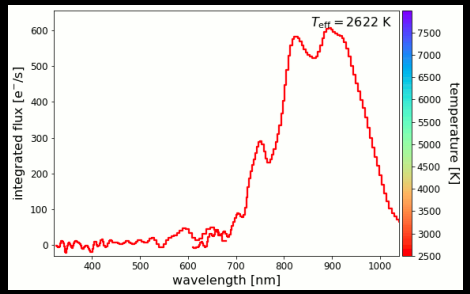
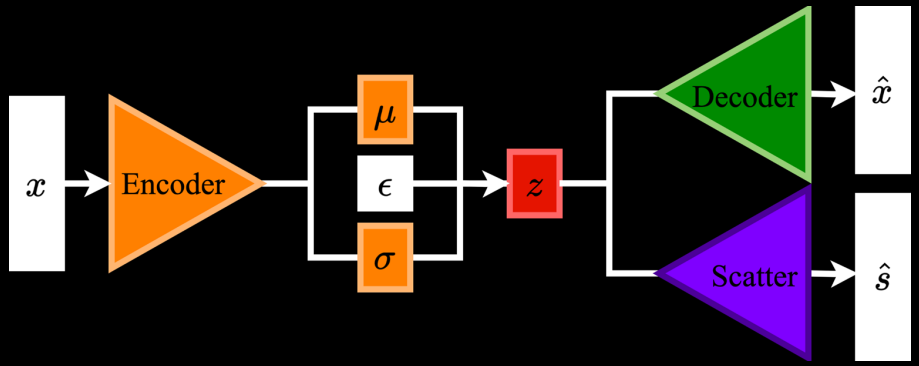
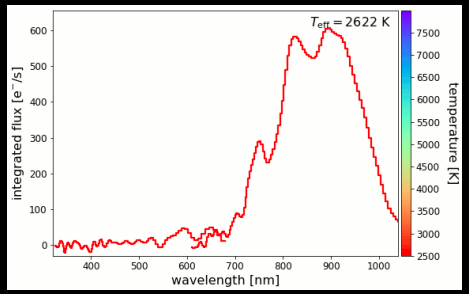






(arXiv:2404.07316)

Towards the discovery of rare stellar populations in the Gaia XP spectra with stellar label independent modeling

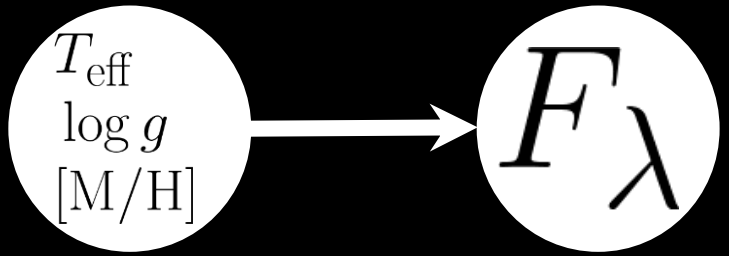
Acknowledging:
Josh Speagle
& Maria Drout

Alex Laroche
University of Toronto
Dunlap Institute

DUNLAP INSTITUTE FOR ASTRONOMY & ASTROPHYSICS

Stellar label dependent model

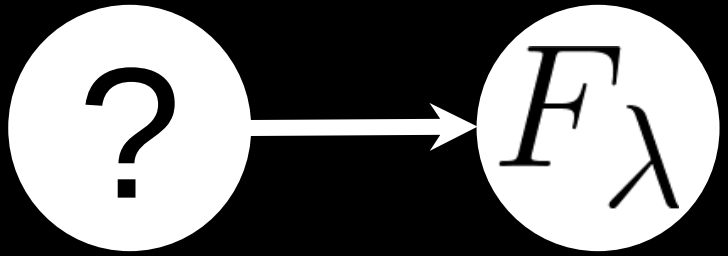
A (generative) stellar label dependent model simulates stellar spectra from ‘stellar labels’



Examples: The Cannon (Ness+15), The Payne (Ting+18), Cycle-Starnet (O’Briain+20), Gaia XP model (Zhang,Green,Rix23), Transformer model (Leung,Bovy23) and many more!

Stellar label independent model

A (generative) stellar label independent model simulates stellar spectra WITHOUT labels



Examples: Not many!
*Our model (**Laroche**,Speagle24),*
also Transformer model (Leung,Bovy23)

Stellar label dependent models suffer from stellar label systematics which decrease model performance:

THE STELLAR LABELS GAP

Stellar label dependent models require ‘good’ stellar labels to train on, but what if…

Stellar label dependent models require ‘good’ stellar labels to train on, but what if…

1. Your training labels are **poorly estimated** (inaccurate synthetic models) X
2. Certain **stellar sub-populations** in your data are not well summarized by labels (e.g. chemically peculiar stars) X
3. You do not have **enough stellar labels** in certain regions of the stellar parameter space your data spans X
4. You have a significant number of **stellar multiples** in your data (binaries, triples, etc.) X

1 + 2 + 3 + 4 = THE STELLAR LABELS GAP

1. Your training labels are **poorly estimated** (inaccurate synthetic models) X
2. Certain **stellar sub-populations** in your data are not well summarized by labels (e.g. chemically peculiar stars) X
3. You do not have **enough stellar labels** in certain regions of the stellar parameter space your data spans X
4. You have a significant number of **stellar multiples** in your data (binaries, triples, etc.) X

1 + 2 + 3 + 4 = THE STELLAR LABELS GAP

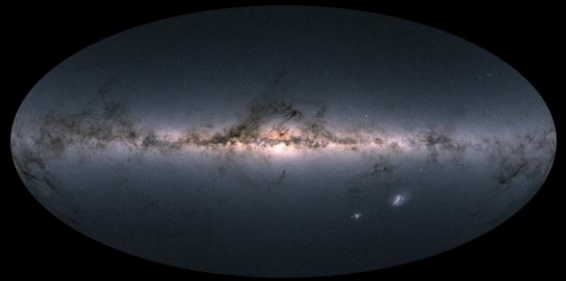
1. Your training labels are **poorly estimated** (inaccurate synthetic models) X
2. Certain **stellar sub-populations** in your data are not well summarized by labels (e.g. chemically peculiar stars) X
3. You do not have **enough stellar labels** in certain regions of the stellar parameter space your data spans X
4. You have a significant number of **stellar multiples** in your data (binaries, triples, etc.) X

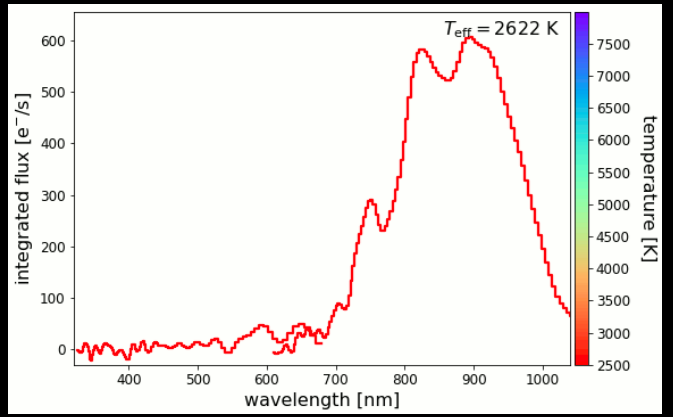
Low-resolution BP/RP (XP) spectra in Gaia DR3

1. Largest spectroscopic survey ever (220+ million spectra)

2. Extremely low resolution (R~100 from ~300 to 1000 nm)

3. Almost certain that rare, undiscovered stellar populations are hiding in the Gaia XP spectra

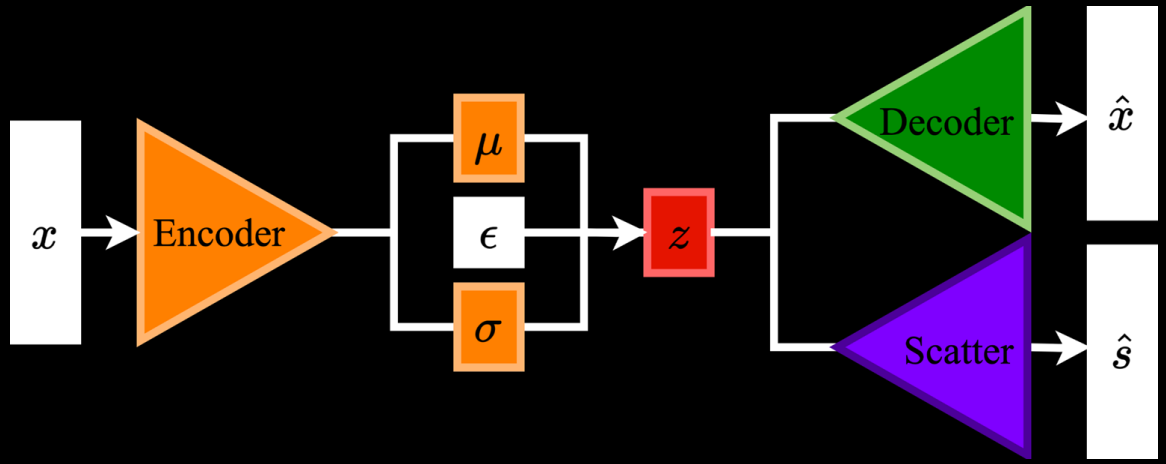




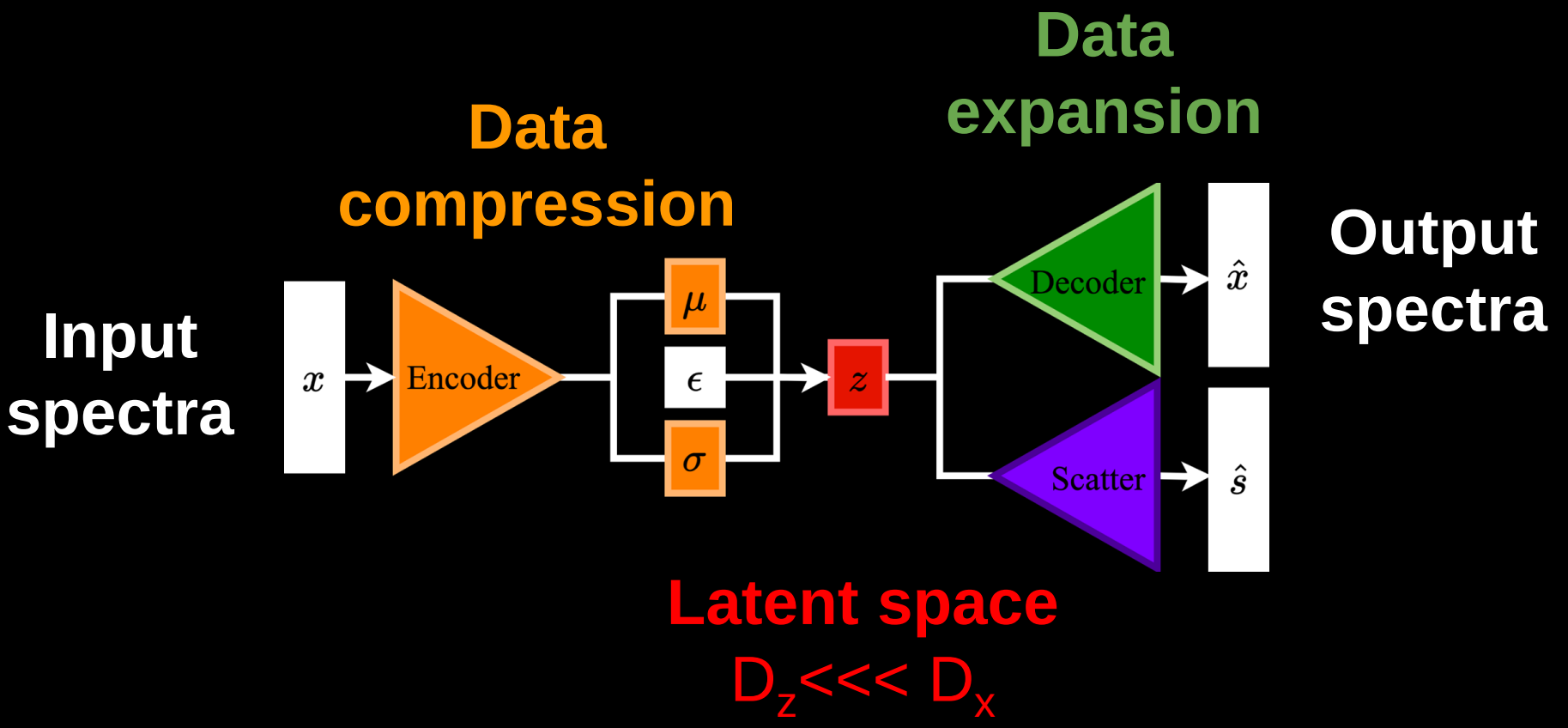
Credit: ESA/Gaia/DPAC

To close the stellar labels gap, we developed a **fully** data-driven model which simulates Gaia BP/RP spectra ***without relying on stellar labels***

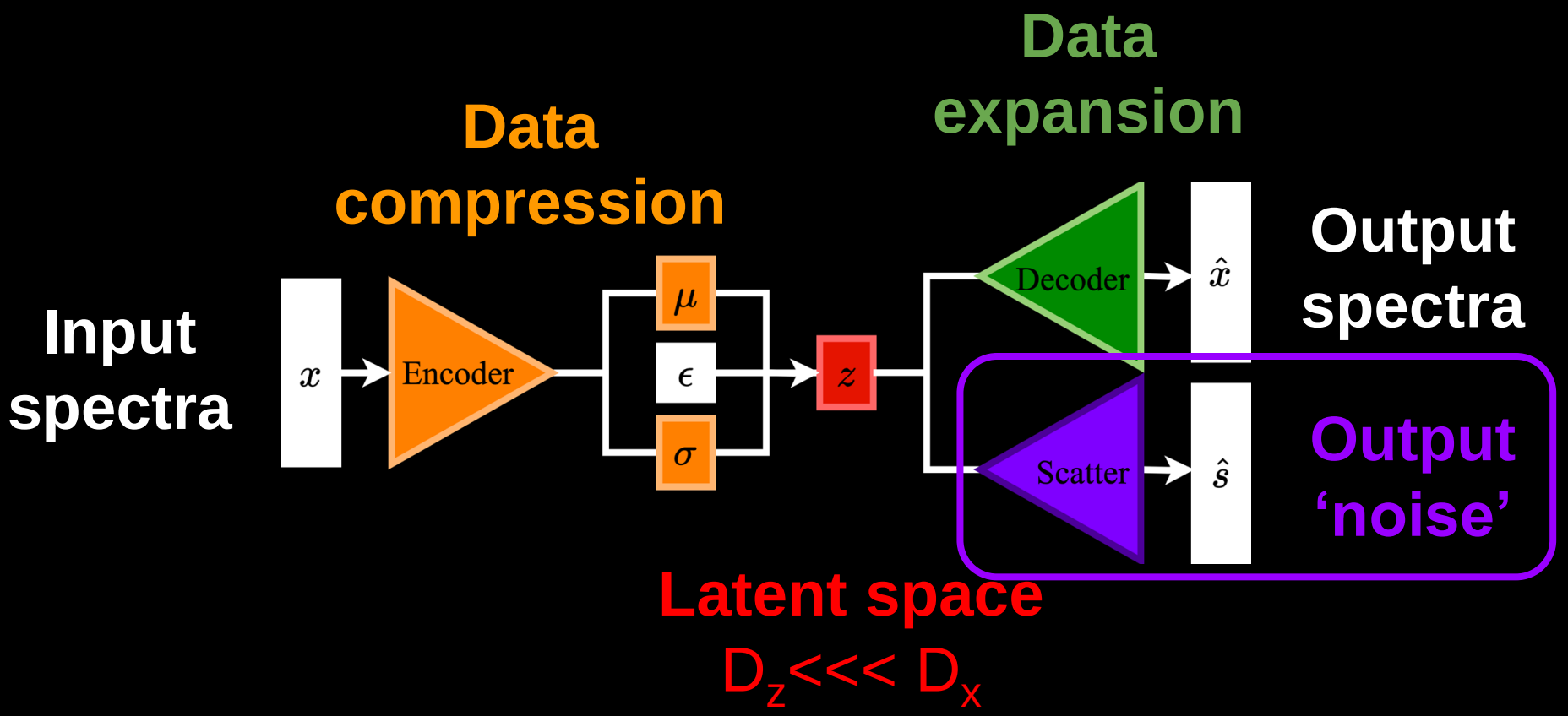
A novel variational autoencoder: ***scatter* VAE**



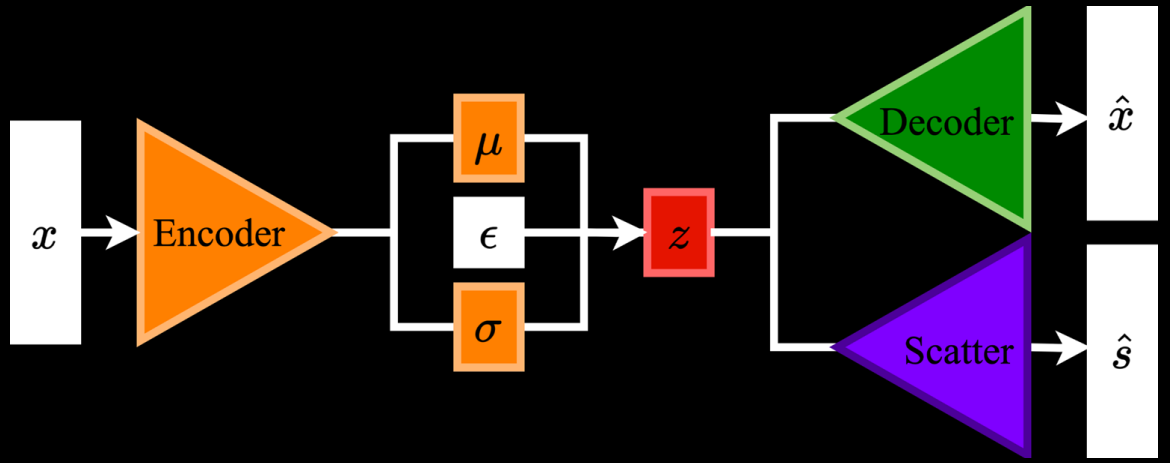
A novel variational autoencoder: ***scatter*** **VAE**



