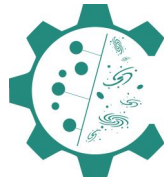


Normalizing Every Stellar Spectrum

Tomasz Różański
04 September 2026



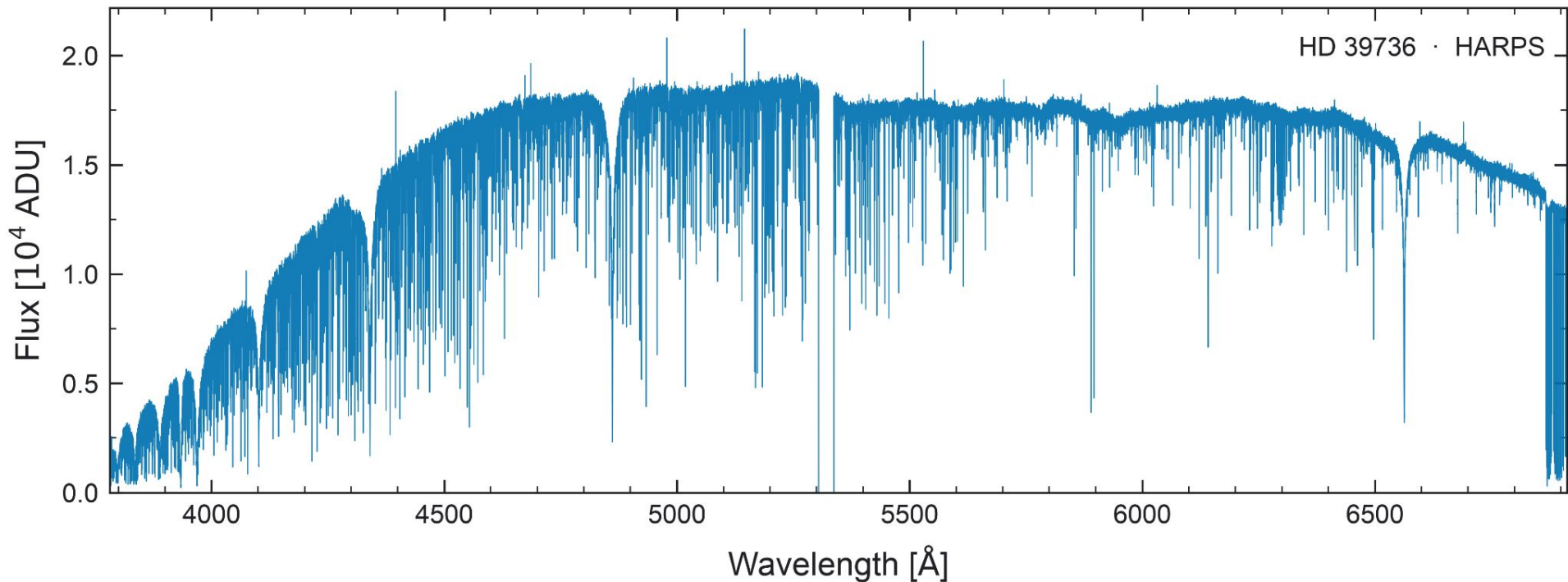
**MACHINE LEARNING
FOR ASTROPHYSICS**
3RD EDITION VALLETTA, 31 AUG - 4 SEP, 2026



Australian
National
University

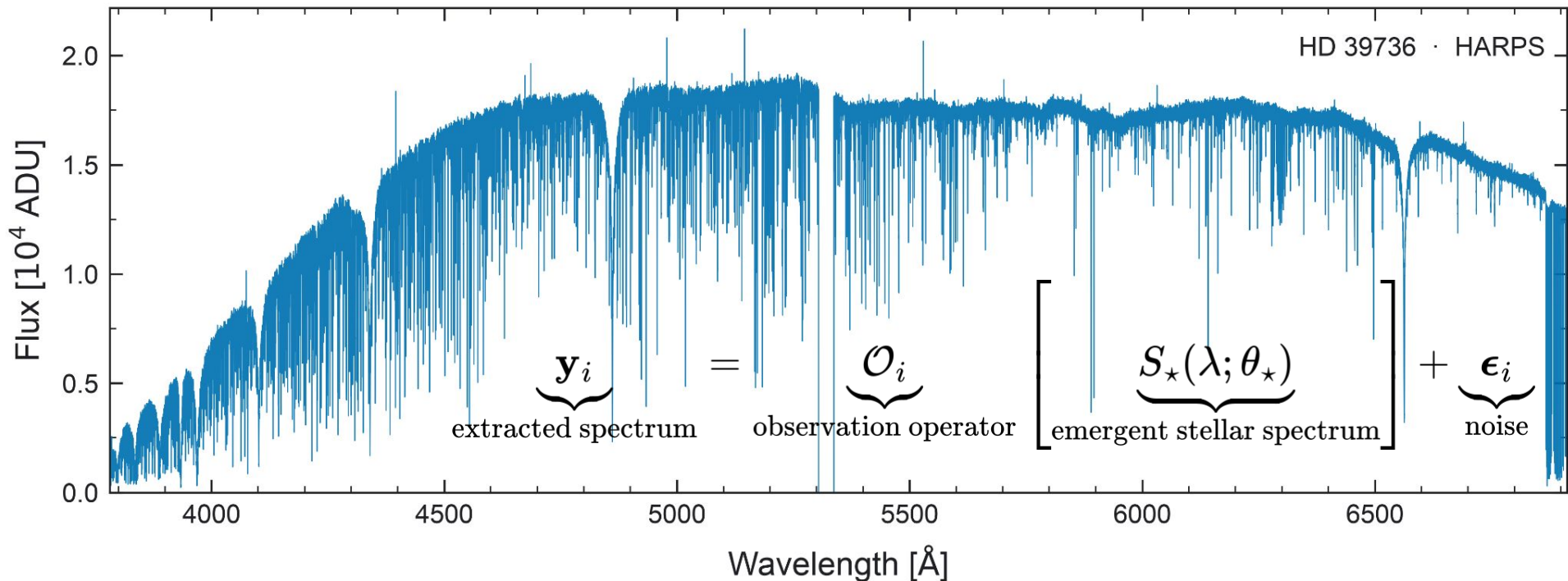
A spectrum is a compressed physical experiment

