# Contrastive Learning for White Dwarf Classification with Gaia

***Xabier Pérez-Couto***
*Universidade da Coruña/CITIC*

***Supervised by:***

*Minia Manteiga*
*Eva Villaver*

***In collaboration with:***

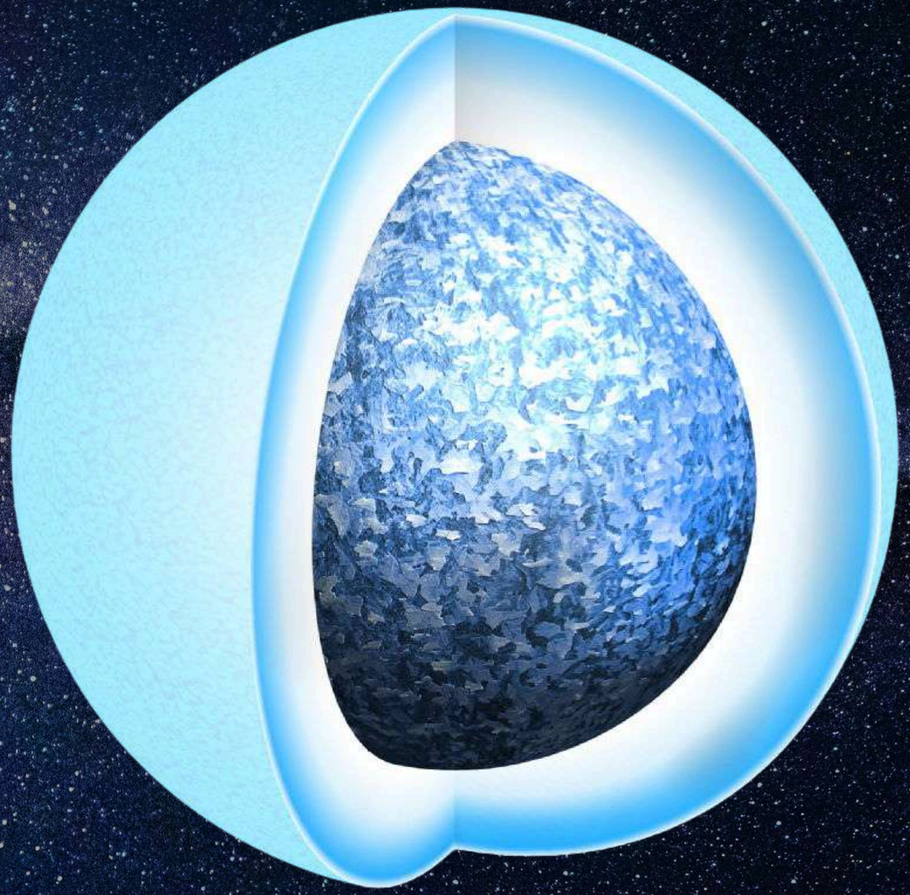
*Clara Valle*
*Andrija Zûpic*
*Ángel Regueiro*
*Lara Pallas-Quintela*
*Carlos Dafonte*
*Manuel Penedo*

