INAF
ISTITUTO NAZIONALE
DI ASTROFISICA

Survey of Surveys: Evaluating Methods for Homogenizing Stellar Parameters Across Spectroscopic Surveys

Aleksandra Avdeeva (INAF - OAA)
E. Pancino and the StarDance team

AI in Astronomy Workshop, Catania 21th May 2025

Star Dance

European Research Council
Established by the European Commission

Motivation

Challenges: Heterogeneity in data from different spectroscopic surveys.

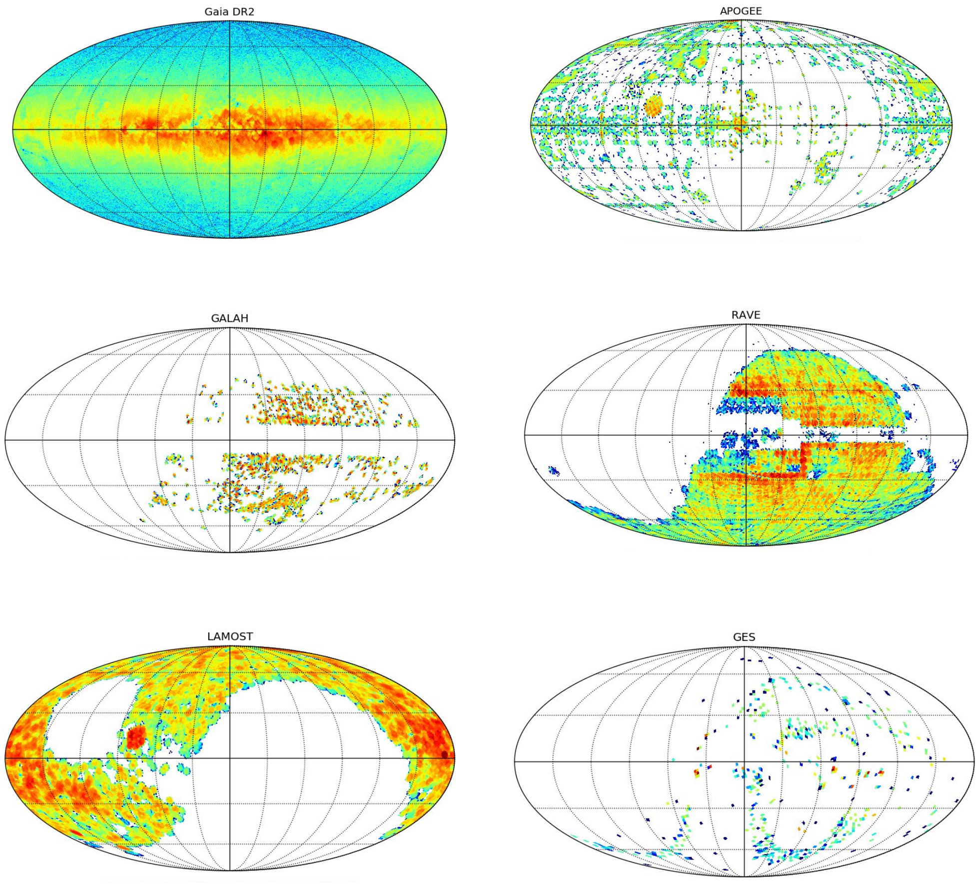
Objective: Create a unified catalogue for stellar parameters, ensuring consistency across datasets.

First results: Survey of Surveys (SoS) DR 1/2

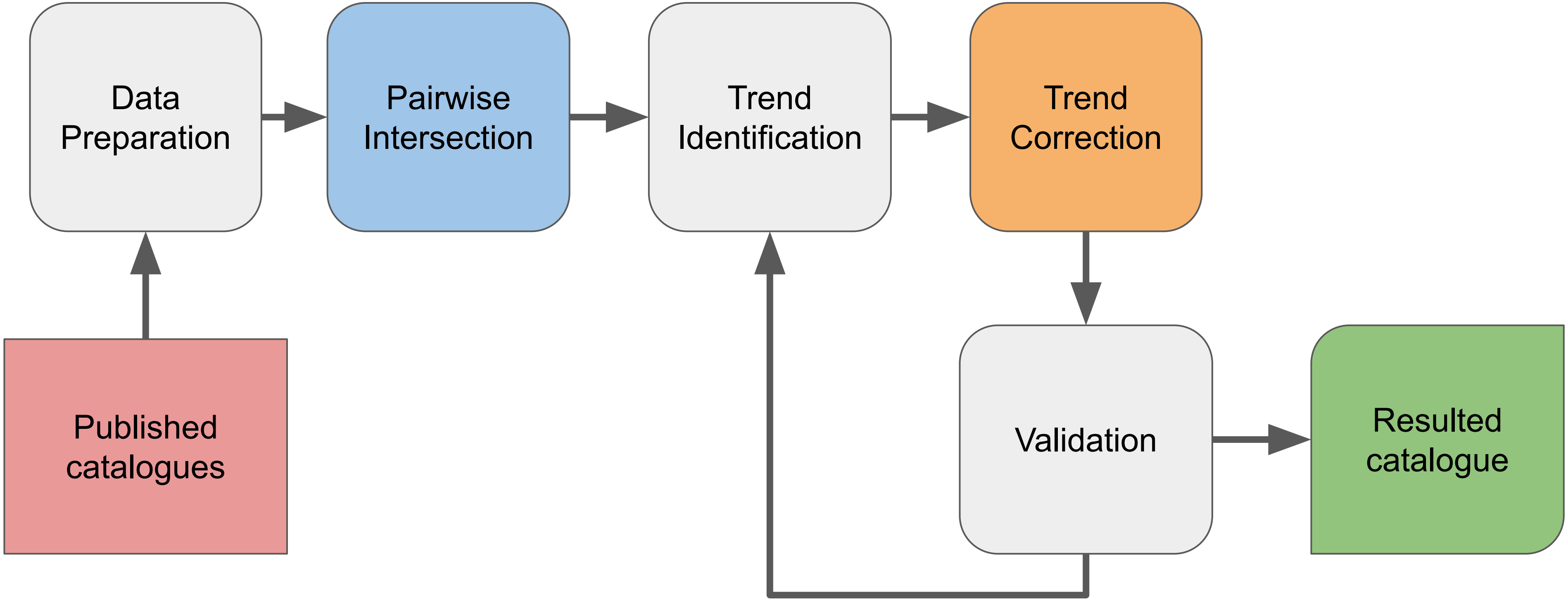
Tsantaki et al., 2022 - Radial velocities.