Extracted 1D stellar spectrum



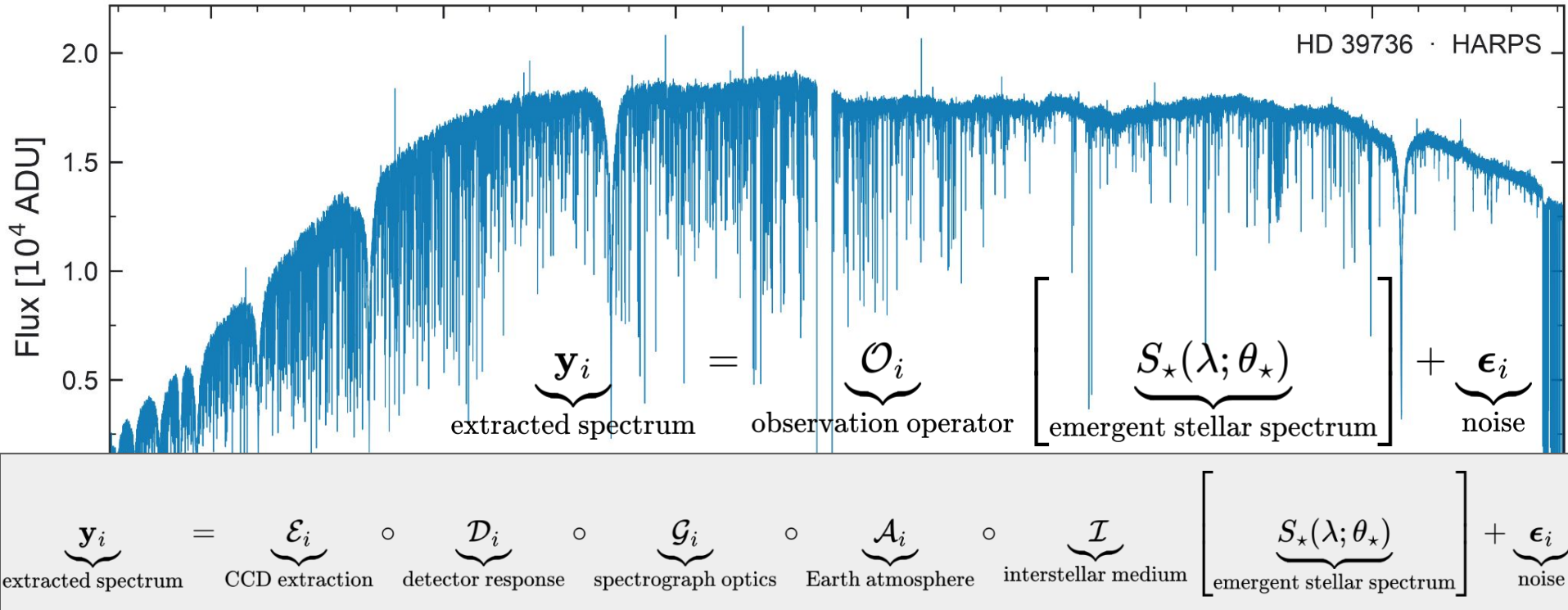
A spectrum is a compressed physical experiment

