Gaia XP spectra
&
data driven analysis

MW-Cost Spectroscopic School

René Andrae

23. Sept 2021

Contents

1) Gaia XP spectra and their formats (30min)

2) usage by DPAC/CU8 (20min)

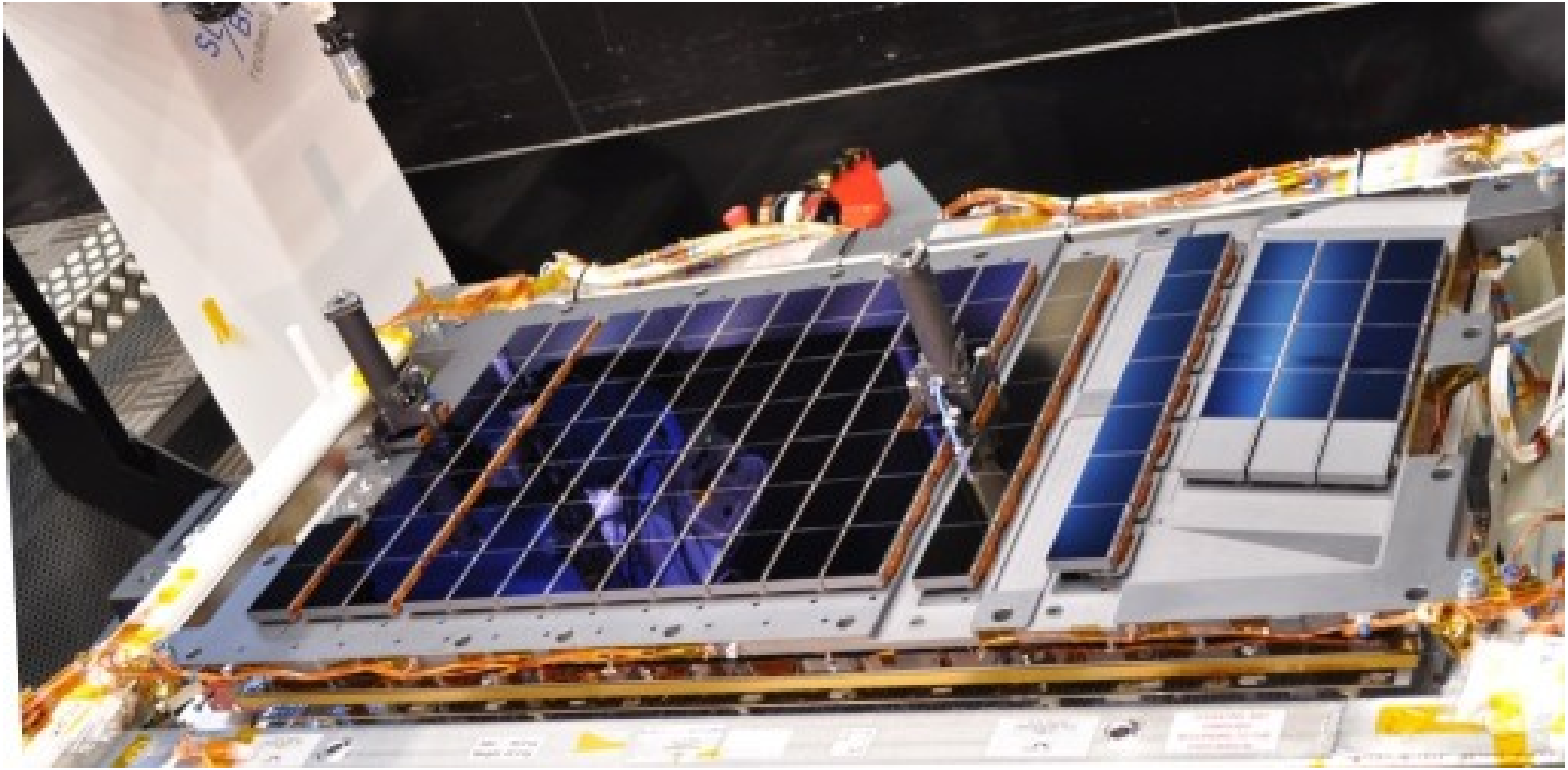
3) constructing empirical training samples (45min)

Gaia XP spectra

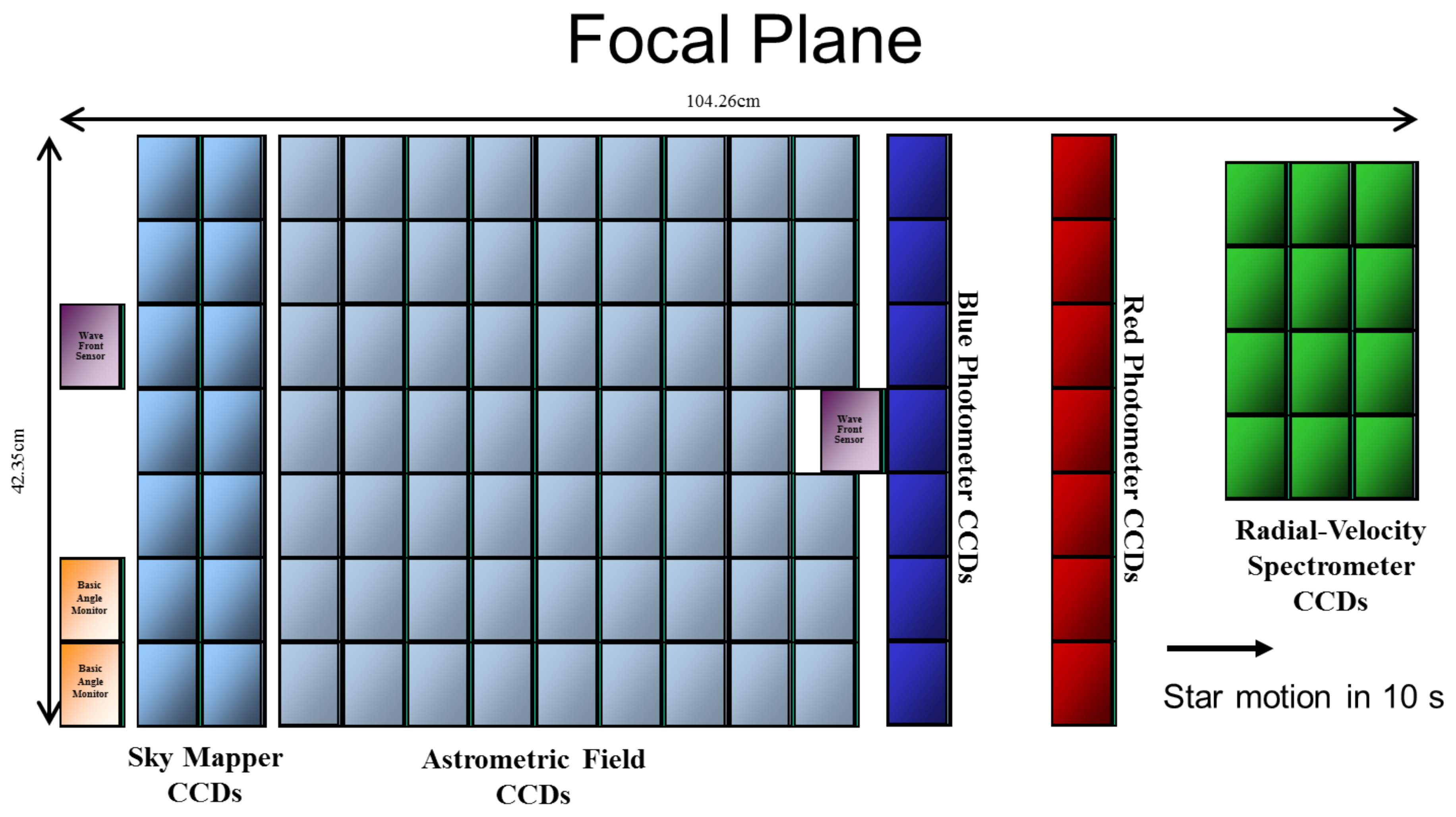
Topics covered:

- XP instrument
- XP mean spectra, continuous basis functions
- sampling XP spectra (GaiaXPy)
- some examples: stars, QSOs, emission lines, etc.

