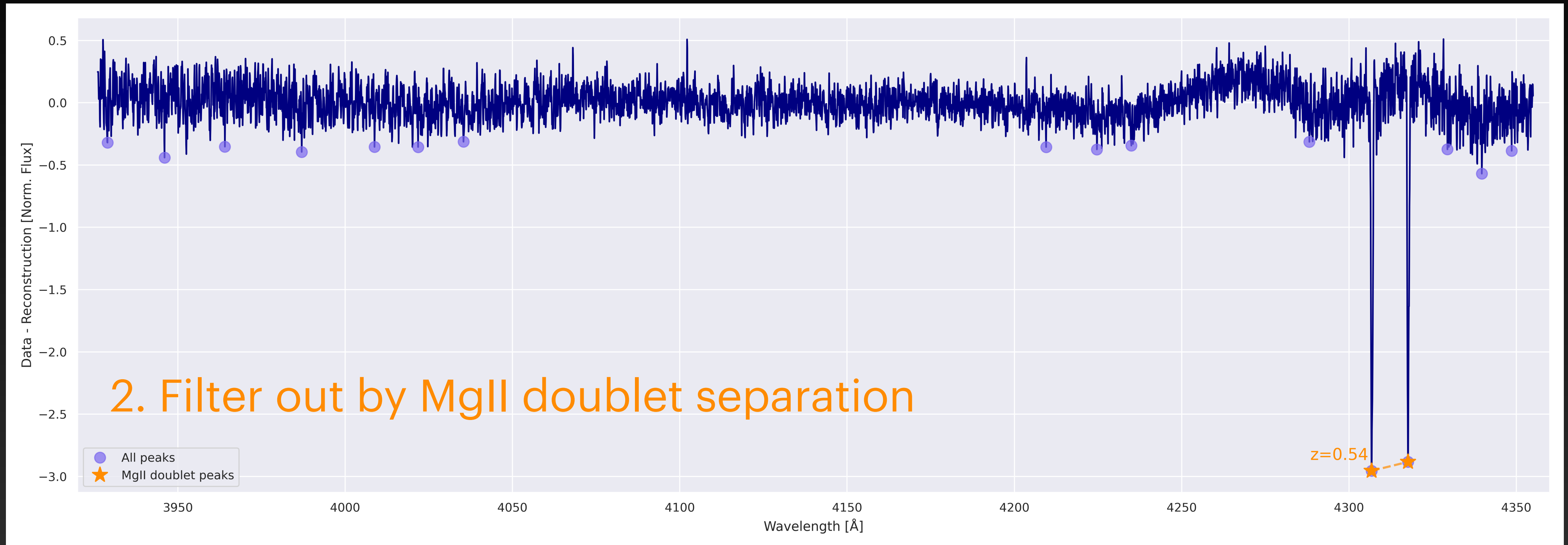
Locating MgII Absorbers



Locating MgII Absorbers

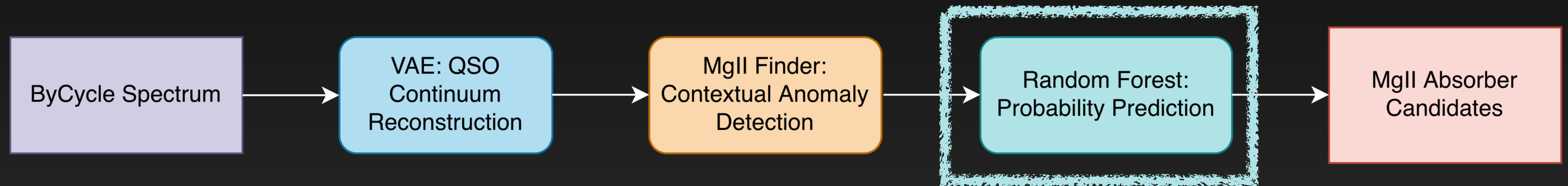


Locating MgII Absorbers

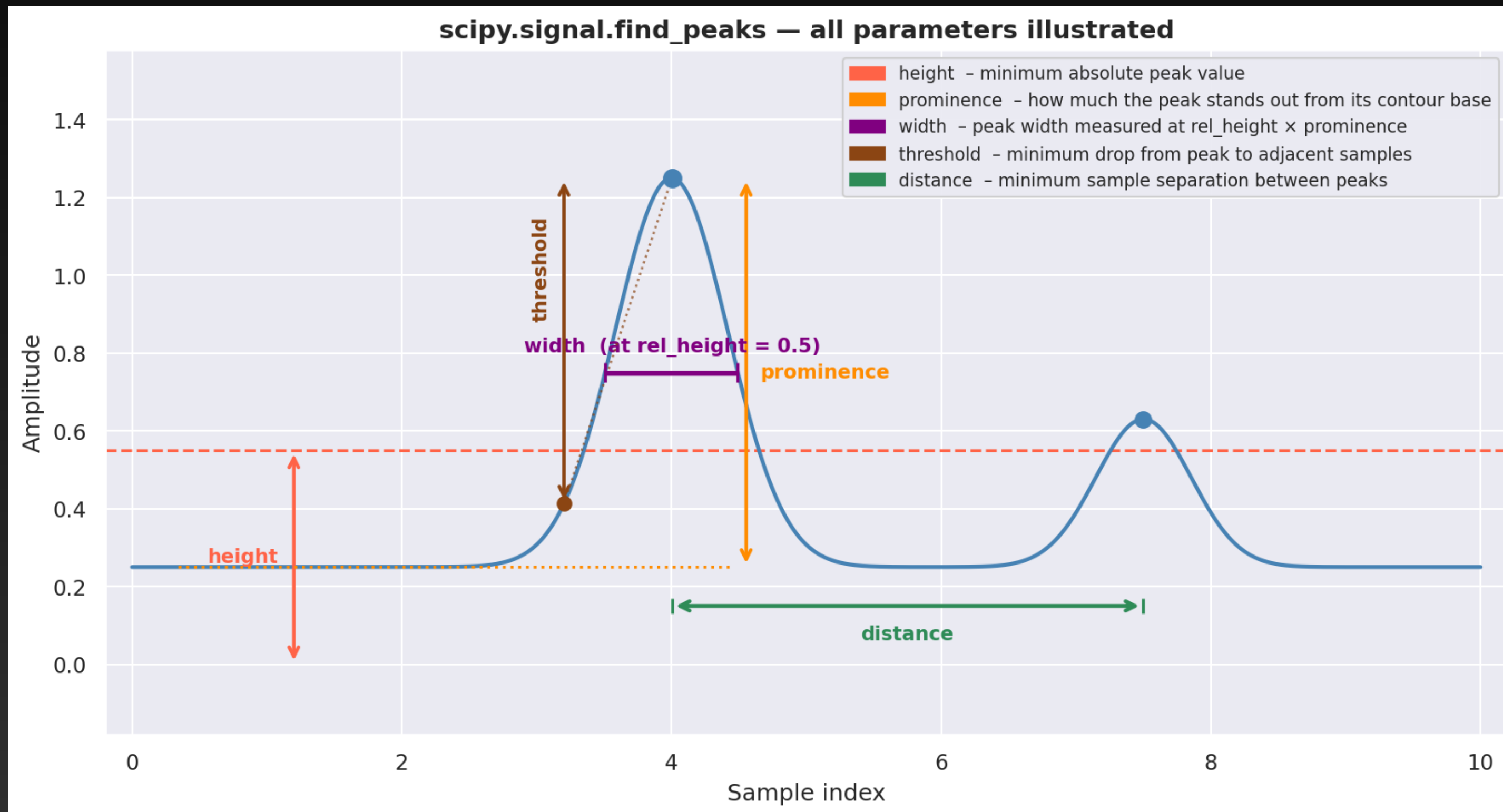


We are able to locate MgII absorbers!

The COOLCAD Pipeline



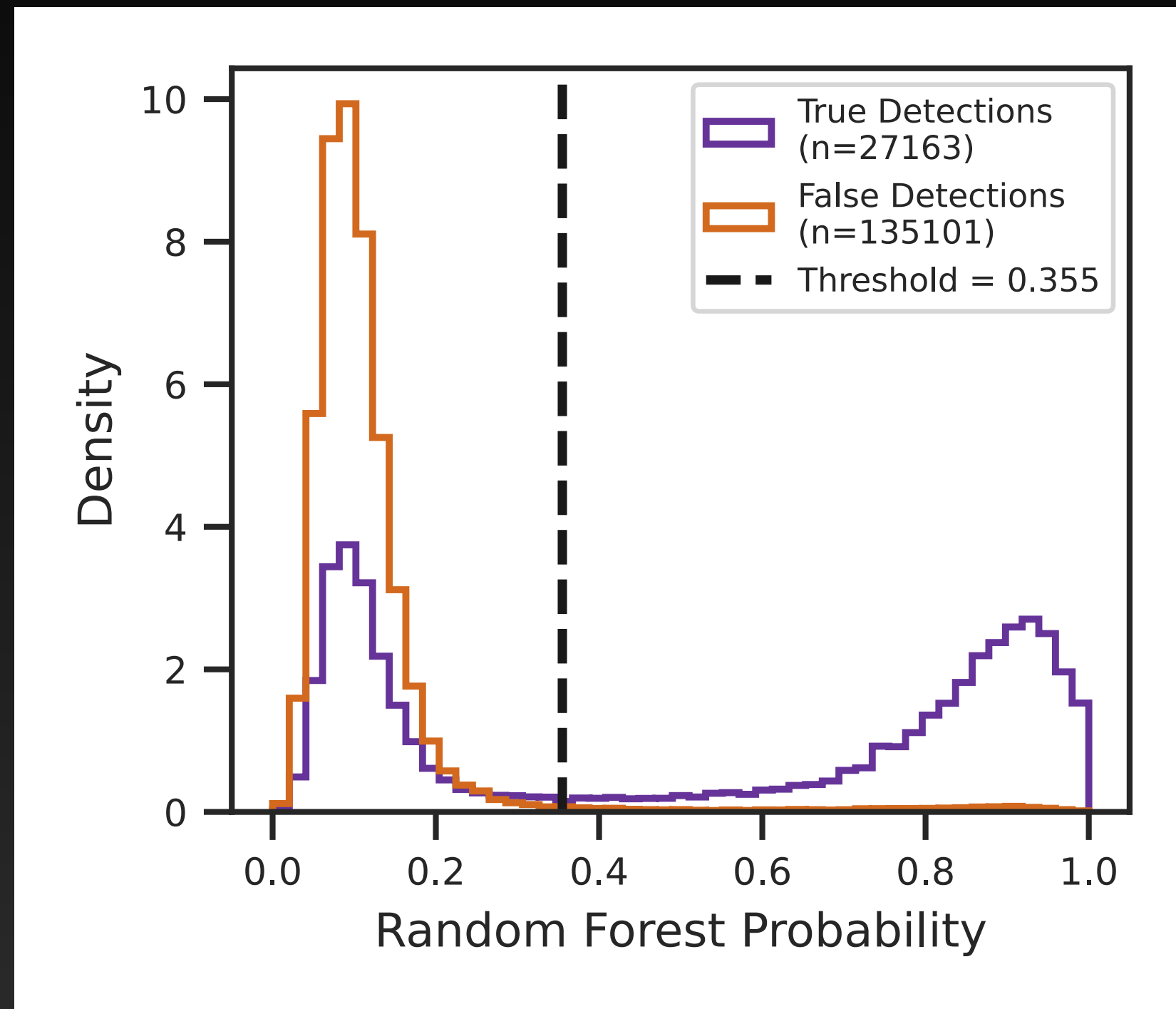
Random Forest: Real vs. False Detections



