Can AI help users find data?
The experience with the ESO Science Archives

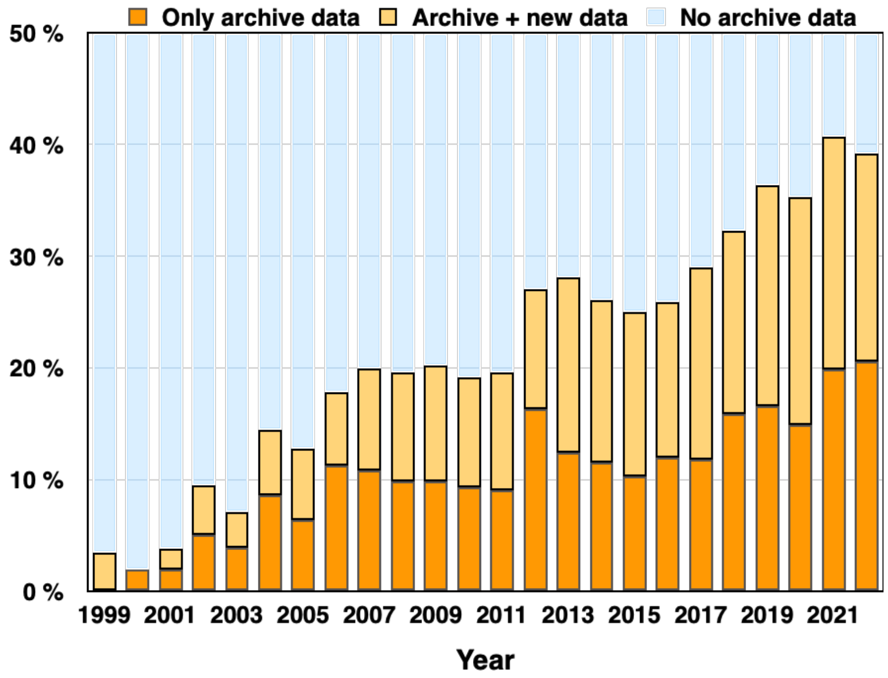
Martino Romaniello (ESO)
Nima Sedaghat , Felix Stoehr, Jon Carrick, FX Pineau
Vojtech Cvrcek, Wolfram Freudling, Pascal Ballester, Radim Sara

Public

CONTEXT: THE PROBLEM

The context:
The ESO Science Archives

- Very substantial contributors to ESO's science output: e.g. 40% of VLT publications
- Millions of science files, tens of millions of metadata items
- It is key to present and characterize the data in a language that speaks to users
  - Sky position
  - Instrument description (setup, …)
  - Data description (SNR, resolution, depth, …)



Source: telbib.eso.org

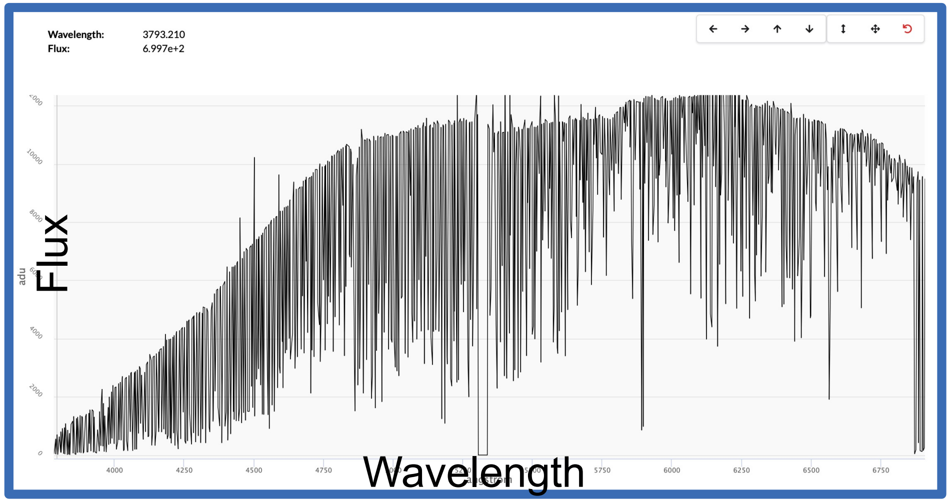
- The next step: characterization by source properties (object type, redshift, chemical composition, …) and/or by similarity

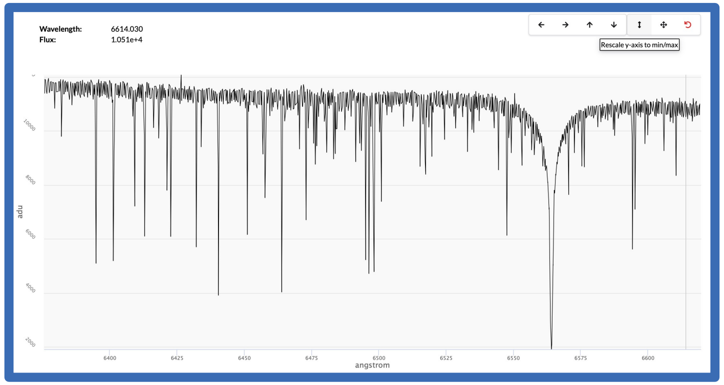
Deep Learning on the ESO Science Archives - I
Goals

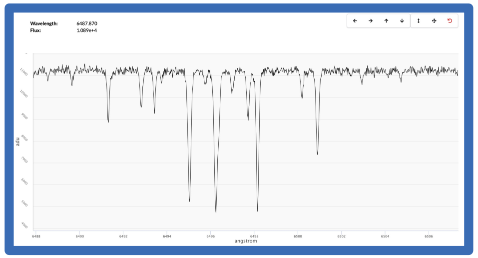
- Assess whether AI is useful in providing users with novel ways to identify data in the ESO Science Archives
  - Starting point: processed data
- Data is very heterogeneous, e.g. La Silla Paranal processed data:
  - 3.6 million files
  - 28 instruments
  - 71 data collections, 56 data providers
  - 3000 PIs, 9000 individual programmes
- Large and varied user base
  - More than half of professional astronomers worldwide
- Strive to limit the imposition of preconceived categories and criteria
- Results should be robust, understood, reproducible and user-friendly

Deep Learning on the ESO Science Archives – II
The HARPS experiment

- Deep Learning analysis of the entire HARPS archive
  - High-resolution, high-stability spectrograph
  - Relatively clean sample: mostly stars in the solar neighborhood
  - Data readily available
    - 1D spectra, processed in physical units (wavelength vs flux) to high accuracy and uniformity
    - ~270k spectra, ~300k wavelengths channels each