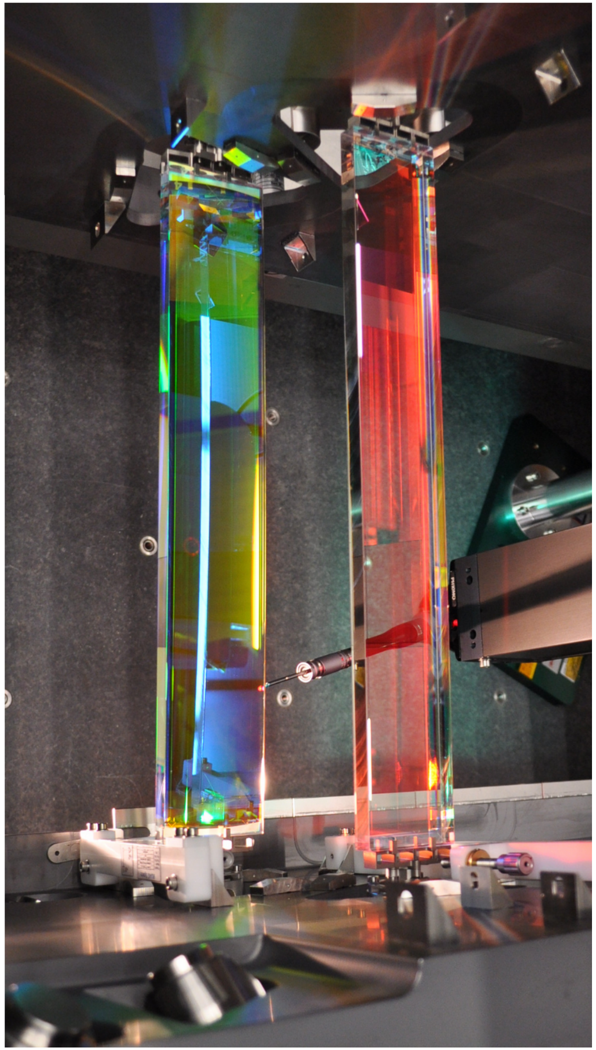
Gaia XP spectra



Gaia XP spectra



Gaia XP spectra



Single epoch observations:

- prism spectra
- 2D CDD image
  (G>11.5 collapsed to 1D onboard)
- each CCD has 60 pixels
- optical: 320-1050nm
- low resolution: R~10-50
- taken for every source seen by Gaia

Gaia XP spectra

- GDR3: ~40 epochs (ultimately ~70 epochs)
- very high signal-to-noise ratio
- will cross different CCD rows
- instrument evolves with time!
- no two epochs have identical wavelength sampling!
- bad news: cannot “just stack” epoch spectra
- “mean spectrum” = continuous function fitted to epochs
- good news: higher effective resolution

Gaia XP spectra

- “mean spectrum” as basis functions (Gauss-Hermite)

$$s(p)=\sum_{i=1}^{55} c_i b_i(p)$$

Carrasco et al. (2021)

$$\vec{s}=D\cdot\vec{c}$$

XP spectrum

design matrix

XP coefficients

- least-squares fit to epoch spectra: c + Cov(c)
- c + Cov(c) = published in GDR3 (few hundred million)
- but there also is GaiaXPy ...

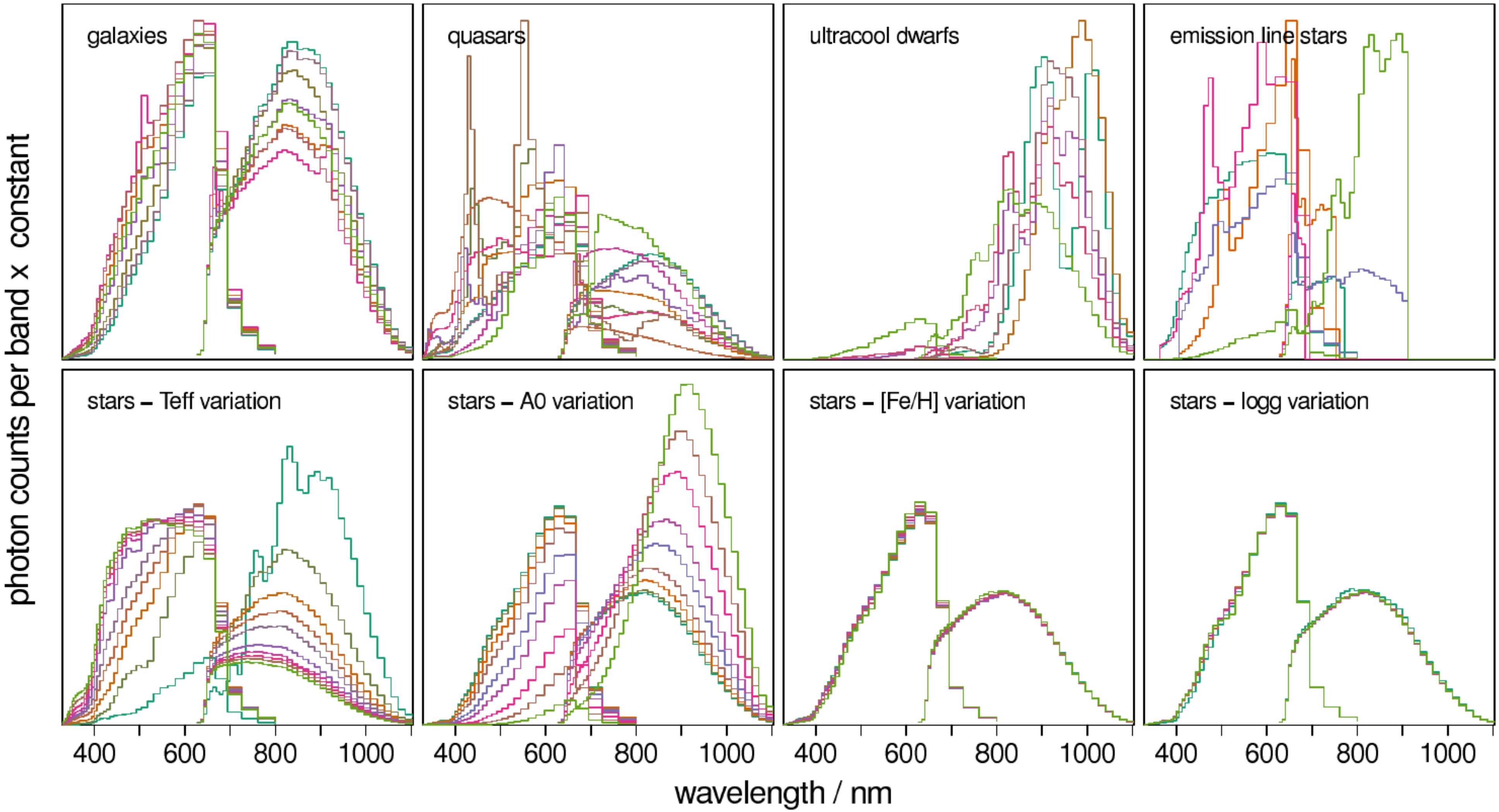
Gaia XP spectra

GaiaXPy:

- Python package provided by DPAC/CU5
- compute “sampled XP spectra” from coefficients (i.e. flux vs pixel from given wavelengths)
- can compute integrated photometry in bands (e.g. UBVRI, SDSS ugriz, etc)
- can simulate XP model spectra from SEDs

Gaia XP spectra

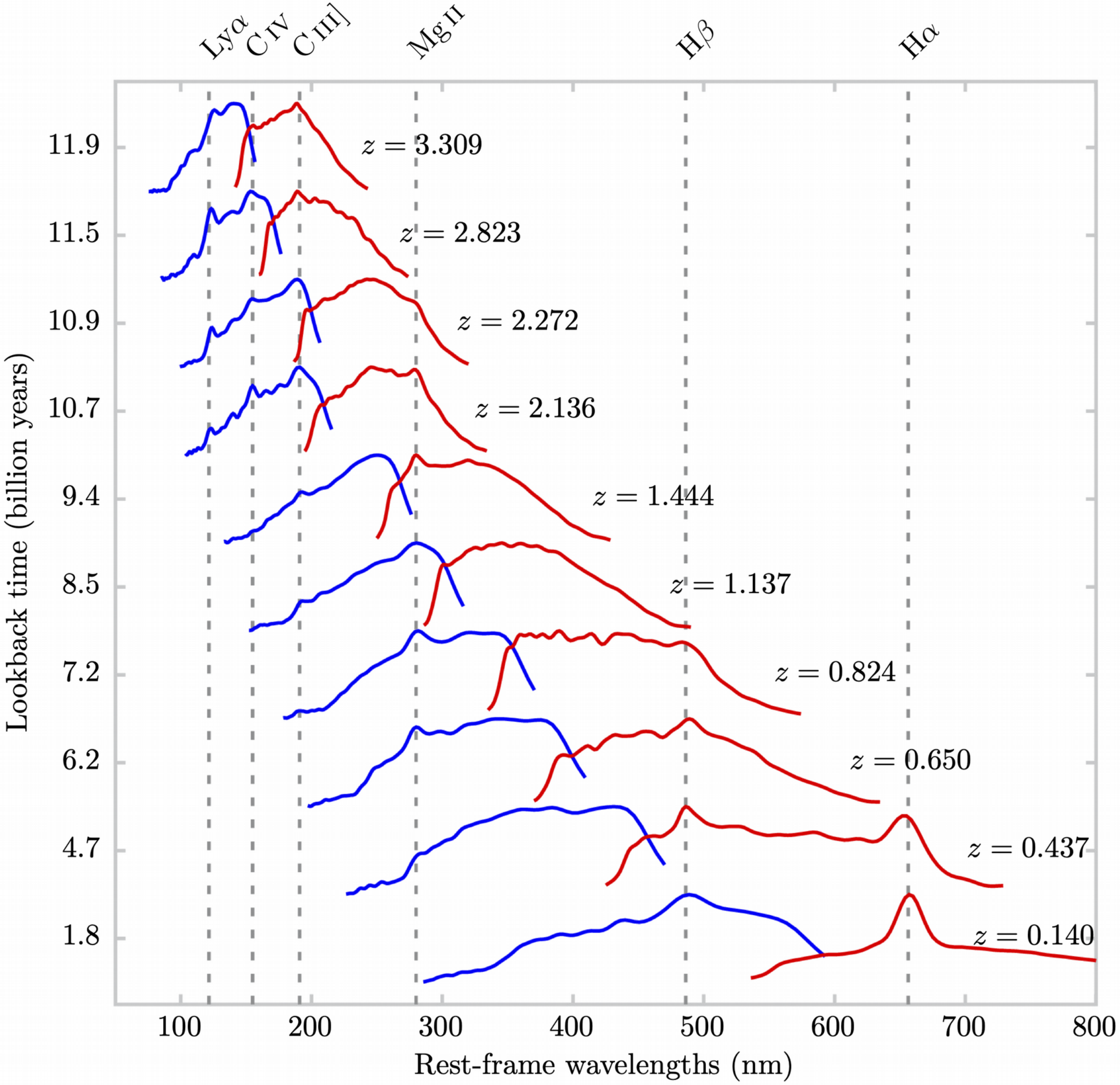
Bailer-Jones et al (2013)