*Grant: PID2024-157964OB-C22*


* *xabier.perez.couto@udc.gal*

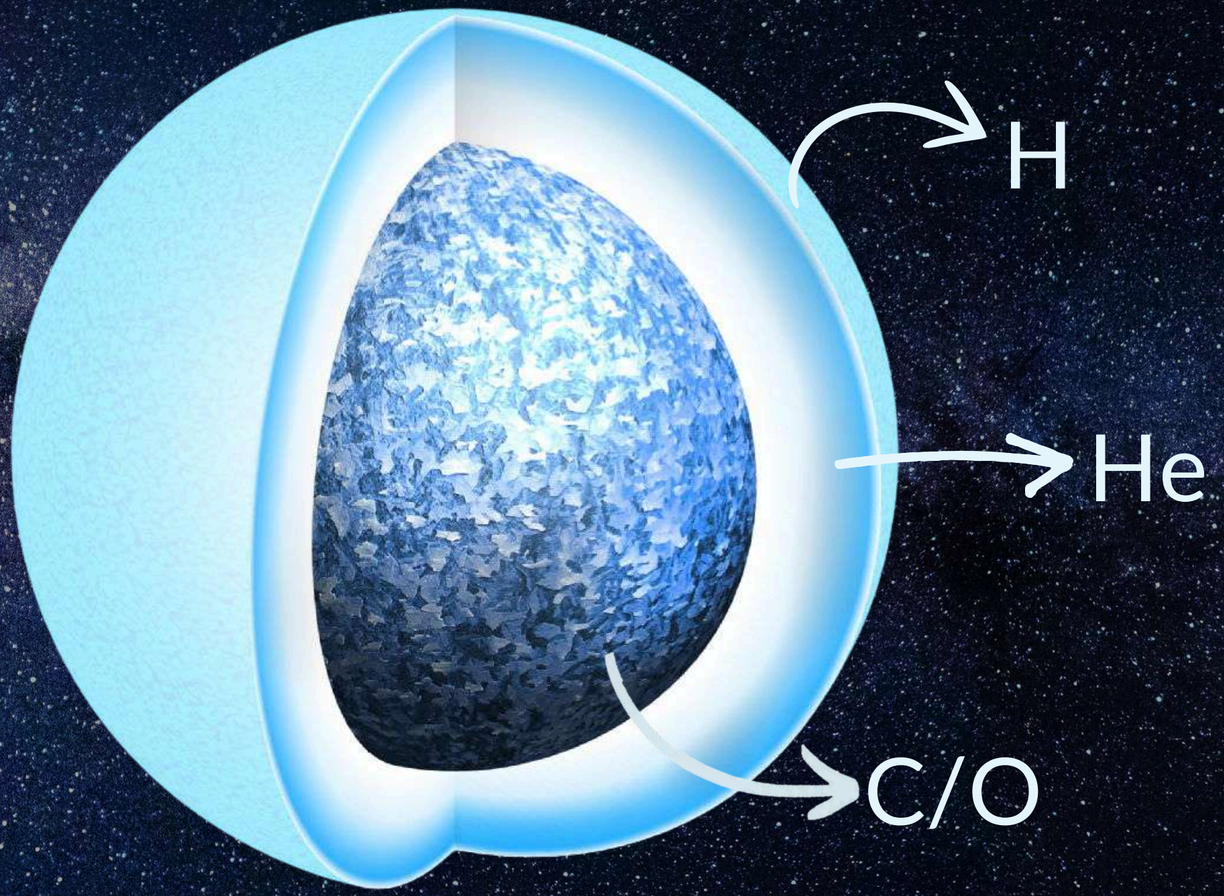
***ML4Astro, Valleta 2026***

# What are white dwarfs made of?



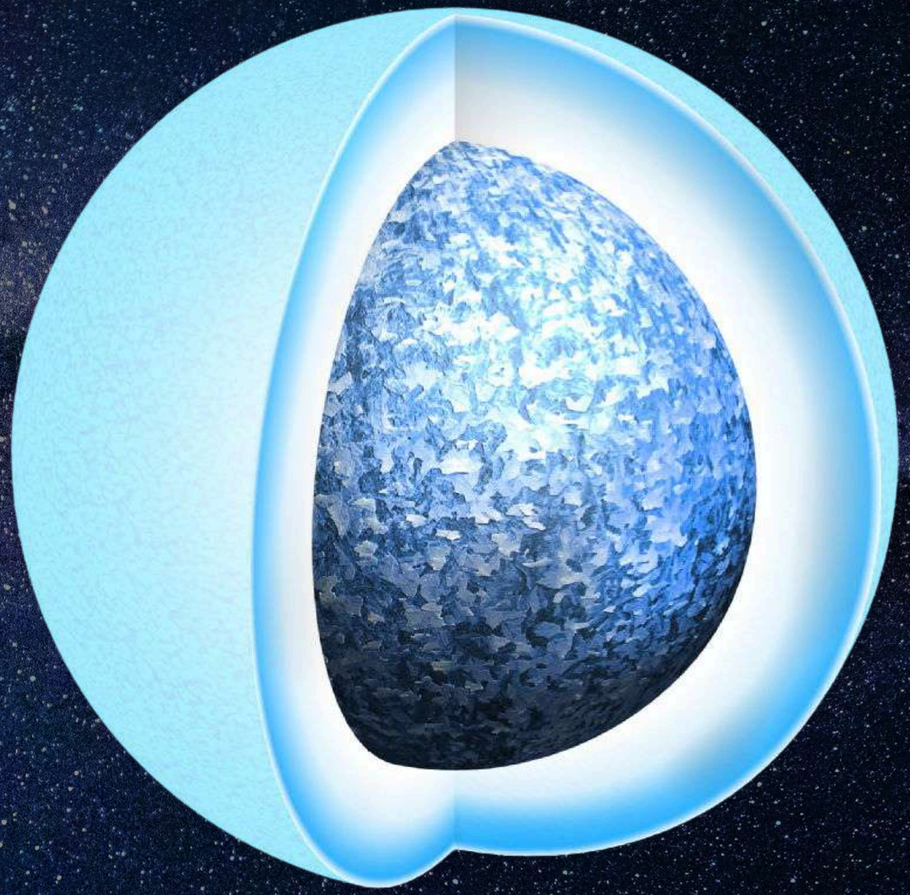
Credit: Mark Garlick
(University of Warwick)

# What are white dwarfs made of?




Credit: Mark Garlick
(University of Warwick)

# What are white dwarfs made of?



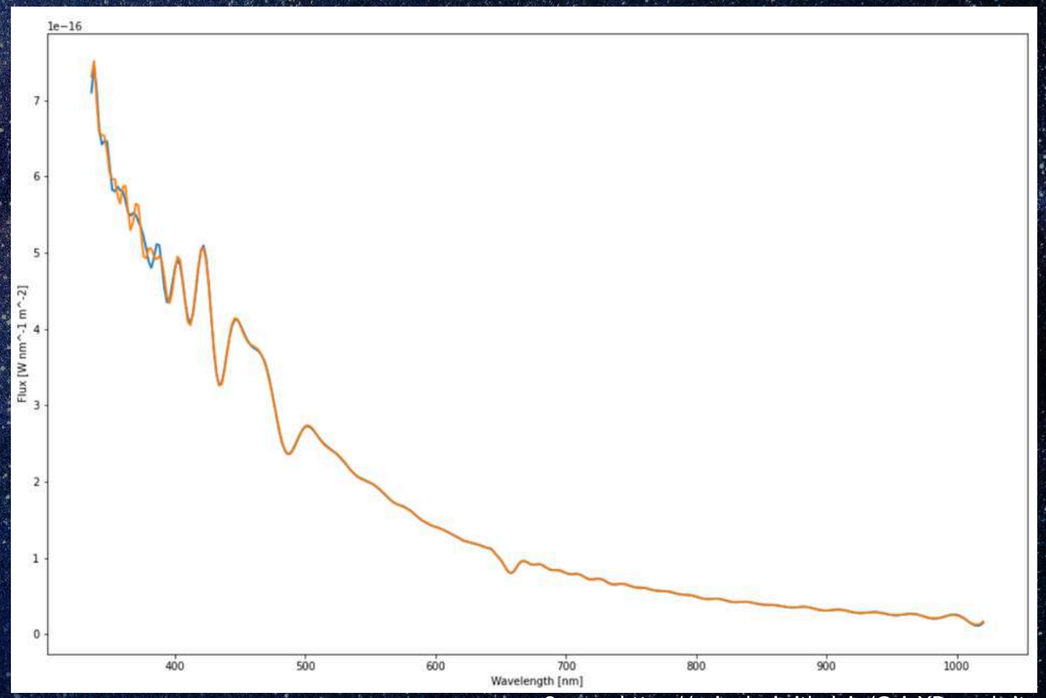
Credit: Mark Garlick
(University of Warwick)