A novel variational autoencoder: *scatter* VAE



Input spectra

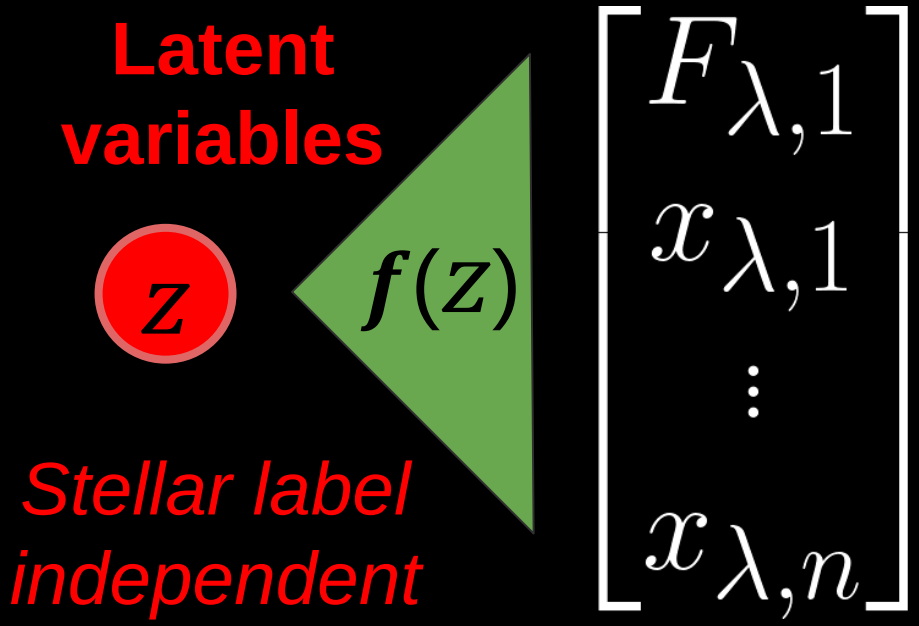


Output spectra

Training procedure:
‘Get out what you put in’

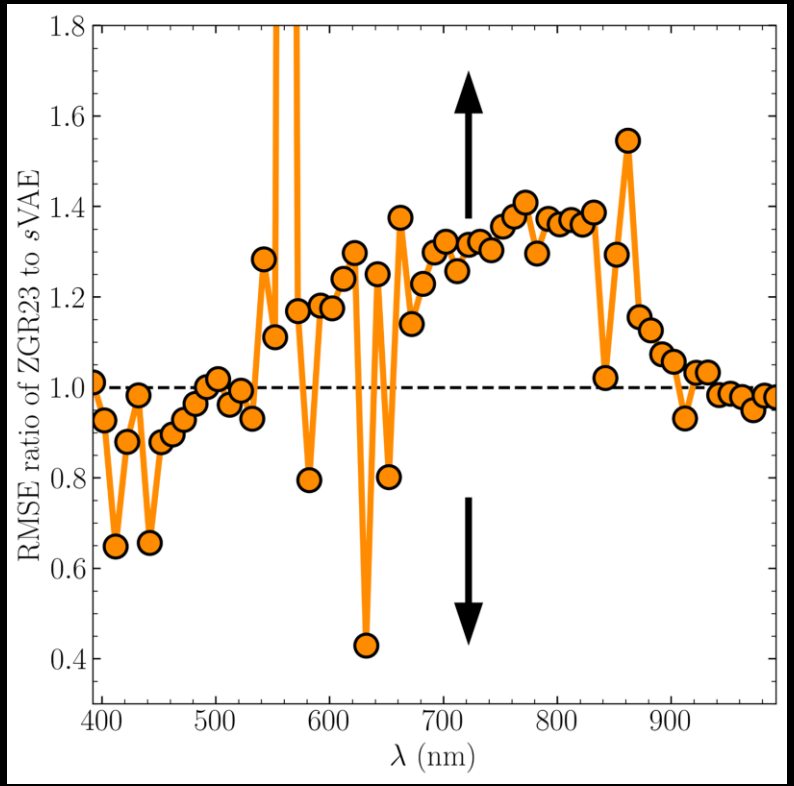
Input = Output

The end result: our model can simulate Gaia XP spectra from the **latent space**, no stellar labels required!



How does our stellar label independent **model performance** compare to stellar label dependent models?

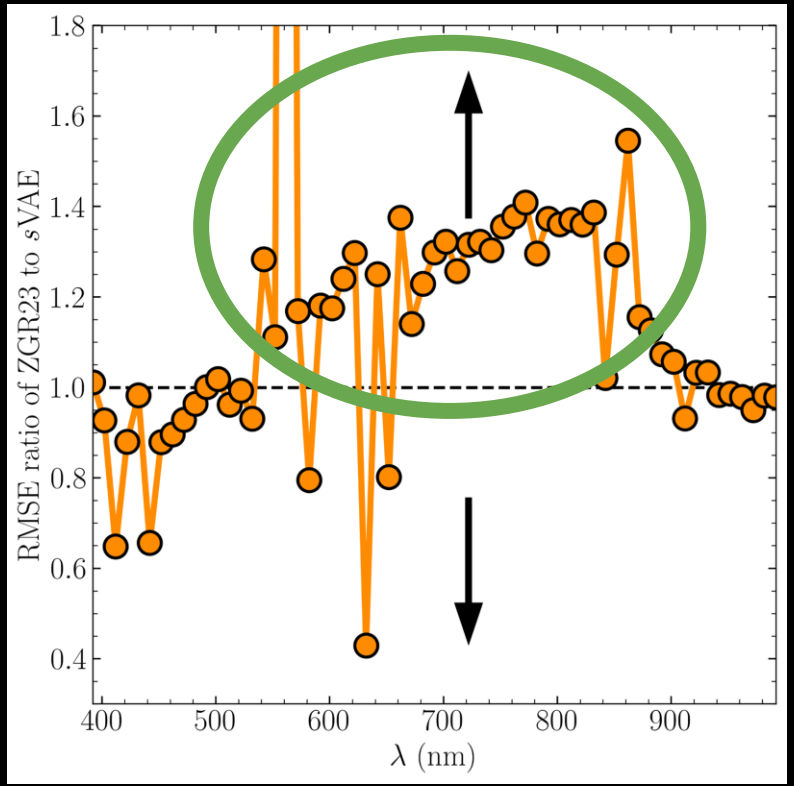
Reconstruction error comparison to the deep stellar label model of **Zhang, Green & Rix (2023)**



Better stellar label **independent** model

Better stellar label **dependent** model

Reconstruction error comparison to the deep stellar label model of **Zhang, Green & Rix (2023)**

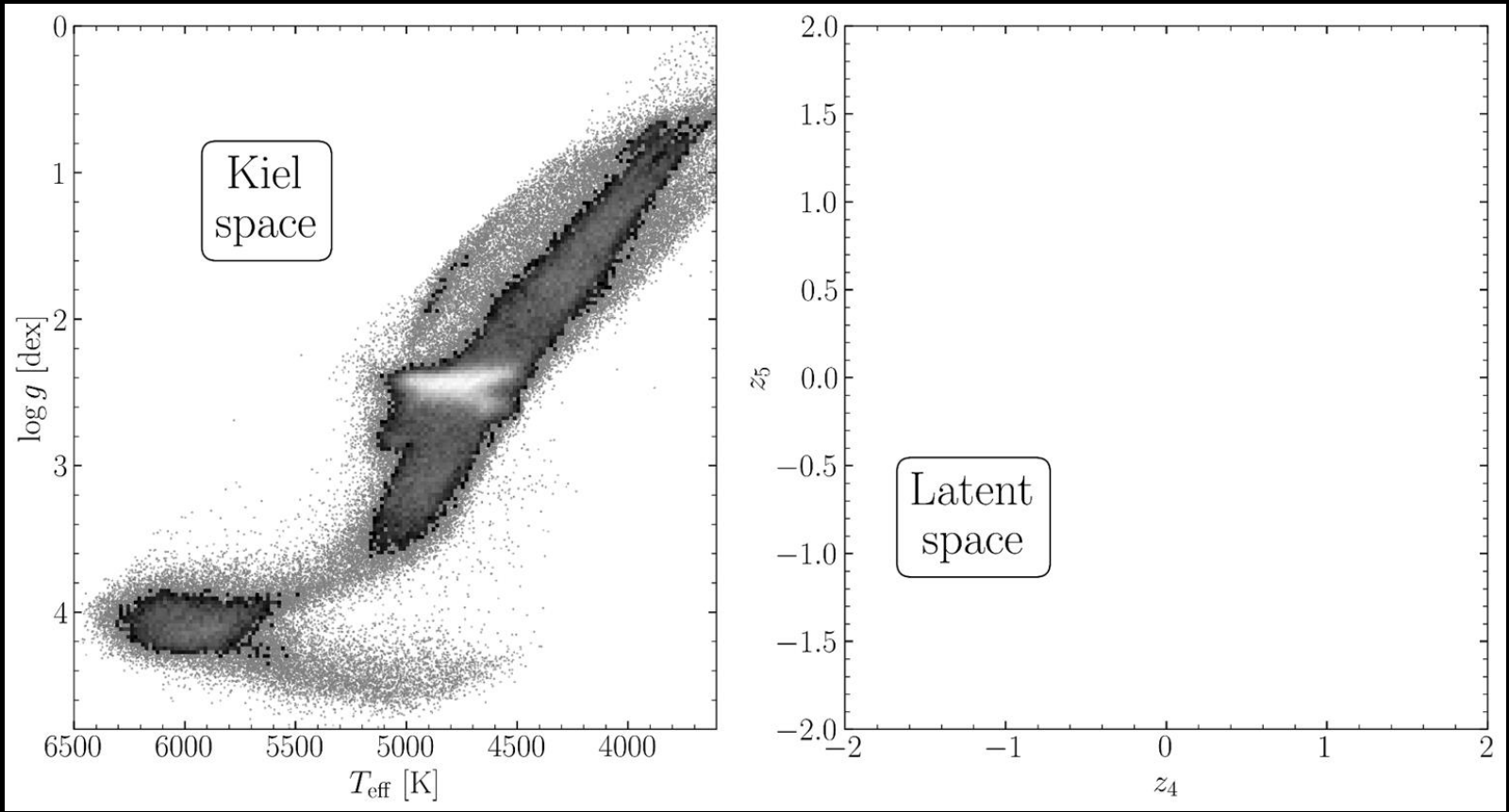


Better stellar label independent model

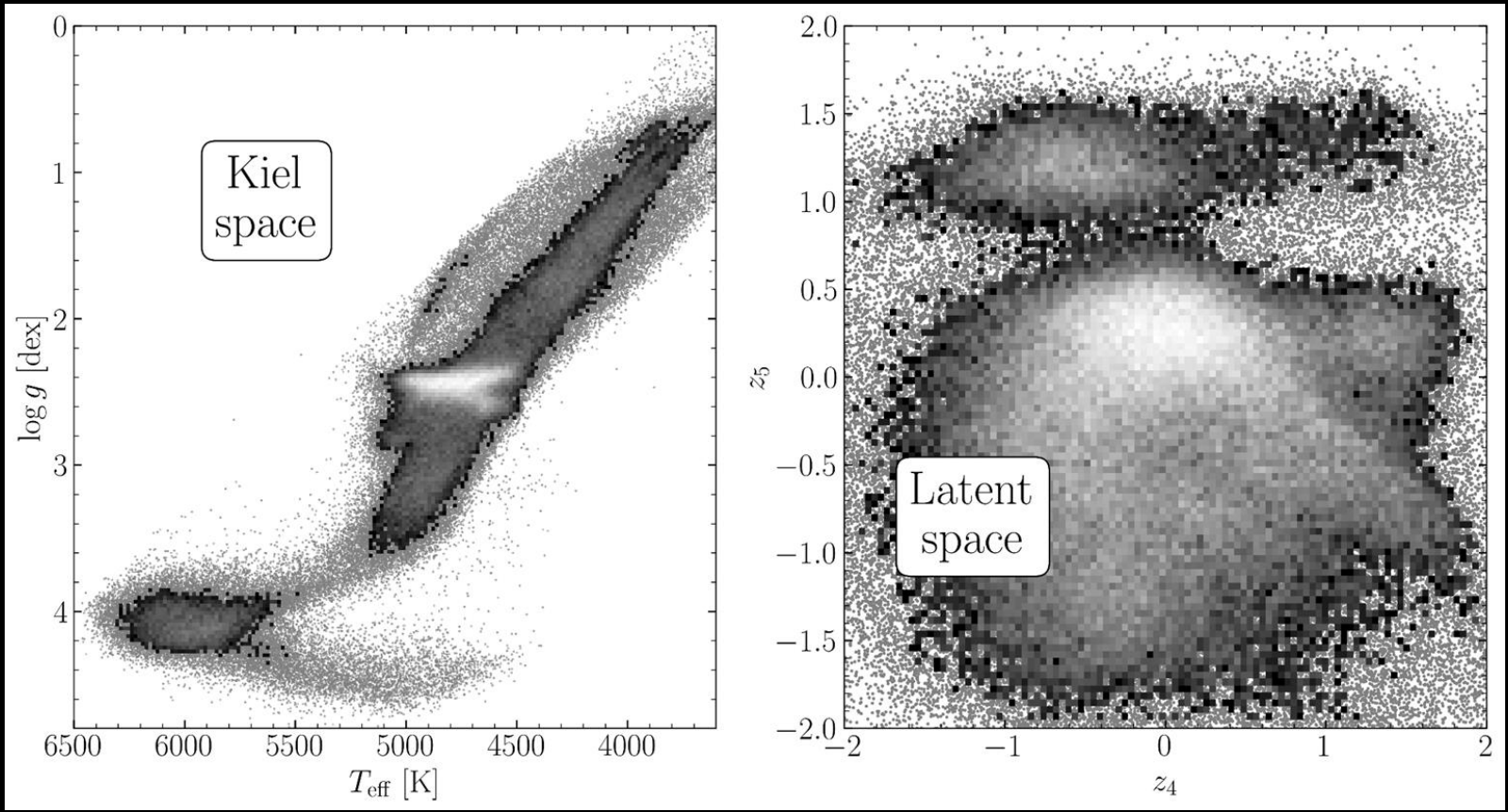
Better stellar label dependent model

What **astrophysical information** has our stellar label independent model **learned**?