Gaia XP spectra

emission lines in real QSOs

https://www.cosmos.esa.int/web/gaia/iow_20201222



Gaia XP spectra

Questions?

Contents

1) Gaia XP spectra and their formats

2) usage by DPAC/CU8

3) constructing empirical training samples

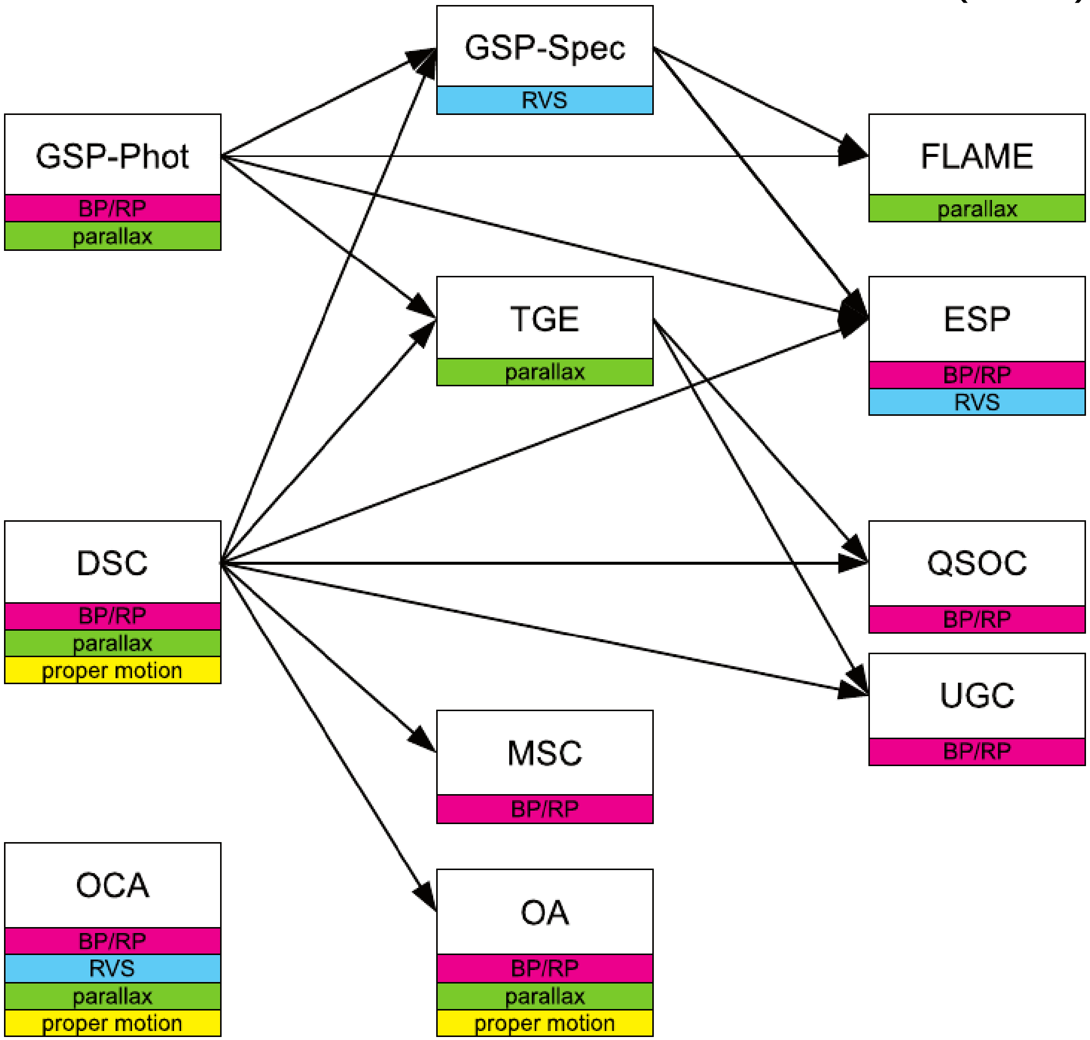
DPAC/CU8 usage of XP spectra

Topics covered:

- What is CU8?
- describe GSP-Phot: highlight forward modelling
- describe DSC: highlight data-driven approach

DPAC/CU8 usage of XP spectra

Bailer-Jones et al (2013)

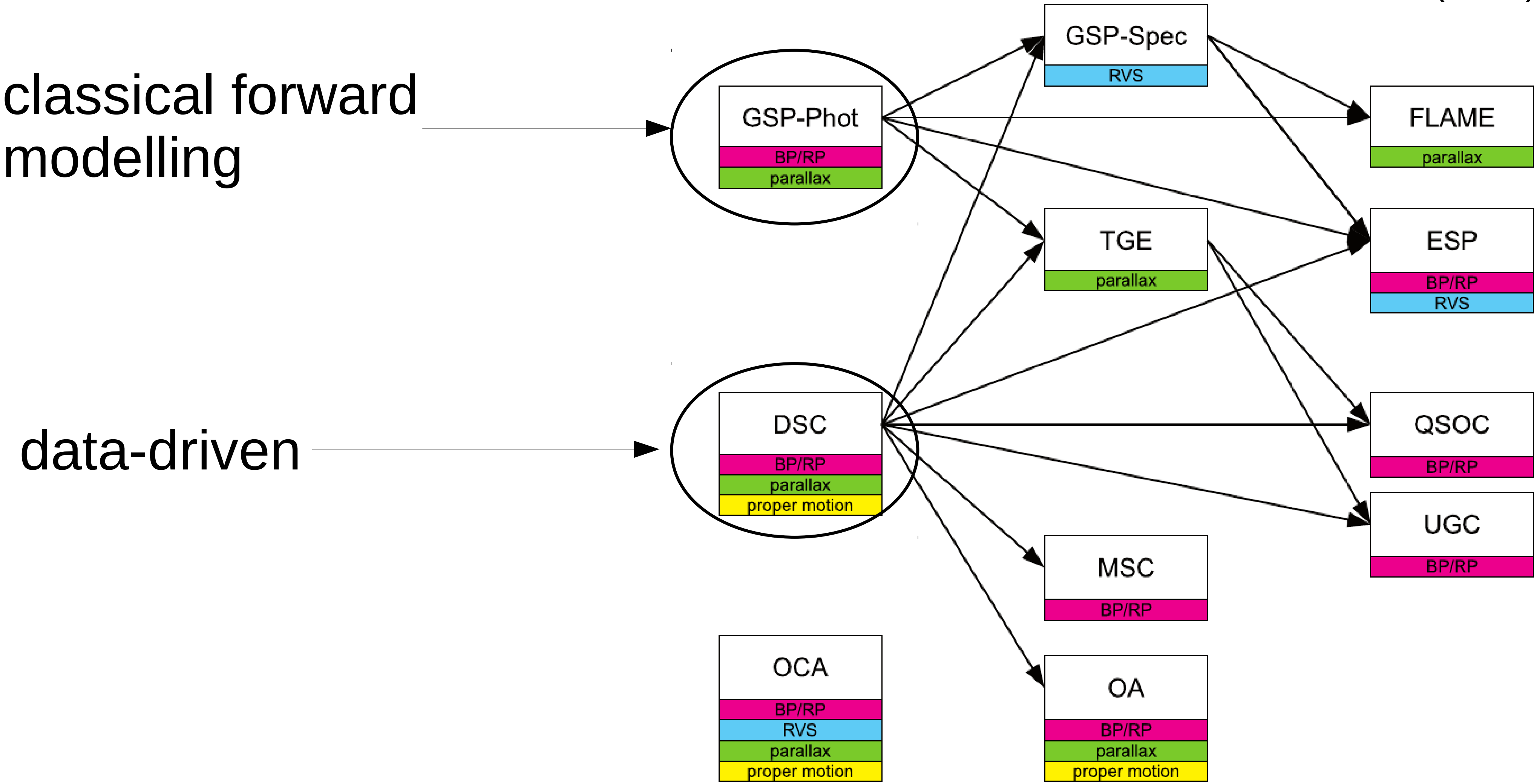


What is CU8:

- part of Gaia DPAC
- characterise sources in terms of astrophysics
- many modules/groups
- results published in GDR3
- "internal use of XP spectra" = scientific validation

DPAC/CU8 usage of XP spectra

Bailer-Jones et al (2013)



DPAC/CU8 usage of XP spectra

<u>Why some CU8 modules use classical forward modelling:</u>

- sometimes good physical models are available (e.g. MARCS, PHOENIX model SEDs for stars)
- can simulate XP spectra from model SEDs
- forward modelling requires good understanding of instrument: “if DPAC/CU8 doesn’t do it, who else?”
- forward modelling allows us to exploit astrophysical knowledge

DPAC/CU8 usage of XP spectra

Exploiting astrophysical knowledge with forward modelling:

- GSP-Phot: stellar parameters for all sources with G<19
- step 1: isochrones provide self-consistent Teff, logg, radius R, absolute M_G magnitude, [M/H]
- step 2: XP spectrum with extinction and amplitude

$$a = (R/d)^2$$

- step 3: predict apparent G magnitude

$$G = M_G + A_G + 5\log_{10} d - 5$$

DPAC/CU8 usage of XP spectra

<u>Downsides of forward modelling:</u>

- instrument model not known perfectly:
  - model XP spectra systematically wrong at some level
  - same applies to derived parameters
- computationally very expensive:
  - complex noise model: data, models, covariances, etc.
  - interpolation of model spectra
  - MCMC sampling
  - hundreds of millions of sources to process

DPAC/CU8 usage of XP spectra

<u>Data-driven approach in CU8:</u>

- DSC: classify all sources into star/QSO/galaxy/binary/WD
- problem: model SEDs for galaxies/binaries exist … but useless since no point-sources for Gaia
- train empirically: real XP spectra + known class type
- Gaussian mixture model + extremely randomised trees

DPAC/CU8 usage of XP spectra

<u>Data-driven approach in CU8:</u>

- DSC: classify all sources into star/QSO/galaxy/binary/WD
- problem: model SEDs for galaxies/binaries exist … but useless since no point-sources for Gaia
- train empirically: real XP spectra + known class type
- Gaussian mixture model + extremely randomised trees
- How do you construct empirical training samples?

DPAC/CU8 usage of XP spectra

Questions?

Contents

1) Gaia XP spectra and their formats

2) usage by DPAC/CU8

3) constructing empirical training samples

Constructing empirical training samples

<u>Topics covered:</u>

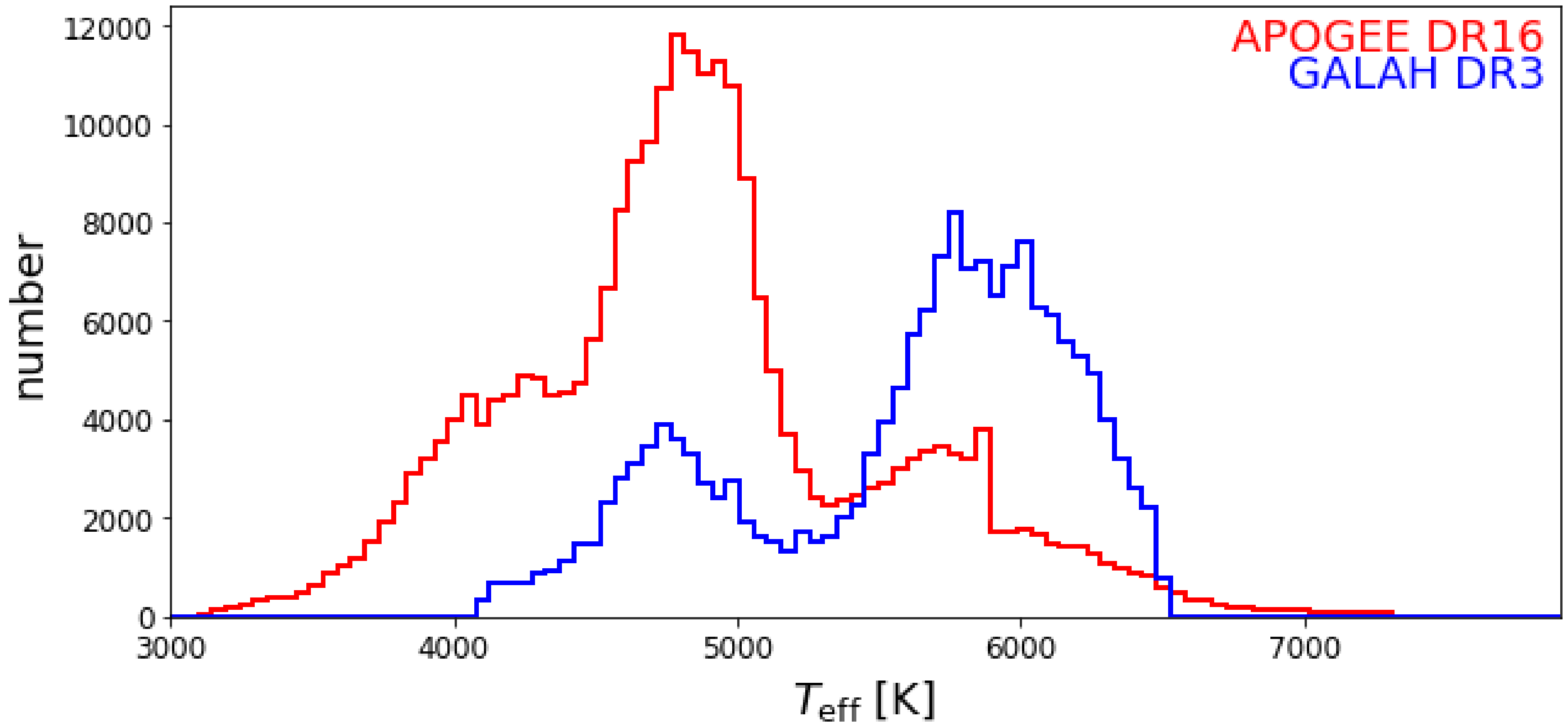
- train/test samples representative of application sample
- DSC prior handling
- XP spectra strongly affected by extinction … not enough literature values

