

Francesco Pistis (NCBJ) and the WEAVE-QSO survey team



Ministero
dell'Università
e della Ricerca

 Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA



Funded by the
European Union
NextGenerationEU

THE AI SPECTROSCOPIST: UNVEILING GALAXY AND IGM PHYSICS IN THE ERA OF MILLION- SPECTRUM SURVEY

WMLQ 2026 — 7-11 SEPTEMBER — WARSAW

STRUCTURE OF THE TALK

- ▶ General introduction to quasar spectroscopy
- ▶ Analysis pipeline
- ▶ Pipeline test

THE COSMIC LIGHTHOUSE

- ▶ Among the brightest persistent sources in the Universe
 - ▶ Ideal background sources
- ▶ Accreting supermassive black holes



Artistic illustration of a quasar. Credit: [NASA](#)

WHY STUDYING THE INTERVENING UNIVERSE?

- ▶ Most baryons are not in stars but in diffuse gas
 - ▶ Difficult direct studies
 - ▶ Absorption studies a natural solution
- ▶ Study of:
 - ▶ Cosmic Web (gas tomography)
 - ▶ Intergalactic medium (IGM)
 - ▶ Circumgalactic medium (CGM)



Illustration of the line of sight of a quasar observation. Credit: ChatGPT

QUASAR ABSORPTION SPECTROSCOPY

- ▶ Intervening gas leaves absorption features on the quasar spectrum
- ▶ Gas clouds absorb photons at specific wavelengths
- ▶ Absorption lines appear at redshifted wavelengths

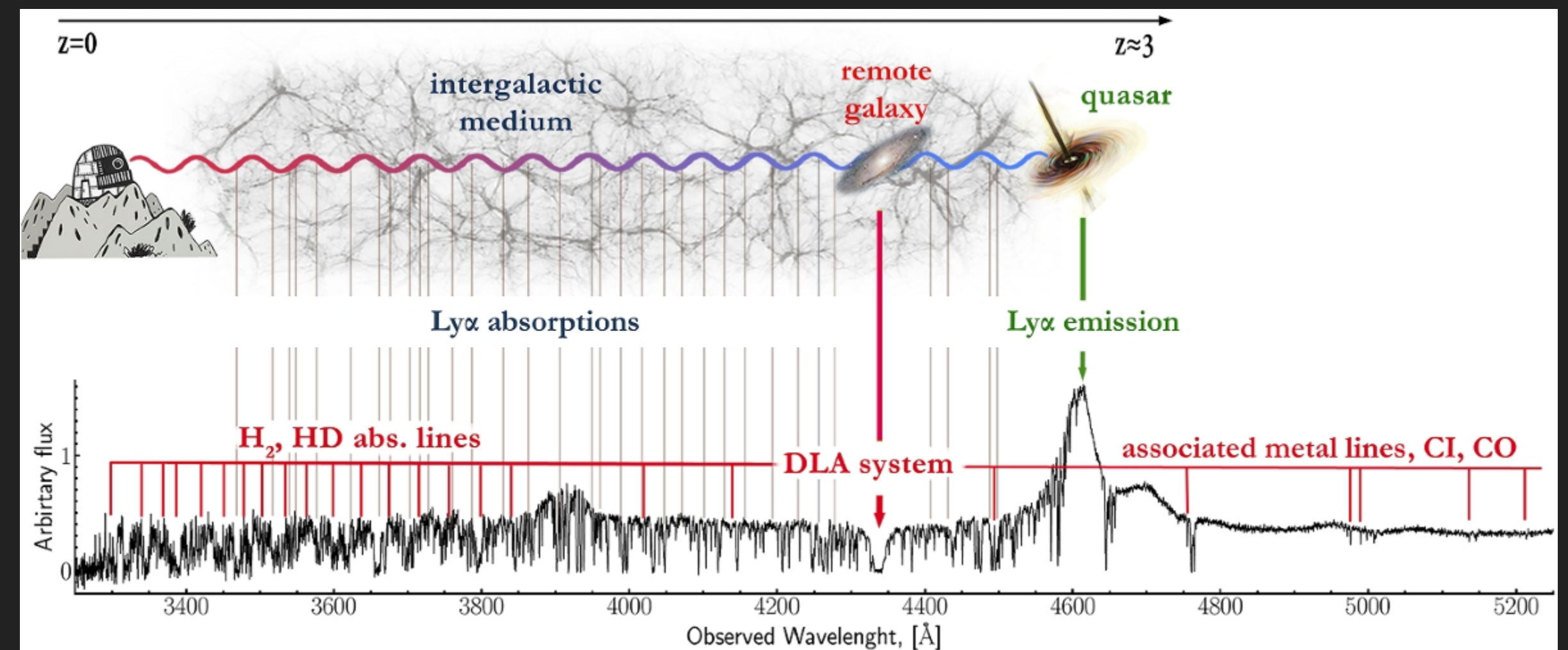


Illustration of absorption features in quasar spectrum.
Credit: [Balashev & Noterdaeme 2022](#)

METAL ABSORPTION SYSTEMS

- ▶ Main metals in exam
 - ▶ CIV
 - ▶ SiIV
 - ▶ MgII
 - ▶ FeII
- ▶ Scientific use
 - ▶ Chemical enrichment
 - ▶ Galactic feedback
 - ▶ Kinematics
 - ▶ CGM studies
 - ▶ Different metals probe different phases of the CGM

CURRENT CHALLENGES AND OPEN QUESTIONS

▶ Technical challenges

- ▶ Toward the million quasar surveys
- ▶ Finding the true continuum levels
- ▶ Detection of low column densities (and, as consequence, low S/N and low EW) lines
- ▶ Classification of absorption lines

▶ Open questions

- ▶ Where are the missing baryons?
- ▶ How galaxies exchange gas with the CGM?
- ▶ How are metals distributed through cosmic time?
- ▶ What is the structure of the multiphase CGM?

ANALYSIS PIPELINE

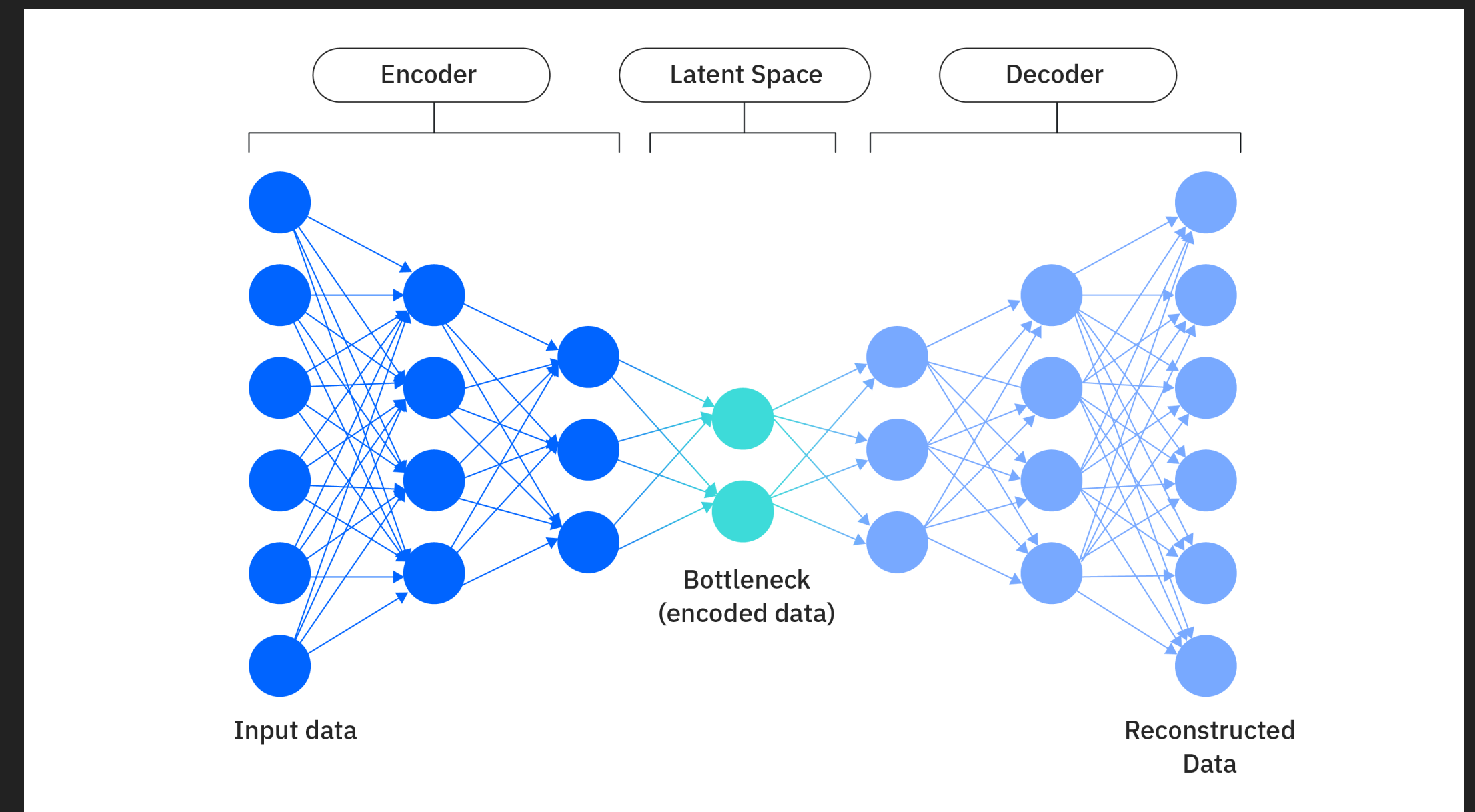
- ▶ Continuum fitting using neural networks
- ▶ Line detection using neural networks
- ▶ PILOT: Physically-Informed Line identificatiOn Tool

WEAVE-LIKE MOCK SAMPLE

- ▶ Pipeline built for the WEAVE (WHT Enhanced Area Velocity Explorer) - QSO survey
- ▶ PCA-based continuum
- ▶ Absorption profiles based on non-negative matrix factorization (NMF) profile maker (NMFP) — total of 24 metal transitions in the sample
- ▶ Absorption clouds distributed randomly along the line of sight

