Parameters of 220 million stars from Gaia BP/RP spectra

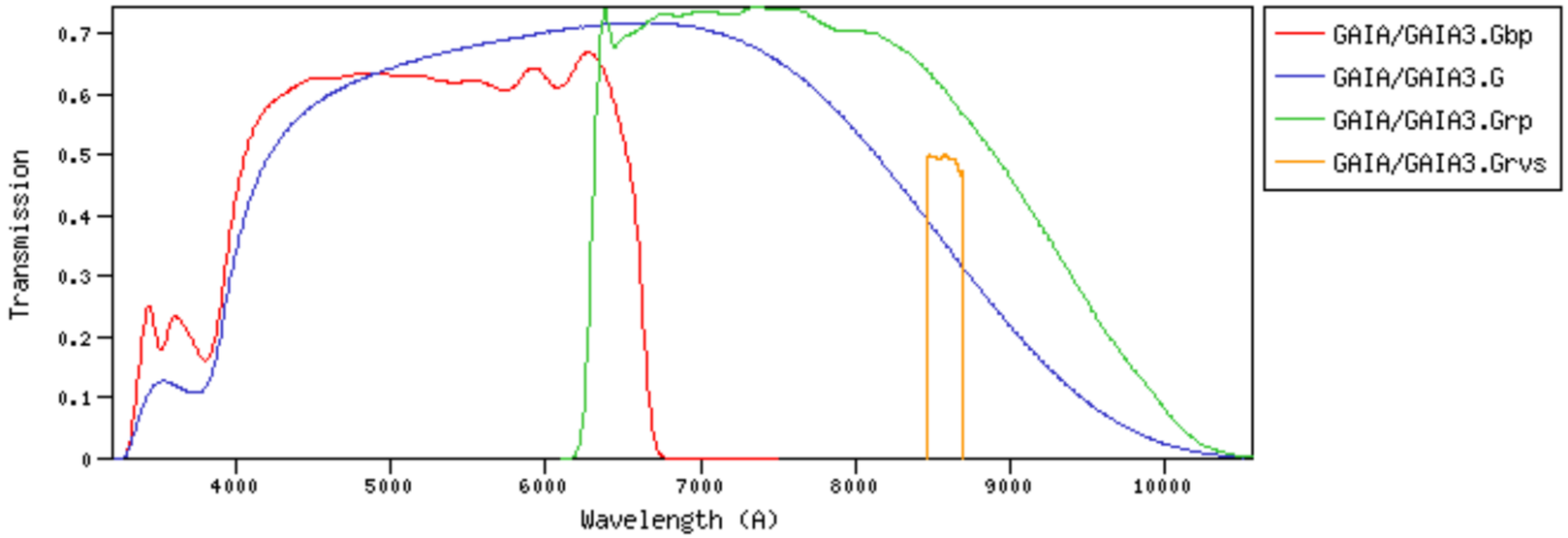
Xiangyu Zhang, Gregory M. Green, Hans-Walter Rix

Reporter: Zheng Yu

Introduction

Gaia Data Release 3 (GDR3, Gaia Collaboration et al. 2022) includes over 220 million flux-calibrated, low-resolution, optical stellar spectra, which provide a unique opportunity to map the properties of stars and dust throughout a large volume of the Milky Way.

- the “Blue Photometer” (BP) 330-680 nm
- the “Red Photometer” (RP) 640-1050 nm
- effective resolution elements 110
- corresponding to a resolution $\sim 50 - 160$



Introduction

How to determine stellar parameters from Gaia XP spectra?

Using physical models
- Predict spectra from fundamental properties of stars (mass, age, abundances, etc.)
- Fit observed XP spectra with isochrone models, spectral atmosphere models, and extinction law
- Drawback: sensitive to accuracy of stellar models, difficult to get information from low-resolution XP spectra

Using empirical forward models
- Learn an empirical forward model from a subset of stars with high-resolution spectra
- Apply this model to all XP spectra, infer stellar types, distances, and extinctions
- Advantage: interpretable, exploit measurement uncertainties, degrade gracefully in low signal-to-noise situations

Using machine learning models
- Train a machine learning model to directly predict stellar parameters from XP spectra
- Similar to the second method, but do not use a forward model
- Advantage: automatic, learn features from spectra
- Drawback: degrade faster in low signal-to-noise situations, hard to explain and validate

Introduction

Empirical forward models

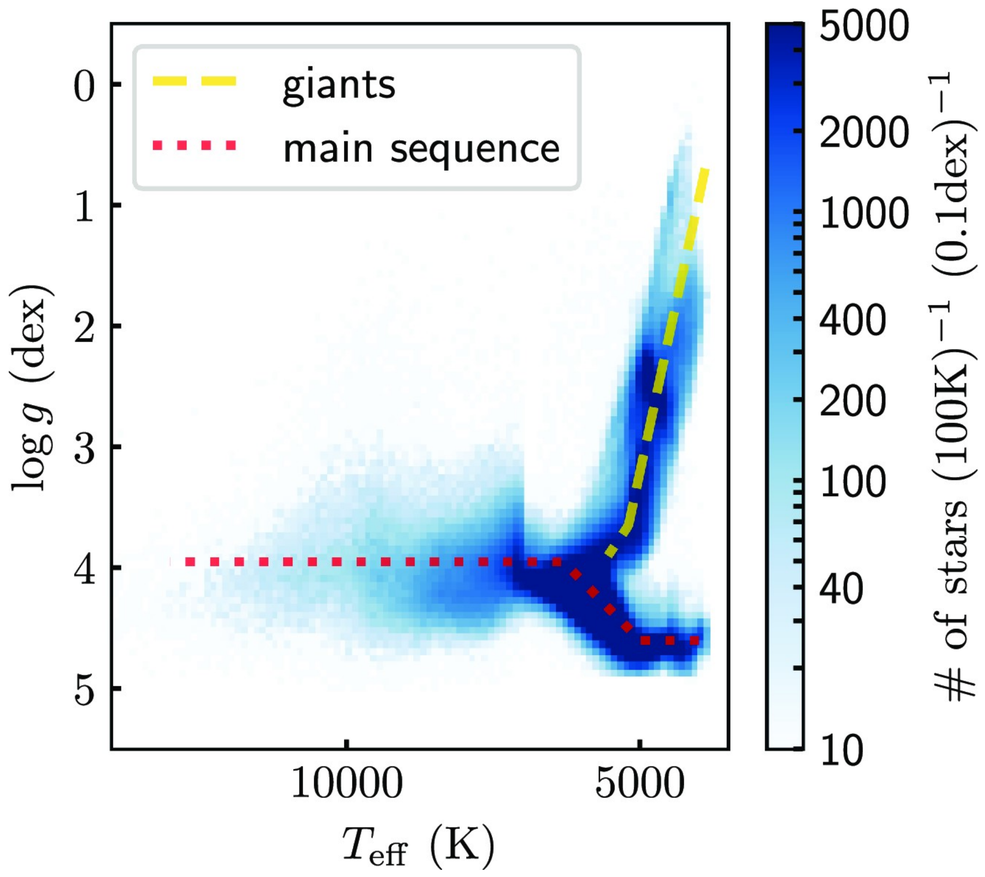
High-quality measured spectra：　LAMOST DR8（1% of *XP*）
Near-infrared photometry : 2MASS(J, H and Ks)　　WISE(W1, W2)

$$f_{\mathrm{pred}}(\lambda \mid \Theta, \varpi, E) = f_{\mathrm{abs}}(\lambda \mid \Theta)\varpi^{2} \exp\left[-E\, R(\lambda)\right]$$

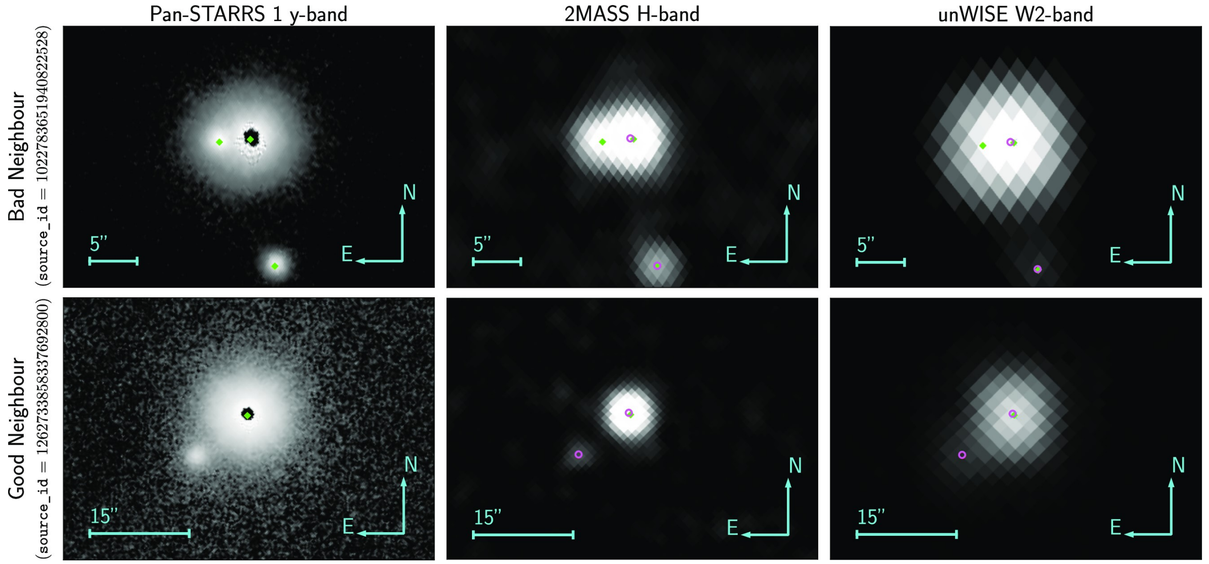
The structure of the model encodes certain reasonable assumptions:
(i) Flux falls with the square of distance.
(ii) Dust imposes a wavelength-dependent optical depth.
(iii) In the absence of dust (and at a standard distance), the stellar spectrum is purely a function of stellar atmospheric parameters.