**Zoo of WD spectral classes:**

DA (Balmer lines)
DB (He I lines)
DO (He II lines)
DQ (C lines)
DZ (heavy metals)
DC (featureless)

MWDD: 44 182 confirmed WDs (Dufour et al. 2017)[1]

Gaia DR3: ~100 000 WD candidates (Gentile-Fusillo et al. 2021) with low-resolution spectra (R ≈ 70)




Source: https://gaia-dpci.github.io/GaiaXPy-website

$$h(x) = \sum_{n=0}^{N} c_n \cdot \varphi_n(x)$$

basis functions (Hermite functions)

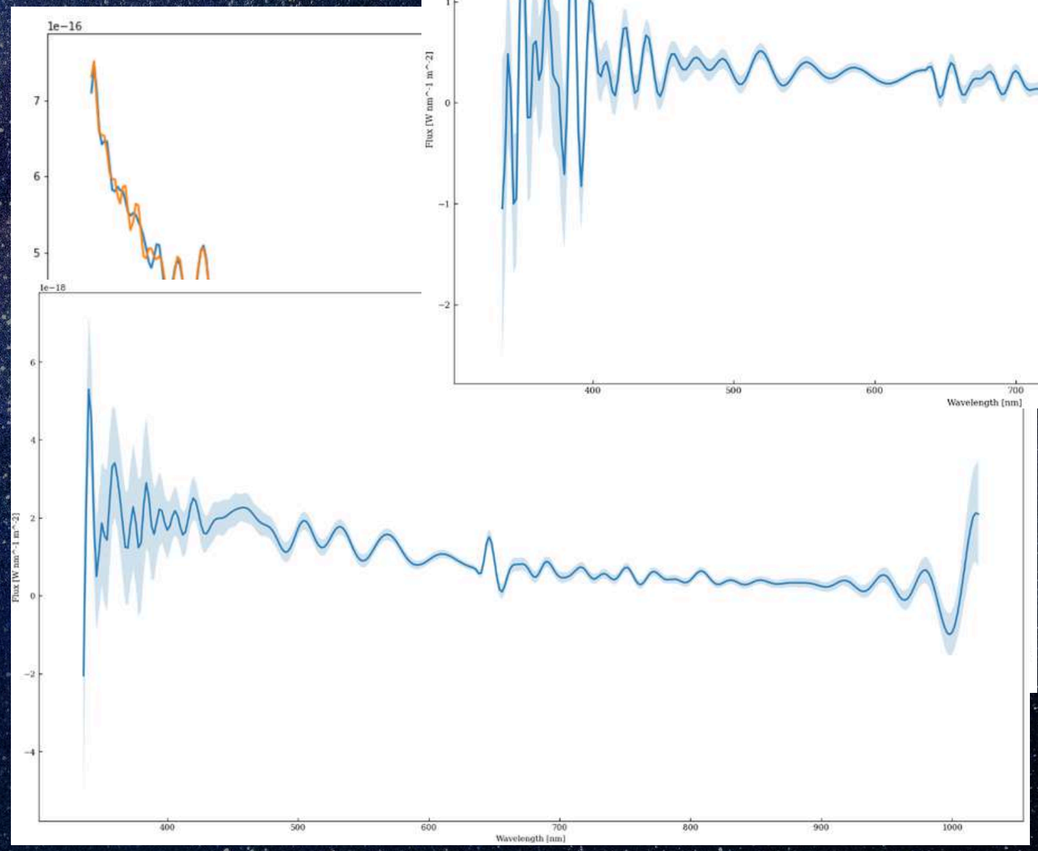
SED

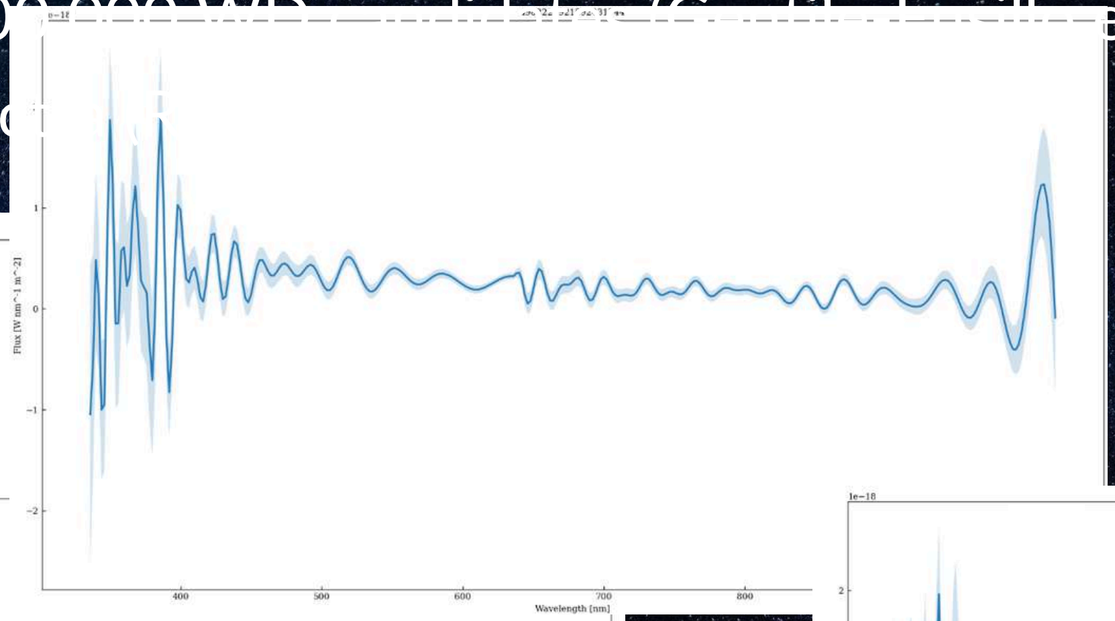
XP coefficients (110 coefficients, 55 for BP and 55 for RP spectra)