Turchi et al. (in preparation) - Homogenized spectroscopic stellar parameters + 19M stellar parameters from photometry with ML.

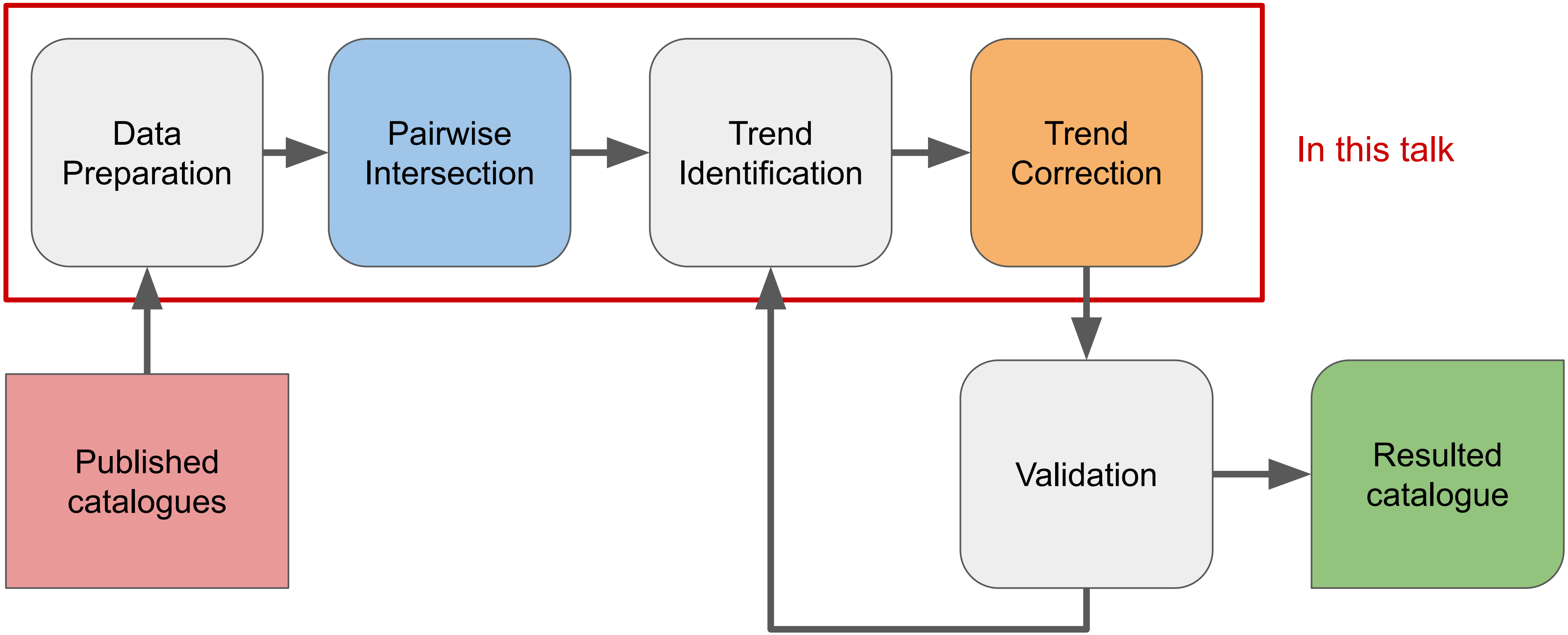
Now: Recent data releases provide improved opportunities for cross-survey calibration.