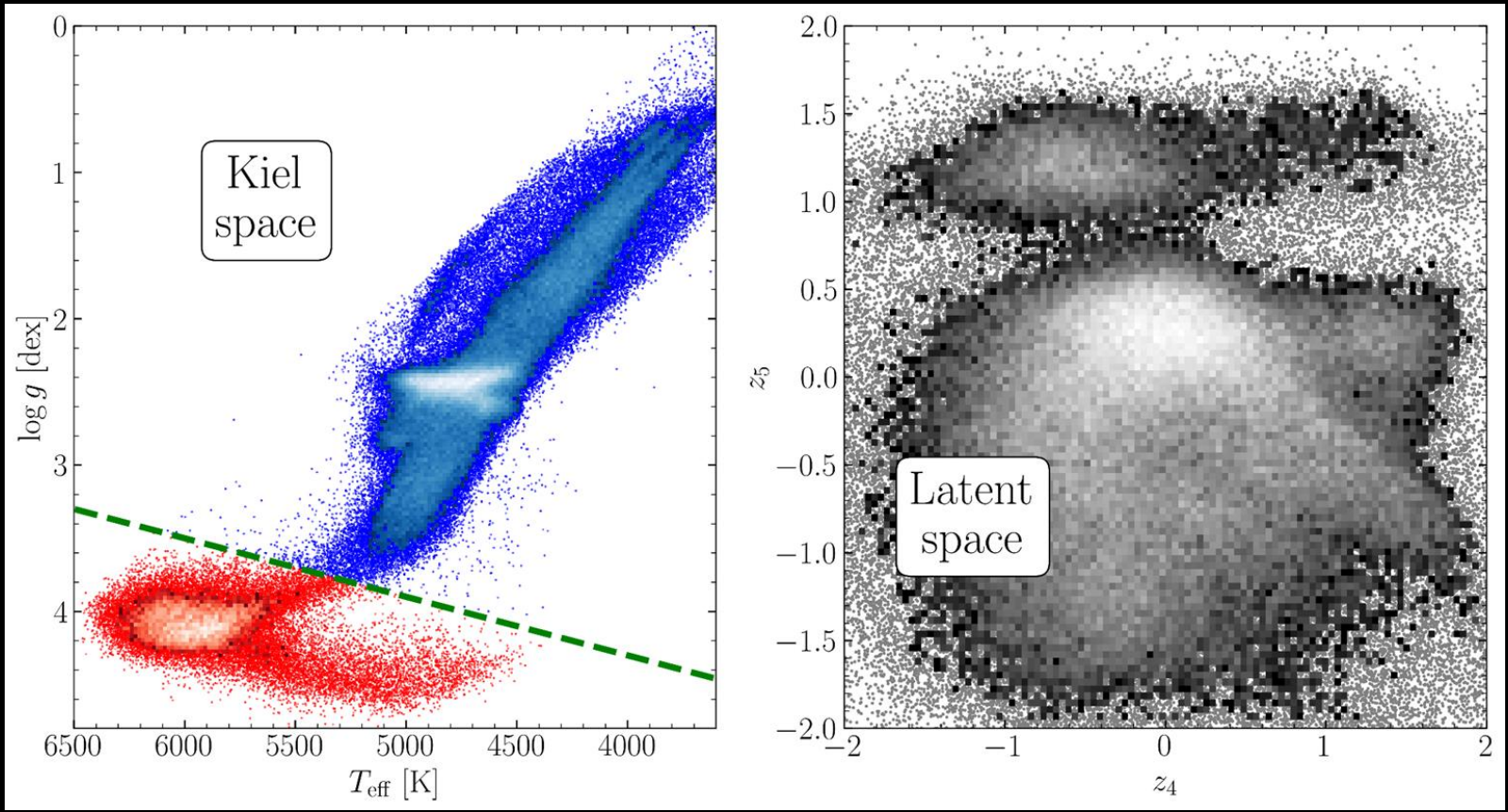
Compare Kiel vs. Latent space



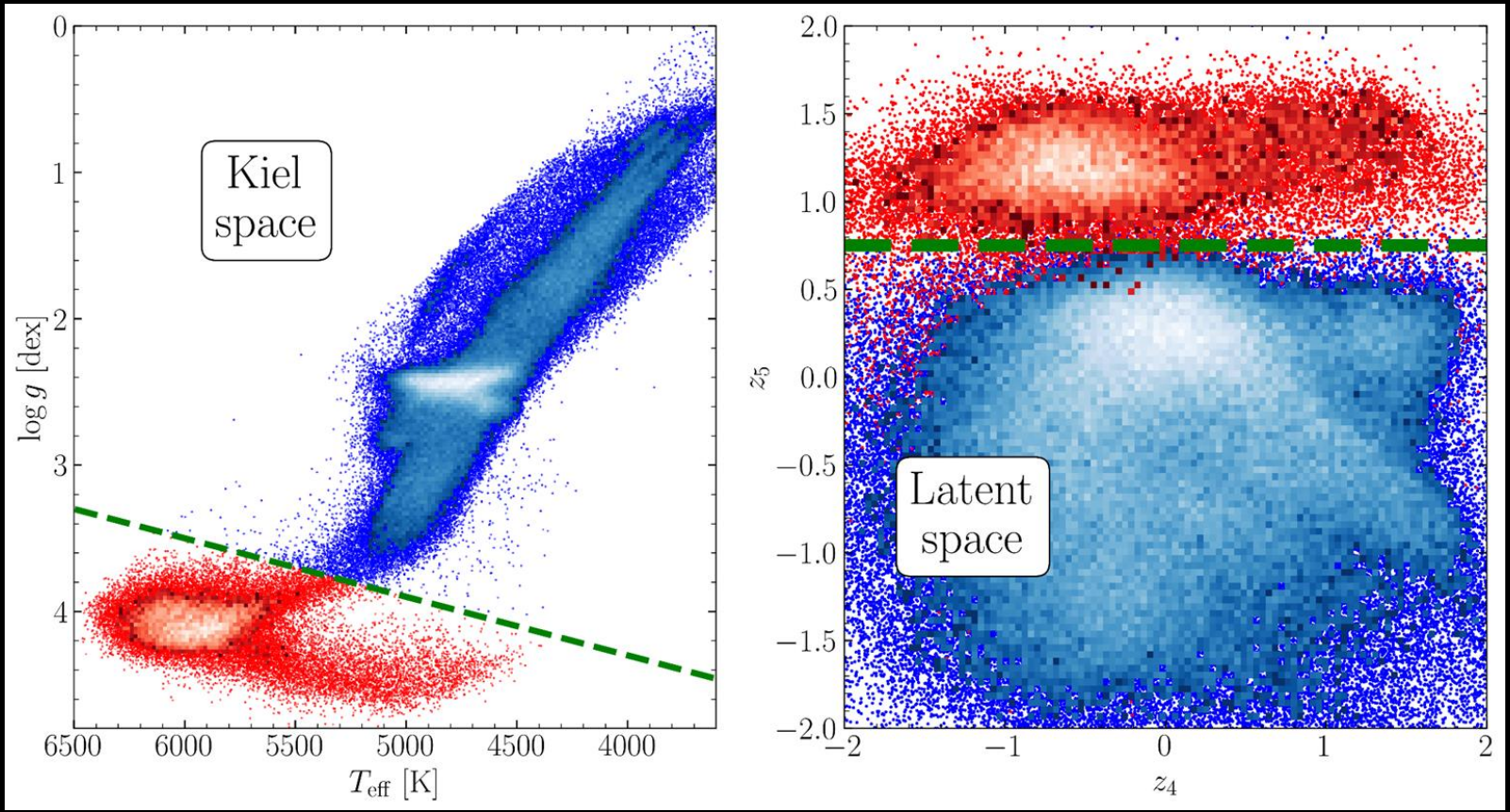
Project into latent space



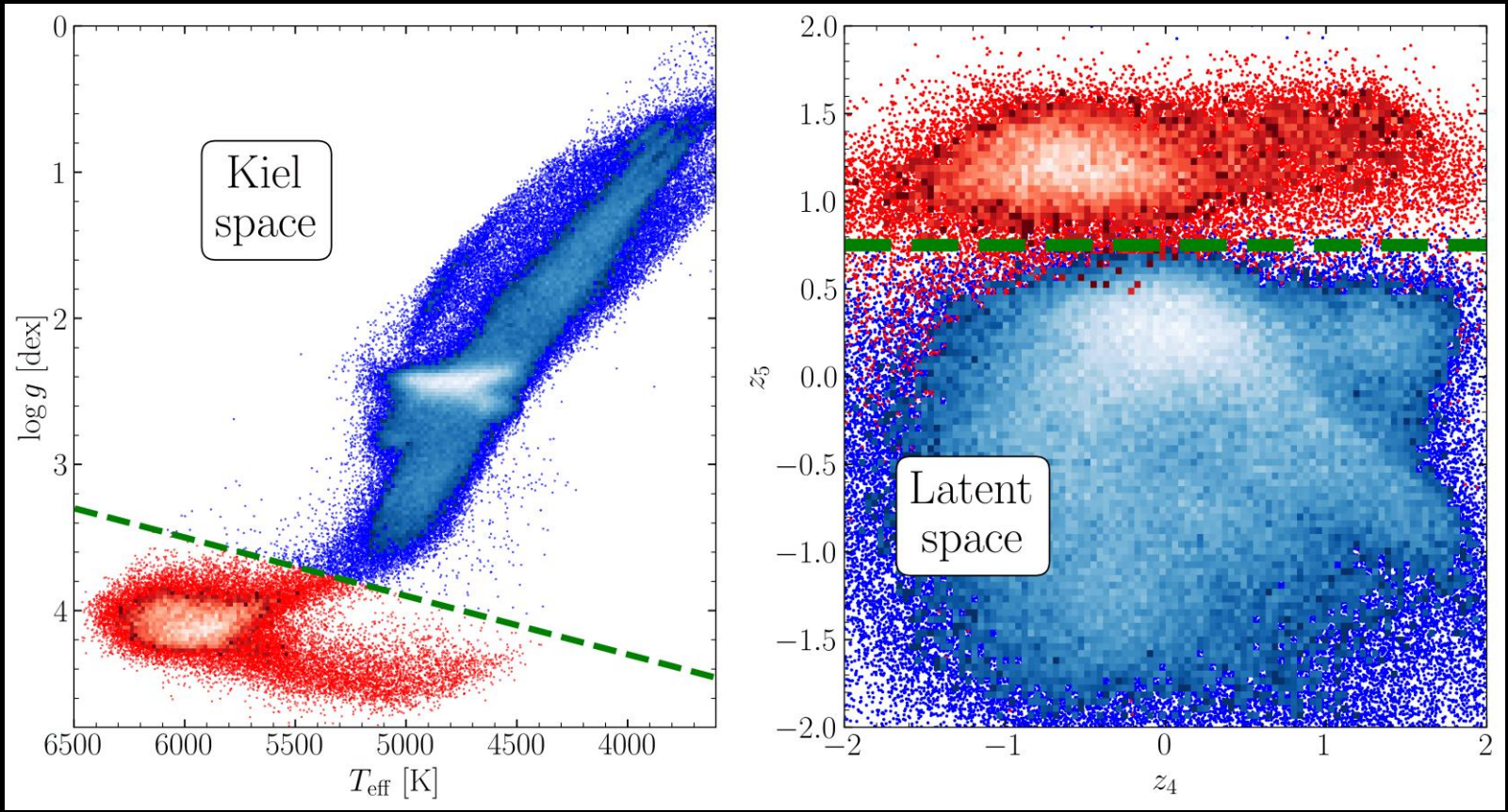
Split the main sequence and giant branch



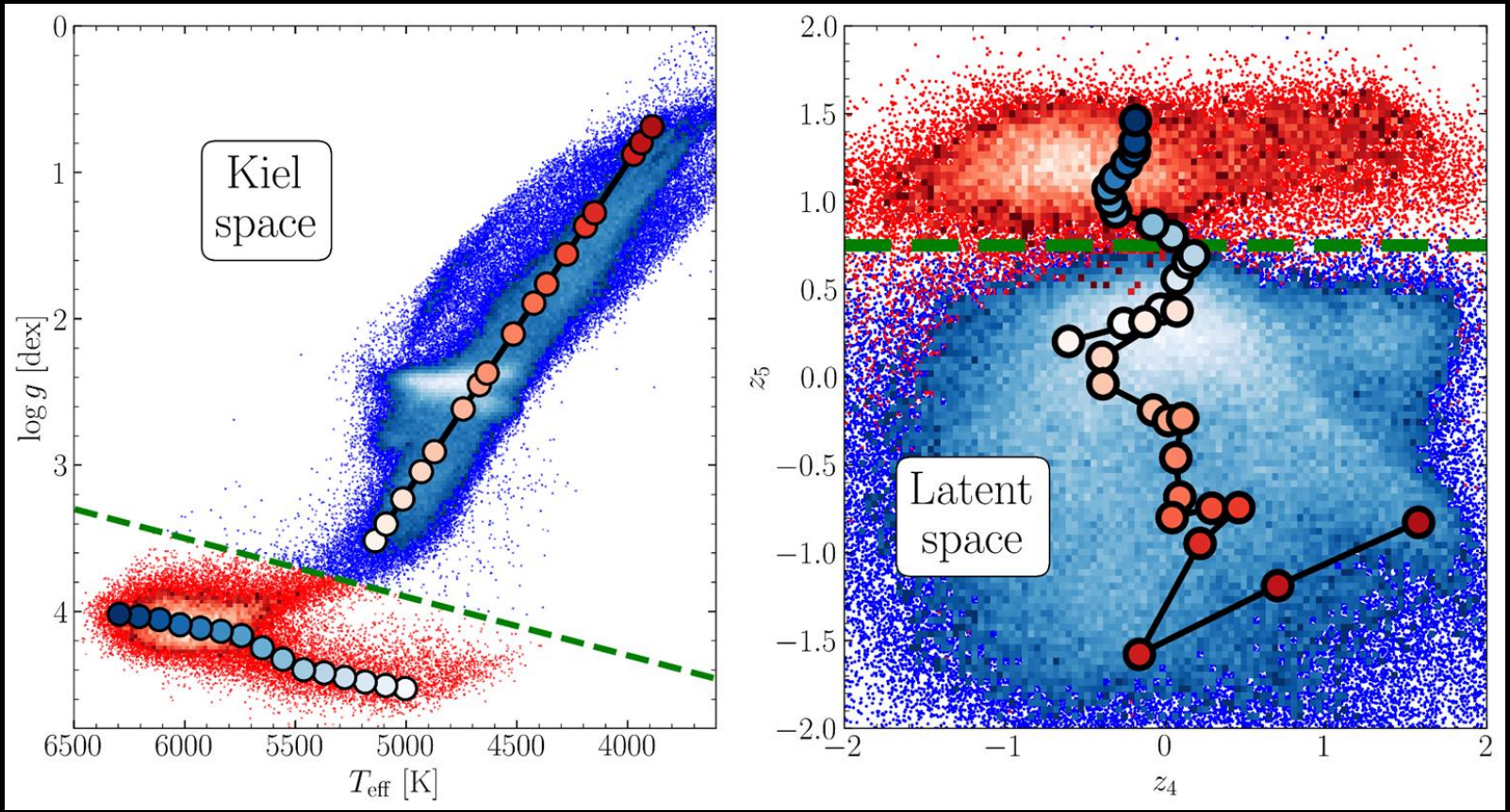
The latent space has learned to classify MS/GB stars



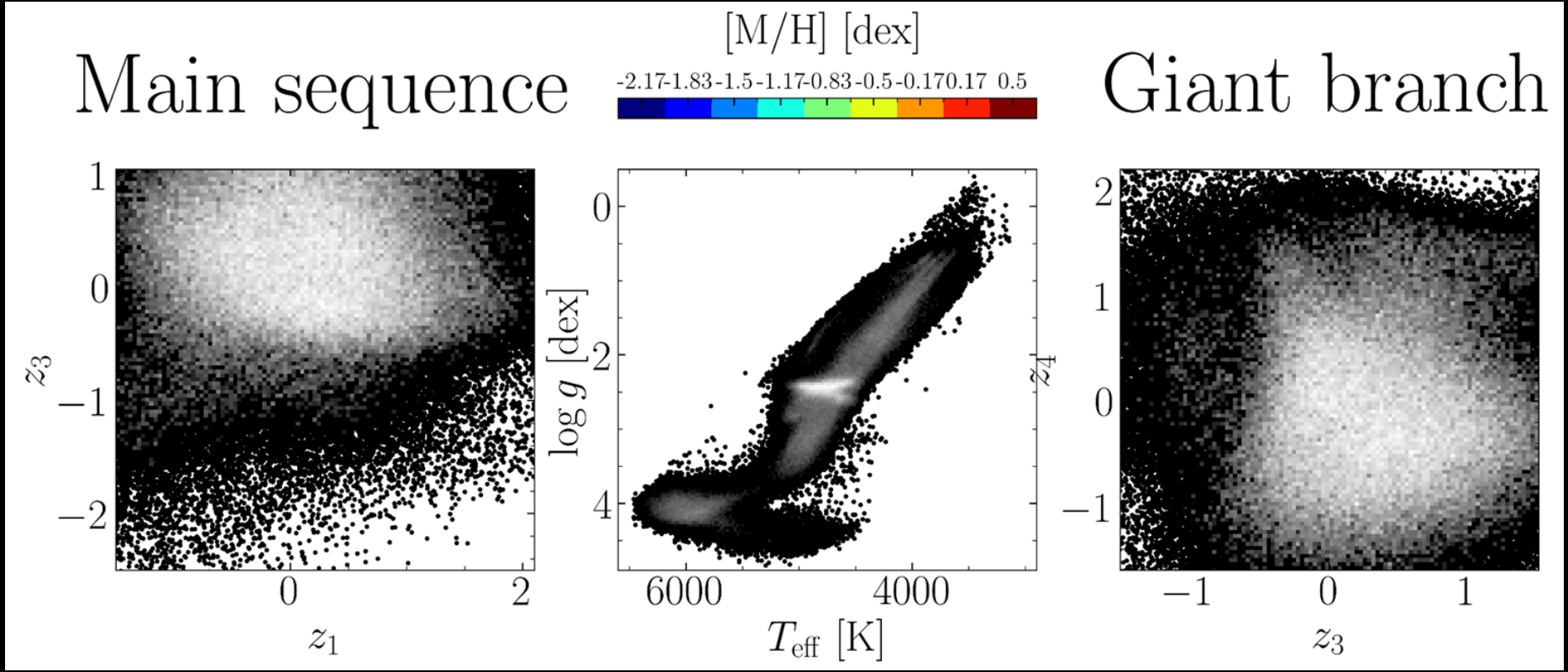
MS/GB stellar ‘evolutionary’ tracks



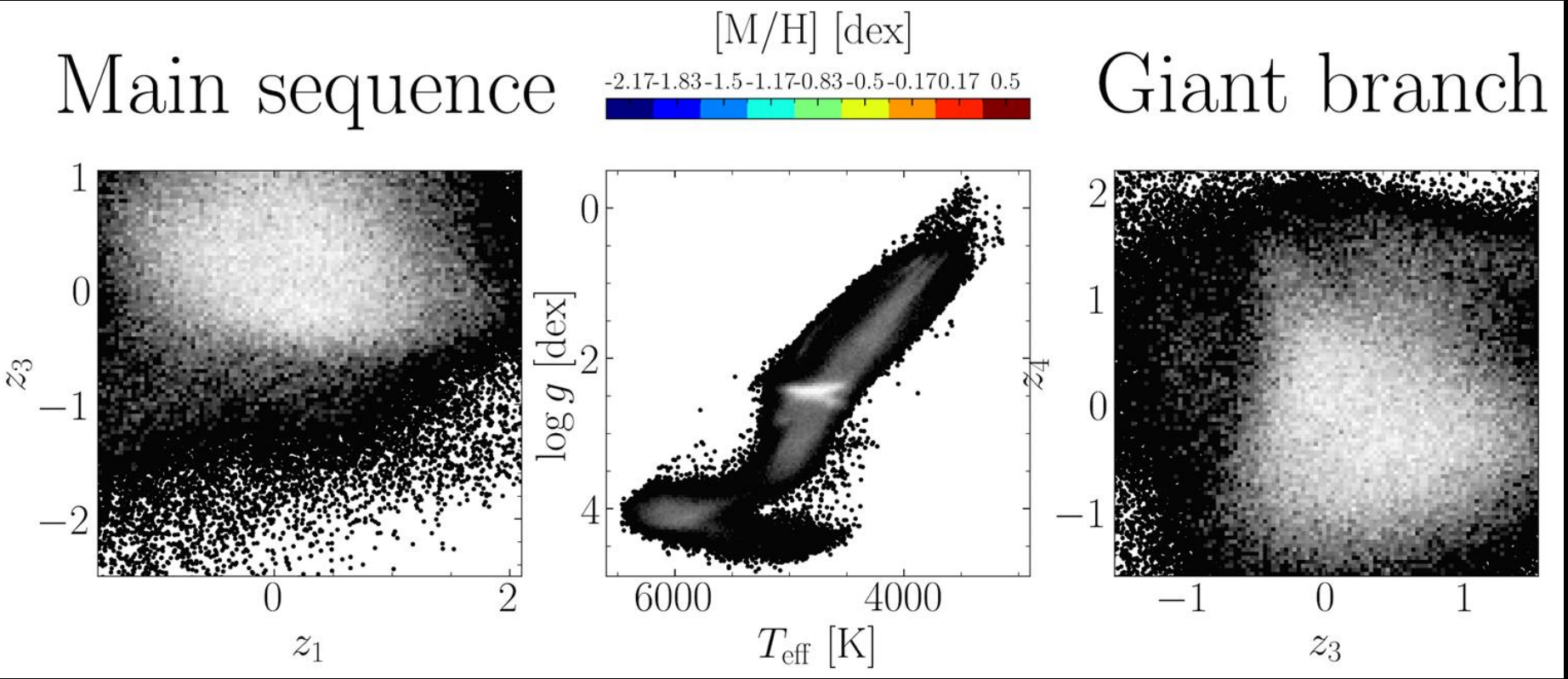
The latent space understands ‘evolution’ along the MS/GB



