Rethinking Gaia XP Spectra for Machine Learning: Truncation, Basis Design, and Functional Methods

3rd Edition of the International Conference on Machine Learning for Astrophysics

Aidas Medžiūnas

Vilnius Univeristy
Faculty of Mathematics and Informatics
Institute of Applied Mathematics

August 31, 2026

Contents

Light spectra

Spectrum is dependant on / decribes:

- temperature and color of a source;
- chemical elements and molecules present;
- physical environment around or inside the object;
- velocity and motion relative to Earth.

High-resolution star spectra

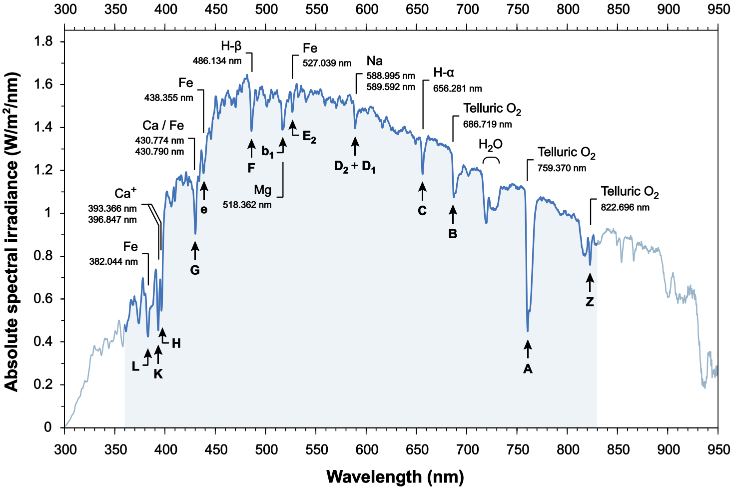


Figure 1: Example high-resolution star spectra.

Gaia mission

From 27 July 2014 to 15 January 2025, Gaia has made more than **three trillion observations of two billion stars and other objects** throughout our Milky Way galaxy and beyond, **mapping their motions, luminosity, temperature and composition**.

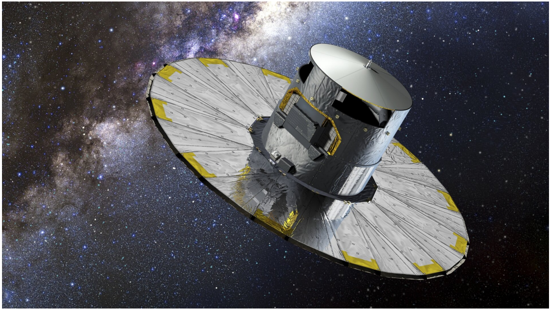


Figure 2: Artist's impression of the Gaia spacecraft

Gaia XP spectra

XP spectra are low-resolution spectrophotometric data produced by two low-resolution photometers:

- Blue Photometer (shorter wavelengths);
- Red Photometer (longer wavelengths).

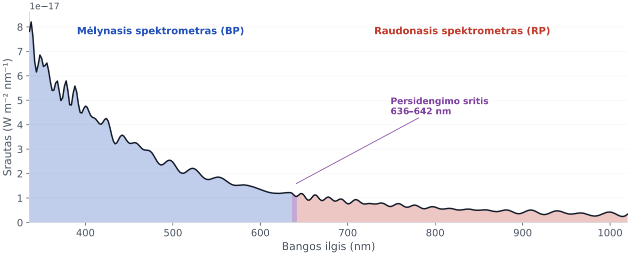


Figure 3: Example star XP spectra divided into BP and RP parts.

Contents

Types of Subdwarfs

Definition

Subdwarfs are stars that are less luminous than main-sequence stars of the same spectral type.

Depending on their physical origins and characteristics there are two types of subdwarfs:

- Cool subdwarfs;
- **Hot subdwarfs:**
  - high temperatures ($20,000 - 50,000$ K);
  - low masses (0.5 solar masses);
  - hydrogen-deficient atmospheres.

Hot subdwarf stars are an advanced evolutionary stage, formed when a red giant loses most of its hydrogen envelope before helium fusion begins.