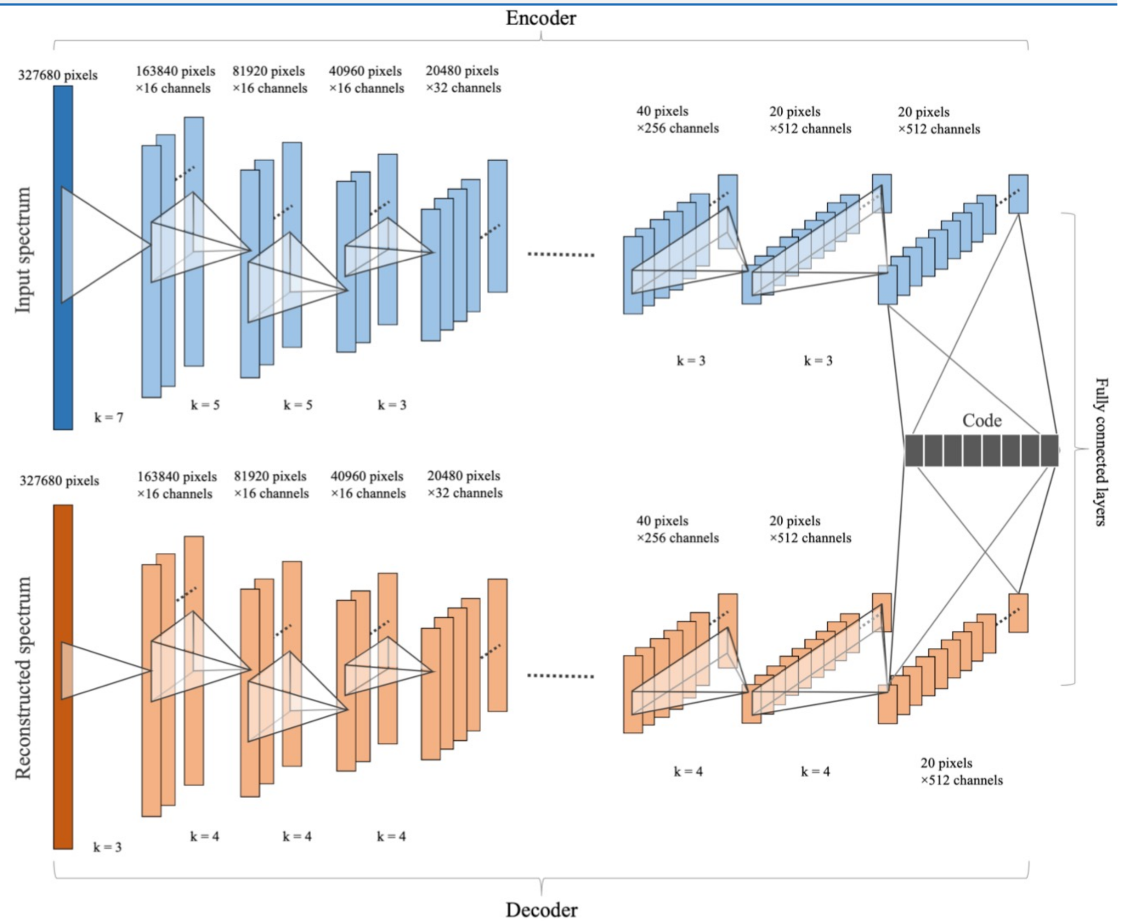
Sedaghat, MR, Carrick, Pineau 2021, MNRAS, 501, 6026

ESCAPE has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement 824064

APPROACH 1: UNSUPERVISED LEARNING

Unsupervised learning: Network schematics

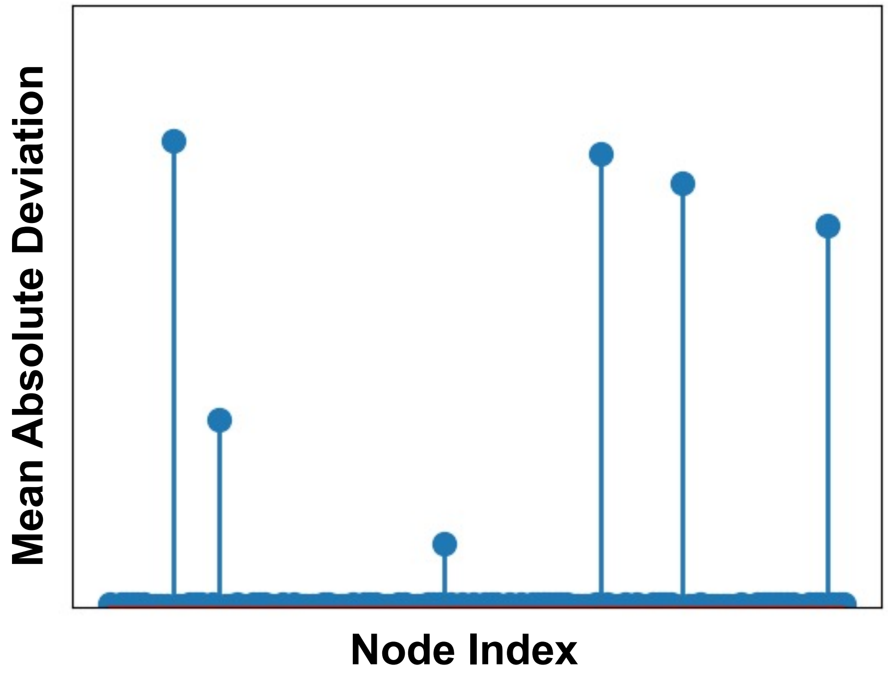
- Variational AutoEncoder
- Loss function: L1 norm
- Disentanglement for interpretation of latent space dimensions
- Understand what the networks ”learned”) in latent space (Code)
- Distances in the latent space for searches based on similarity