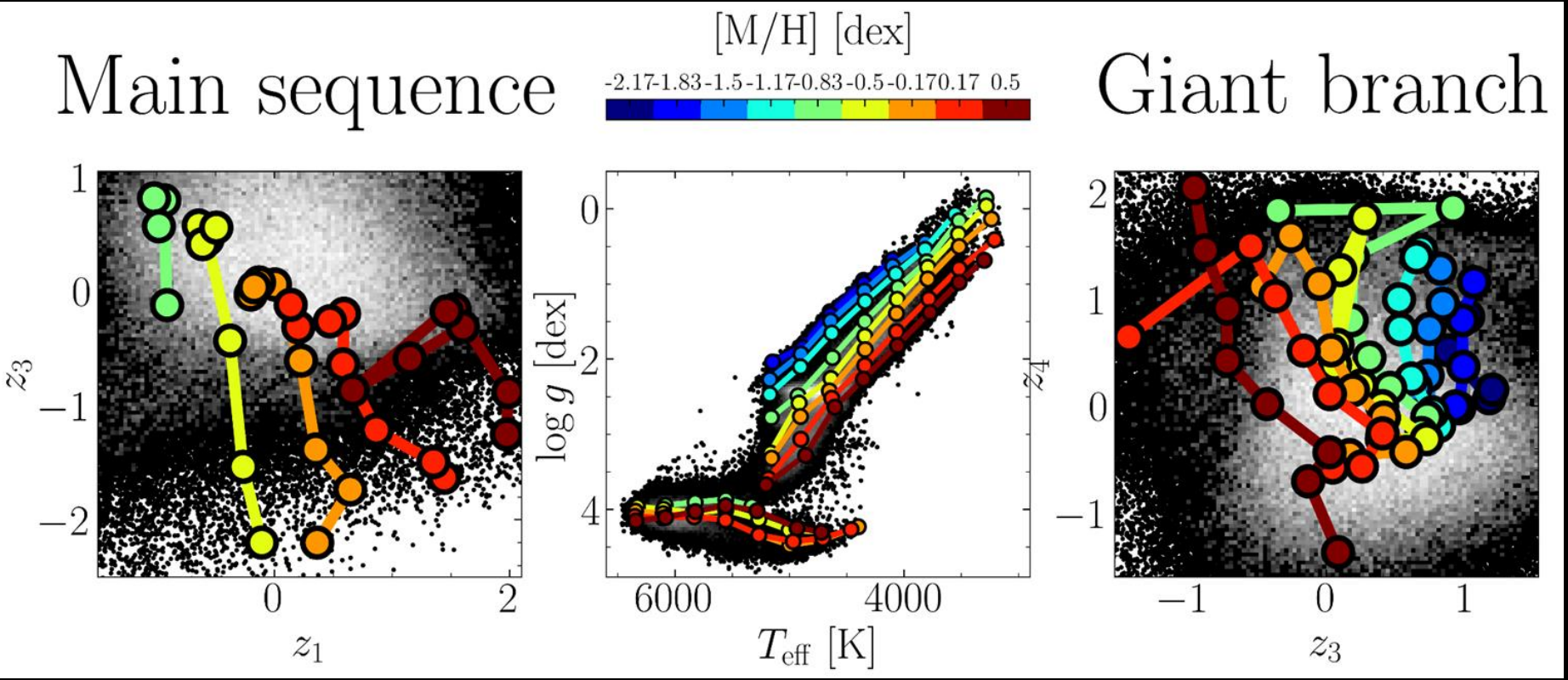
What about metallicity?



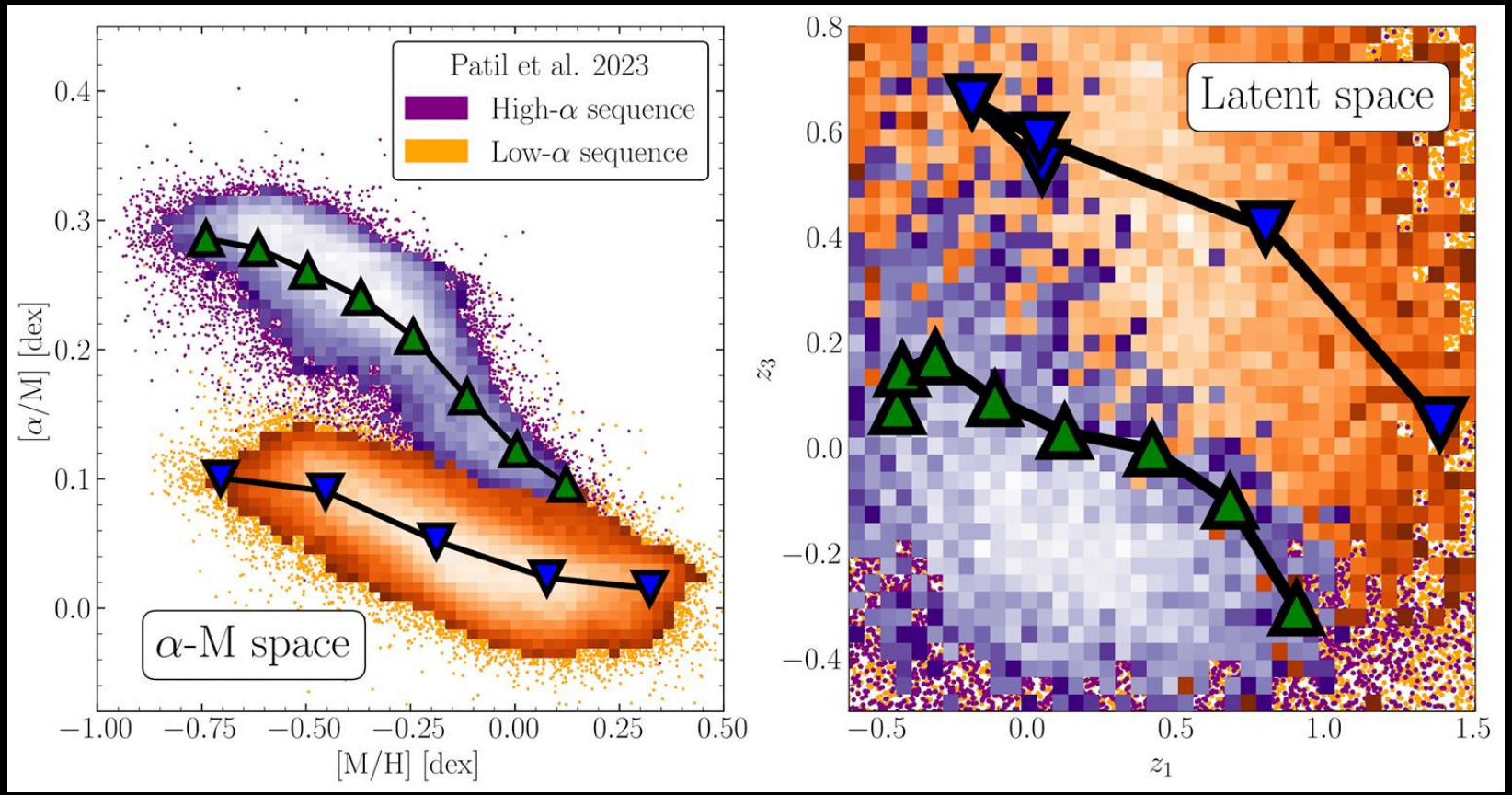
Vary [M/H] along the MS and GB



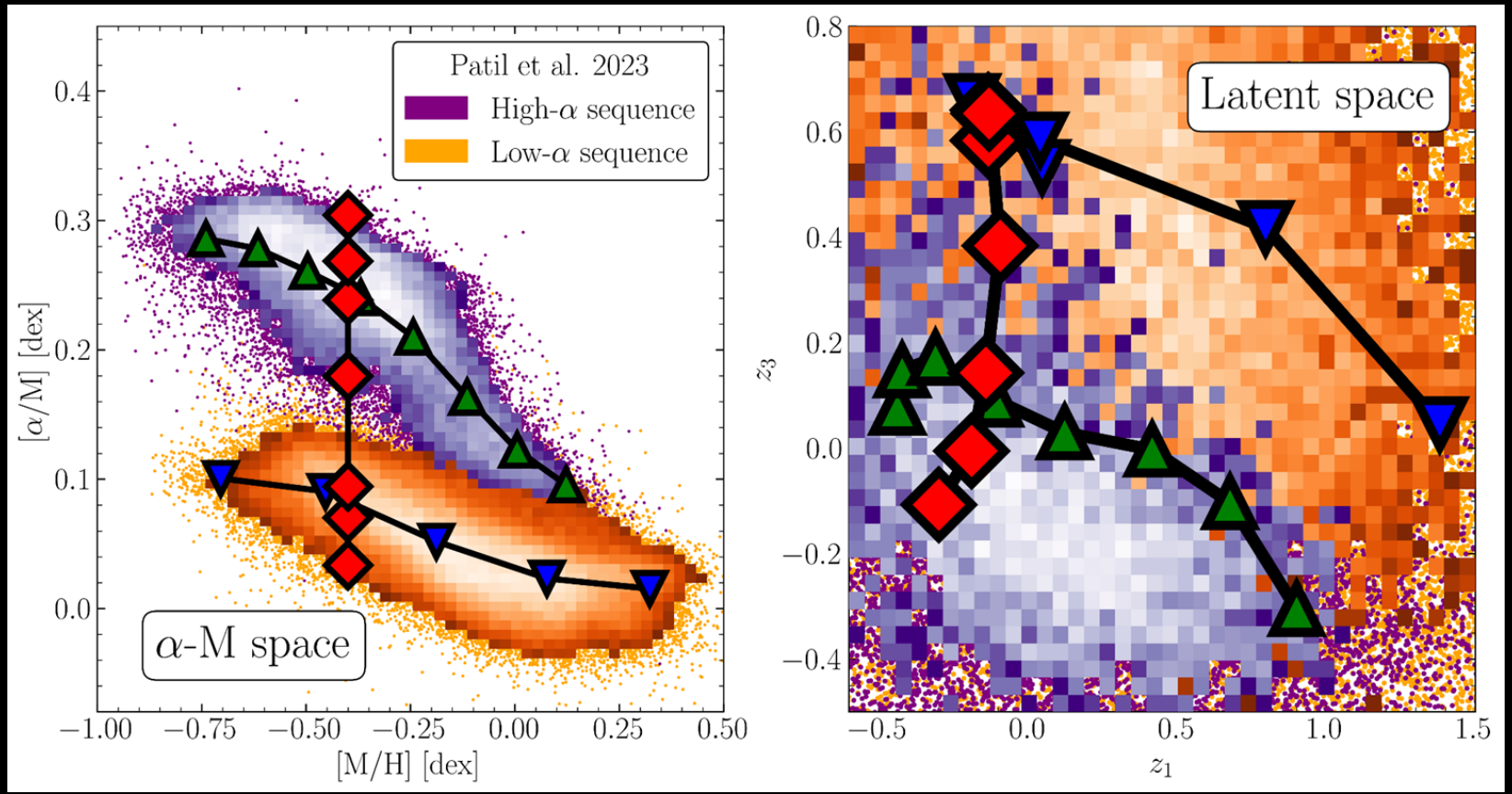
The latent space has also learned a metallicity ‘gradient’



In brief, main result of Laroche & Speagle (2024): Stellar label independent evidence for α-information in Gaia XP spectra



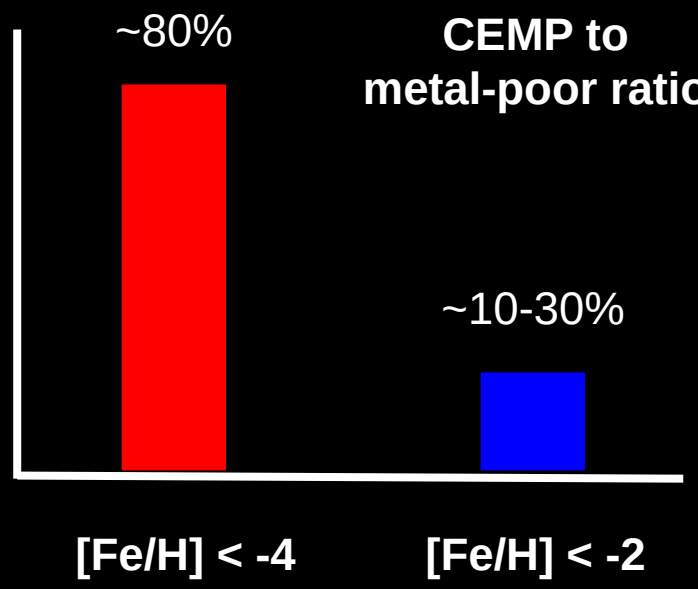
In brief, main result of Laroche & Speagle (2024): Stellar label independent evidence for α-information in Gaia XP spectra



Stellar label independent modeling:
A promising tool for discovering
rare stellar populations in large-scale
spectroscopic surveys

Carbon-enhanced metal poor stars

- Metal poor stars serve as ‘fossil records’ of the early Universe
- Metal-poor star surveys have found a counter-intuitive chemical peculiarity:
  - Metal poor stars with carbon enhancement [C/Fe] > +0.7
  - Referred to as CEMPs (carbon-enhanced metal poor)
- The relative CEMP fraction is *inversely correlated* with [Fe/H]

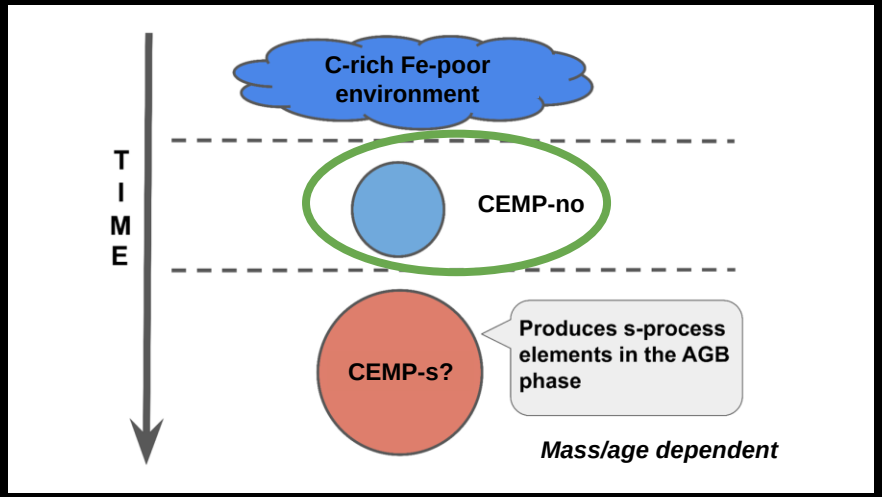


Lucatello+06, Lee+13, Placco+14, Yoon+18

Adapted from Goswami+21

CEMP formation (2 of many)

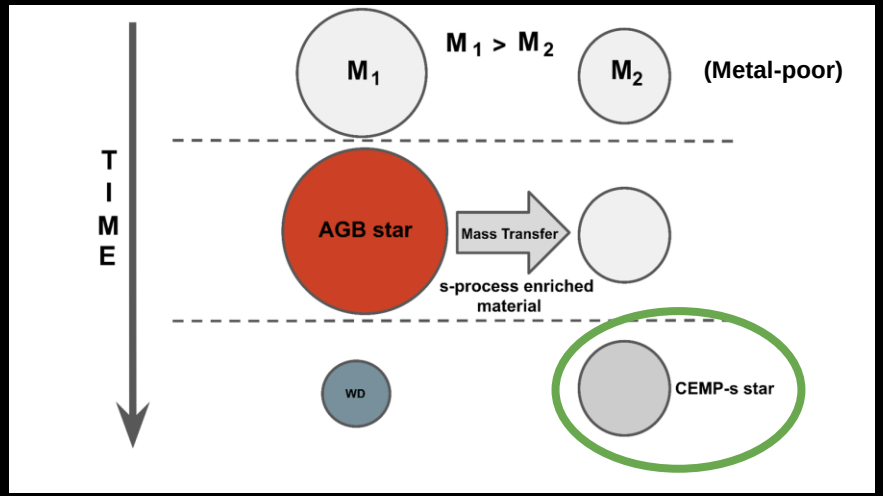
Nature



CEMP-no star

Single star evolution

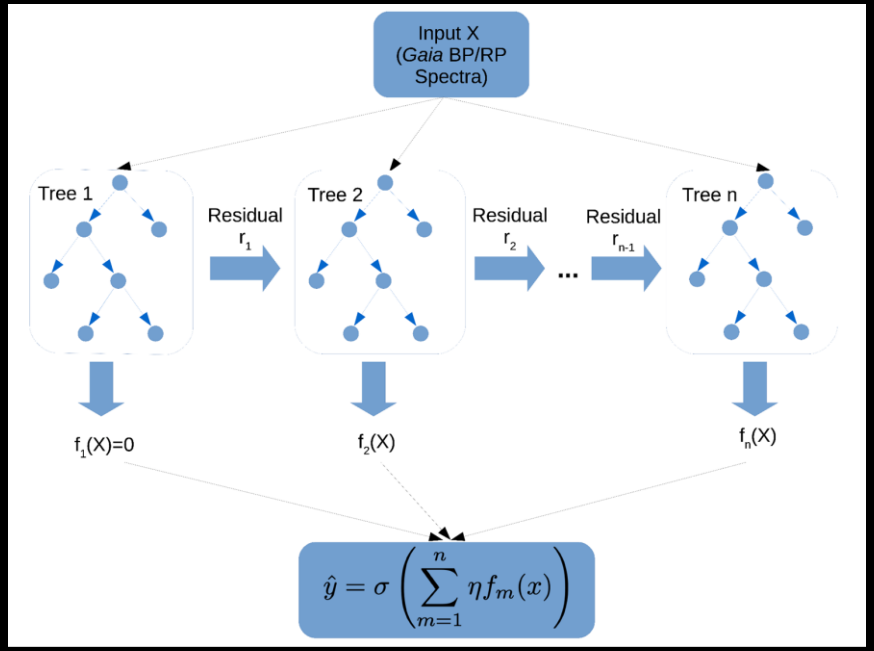
Nurture



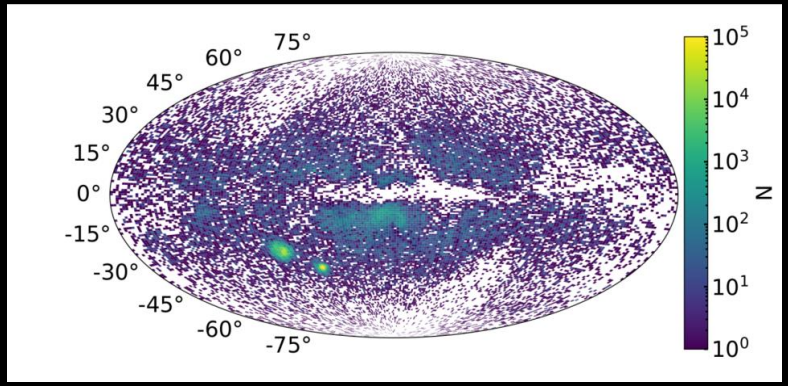
CEMP-s star

Binary evolution

Gaia XP CEMP candidates across the Milky Way (Lucey+23)



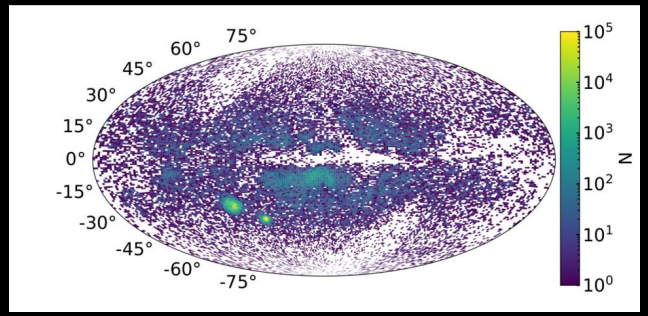
XGBoost trained on confirmed CEMPs



Largest all-sky CEMP candidate catalog to date (58,872 candidates)

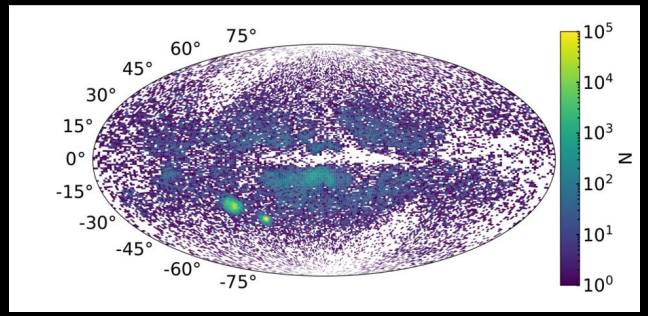
What lurks within the Lucey CEMP candidate sample?