Data

Gaia&LAMOST(2,575,354 stars)　177 mas × 59 mas　2″ × 2″　FWHM 6 .1″6 .4″



the “Hot Payne” (Xiang et al. 2022)



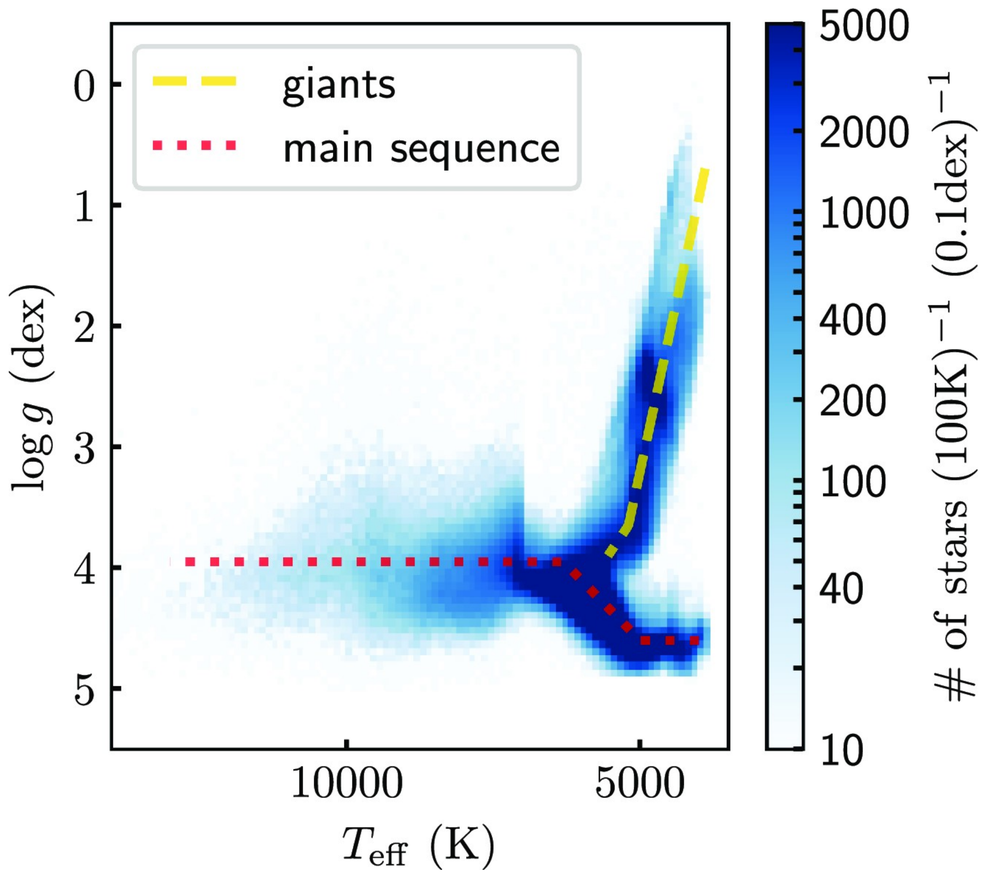
norm_dg > −5　norm_dg > −10

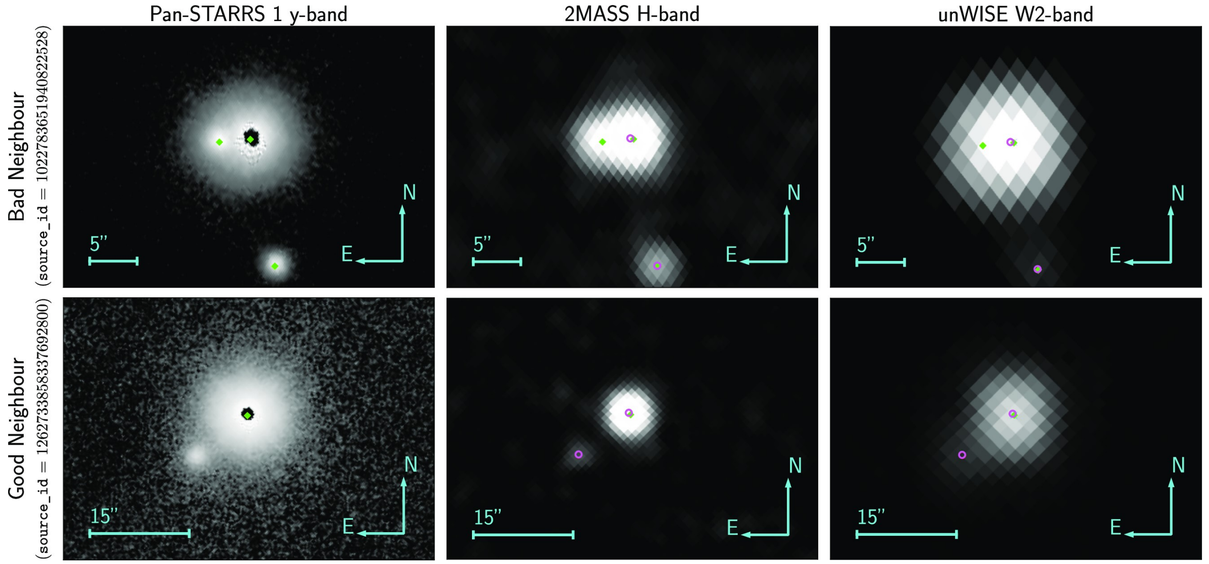
$$max\{\frac{\Delta G_i}{\text{mag}} - \frac{\theta_i}{\text{arcsec}}\}$$

a measure of crowding, with larger values indicating the presence of closer, brighter neighbors.

Data

Gaia&LAMOST(2,575,354 stars) 177 mas × 59 mas 2 ″ × 2 ″ FWHM 6 .1 ″6 .4 ″



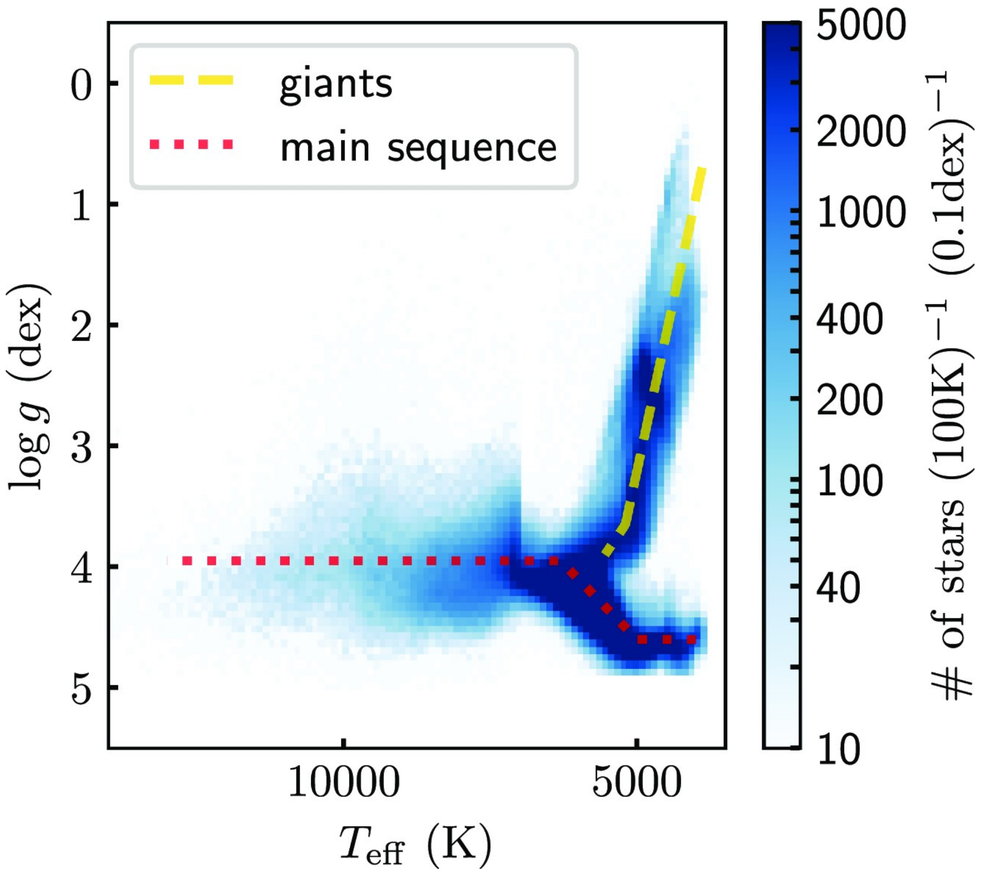


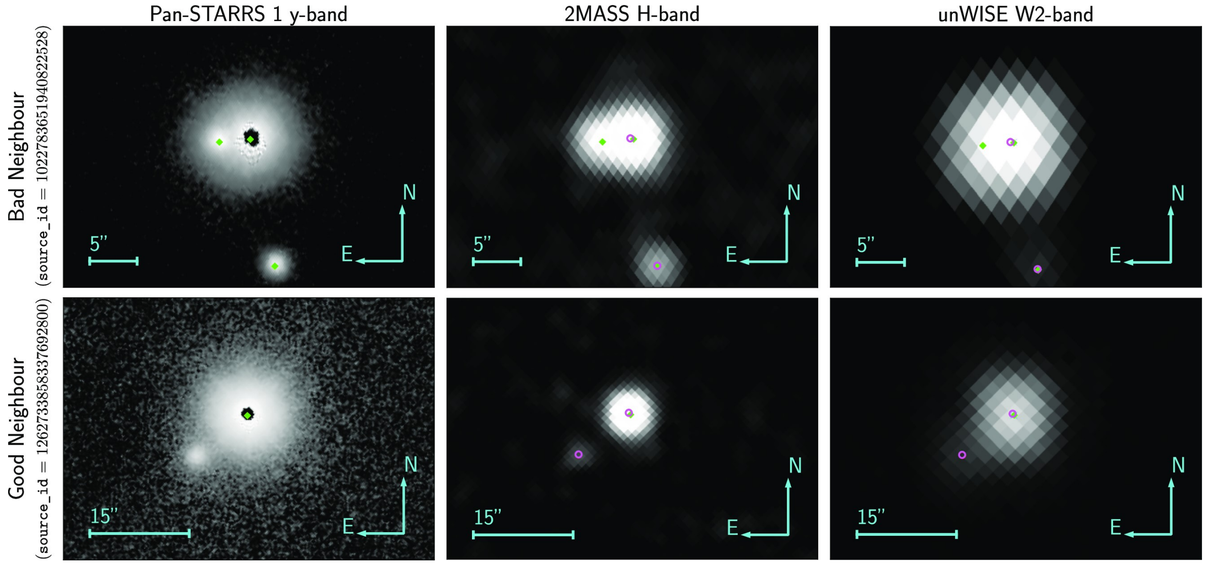
$$f_\lambda = (3631\,\mathrm{Jy})\, c\lambda_0^{-2} 10^{-0.4(m+\Delta m)}$$

m is the reported Vega magnitude;
c is the speed of light;
Δm is the AB-Vega offset (J 0.91, H 1.39 and Ks 1.85 mag)
Lamda is the central wavelength (1.235, 1.662 and 2.159 μm

