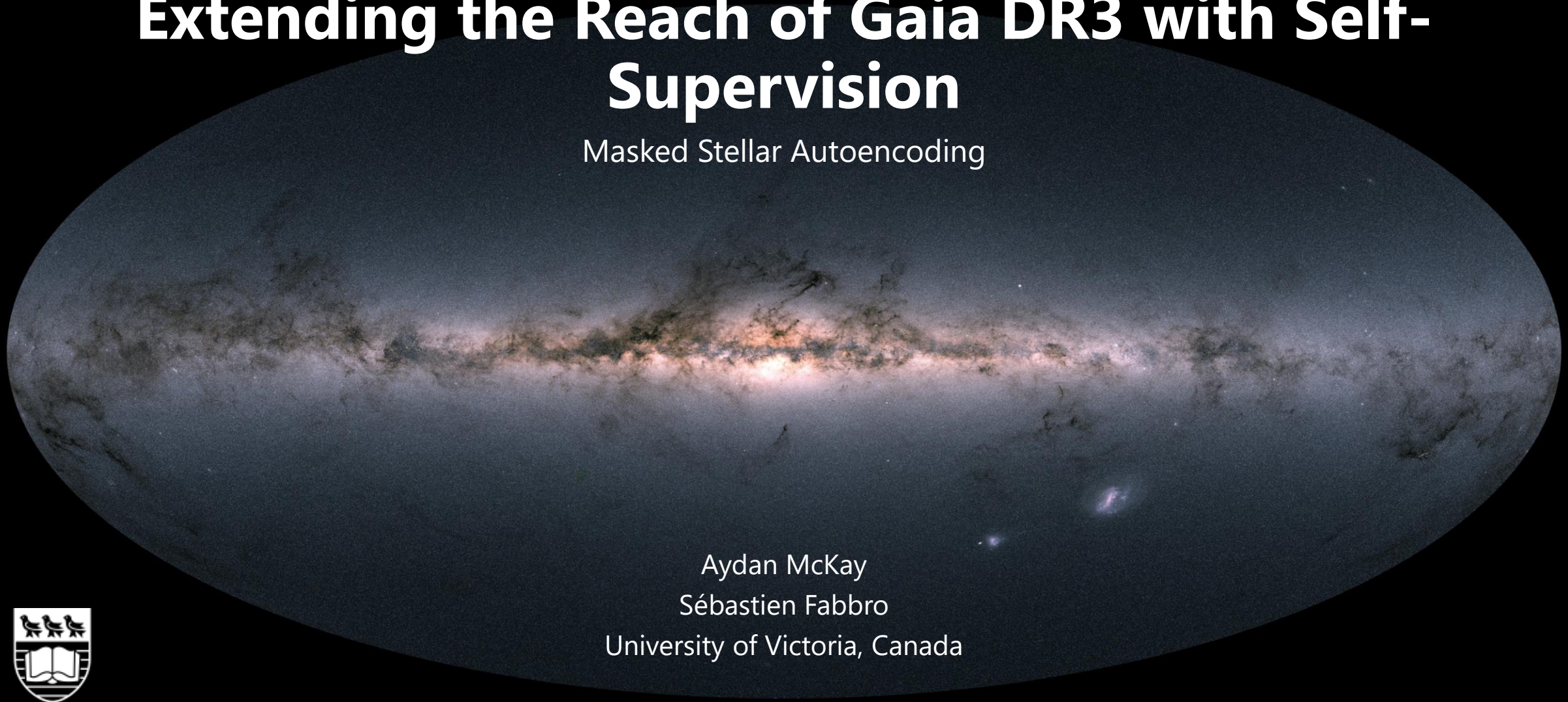
Extending the Reach of Gaia DR3 with Self-Supervision