Sedaghat, MR, Carrick, Pineau 2021, MNRAS, 501, 6026

Unsupervised learning: Results - I

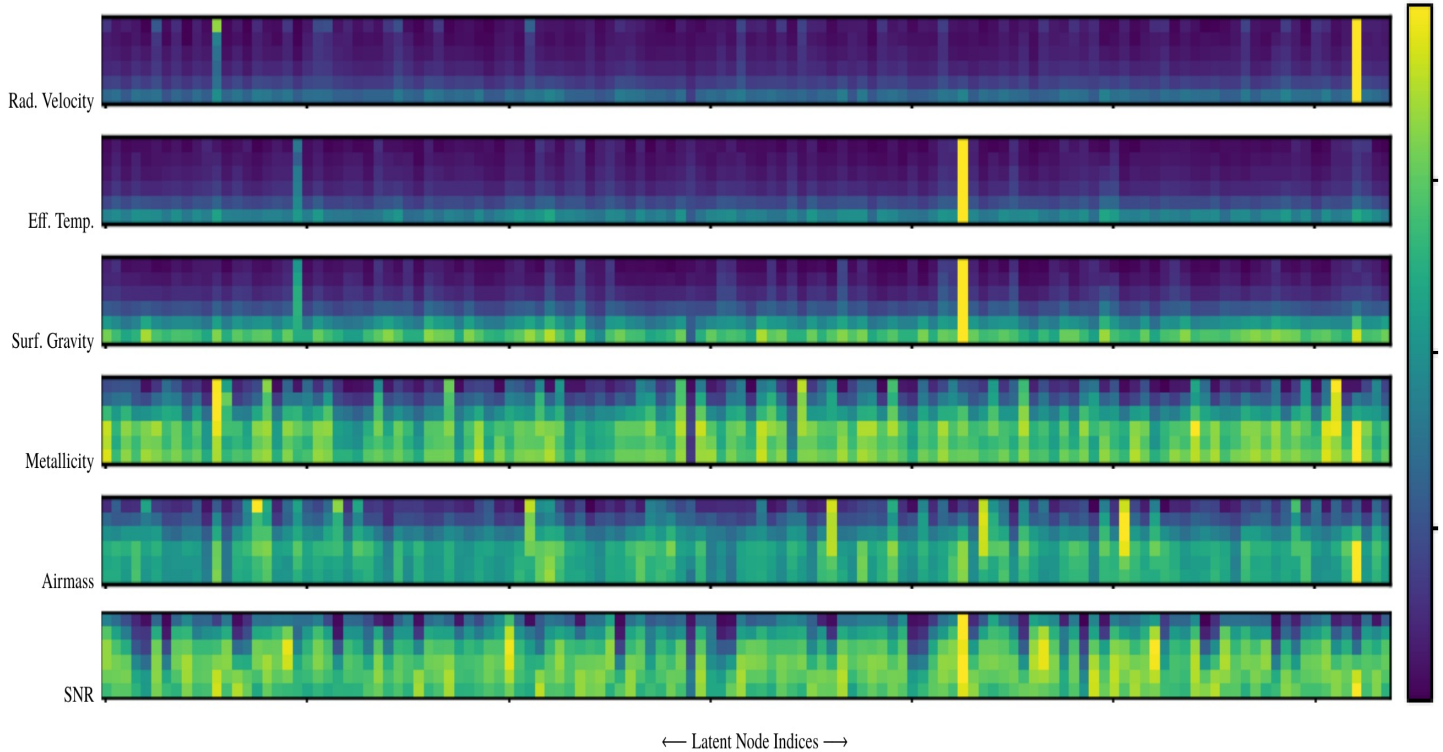
- 128 latent dimensions needed for good reconstruction
- Not all of them carry significant information
  - In fact, only 6 out of 128 do
- So, what are they? Do they have a physical interpretation?



Sedaghat, MR, Carrick, Pineau 2021, MNRAS, 501, 6026

Unsupervised learning: Results - II

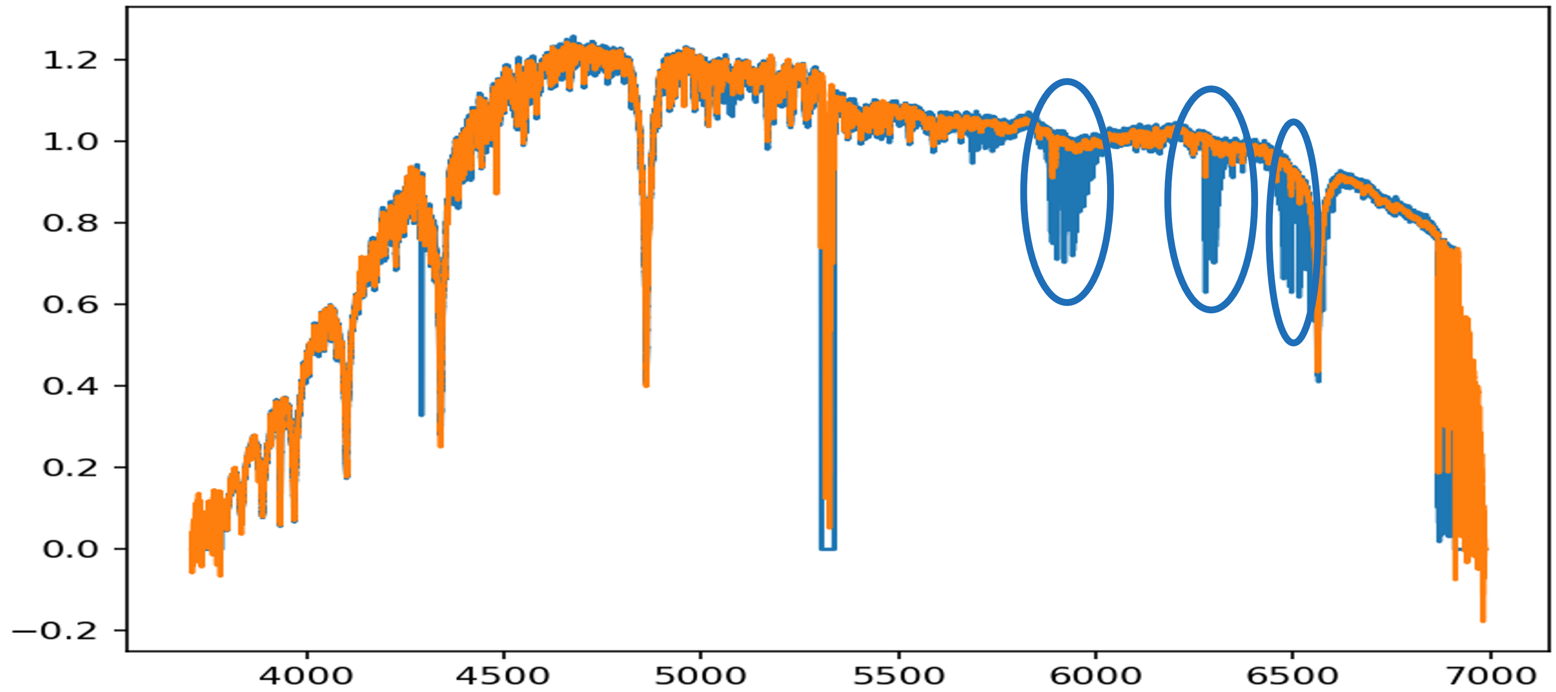
- Mutual Information w/ stellar parameters from SIMBAD@CDS
- Success!
  - Radial velocity (horizontal shift)
  - Temperature
  - Surface gravity
- Not so much so …
  - Chemical composition (metallicity)