Data

Gaia&LAMOST(2,575,354 stars) 177 mas × 59 mas 2″ × 2″ FWHM 6.1″6.4″





$$f_\lambda = (3631\,\mathrm{Jy})\, c\lambda_0^{-2} 10^{-0.4(m+\Delta m)}$$

m is the reported Vega magnitude;
c is the speed of light;
Δm is the AB-Vega offset (2.699 and 3.339 mag)
Lamda is the central wavelength (3.3526 and 4.6028 μm

Data

Gaia & LAMOST & 2MASS & unWISE

- LAMOST SNR of greater than 20 in -, -, and -bands.
- LAMOST uncertainties of less than 500 K in eff, and less than 0.5 dex in [Fe/H] and log .
- Well constrained parallaxes: parallax_over_error > 3.
- Reliable Gaia astrometry: fidelity_v2 > 0.5.
- Low BP/RP flux excess: bp_rp_flux_excess < 1.3.

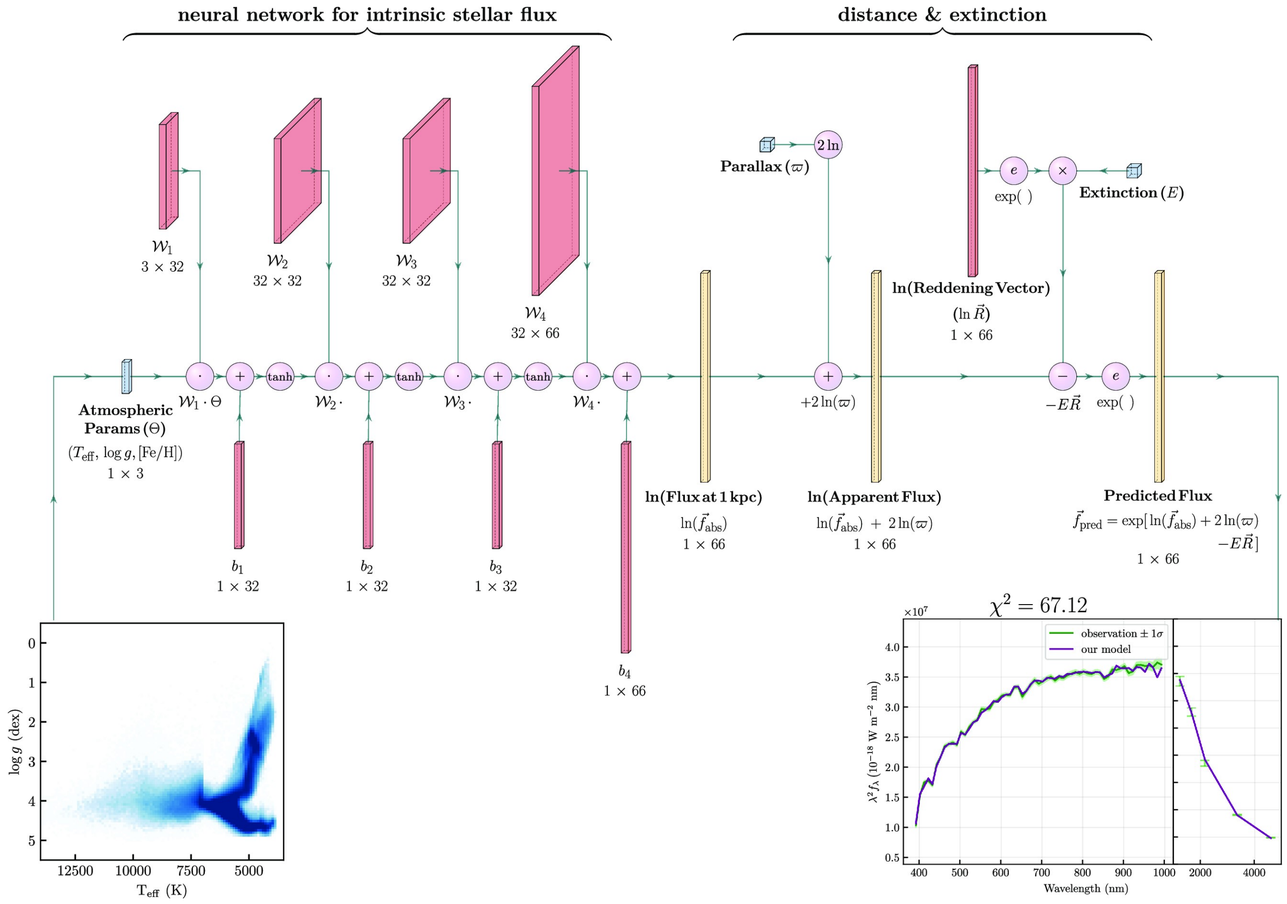
2,575,354 sources

a training set (80% of sources) and a validation set (20% of sources)

METHOD

$$\vec{f}_{\text{pred}} = \vec{f}_{\text{abs}}(\Theta, \mathcal{W}, b)\varpi^2 \exp\left(-E\vec{R}\right)$$

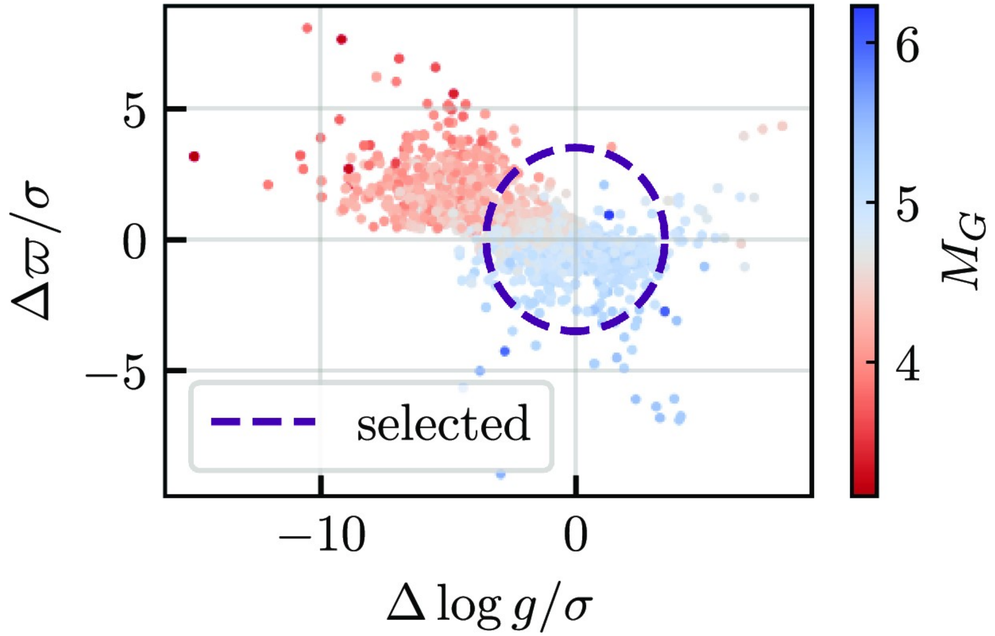
W and b represent all trainable neural-network weights and biases, respectively, in the absolute flux model

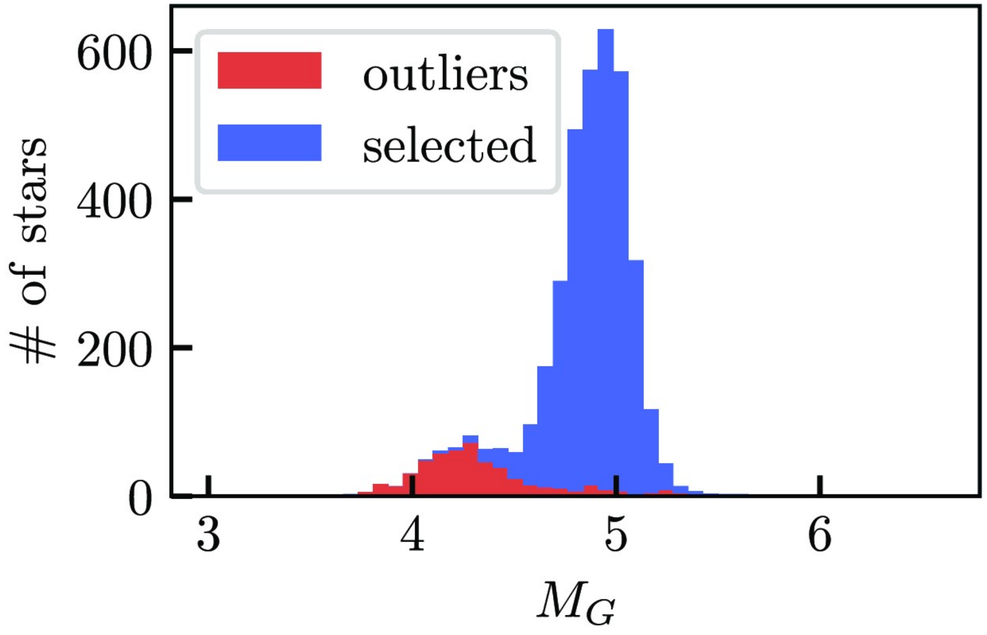


METHOD

$$\vec{f}_{\text{pred}} = \vec{f}_{\text{abs}}(\Theta, \mathcal{W}, b)\varpi^2 \exp\left(-E\vec{R}\right)$$