Masked Stellar Autoencoding

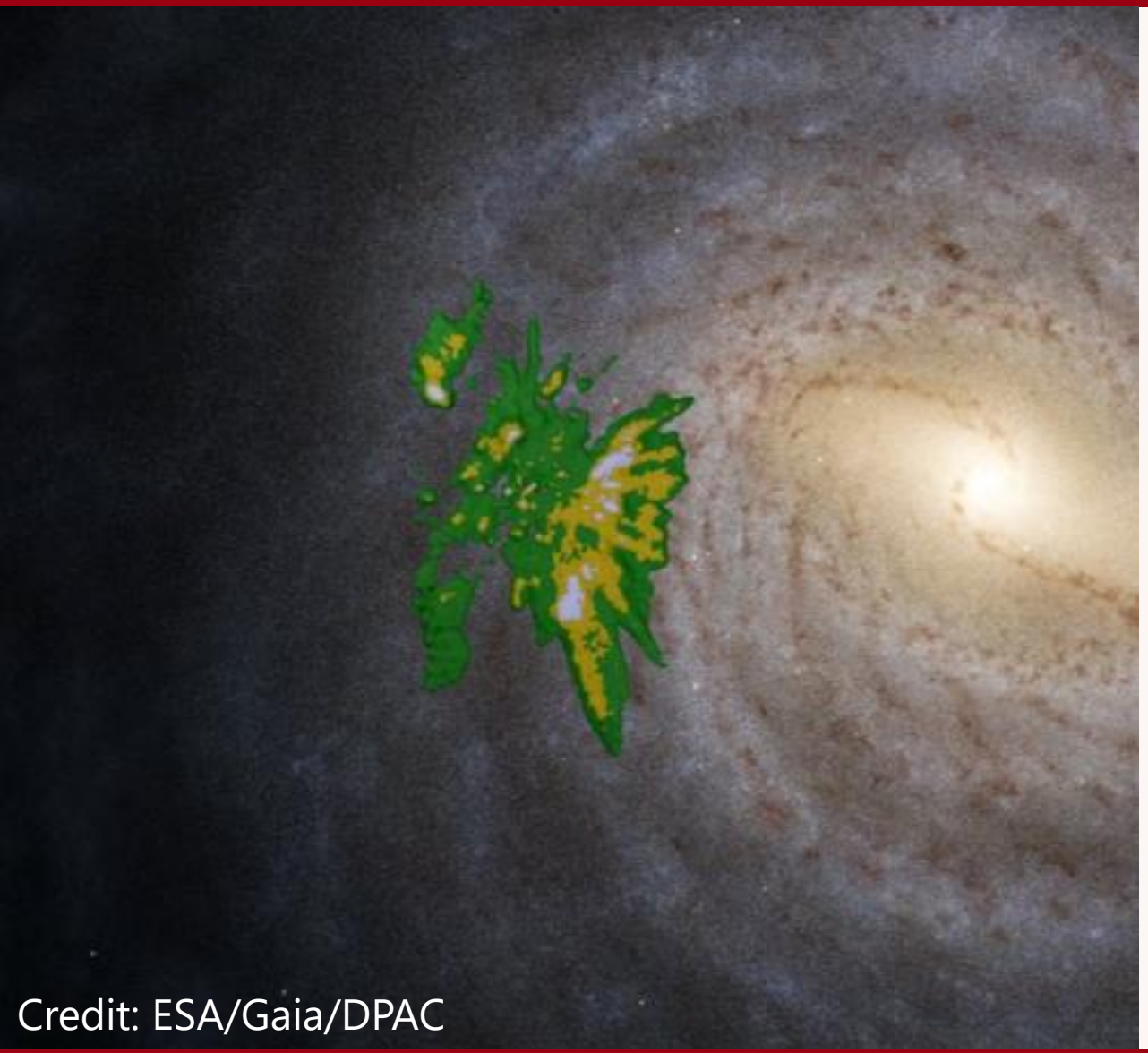


Aydan McKay
Sébastien Fabbro
University of Victoria, Canada

University of Victoria

Image Credit: ESA/Gaia/DPAC

The Current Data Swamp



Credit: ESA/Gaia/DPAC