[1]https://www.montrealwhitedwarfdatabase.org/

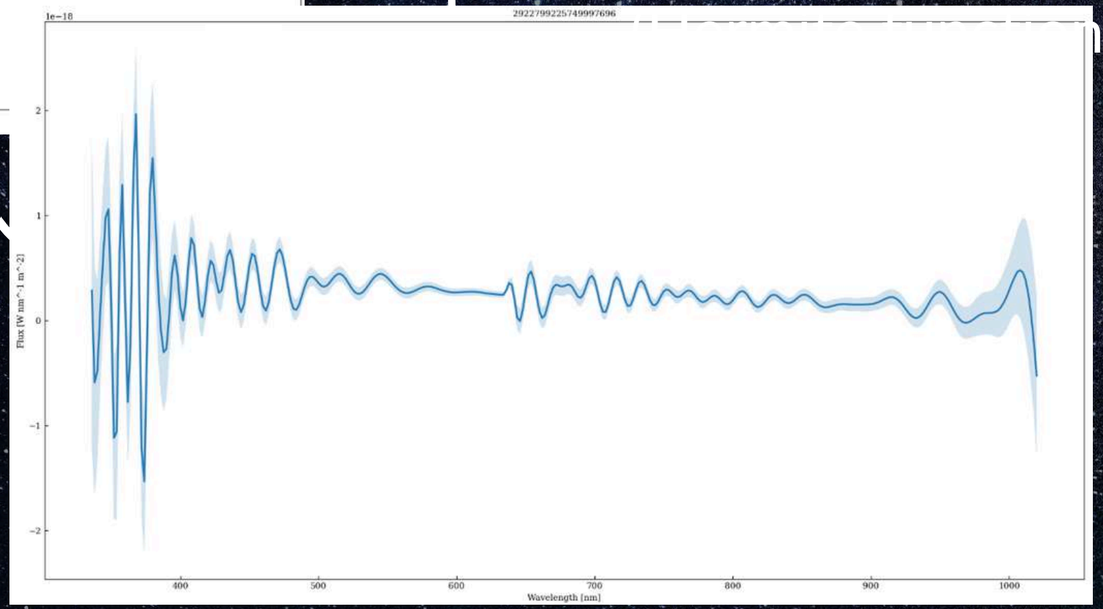
MWDD: 44 182 confirmed WDs (Dufour et al. 2017)[1]

Gaia DR3: ~100 000 WD candidates (Gentile Fusillo et al. 2021) with low-resolution spec



