PhotoZ estimation with Gaussian Processes

Delight in DESC-DC2

Bayesian statistics, Maths and cosmological physics

May 27th 2021

Ref:

Data-driven, Interpretable Photometric Redshifts Trained on Heterogeneous and Unrepresentative Data

Boris Leistedt[1,4] and David W. Hogg[1,2,3]

The Astrophysical Journal, 838:5 (14pp), 2017 March 20

https://doi.org/10.3847/1538-4357/aa6332

Sylvie Dagoret-Campagne

IJCLab/IN2P3/CNRS

Bayesian Hierarchical Modeling of photometric redshift Using PZ Gaussian Process prediction in LSST DC2 context

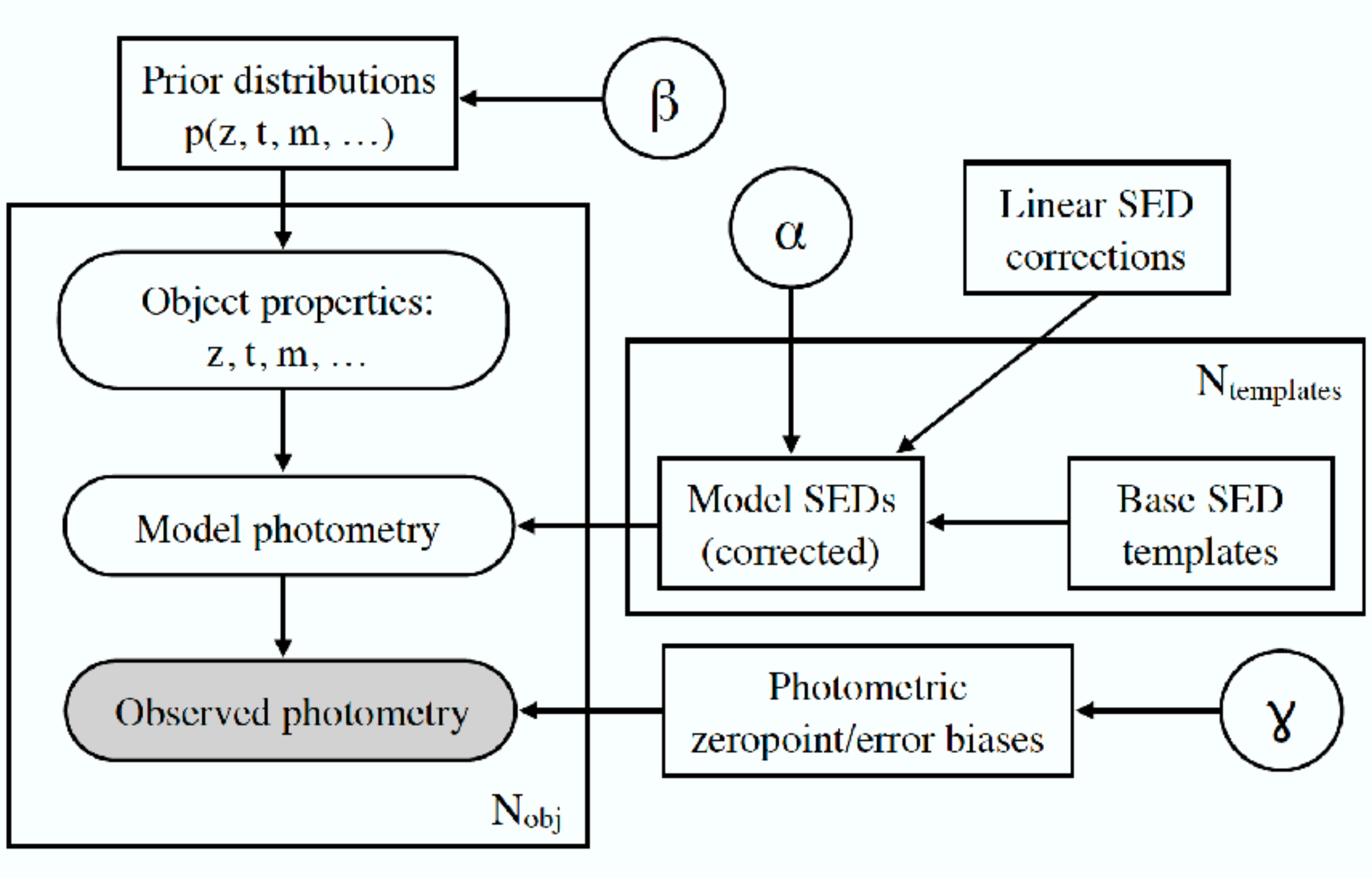
Use Delight (https://github.com/ixkael/Delight or https://github.com/LSSTDESC/Delight)