CEMP candidates



What lurks within the Lucey CEMP candidate sample?

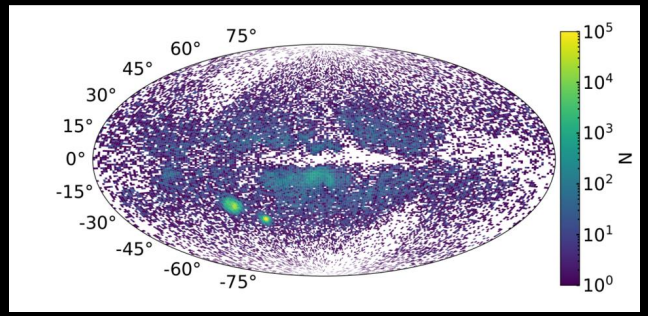
CEMP candidates



Train an ‘expert’ model to learn a latent representation

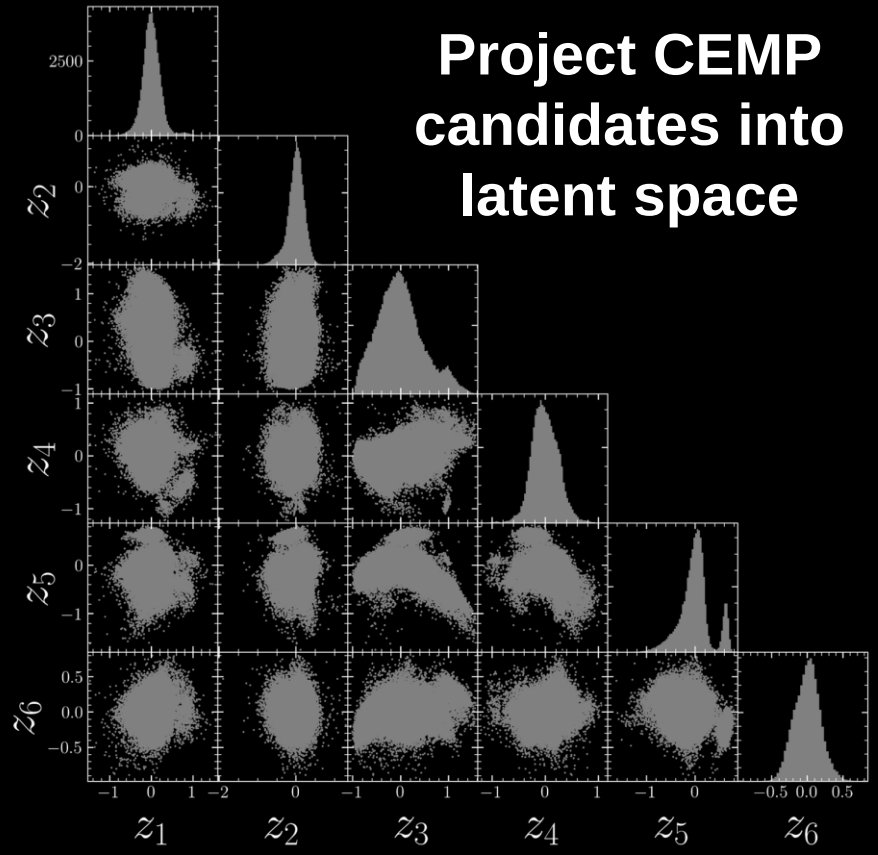
What lurks within the Lucey CEMP candidate sample?

CEMP candidates



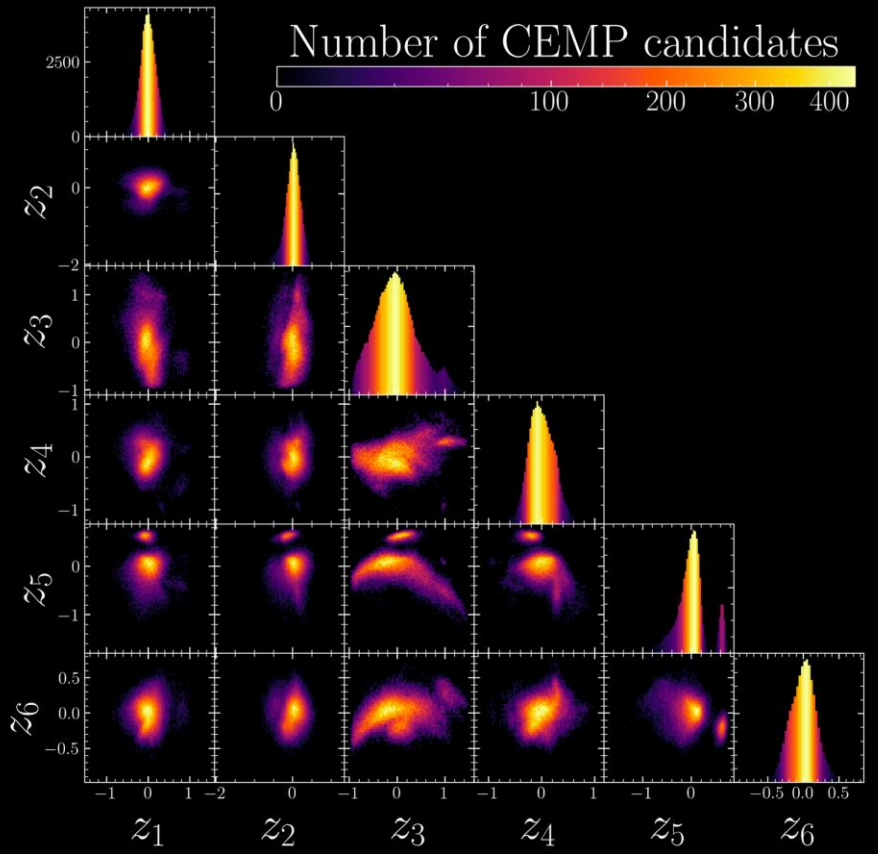
Train an ‘expert’ model to learn a latent representation

Project CEMP candidates into latent space



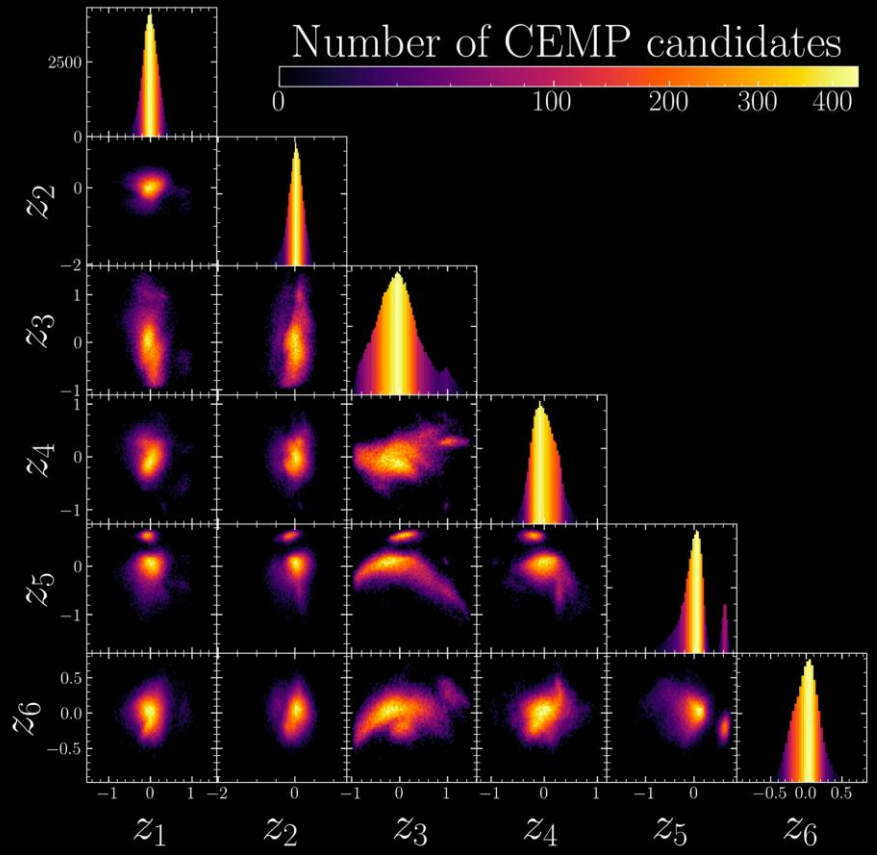
Lucey CEMP catalog latent space representation

- Fully data-driven low-dimensional representation
- Clear that the latent space is ‘structured’
  - Non-gaussian
  - Latent ‘islands’
- Already suggests that the Lucey catalog contains multiple sub-populations



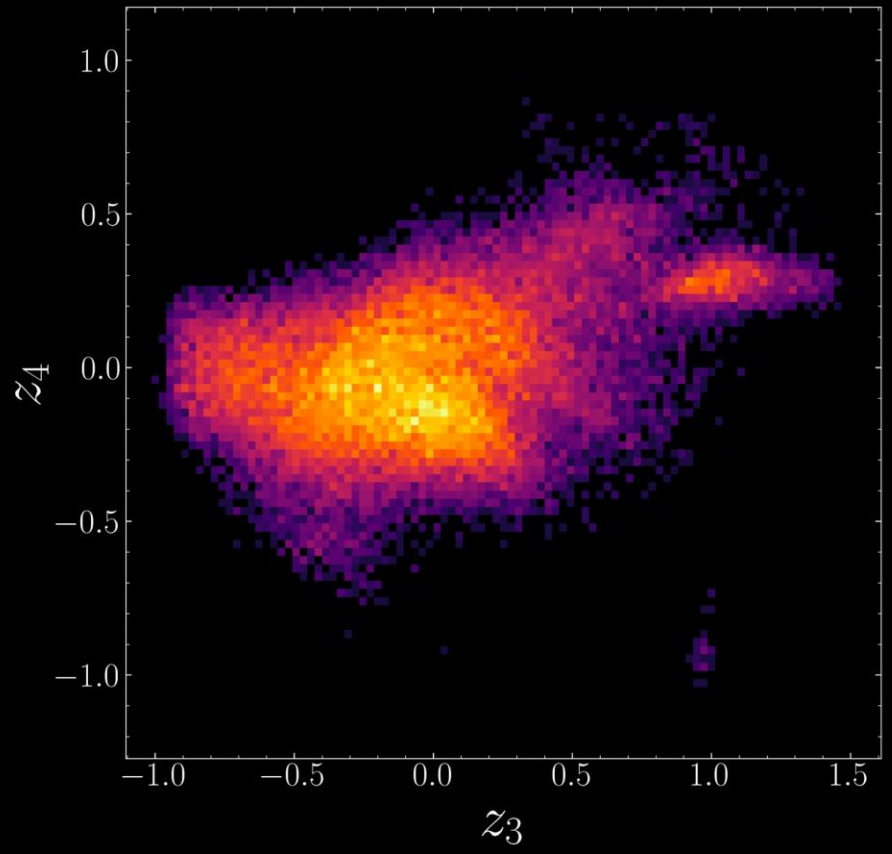
Lucey CEMP catalog latent density distribution

- Can we separate sub-populations in the latent space?
- What are structured low-dimensional data representations good for?
    - Clustering algorithms!



Discovering hidden sub-populations in the Lucey CEMP catalog

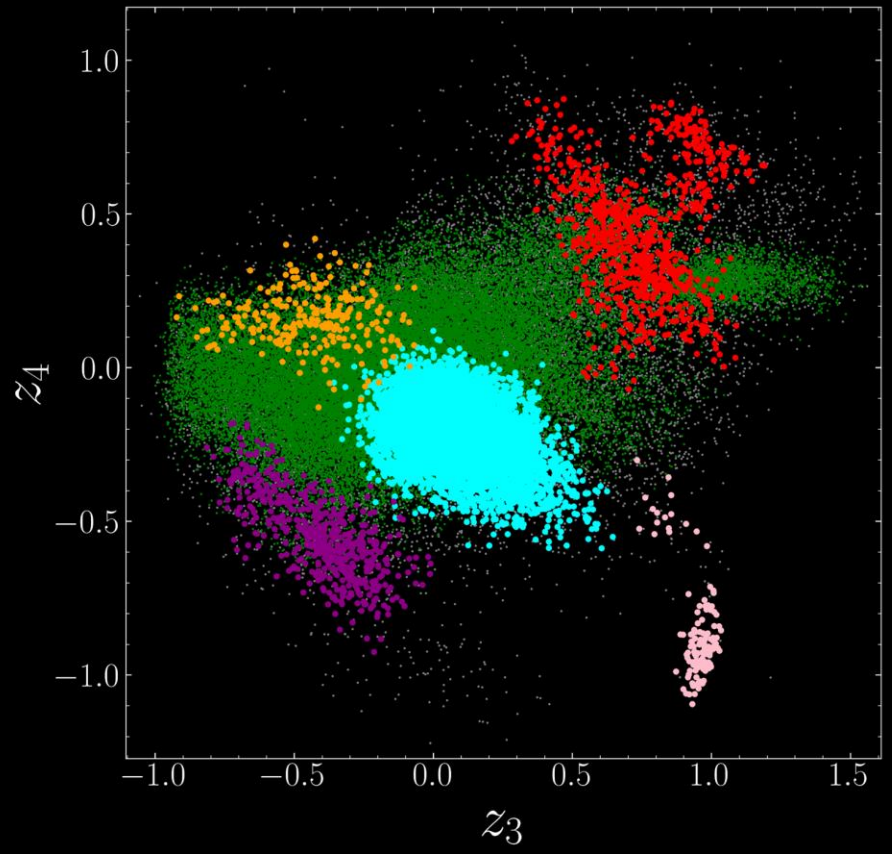
- Apply HDBSCAN (a popular clustering algorithm) to Lucey CEMP catalog latent vectors



2D marginal distribution visualized, but clustering is based on the entire latent space

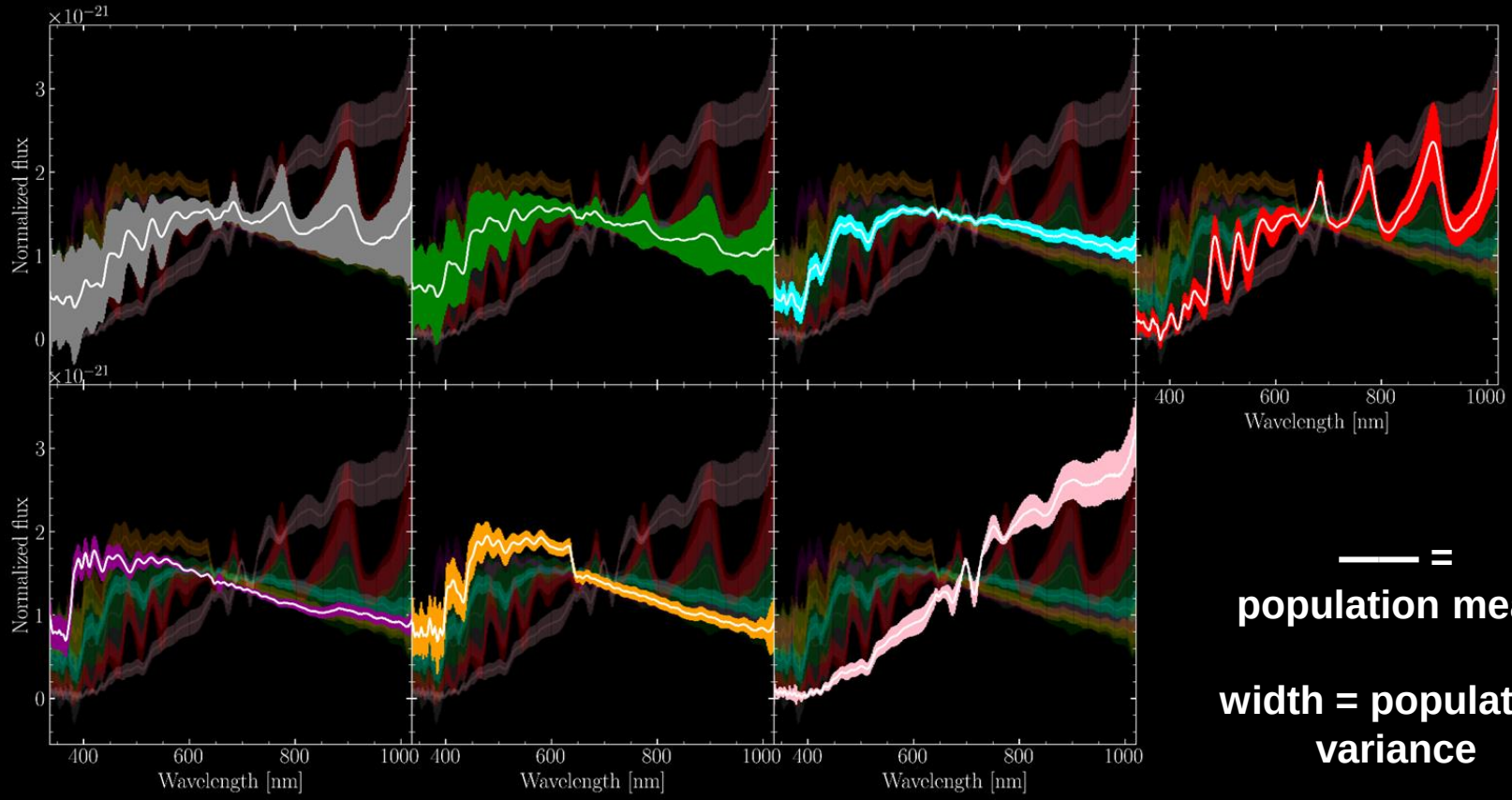
Discovering hidden sub-populations in the Lucey CEMP catalog

- Apply HDBSCAN (a popular clustering algorithm) to Lucey CEMP catalog latent vectors
- Discover several (7) populations
- Are these populations truly distinct?
  - Spectroscopically?
  - Photometrically?
  - Spatially?