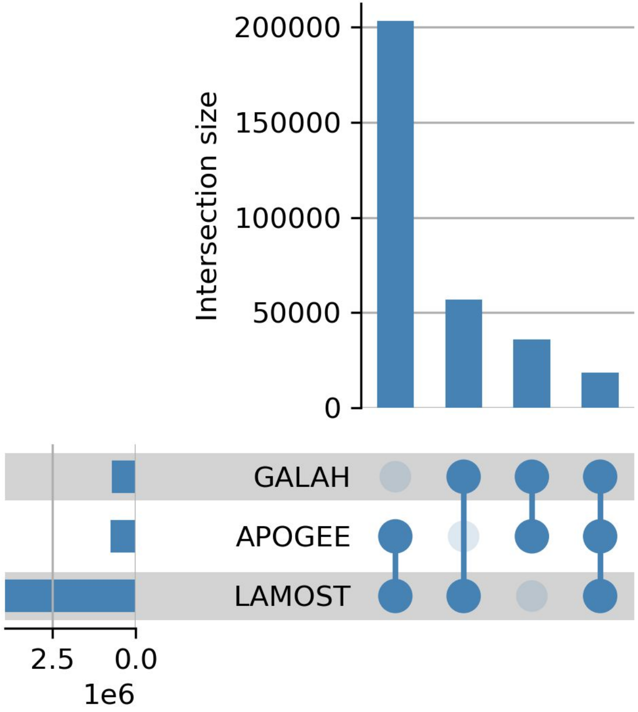
Strategy Overview



Strategy Overview

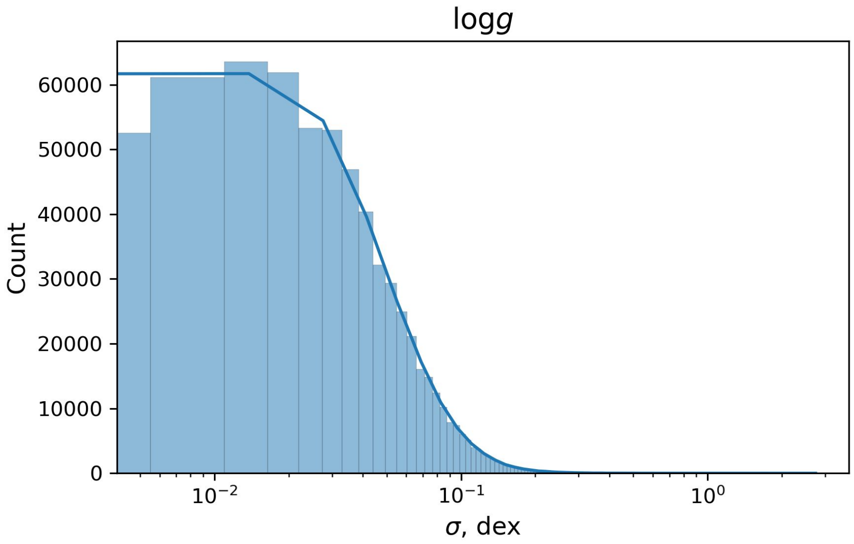


Data Selection and Preprocessing

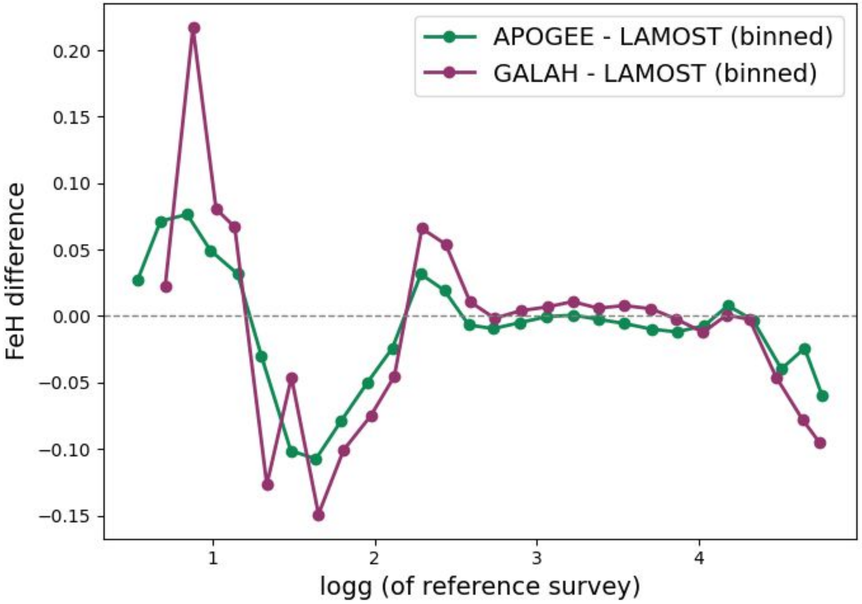


Surveys: APOGEE, GALAH, LAMOST

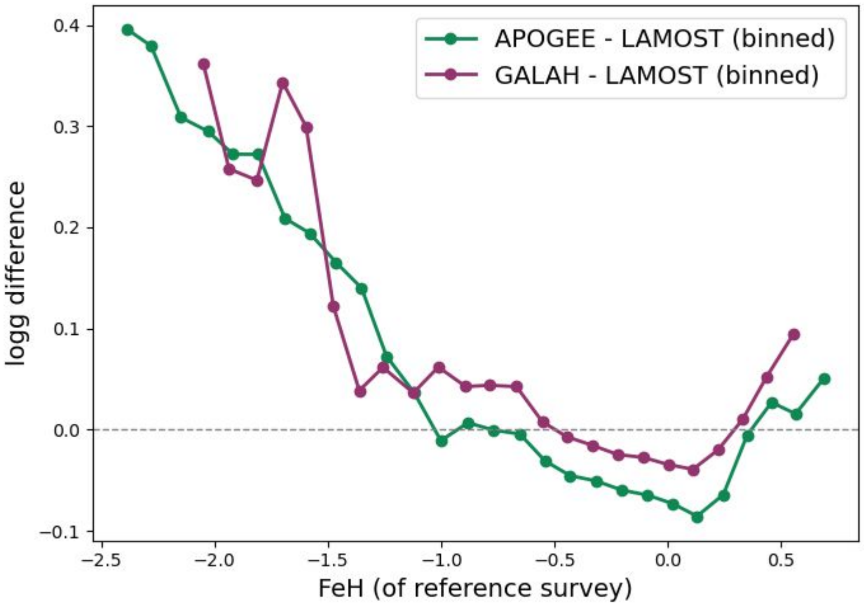
Filters: quality flags, duplicates with large parameter spread.