- Gaussian Processes for Redshift estimation —> hybrid method between Template fitting and ML
  - Based on Flux-Redshift cosmological model also referred as Templates
  - Add flexibility to the model by adding parameters:
    - hyper parameters α, β, γ in different stage of the model (with/without priors) will require optimisation
    - Nuisance parameters that pdf will be marginalized
  - Smaller training set required but allowed not to be fully representative of truth

Bayesian inference of posteriors

$$p(z/\hat{\boldsymbol{F}}) = \int dt\, p(\hat{\boldsymbol{F}}/z,t)p(z,t)$$

Posterior — Likelihood on fluxes using GP — Prior



Delight Code works with two tasks :
- **1) Bayesian Template Fitting**
  - ➡No training dataset
  - ➡ PZ Estimation only
- **2) Bayesian GP fixing with flux prediction in likelihood done by a Gaussian Process**
  - ➡Training and Validation dataset
  - ➡**PZ Learning** / **PZ Estimation (prediction)**

Delight works on two data types:
- Internal mock dataset (internal control)
- External data (by example DC2 - 5 Years 300°^2)

1) Leistedt, Boris & Hogg, David (**Results from Delight code on SDSS data**)
Data-driven, Interpretable Photometric Redshifts Trained on Heterogeneous and Unrepresentative Data
2) Leistedt, Boris & Hogg, David & Wechsler, Risa & DeRose, Joe. (2019).
Hierarchical modeling and statistical calibration for photometric redshifts.

Bayesian inference of redshift *z* from noisy fluxes : $\hat{F} = (\hat{F}_1, \ldots \hat{F}_b, \ldots \hat{F}_{N_b})$

For each target galaxy :

$$p(z/\hat{\boldsymbol{F}}) = \int dt\, p(\hat{\boldsymbol{F}}/z, t)p(z, t) \simeq \sum_i \underbrace{p(\hat{\boldsymbol{F}}/z, t_i)}_{\substack{\text{Likelihood based on} \\ \text{Flux-Redshift model} \\ \text{at } z_i \text{ for template } t_i}} \overbrace{p(z/t_i)p(t_i)}^{\text{Prior on galaxy template } t_i \text{ and its 2D redshift distribution } p(z, t_i)}$$

	Template Fitting	Gaussian processes
Prior $p(z_i, t_i)$	The redshift priors on SED templates	The redshift priors on redshift taken to be a gaussian at **each training galaxy of redshift z_i**
Likelihood $p(\hat{\boldsymbol{F}}/z, t_i)$	Use the analytical Flux-Redshift model	Use GP formula to predict • average flux $F^*(z)$ • covariance. $\Sigma^*_F(z)$ See later