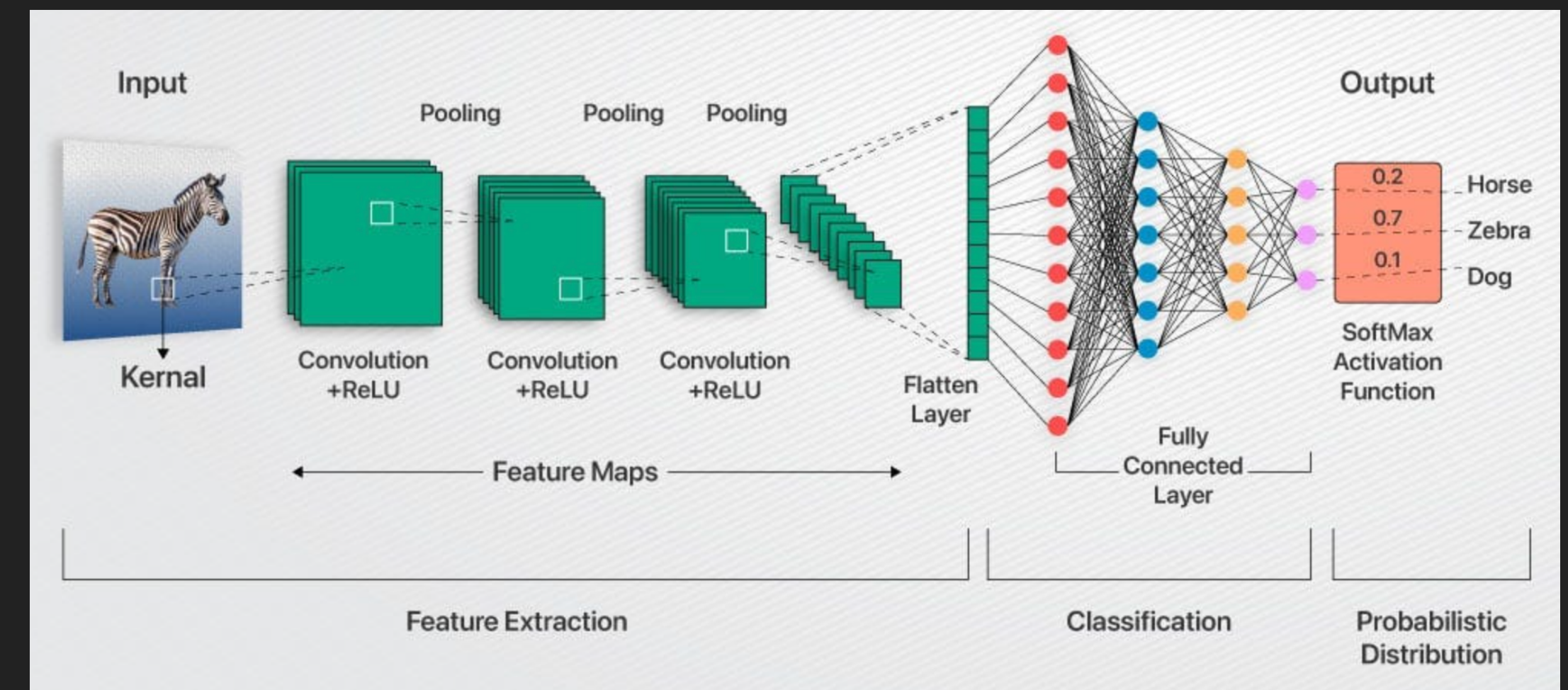
CONTINUUM FITTING

- ▶ All architectures trained in supervised mode
- ▶ Autoencoders
 - ▶ Learn a low-dimensional representation (latent space) of the data
 - ▶ Fast and simple



CONTINUUM FITTING

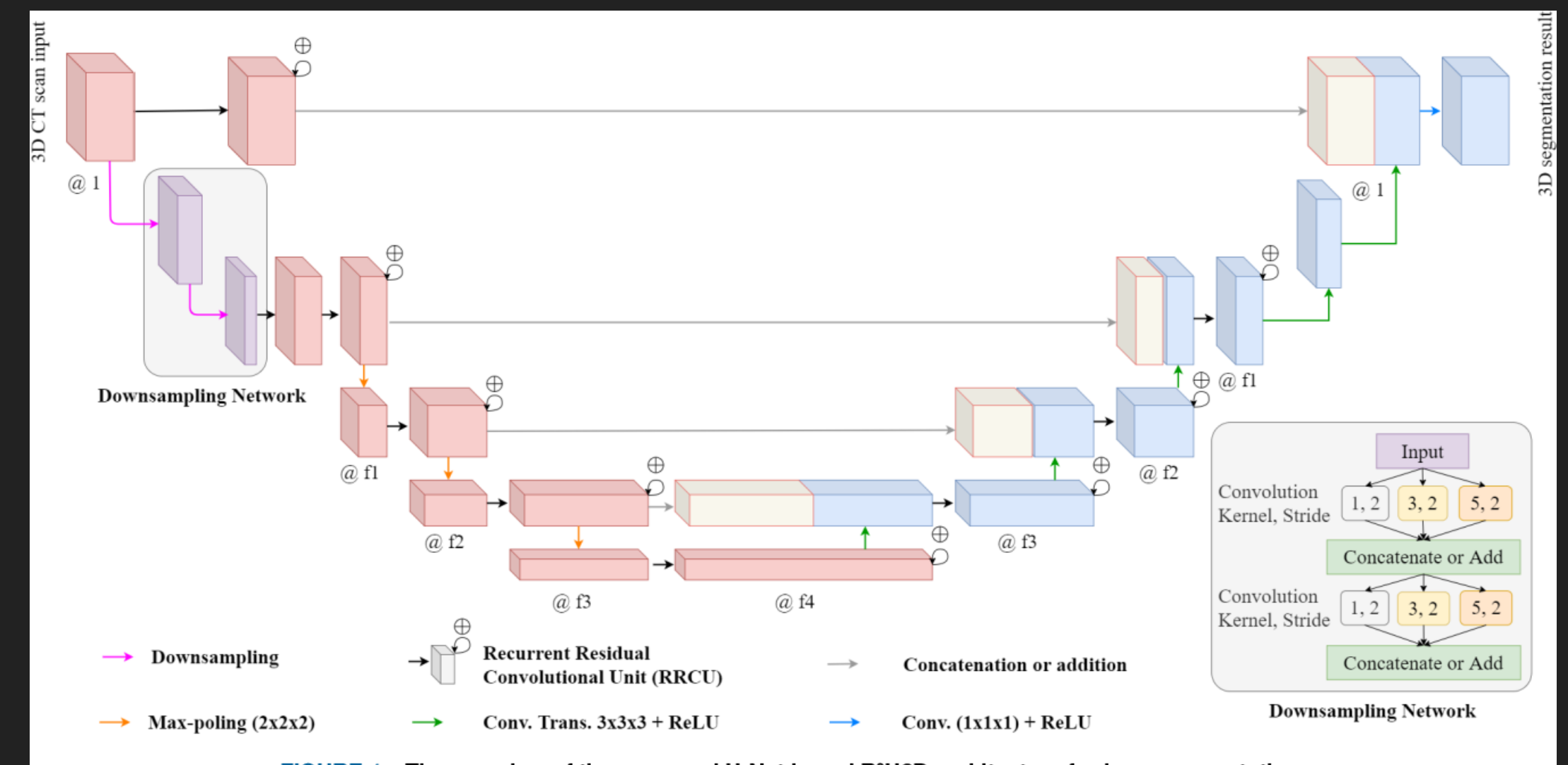
- ▶ All architectures trained in supervised mode
- ▶ Autoencoders
 - ▶ Learn a minimal representation (latent space) of the data
 - ▶ Fast and simple
- ▶ CNN
 - ▶ Convolution layer: detect features in the data
 - ▶ Local connectivity: each neuron focus on a small region



Scheme of a CNN. Credit: [Medium](#)

CONTINUUM FITTING

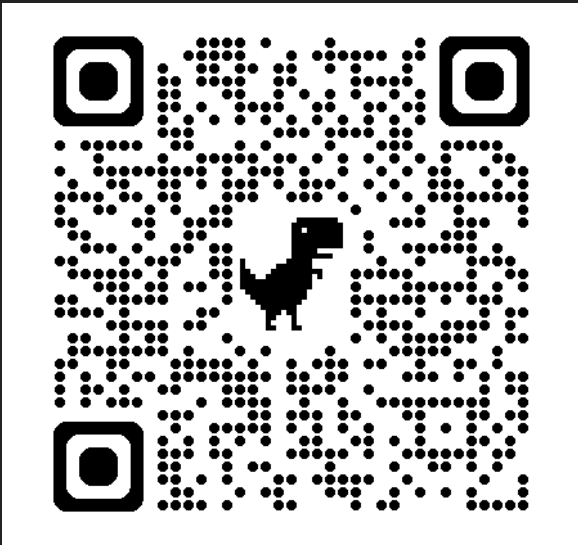
- ▶ All architectures trained in supervised mode
- ▶ Autoencoders
 - ▶ Learn a minimal representation (latent space) of the data
 - ▶ Fast and simple
- ▶ CNN
 - ▶ Convolution layer: detects features in the data
 - ▶ Local connectivity: each neuron focuses on a small region
- ▶ U-Net
 - ▶ Encoder + decoder of convolutional layer
 - ▶ Skip connection to preserve global features



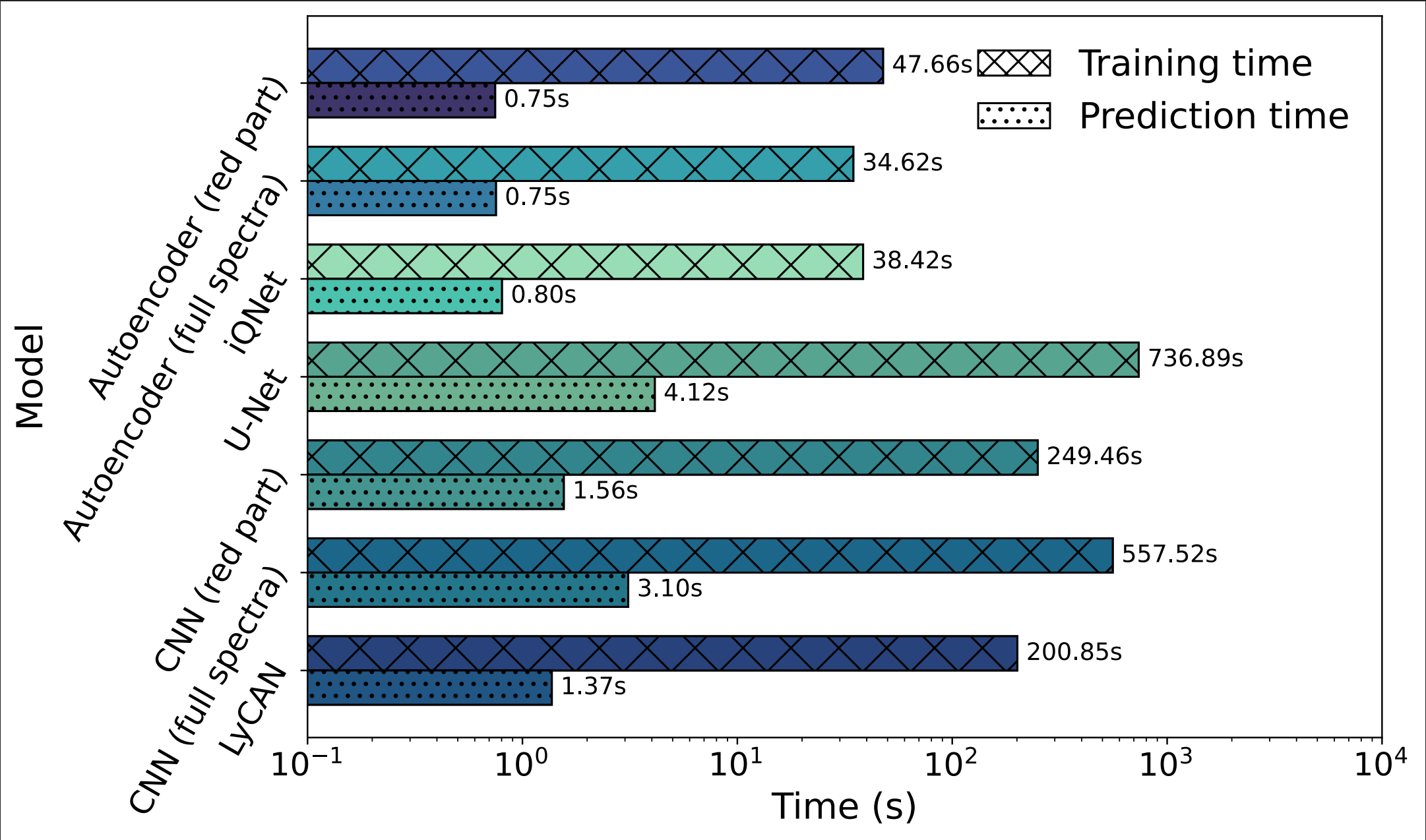
Scheme of a U-Net. Credit: [Medium](#)