Constructing empirical training samples

<u>Things to watch out for:</u>

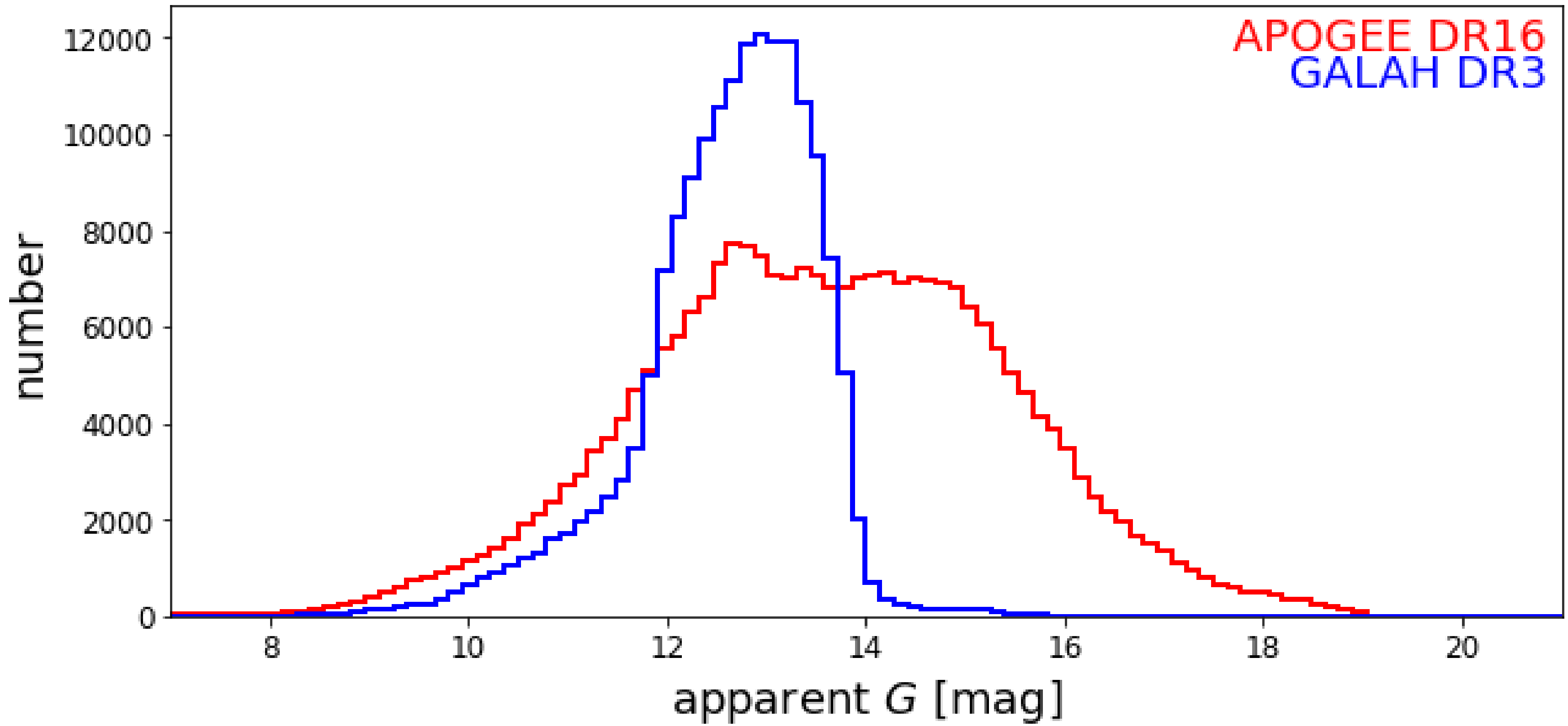
- classification: use class labels from literature
- regression: use stellar APs from literature
- inherit all systematic errors from literature (e.g. Teff offsets)
- distribution of literature targets usually not representative (e.g. SDSS definition of “QSO”)
- cross-match errors

Constructing empirical training samples



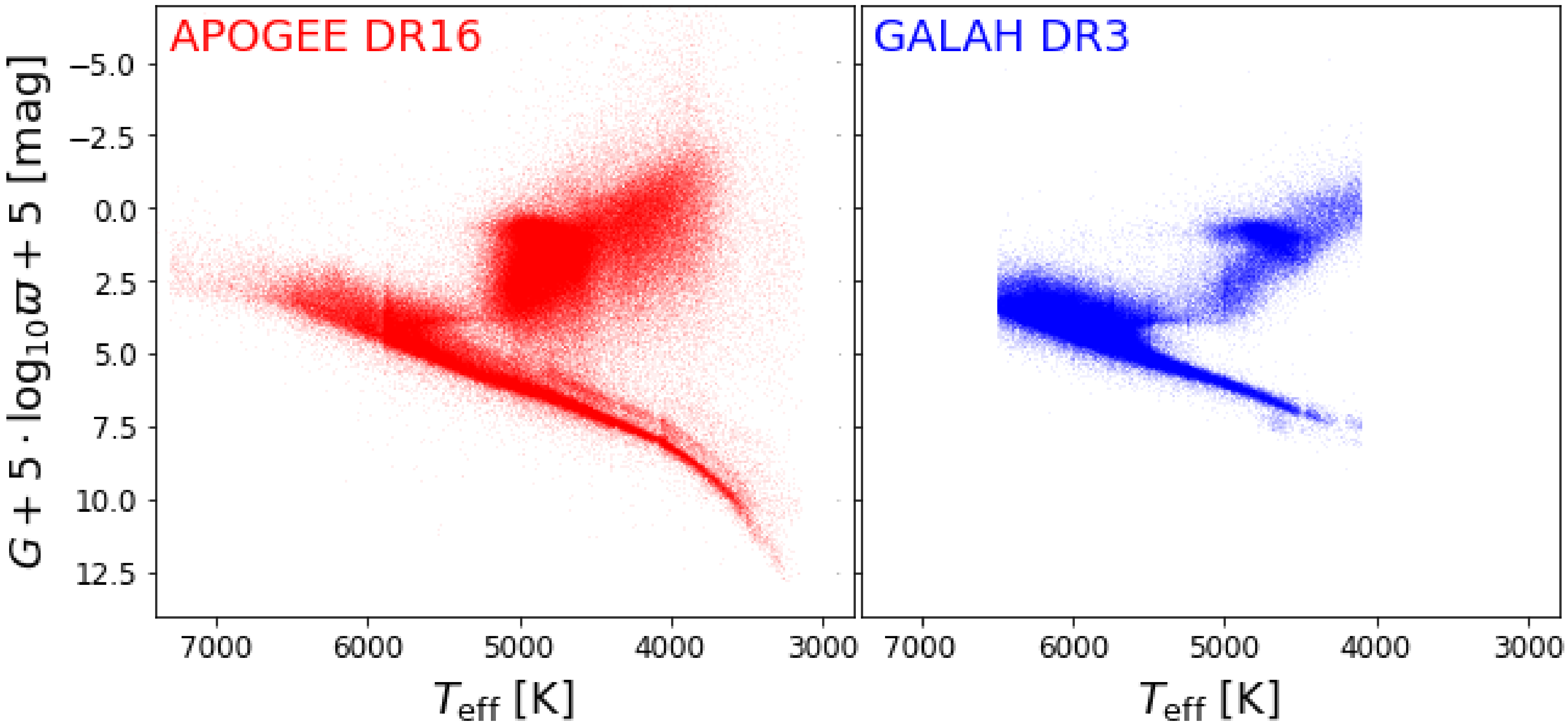
- Teff distributions of APOGEE DR16 and GALAH DR3 are very different

Constructing empirical training samples



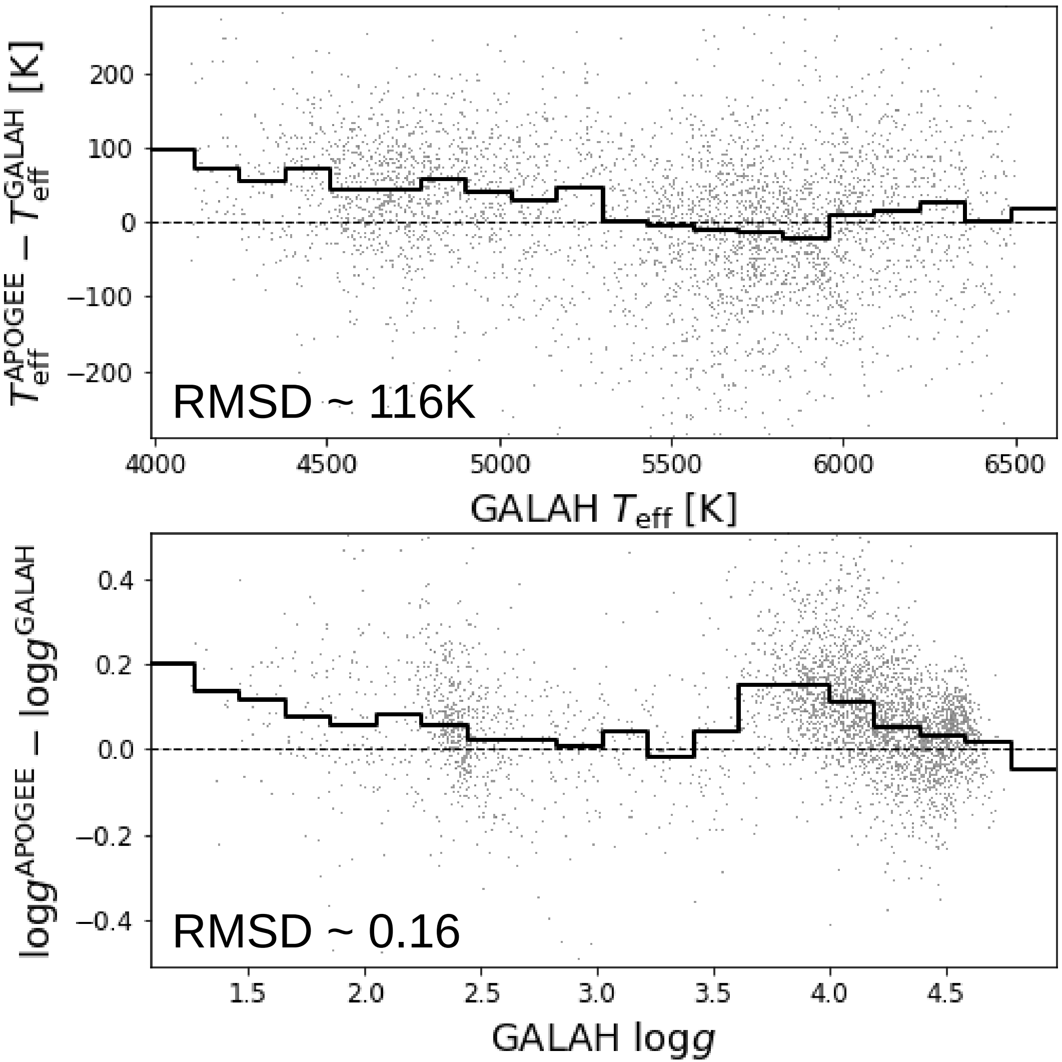
- APOGEE DR16 much fainter than GALAH DR3 (e.g. higher extinctions, lower SNR in XP spectra)

Constructing empirical training samples



- APOGEE DR16 and GALAH DR3 probe different populations

Constructing empirical training samples



- 4000 stars in both APOGEE DR16 and GALAH DR3
- Teff and logg show random differences (fundamental limit of performance test)
- also show systematic differences

Constructing empirical training samples

- “good model” trained on GALAH DR3 will perform very poorly when applied to APOGEE DR16 (and vice versa)
- need following concepts:
  - training sample
  - test sample
  - application sample
- random subsets of literature values for training and testing (cross-validation, bootstrapping)

If application sample has different distributions,
good performance on test sample becomes meaningless!