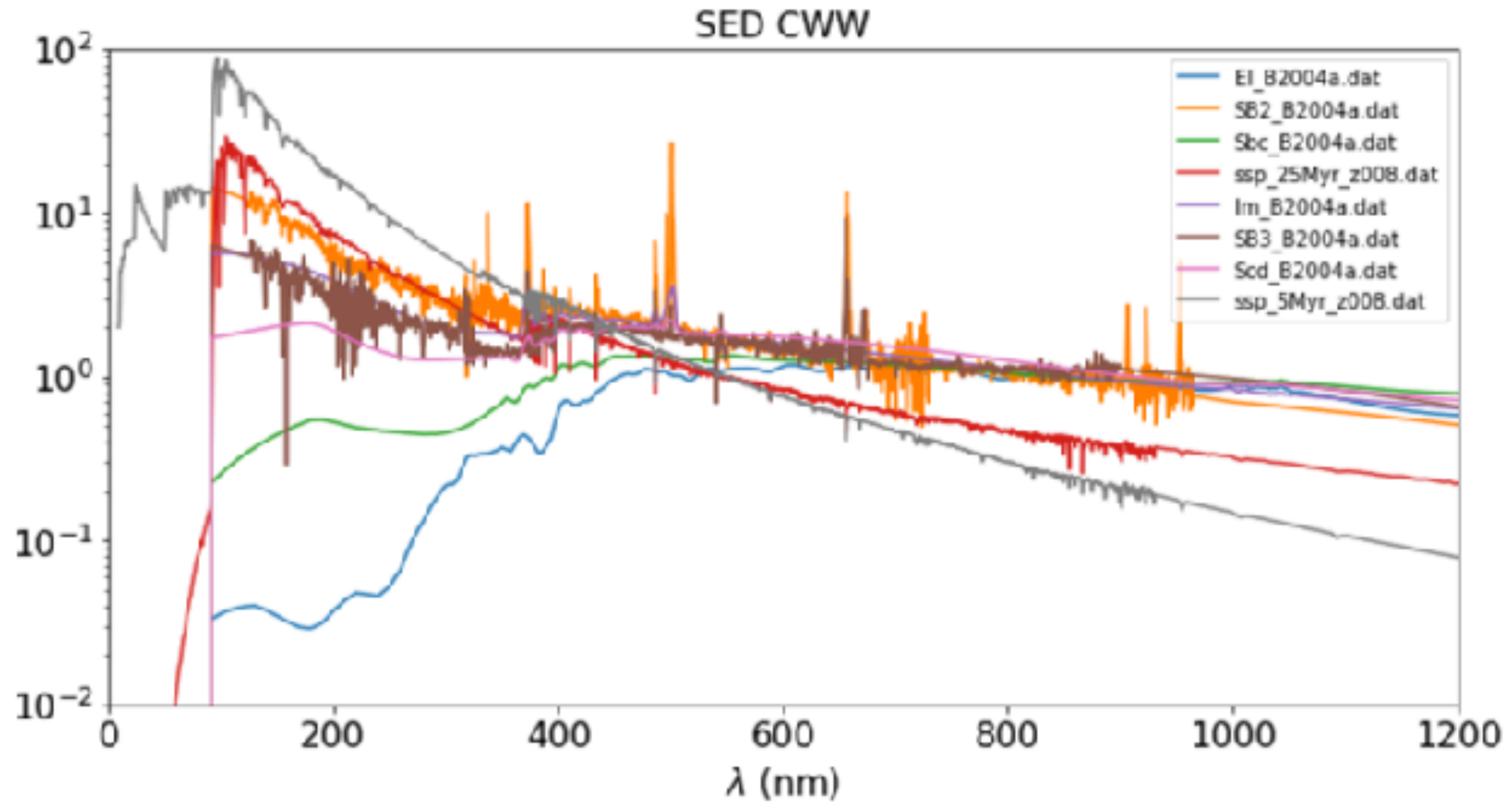
$$p(\underbrace{\hat{\boldsymbol{F}}/z}_{target}, t_i) = p(\hat{\boldsymbol{F}}/z, \underbrace{z_i, \hat{F}_i}_{training}) = \int d\boldsymbol{F}\, p(\hat{\boldsymbol{F}}/\boldsymbol{F})\, p(\boldsymbol{F}/z, z_i, \hat{F}_i)$$

$$p(\boldsymbol{F}/z, z_i, \hat{F}_i) = \mathcal{N}(\boldsymbol{F} - \boldsymbol{F}(z)^*; \boldsymbol{\Sigma}^*_{\boldsymbol{F}}(z))$$