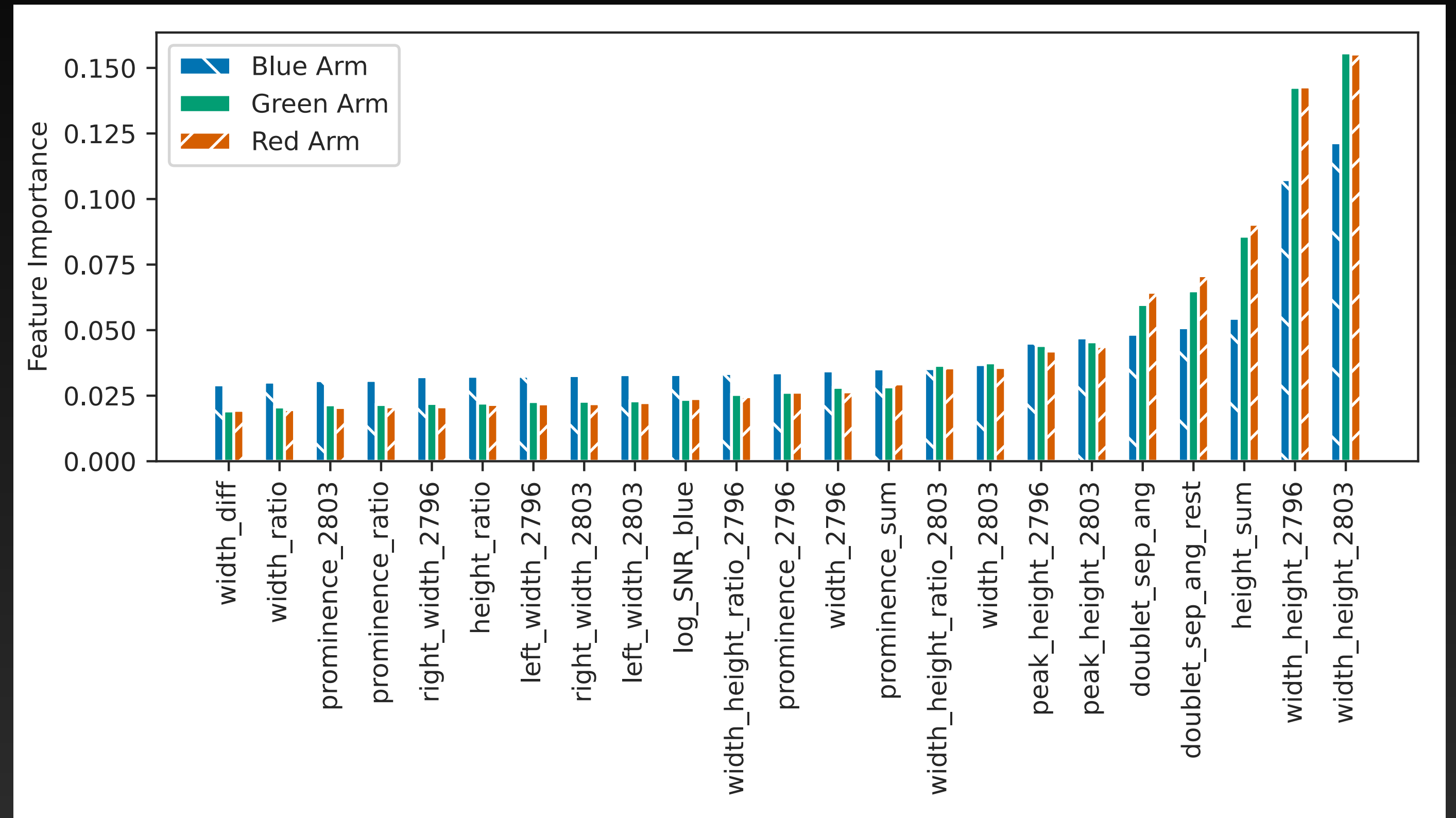
We train and optimise RFs on the measurements of the peaks (from `scipy.signal.find_peaks`)

Random Forest: Real vs. False Detections

Blue Arm



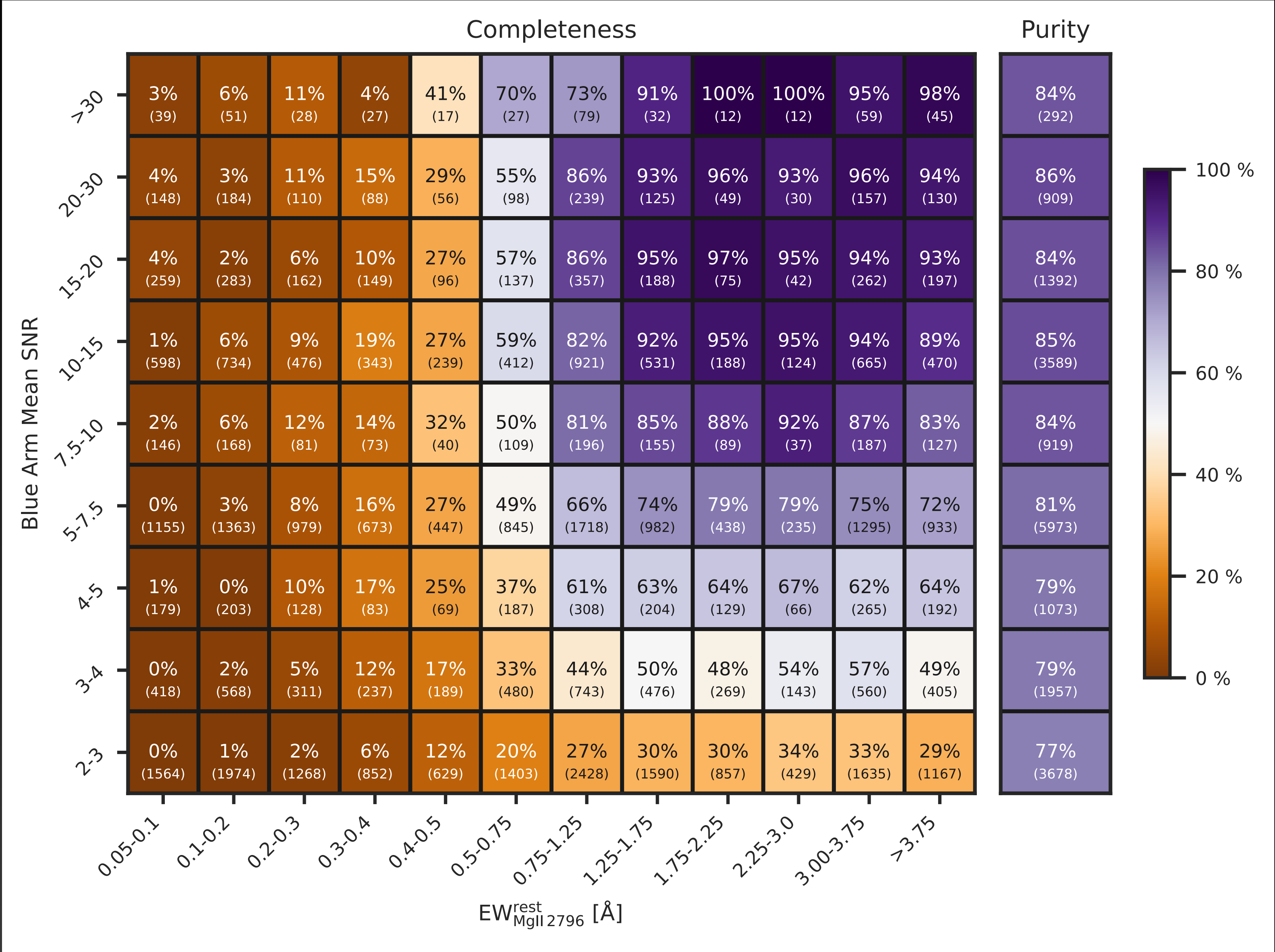
Our RFs can effectively separate true absorbers from false detections



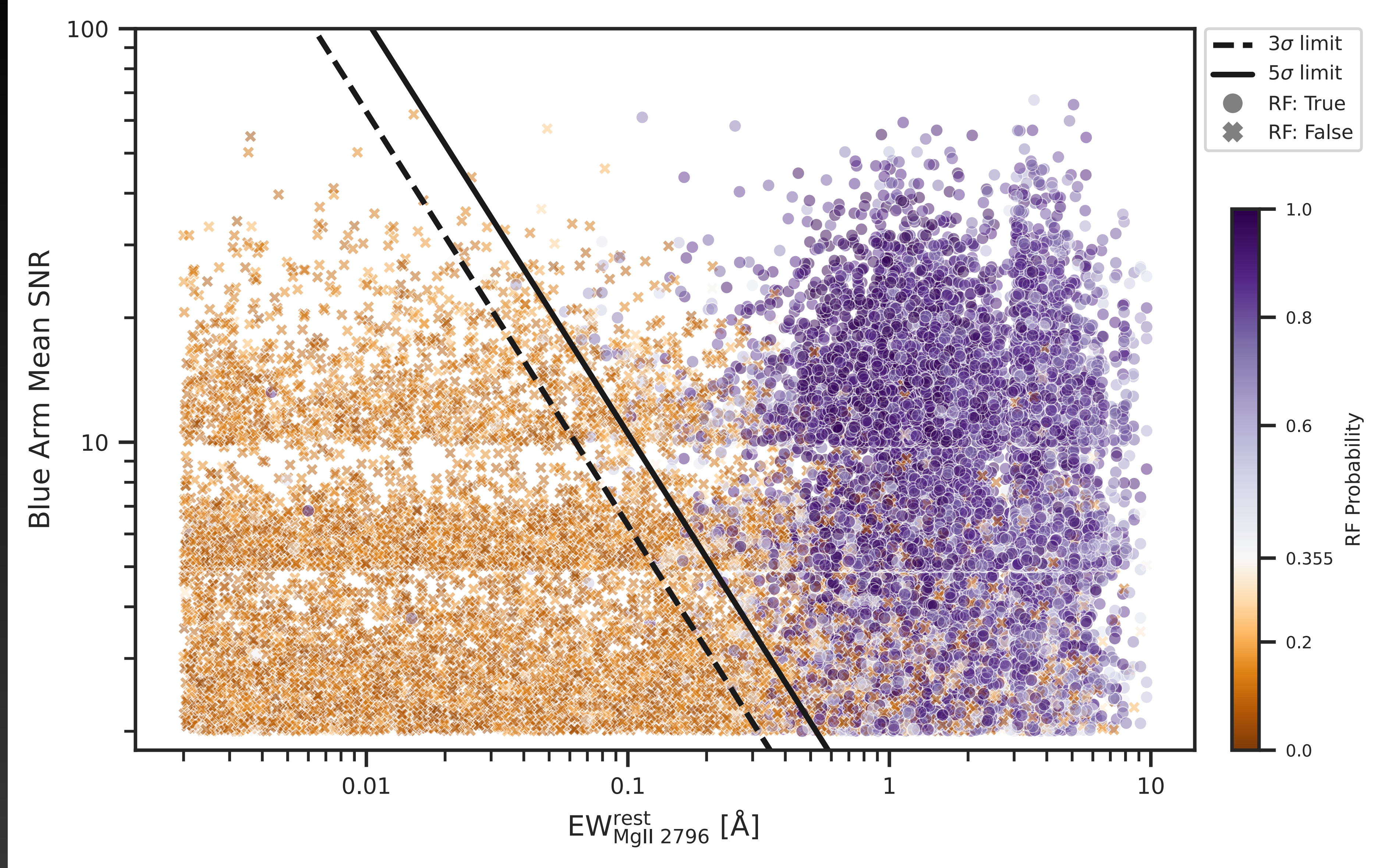
Feature Importance:

Shape (width vs. height) of false detections **do not follow a Voigt profile**

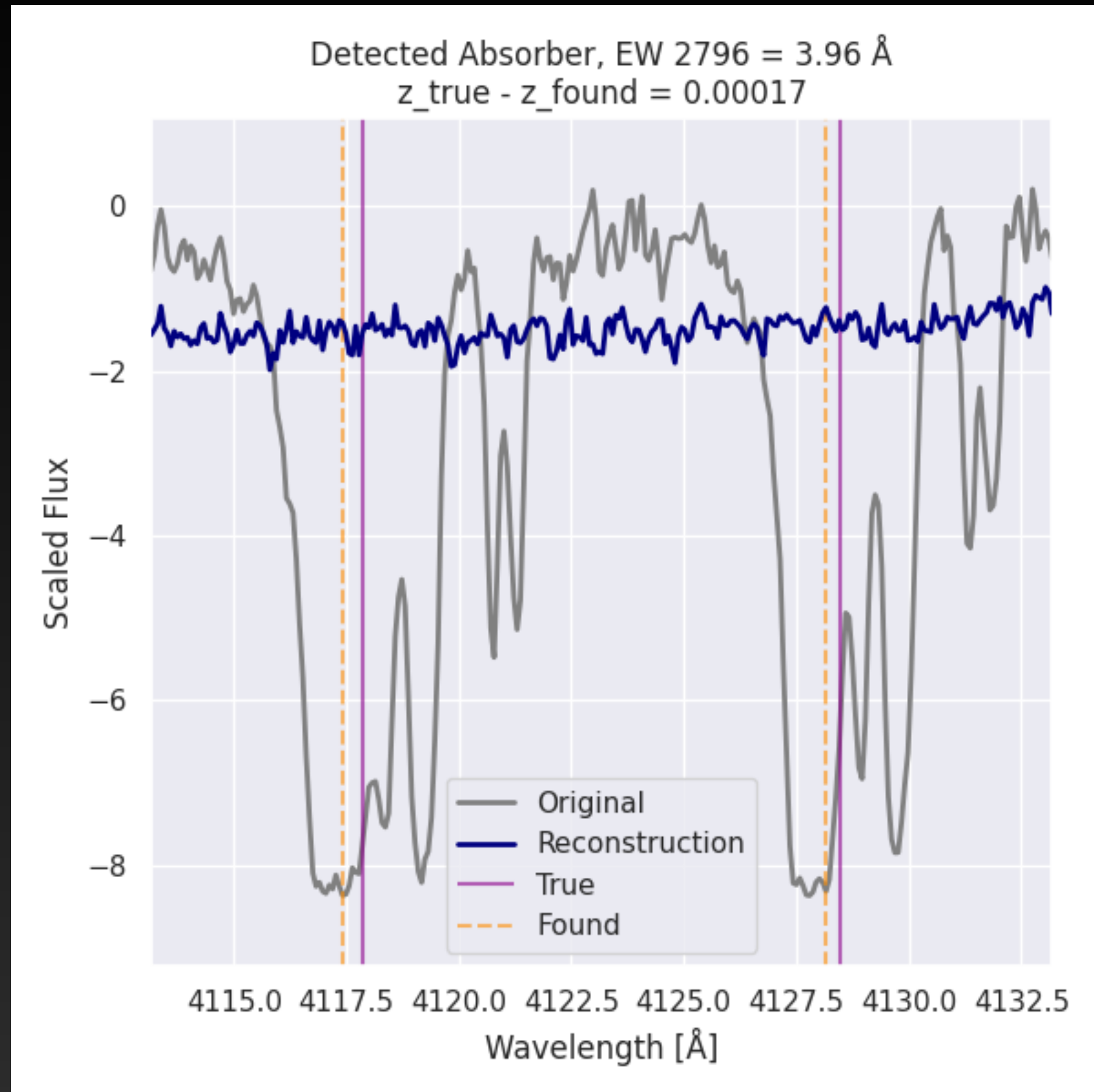
Completeness vs. Purity: **blue** arm



Minimum EW vs SNR: **blue** arm



SNR > 30



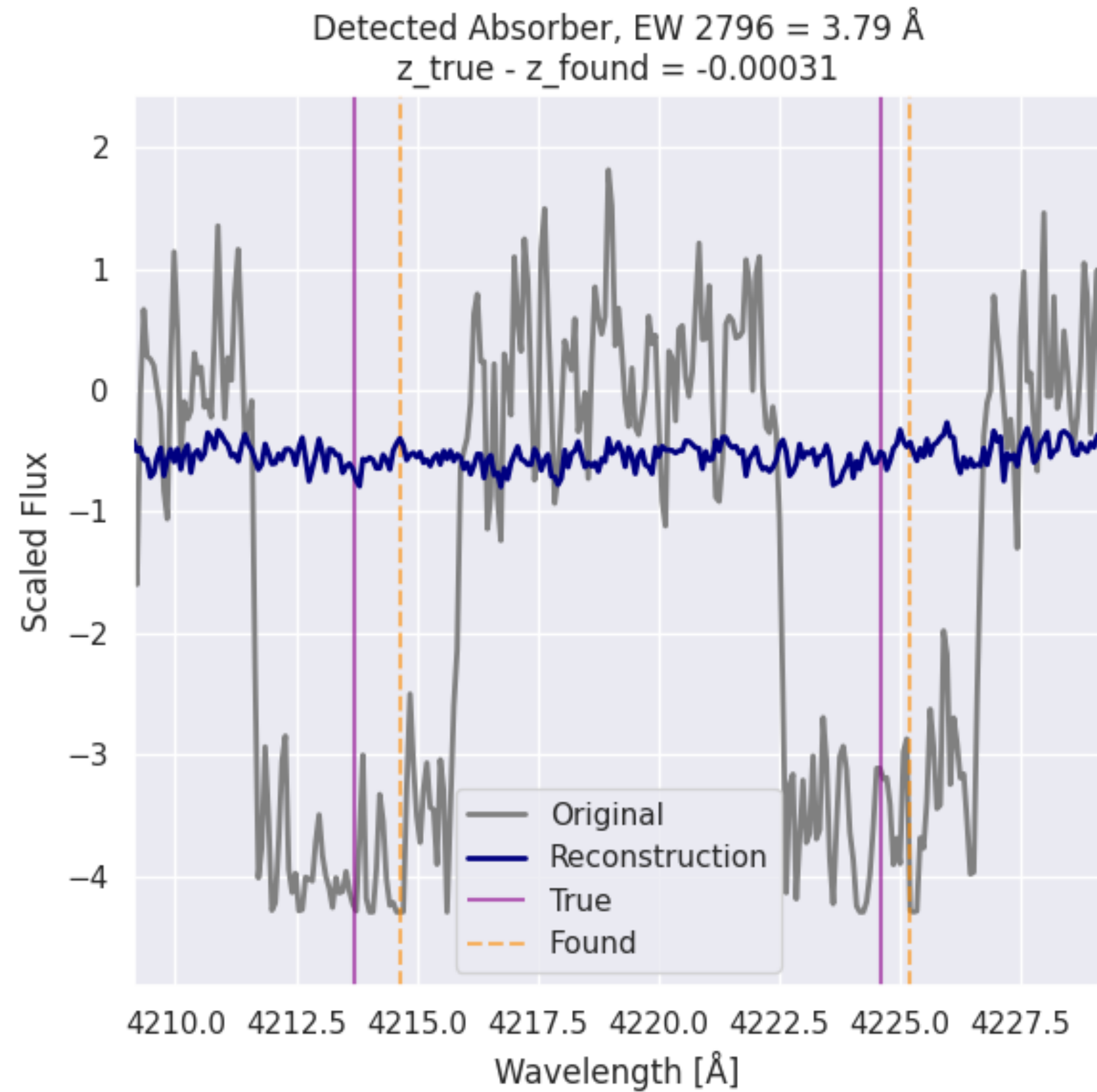
SNR > 30



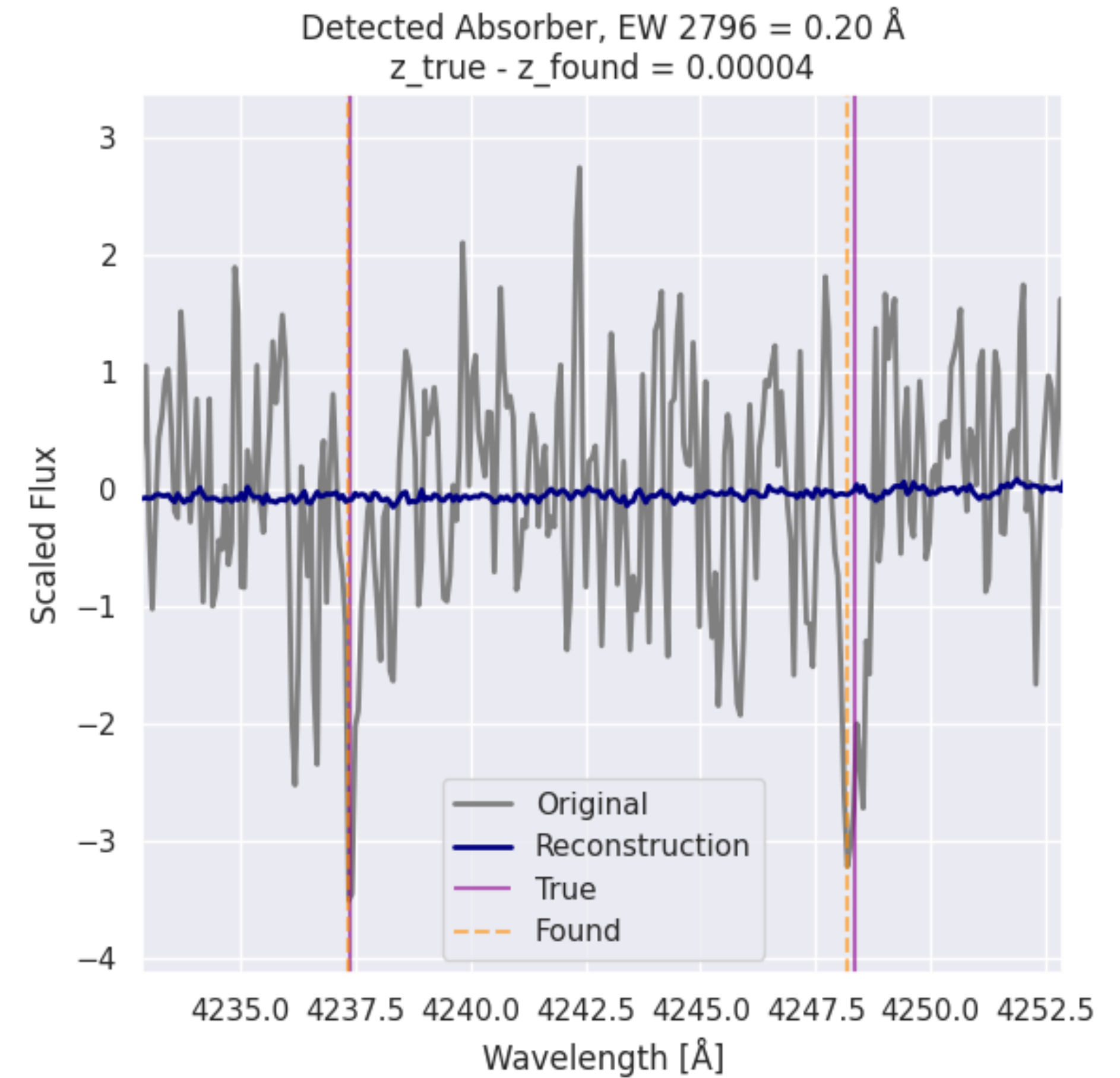
Good detections:

strong & weak absorbers in good SNR

4.0 < SNR < 5.0



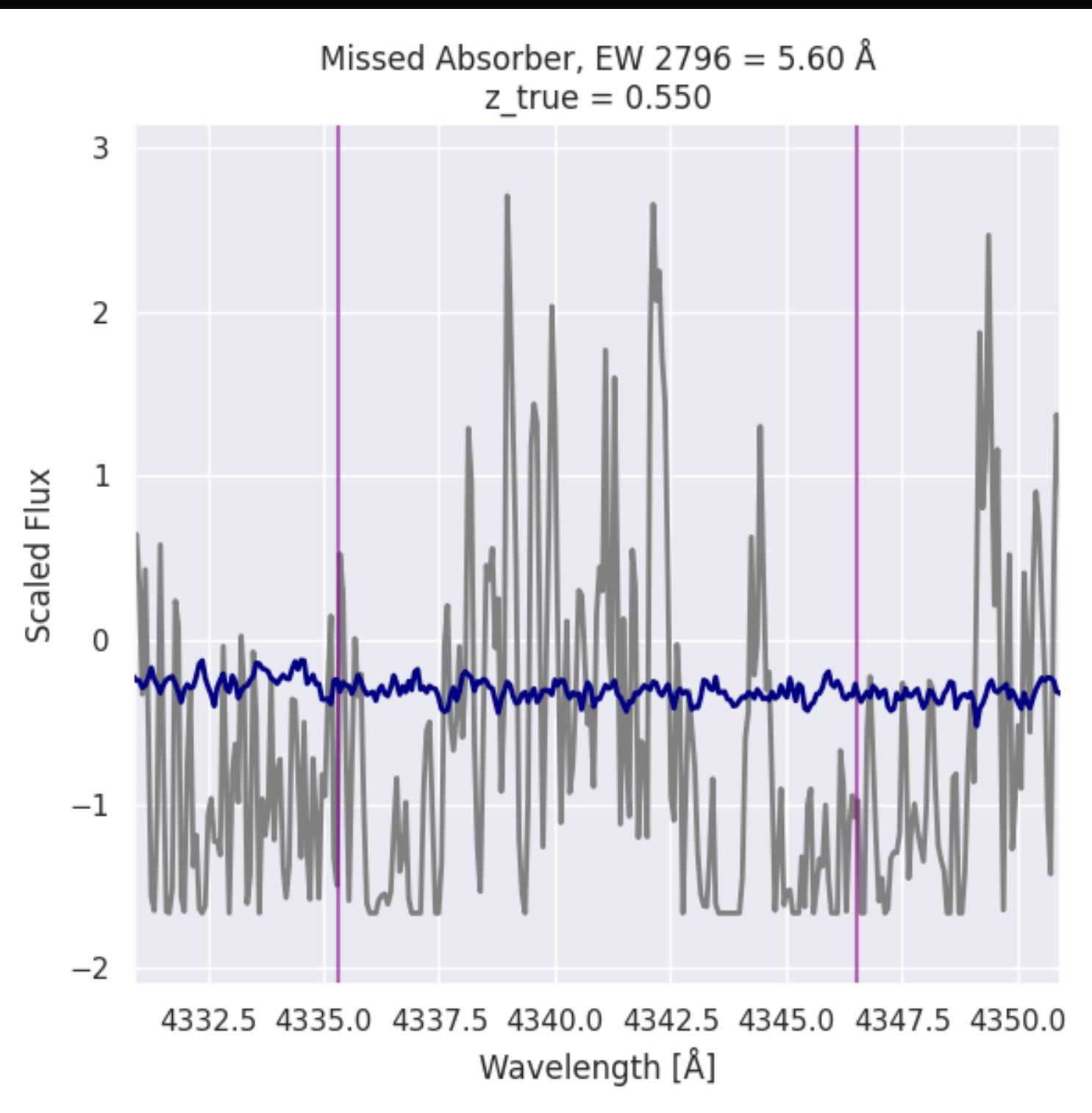
2.0 < SNR < 3.0



Good detections:

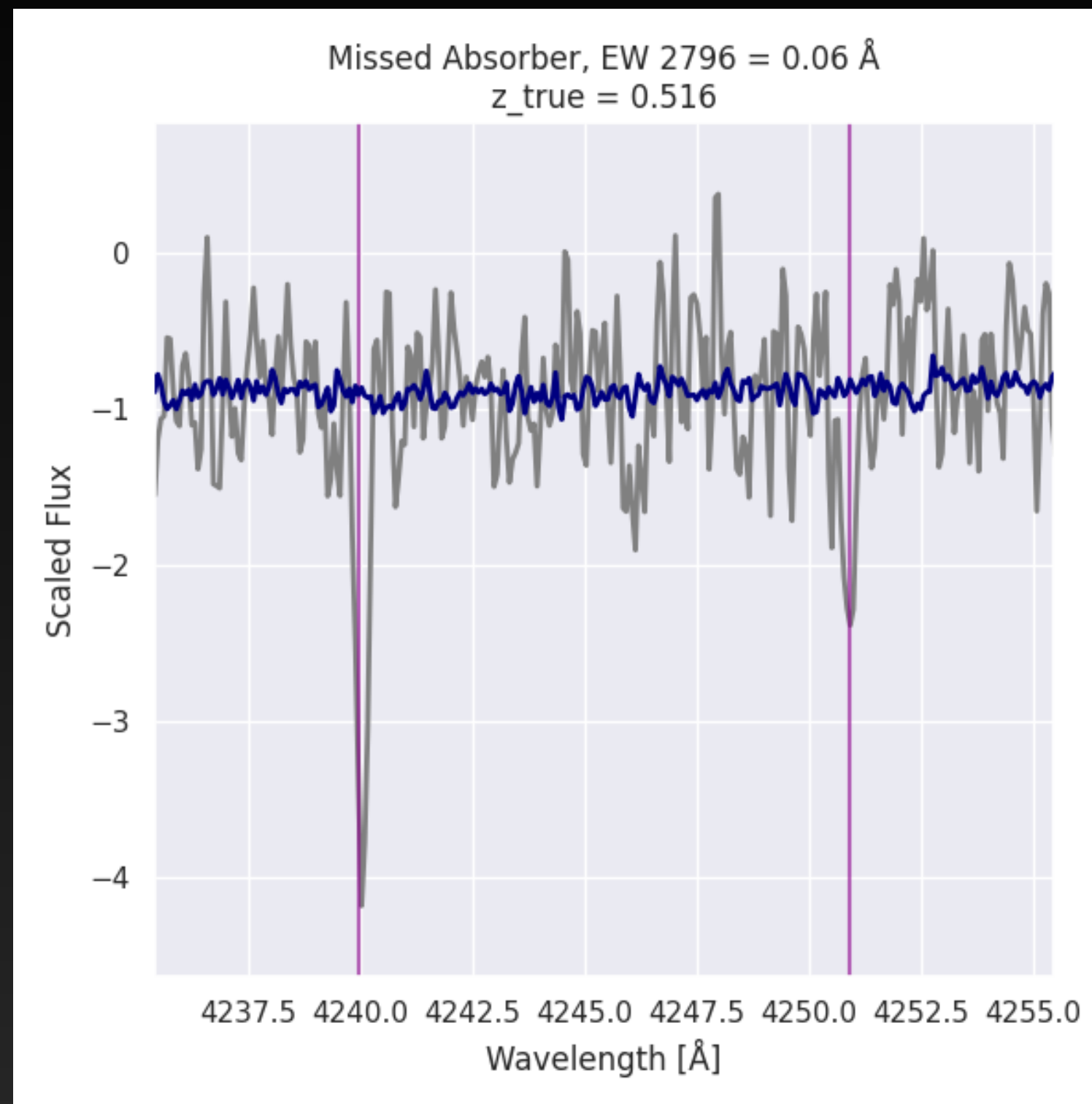
strong & weak absorbers in low SNR

$4.0 < \text{SNR} < 5.0$



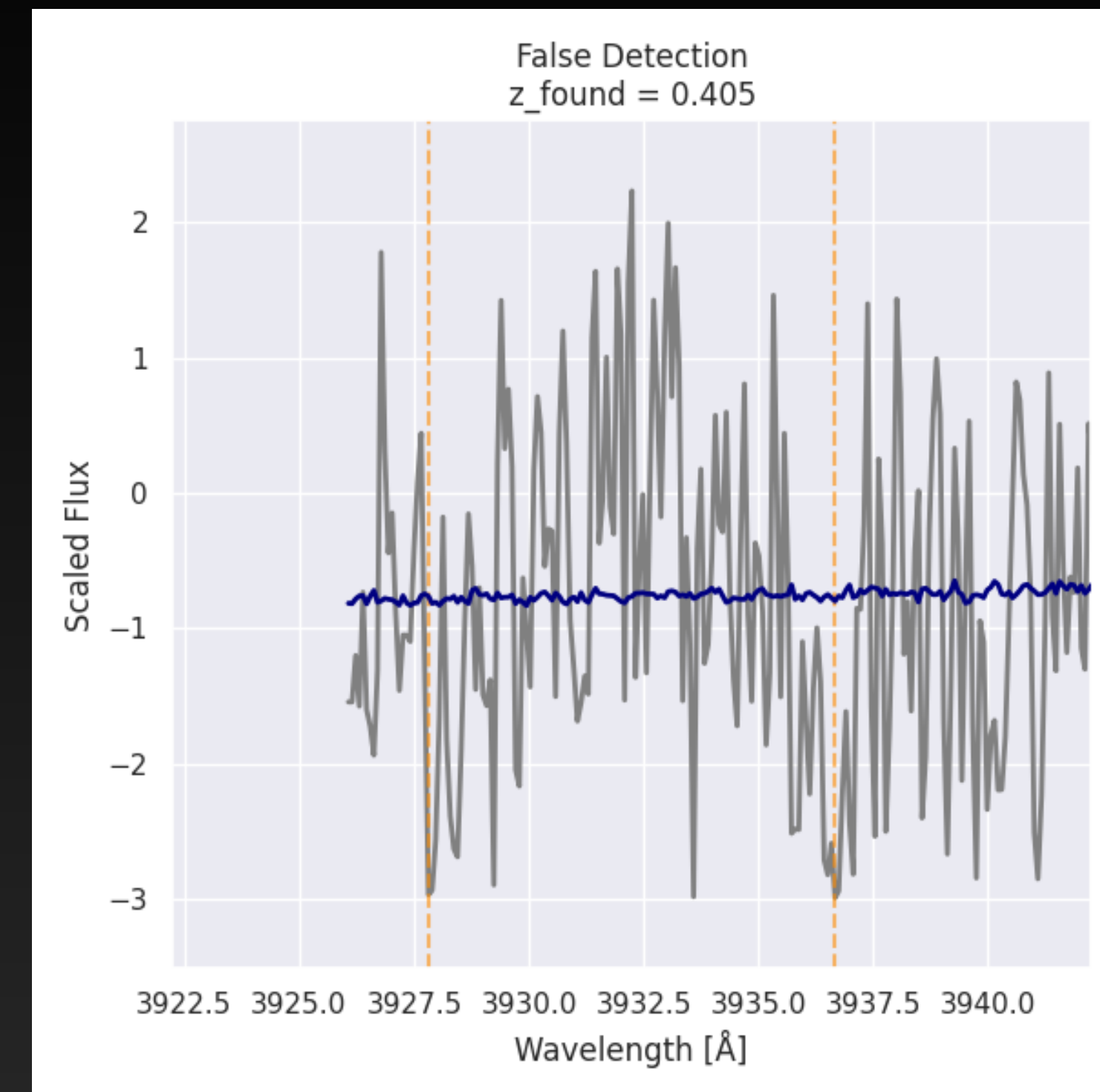
Missed **strong** absorber
in **low** SNR

$\text{SNR} > 30.0$



Missed **weak** absorber in
high SNR

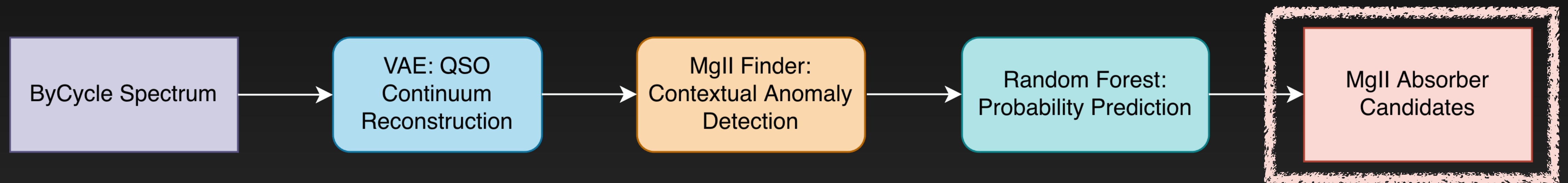
$\text{SNR} < 5.0$



False Positive

Wrong detections

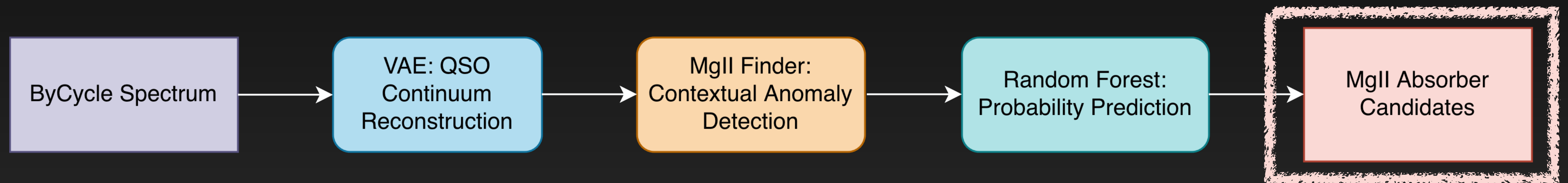
The COOLCAD Pipeline



Absorber catalogue + Completeness corrections

The COOLCAD Pipeline

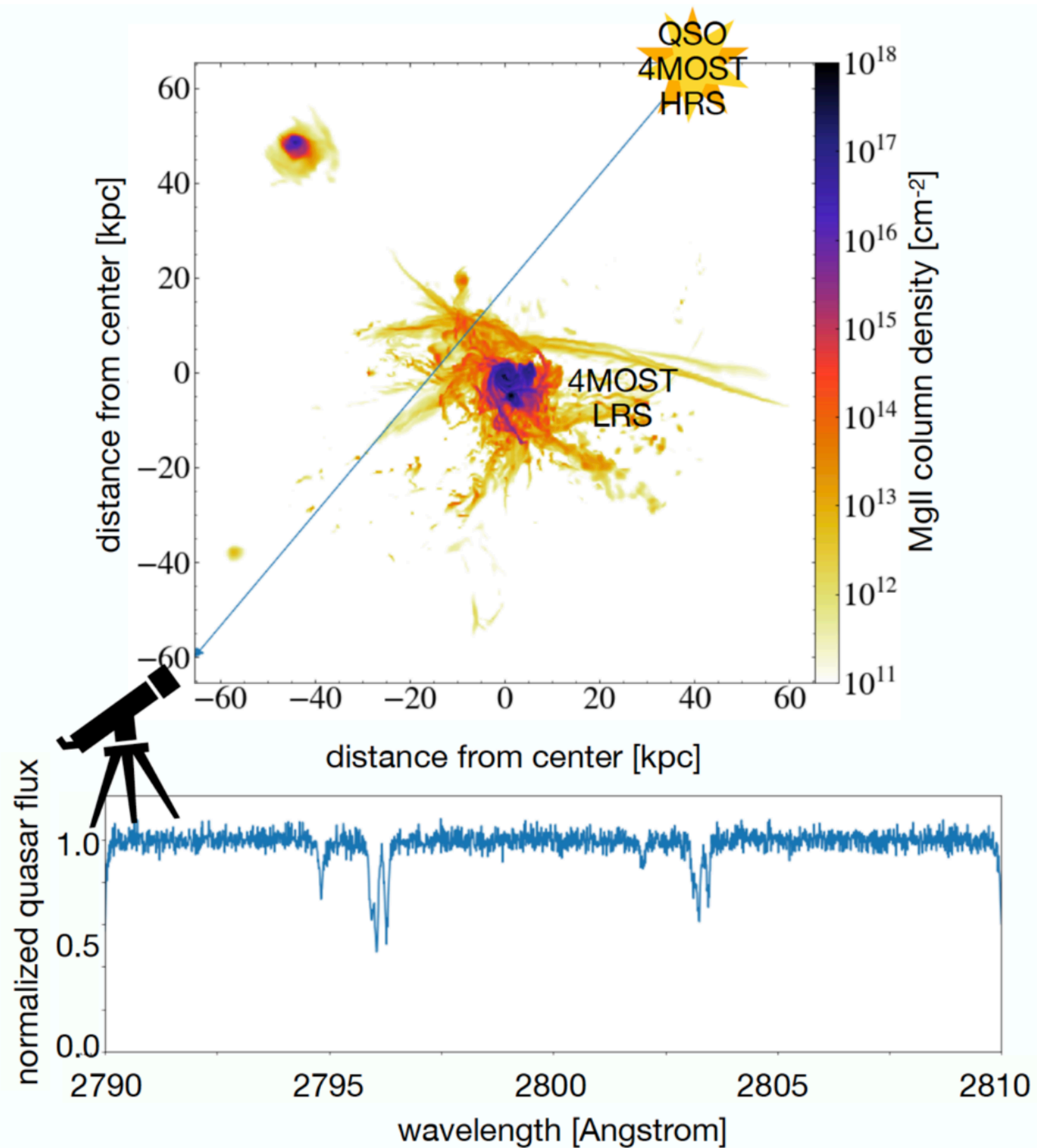
Absorber catalogue + Completeness corrections