Subdwarfs

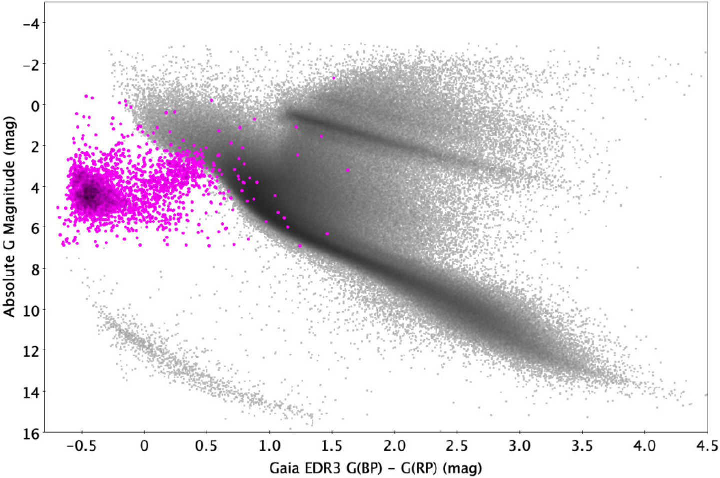


Figure 4: Hot-Subdwarfs location (pink) in Hertzsprung–Russell diagram.

Hot Subdwarfs

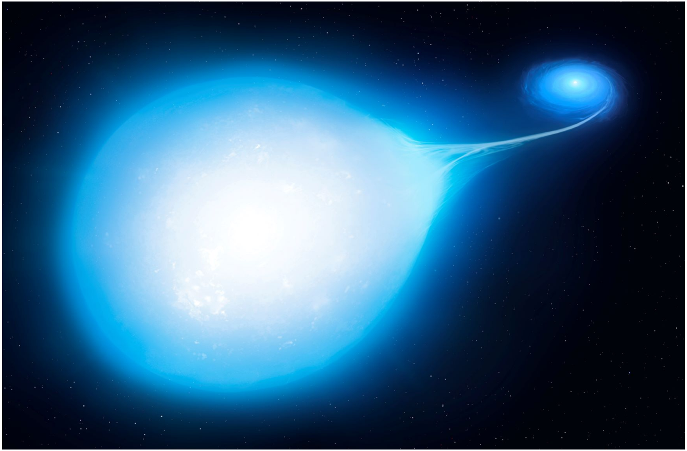


Figure 5: Mass transfer in binary star system (Mark Garlick, University of Warwick). Larger star represents the hot subdwarf star.

Investigation of the problem

Our Problem

Using Gaia XP spectra to identify binarity of hot subdwarf stars.

Approach

Apply supervised machine learning algorithms within a multivariate framework to perform binary classification.

Investigation of the problem

Our Problem

Using Gaia XP spectra to identify binarity of hot subdwarf stars.

Approach

Apply supervised machine learning algorithms within a multivariate framework to perform binary classification.

Realization #1

Use raw fluxes as input features to predict the class label.

Classifiers adapted from multivariate framework

In real-world applications, functional data are usually **not observed as continuous functions**, but rather as **discrete samples** representing those functions.
For example

$$X = \begin{pmatrix} x_{1,t_1} & \cdots & x_{1,t_K} \\ \vdots & \ddots & \vdots \\ x_{L,t_1} & \cdots & x_{L,t_K} \end{pmatrix}. \tag{1}$$

Due to the **highly correlated** nature of functional data, these discrete samples cannot be directly used as features in machine learning algorithms.

Classifiers adapted from multivariate framework

Realization #2 (general)

We approximate each function using a linear combination of orthogonal basis functions $\phi_k(t)$:

$$X_i(t) \approx \hat{X}_i(t) = \sum_{k=0}^{N-1} a_{ik}\, \phi_k(t), \tag{2}$$

where a_{ik} are coefficients for the i-th function.
Then we apply multivariate classifiers to the coefficient matrix

$$A_X = \begin{pmatrix} a_{1,0} & \cdots & a_{1,N-1} \\ \vdots & \ddots & \vdots \\ a_{L,0} & \cdots & a_{L,N-1} \end{pmatrix}. \tag{3}$$

Hermite functions

Definition

Let

$$\varphi_0(\theta) = \pi^{-\frac{1}{4}} e^{-\frac{\theta^2}{2}},$$

$$\varphi_1(\theta) = \sqrt{2}\pi^{-\frac{1}{4}} \theta e^{-\frac{\theta^2}{2}}$$

then Hermite functions satisfy the following recurrence relation

$$\varphi_n(\theta) = \sqrt{\frac{2}{n}}\,\theta\,\varphi_{n-1}(\theta) - \sqrt{\frac{n-1}{n}}\,\varphi_{n-2}(\theta). \tag{4}$$

Basically, the Hermite functions arise from Hermite polynomials by multiplication with a Gaussian function.