Extracted 1D stellar spectrum

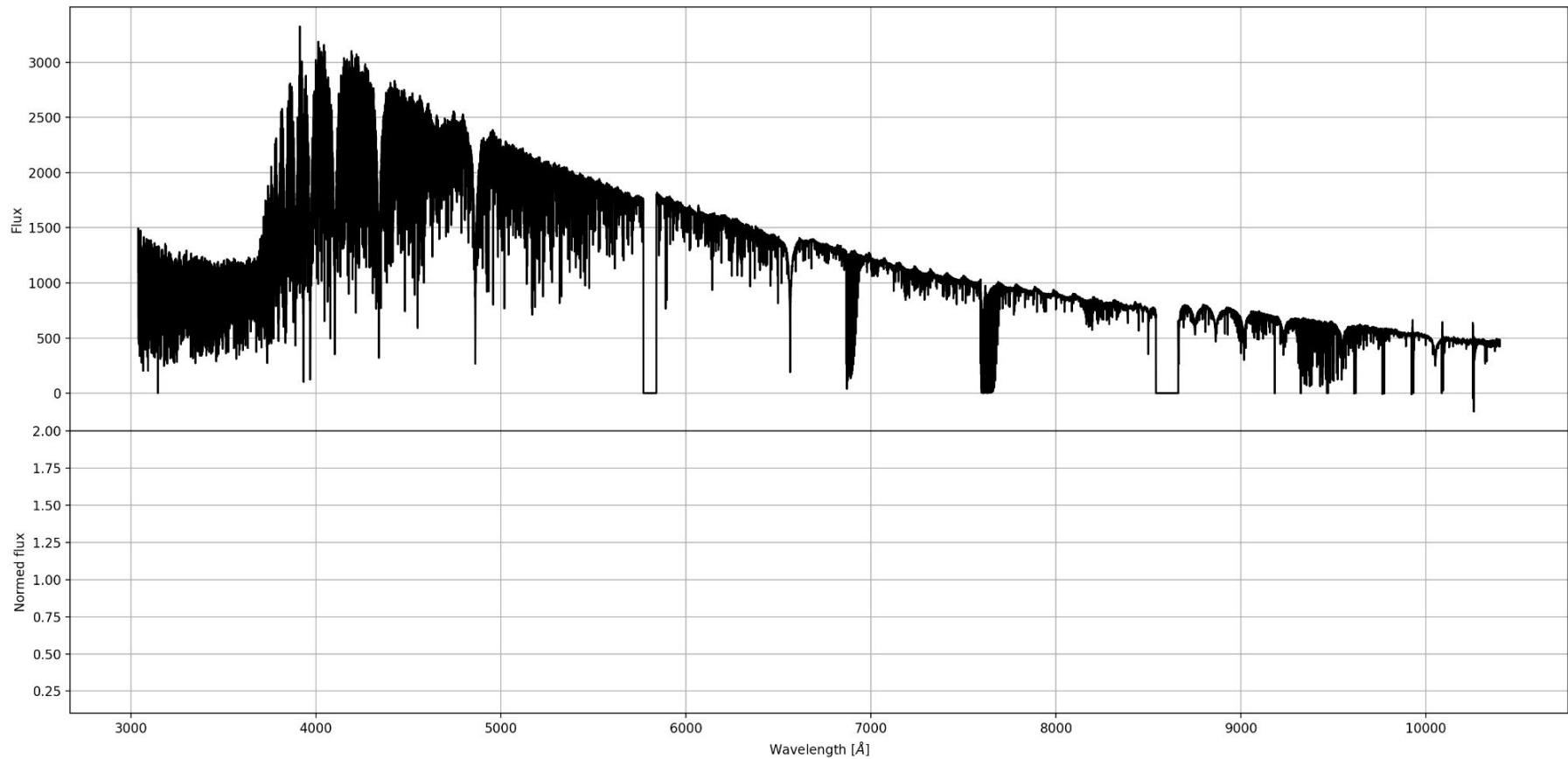


A spectrum is a compressed physical experiment

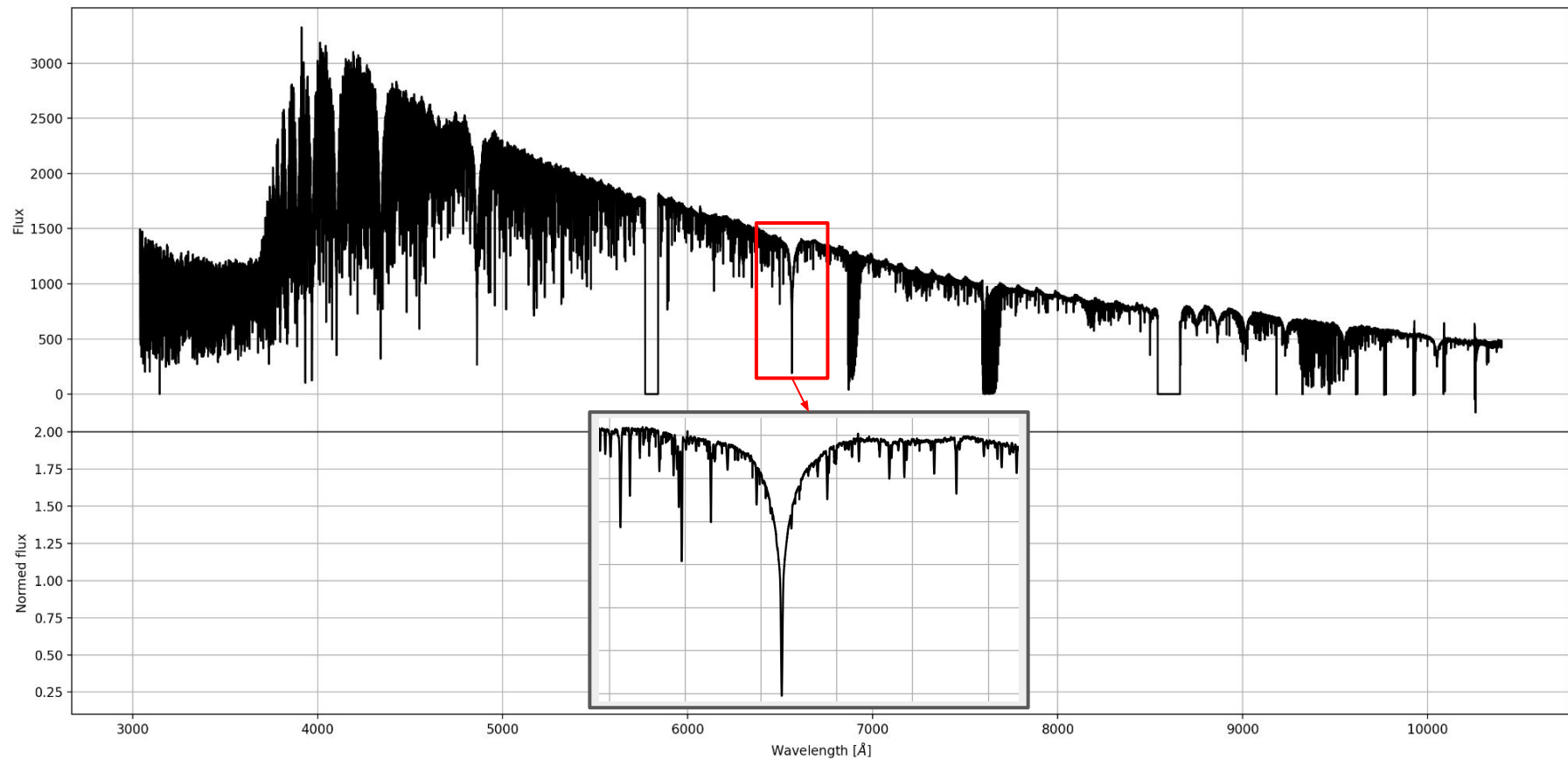
Extracted 1D stellar spectrum