2D marginal distribution visualized, but clustering is based on the entire latent space

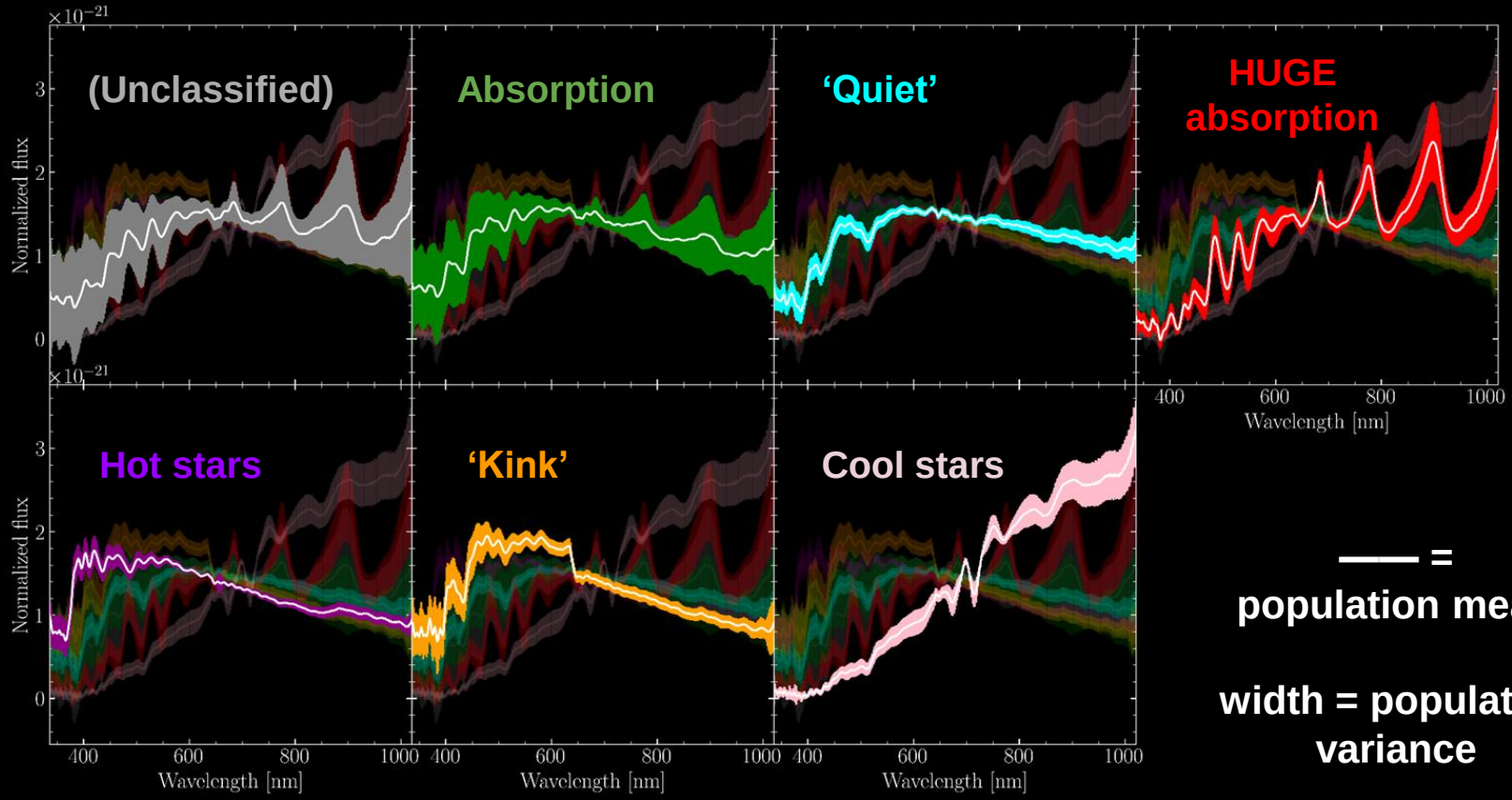
CEMP sub-populations are spectroscopically distinct



—— = population mean

width = population variance

CEMP sub-populations are spectroscopically distinct

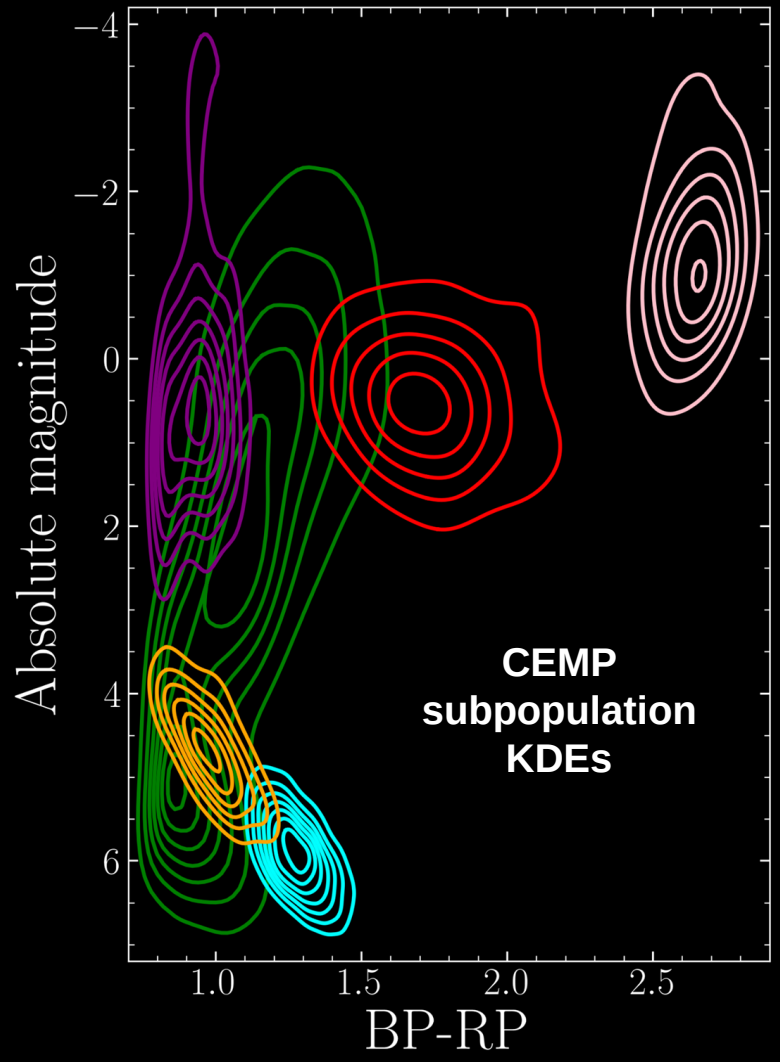


—— = population mean

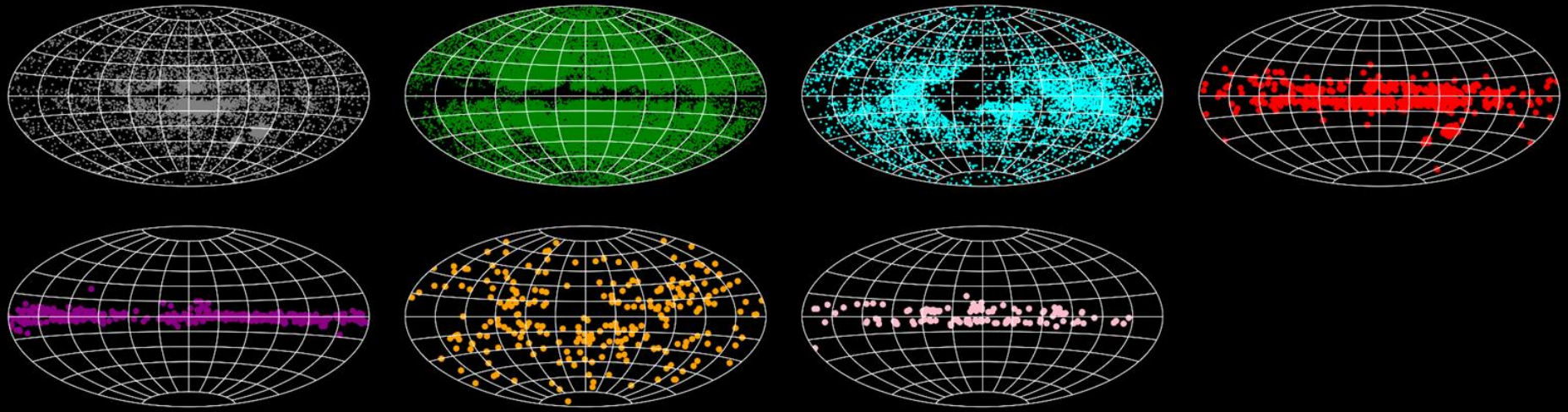
width = population variance

CEMP sub-populations are less distinct photometrically

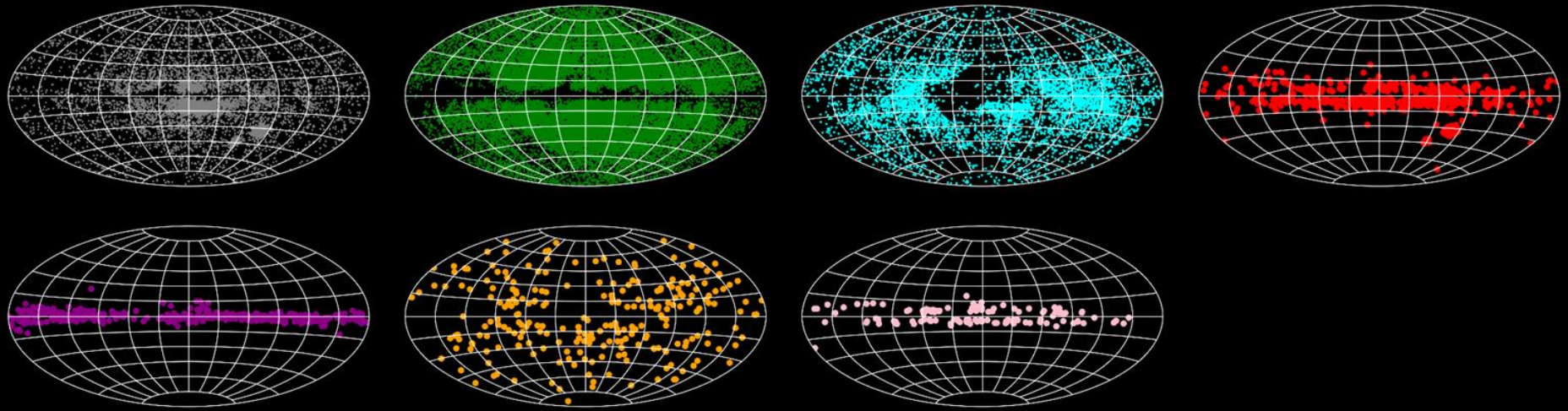
- Most populations are overlapping in the Gaia HRD
- Highlights the strength of the Gaia XP spectra
  - Numerous
  - Rich in stellar information
    - Despite low resolution
- A Gaia photometry search could miss sub-populations which are not well separated photometrically



CEMP sub-populations are spatially distinct



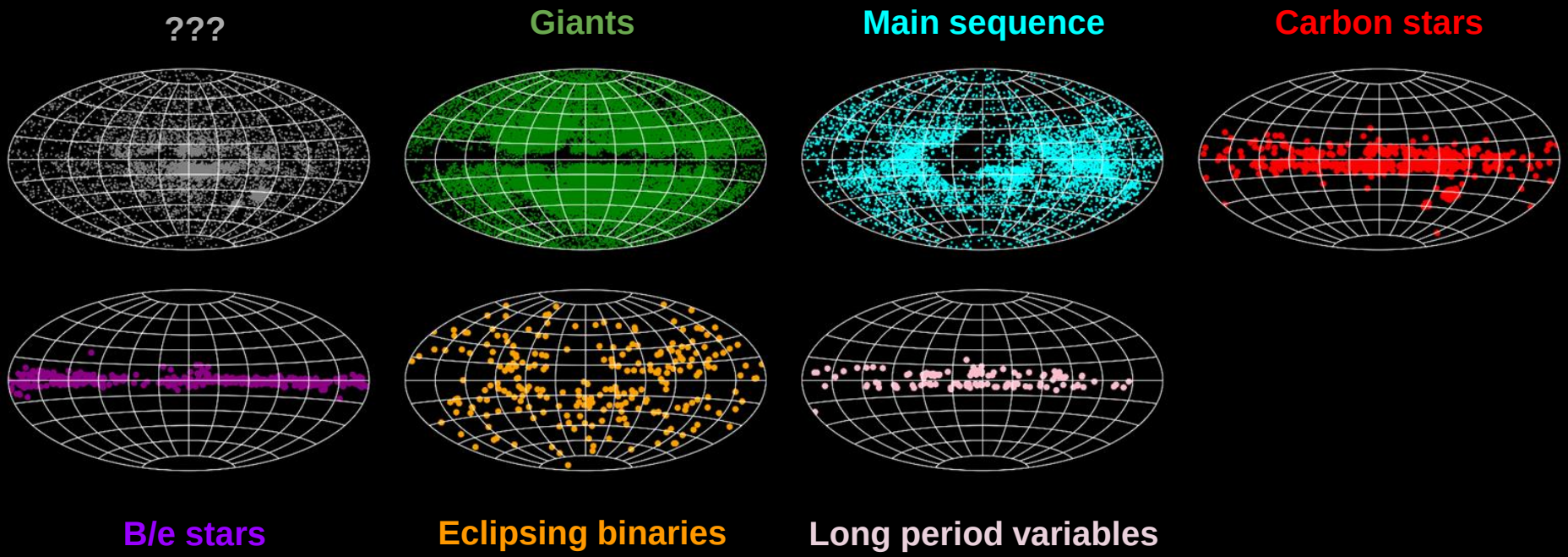
CEMP sub-populations are spatially distinct



Stellar label independent model has never 'seen' distance, position or brightness only (flux normalized) spectra

Additional evidence that these sub-populations are truly distinct…different CEMP formation channels?