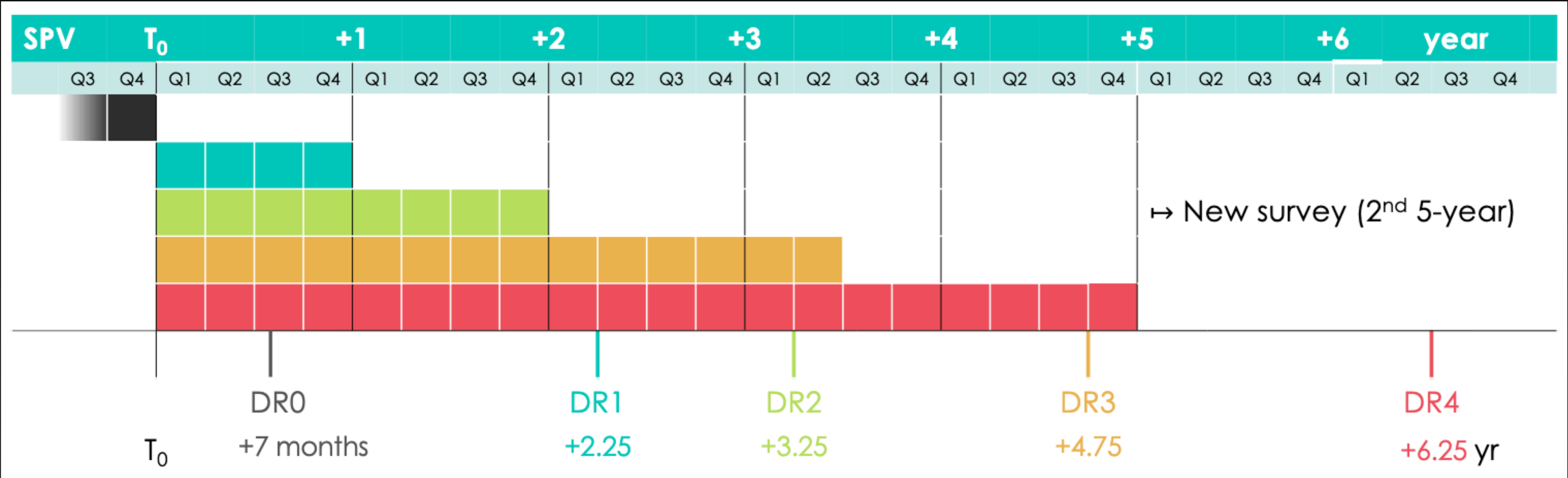
- * Testing on **SPV** in progress! (will we need transfer learning?)
- * Latent spaces to be explored in more detail

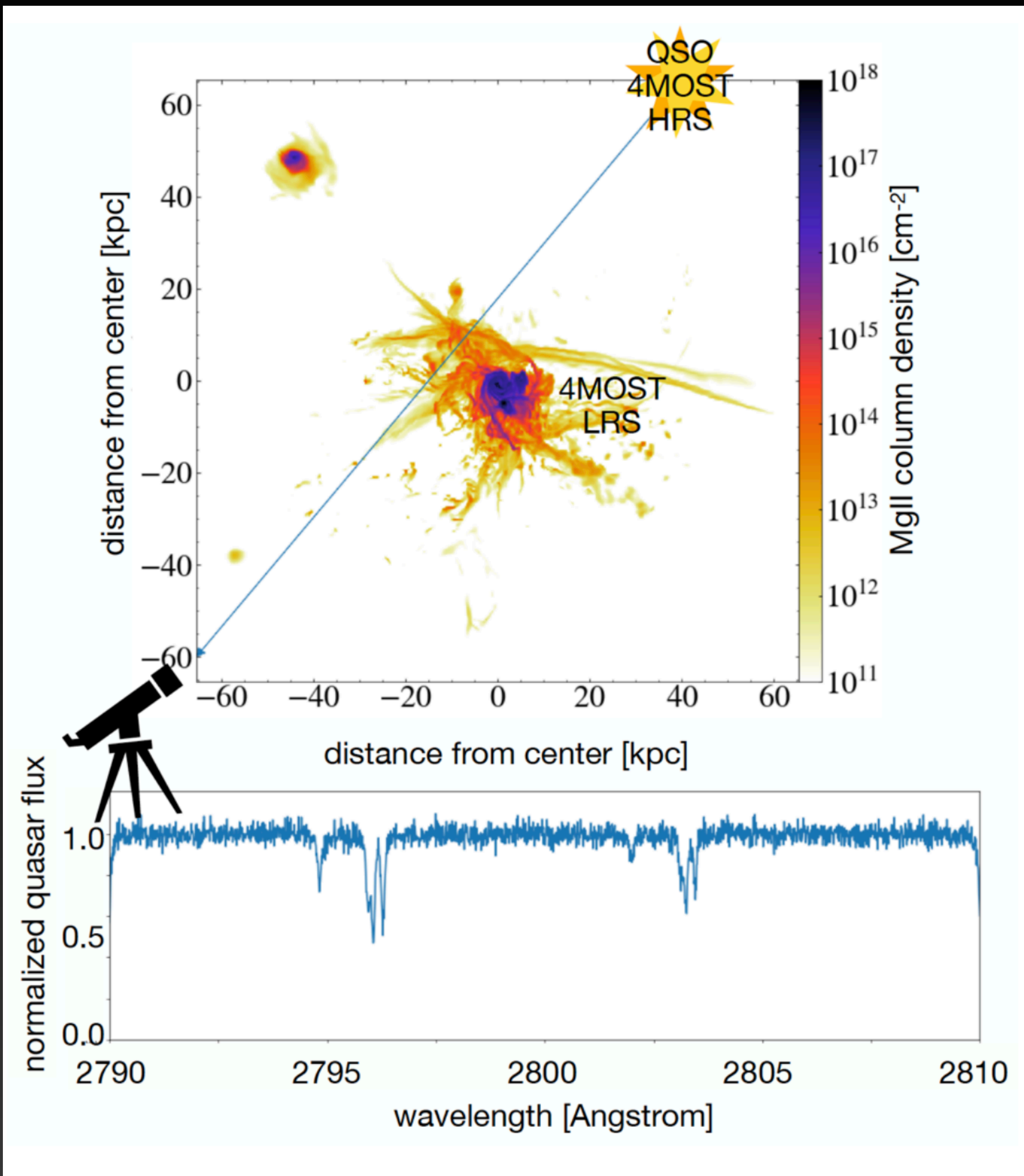
Take-Home Message

ByCycle can find MgII absorbers
with good completeness and purity
by using machine learning



4MOST Public Data Release Plan





Credit: Péroux et al.

Take-Home Message

ByCycle can find MgII absorbers
with good completeness and purity
by using machine learning



Thank you! 🌈🚲

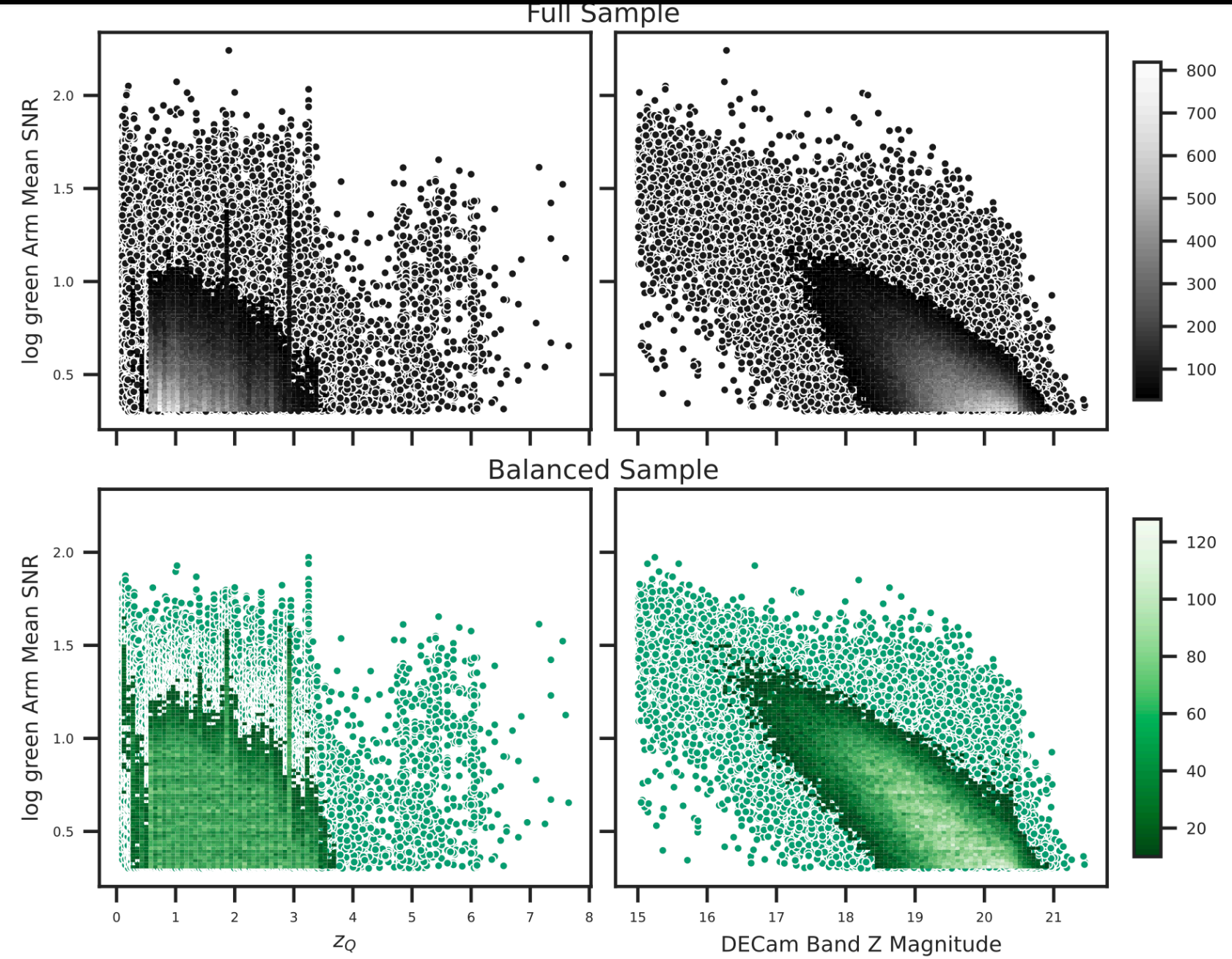
nicolas.guerravaras@eso.org

Training Set Sizes

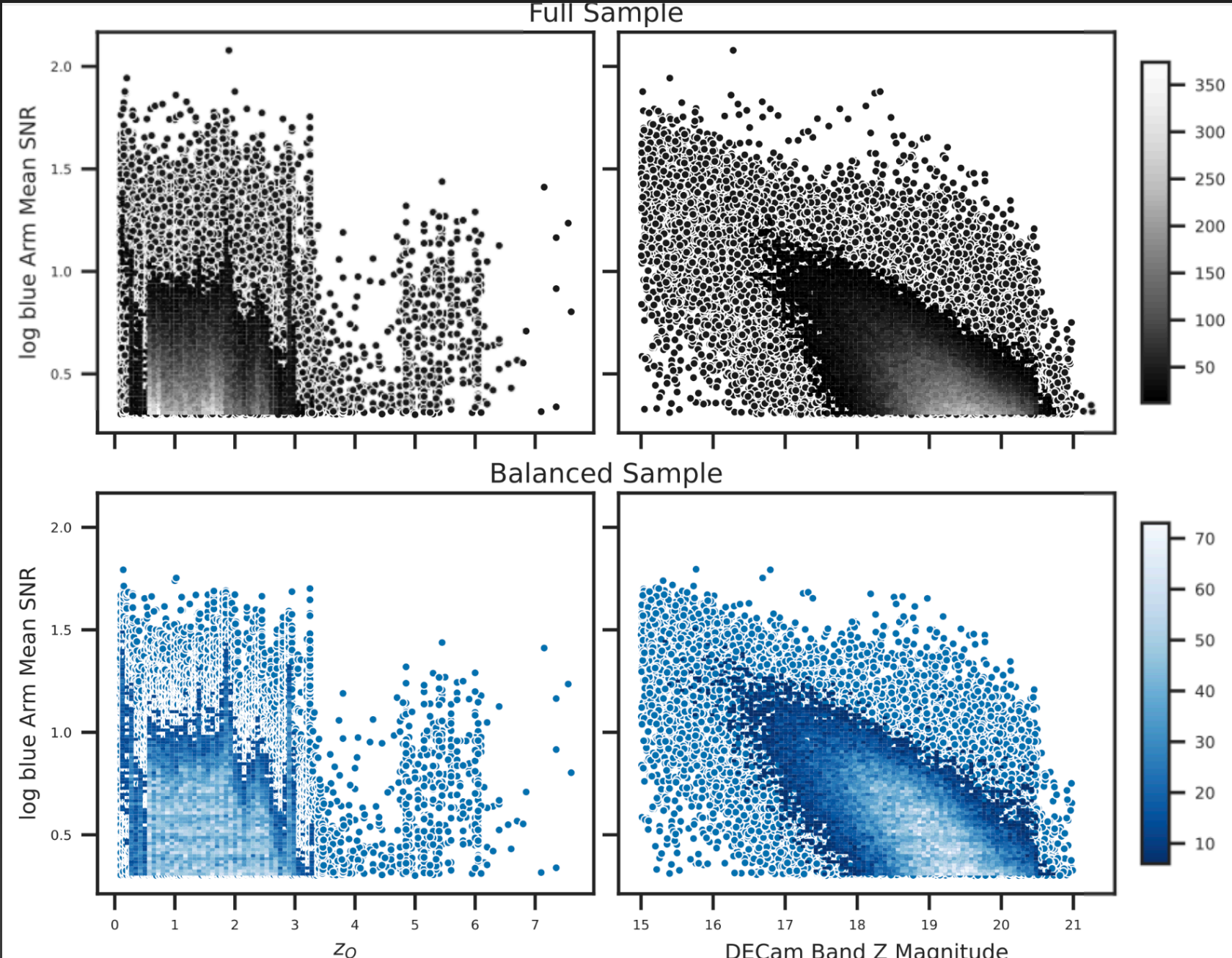
4MOST Spectral Arm	Original	Balanced
Blue	108,756	65,609
Green	220,975	93,102
Red	257,609	98,507