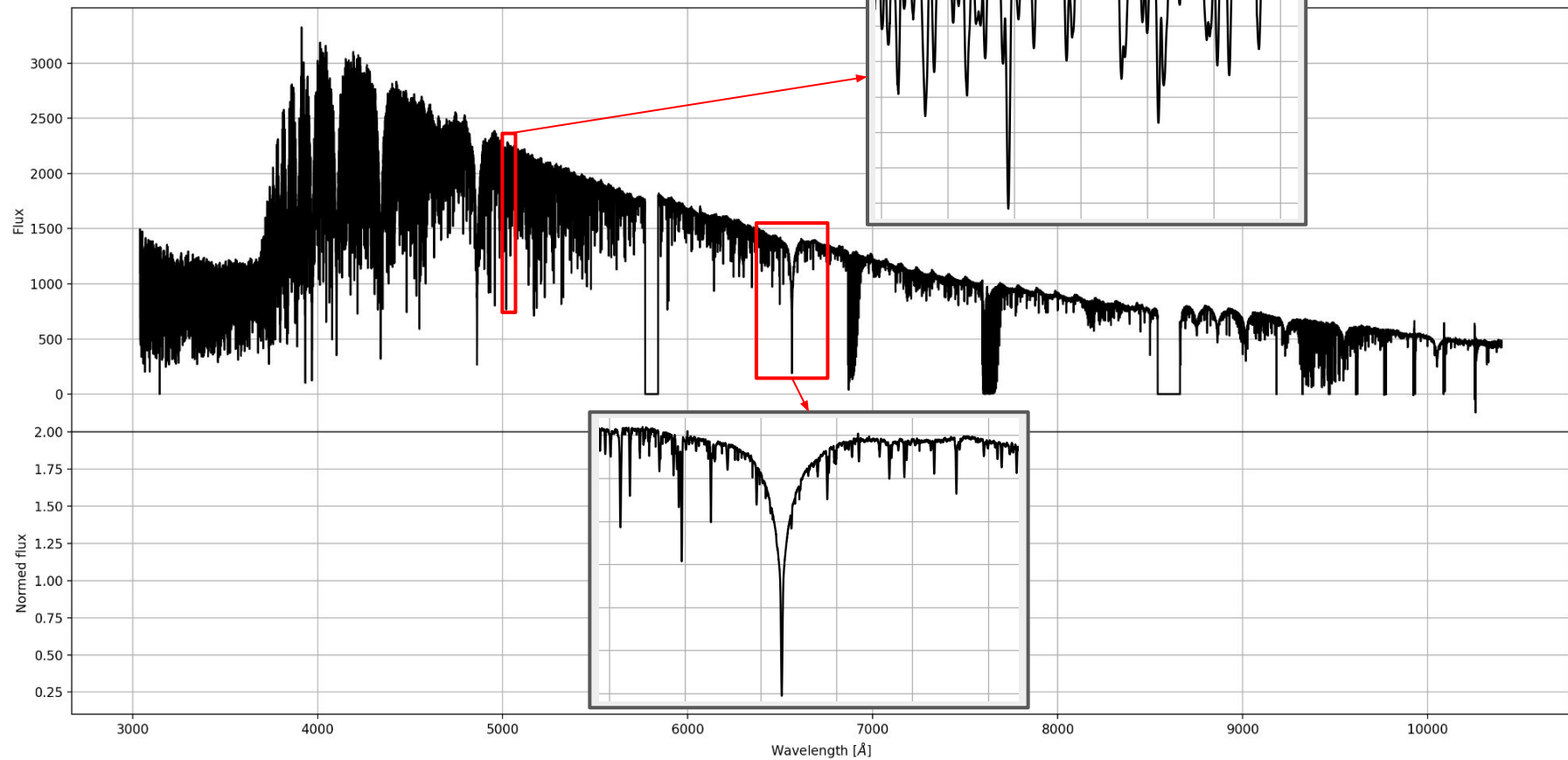
Normalization of stellar spectra



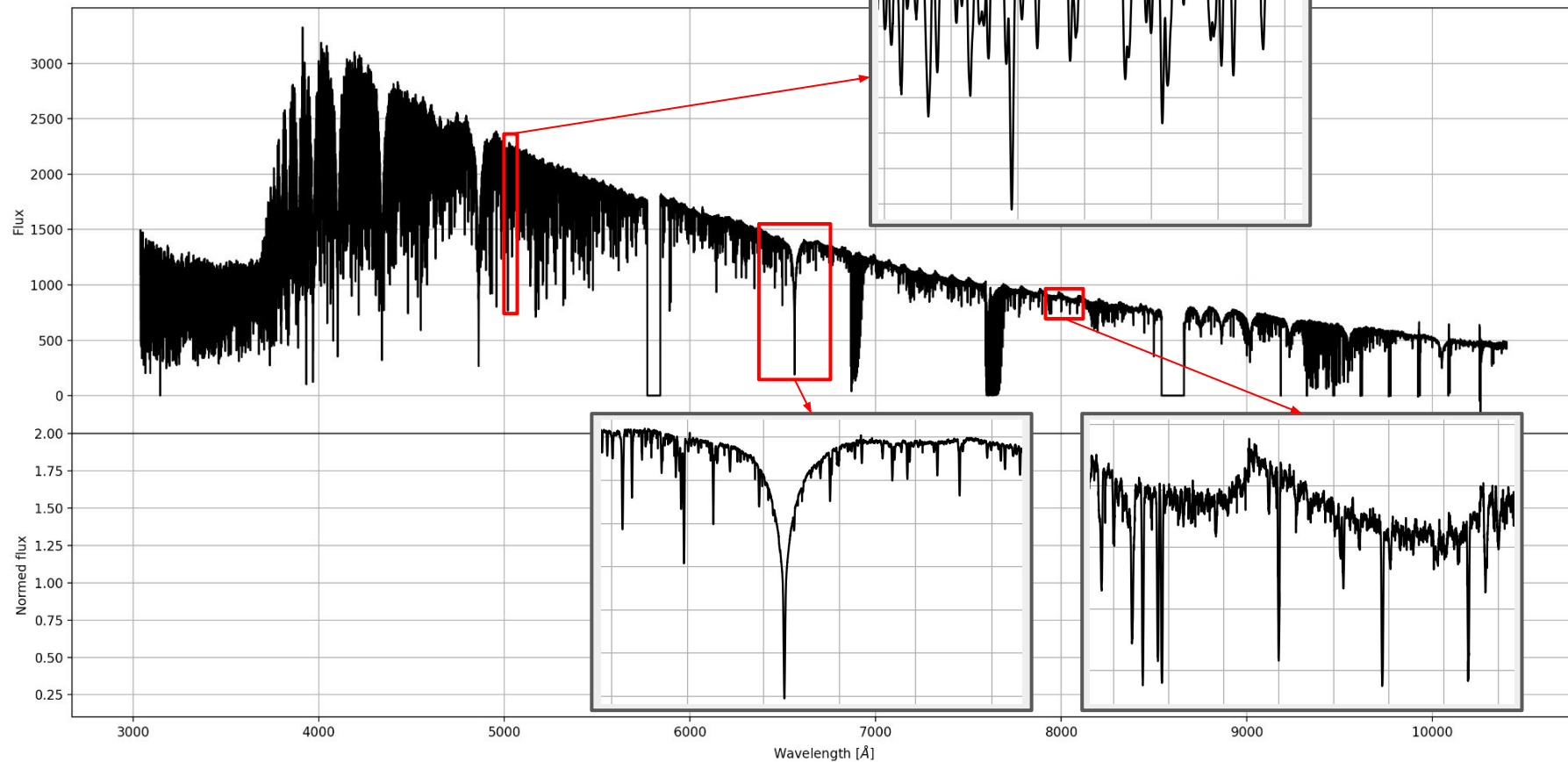
Normalization of stellar spectra



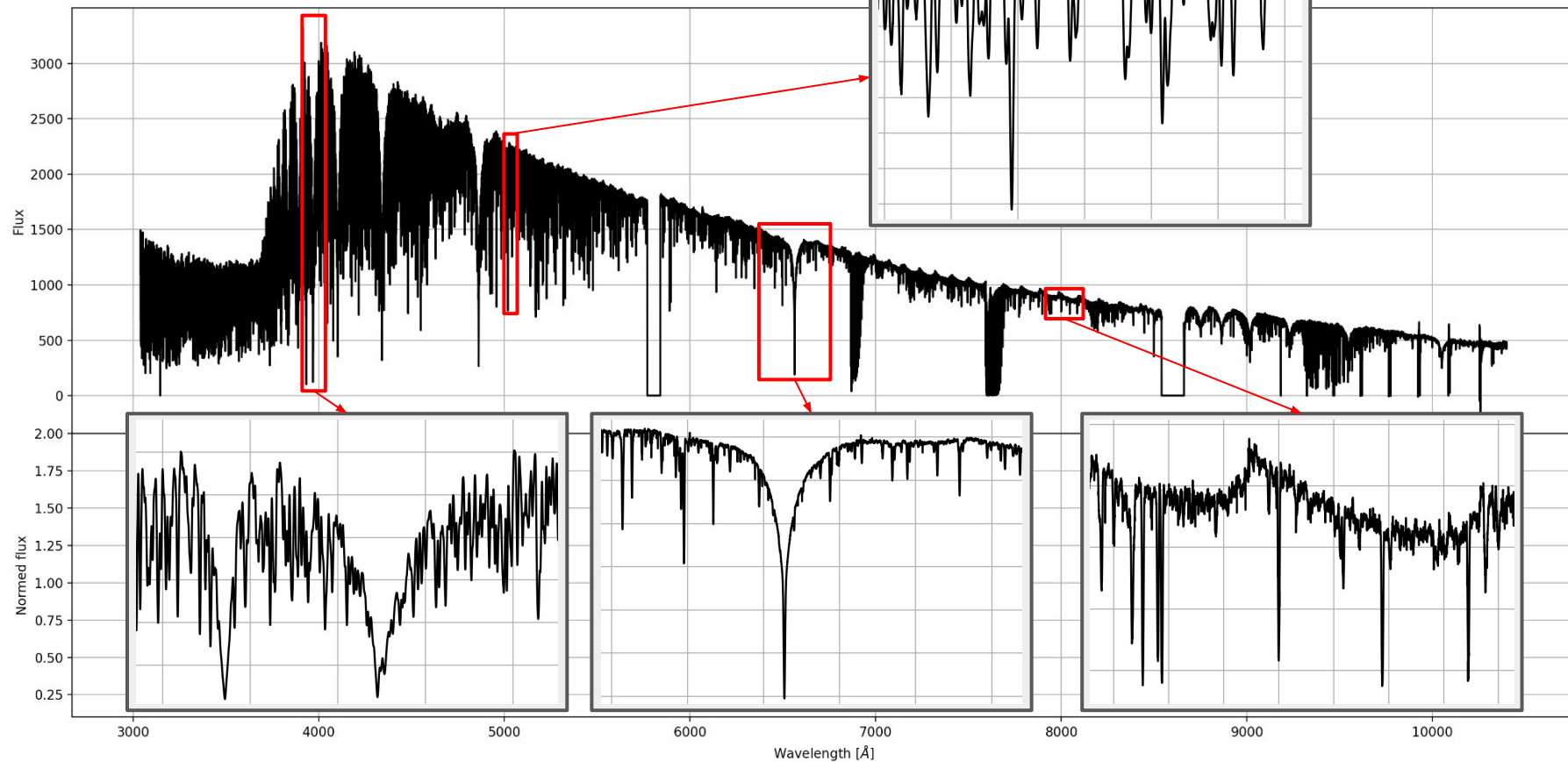
Normalization of stellar spectra