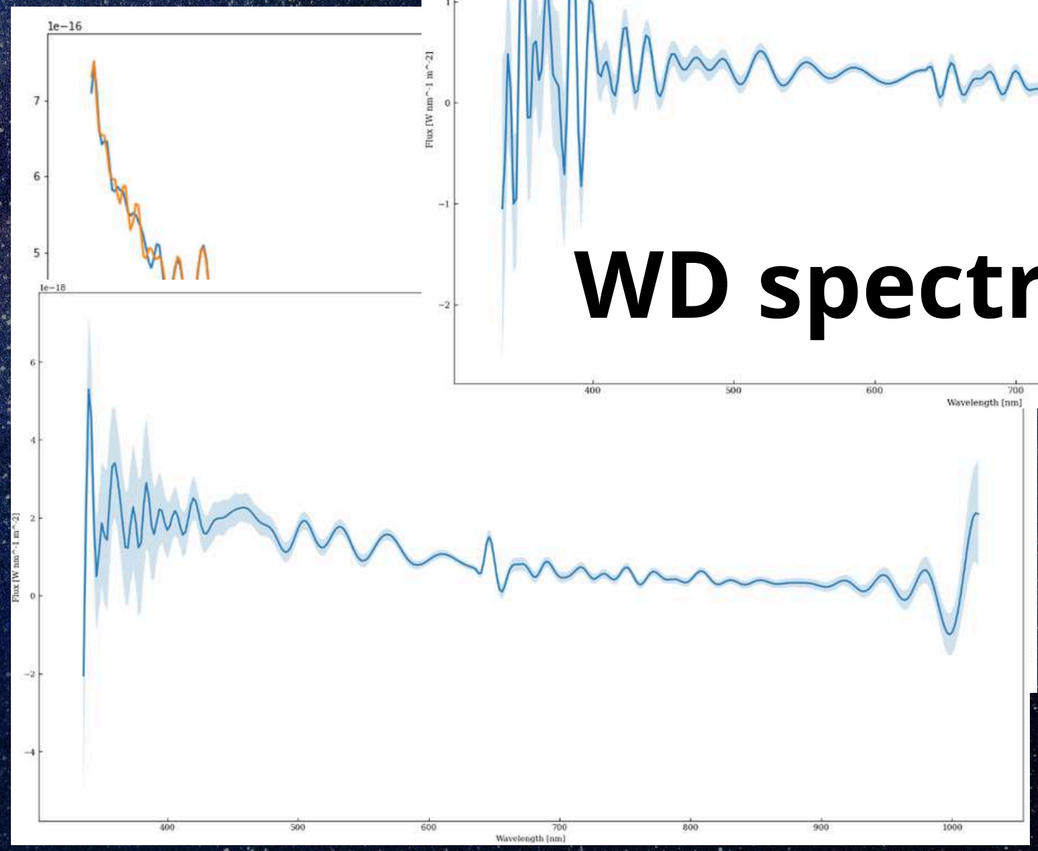



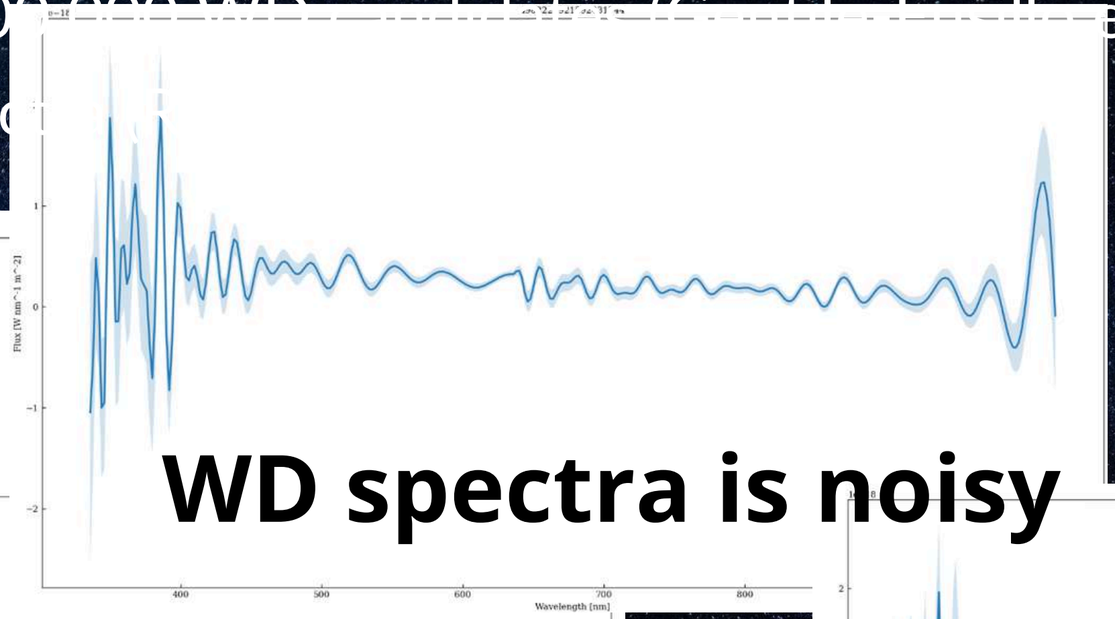

basis functions



MWDD: 44 182 confirmed WDs (Dufour et al. 2017)[1]

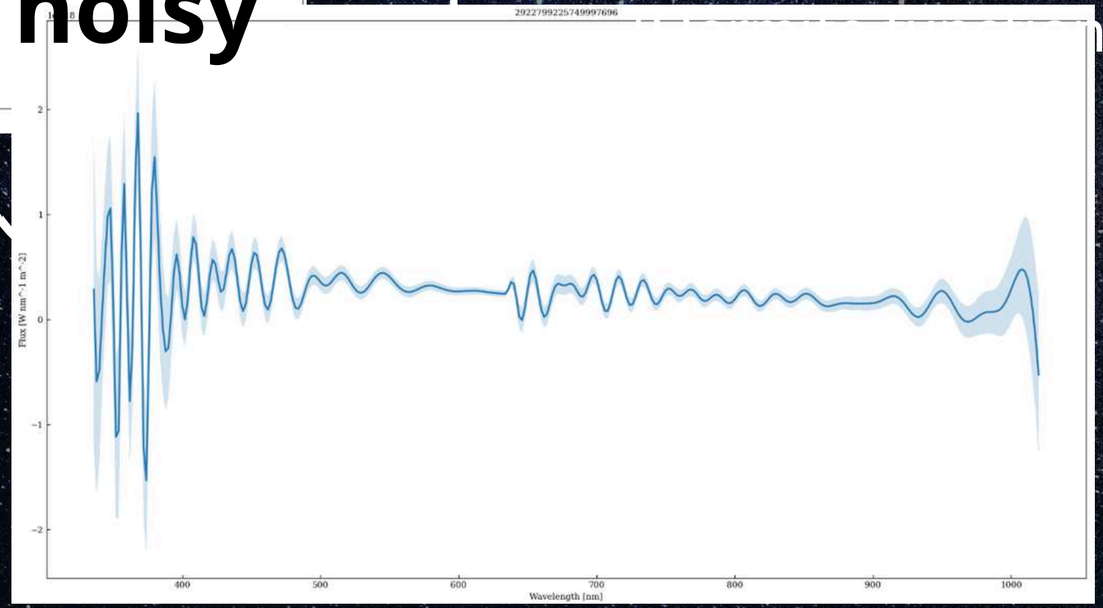
Gaia DR3: ~100 000 WD candidates (Gentile Fusillo et al. 2021) with low-resolution spec




**WD spectra is noisy**

basis functions



MWDD: 44 182 confirmed WDs (Dufour et al. 2017)[1]

Gaia DR3: ~350 000 WD candidates (Gentile-Fusillo et al. 2021)

Supervised ML Classification:

- Gradient Boosted Decision Trees: Vincent et al. (2024)
- Random Forest: García-Zamora et al. (2023)