Sedaghat, MR, Carrick, Pineau 2021, MNRAS, 501, 6026

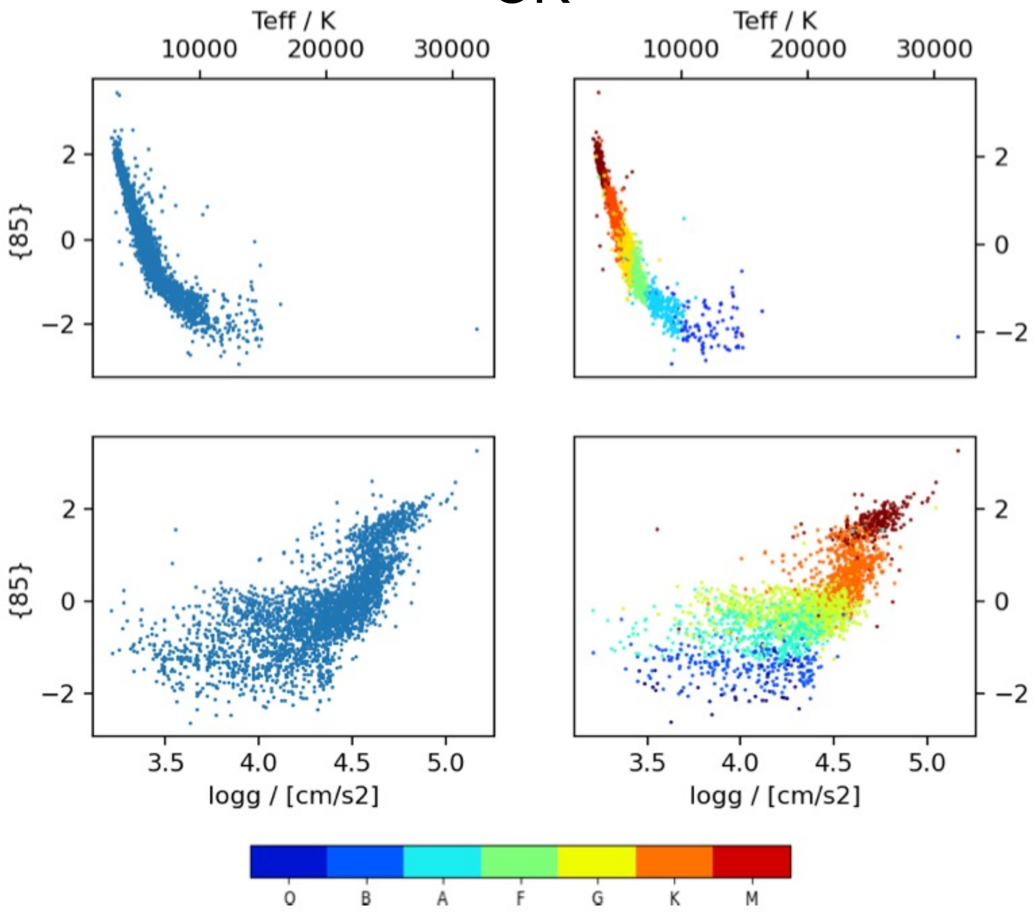
Unsupervised learning: Tantalizing hints

- Tendency to separate stellar vs Earth atmosphere features
  - In any reference frame, except topocentric

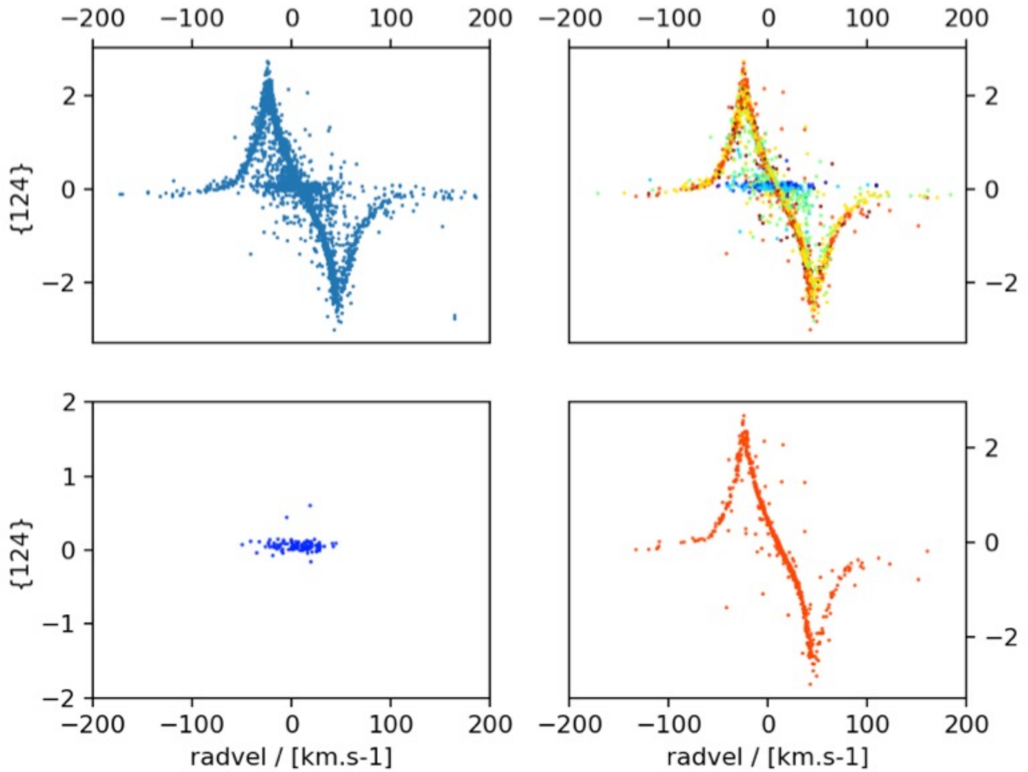


Unsupervised learning: Puzzling results - I

OK



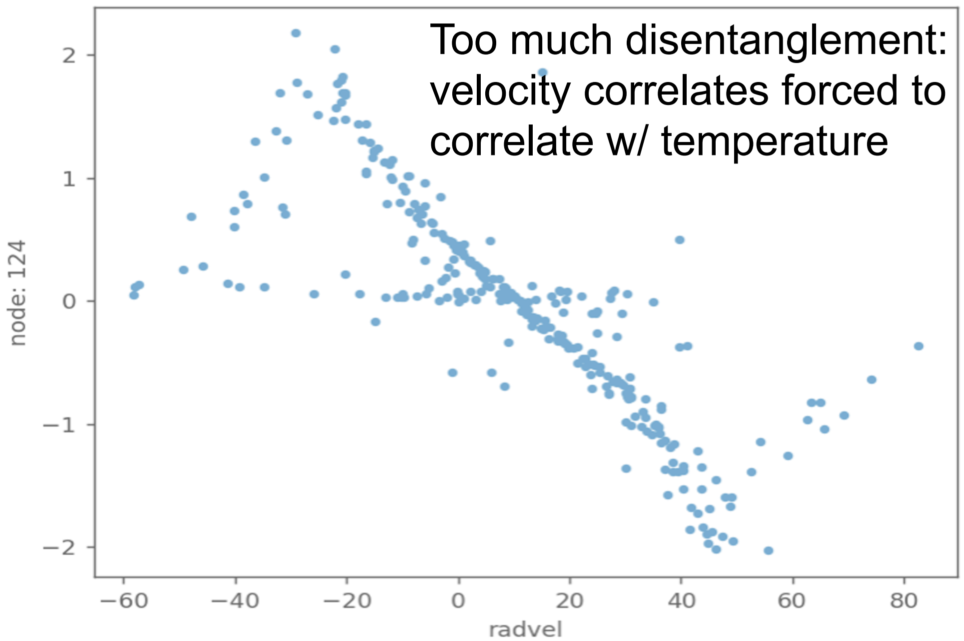
Non-monotonic?!?