CONTINUUM FITTING

► ~ 1s for about 10,000 spectra



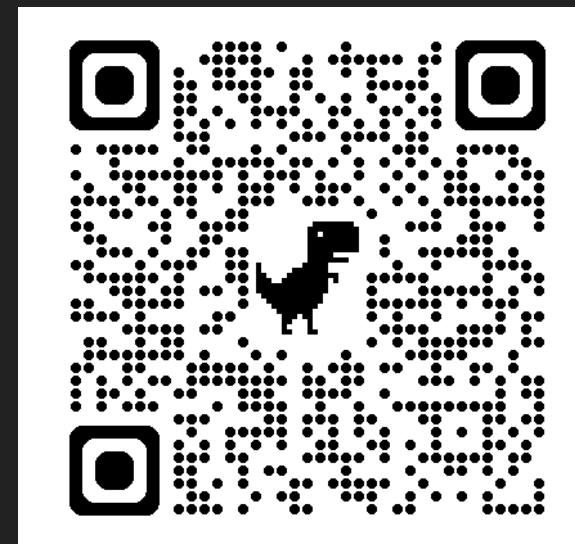
Scan for abstract



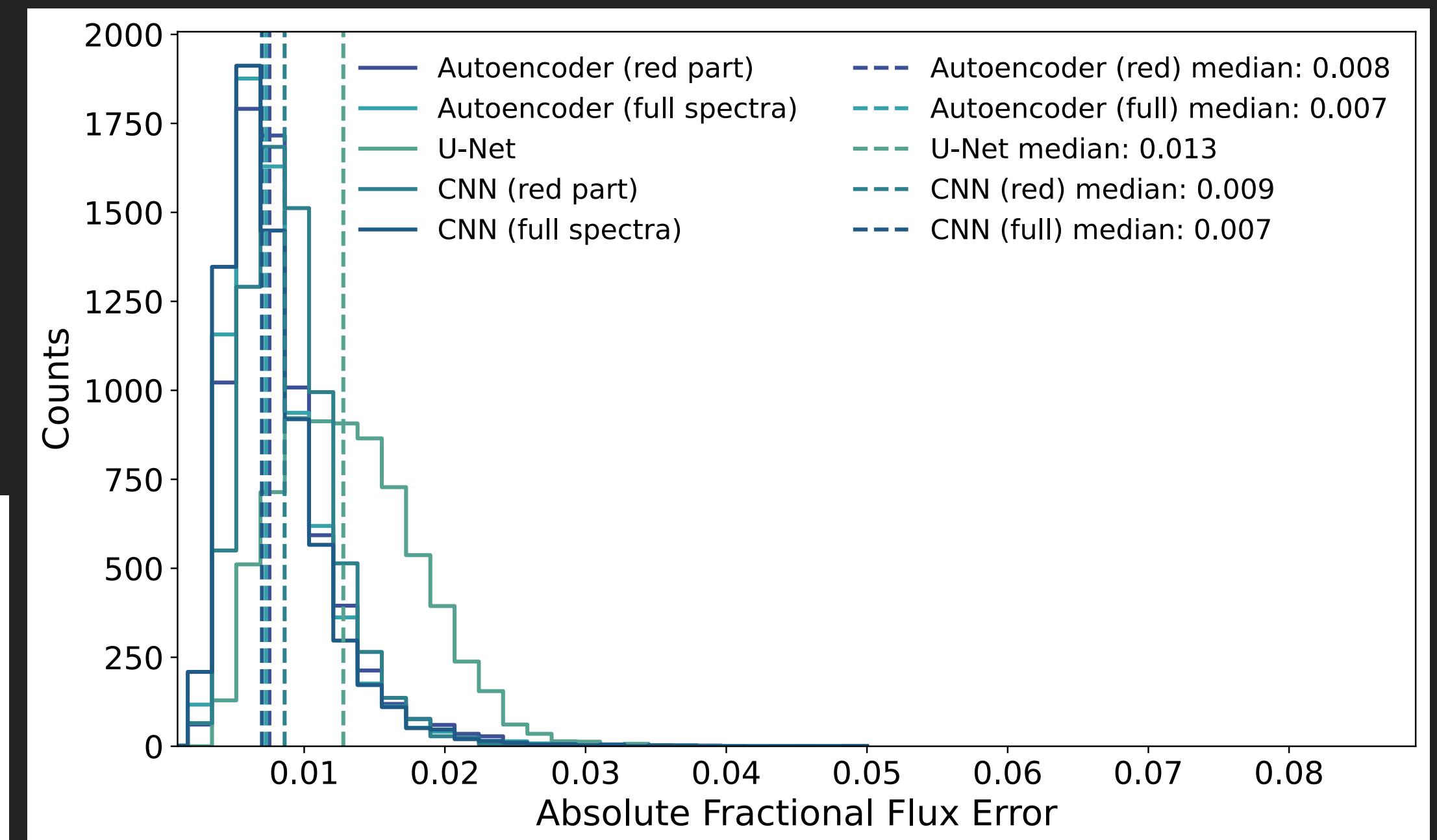
Computational time for different architectures. Credit: [Pistis et al. 2025](#)

CONTINUUM FITTING

- ▶ ~ 1 s for about 10,000 spectra
- ▶ $\sim 1\%$ average error for the whole spectrum



Scan for abstract



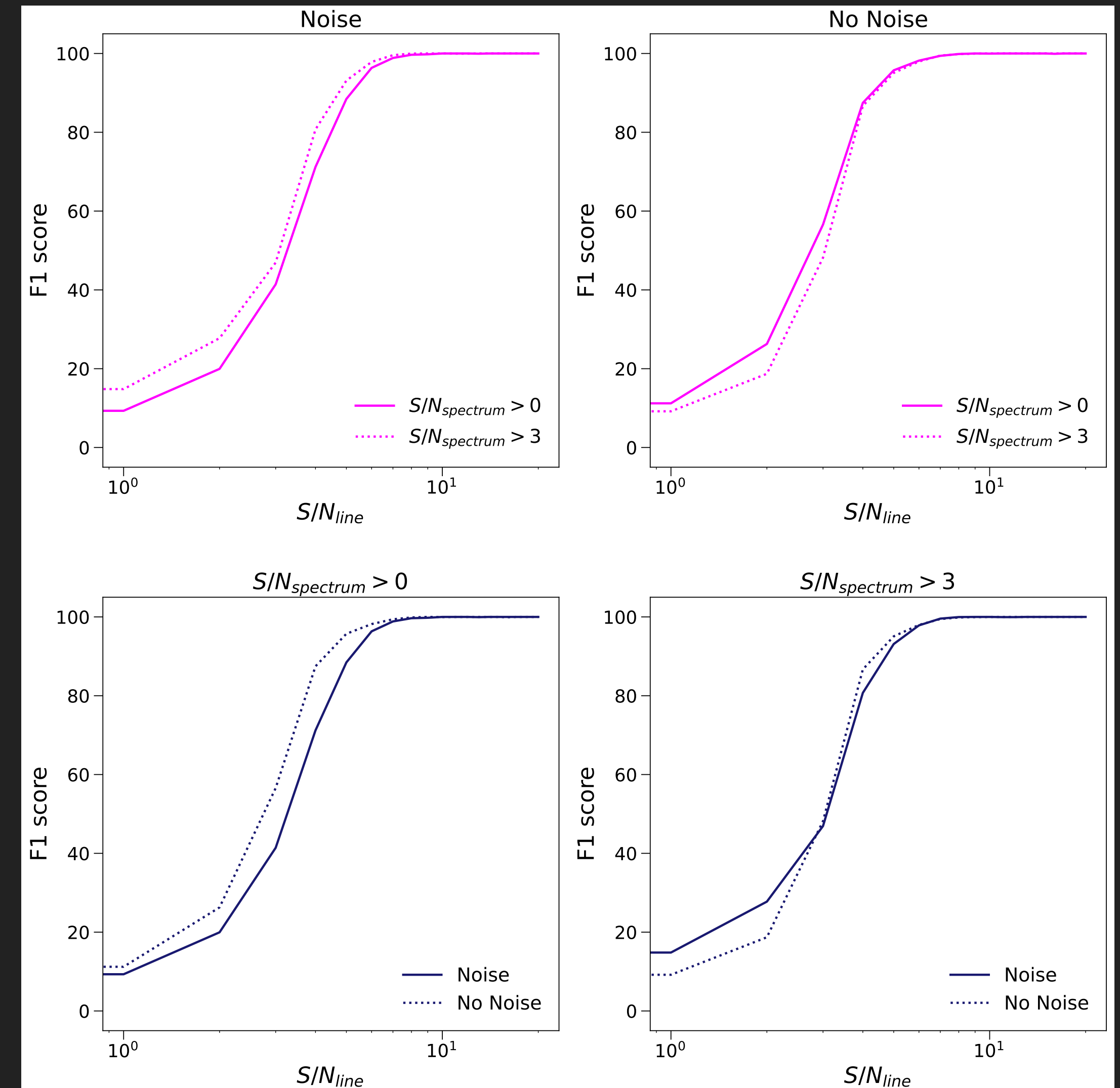
Distribution of the errors on the continuum recovery.
Credit: [Pistis et al. 2025](#)