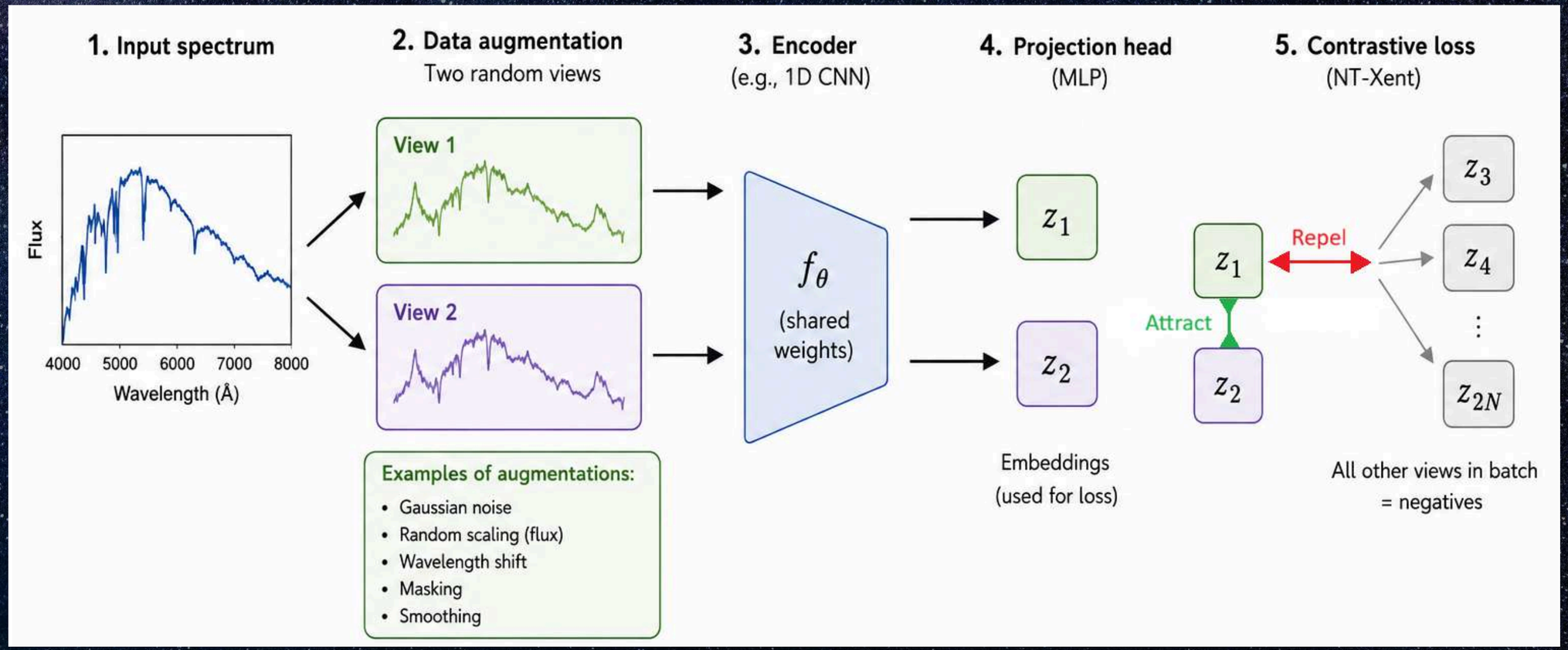
Unsupervised ML Classification:

- UMAP: Kao et al. (2024)
- SOM: Pérez-Couto et al. (2024), Pérez-Couto et al. (2025)

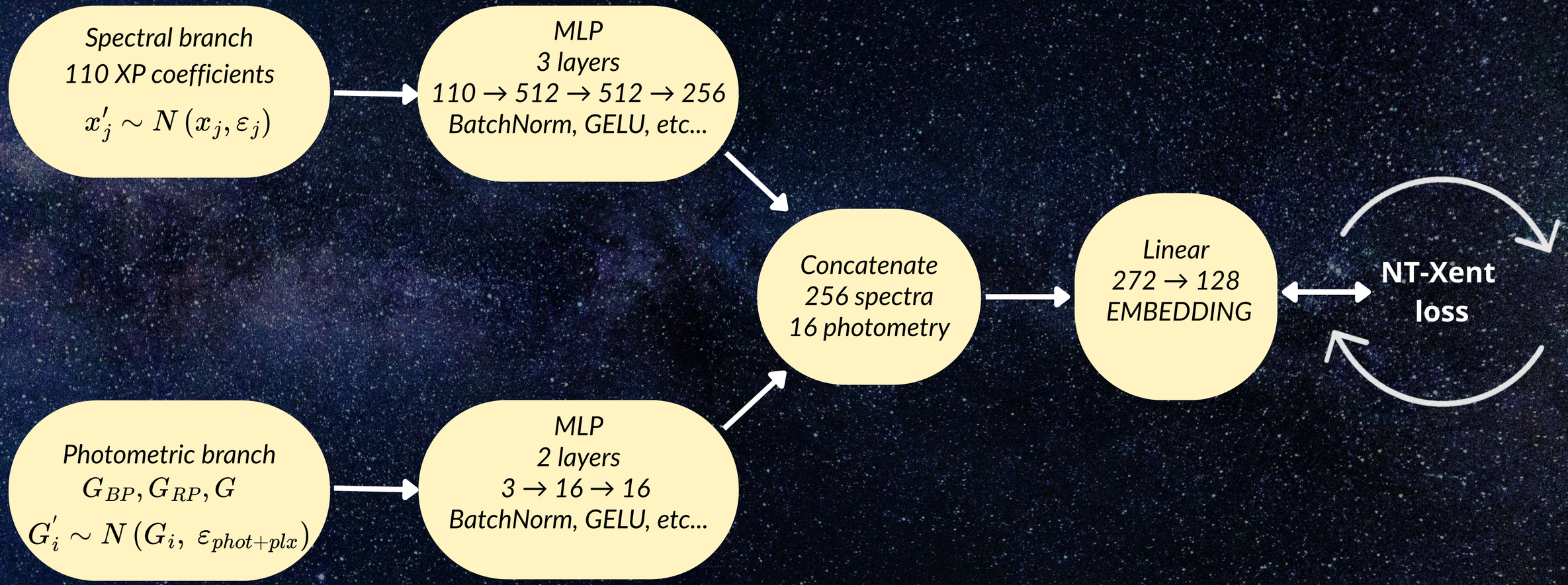
**Can a NN learn a better representation of the XP spectra?**

## SimCLR in one idea

- Take a star. Make **two corrupted copies** of it ("views") by adding noise
- Push both through a network to get two vectors in the latent space
- Train the network so those two vectors are **close to each other** and **far from the vectors of every other star**