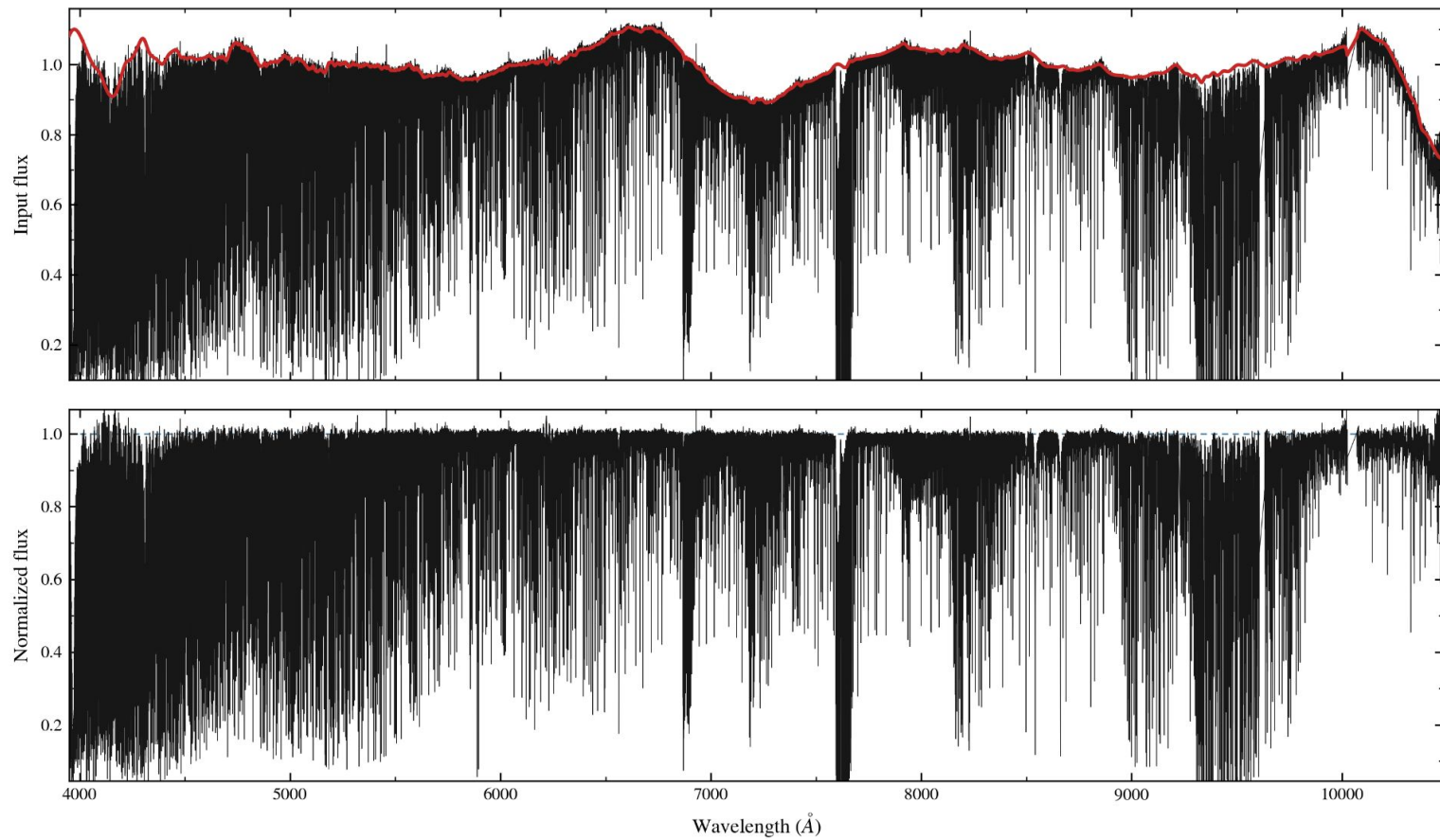
Normalization of stellar spectra



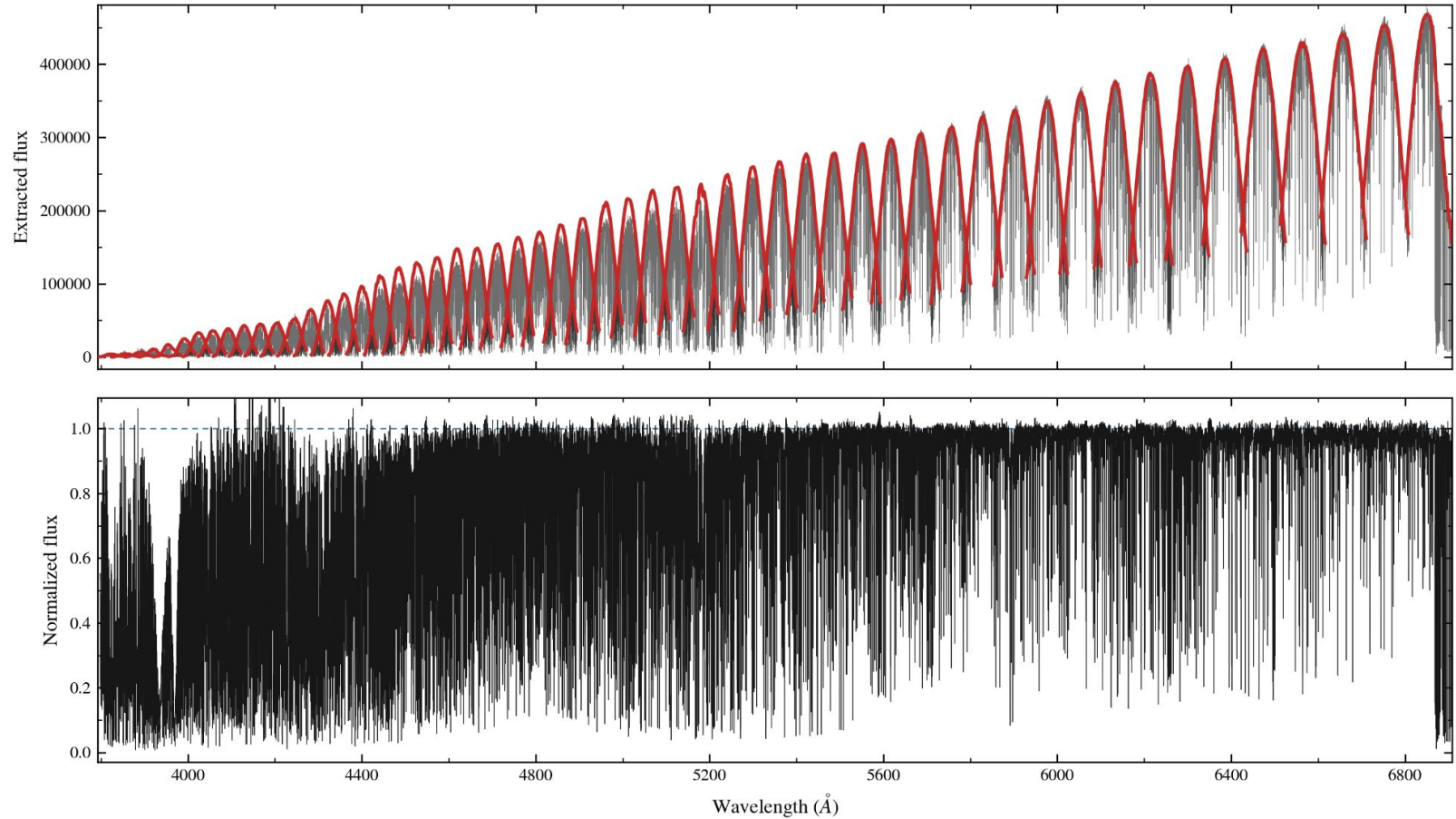
Normalization of stellar spectra



Normalization of stellar spectra (I)