STANDARD METHODOLOGY: LINE DETECTION AND CLASSIFICATION

- ▶ Focus on doublets
- ▶ Continuum-normalized spectra are convolved with a double Gaussian — mimicking the doublet
- ▶ Potential absorbers identified by an adaptive average S/N threshold (2σ) — standard deviation between ± 200 pixels
- ▶ Double Gaussian fit with varying amplitudes, line widths, and line centers
- ▶ Limited to doublets only
- ▶ Slow — the procedure has to be repeated for each class of doublets
- ▶ The line profile is not Gaussian — it is a Voigt profile
- ▶ Detection of only doublets where both transitions pass the S/N threshold selection — one transition is weaker (S/N ratio of $\sim 2:1$)

LINE DETECTION

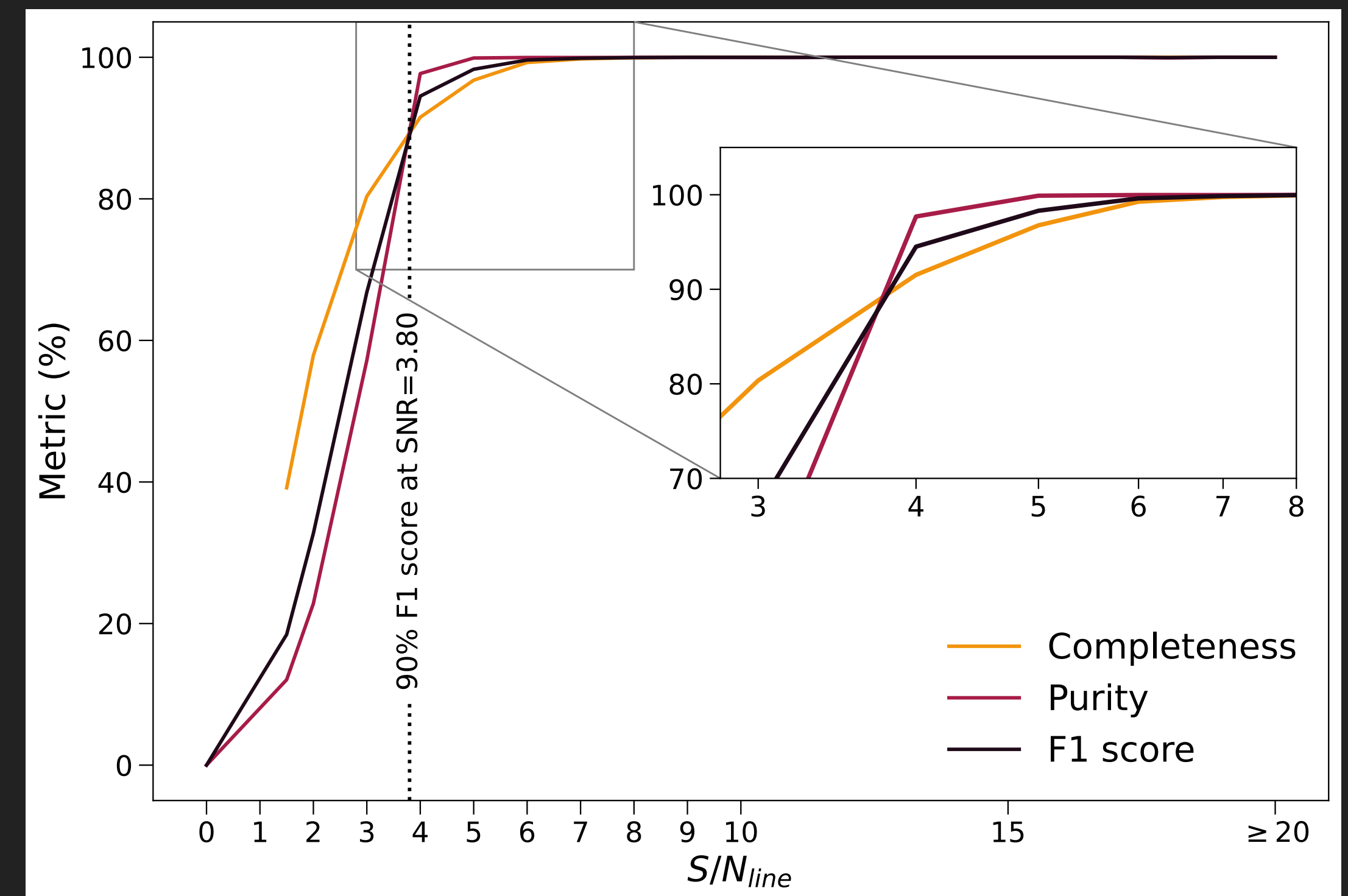
- ▶ Back to U-Net architecture
- ▶ Continuum: 0 — Absorption: 1
- ▶ Comparison between 2 models
 - ▶ Model A: 1-channel input — flux
 - ▶ Model B: 2-channel input — flux + noise
- ▶ Best model: A trained over all S/N range (spectra)



Comparison between models A and B. Credit: Mangola, Pistis et al., submitted

LINE DETECTION

- ▶ Continuum: 0 — Absorption: 1
- ▶ Comparison between 2 models
 - ▶ Model A: 1-channel input — flux
 - ▶ Model B: 2-channel input — flux + noise
- ▶ Best model: A trained over all SNR (spectra)
 - ▶ 90 % on both completeness and purity around S/N (line) of 3.90
- ▶ Negligible difference in metrics if the continuum fitting and line detection are applied consecutively



Metric curves in function of the S/N of the lines. Credit: Mangola, Pistis et al., submitted

PILOT: PHYSICALLY-INFORMED LINE IDENTIFICATION TOOL

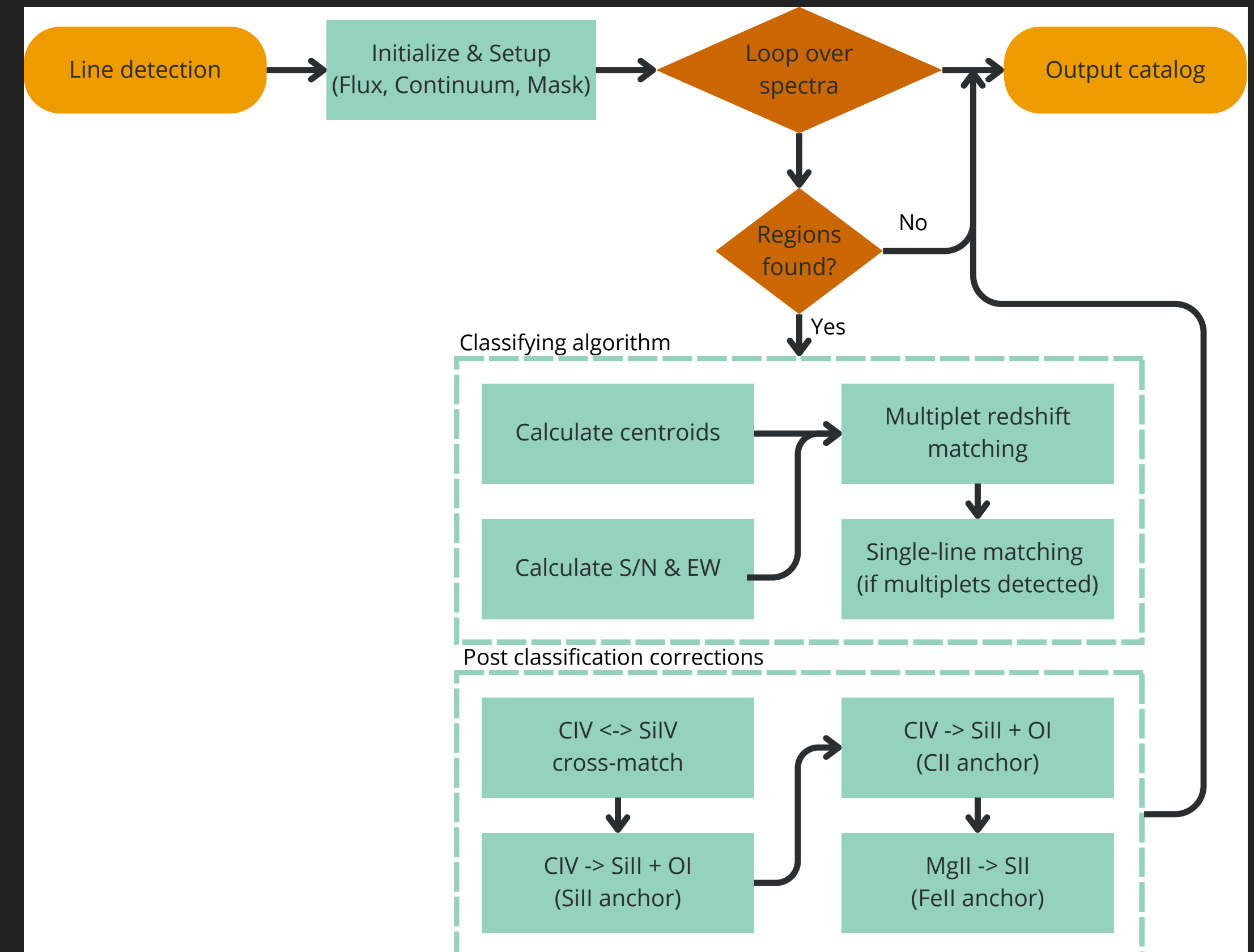
- ▶ The light travels through multiple gas clouds
- ▶ Different clouds have different amounts of heavy elements (metals)
 - ▶ Each cloud can leave multiple absorption lines in the quasar spectrum (different metals)
 - ▶ Each cloud leaves its absorption lines at a different position (due to its redshift)