# The architecture blueprint




*Pérez-Couto et al. in prep*

# Results

**Training set:**

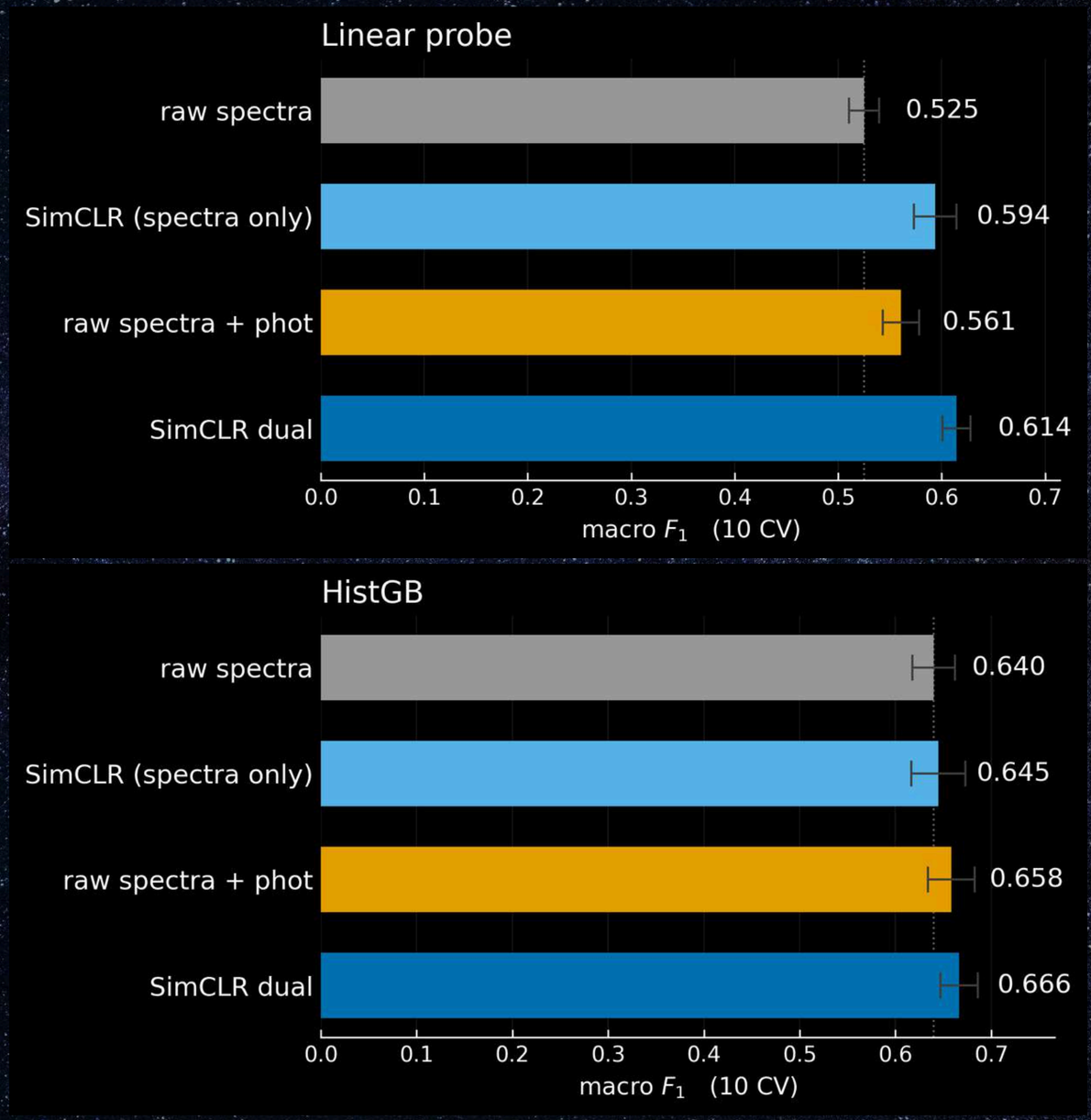
101 702 WDs with Gaia XP spectra
12 348 labeled (MWDD)

| | | |
|---|---|---|
| DA | 9597 | 78% |
| DAH | 214 | 2% |
| DB | 1004 | 8% |
| DC | 861 | 7% |
| DQ | 331 | 3% |
| DZ | 341 | 3% |

Logistic Regression and Histogram-based Gradient Boosting as downstream classifiers