"Self-cleaning" process



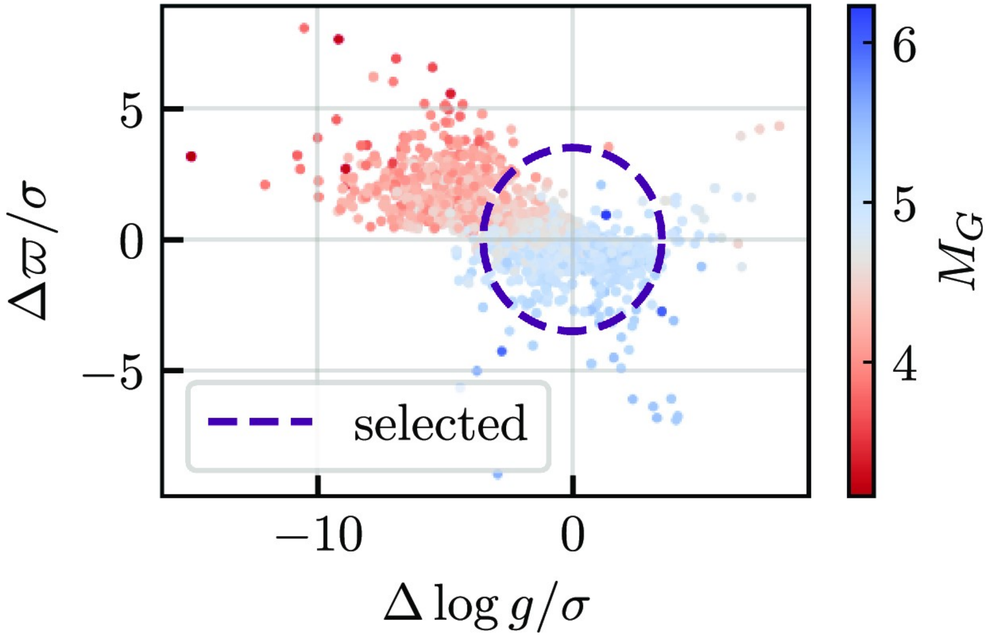


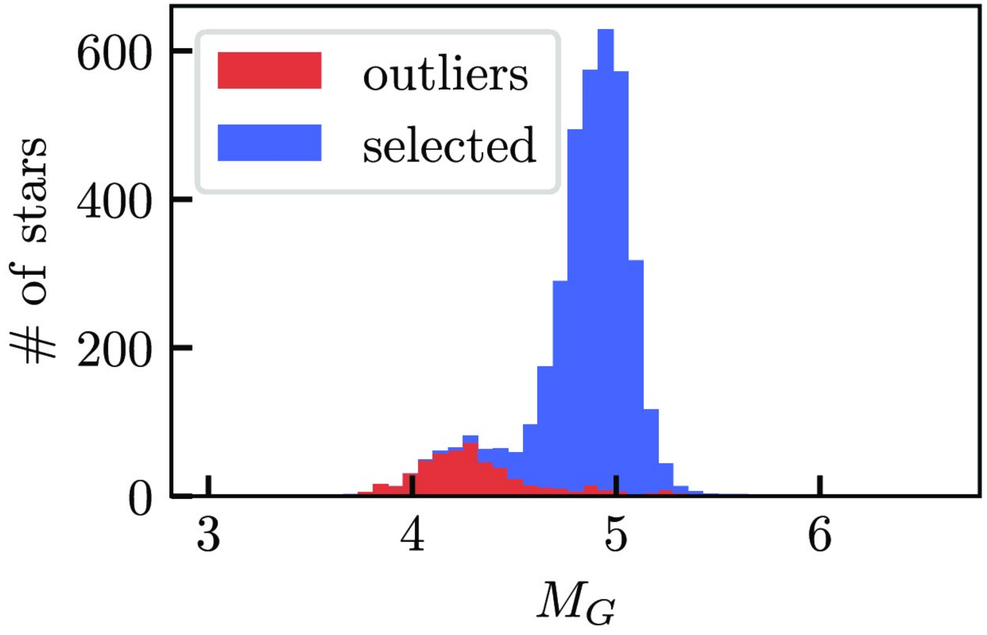
1. $(\Delta \log g/\sigma)^2 + (\Delta\varpi/\sigma)^2 > 3.5^2$

METHOD

$$\vec{f}_{\rm pred} = \vec{f}_{\rm abs}(\Theta, \mathcal{W}, b)\varpi^2 \exp\left(-E\vec{R}\right)$$

"Self-cleaning" process





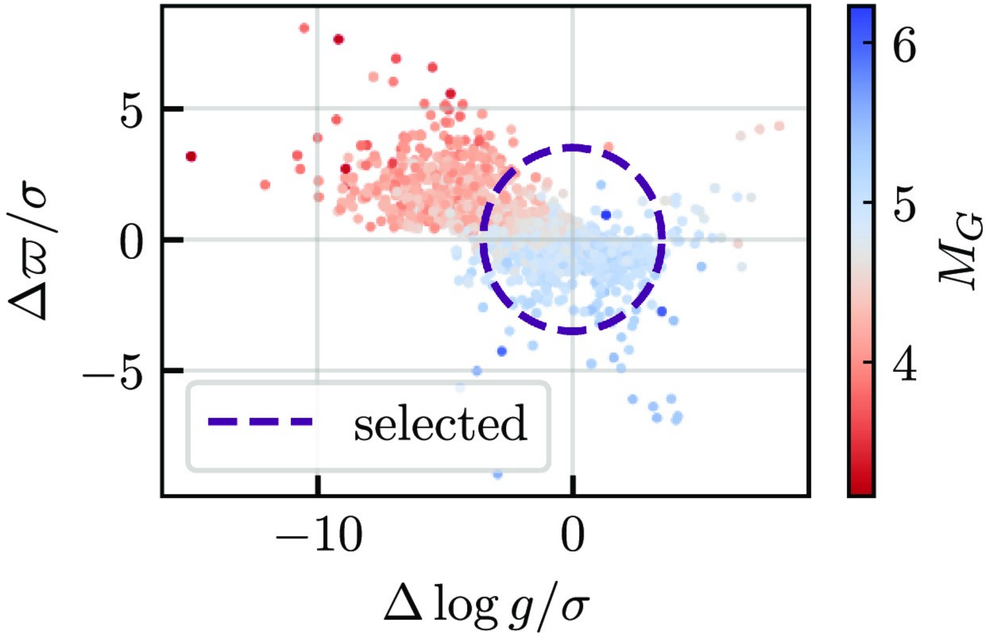
2. Identifying outliers

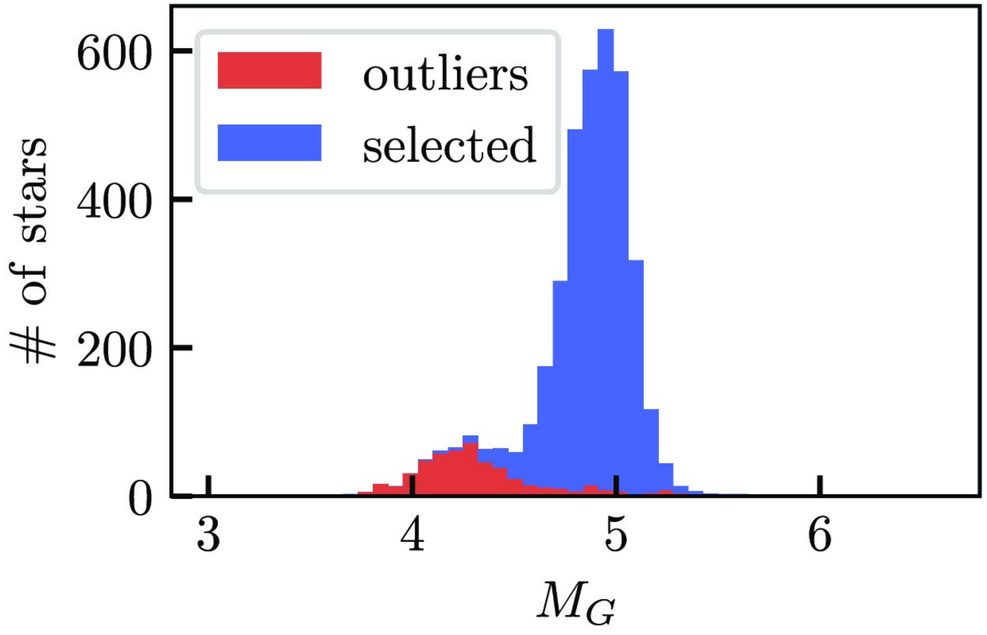
the residual between the observed and predicted flux at any wavelength > 4sigma

METHOD

$$\vec{f}_{\rm pred} = \vec{f}_{\rm abs}(\Theta, \mathcal{W}, b)\varpi^2 \exp\left(-E\vec{R}\right)$$