Balanced in redshift and SNR

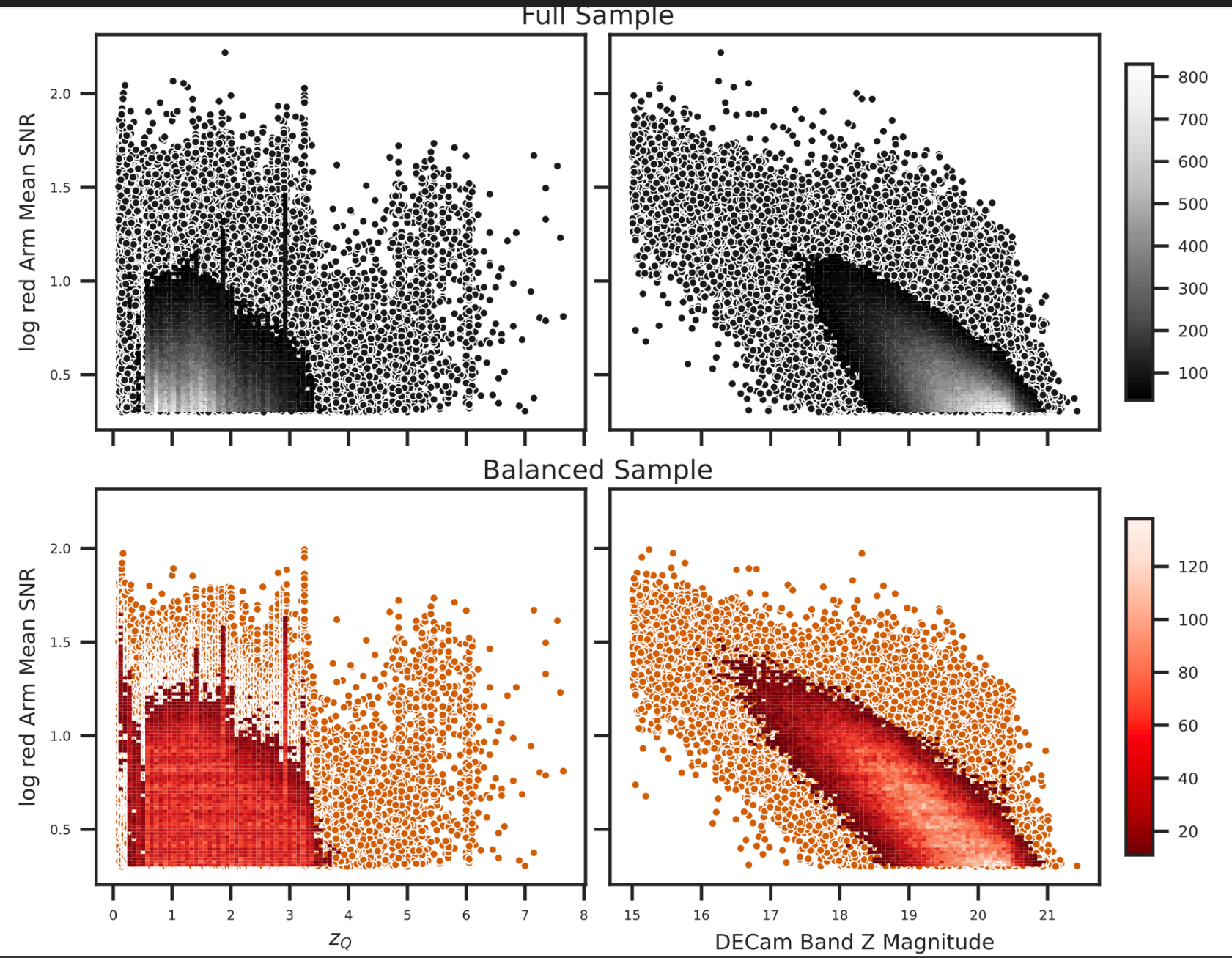
Green Arm



Blue Arm

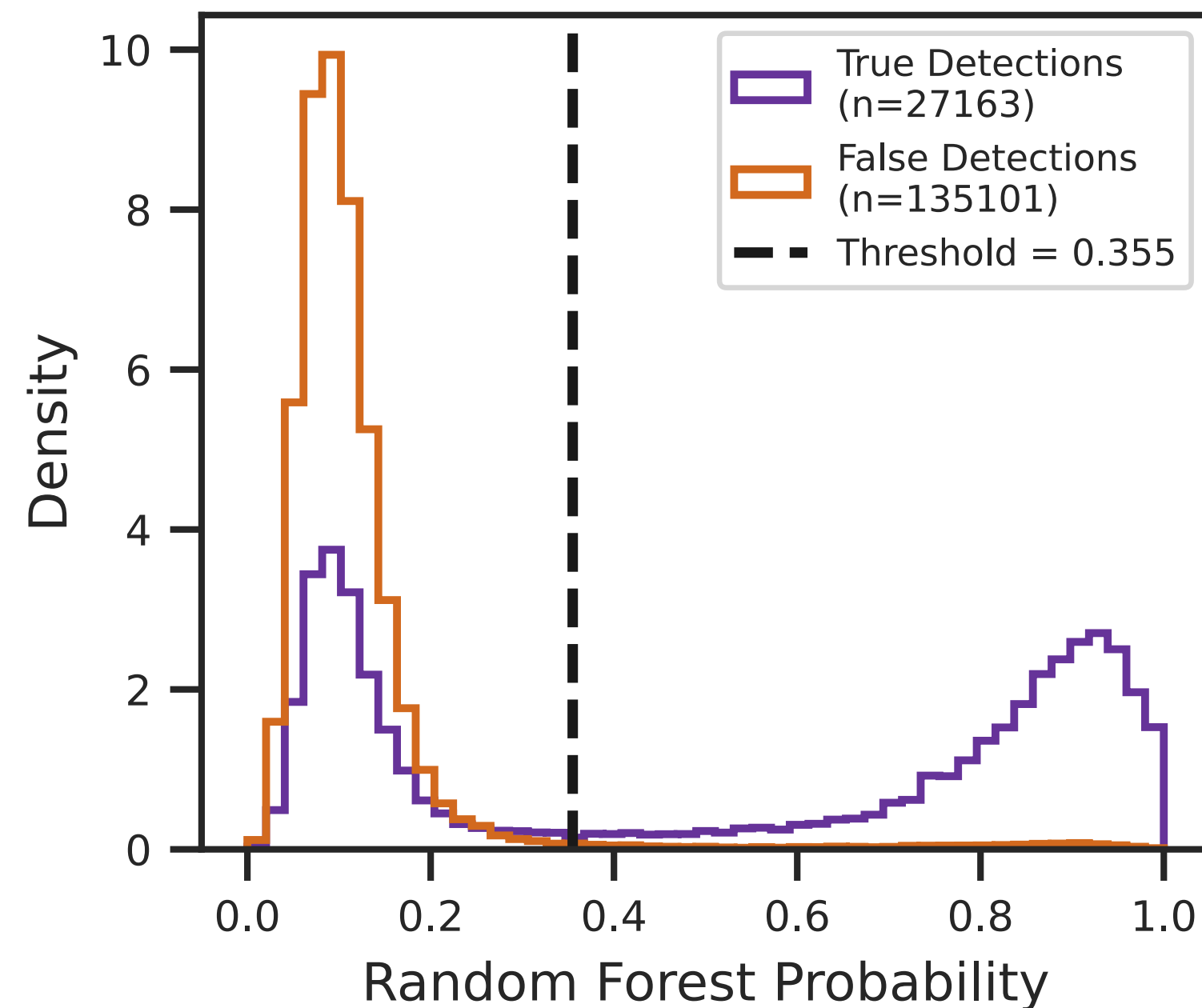


Red Arm

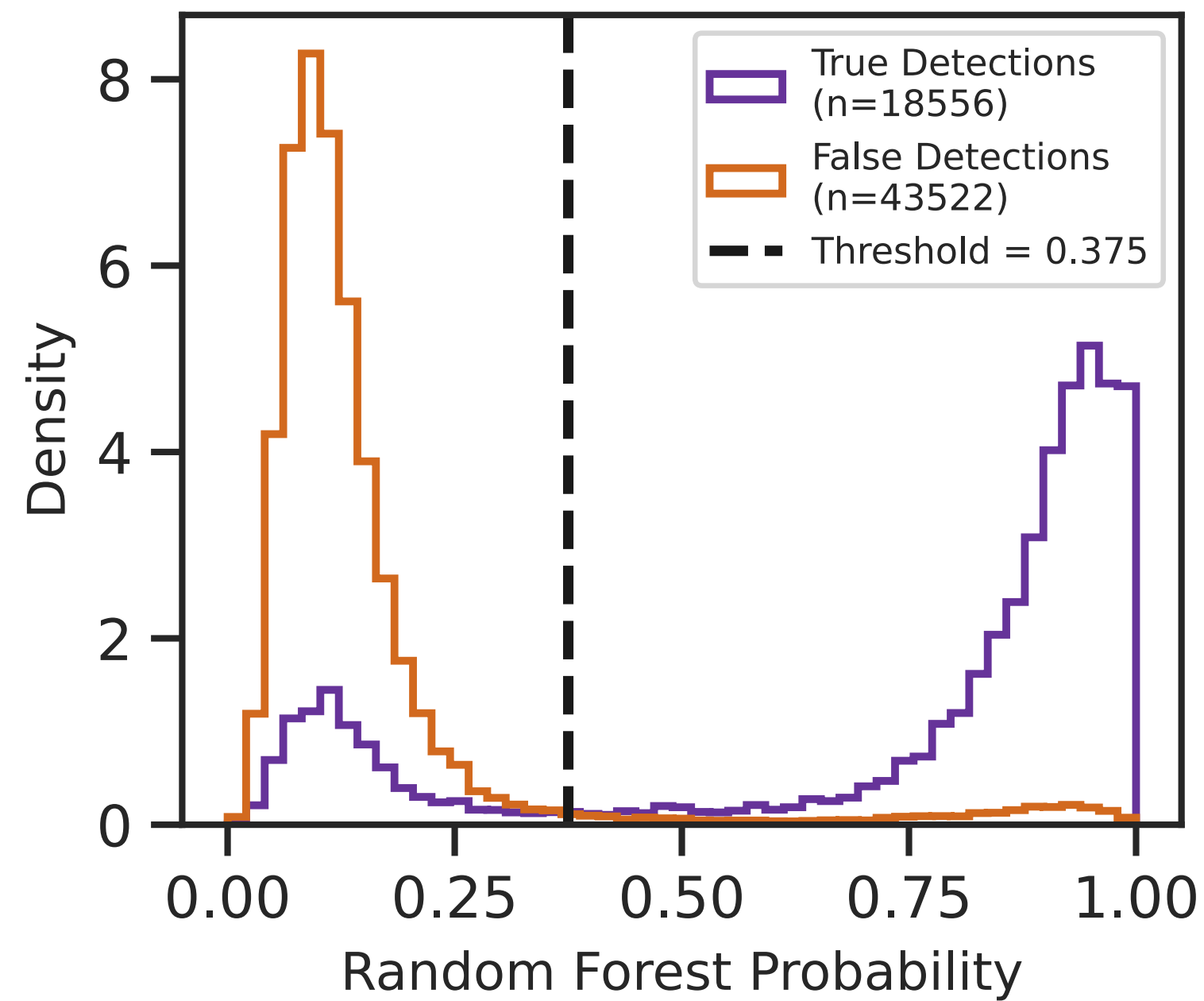


Random Forest: Real vs. False Detections

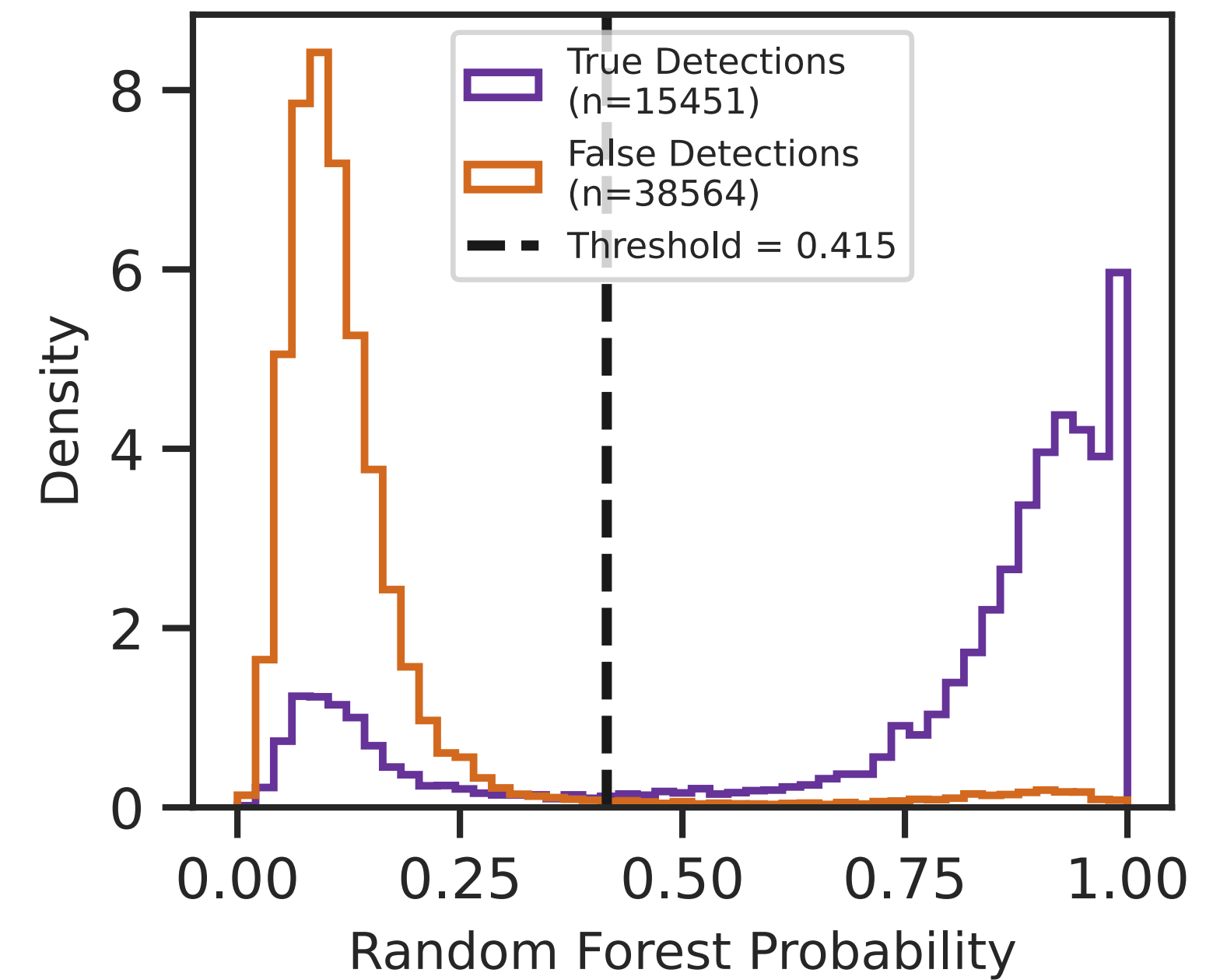
Blue Arm



Green Arm

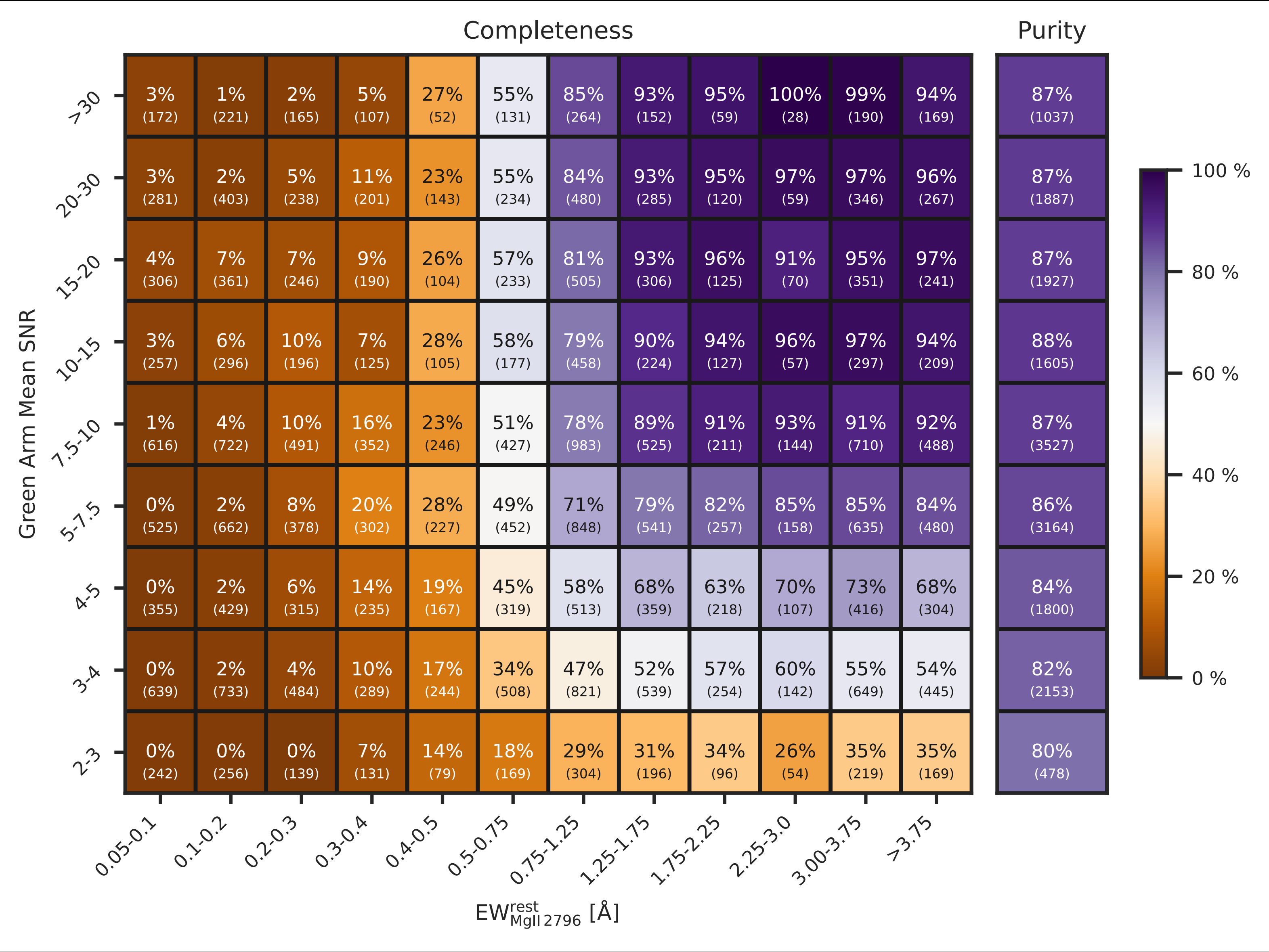


Red Arm

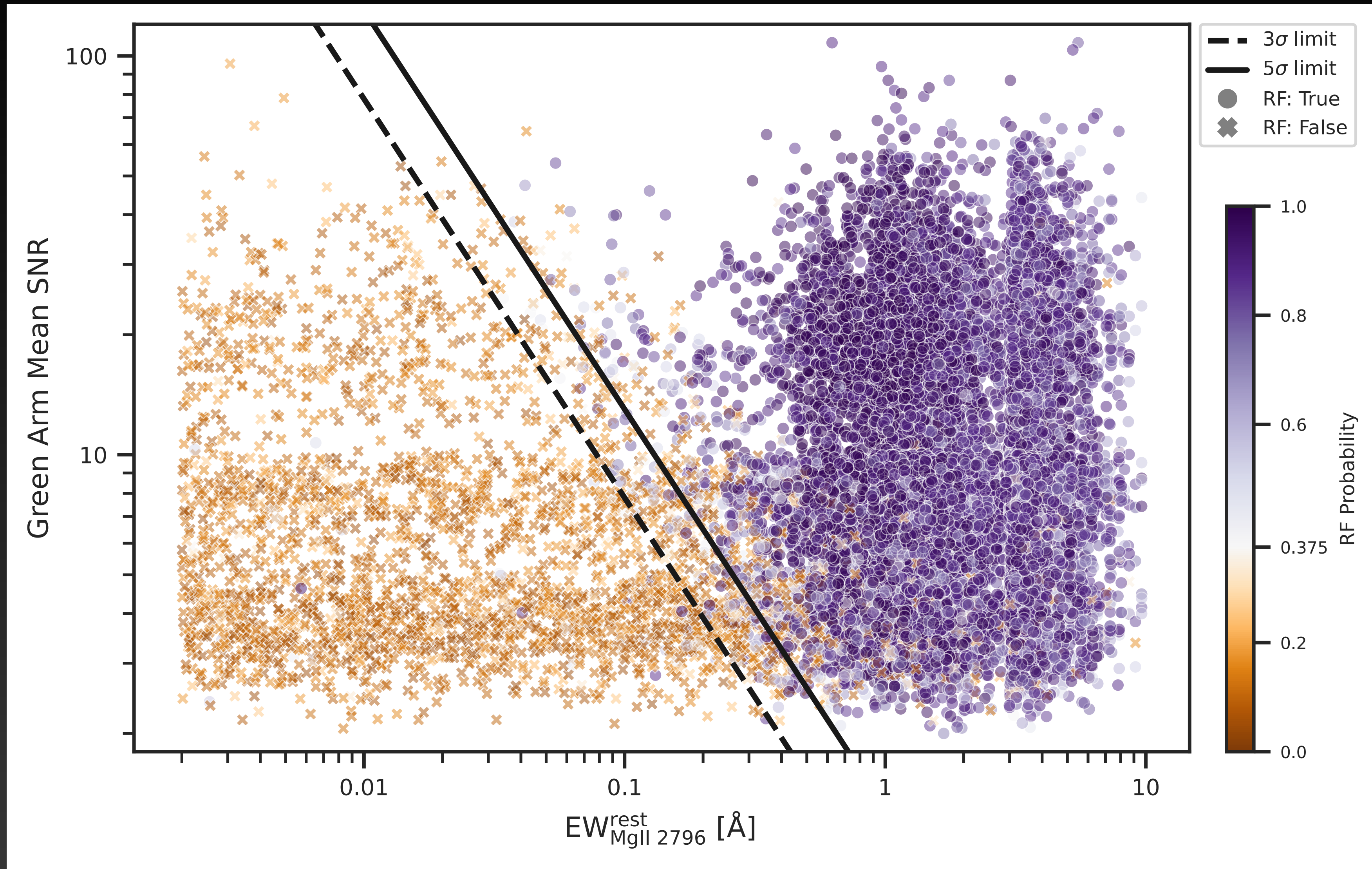