Illustration of the line of sight of a quasar observation. Credit: ChatGPT

PILOT

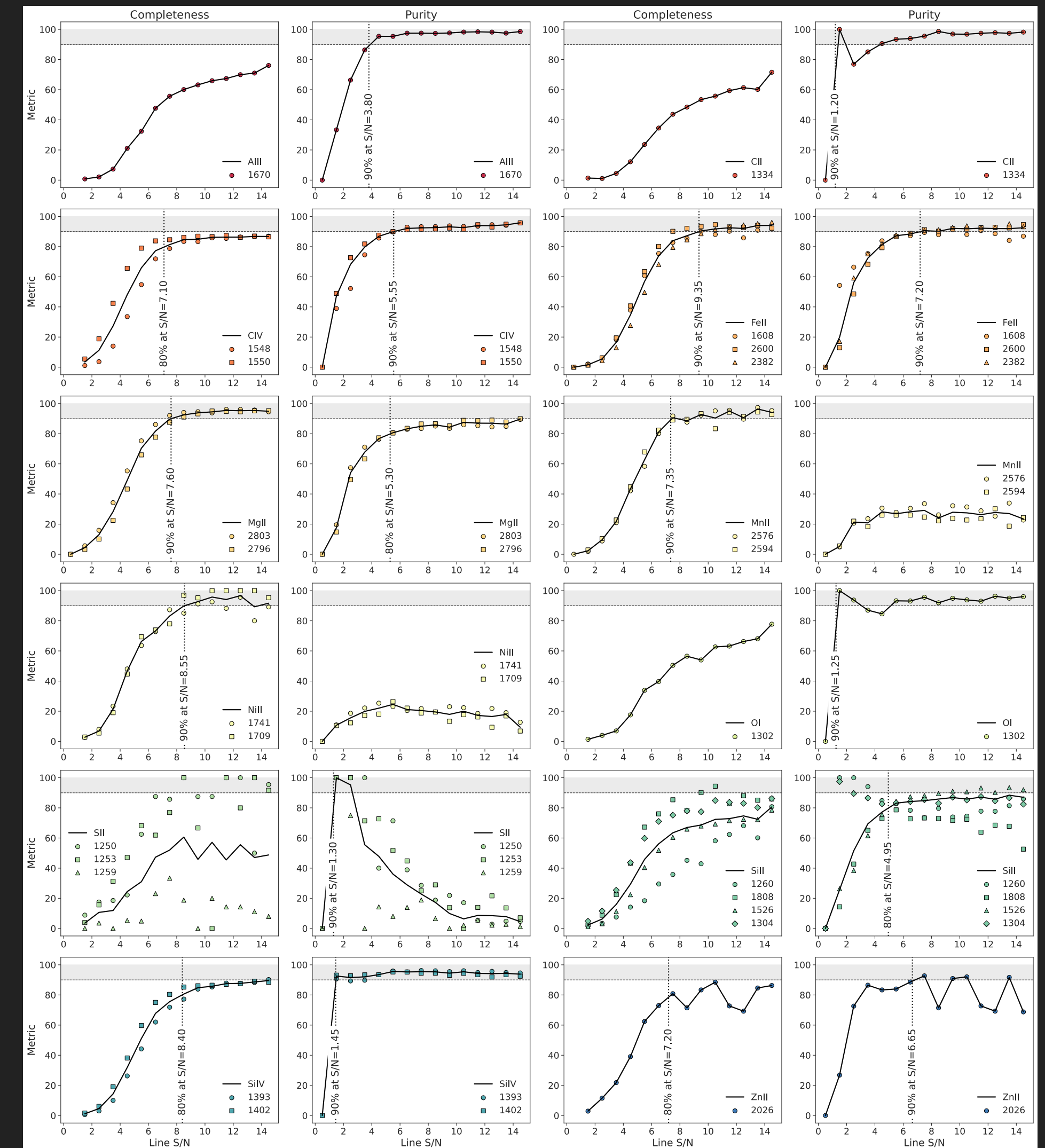
- ▶ **Line centroids $\rightarrow$ redshift (for all possible transitions)**
- ▶ **Exploiting doublets to find absorbing systems (redshift match within a given threshold for transitions of the same metal)**
- ▶ **Looking for singlets and multiplets for the same absorbing systems.**
- ▶ **Post-classification corrections**



Workflow of the line classifier PILOT. Credit: Pistis et al., submitted

PILOT

- ▶ Line centroids → redshift (for all possible transitions)
- ▶ Exploiting doublets to find absorbing systems (redshift match within a given threshold for transitions of the same metal)
- ▶ Looking for singlets and multiplets for the same absorbing systems.
- ▶ Post-classification corrections
- ▶ ~90% metrics for main metals (CIV, SiIV, MgII, FeII)



Completeness and purity curves for metal lines. Credit: Pists et al., submitted

PILOT TESTING: DESI DATA (DR1)

- ▶ No optimization on DESI data; “blind” and direct application — sample of 130,000 spectra
- ▶ Total of 18 million absorptions detected
 - ▶ 5M+ at S/N>4
- ▶ 200,000+ CIV 1548 Å with S/N>1.5
 - ▶ This work: 1.5 absorbers per spectrum
 - ▶ Literature: 0.3 absorbers per spectrum
 - ▶ 5 times more absorbers than the standard method

Table 1 Sample Statistics for Different Catalogs			
Object	DESI (DR2)	DESI (DR1)	SDSS (DR7)
Total quasars	787,166 ^a	241,862	48,260 ^b
Selected quasars ^c	300,637	94,986	26,030
C IV absorbers ^d	101,487	32,321	14,772
C IV (EW ₁₅₄₈ > 0.4 Å)	73,010	23,385	13,514

Notes. References: DESI (Current), Sloan Digital Sky Survey (SDSS): K. L. Cooksey et al. (2013).

^a Selection cuts: $130 < \text{TSNR}^2 < 400$ and $z \geq 1.4$.

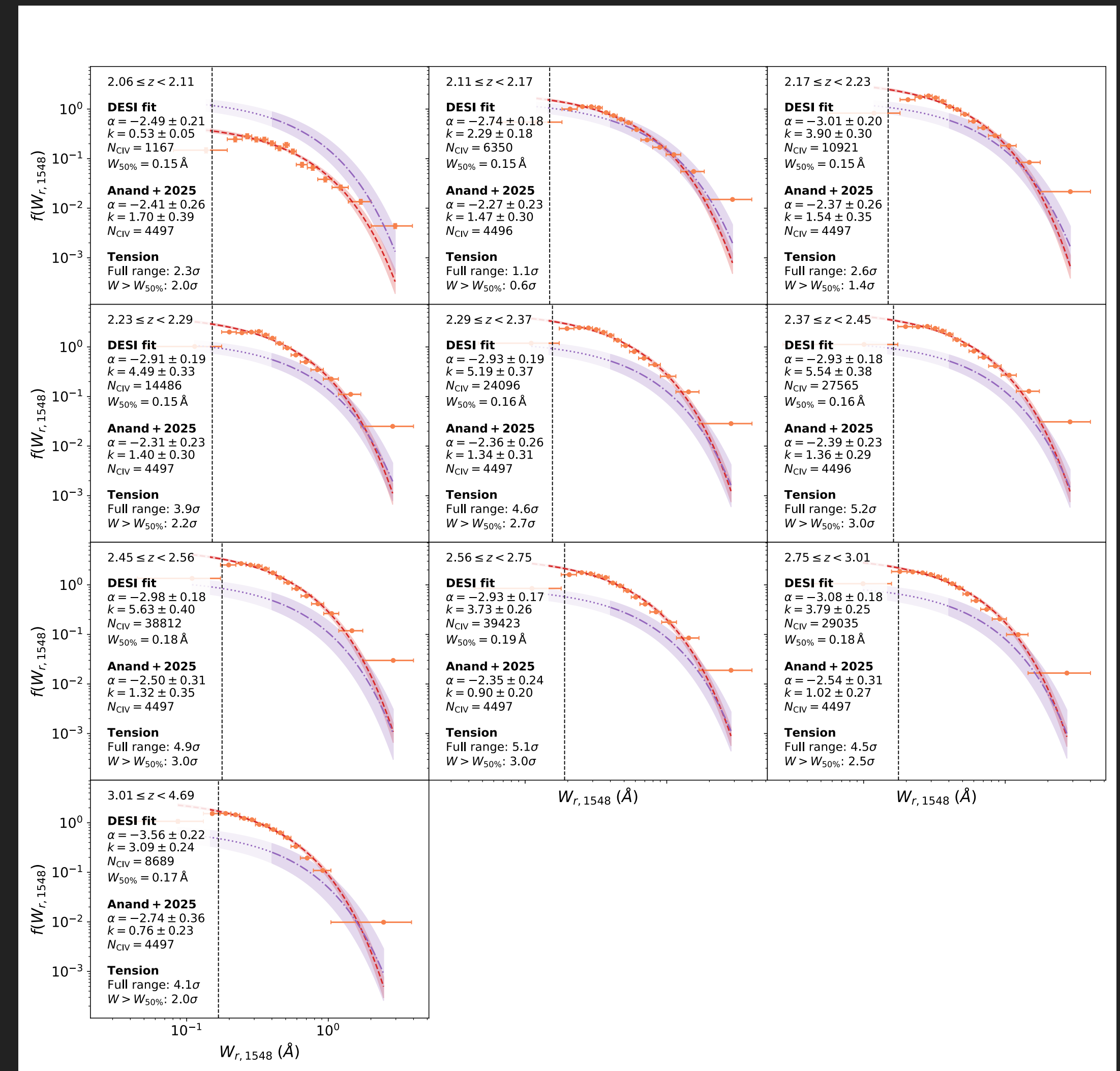
^b For SDSS, $z \geq 1.7$.

^c For DESI, $\langle \text{SNR} \rangle \geq 3 \text{ pix}^{-1}$; for SDSS, $\langle \text{SNR} \rangle \geq 4 \text{ pix}^{-1}$.

^d In selected quasars.

PILOT TESTING: DESI DATA (DR1)

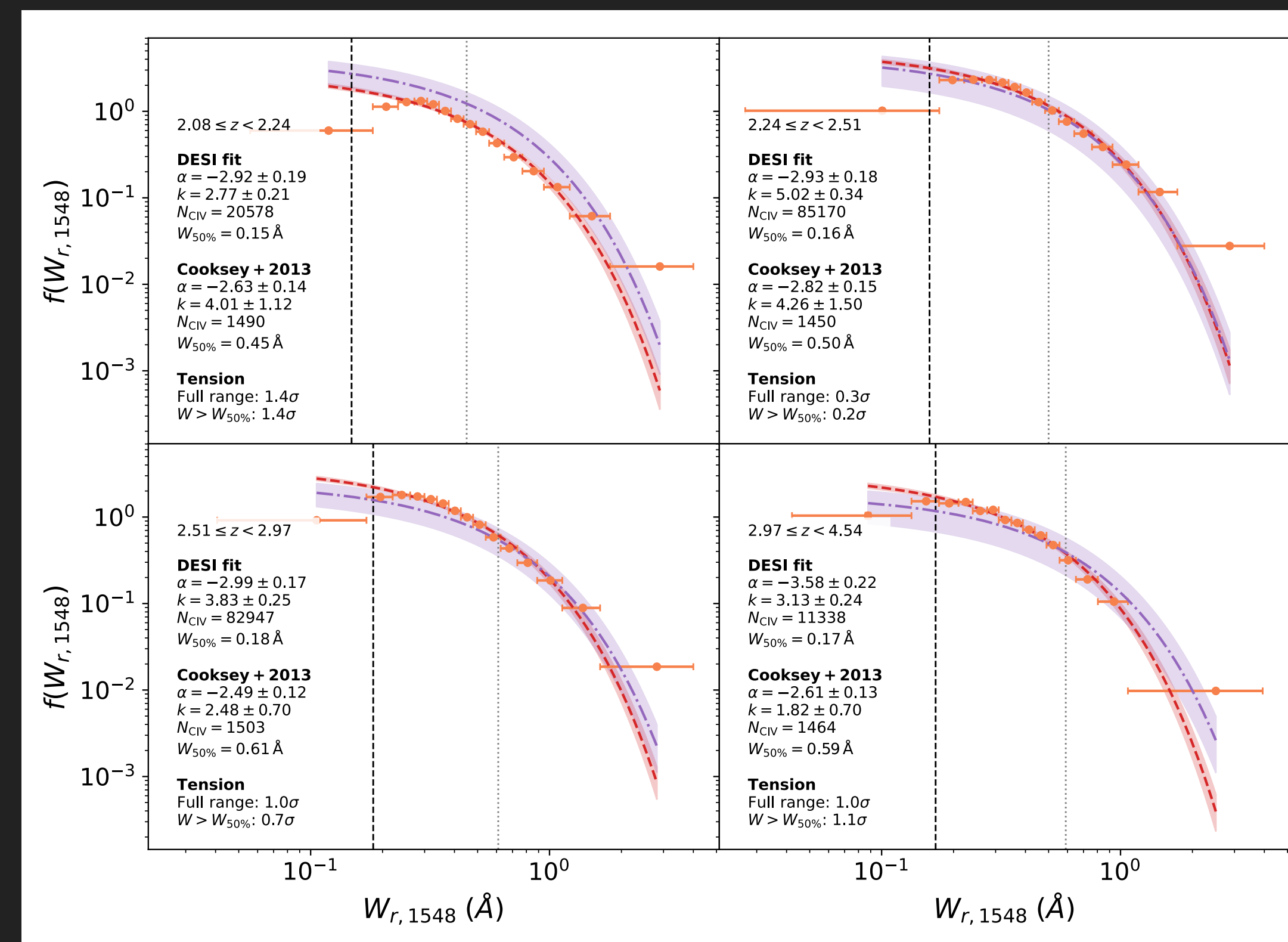
- ▶ Total of 17 million absorptions detected
 - ▶ 5M+ at $S/N > 4$
- ▶ 200,000+ CIV 1548 Å with $S/N > 1.5$
- ▶ Test on computing the rest-frame equivalent width distribution function of CIV absorbers
 - ▶ Tension with [Anand et al. 2025](#) $> 2\sigma$, up to 4σ , both on DESI data