Constructing empirical training samples

Simple example from QSO classification:

Bailer-Jones et al. (2019)

- classifier trained on balanced training set (equal fractions of QSO, galaxy, star)

	STAR	QSO	GALAXY
true STAR	0.9982	0.0016	0.0002
true QSO	0.4170	0.5815	0.0015
true GALAXY	0.2603	0.0123	0.7274

Constructing empirical training samples

<u>Simple example from QSO classification:</u> Bailer-Jones et al. (2019)

- classifier trained on balanced training set (equal fractions of QSO, galaxy, star)

	STAR	QSO	GALAXY
true STAR	0.9982	0.0016	0.0002
true QSO	0.4170	0.5815	0.0015
true GALAXY	0.2603	0.0123	0.7274

- 58.15% of QSOs correctly identified, 0.16% of stars contaminate into QSO class

- BUT: stars are ~500x more common than QSOs

Constructing empirical training samples

<u>Simple example from QSO classification:</u> Bailer-Jones et al. (2019)

	STAR	QSO	GALAXY
true STAR	0.9982	0.0016	0.0002
true QSO	0.4170	0.5815	0.0015
true GALAXY	0.2603	0.0123	0.7274

$$\frac{500\times0.0016}{1\times0.5815+500\times0.0016}\approx0.579$$

- ~57.9% of sources classified as “QSO” are actually stars
- because distribution of QSOs in application sample (1 in 500) is different than in training sample (1 in 3)

Constructing empirical training samples

- correction of different distributions in training and application samples
- classification (Bailer-Jones et al. 2019):

1/500 for QSO

$$p_{app}(class|data) \propto \frac{p_{app}(class)}{p_{train}(class)} p_{train}(class|data)$$

1/3 for QSO

Constructing empirical training samples

- correction of different distributions in training and application samples

- classification (Bailer-Jones et al. 2019):

1/500 for QSO

$$p_{app}(class|data) \propto \frac{p_{app}(class)}{p_{train}(class)}\, p_{train}(class|data)$$

1/3 for QSO

- regression:

expected distribution

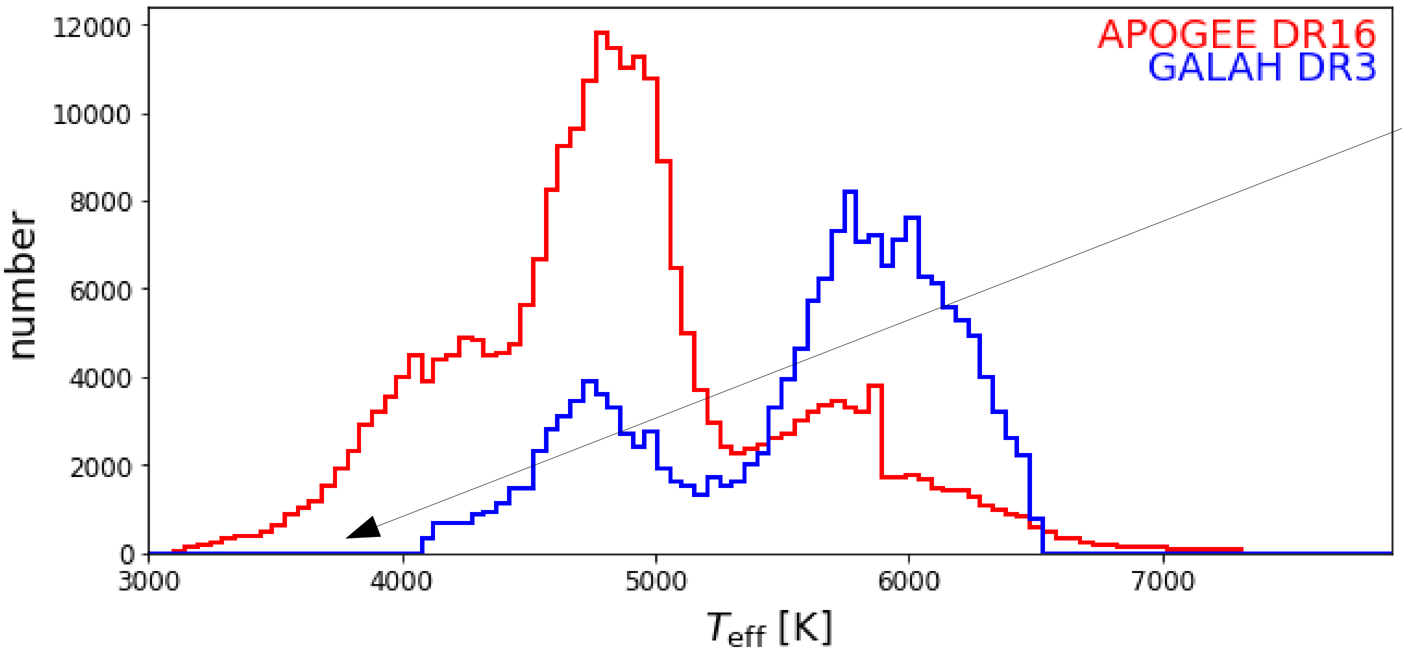
$$p_{app}(Teff|data) \propto \frac{p_{app}(Teff)}{p_{train}(Teff)}\, p_{train}(Teff|data)$$

training distribution

Constructing empirical training samples

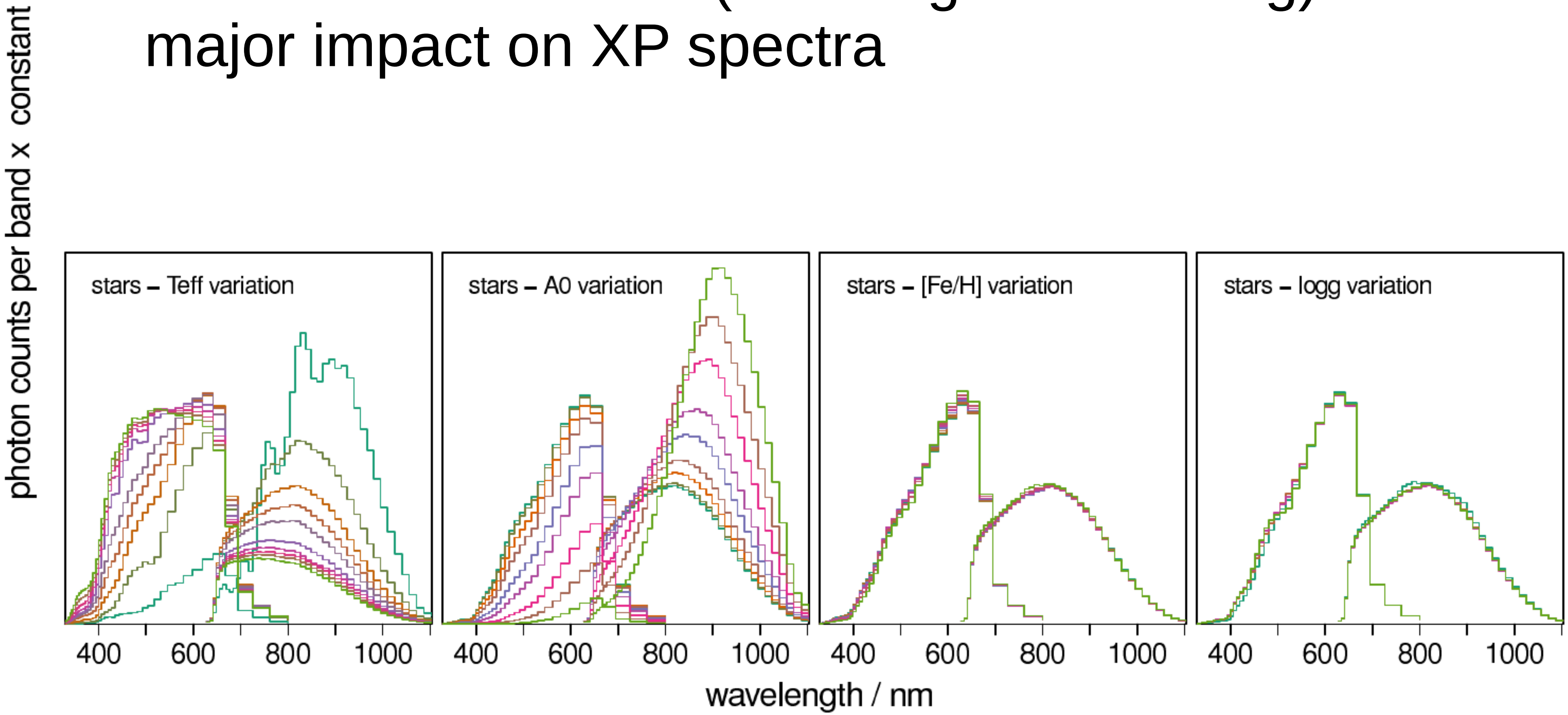
- problem 1: prior often too weak to change regression result
- problem 2: $$p_{app}(Teff|data) \propto \frac{p_{app}(Teff)}{p_{train}(Teff)} p_{train}(Teff|data)$$