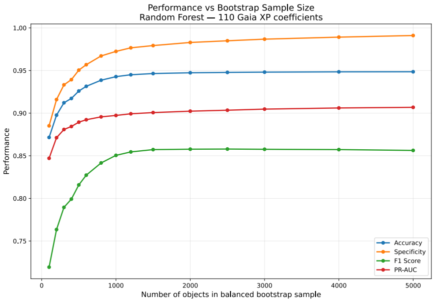
Hot-Subdwarf Binarity Identification

Algorithm 1 Gaia XP binary classification

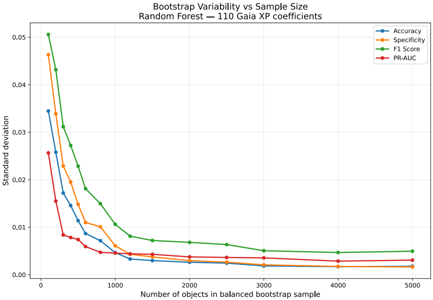
Require: Culpan-classified hot subdwarfs + Gaia DR3 XP spectra

1: **Label:** single / binary $\longrightarrow$ $y \in \{0, 1\}$
2: **Split:** training set | test set
3: **Basis representation:** $S(\lambda) \approx \sum_{k=0}^{109} c_k H_k(\lambda)$
4: **Features:** $\mathrm{x} = (c_0, \ldots, c_{109})$
5: **ML classifier:** $\mathrm{x} \longrightarrow P(\text{binary} \mid \mathrm{x})$
6: **Evaluation:** $\text{Sensitivity}, \ \text{Specificity}, \ \text{Accuracy}, \ \text{PR}-\text{AUC}$
7: **return** Binary candidates

Investigation of bootstrapping stability



(a) Performance metric mean



(b) Performance metric standard deviation

Figure 6: Convergence analysis of bootstrapping for `RandomForest`.

Investigation of model stability

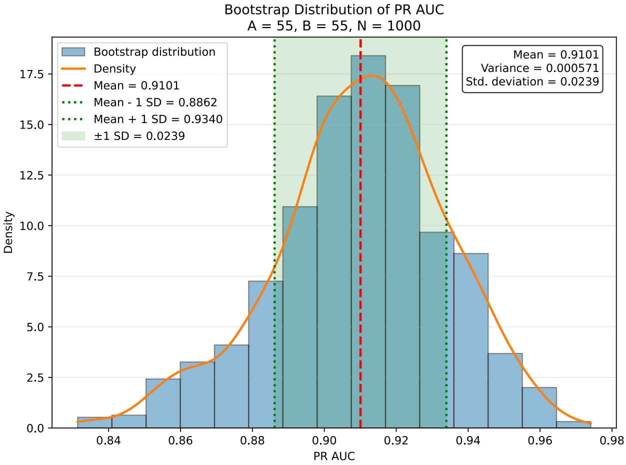
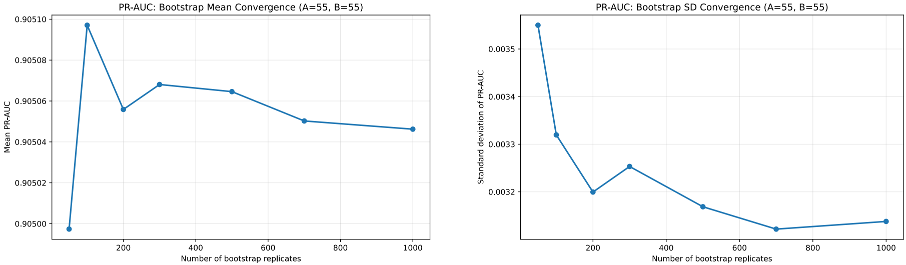


Figure 7: Bootstrap distribution of PR-AUC for the `RandomForest` classifier using 55 coefficients from each Gaia XP component.

Investigation of model stability



(a) PR-AUC Mean

(b) PR-AUC Standard deviation

Figure 8: Convergence analysis of bootstrapped distribution.