Trend Detection



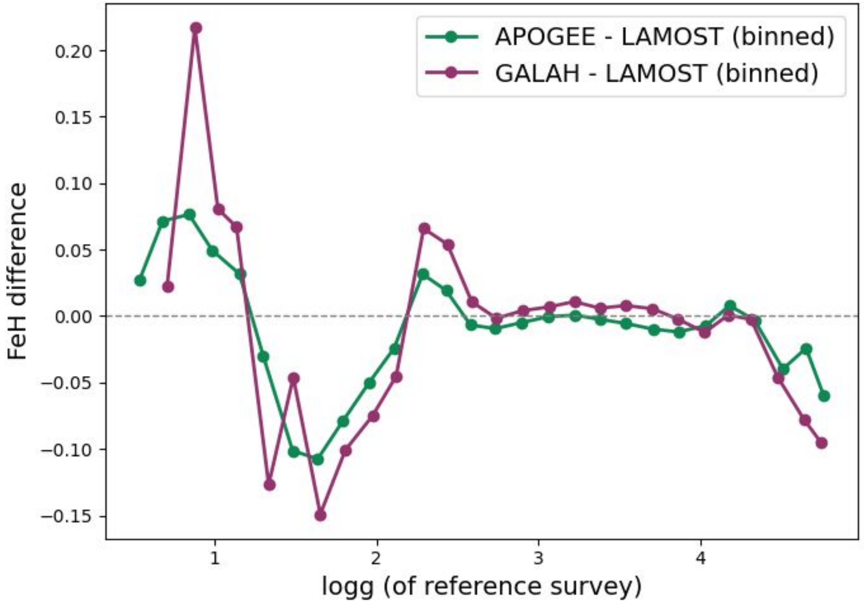
On y-axis: the difference between LAMOST and APOGEE/GALAH [Fe/H] measurements.



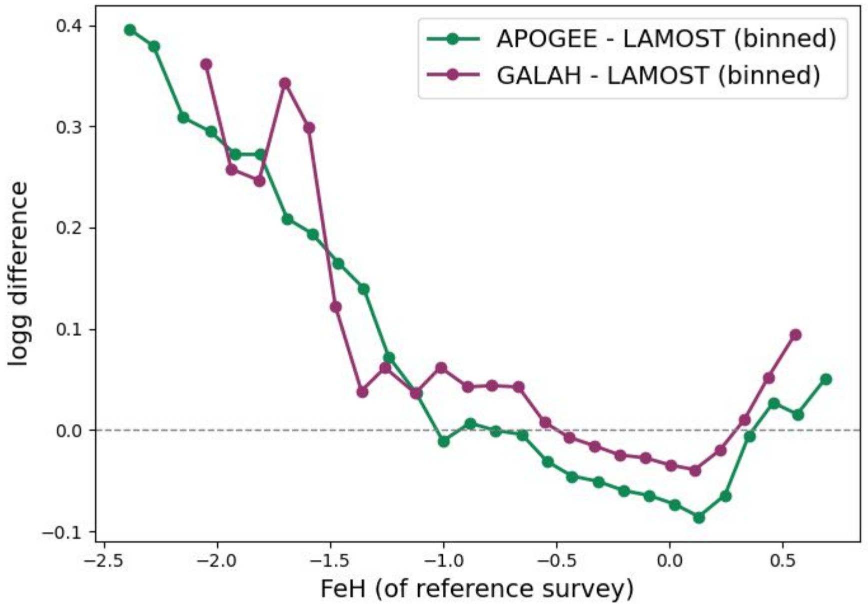
On y-axis: the difference between LAMOST and APOGEE/GALAH logg measurements.

Trend Detection

In this talk



On y-axis: the difference between LAMOST and APOGEE/GALAH [Fe/H] measurements.



Inputs to the models are normalized effective temperature (Teff/5040 K), surface gravity (logg), and metallicity ([Fe/H]).

We use the union of GALAH and APOGEE as the reference because LAMOST shows a consistent metallicity trend when compared to both surveys.

Method Comparison: Goals & Expectations

Complexity ↓

Polynomial fitting – easy to implement, interpretable, and fast, but relies on my intuition or … patience?

Symbolic Regression (in PySR implementation) – uses a genetic/evolution algorithm to find explicit analytical expressions.

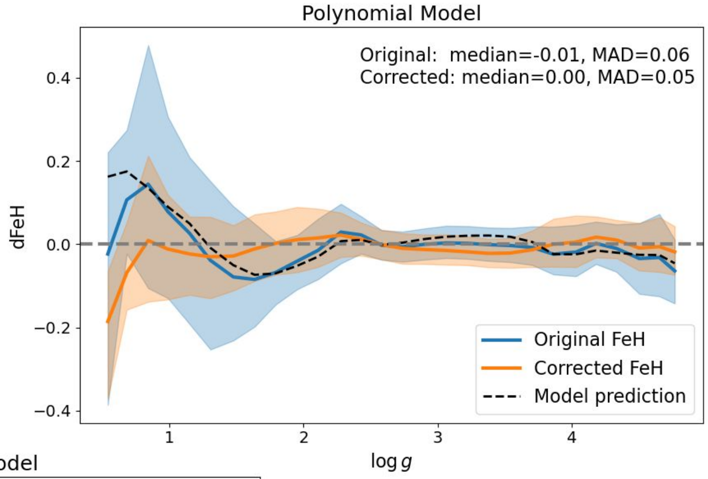
XGBoost – a more flexible, high-performance classical machine learning model, but less transparent.