Comparison with the literature fit by [Anand et al. 2025](#)

PILOT TESTING: DESI DATA (DR1)

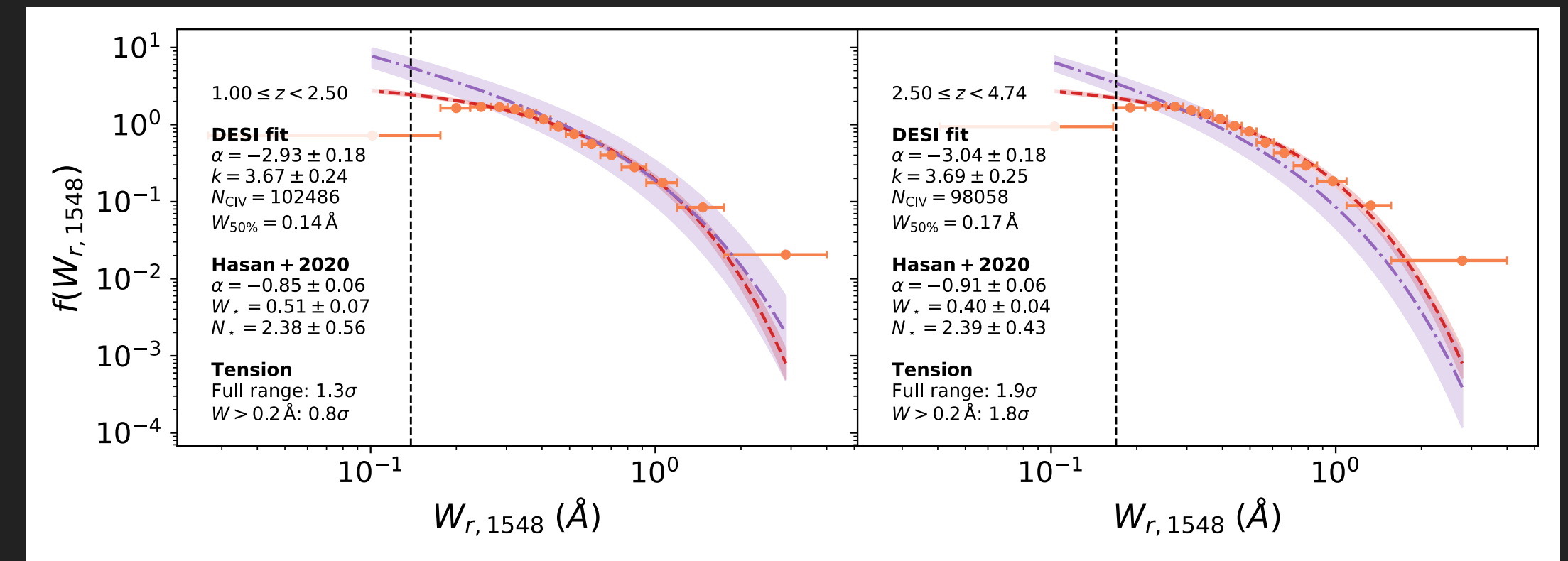
- ▶ Total of 17 million absorptions detected
 - ▶ 5M+ at $S/N > 4$
- ▶ 200,000+ CIV 1548 Å with $S/N > 1.5$
- ▶ Test on computing the rest-frame equivalent width distribution function of CIV absorbers
 - ▶ Tension with [Anand et al. 2025](#) $> 2\sigma$, up to 4σ , both on DESI data
 - ▶ Tension with [Cooksey et al. 2013](#) $\sim 1\sigma$, literature on SDSS data



Comparison with the literature fit by [Cooksey et al. 2013](#)

PILOT TESTING: DESI DATA (DR1)

- ▶ Total of 17 million absorptions detected
 - ▶ 5M+ at $S/N > 4$
- ▶ 200,000+ CIV 1548 Å with $S/N > 1.5$
- ▶ Test on computing the rest-frame equivalent width distribution function of CIV absorbers
 - ▶ Tension with [Anand et al. 2025](#) $> 2\sigma$, up to 4σ , both on DESI data
 - ▶ Tension with [Cooksey et al. 2013](#) $\sim 1\sigma$, literature on SDSS data
 - ▶ Tension with [Hasan et al. 2020](#) $< 2\sigma$, high-resolution spectra from Keck and VLT observations



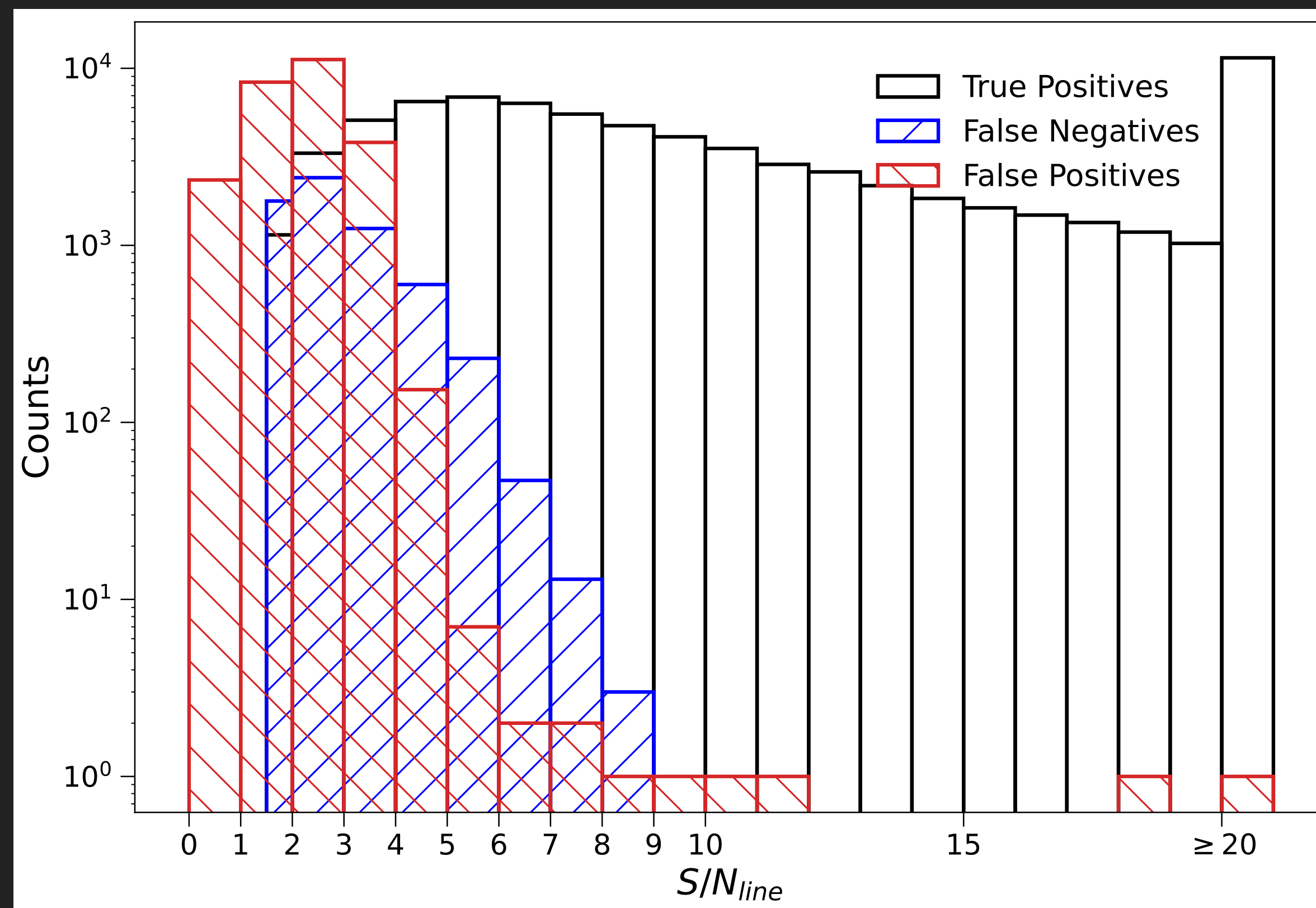
Comparison with the literature fit by [Hasan et al. 2020](#)