Contents

Gaia XP Spectra 1-D Truncation

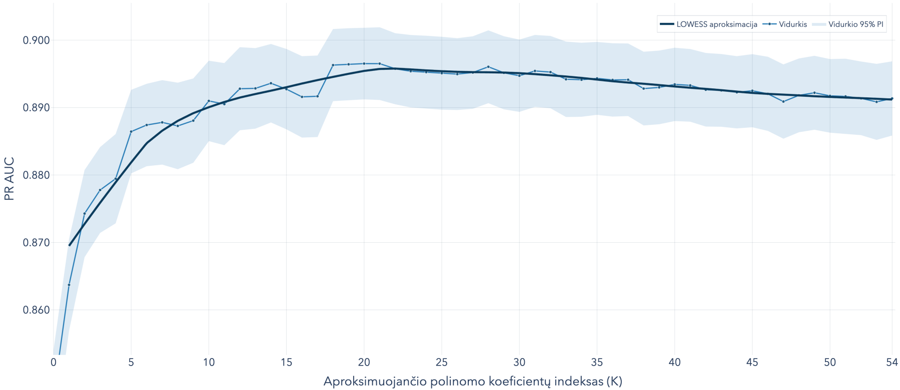


Figure 9: Dependence of the random forest PR-AUC on the number of coefficients K in the approximating polynomial.

Gaia XP Spectra 2-D Truncation

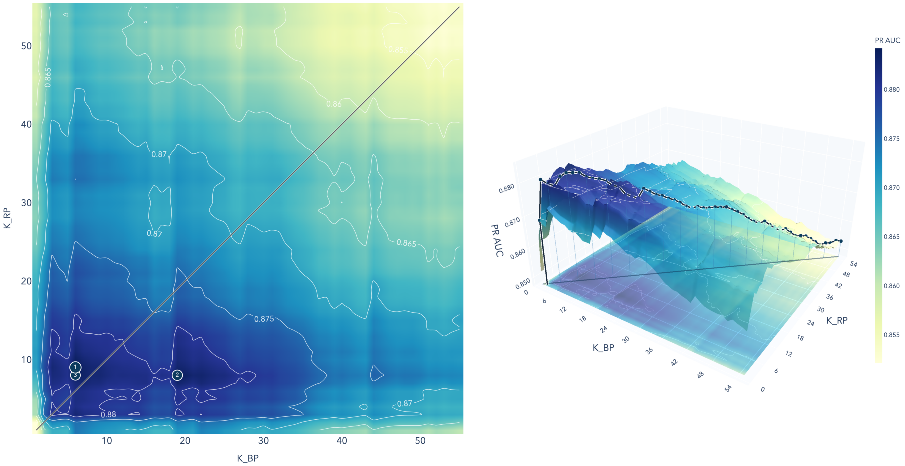


Figure 10: Dependence of the `RandomForrest` PR-AUC on the number of BP and RP coefficients in the approximating polynomials.

Contents

Mean Coefficient Spectrum

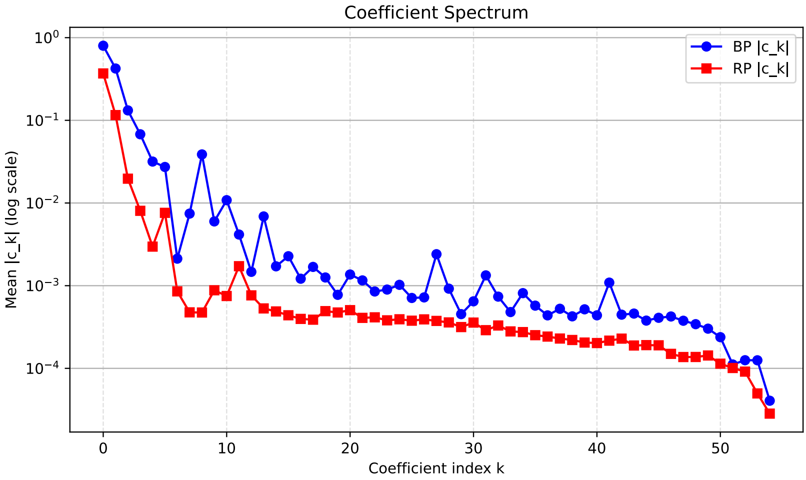


Figure 11: Mean Coefficient Spectrum of BP and RP.