Large surveys like Euclid and LSST will provide tremendous amounts of data in the coming years.

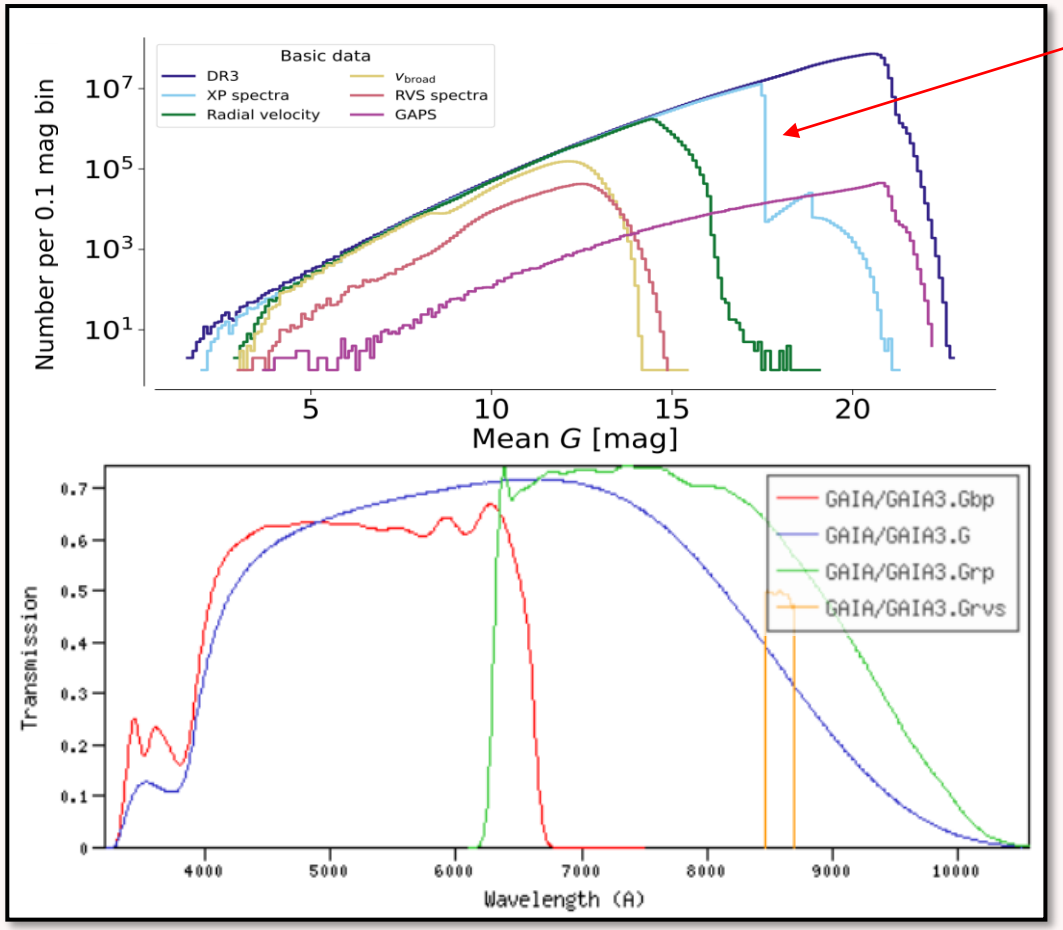
No need to wait for them, surveys such as Gaia, DELVE, etc. are extensive and readily available.