SUMMARY

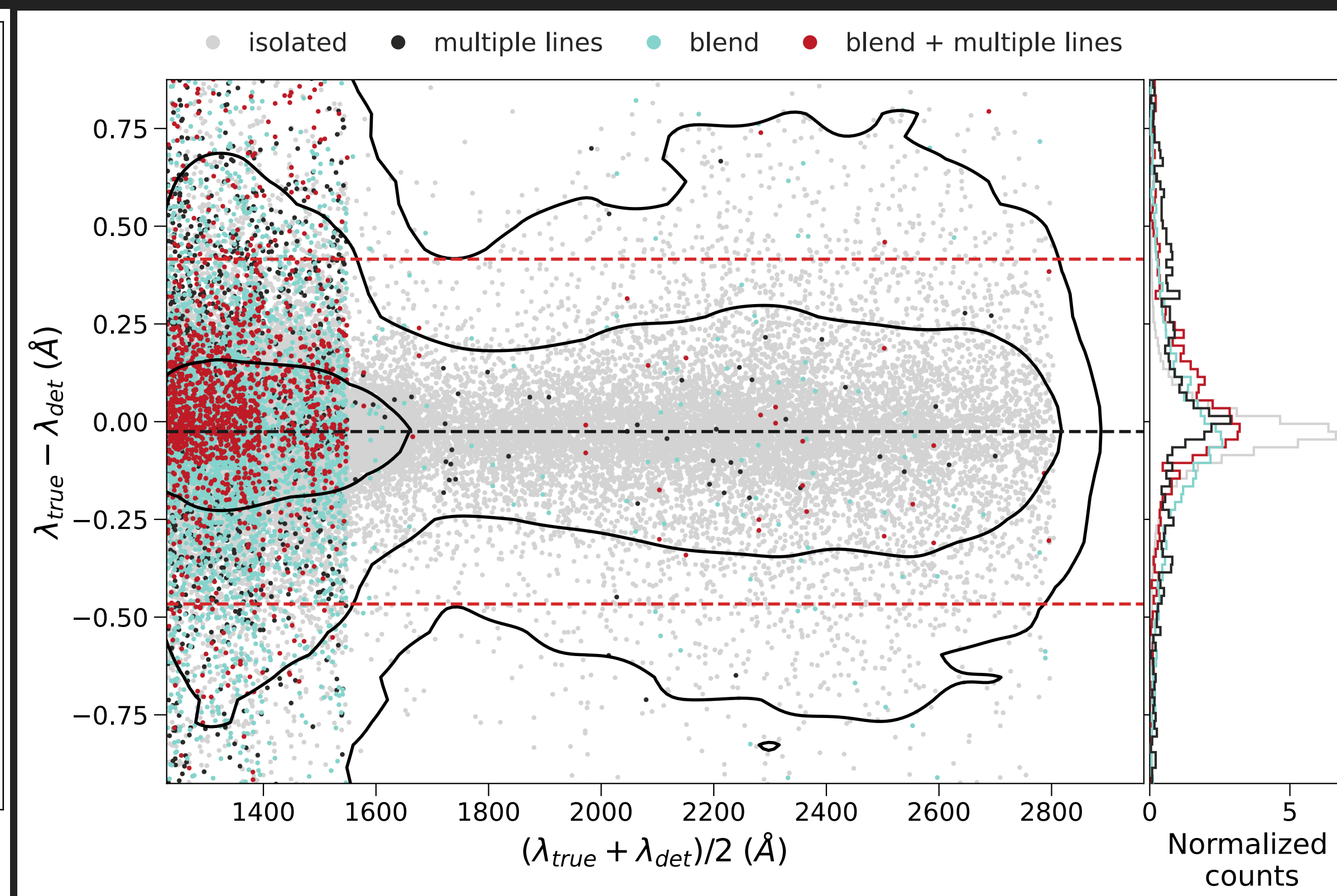
- ▶ ~1% error on continuum recovery within 1s for 10,000 spectra
- ▶ Detection of absorption lines with ~90% completeness and purity at S/N~4 of the lines
- ▶ Classification with ~90% completeness and purity for main metal lines (CIV, SiIV, MgII, FeII)
 - ▶ Recovery of the rest-frame equivalent width frequency distribution
- ▶ Main performance bottleneck due to absorption line detection

**THANK YOU FOR
YOUR ATTENTION!**

LINE DETECTION

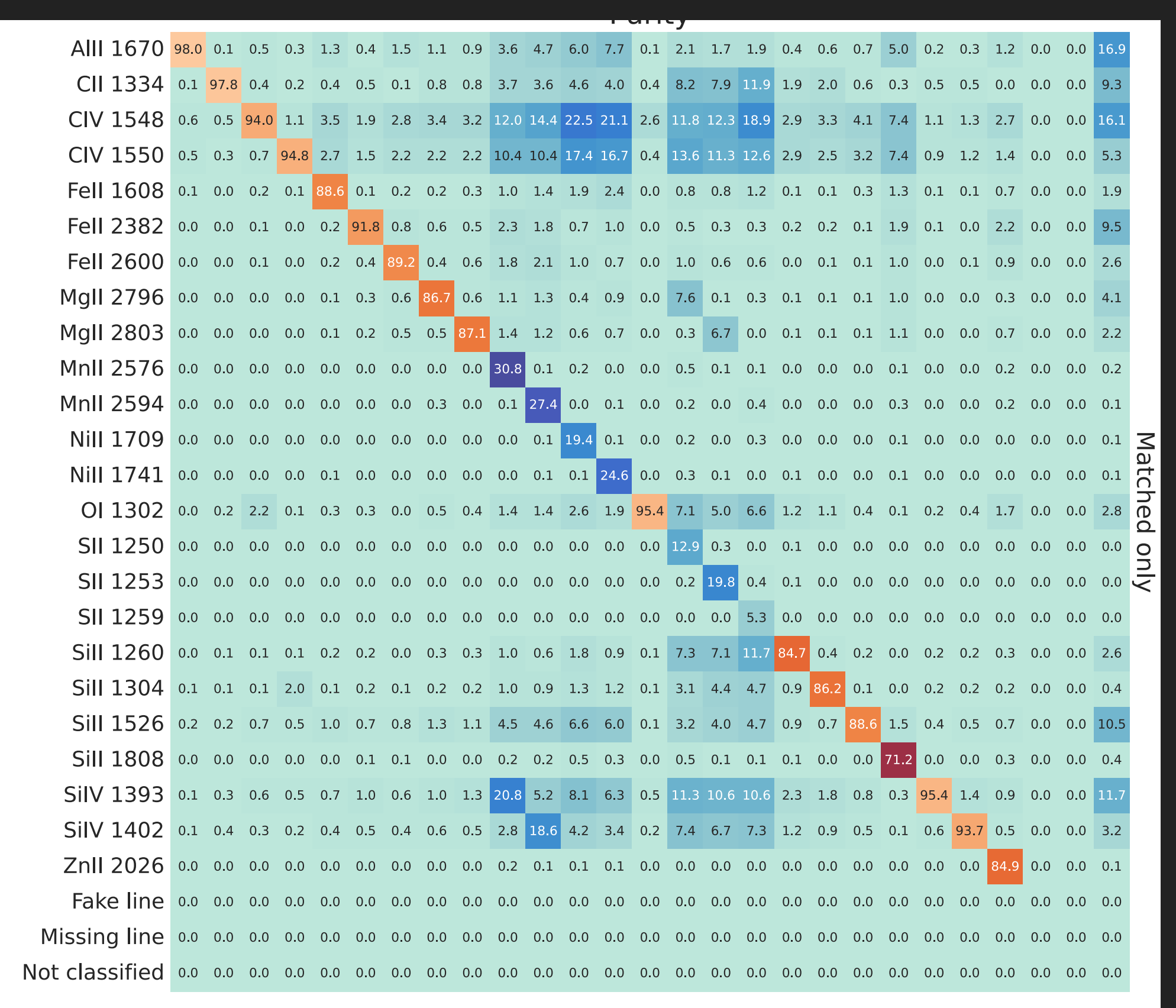
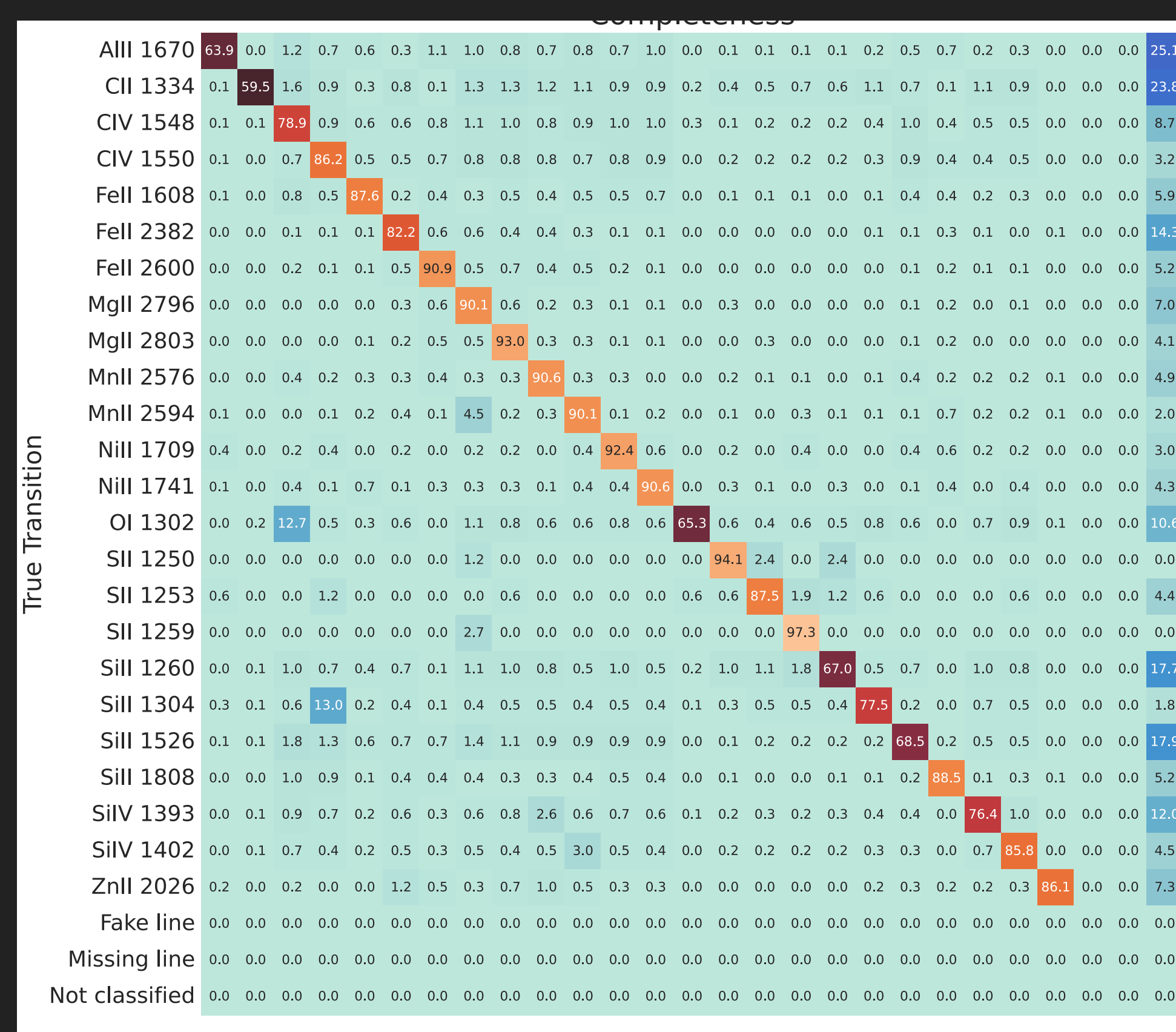