Preliminary characterizations with ‘majority vote’ (SIMBAD cross-match)



Next: false positive identification, CEMP binarity estimates, formation channels…

(arXiv:2404.07316)

Towards the discovery of rare stellar populations in the Gaia XP spectra with stellar label independent modeling

- The **stellar labels gap** negatively impacts stellar label dependent model searches for rare stellar populations in large scale surveys

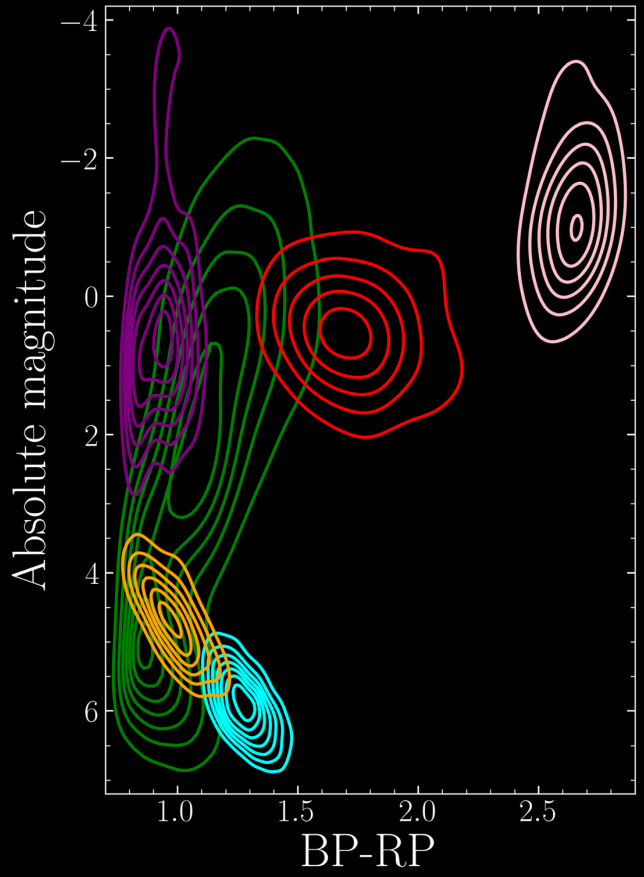
- Our **stellar label independent model** can discover hidden populations in large-scale surveys
  - Such as **carbon-enhanced metal poor stars**
  - Future work:
    - CEMP binary fraction estimates
    - Constrain formation channels

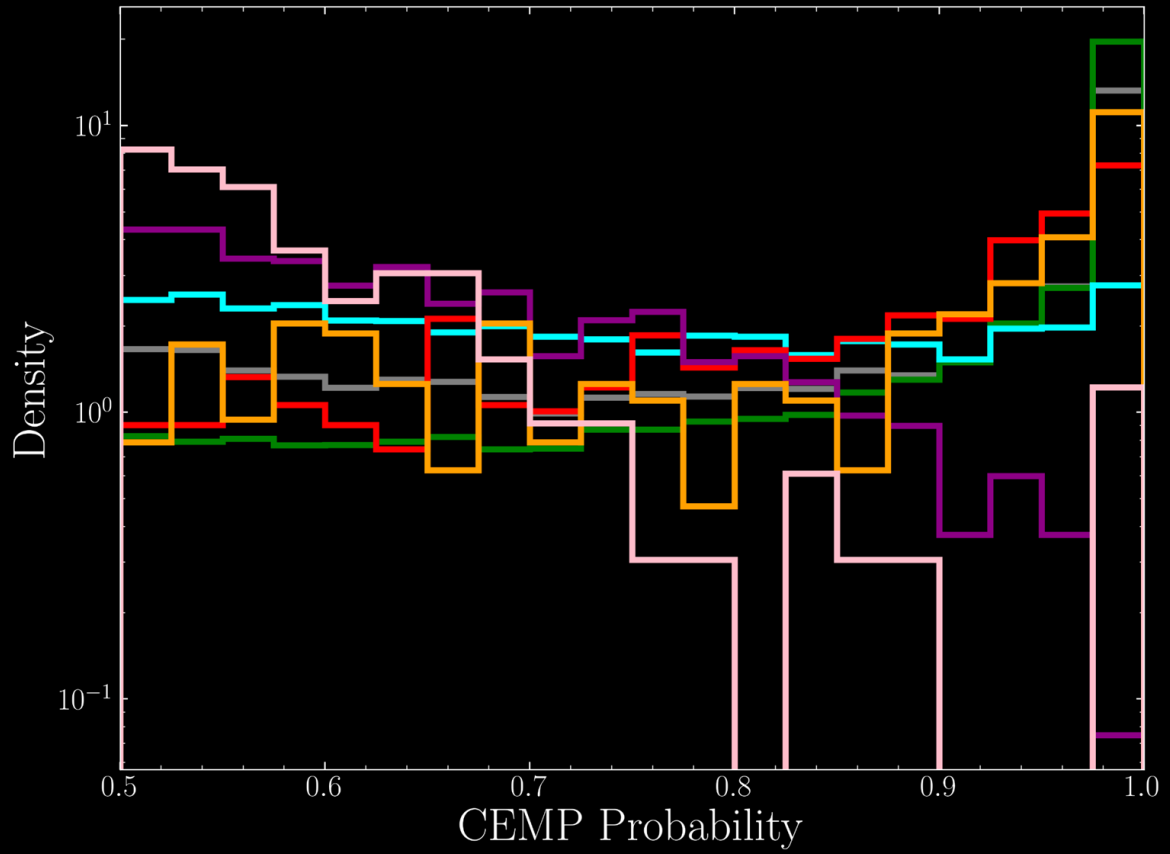
- Do **you** have Gaia XP spectra you are trying to characterize?
- Or, do **you** have a completely different spectroscopic survey to analyze?

If so, **let's talk!**

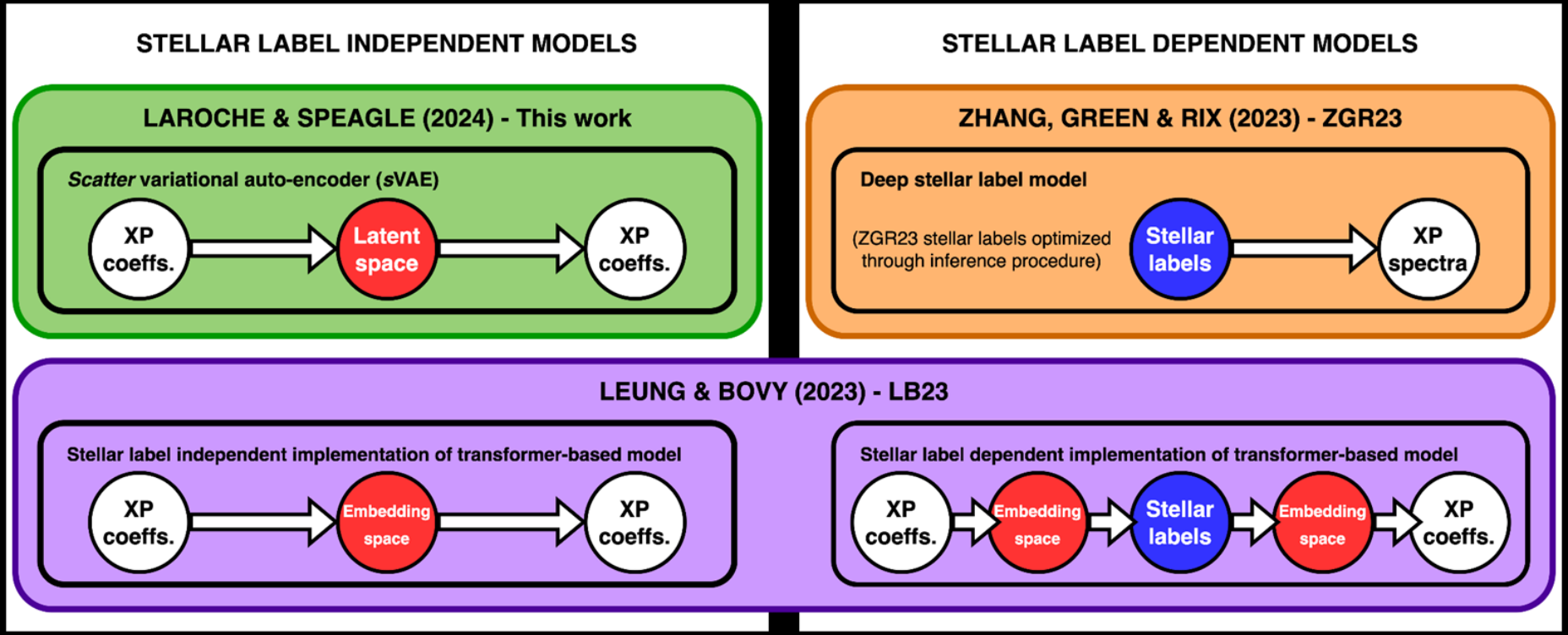
Backup Slides

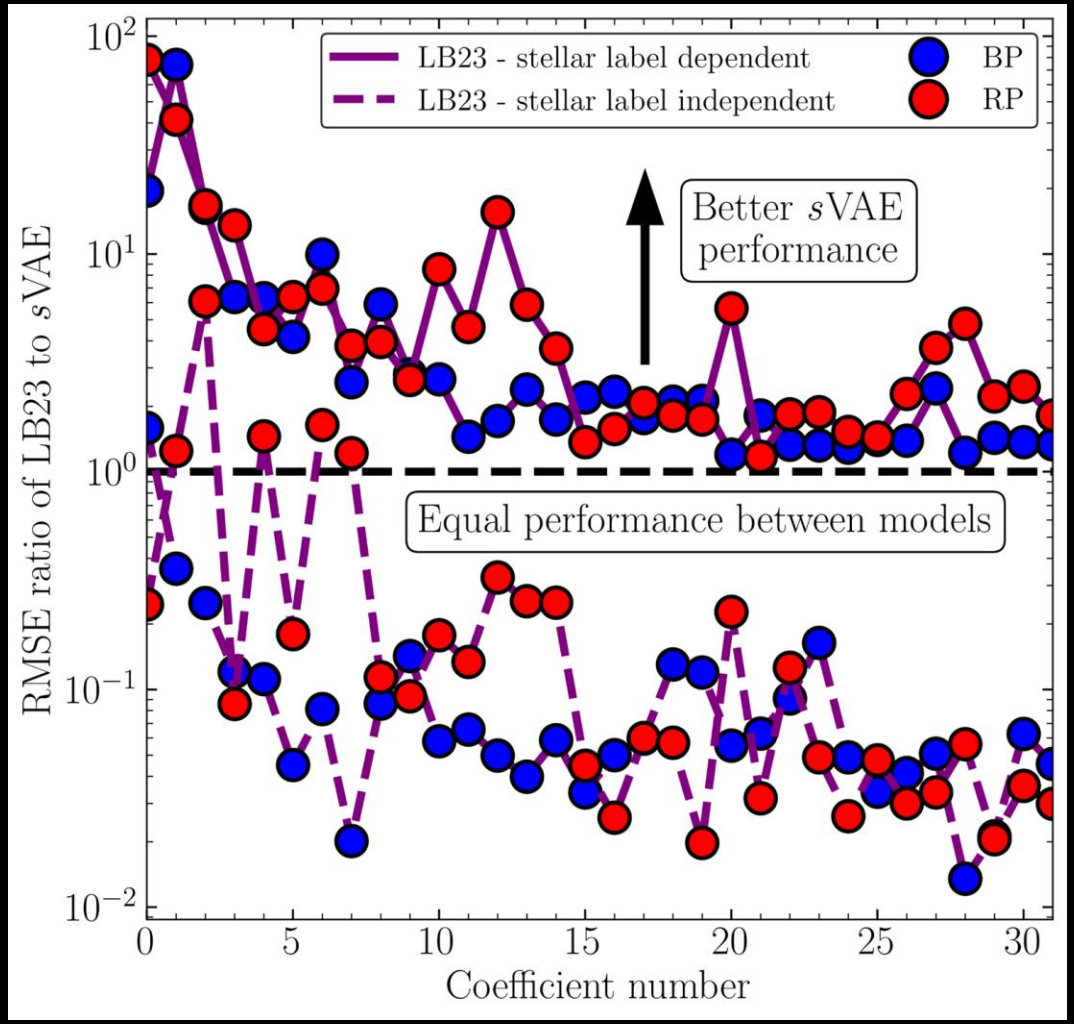
CEMP probability distributions across sub-populations





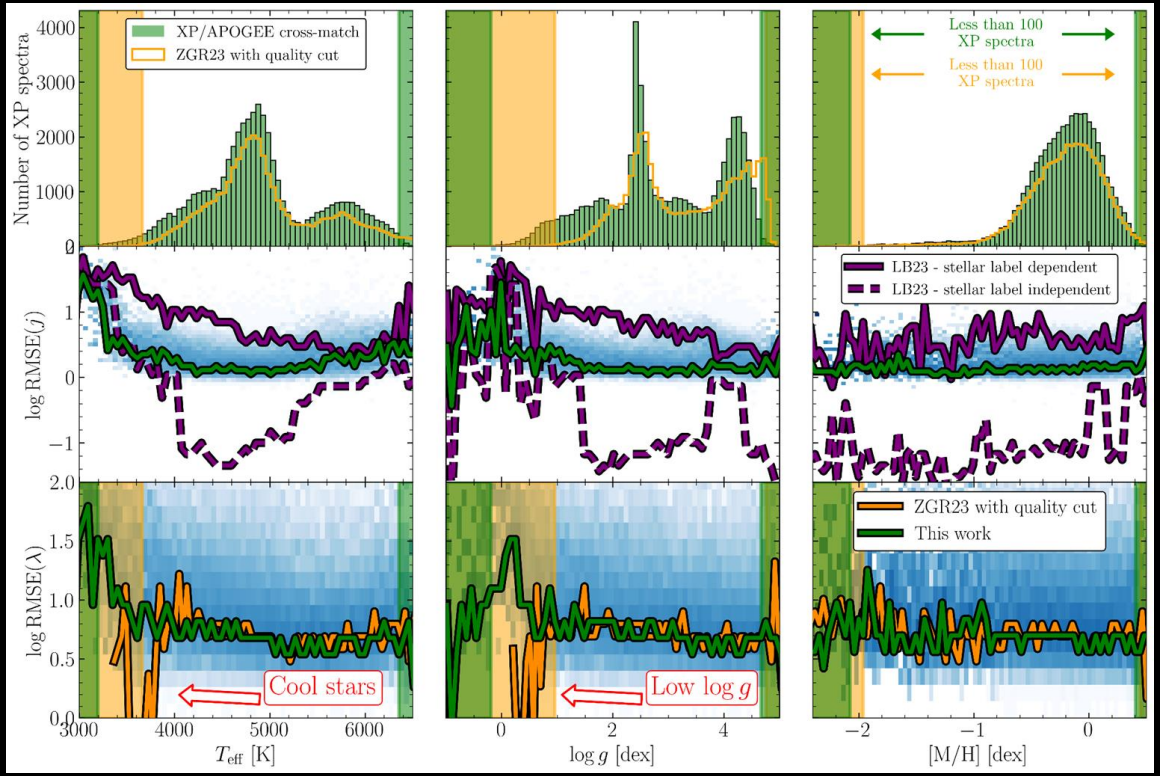
Stellar label independent and dependent models referenced in this work.
Our model is stellar label independent, whereas **ZGR23** is a stellar dependent model.
LB23 can do both (but their embedding space does not perform data compression)



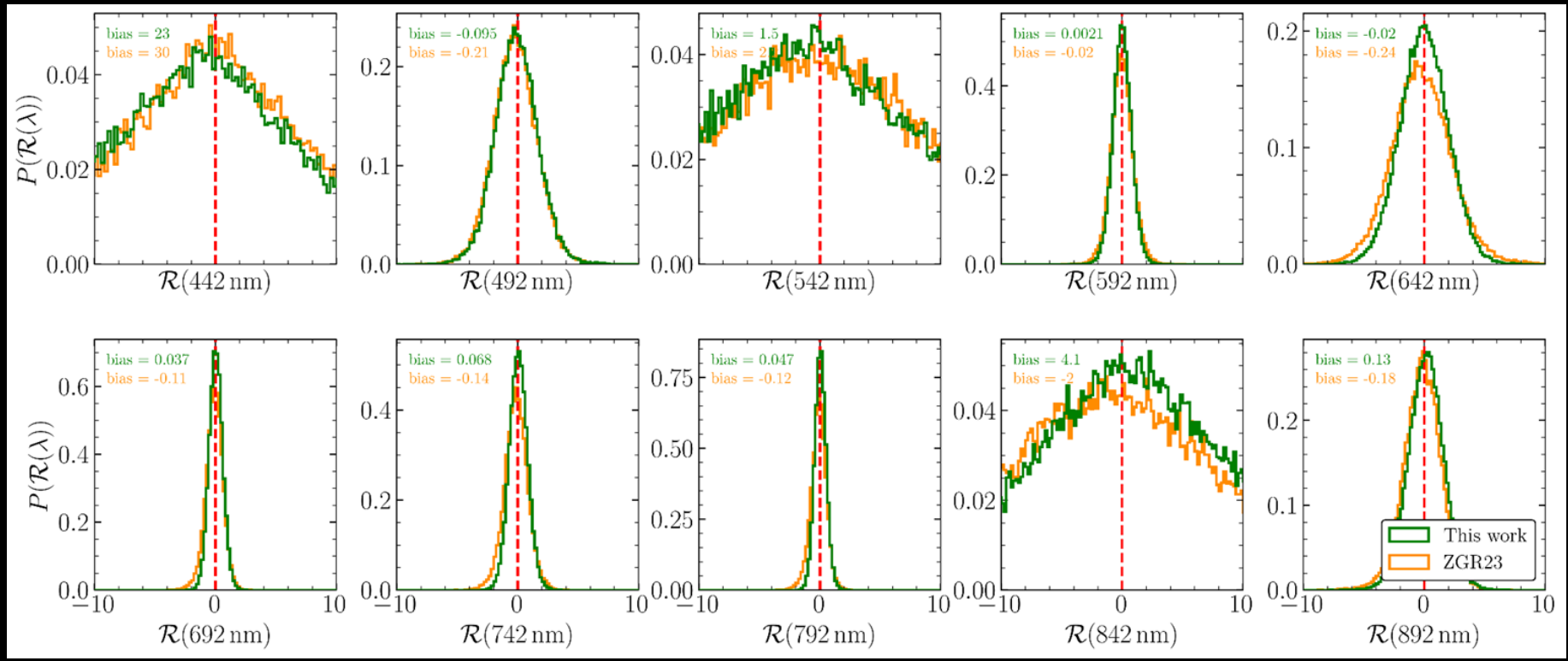


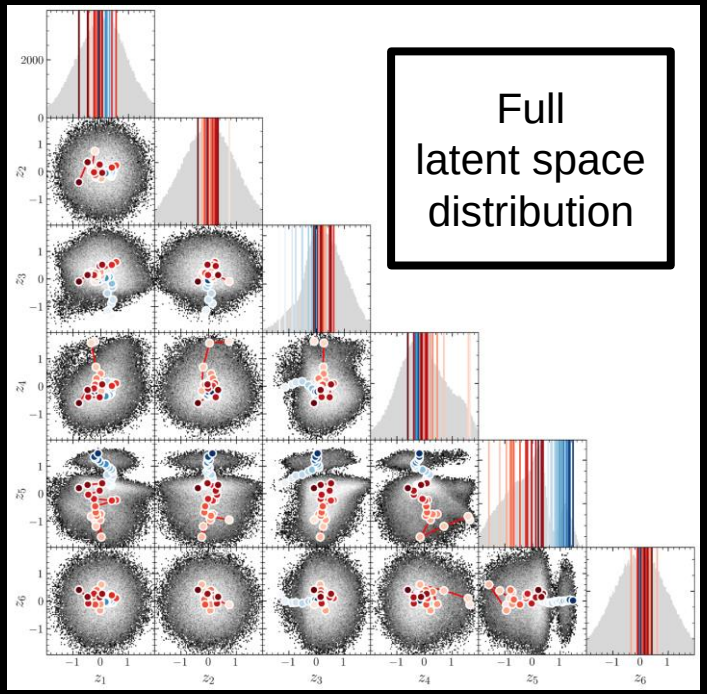
Comparison to **Leung & Bovy (2023)**, with two implementations of their model: stellar label independent and dependent