What can we do now with this information and how can we prepare ourselves for the onslaught of new data

How can we maximize the use of all these data sets for stellar astrophysics?

Gaia DR3

Vallenari et al. 2022



The SVO Filter Profile Service. Rodrigo, C., Solano, E., 2020

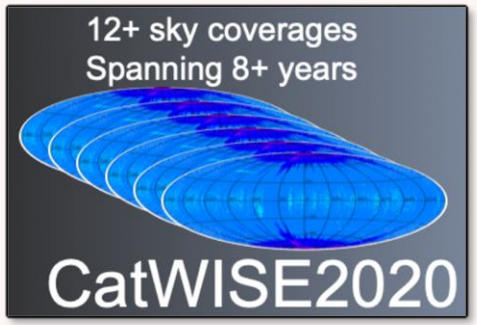
XP drop-off from limitation

Gaia DR3:

- **1.8 billion stars with photometry** + astrometry
  - Limited to $\lesssim$ 21 mag in Gaia G-band
- **220 million stars** with low-resolution (R~50) **spectro-photometry** from BP and RP instruments
  - Sampled, continuous, or coefficients*
  - **Limited to $\lesssim$ 17.65 mag** in Gaia G-band
- Will be used as the reference catalogue for merging datasets

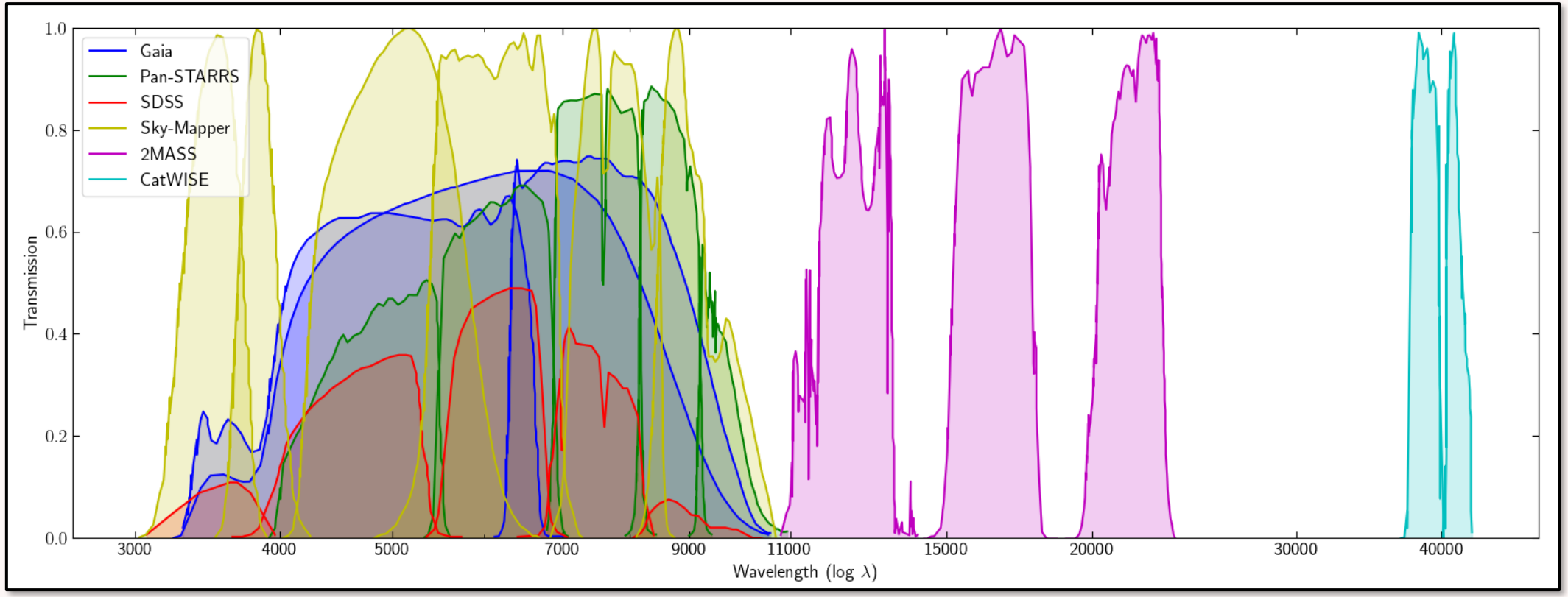
Considered Photometric Surveys

Survey	SDSS 13	Sky-Mapper DR2	Pan-STARRS1	2MASS	CatWISE2020	Gaia DR3 XP Spectra
Bands	u', g', r', i', z'	u, v, g, r, i, z	$g_{P1}, r_{P1}, i_{P1}, z_{P1}, y_{P1}$	J, H, K_S	W_1, W_2	110 Coefficients
Depth	$g' < 23.13$	$g < 21.7$	$g_{P1} < 23.5$	$J < 15.8$	$W_1 < 17.7$	$G < 17.65$
Sources in Gaia DR3	120 M	440 M	950 M	460 M	670 M	220 M



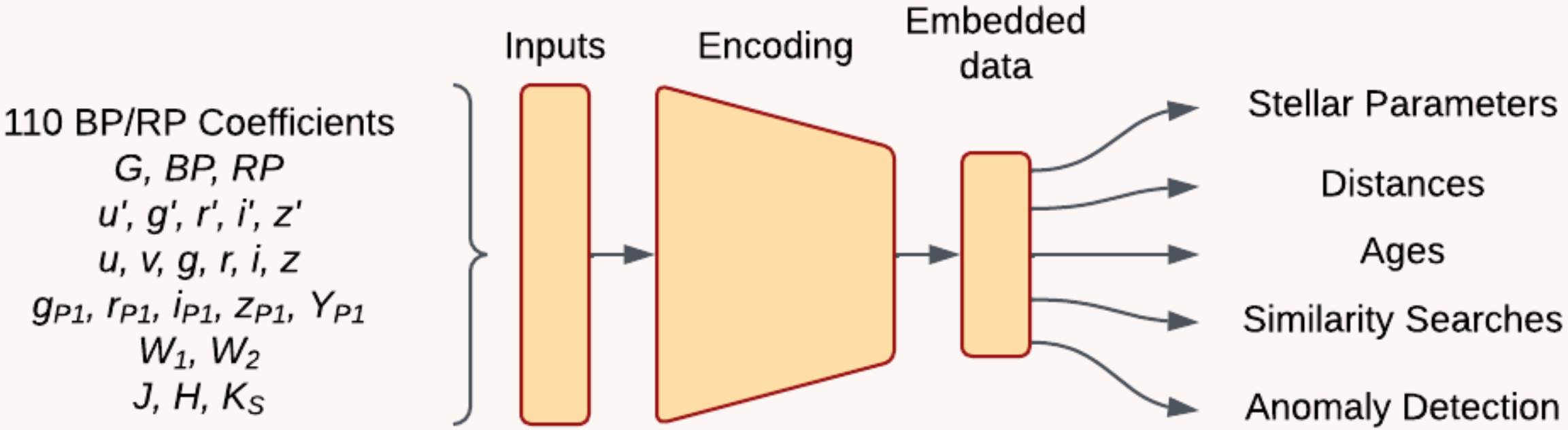
The Overlap in the Data

- The slight differences and inconsistencies in filters may unveil differences in stellar spectra