Histogram of true positives, false negatives, and false positives



Analysis of the precision for the centroid measurement

LINE CLASSIFICATION



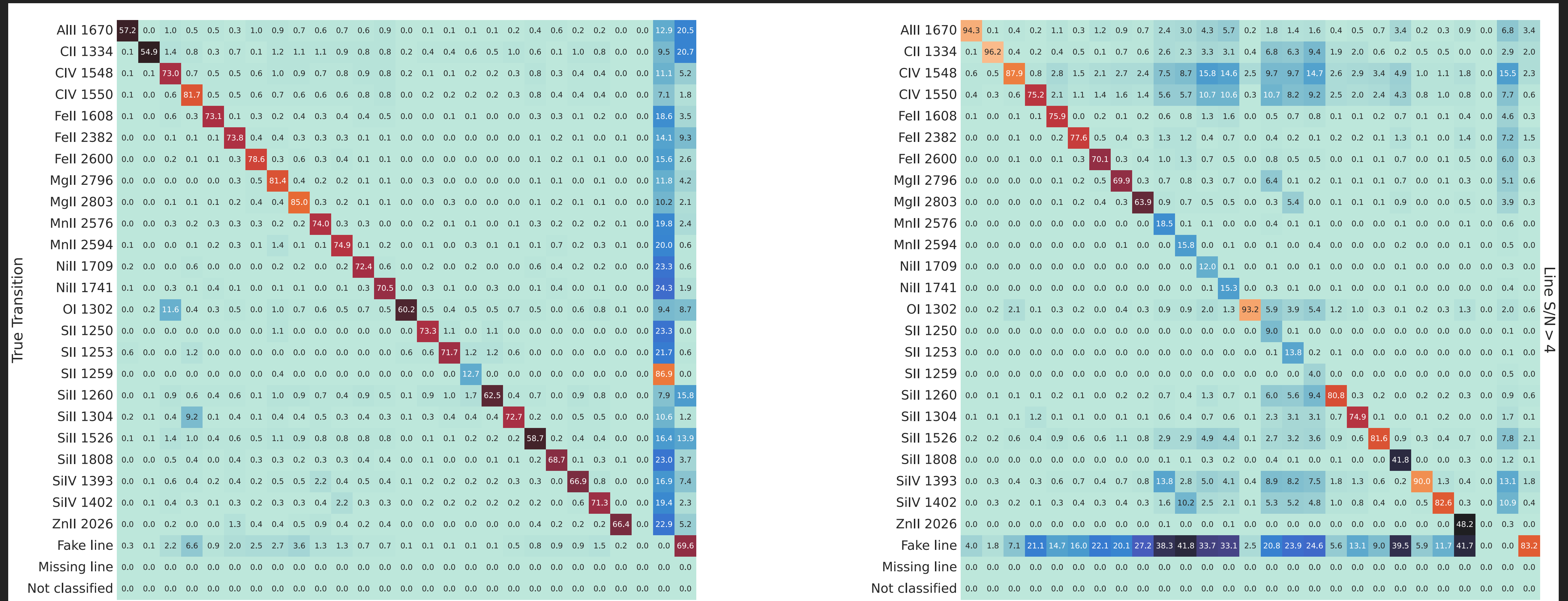
Confusion matrix for only true detections

True Transition



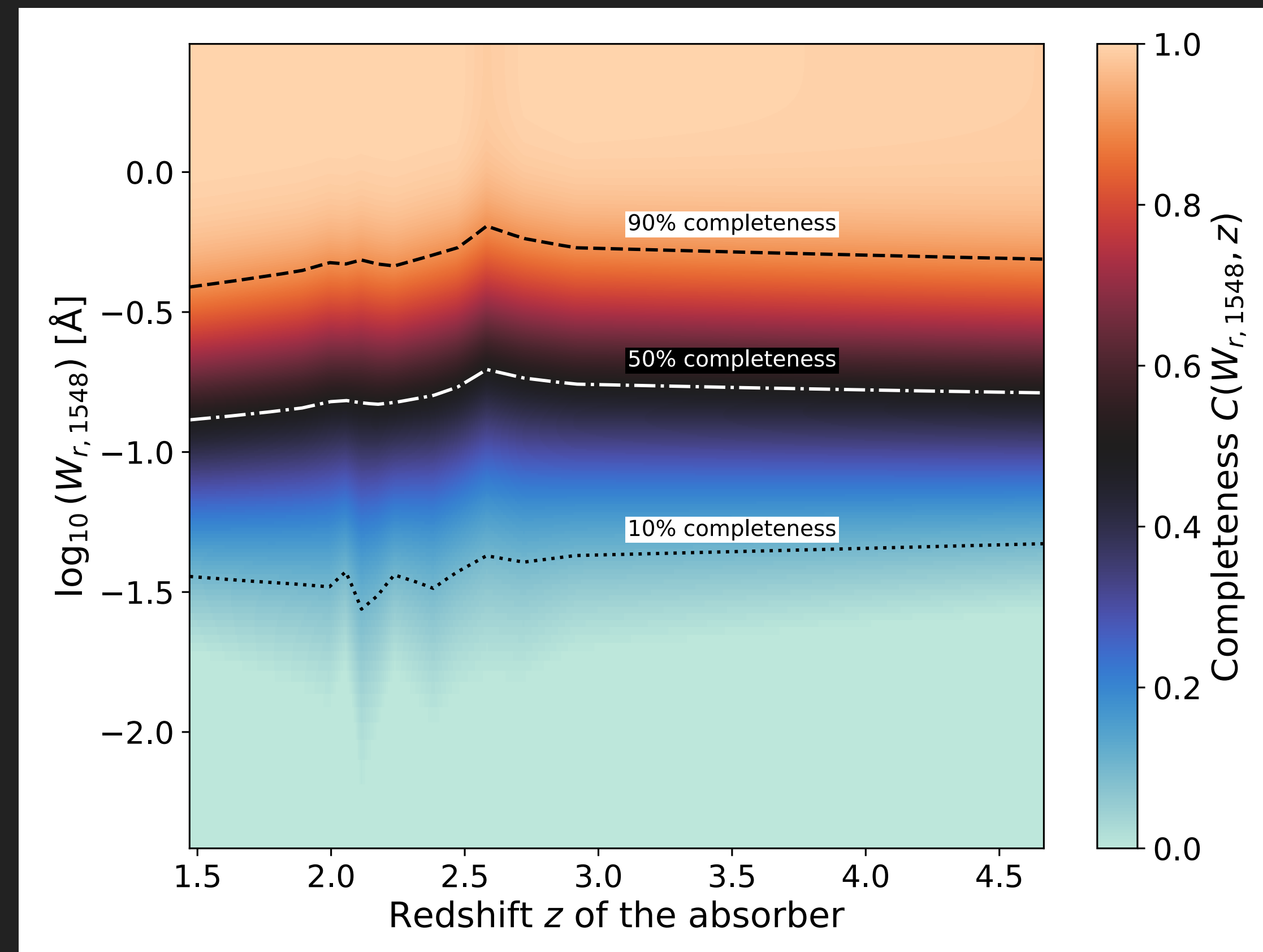
Confusion matrix for all lines

LINE CLASSIFICATION



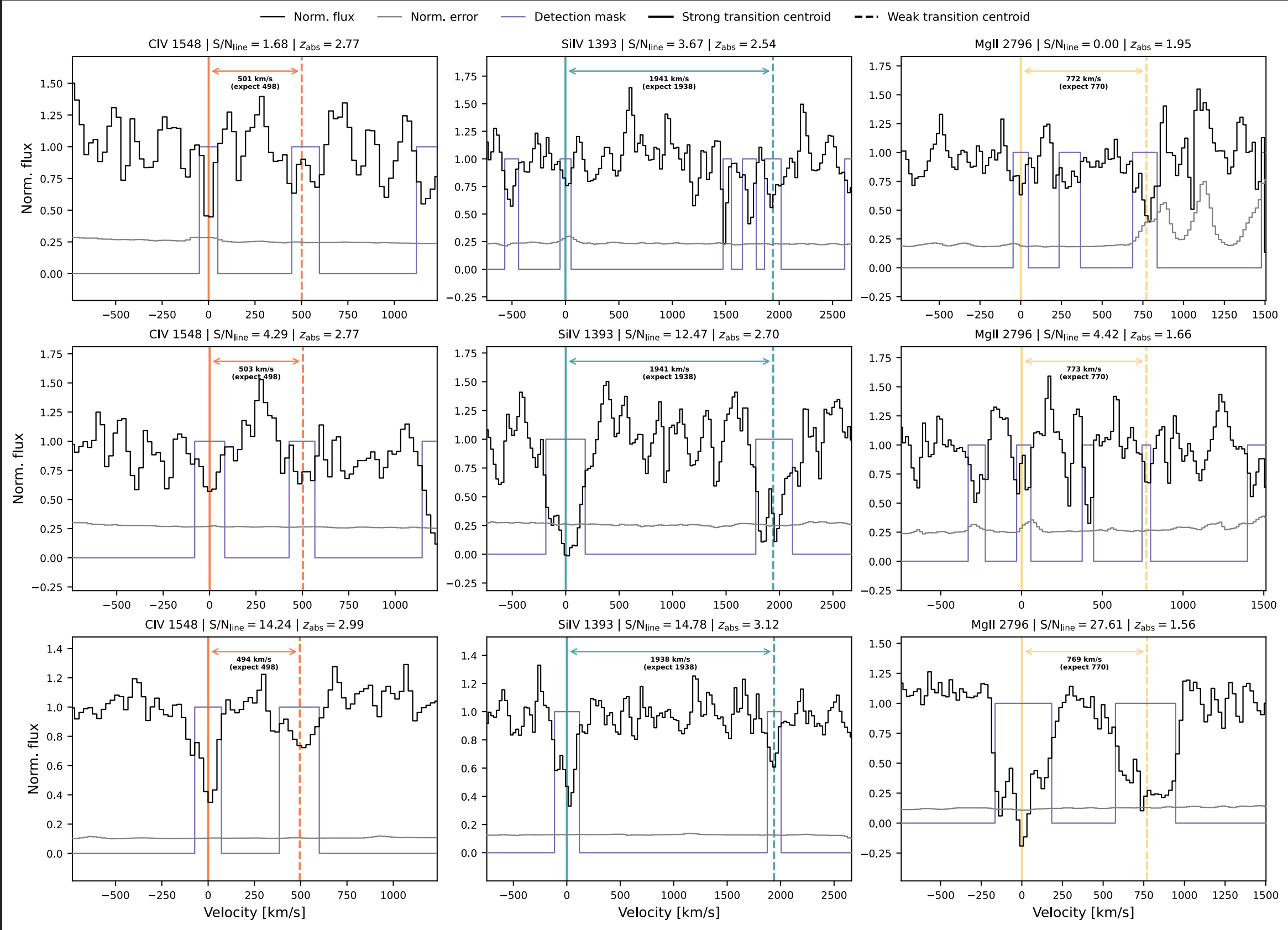
Confusion matrix for all lines with S/N>4

LINE CLASSIFICATION



Completeness in equivalent width - redshift bins

LINE CLASSIFICATION



Classification examples at different S/N