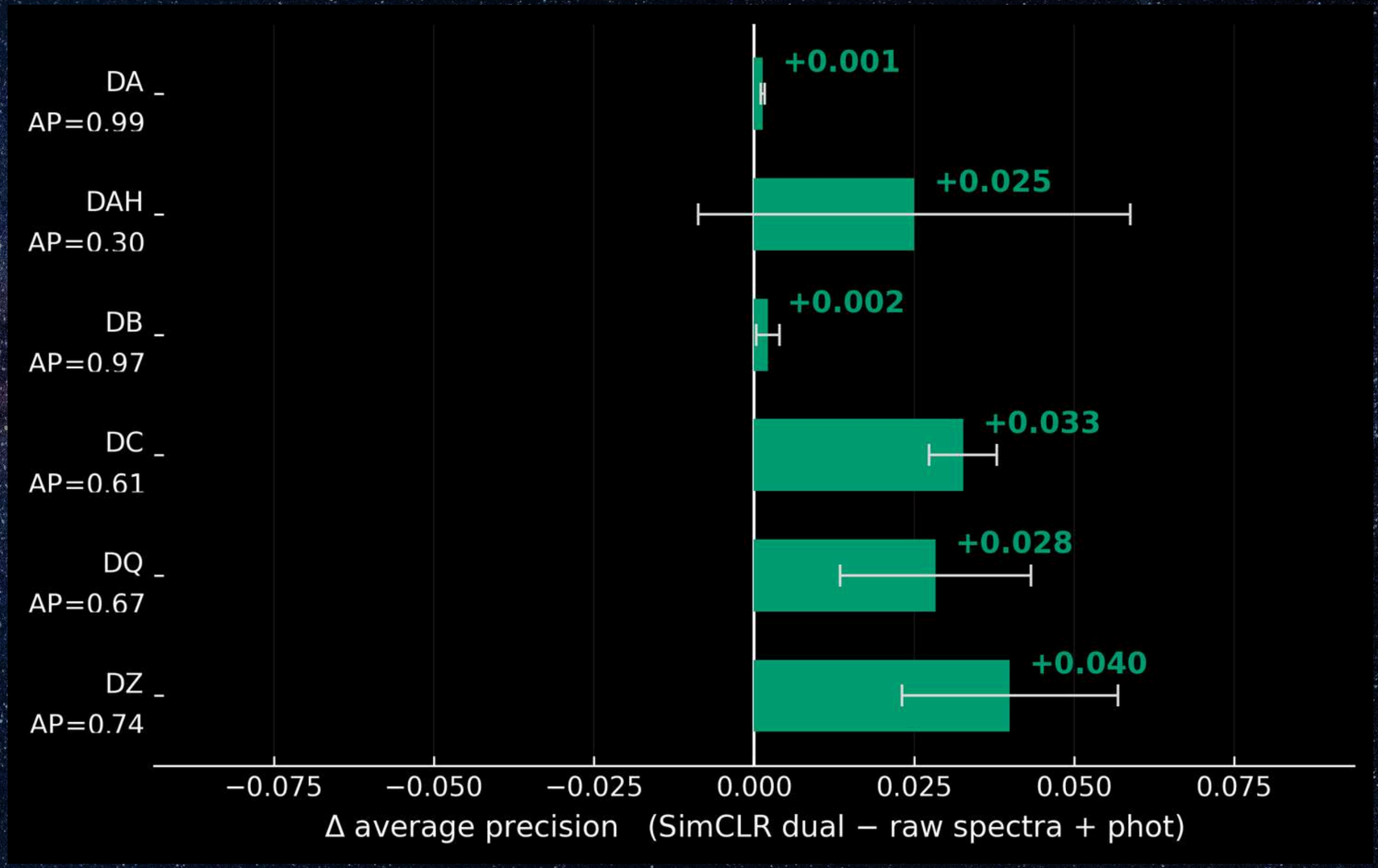
Stratified K-fold 9:1

Macro F1 as the diagnostic metric



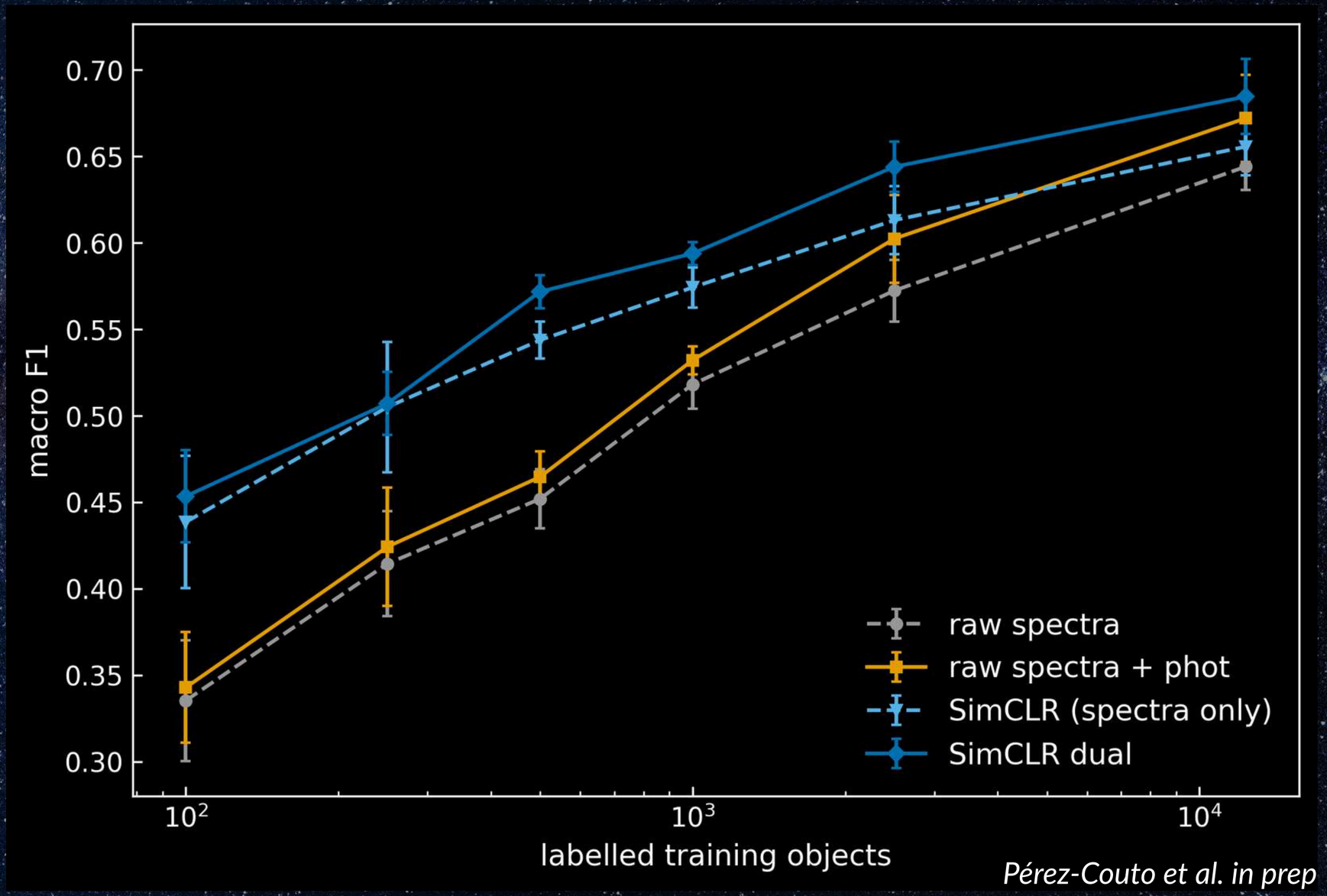

*Pérez-Couto et al. in prep*

# Results




*Pérez-Couto et al. in prep*

# Results




*Pérez-Couto et al. in prep*

# Take-aways

- Contrastive pretraining doesn't add information:  it reorganises it
- The learned embedding significantly improved the macro F1 of linear classifiers, while the upgrade is meaningless when using gradient boosting decision trees
- Minority classes (DQ, DZ) show the biggest improvement
- Label efficient: 3x fewer labels for the same macro F1
- Next: more sophisticated augmentations, external survey data, Gaia DR4...

**Thanks! - Grazzi!**

*xabier.perez.couto@udc.gal*