"Self-cleaning" process





2. Identifying outliers

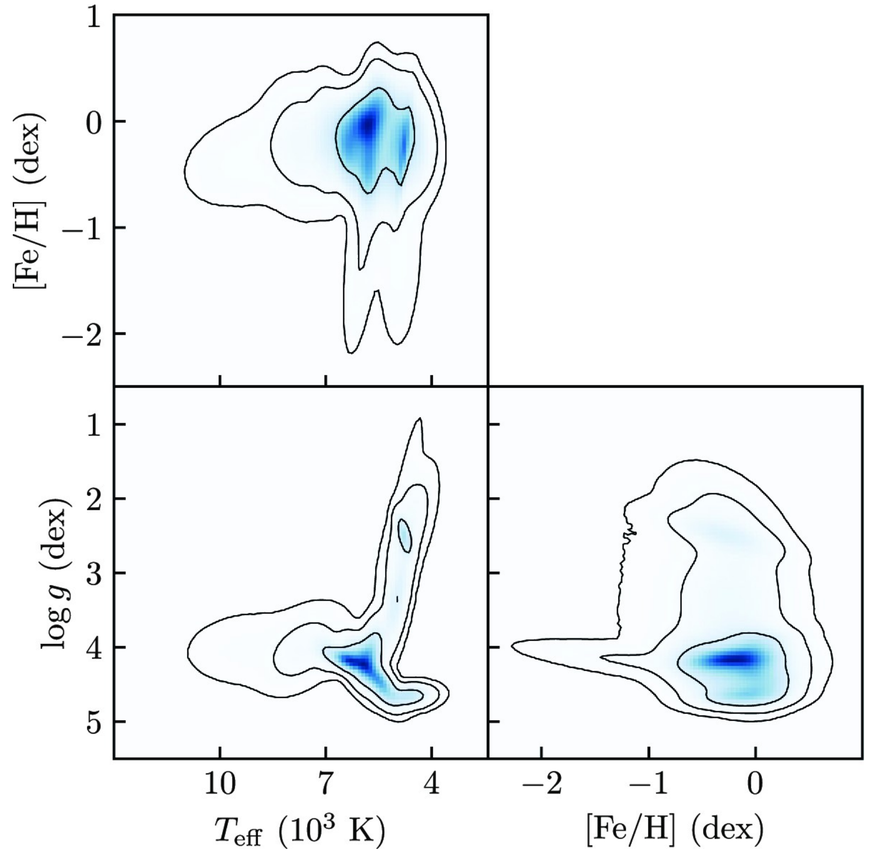
1,556,666 sources, or 75.3% of the training set

the residual between the observed and predicted flux at any wavelength > 4sigma

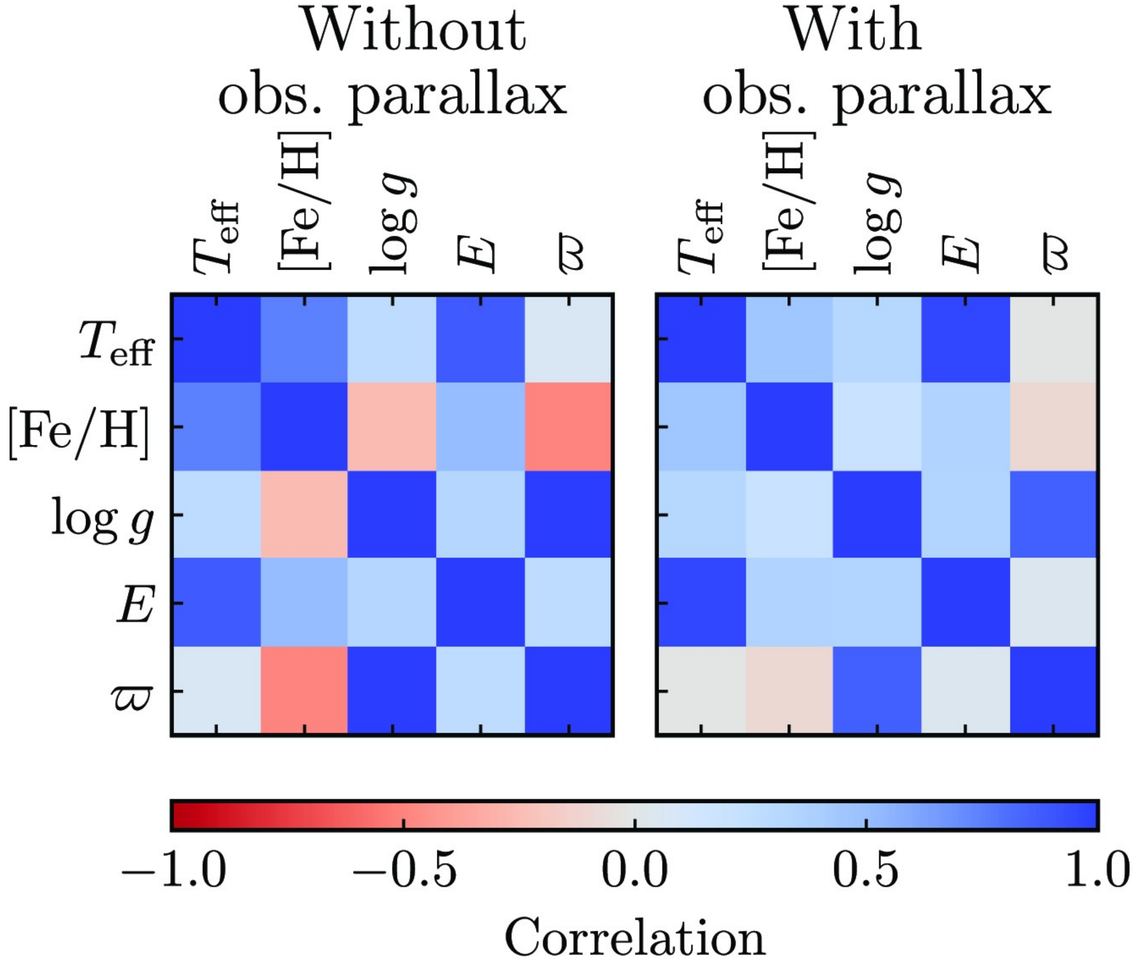
METHOD

$$\vec{f}_{\rm pred} = \vec{f}_{\rm abs}(\Theta, \mathcal{W}, b)\varpi^2 \exp\left(-E\vec{R}\right)$$

The prior assumption

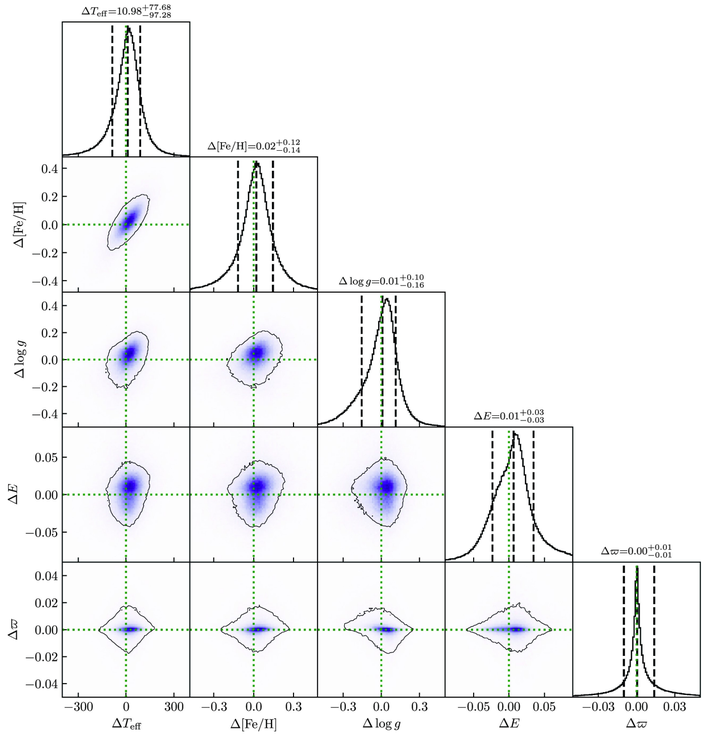


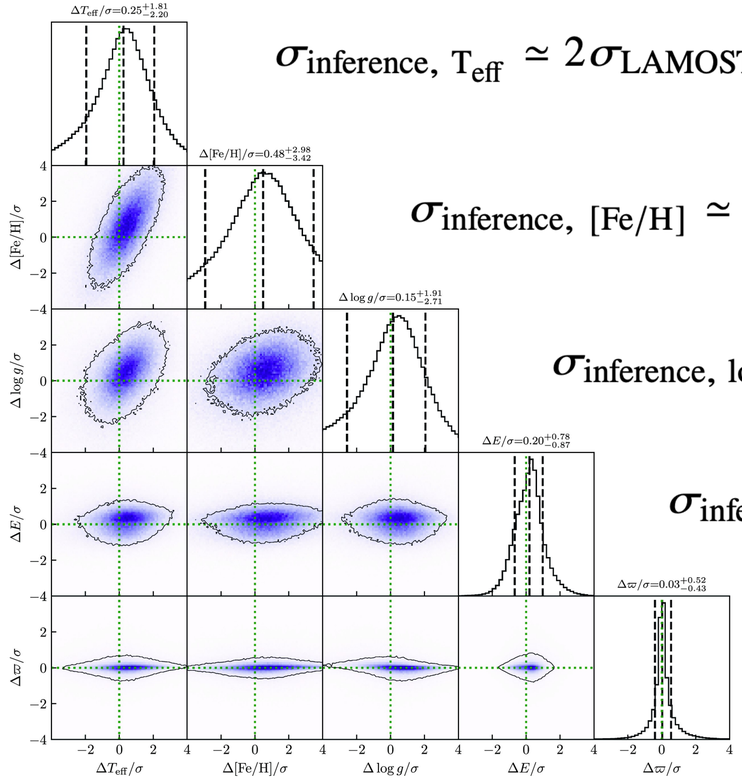
a Gaussian mixture model with N=16



Result

$$\vec{f}_{\text{pred}} = \vec{f}_{\text{abs}}(\Theta, \mathcal{W}, b)\varpi^2 \exp\left(-E\vec{R}\right)$$





$\sigma_{\text{inference, T}_{\text{eff}}} \simeq 2\sigma_{\text{LAMOST, T}_{\text{eff}}}$