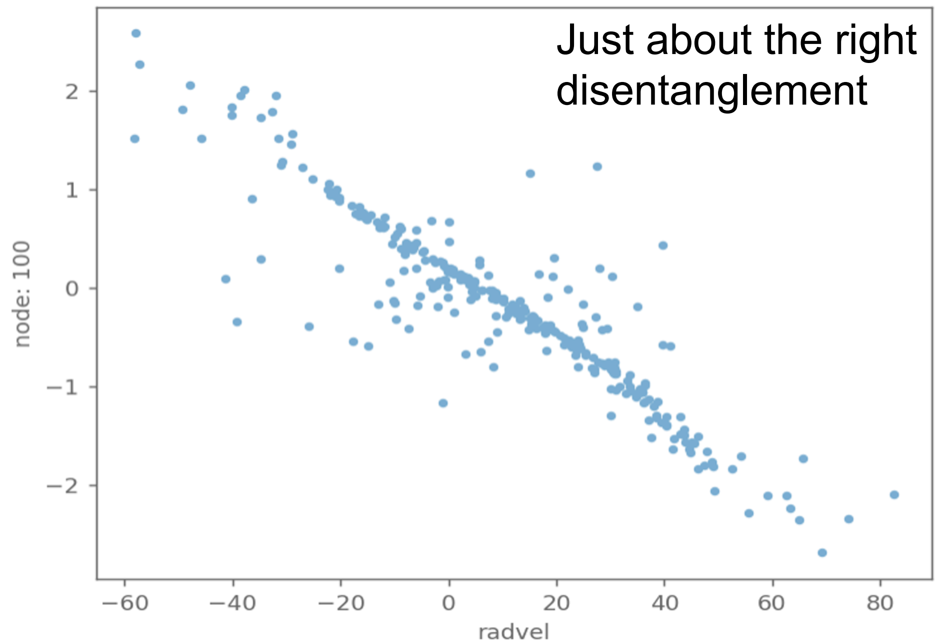


Sedaghat, MR, Carrick, Pineau 2021, MNRAS, 501, 6026

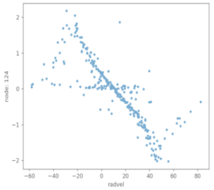
Unsupervised learning: Fine tuning disentangling - I

- Too high disentanglement penalty: latent nodes forced to correlate with multiple uncorrelated physical variables
- Too low disentanglement penalty: nodes are entangled