- We need a method to combine all this data in informative embeddings for stars

Embedding the Data



The Goals

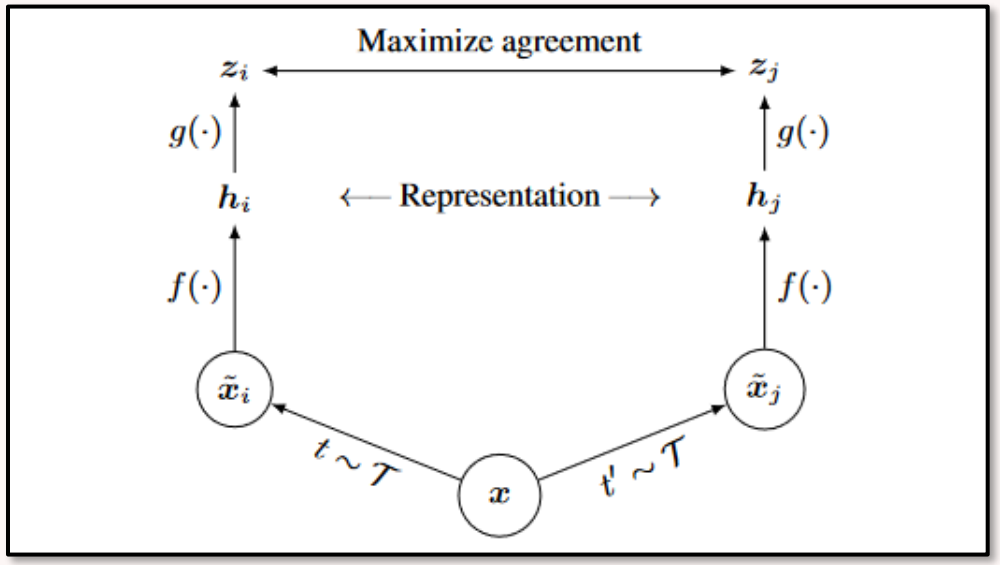
The model will:

- Encode the data into powerful embeddings while natively **imputing the information where missing** from the input

- **Extend the depth of XP spectra** beyond their limiting magnitude using these informative embeddings

- **Mitigate selection biases in surveys**

Self-Supervised in Computer Vision

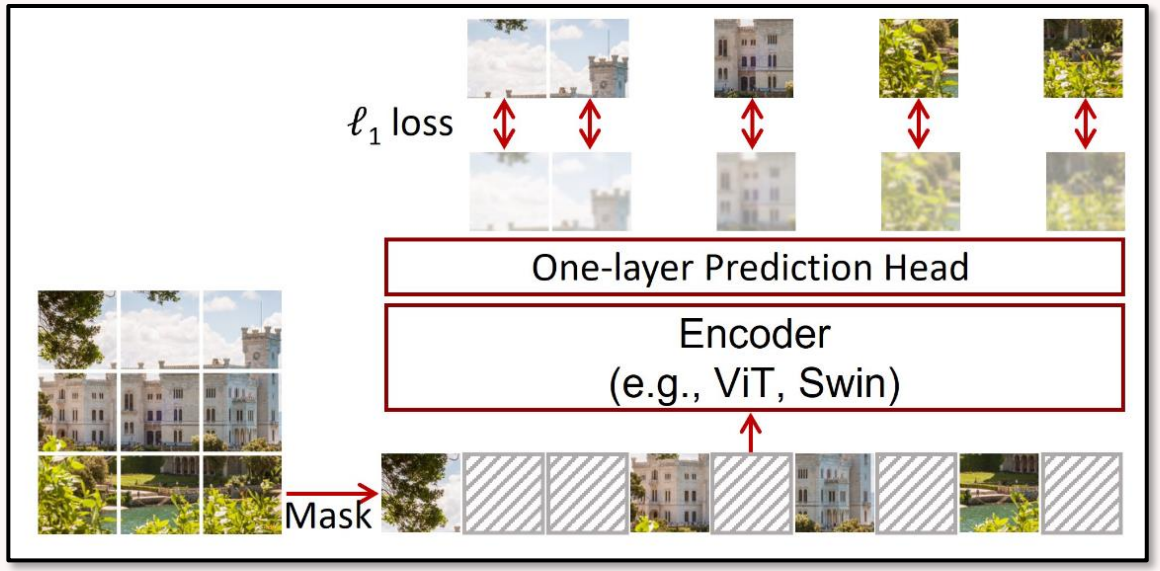


Contrastive Learning

- **Maximize agreement** between two similar views
- **Minimize disagreements** between two dissimilar views
- See: SimCLR (Chen+ 2020)