Noise-based Thresholding of Coefficient Spectrum

We define mean adjusted coefficient spectrum:

$$\mathsf{spec}_k = \langle |c_k| \rangle - \mathrm{mean}\langle |c_k| \rangle.$$

Noise level is estimated from the tail:

$$\sigma = \mathrm{median}\left(\mathsf{spec}_{\text{last 20\% of coefficients}}\right).$$

We introduce SNR-based thresholds:

$$T_{\text{weak}} = 2\sigma, \quad T_{\text{strong}} = 5\sigma.$$

Classification:

$$\text{strong} : \mathsf{spec}_k > 5\sigma, \quad \text{noise} : \mathsf{spec}_k \leq 2\sigma$$

Noise-based Thresholding of Coefficient Spectrum

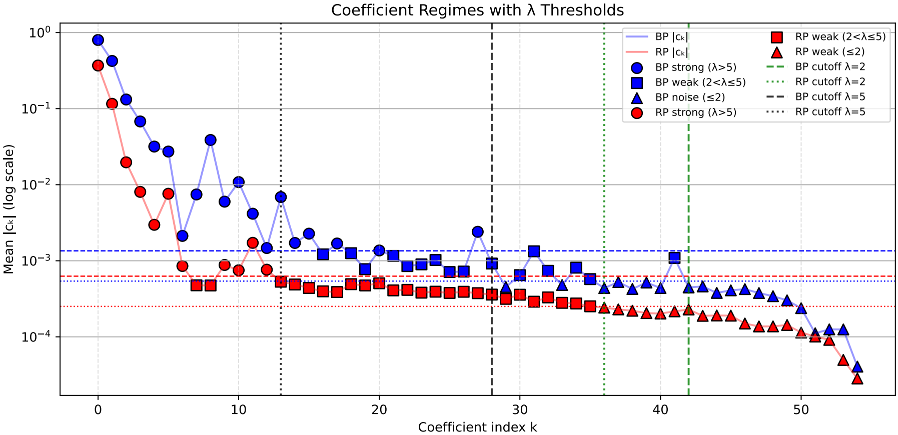


Figure 12: Coefficient spectrum and noise-based thresholds.

Eigen Value Energy Spectrum

Definition

Given a data matrix $X \in \mathbb{R}^{n \times d}$, we center it and compute the covariance matrix:

$$C = \mathrm{cov}(X)$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be the eigenvalues of C.
The *energy spectrum* is defined as the normalized eigenvalue distribution:

$$E_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}.$$

Eigen Value Energy Spectrum

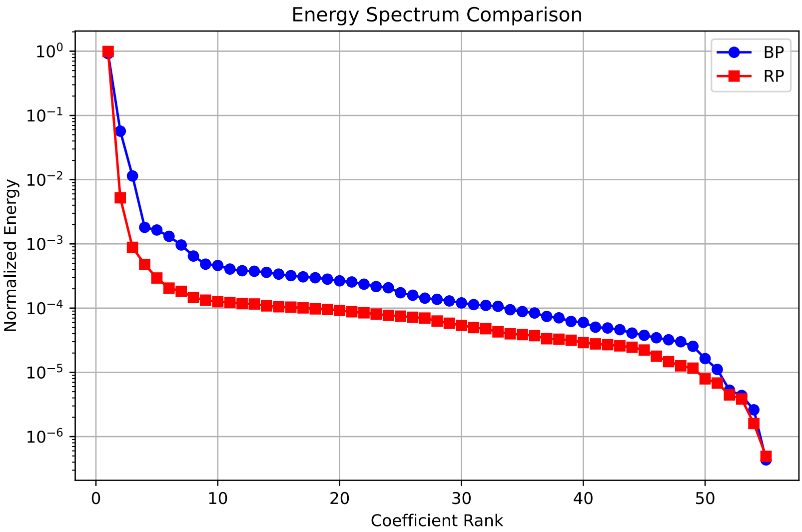


Figure 13: Eigen value energy spectrum of BP and RP coefficients.