$\sigma_{\text{inference, [Fe/H]}} \simeq 3\sigma_{\text{LAMOST, [Fe/H]}}$

$\sigma_{\text{inference, log g}} \simeq 2\sigma_{\text{LAMOST, log g}}$

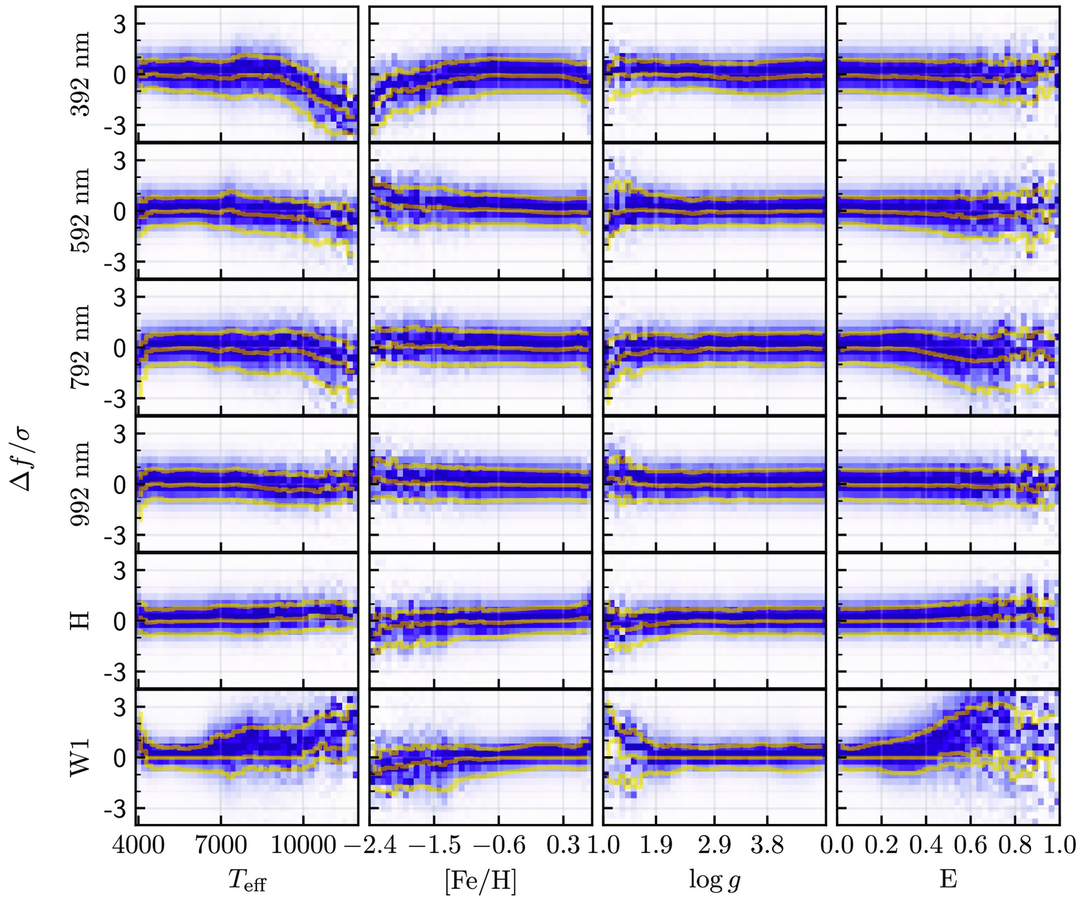
$\sigma_{\text{inference, E}} \simeq 0.8\sigma_{\text{Bayestar19, E}}$