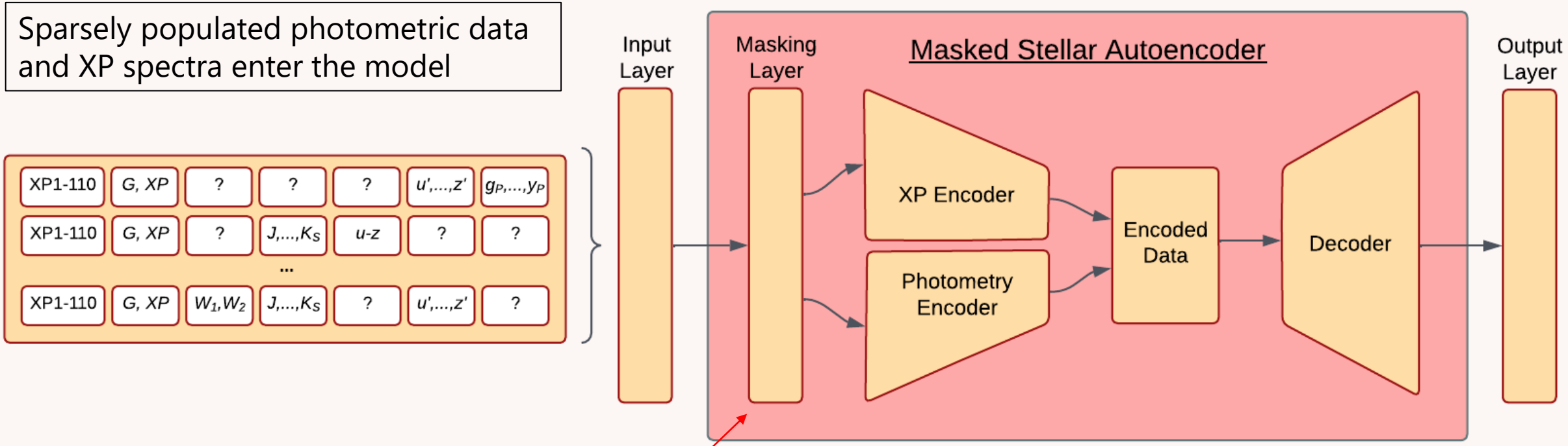
Masked Image Modelling

- **Reconstructs** artificially corrupted data
- **Masks inputs** and encodes entire vector
- Use transformers and **scale**
- See: MAE (He+ 2022), SimMiM (Xie+ 2022)



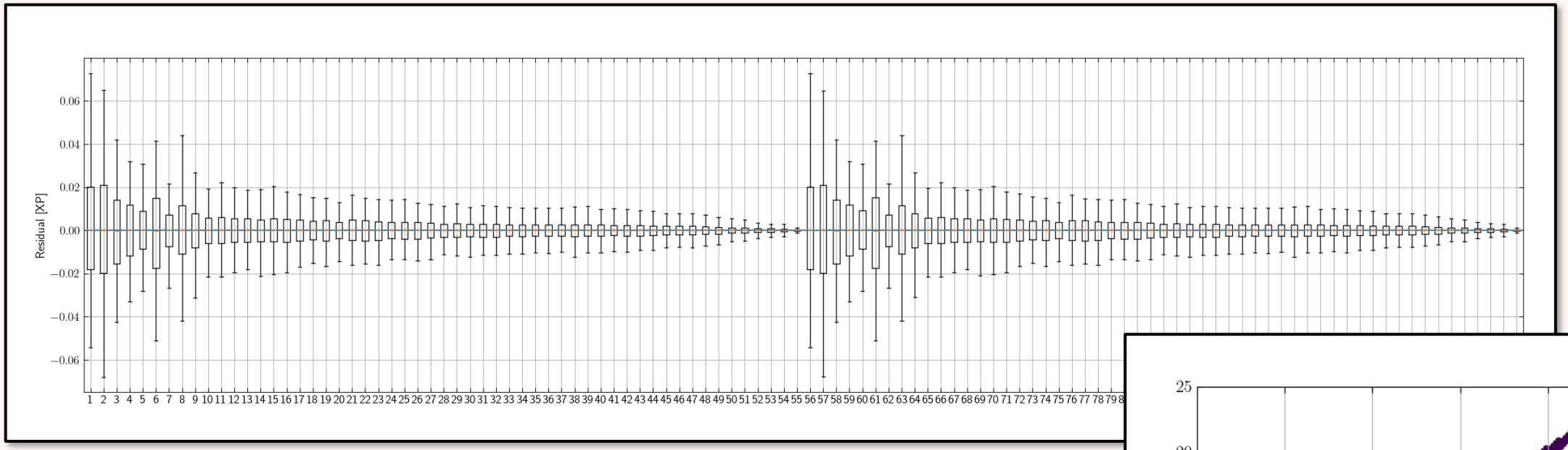
Masked Stellar Autoencoder - MSA

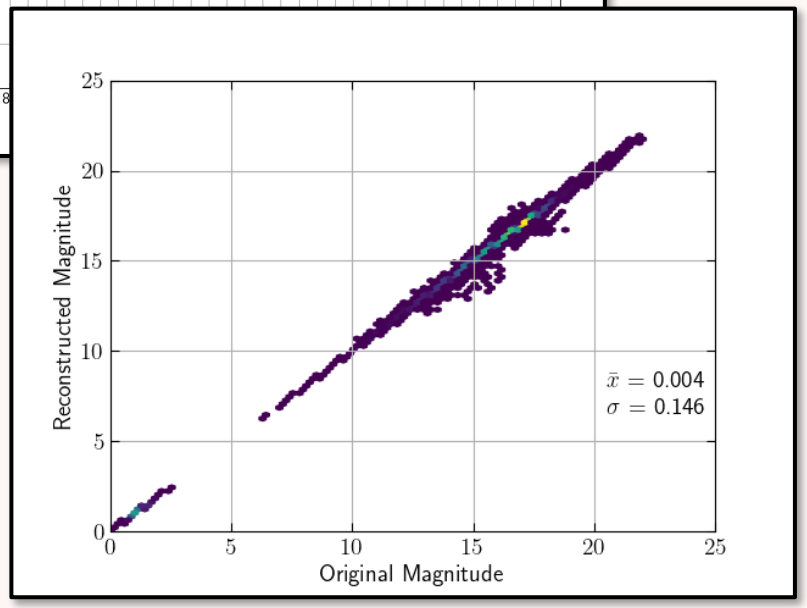
Sparsely populated photometric data and XP spectra enter the model



Missing values, and random XP spectra are masked either entirely or not at all.