$$p_{train}(Teff)=0$$



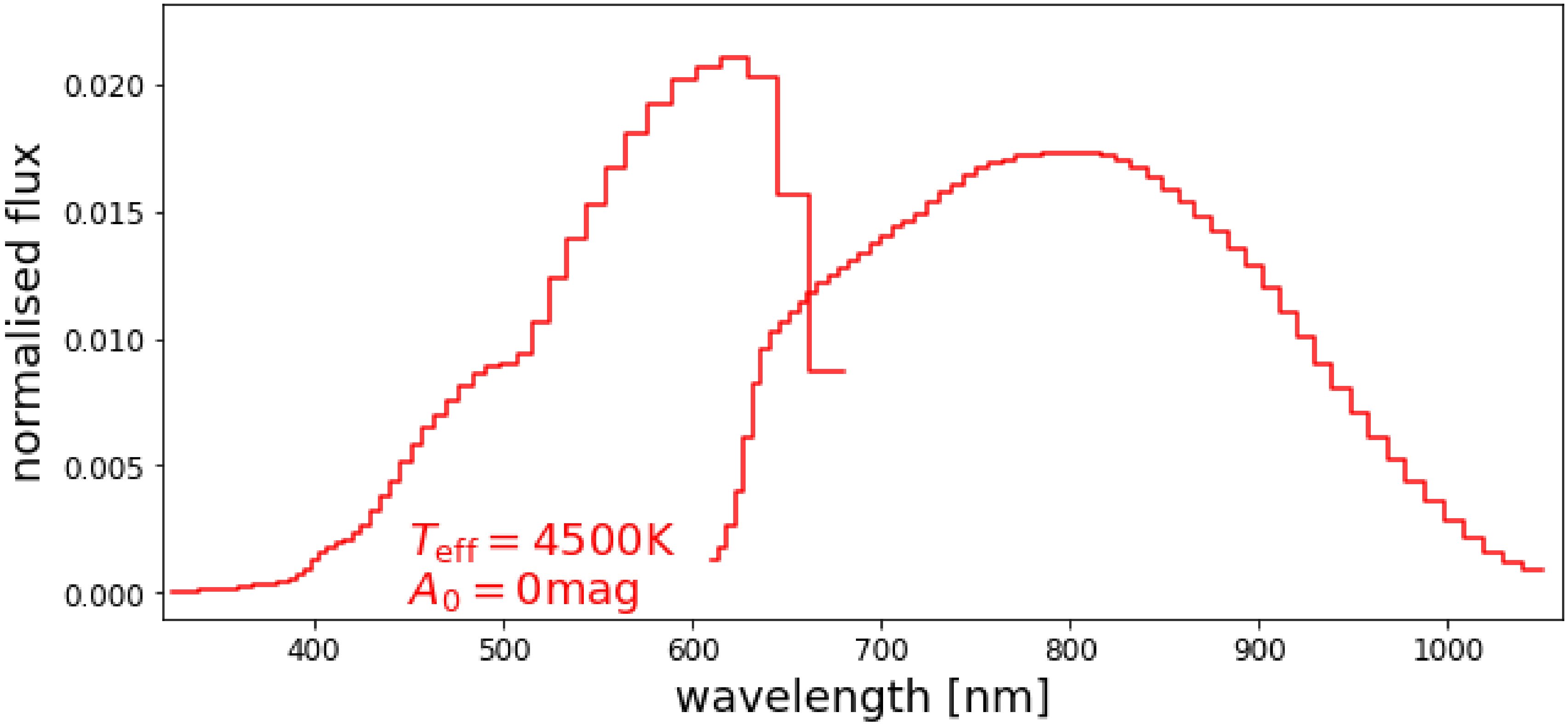
Constructing empirical training samples

- interstellar extinction (dimming + reddening) has major impact on XP spectra



Constructing empirical training samples

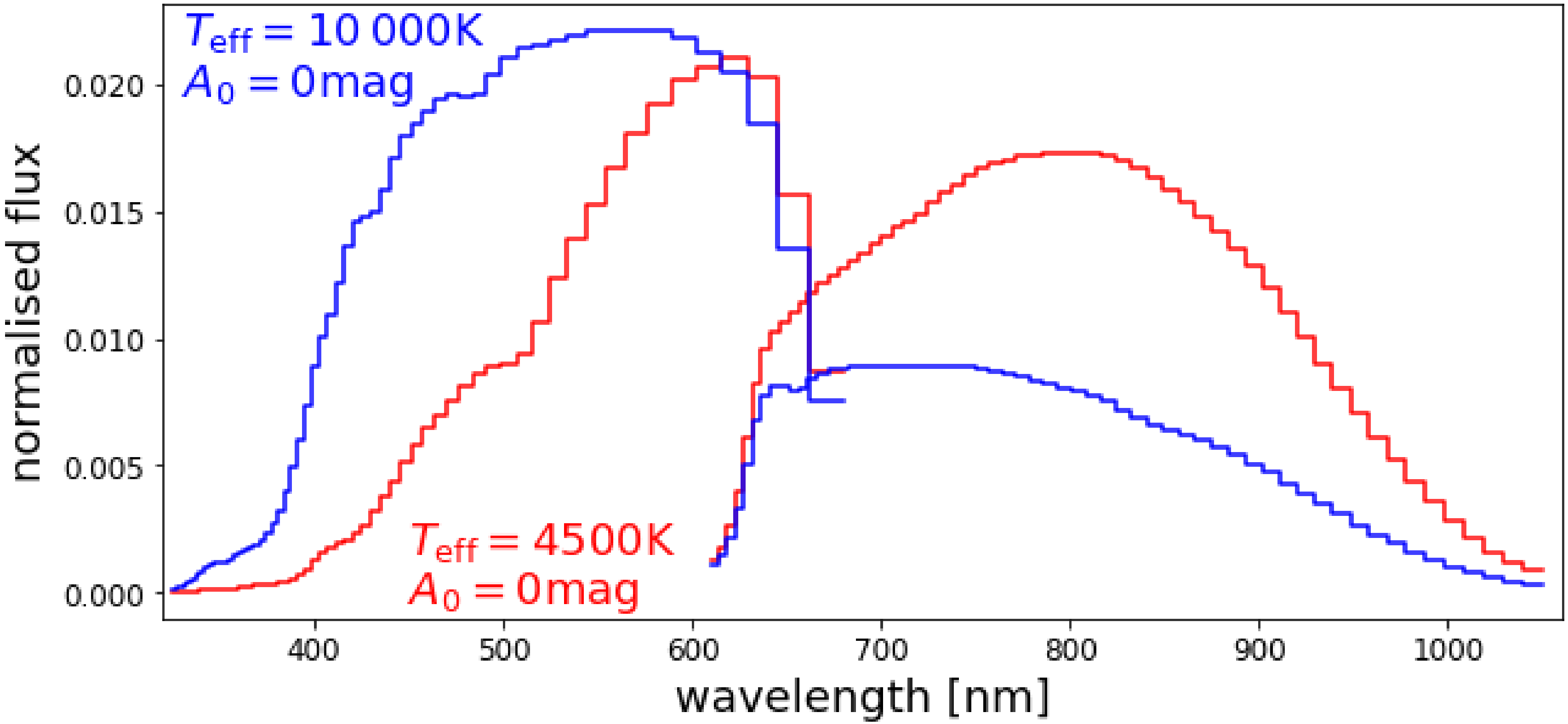
- extinction and temperature are highly degenerate



(PHOENIX models, logg = 4, [M/H] = 0)