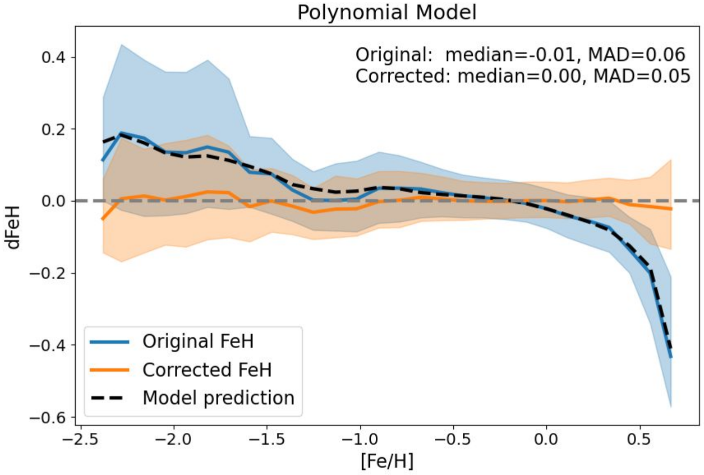
Transparency of the results ↑

The goal: to reduce systematic trends in [Fe/H] across the parameter space **without introducing artifacts**, ensuring smooth, consistent calibration.

Test case of FeH: Polynomial Fitting (4th order)

x-axis: logg





x-axis: [Fe/H]

The 4th-order polynomial with crossed terms effectively eliminates the trend dependent on [Fe/H].

Residual offsets remain at low logg (< 1 dex) and high temperatures (> 7000 K).

Binned plots show the 16th-84th percentile range as shaded areas.

Test case of FeH: Symbolic Regression (PySR)

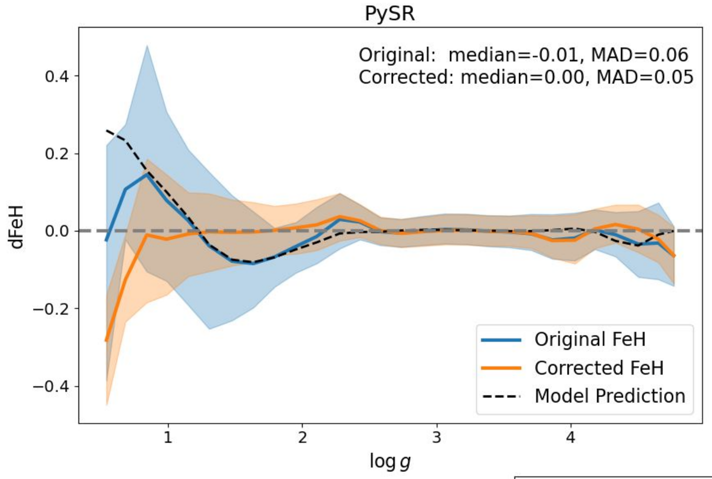
Star Dance

The PySR symbolic regression method shows similar performance to the 4th-order polynomial.

It also struggles with extreme values, especially at low logg (< 1 dex) and high Teff (> 7000 K). Moreover, it fails at the edges of metallicity ([Fe/H]<−2 dex or [Fe/H]>0.2 dex).

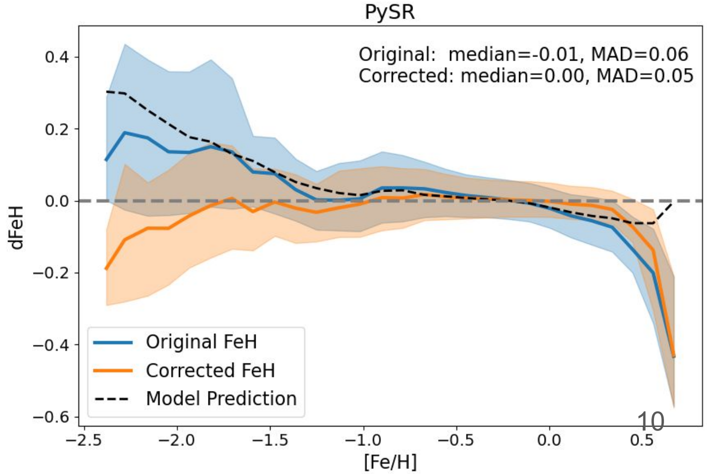
While it reduces the trend, it does not significantly improve calibration in these regions.

Or it needs significantly more computational time to converge.

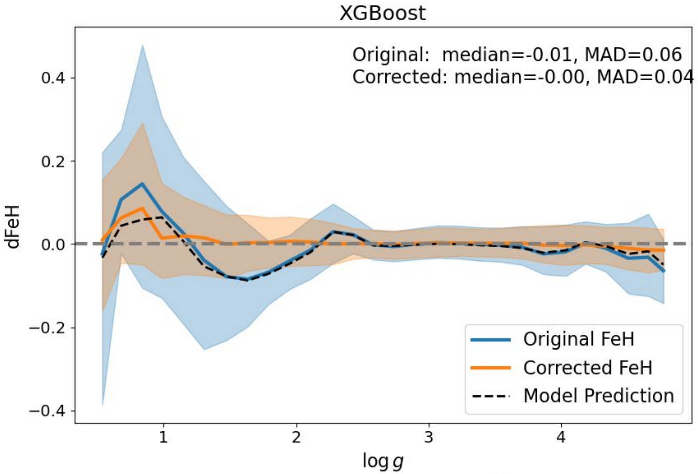


x-axis: logg

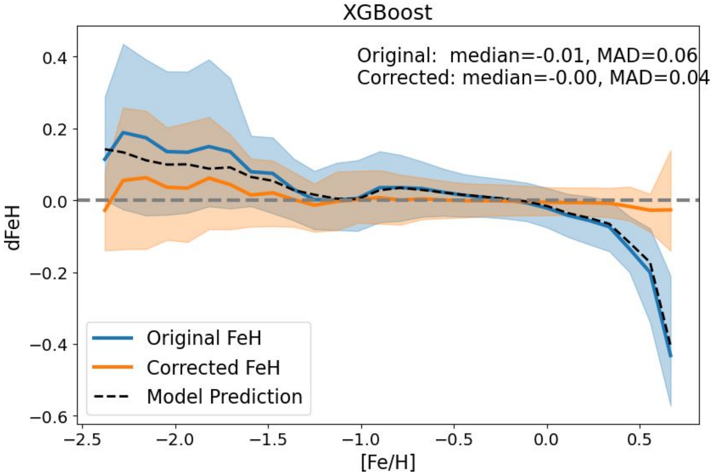
x-axis: [Fe/H]