$\sigma_{\text{inference, }\varpi} \simeq 0.5\sigma_{\text{GDR3, }\varpi}$

uncertainty	Teff.	Log g	[Fe/H].	E
	90k.	0.15dex.		0.03mag

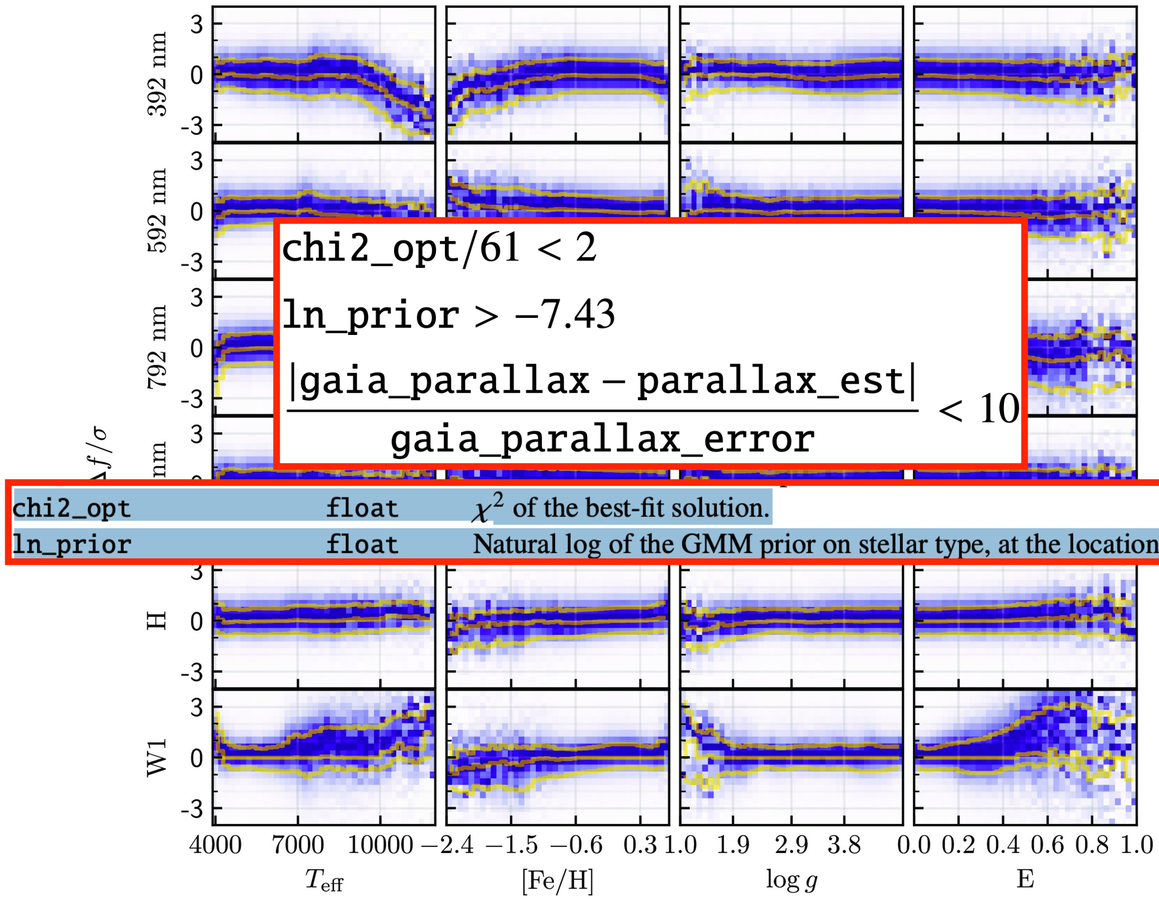
Result

$(\triangle f/\sigma \equiv (f_{\rm pred} - f_{\rm obs})/\sigma_f)$



Result

$$(\triangle f/\sigma \equiv (f_{\rm pred} - f_{\rm obs})/\sigma_f)$$

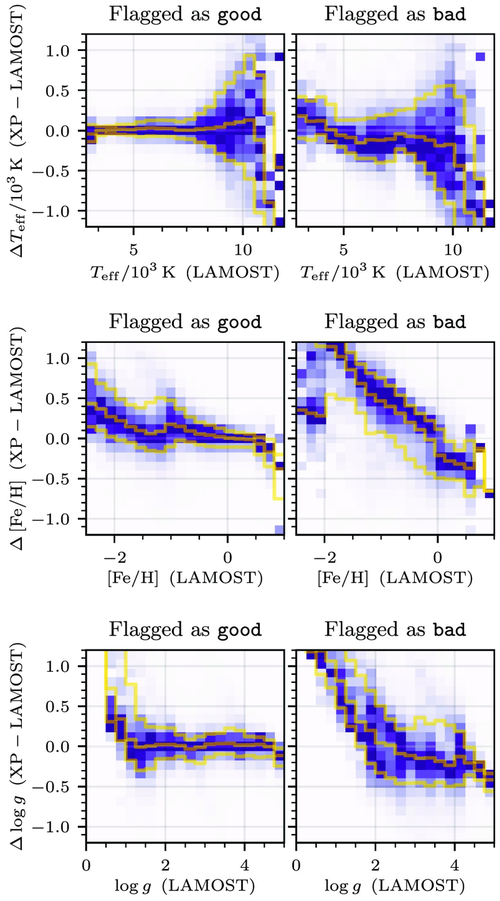


chi2_opt/61 < 2

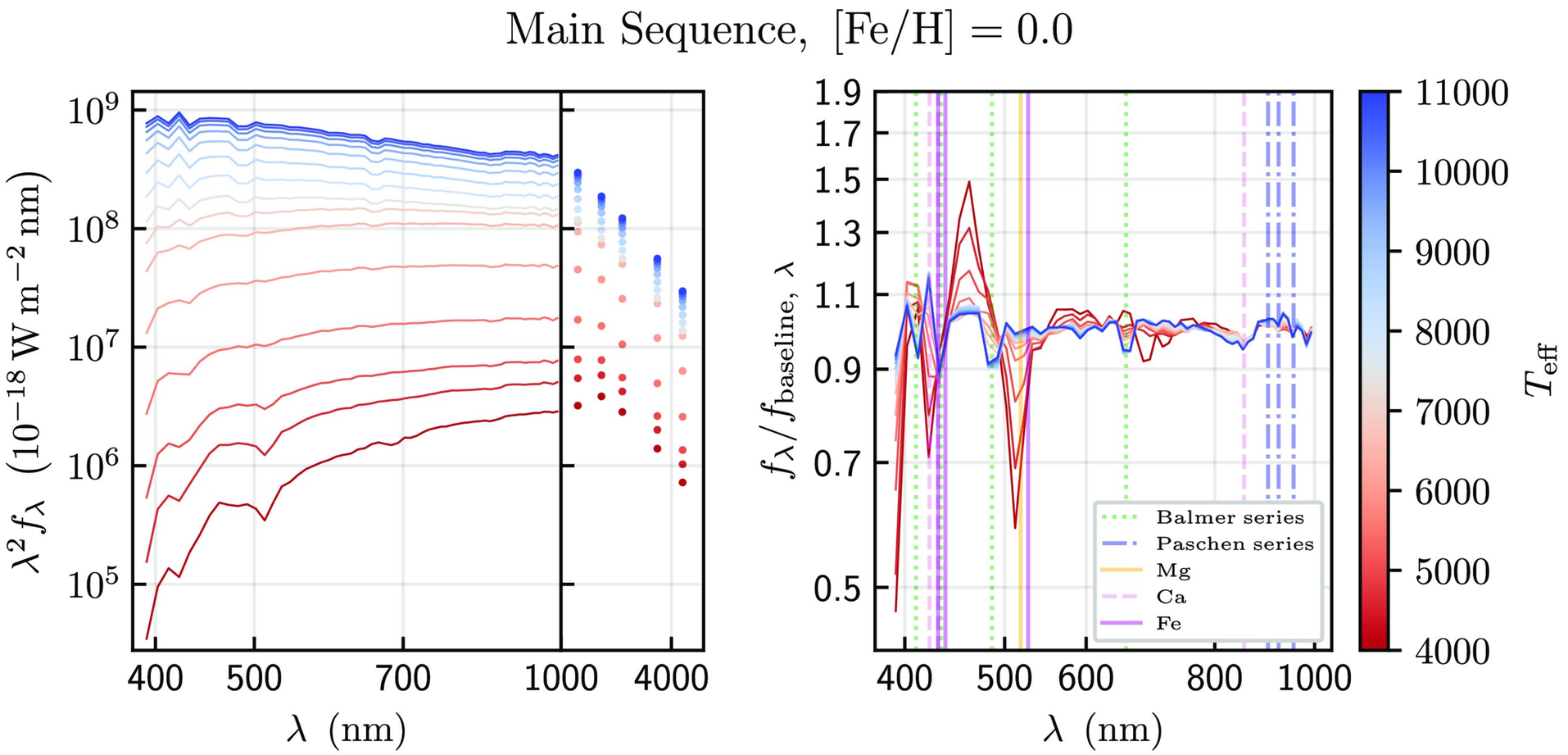
ln_prior > −7.43

$$\frac{|\texttt{gaia\_parallax} - \texttt{parallax\_est}|}{\texttt{gaia\_parallax\_error}} < 10$$

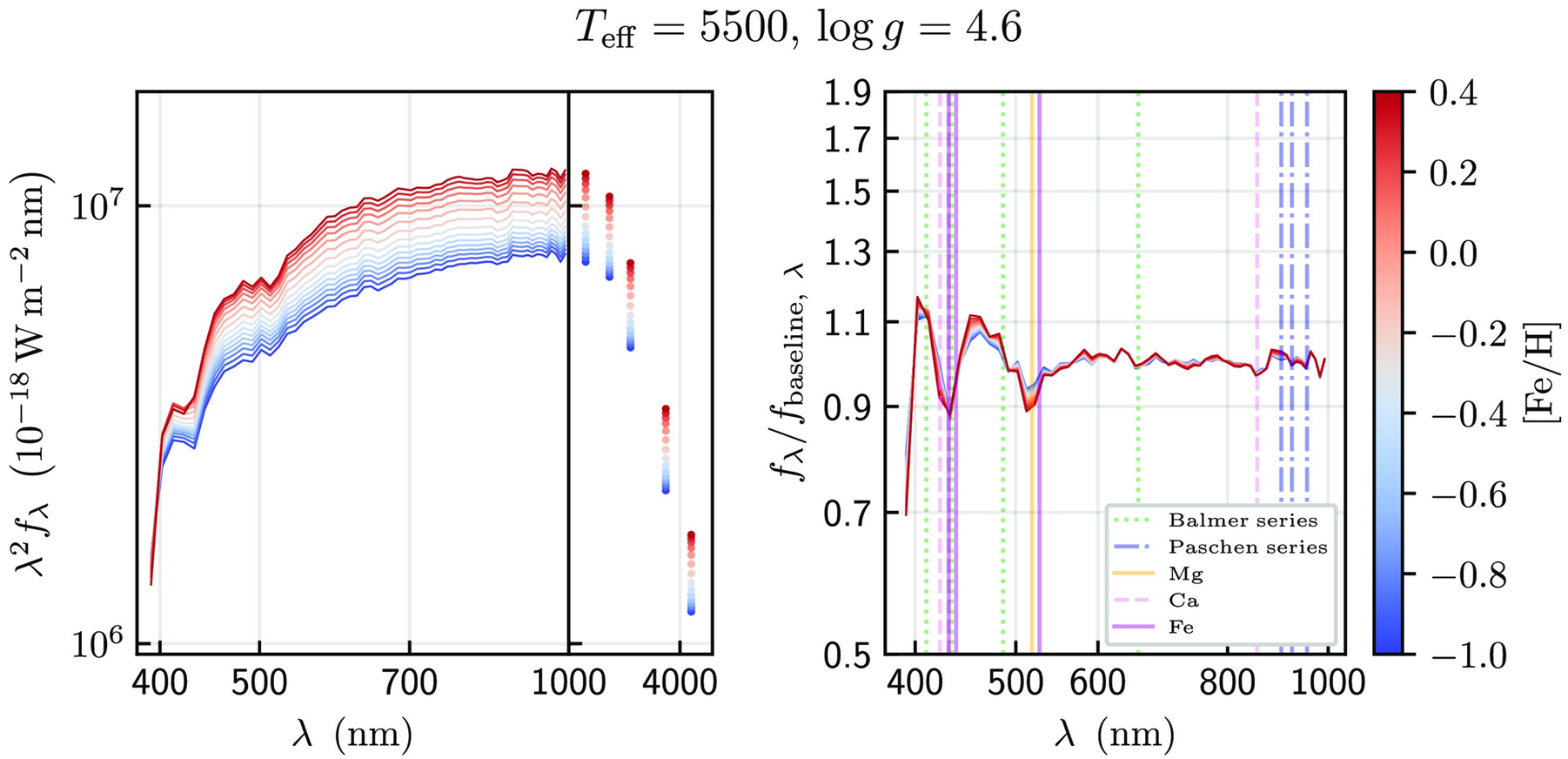
chi2_opt	float	χ^2 of the best-fit solution.
ln_prior	float	Natural log of the GMM prior on stellar type, at the location of the optimal solution.



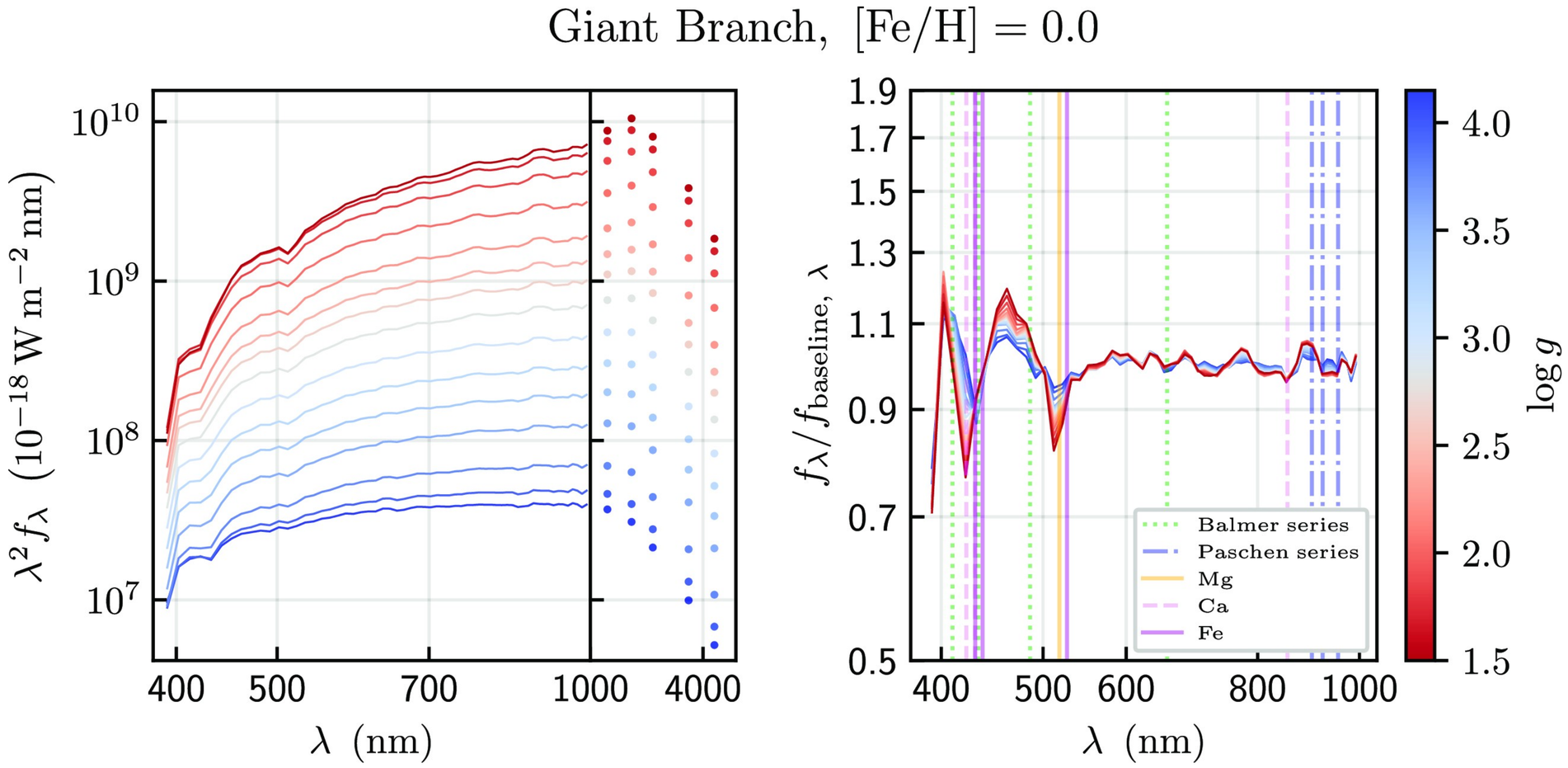
Result


Main Sequence, [Fe/H] = 0.0
$\lambda^2 f_\lambda$ (10^{-18} W m^{-2} nm)
10^9
10^8
10^7
10^6
10^5
400
500
700
1000
4000
λ (nm)
$f_\lambda / f_{\mathrm{baseline},\,\lambda}$
1.9
1.7
1.5
1.3
1.1
0.9
0.7
0.5
Balmer series
Paschen series
Mg
Ca
Fe
400
500
600
800
1000
λ (nm)
11000
10000
9000
8000
7000
6000
5000
4000
T_{eff}