Constructing empirical training samples

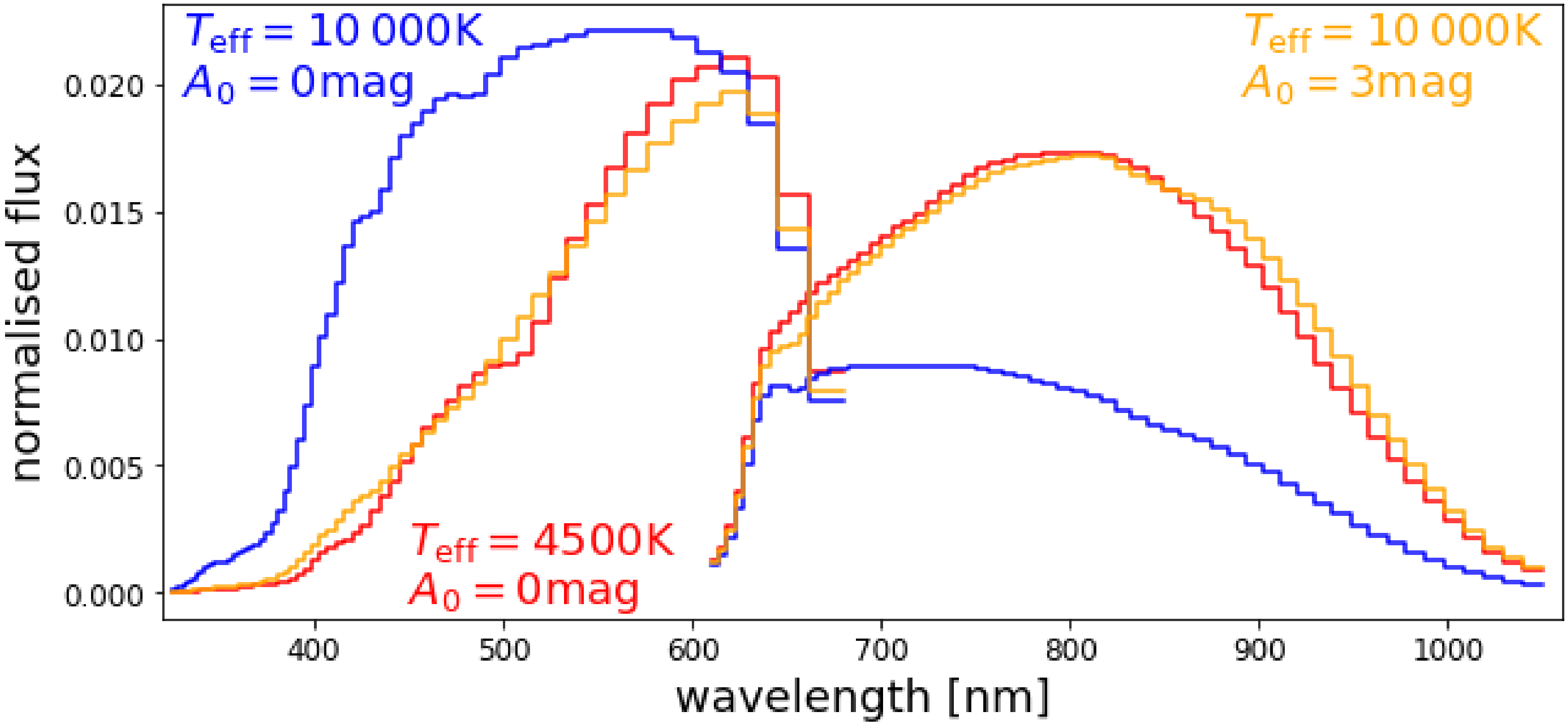
- extinction and temperature are highly degenerate



(PHOENIX models, logg = 4, [M/H] = 0)

Constructing empirical training samples

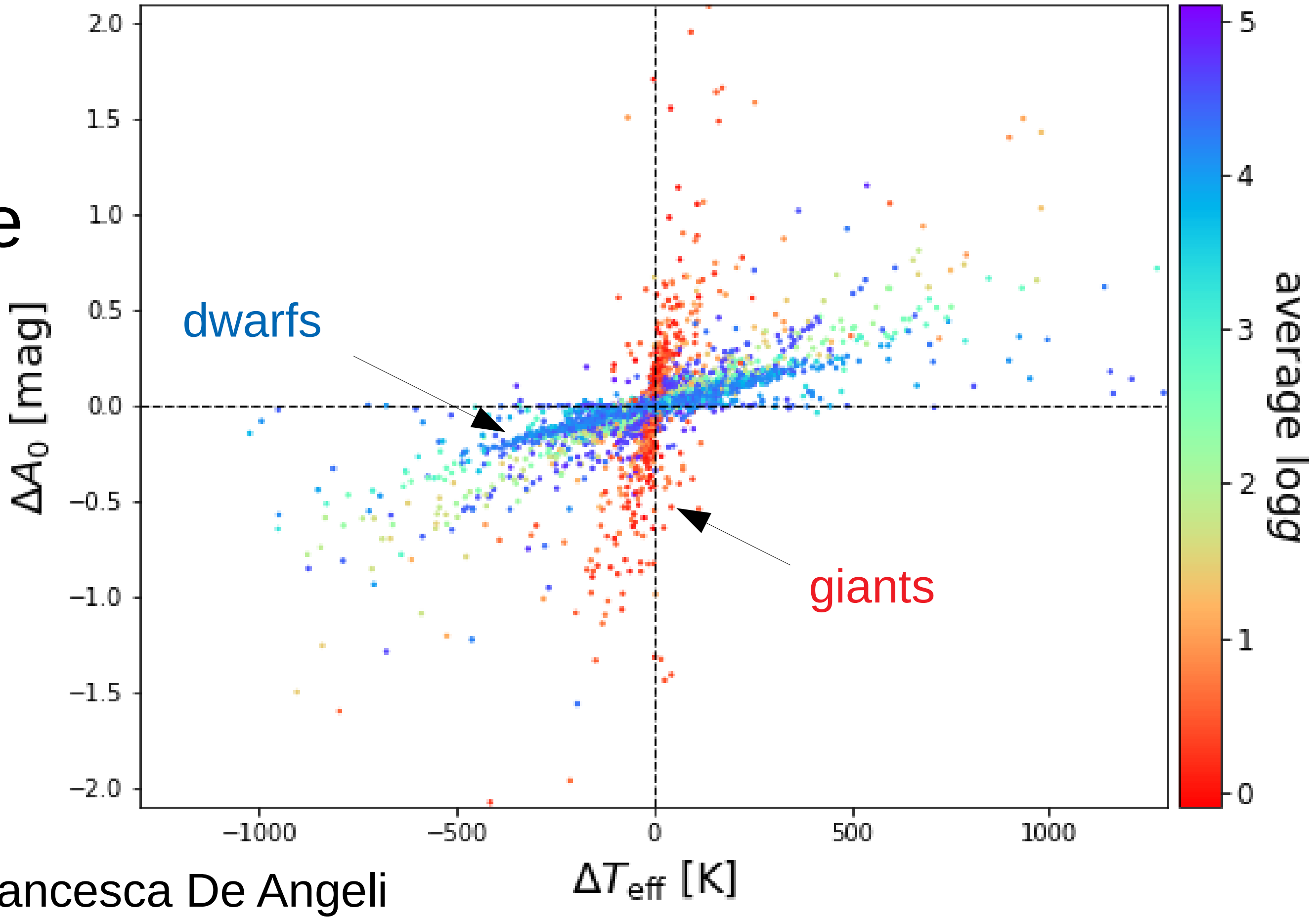
- extinction and temperature are highly degenerate



(PHOENIX models, logg = 4, [M/H] = 0)

Constructing empirical training samples

- extinction and temperature are highly degenerate
- differences in Teff and A0 estimates from sources whose epoch spectra were randomly split to form two independent mean XP spectra



courtesy: Francesca De Angeli

Constructing empirical training samples

- extinction has major impact on XP spectra
- limited supply of extinction labels for empirical training:
  BayesStar
  StarHorse

Constructing empirical training samples

- extinction has major impact on XP spectra
- limited supply of extinction labels for empirical training:
  BayesStar
  StarHorse
- beware of different definitions of “extinction”:
  - A(V) = A_0 = monochromatic at 547.7nm
    (parameter in extinction law)
  - A_V = extinction in Johnson V band
  - E(B-V) = parameter in extinction law (= A_0/3.1)
    or reddening in Johnson B-V colour
  - different extinction laws (e.g. Cardelli, Fitzpatrick)

Summary

- XP spectra: low-resolution, high-SNR, optical spectra
- GaiaXPy: sampling, simulation, integrated photometry
- GDR3 will include stellar parameters derived from XP
- choose wisely: data-driven vs “classical” forward-modelling
- XP spectra strongly affected by interstellar extinction
- beware of different parameter distributions in training and application samples