Cumulative Energy

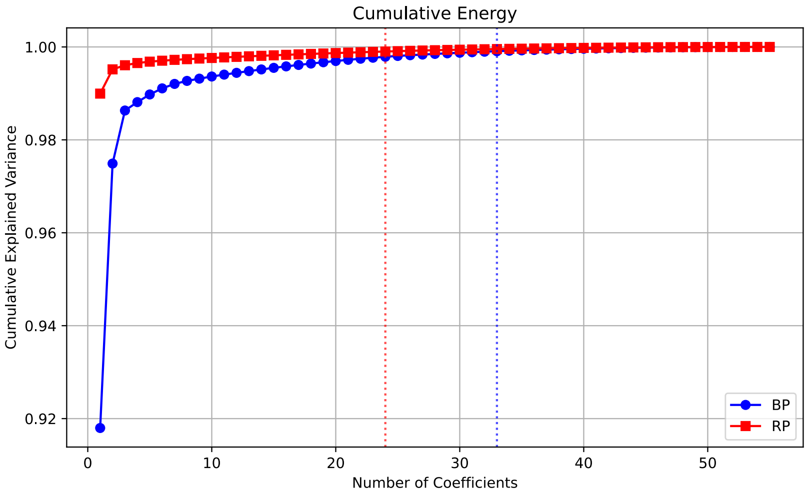


Figure 14: Cumulative energy of BP and RP coefficients eigen value spectrum.

Tail Flattening Elbow Detection

To detect the elbow, we analyze local decay:

$$s_i = |E_{i+1} - E_i|.$$

A tail region is defined as the last fraction of modes, where:

$$\bar{s}_{\text{tail}} = \mathrm{mean}(s_{i \in \text{last } 20\ \%})$$

The elbow index k^* is the first position where:

$$s_k \leq \alpha\, \bar{s}_{\text{tail}}, \quad \alpha = 3.$$

Tail Flattening Elbow Detection

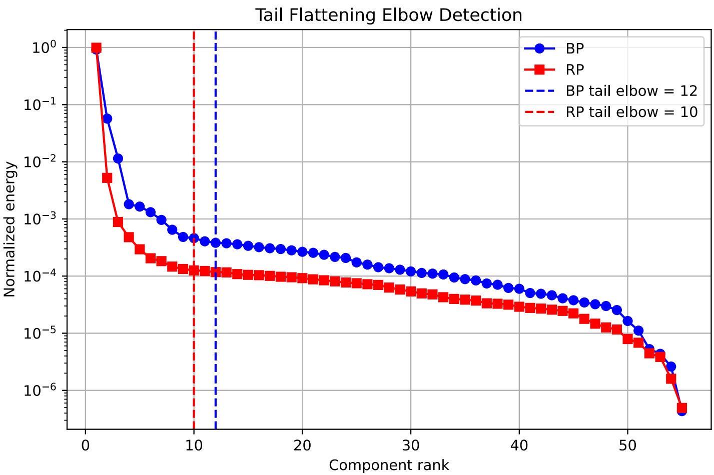


Figure 15: Elbows of cumulative energy of BP and RP coefficients eigen value spectrum.

Meta-Model for Predicting Optimal Cutoffs

Feature construction:

- **Cumulative energy cutoff:**

$$k_E : \sum_{i \le k_E} E_i \ge 0.999$$

- **Noise-based spectrum thresholds:**

$$T_\lambda = \lambda\sigma, \quad \lambda \in \{2, 5\}$$

  defining weak/strong regimes from coefficient magnitudes.

- **Tail flattening elbow:**

$$s_k \le \alpha\, \bar{s}_{\text{tail}}, \quad \alpha = 3$$

  detecting transition to flat spectrum.

Model:

$$(\text{best}_A, \text{best}_B) = f(\text{energy}, \text{threshold}, \text{elbow})$$

using `HistGradientBoostingRegressor` in a multi-output setting.

Regression Performance Summary

Metric	Best BP	Best RP
MAE	3.1789	5.4700
RMSE	3.9380	8.0929
MSE	15.5078	65.4957
R^2	0.9605	0.8599

Table 1: Evaluation metrics for multi-output regression model

Contents

Conclusion

- Gaia XP spectra contain sufficient information to detect binarity of hot subdwarfs;
- Truncation of Gaia XP Hermite function coefficients representation improves model performance;
- Truncation of Gaia XP Hermite function coefficients representation is asymmetric;
- Meta-Model based on cumulative energy cutoff, noise-based spectrum thresholds and tail flattening elbow is fast and effective tool for predicting optimal cutoffs.

Additional Remarks

This talk (in part) was based on paper:

- Ambrosch, M. and Viscasillas Vázquez, C. and Solano, E. and Ulla, A. and Pérez-Couto, X. and Pérez-Fernández, E. and Medžiunas, A. and others, "Detection of hot subdwarf binaries and sdB stars using machine learning methods and a large sample of Gaia XP spectra", *Astronomy & Astrophysics*, 2026.

Research is conducted as part of the project *Unveiling the Nature of Hot Subdwarfs through Spectroscopy and Machine Learning*, supported by the Vilniaus universiteto mokslo skatinimo fondas.