Test case of FeH: XGBoost



x-axis: logg



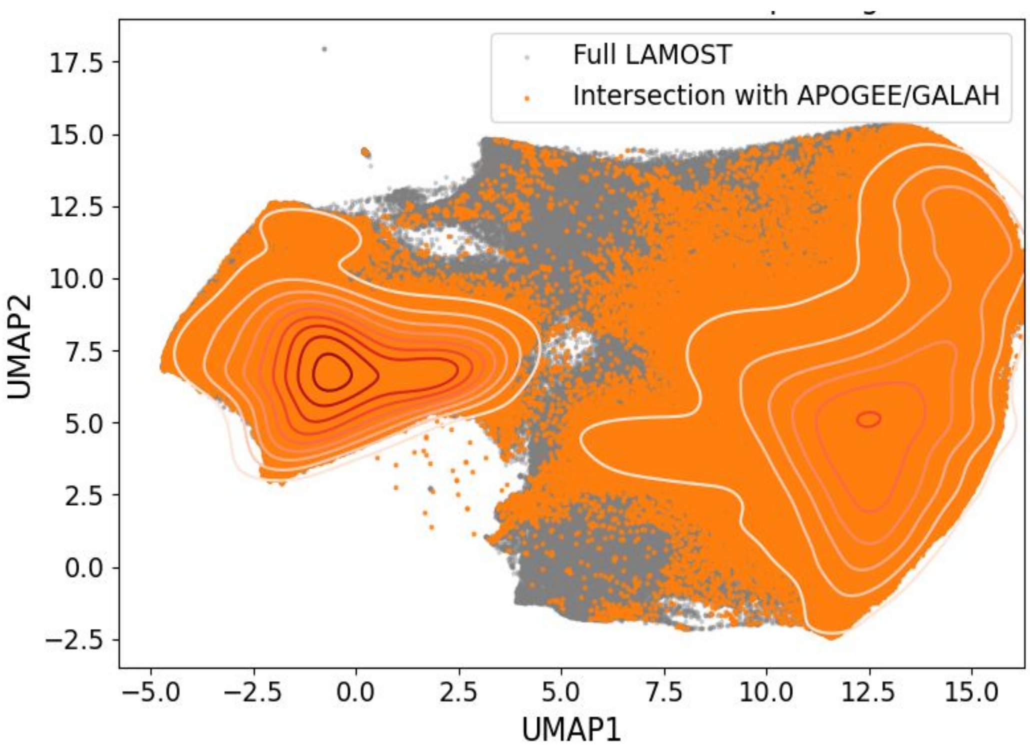
x-axis: [Fe/H]

XGBoost provides superior performance in correcting trends across the parameter space, including in regions with extreme [Fe/H], low logg (< 1), and high Teff (> 7000 K).

While the quartiles are higher at the edges, the median is well-calibrated across all parameters, showing consistent improvements in overall performance.

Star Dance

UMAP (Uniform Manifold Approximation and Projection): Coverage Exploration



UMAP exploration used the following parameters:

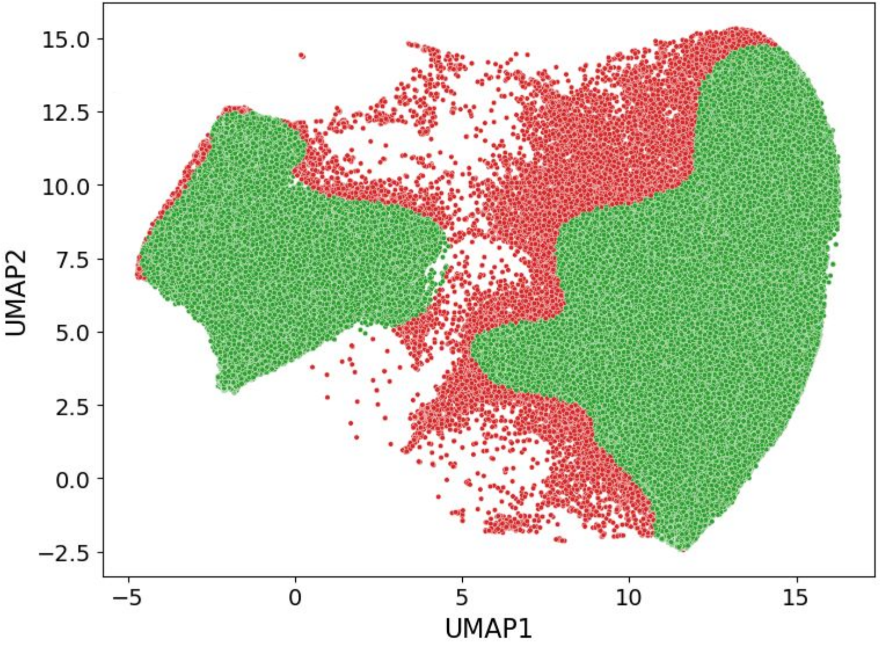
- 8 parameters: Teff, logg, [Fe/H], SNR, radial velocity, Gaia magnitude, distance and the estimation of extinction