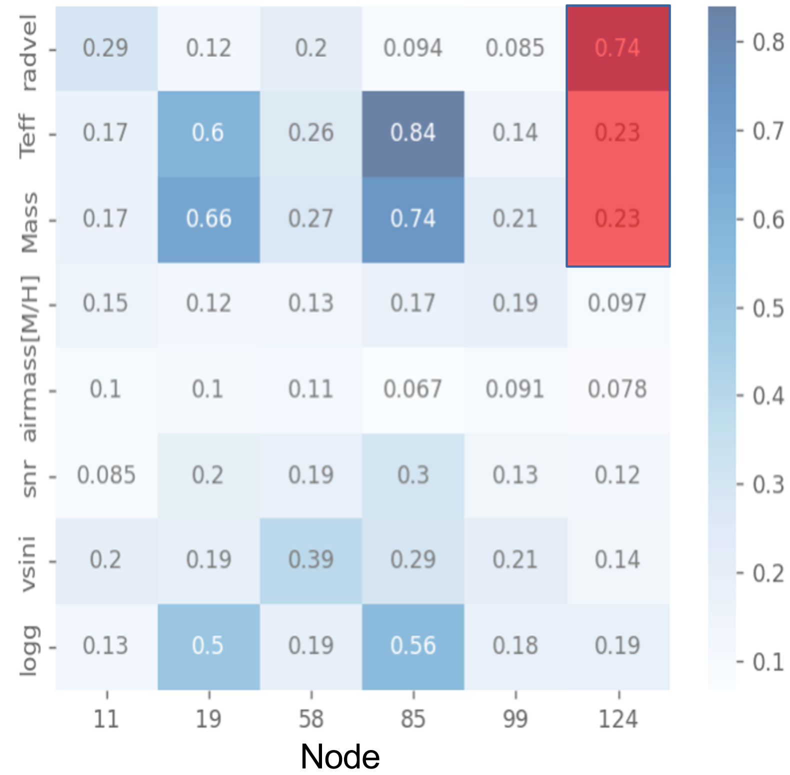


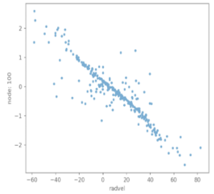


Unsupervised learning: Fine tuning disentangling - II

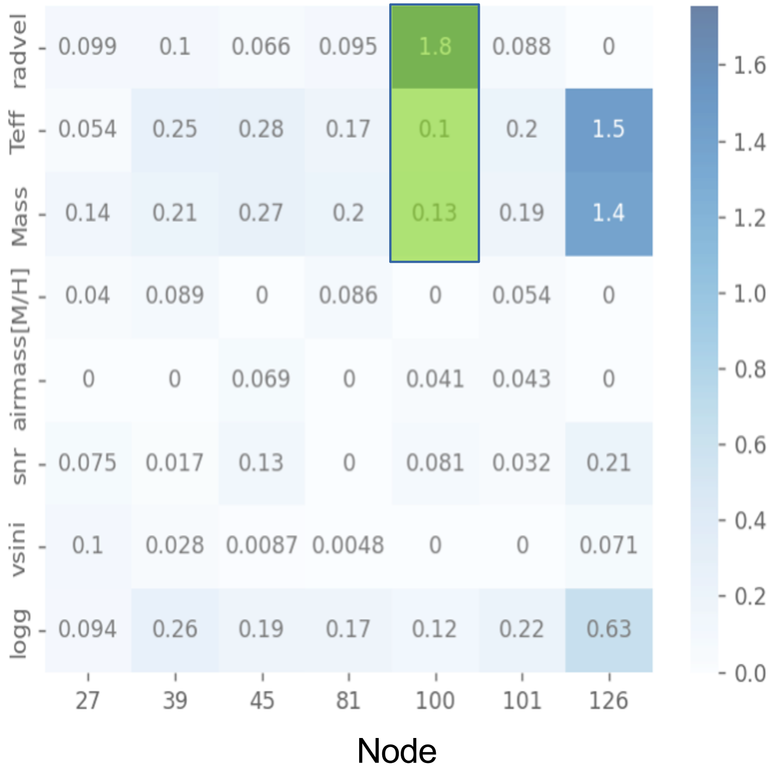


Too high disentanglement





Right level of disentanglement



Unsupervised learning: Provisional summary

- No labels used in the training, checked a-posteriori for interpretability
- Only a handful of latent space dimensions carry significant information
  - Some of them relate directly to physical parameters of the stars …
    - Effective Temperature, surface gravity, radial velocity
  - … but NOT ALL
    - No Mutual Information between chemical composition and nodes; some nodes unexplained
- Disentanglement needs tuning to be effective
- Physically correlated quantities remain so in the latent space (e.g., effective temperature, surface gravity, mass)
  - Problem for interpretability of, e.g., archive queries based on similarity
- Interesting features, but not quite ready for primetime

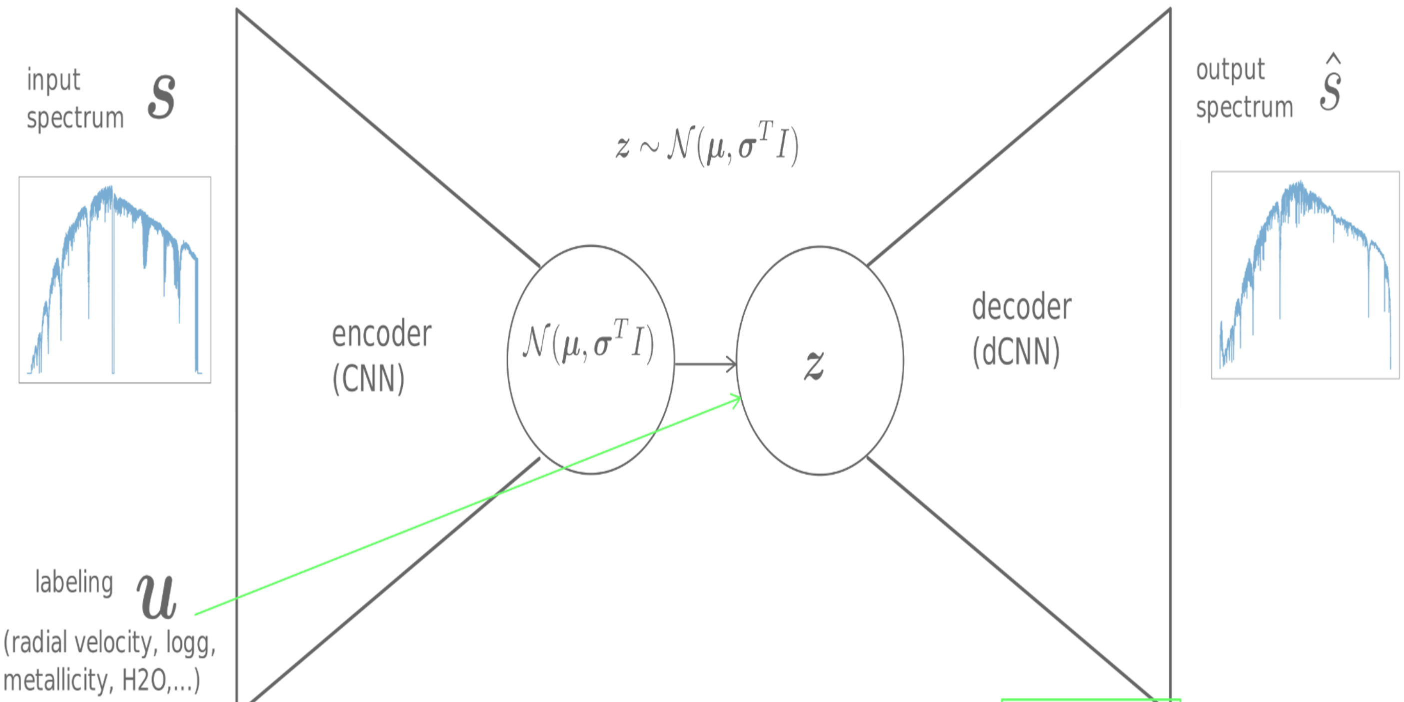
Very much work in progress …

APPROACH 2: WEAKLY SUPERVISED LEARNING

Weakly supervised learning: Network schematics

- Labels used for training, alongside reconstruction losses

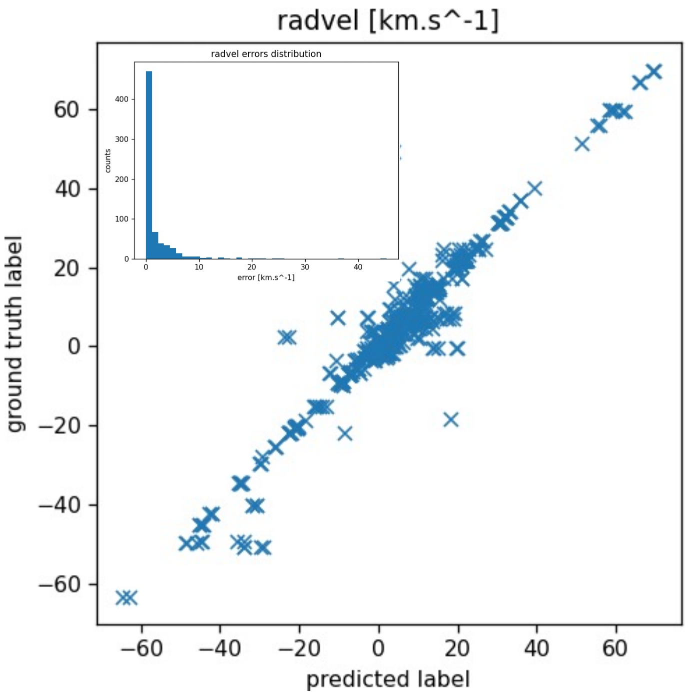
Labels include source and Earth atmosphere parameters