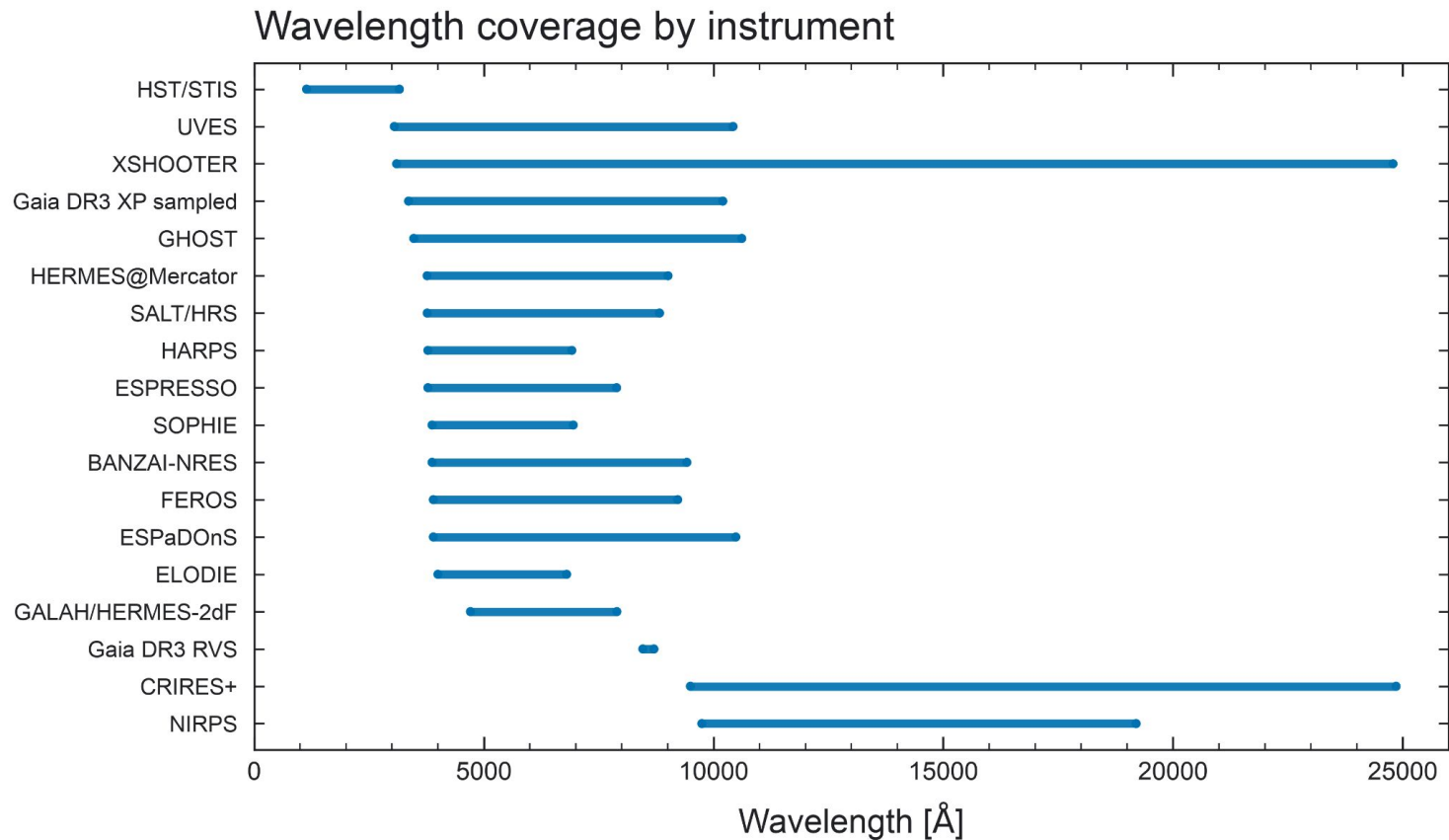
Normalization of stellar spectra (II)



Broad spectral dataset

Instrument / dataset	Spectra represented	Instrument / dataset	Spectra represented	Instrument / dataset	Spectra represented
Gaia DR3 XP	219,197,643	GALAH DR4	~1,085,520	Gaia DR3 RVS	999,645
IUE	67,063	UVES	49,636	FEROS	33,881
HST/STIS	29,681	HERMES@Mercator	24,900	HARPS	18,666
ESPaDOnS	15,350	SOPHIE	14,713	ELODIE	11,151
XSHOOTER	9,173	FUSE	5,729	BANZAI-NRES	4,535
CRIRES+	3,649	ESPRESSO	2,727	GHOST	1,920
NIRPS	1,565	StarCAT	612		
SALT/HRS	456	ASTRAL	56		

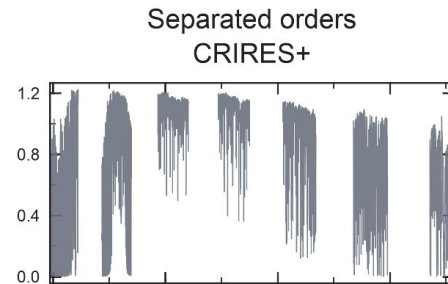
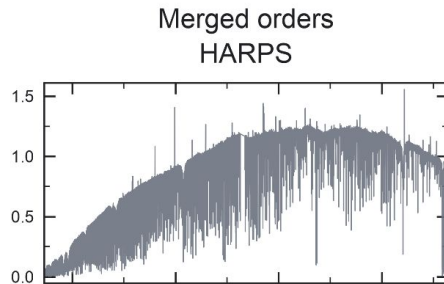
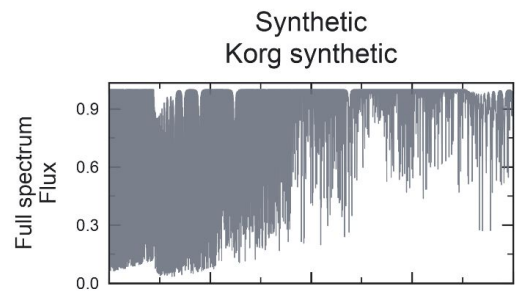
Broad spectral dataset



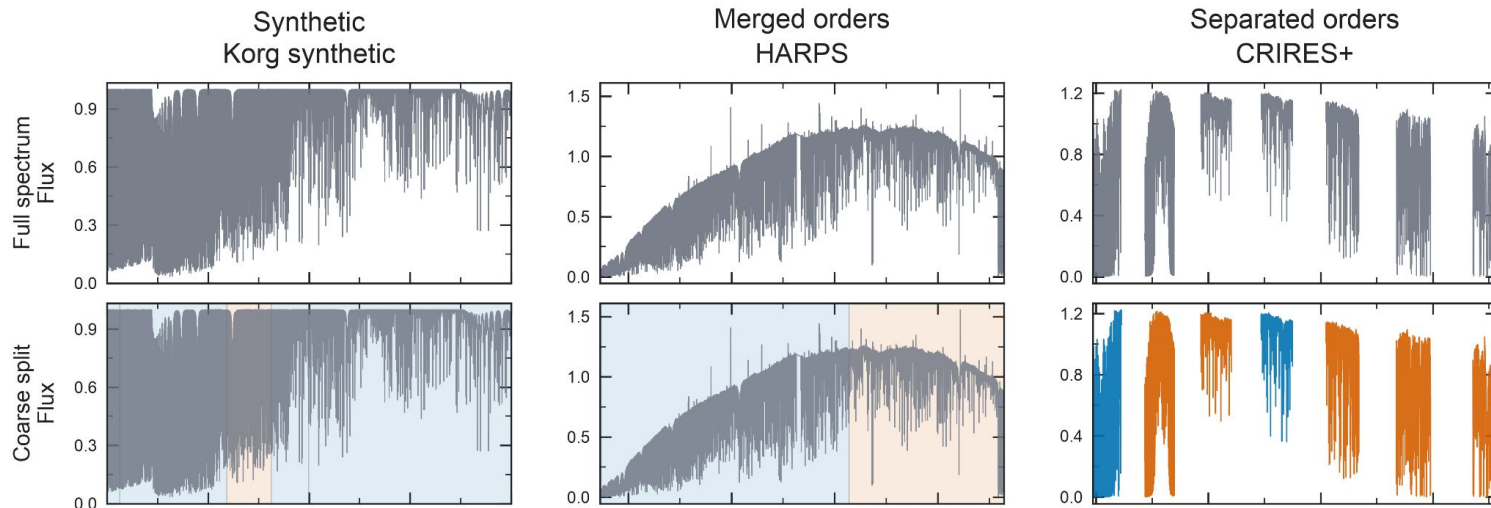
Training base model

- Transformer-based regression model.
- Each token represents:
 - Wavelength, flux, and flux uncertainty,
 - Order index and the task associated with the given token (e.g. condition token, prediction token, ...),
 - Special tokens, such as instrument tokens.
- Synthetic and observed spectra are used together during training.
- Heavy augmentation of synthetic spectra:
 - Radial velocity shifts, resolution changes, cosmic rays, and noise,
 - A model of imperfections introduced by the echelle reduction pipeline (“wiggles”),
 - A simple model of the echelle spectrograph response.
- Masked autoencoder training.