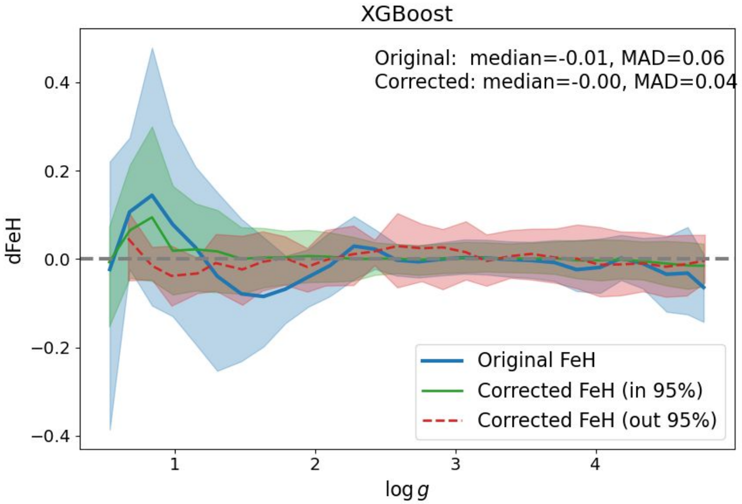
Key finding:

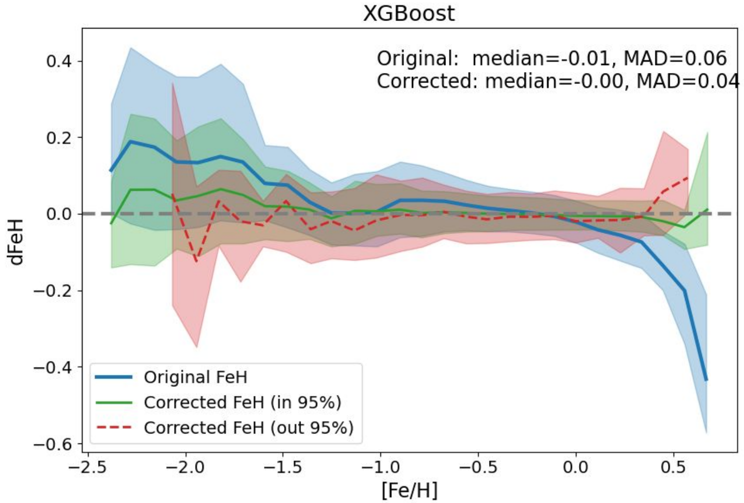
- Only part of the parameter space is well-covered by the LAMOST-GALAH/APOGEE intersection

Generalization Challenges

We divide intersection in two groups: most populated in the parameter space 95% (in green) and least populated 5% (in red):