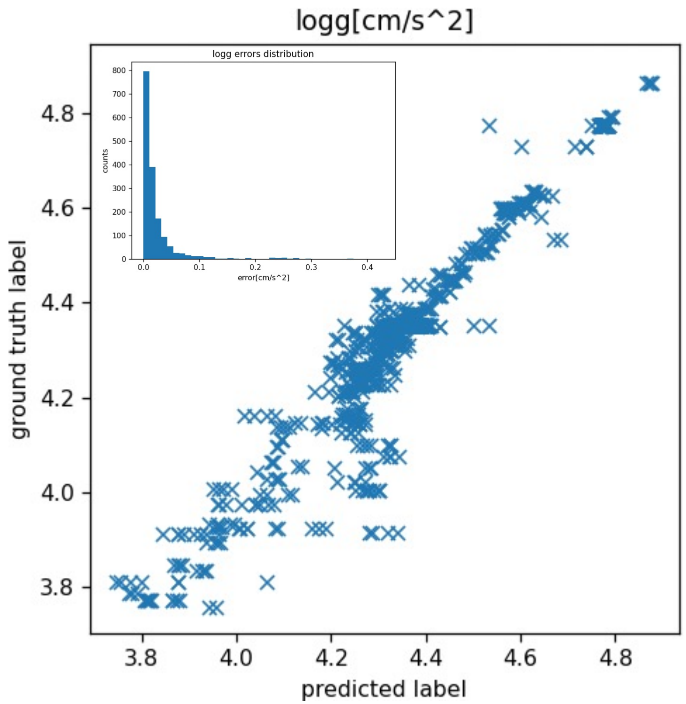


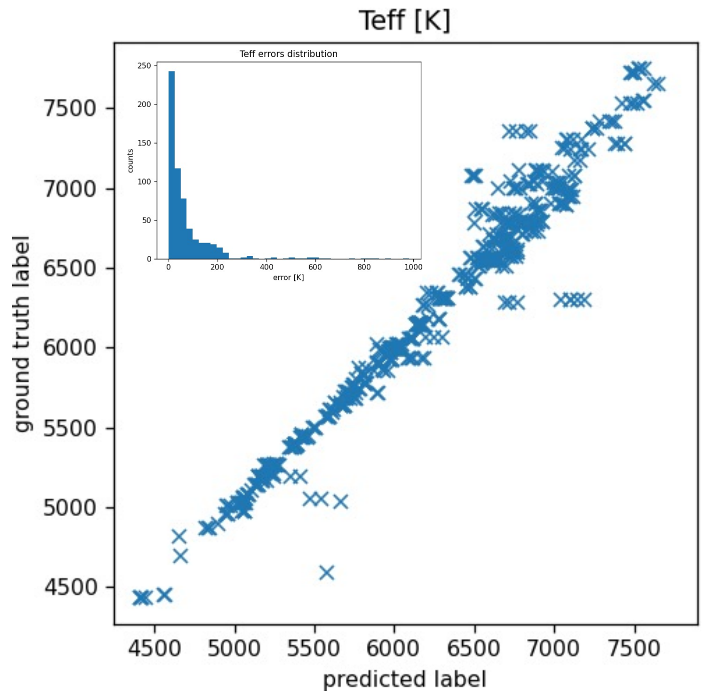
$$L(s,\hat{s},z,u;\lambda_1,\lambda_2) = \frac{1}{N}|s-\hat{s}| + \lambda_1 D_{KL}(q(z \mid s)||p(z)) + \boxed{\lambda_2 \sum_i |z_i - u_i|}$$