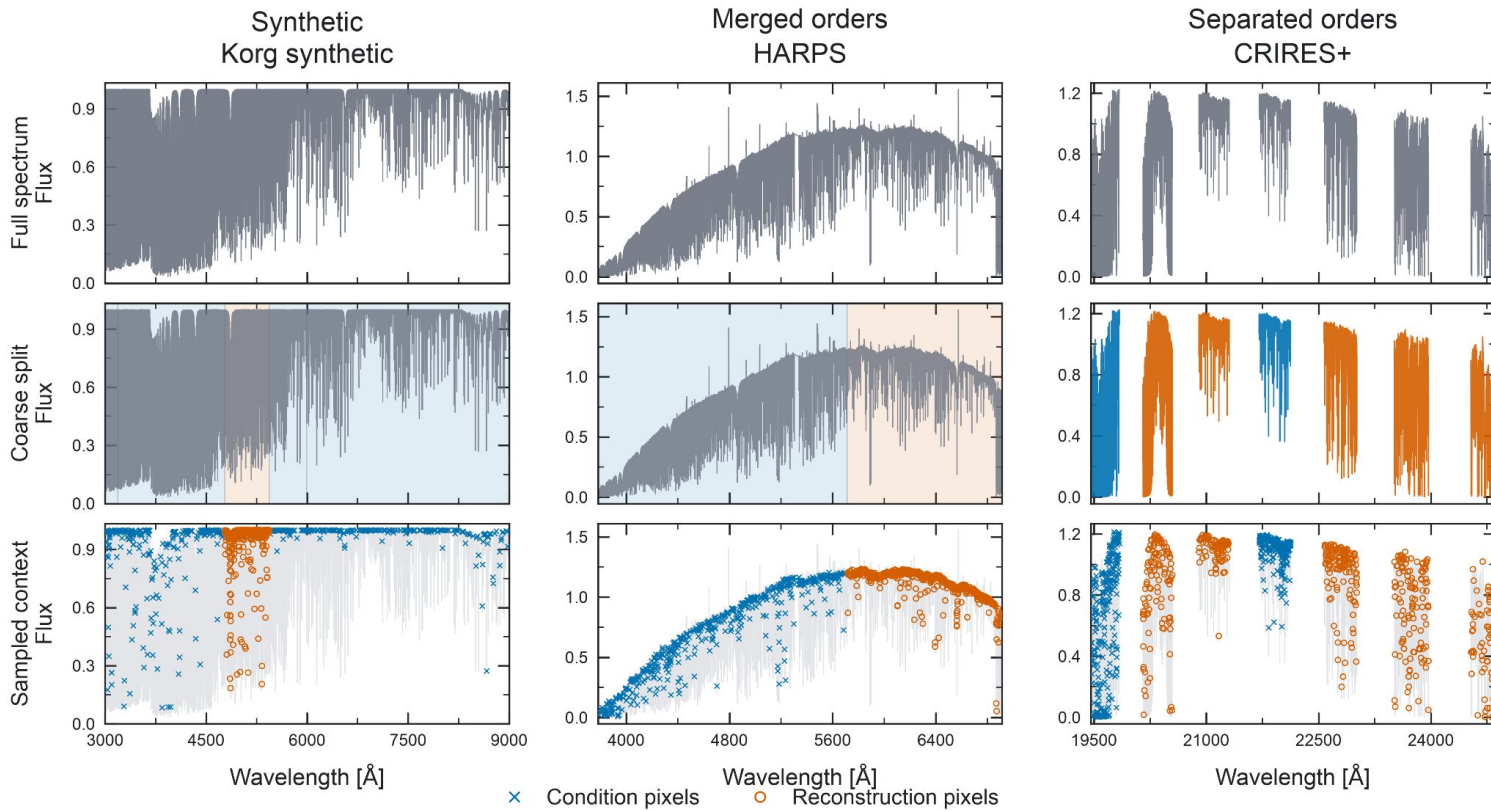
Masked autoencoder training across dataset



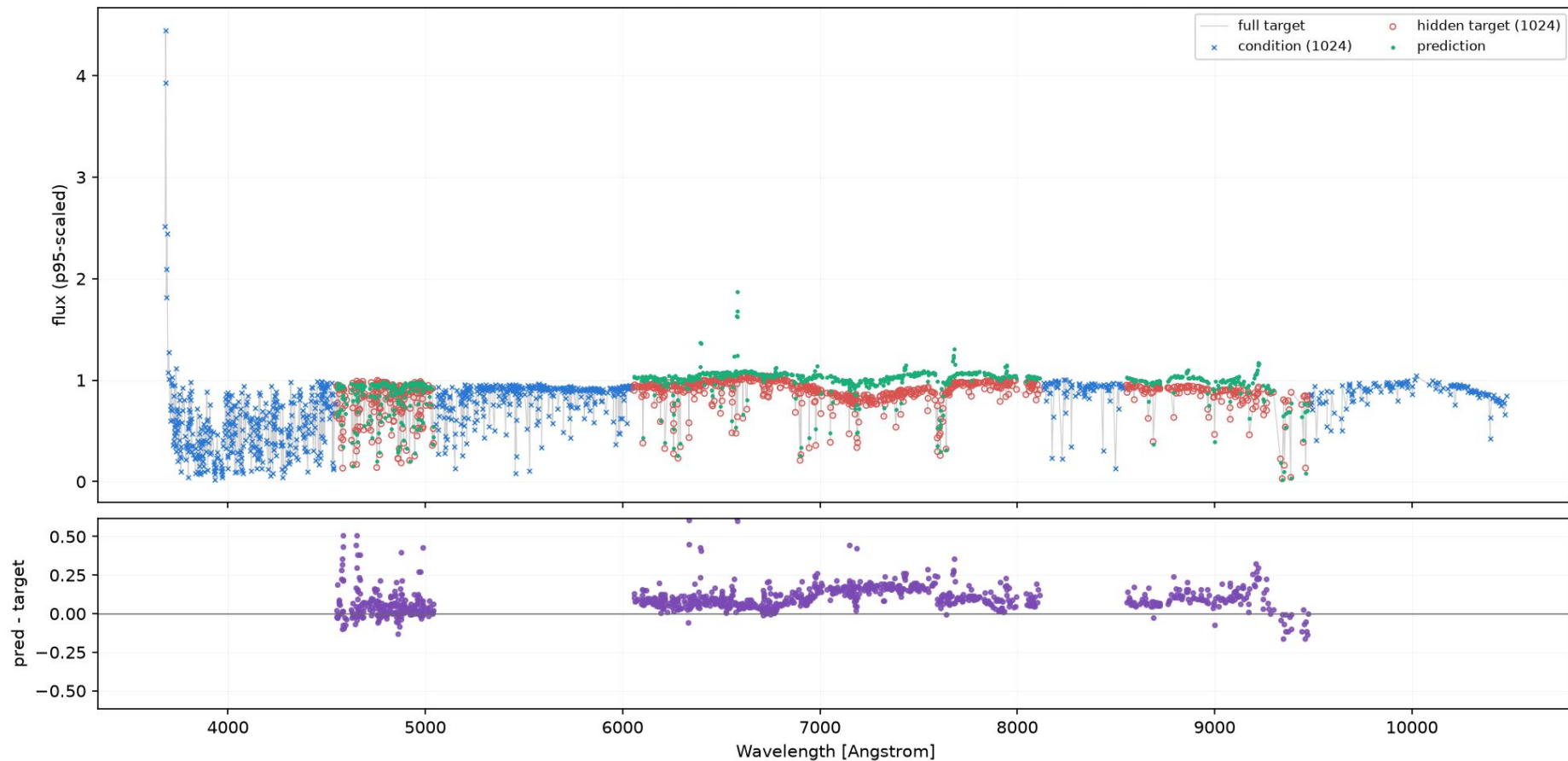
Masked autoencoder training across dataset