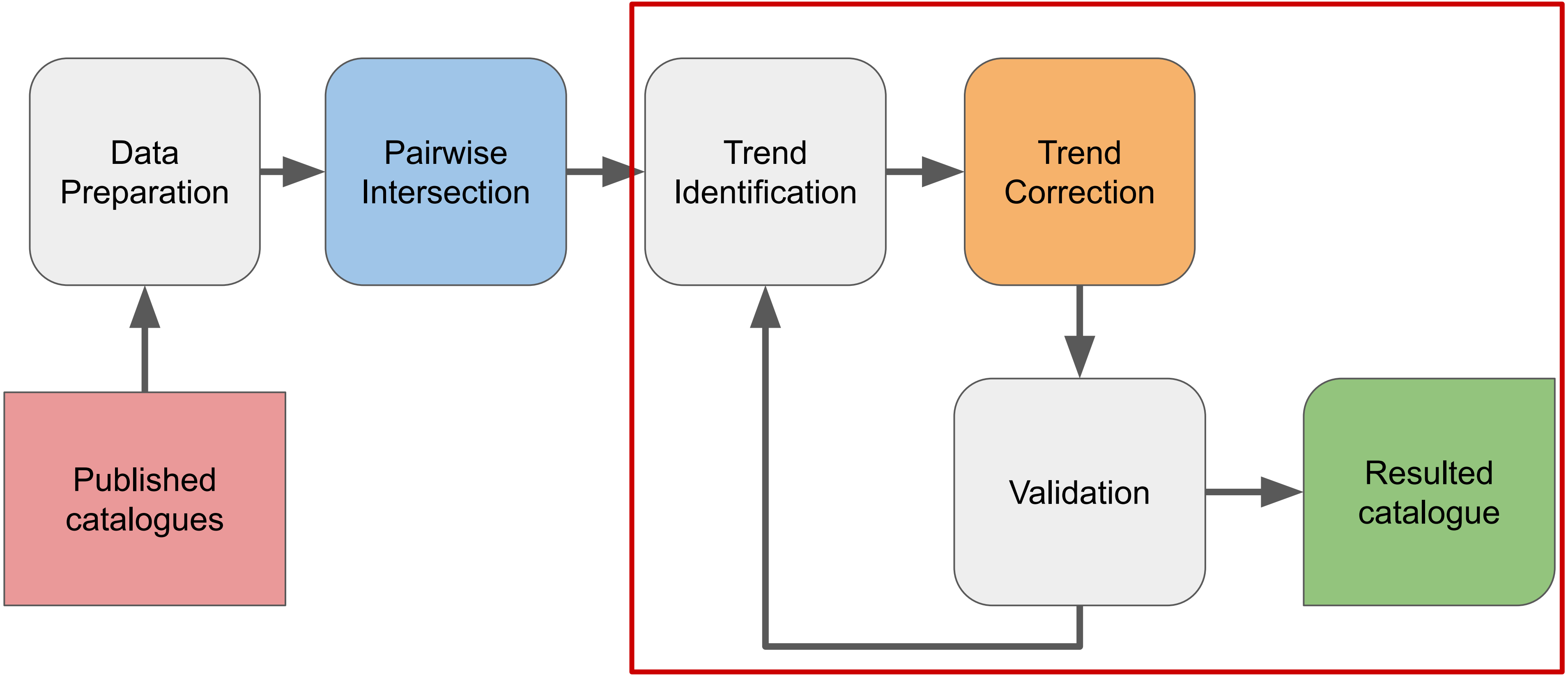


Summary

Star Dance

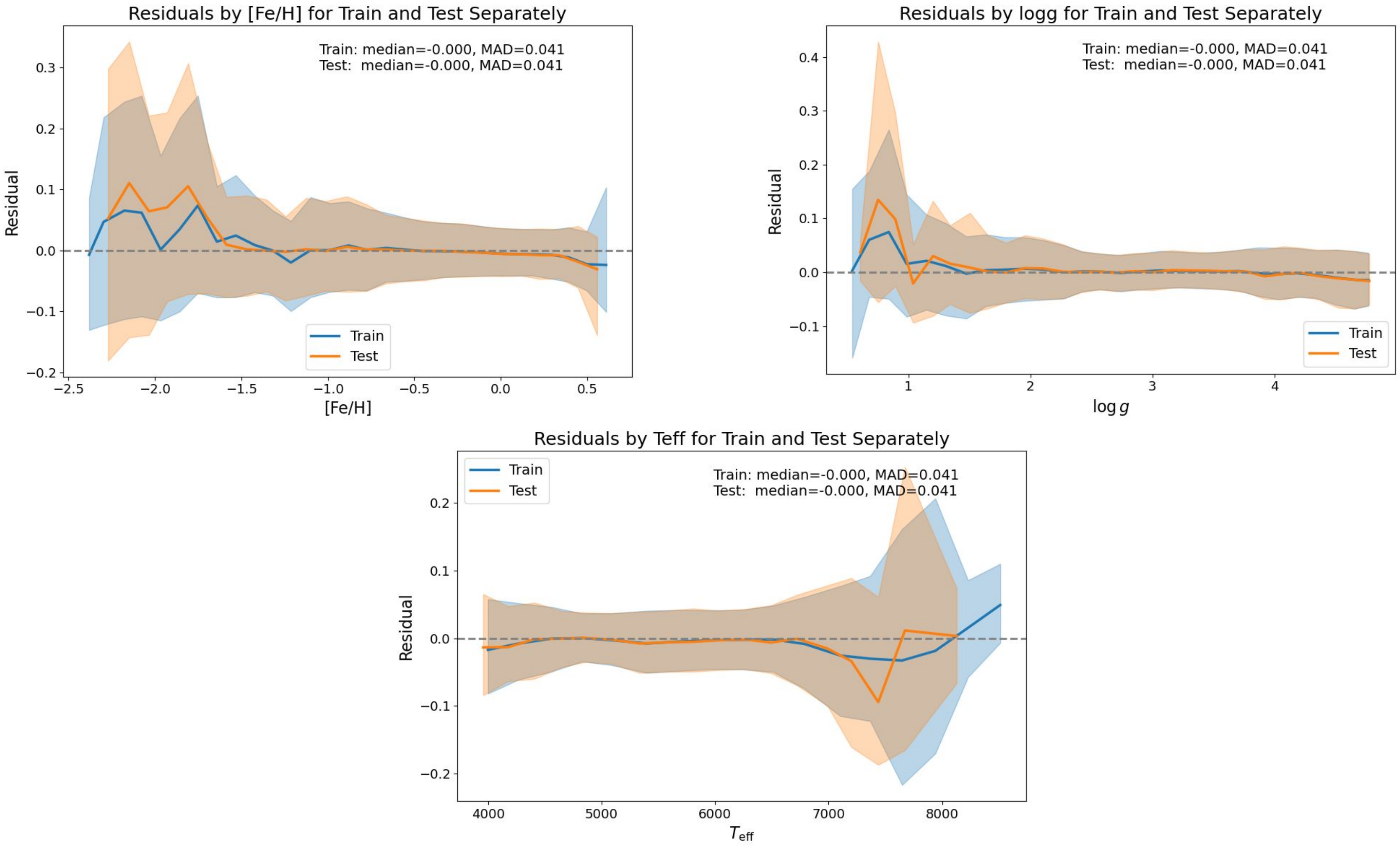
- ➔ LAMOST metallicities show systematic trends when compared with high-resolution surveys (GALAH/APOGEE)
- ➔ Polynomial correction (4th order with cross-terms): Removes most trends in [Fe/H], but struggles at low logg and high Teff.
- ➔ Symbolic Regression (PySR): Interpretable formulas, but less effective at edges of the parameter space. Also requires a lot of time to find the right formula.
- ➔ XGBoost: Best correction performance across full parameter range, effective in low-density / edge regions.
- ➔ UMAP exploration reveals that ~33% of LAMOST stars lie outside well-calibrated intersection. Generalization to full dataset needs further validation.

Future Work



Data Preparation
Pairwise Intersection
Trend Identification
Trend Correction
Published catalogues
Validation
Resulted catalogue

Thank you!
Questions?


Residuals by [Fe/H] for Train and Test Separately
Train: median=-0.000, MAD=0.041
Test: median=-0.000, MAD=0.041
Residual
[Fe/H]
Train
Test
Residuals by logg for Train and Test Separately
Train: median=-0.000, MAD=0.041
Test: median=-0.000, MAD=0.041
Residual
log g
Train
Test
Residuals by Teff for Train and Test Separately
Train: median=-0.000, MAD=0.041
Test: median=-0.000, MAD=0.041
Residual
T_{eff}
Train
Test