Weakly supervised learning: Results - I

- Rather good reconstruction of the labels

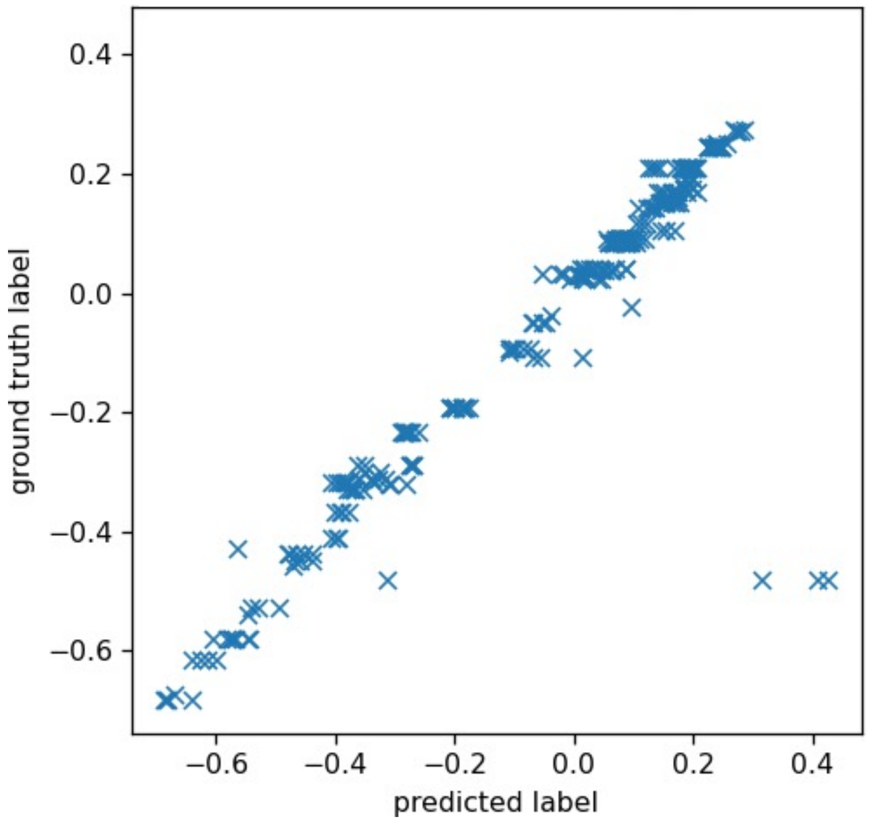




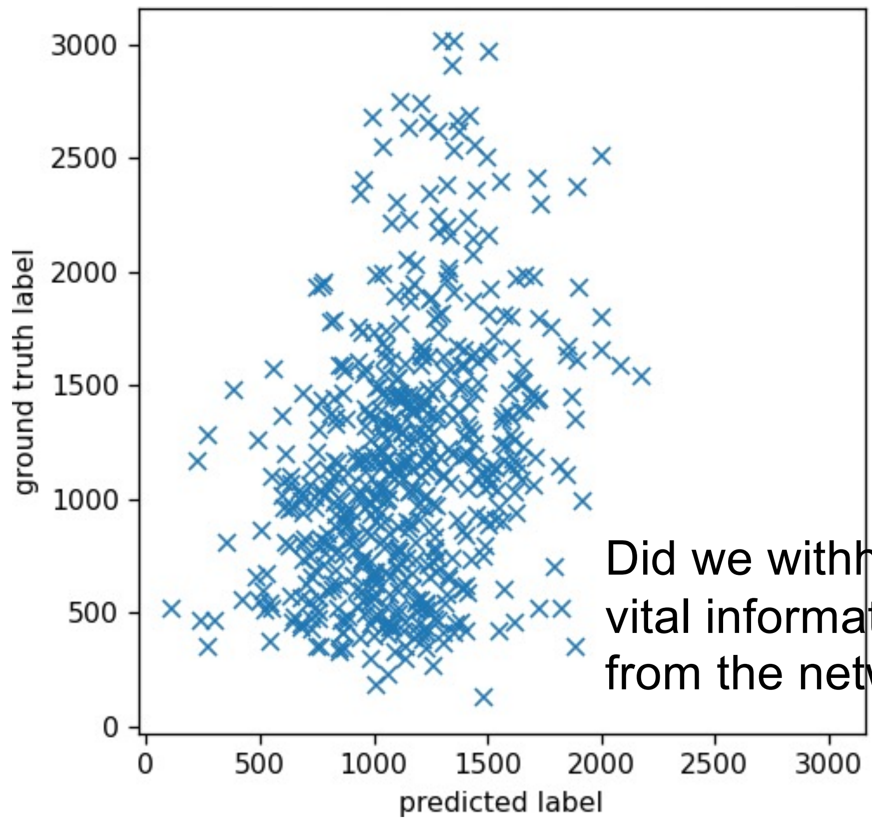


Weakly supervised learning: Results - II

Stars' chemical composition



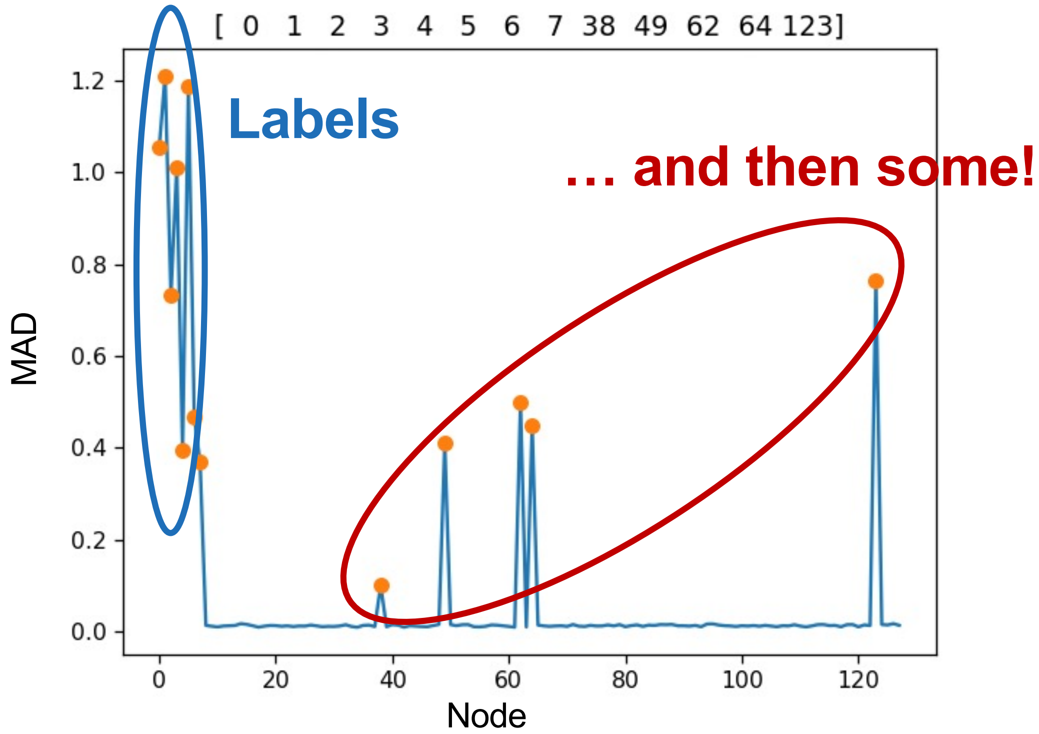
Earth's atmosphere



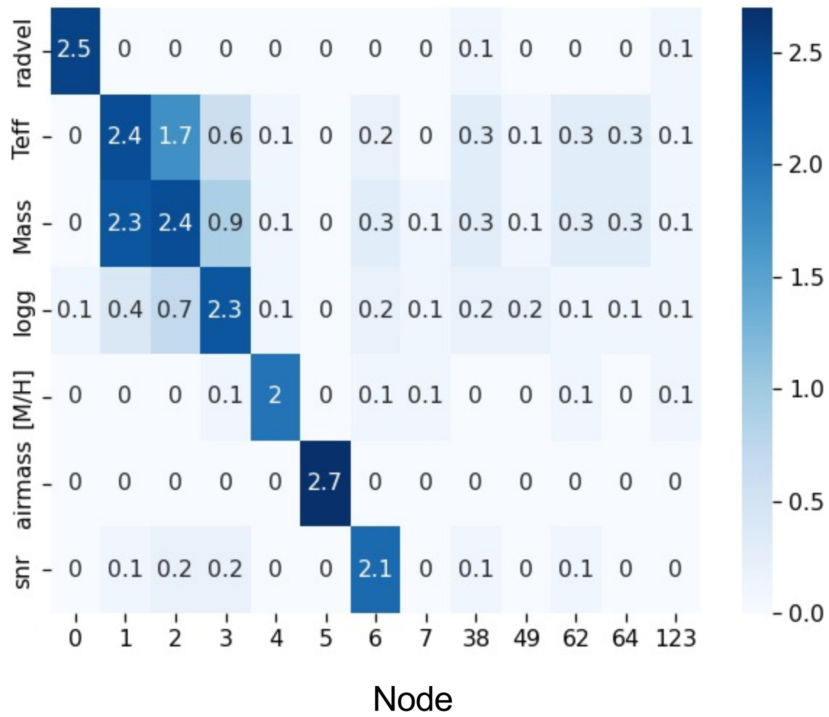
Did we withhold vital information from the network?

Weakly supervised learning: Results - III

- The network learns the labels, and then some



- Label nodes are well disentangled



Weakly supervised learning: Provisional summary

- Labels use in training alongside reconstruction losses
- Reconstruction of labels is solid (“supervised is easy”, ©Maggie Lieu)
  - BUT, several significant dimensions in addition: what are they?
- Do we have reliable sources of labels for all the cases?
- What to do for diverse samples with disjoint sets of parameters, e.g., stars and galaxies and QSOs and …,?
  - Pre-classification? (Cf. Caroline Heneka)
  - Spars(er) label matrix?
- Simulations may help
  - We are after physics, after all
  - Reinforcement learning (Cf. Maxime Quesnel’s talk with simulator as decoder)
  - Domain adaptation

Provisional conclusions

- We are running an experiment to extract physical parameters from massive dataset to build new query capabilities for archive research
  - Still very much work in progress
- The purely unsupervised approach has issues if interpretability in terms of the object's physical parameters is desired
  - Interpretability is important to present results to the intended broad and diverse audience of archive users
- The weakly supervised approach is promising in that sense, but brings the question of quality and availability of labels
  - Which anyhow affects the unsupervised approach, where labels are needed to validate the interpretability
- Simulations may help both approaches
  - WIP …