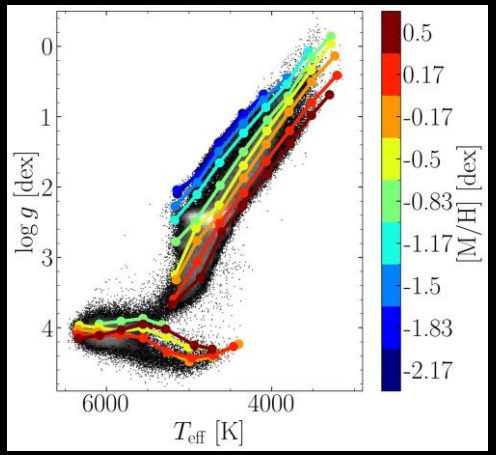
Model reconstruction errors in comparison to **LB23** **and** **ZGR23** *as a function of stellar labels*. Our stellar label independent model does not suffer from reduced stellar label coverage issues for both cool and low surface gravity stars

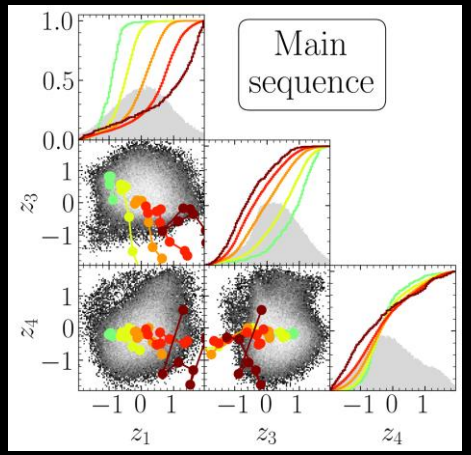


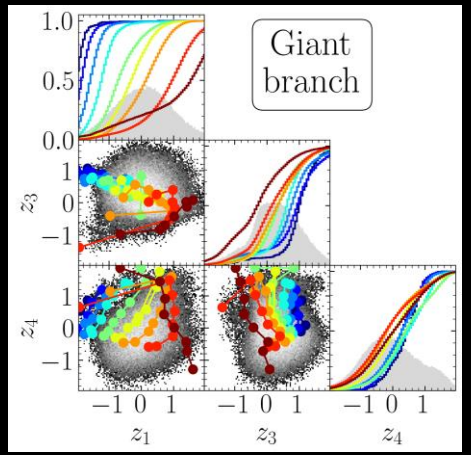
Model reconstruction errors at specific wavelengths across the Gaia XP wavelength range. Our **stellar label dependent model** is less biased than **ZGR23** from roughly 450 to 850 nm





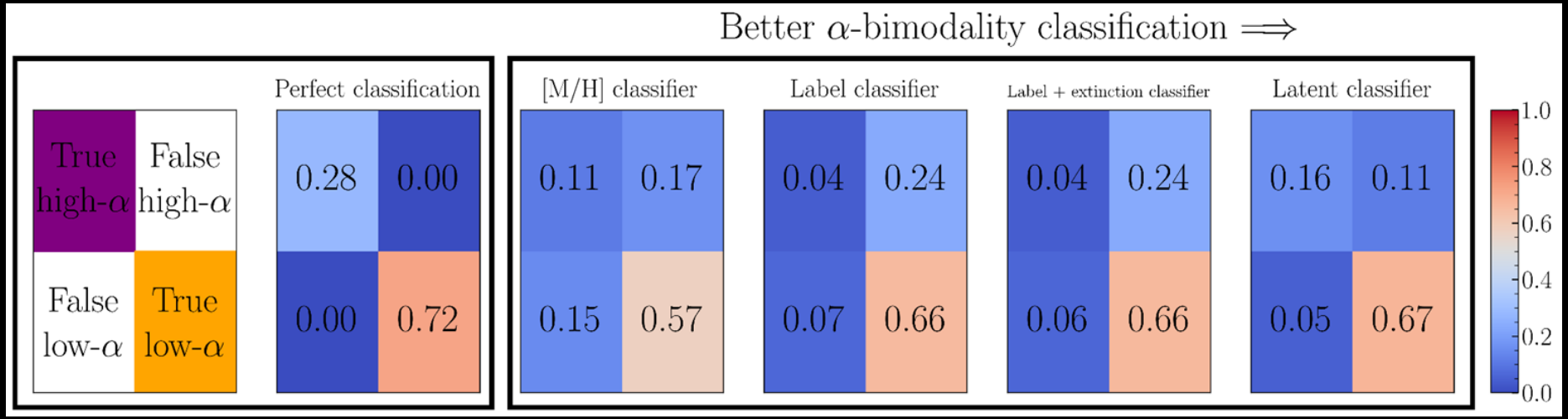






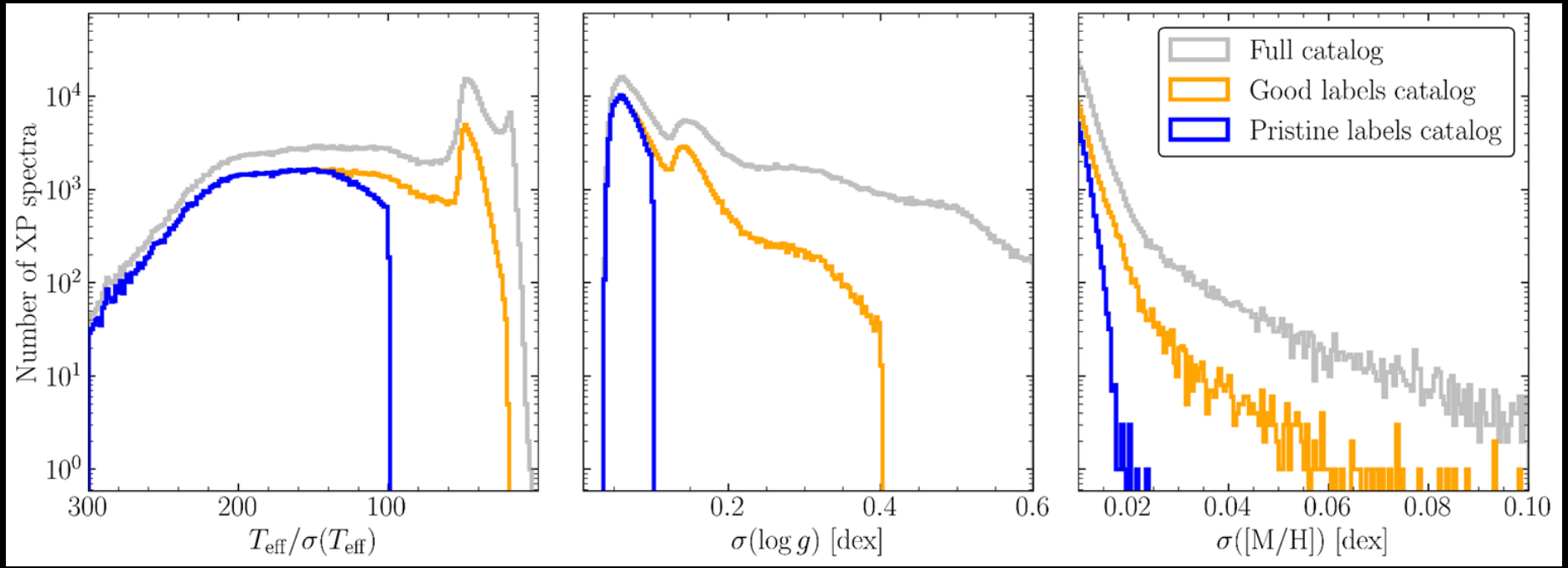
Astrophysical information is not encoded into a single latent dimension. Rather, the information is shared across the entire latent space.

We train our latent space to classify the α-bimodality. In comparison to several stellar label based classifiers (not including [α/M]), our latent space achieves better classification.

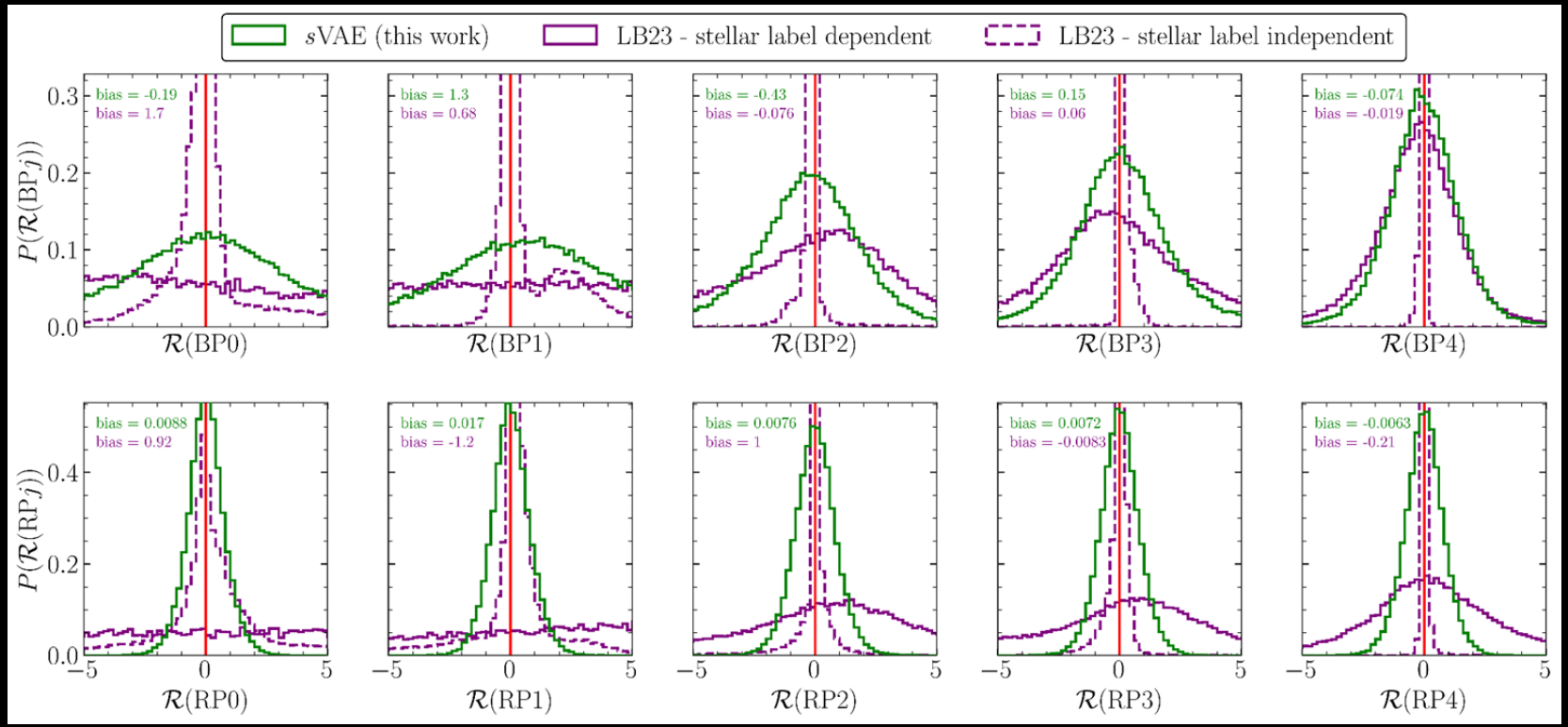


This simultaneously demonstrates that (i) our stellar label independent model has learned genuine α-information and (ii) the Gaia XP spectra contain α-information (without relying on stellar label correlations)

The stellar label error distributions for the catalogs we use in this work.
Full = no cuts, **good labels = some cuts**, **pristine labels = harsh cuts**

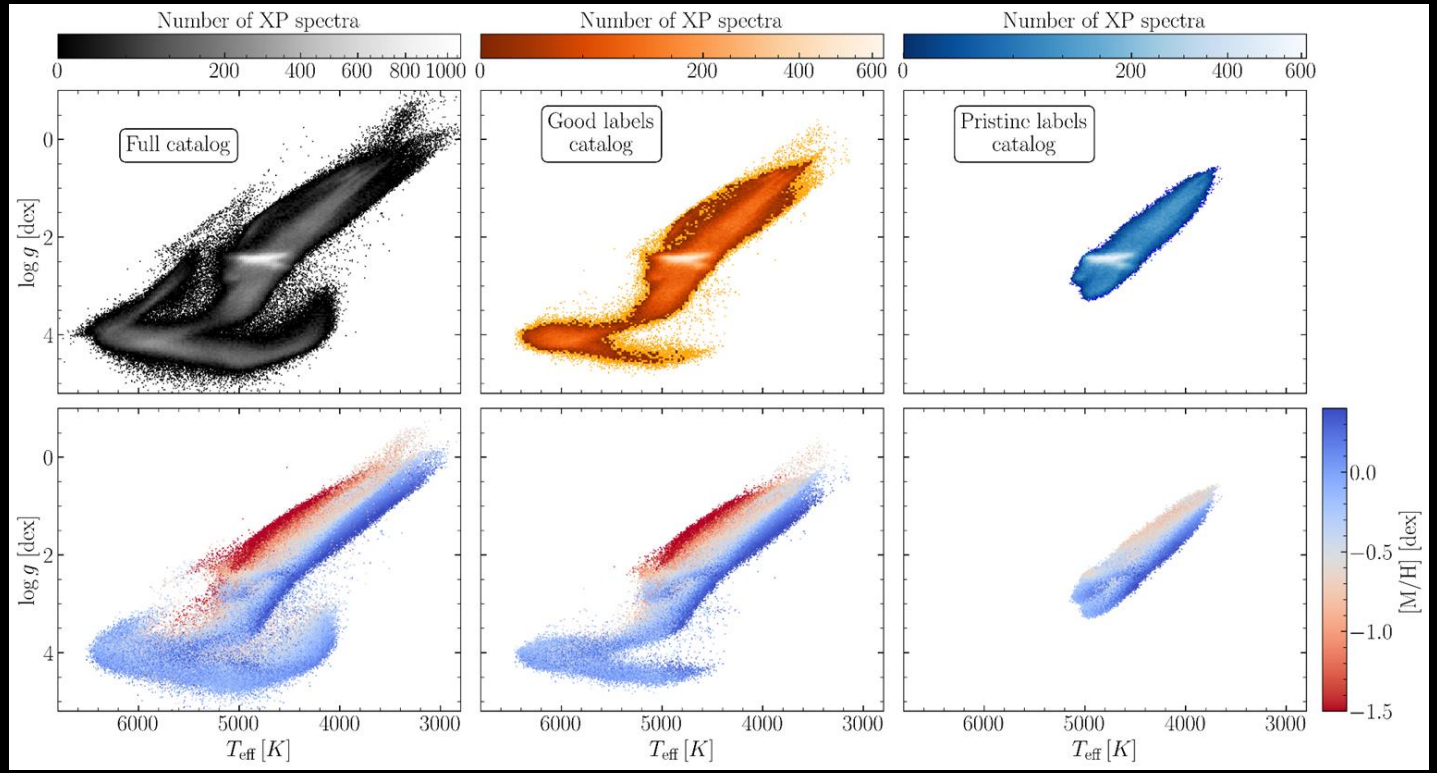


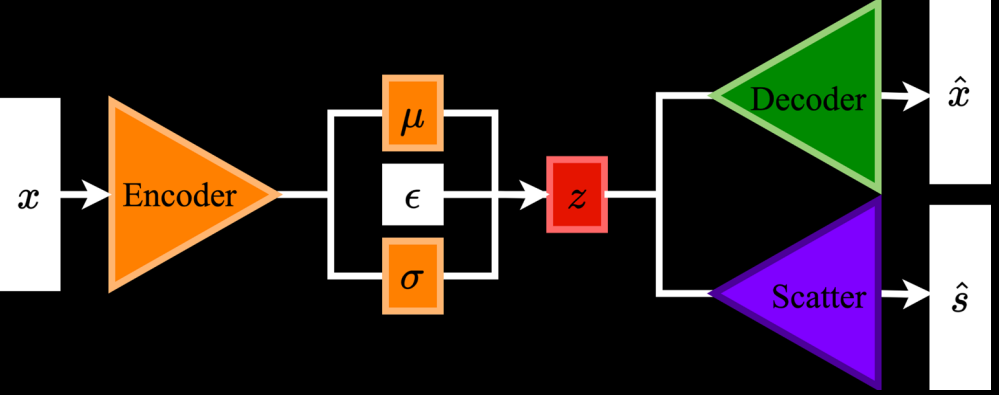
Model reconstruction errors over the first 5 BP/RP coefficients.
Our **stellar label dependent model** is more (less) accurate than the stellar label dependent (independent) model of **LB23**, the latter due to data compression



Catalogs used in this work, in Kiel space.

Full = no cuts, good labels = some cuts, pristine labels = harsh cuts





$$\mathcal{L} = \tilde{\chi}^2(x, \hat{x}, \sigma, \hat{s}) + D_{\mathrm{KL}}(\mu, \sigma'). \qquad (1)$$

$$\tilde{\chi}^2 = \chi^2(x, \hat{x}, \sigma, \hat{s}) + P(\sigma, \hat{s}). \qquad (2)$$

$$D_{\mathrm{KL}}(\mu, \sigma') = \frac{1}{6N} \sum_{i=1}^{N} \sum_{j=1}^{6} \mu_i^2 + \sigma_i'^2 - \left(1 + \log \sigma_i'^2\right), \qquad (5)$$

Balancing act
between
latent space structure
and
reconstruction error

$$P(\sigma, \hat{s}) = \frac{1}{110N} \sum_{i=1}^{N} \sum_{j=1}^{110} \log\left(\sigma_{ij}^2 + \hat{s}_{ij}^2\right), \qquad (4)$$

$$\chi^2(x, \hat{x}, \sigma, \hat{s}) = \frac{1}{110N} \sum_{i=1}^{N} \sum_{j=1}^{110} \frac{(x_{ij} - \hat{x}_{ij})^2}{\sigma_{ij}^2 + \hat{s}_{ij}^2}, \qquad (3)$$