Reconstruction of Magnitudes





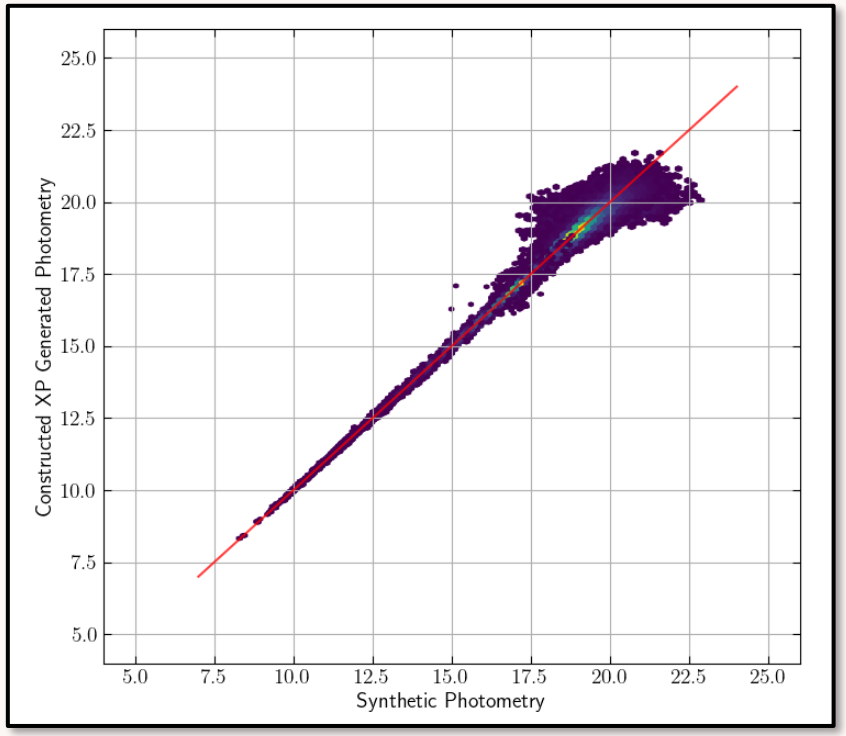
Construction of known magnitudes and XP coefficients using the MSA

- **Low residuals** in the coefficients
  - Even when masked in the input
- Great agreement between the magnitudes
  - Expected for **no loss of information**

Adding Consistency Constraints

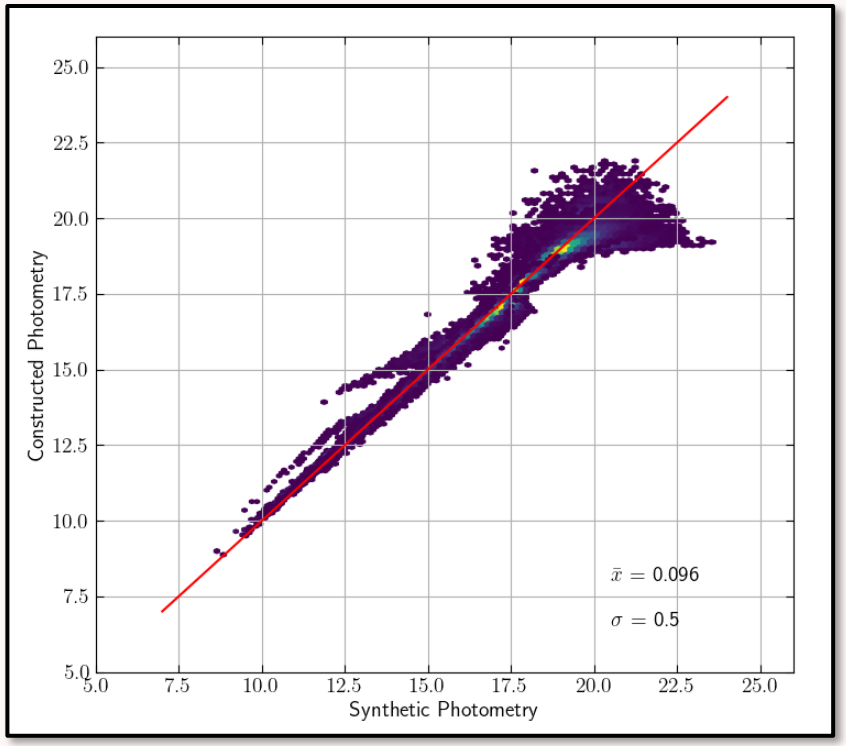
Constructing Magnitudes that didn't exist in their proper surveys in the first place:

- **Constraining** both constructed XP coefficients and magnitudes with synthetic magnitudes generated from XP Spectra (GaiaXPy)
- Mitigating the hallucinated magnitudes with **synthetic photometry**



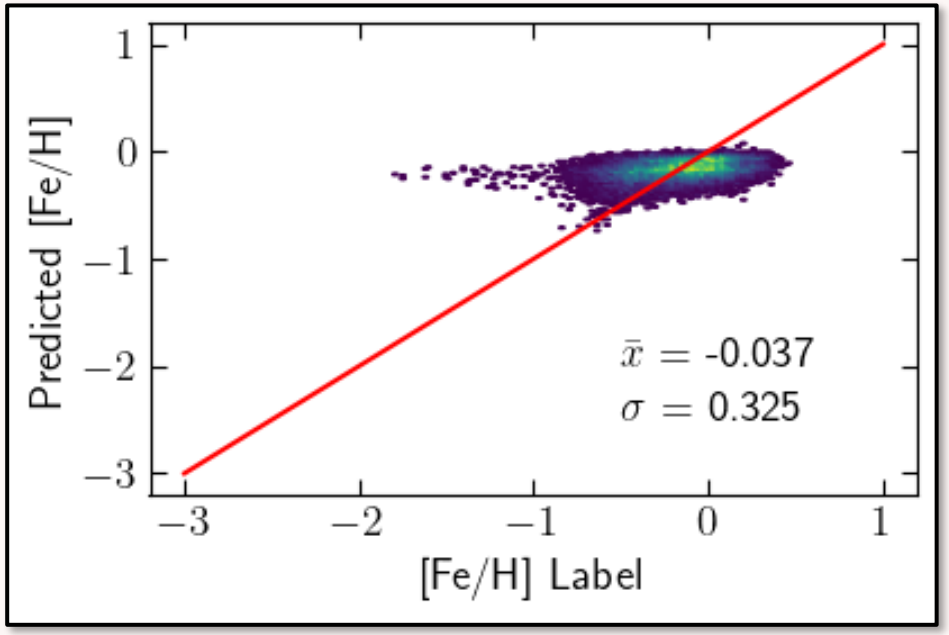
Magnitudes Generated with Reconstructed XP Coefficients

Magnitudes Reconstructed by the MSA

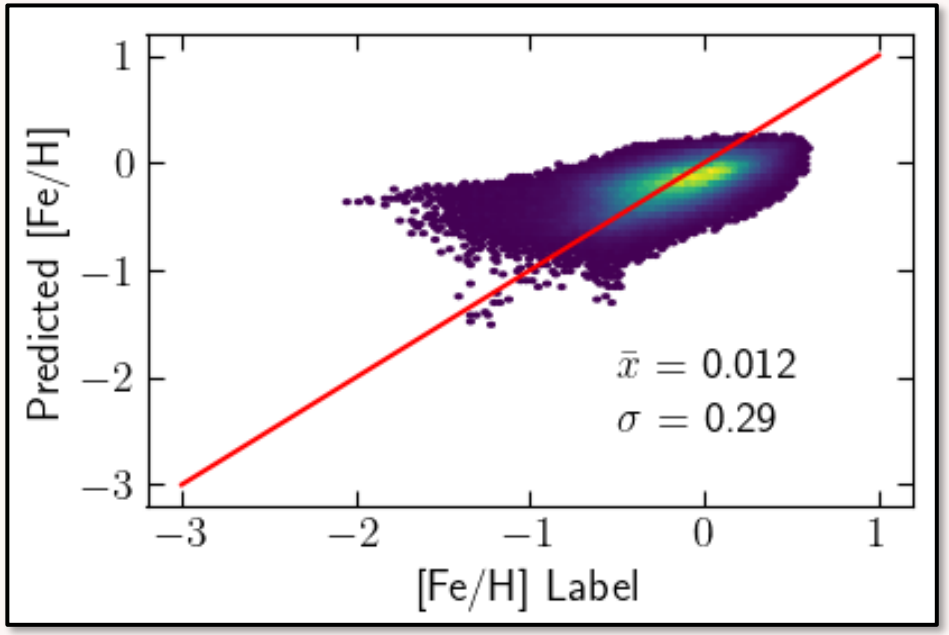


Astrophysical Fine Tuning when Scaling

Example astrophysical task: the prediction of metallicity using APOGEE+LAMOST labels



Adding More Data



Summary

- With the abundance of data already existing, we can **combine** all the different photometric filters **to infer spectroscopic parameters**
- The **Masked Stellar Autoencoder** (MSA) is used to create extremely informative encodings that are **more powerful than magnitudes by themselves**.
- **With more data**, the model becomes better at mitigating the hallucinations in magnitudes and XP spectra coefficients
- Work in Progress! Converging on results -> early 2024