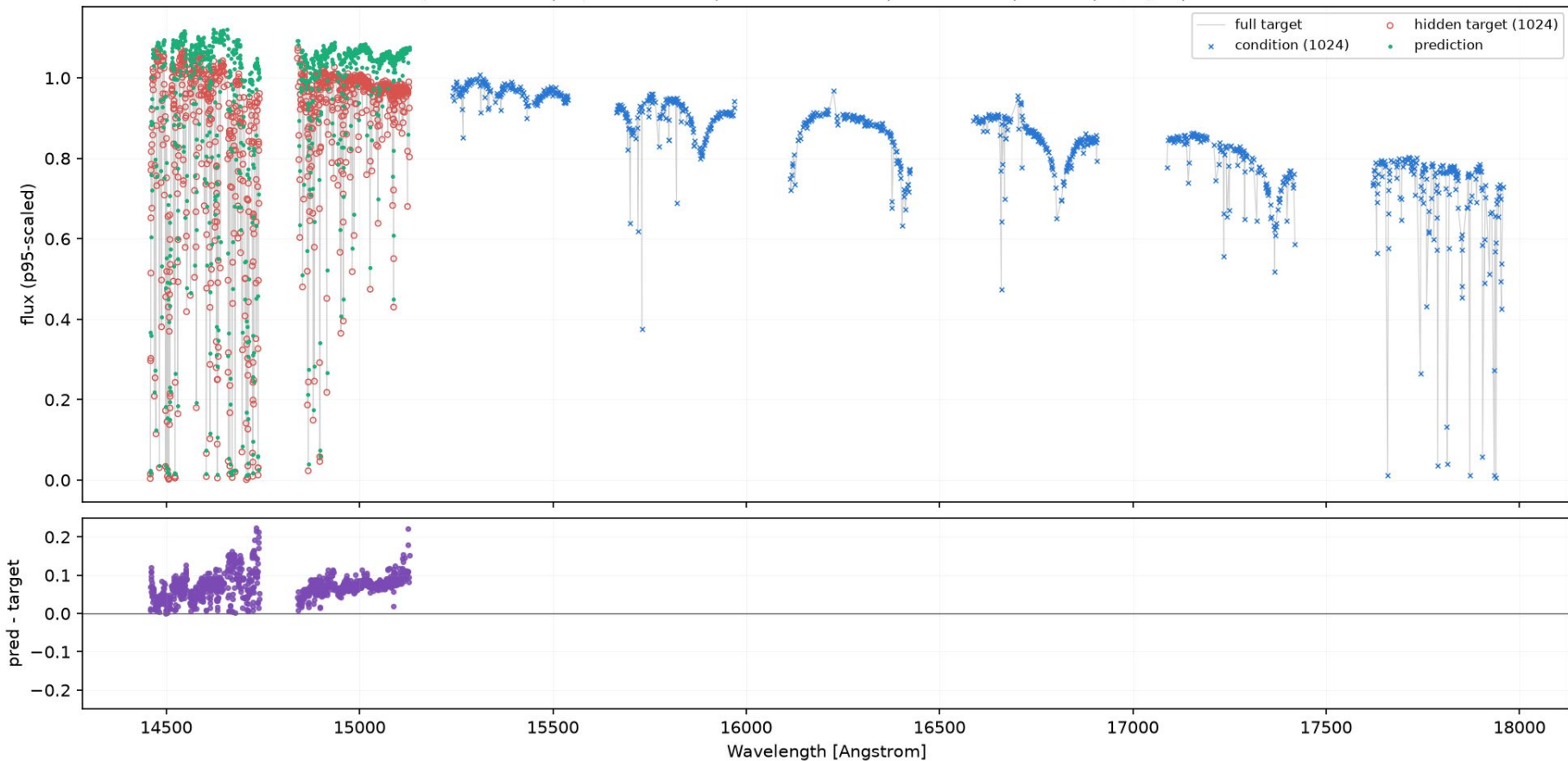
Masked autoencoder training across dataset



Example prediction in validation set

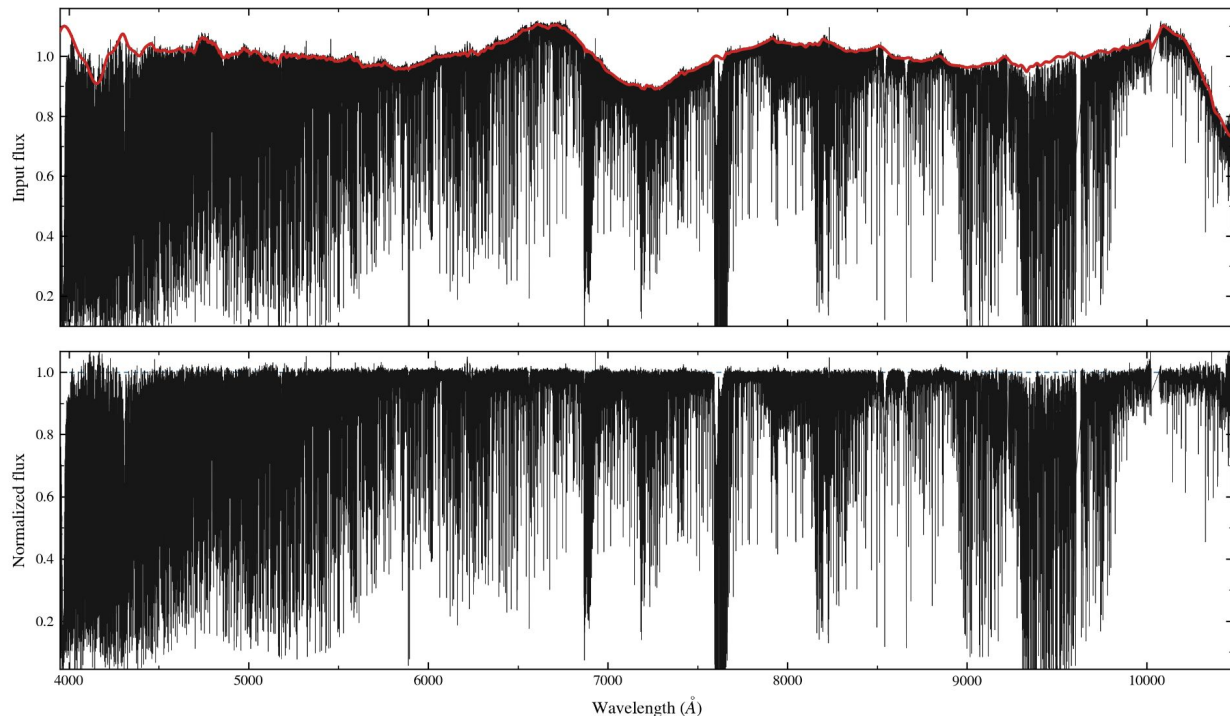


Example prediction in validation set



Fine-tuning for normalization task

- Adding a new “task token”: a normalization token.
- Fine-tuning the pretrained model using synthetic spectra only, with known continua.



Summary

- Broad, diverse multi-instrument dataset.
- One base model across spectral resolutions.
- First downstream task: spectrum normalization.
- Next: run ablations and release the tool.

Please reach out to me at Tomasz.Rozanski1@anu.edu.au if you'd like to include your dataset in the training, or if you have an instrument in mind that I should consider.