Our RF can effectively separate true absorbers from false detections

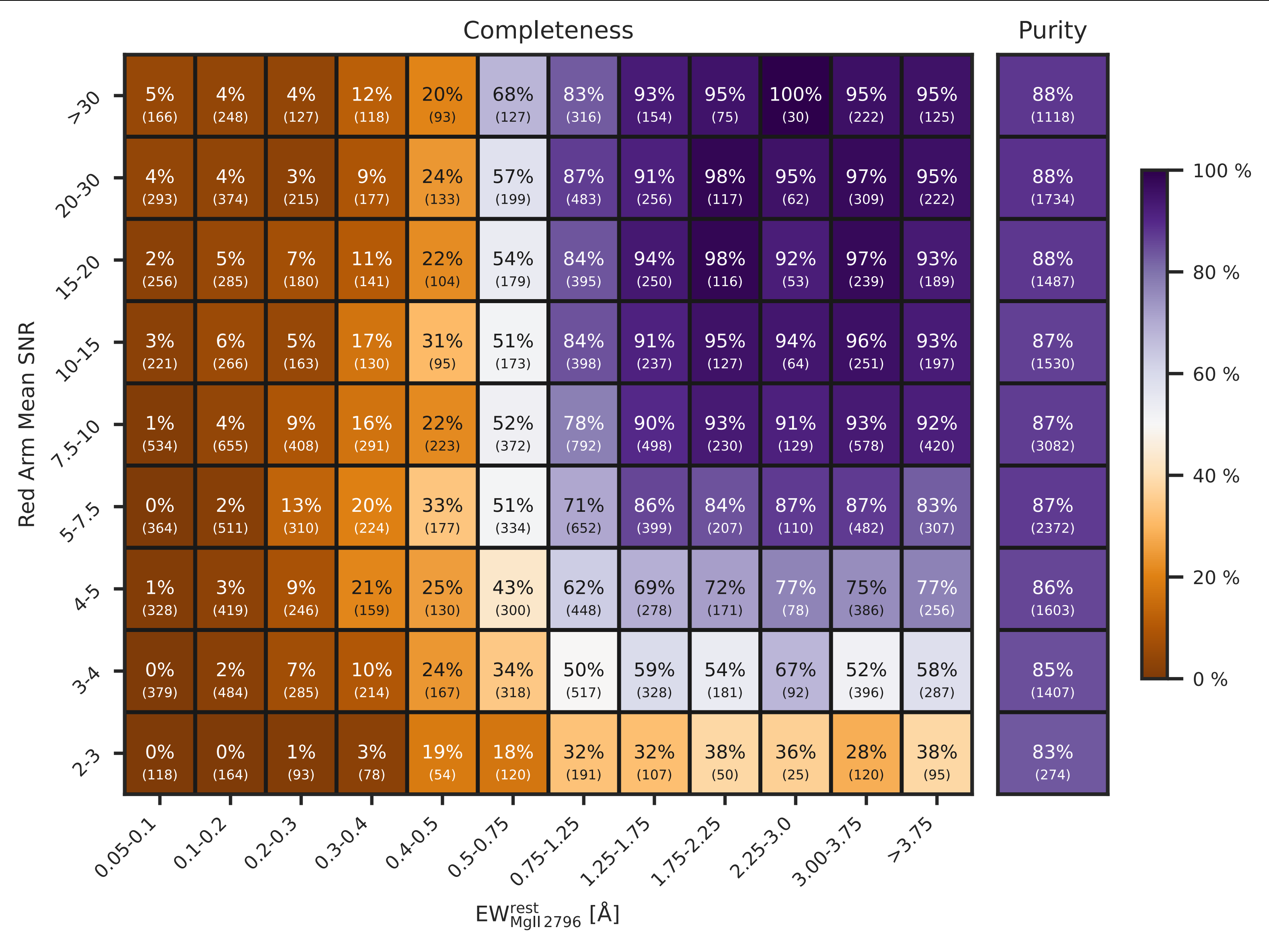
Completeness vs. Purity: green arm



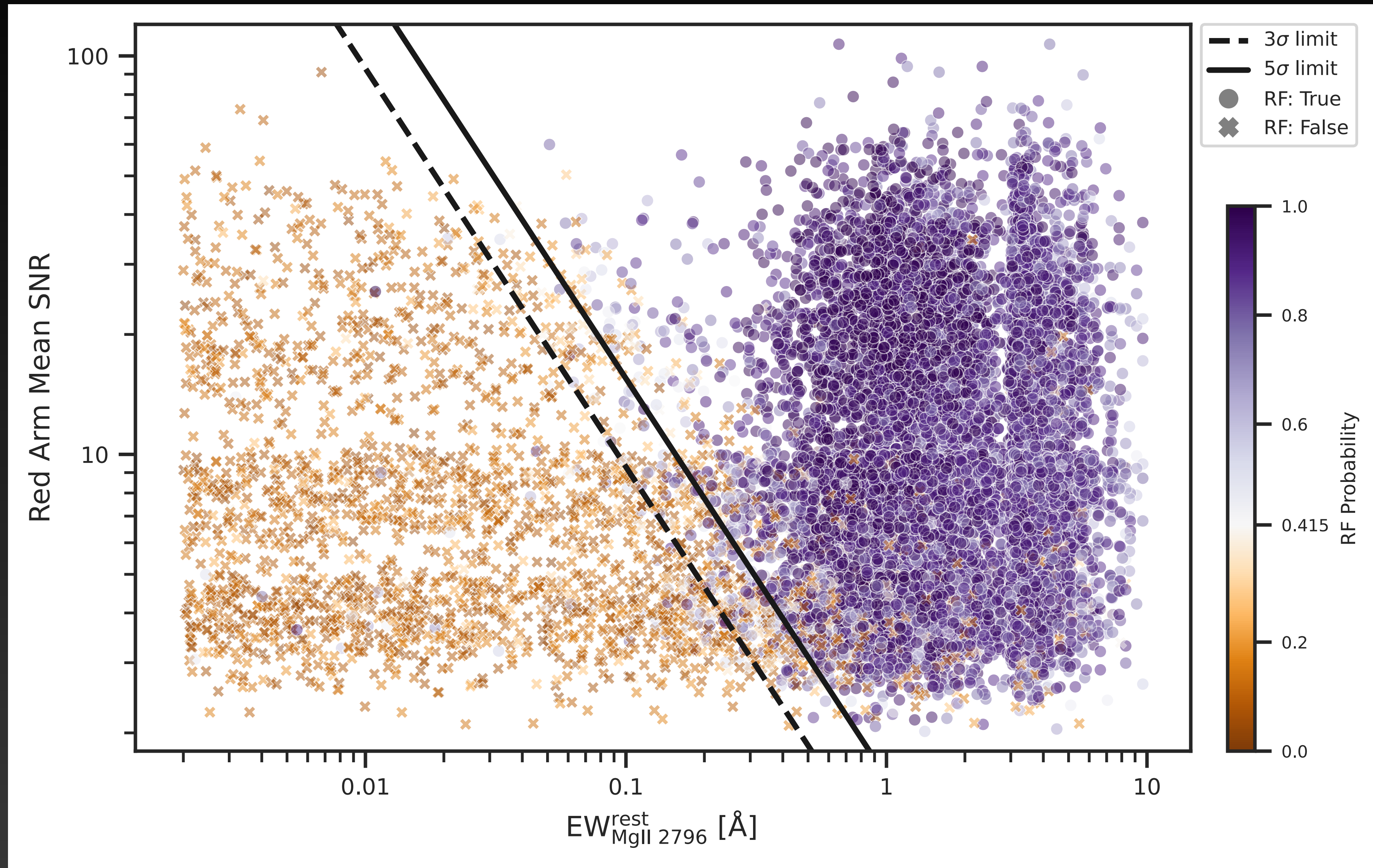
Minimum EW vs SNR: **green** arm



Completeness vs. Purity: red arm



Minimum EW vs SNR: **red** arm



In a spectrum with a resolution $FWHM$, the minimum detectable equivalent width with a $n\sigma$ significance is:

$$W_{\min} = n \times FWHM \times (S/N)^{-1}. \quad (1)$$

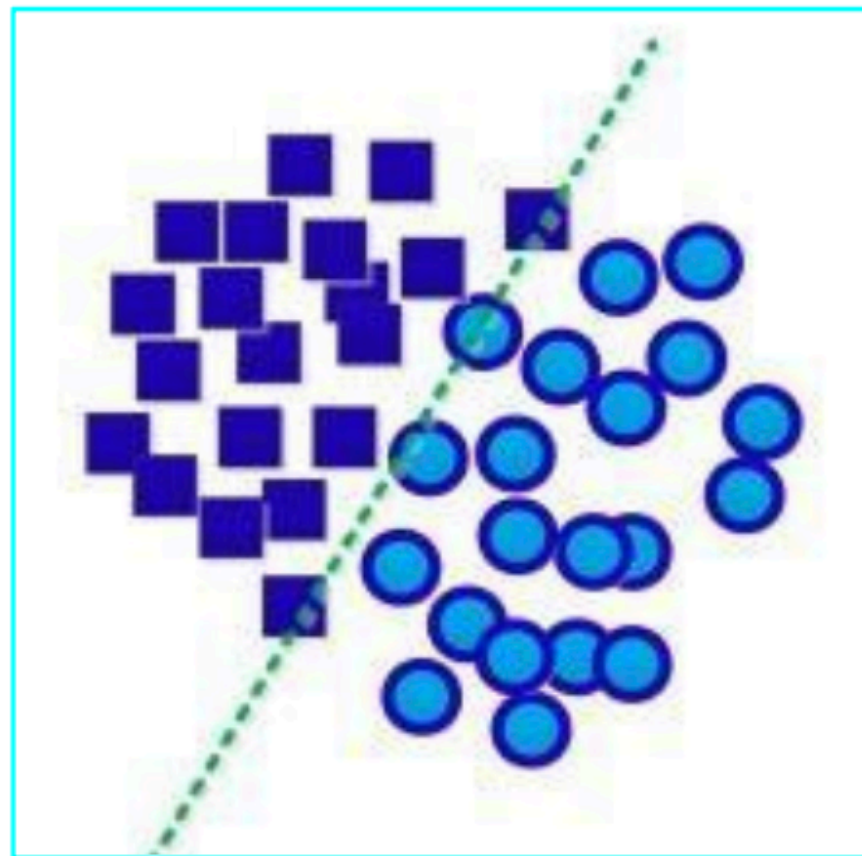
Ménard & Péroux 2003

Will the models need to be re-trained once real data arrives?

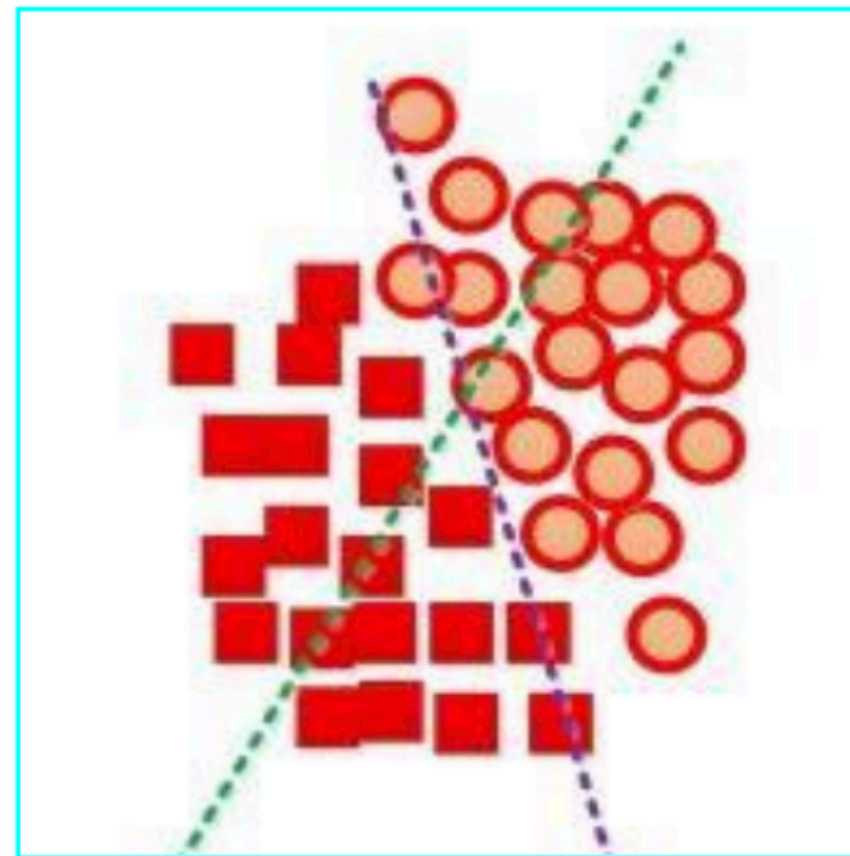
Maybe

Domain Adaptation

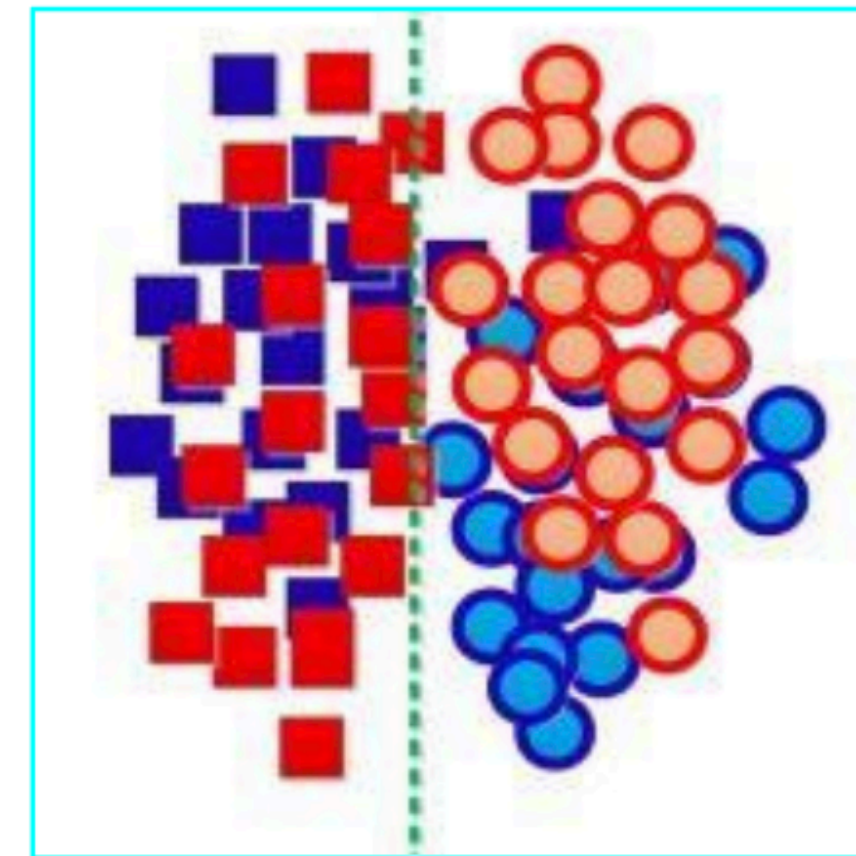
Source Domain



Target Domain



Domain Alignment



Credit: Aleksandra Ćiprijanović