Result

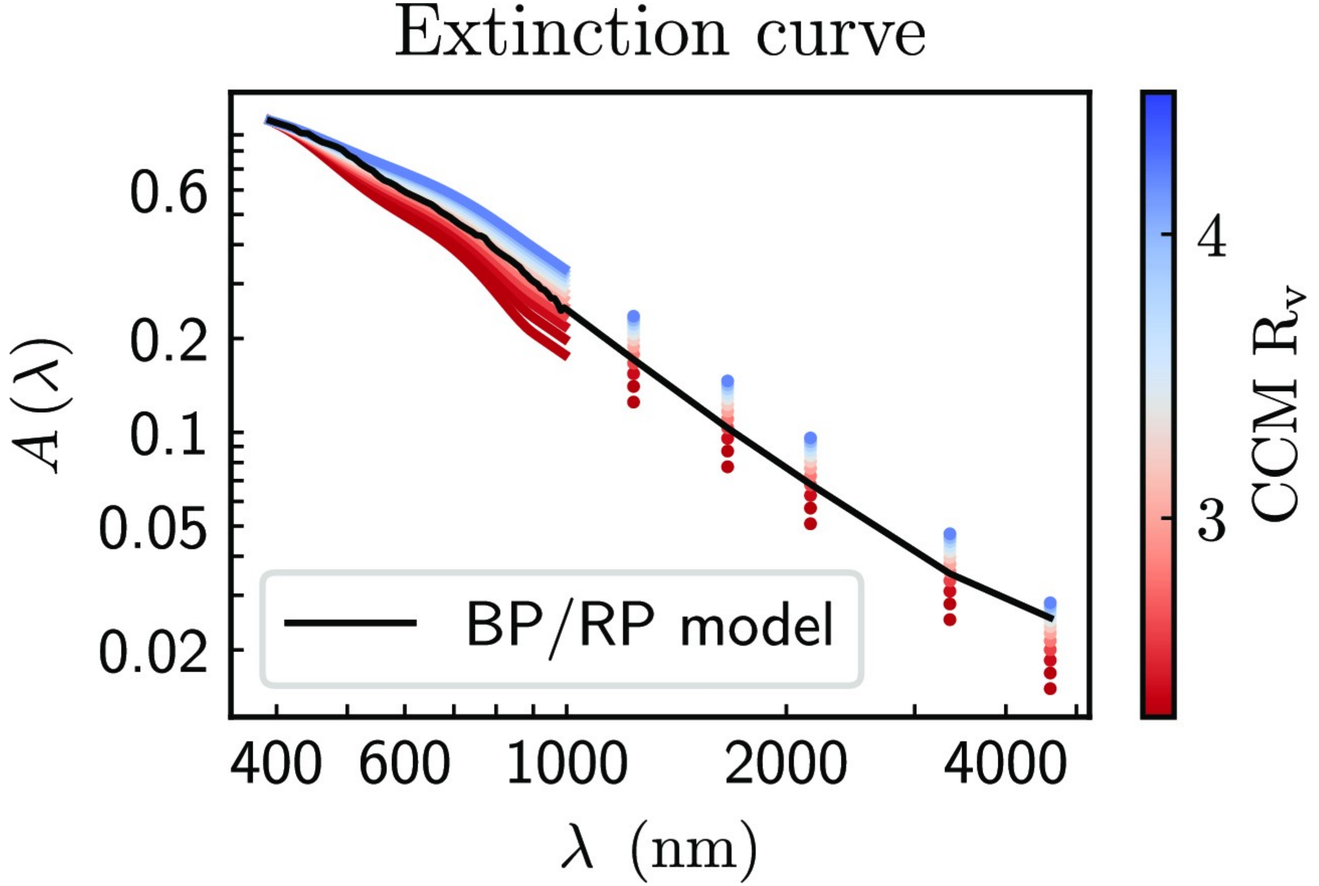

$T_{\rm eff} = 5500,\ \log g = 4.6$
$\lambda^2 f_\lambda$ $(10^{-18}\,\mathrm{W\,m^{-2}\,nm})$
10^7
10^6
400
500
700
1000
4000
λ (nm)
$f_\lambda / f_{\rm baseline,\ \lambda}$
1.9
1.7
1.5
1.3
1.1
0.9
0.7
0.5
400
500
600
800
1000
λ (nm)
Balmer series
Paschen series
Mg
Ca
Fe
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
[Fe/H]

Result


Giant Branch, [Fe/H] = 0.0
$\lambda^2 f_\lambda$ (10^{-18} W m^{-2} nm)
10^{10}
10^9
10^8
10^7
400
500
700
1000
4000
λ (nm)
$f_\lambda / f_{\mathrm{baseline},\,\lambda}$
1.9
1.7
1.5
1.3
1.1
0.9
0.7
0.5
400
500
600
800
1000
λ (nm)
Balmer series
Paschen series
Mg
Ca
Fe
4.0
3.5
3.0
2.5
2.0
1.5
log g

Result

Compared with the CCM model (colored curves; Cardelli et al. 1989)



roughly follows CCM models with $\sim 2.8 - 3.2$

CONCLUSIONS

The authors develop a data-driven model to estimate stellar parameters, distances and extinctions for 220 million stars with Gaia XP spectra.

The model uses a neural network to predict the absolute flux of a star as a function of its atmospheric parameters, and incorporates 2MASS and WISE photometry to reduce degeneracies.

The model is trained on a subset of stars with LAMOST spectra, and validated on a separate subset. The model also infers the extinction curve at the resolution of the XP spectra.

The model is applied to all stars with XP spectra, and the resulting catalog of stellar parameters and extinctions is provided. The catalog reveals the 3D distribution of stellar types and dust in the Milky Way.

$$\log g_{\rm MS} = \begin{cases} 4.6\,, & T_{\rm eff} < 5000 \\ 4.6 - 5.0\times10^{-4}(T_{\rm eff} - 5000)\,, & 5000 \le T_{\rm eff} < 6300 \\ 3.95\,, & T_{\rm eff} \ge 6300 \end{cases} \tag{28}$$

$$T_{\rm eff,GB} = \begin{cases} 5200 - 441.86(3.65 - \log g)\,, & \log g < 3.65 \\ 5900 - 1400(4.15 - \log g)\,, & \log g \ge 3.65 \end{cases}$$

- **使用物理模型**：这种方法是基于恒星的基本属性，如质量、年龄和元素丰度，来预测恒星的光谱。例如，Andrae et al. (2022)使用等时线模型、四种理论的光谱大气模型和平均消光定律来拟合观测到的XP光谱。这种方法的缺点是它对底层恒星模型的准确性非常敏感，而XP光谱的低分辨率使得从单个吸收线获取信息变得困难。
- **使用经验前向模型**：这种方法是利用一小部分有高分辨率光谱的恒星，来学习一个经验的前向模型，即XP光谱作为恒星参数的函数。然后，我们可以将这个模型应用到所有220百万个光谱，来推断恒星类型、距离和消光，使用标准的贝叶斯前向建模技术。这种方法的优点是它相对可解释，因为它使用了前向模型来预测数据应该是什么样子的，允许探索残差和发现新的系统误差和解释变量。此外，这种方法可以处理缺失数据（通过将相应的不确定性设置为无穷大），并且应该在低信噪比的情况下优雅地退化。在本文中，我们采用了经验前向建模的方法。
- **使用机器学习模型**：这种方法是训练一个机器学习模型，直接从XP光谱预测恒星参数（即监督学习）。这与第二种方法类似，因为它也利用了一小部分有高分辨率光谱的恒星（例如，Rix et al. 2022; Andrae et al. 2023a）。[1]主要的区别是，监督学习不使用前向模型，而是直接找到观测光谱中表明恒星参数的特征。这种直接的机器学习方法在低信噪比的情况下应该退化得更快，因为它没有充分利用可用的测量不确定性。然而，这种方法也可能学会正确地忽略光谱中与感兴趣的参数无关的特征（如系统误差和依赖于未建模的恒星参数的光谱特征）。相比之下，前向建模方法必须显式地对所有对观测光谱有显著影响的参数进行建模（或者至少引入误差项来考虑它们）。这种前向建模的缺点也可以被视为优点，因为它揭示了数据中的系统误差和未建模变量的特征。因此，前向建模和监督学习方法都有其优点。

前向模型和机器学习模型是两种不同的方法来从Gaia XP光谱中推断恒星参数。前向模型的思想是，根据恒星的基本属性（如温度、重力、金属丰度等），预测恒星的光谱应该是什么样子的，然后用观测到的光谱和预测的光谱进行比较，找到最符合的恒星参数。前向模型的优点是，它可以利用观测数据的不确定性，提供恒星参数的误差估计，而且可以检测数据中的系统误差和未建模的变量。前向模型的缺点是，它需要一个准确的物理模型来预测恒星光谱，而这个模型可能存在一些假设和局限性。

机器学习模型的思想是，利用一小部分已经有高分辨率光谱和恒星参数的样本，训练一个算法来直接从Gaia XP光谱中提取恒星参数。机器学习模型的优点是，它可以自动学习光谱中的特征，而不需要依赖物理模型。机器学习模型的缺点是，它在低信噪比的情况下可能退化得更快，因为它没有充分利用观测数据的不确定性，而且它可能难以解释和验证。

Fig1 LAMOST的标准参数估计方法在8000 K ≤ Teff ≤ 12000 K的范围内有困难，是因为这些恒星的巴尔末线（Balmer lines）很难用理论模型来拟合。巴尔末线是氢原子的光谱线，反映了恒星的温度和表面重力。然而，这些线的形状和强度也受到其他因素的影响，例如恒星的金属丰度，微湍流速度，非局部热平衡效应等

FWHM full width of half maximum

帕申系
物理学上氢原子的电子从主量子数n大于或等于4跃迁至n= 3的一系列光谱线
巴尔莫
来自氢原子所发射的光谱线在可见光有4个波长：410纳米、434纳米、486纳米和656纳米。它们是吸收光子能量的电子进入受激态后，返回主量子数n等于2的量子状态时释放出的谱线。