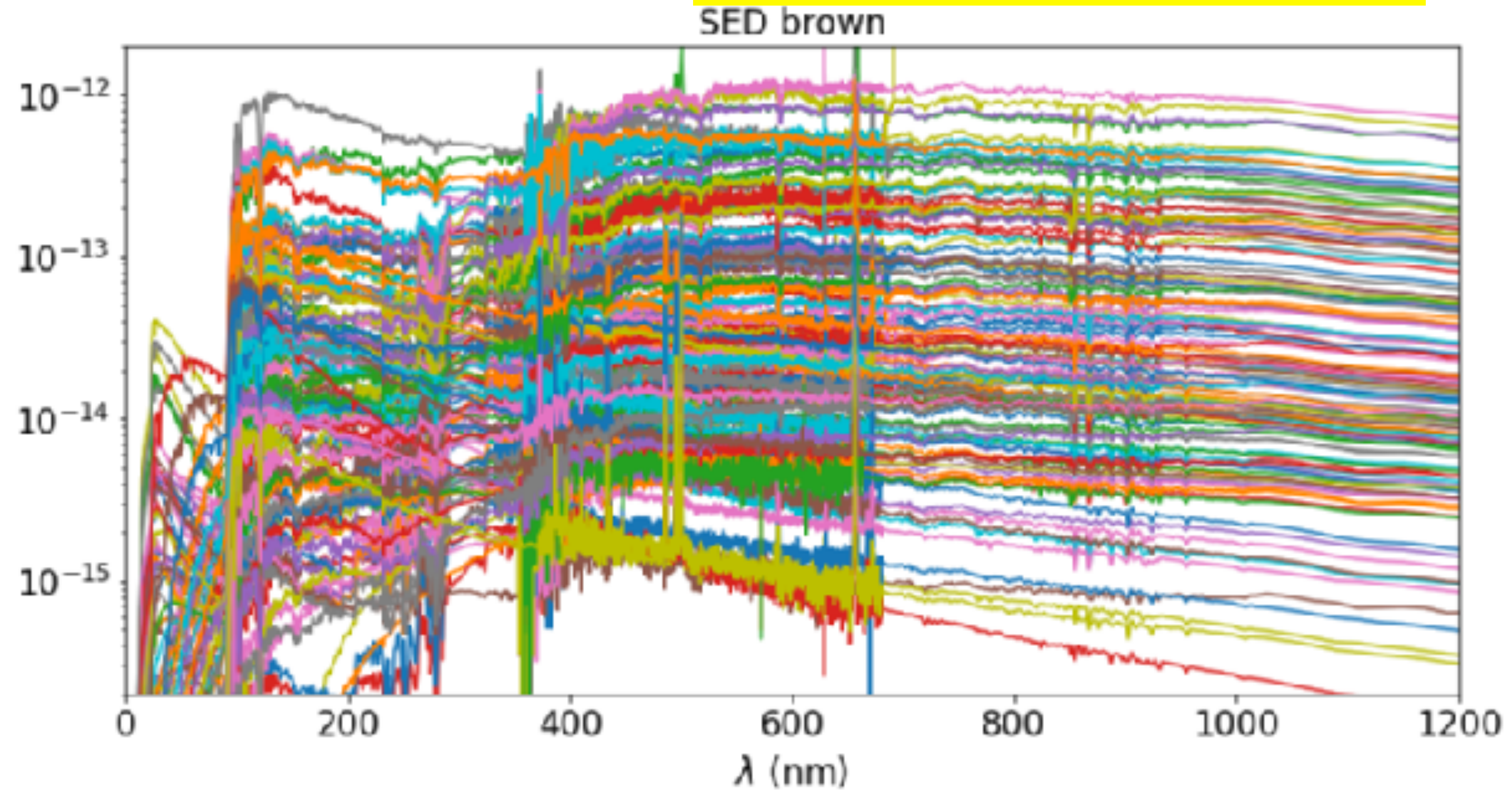
Base Flux-redshift model

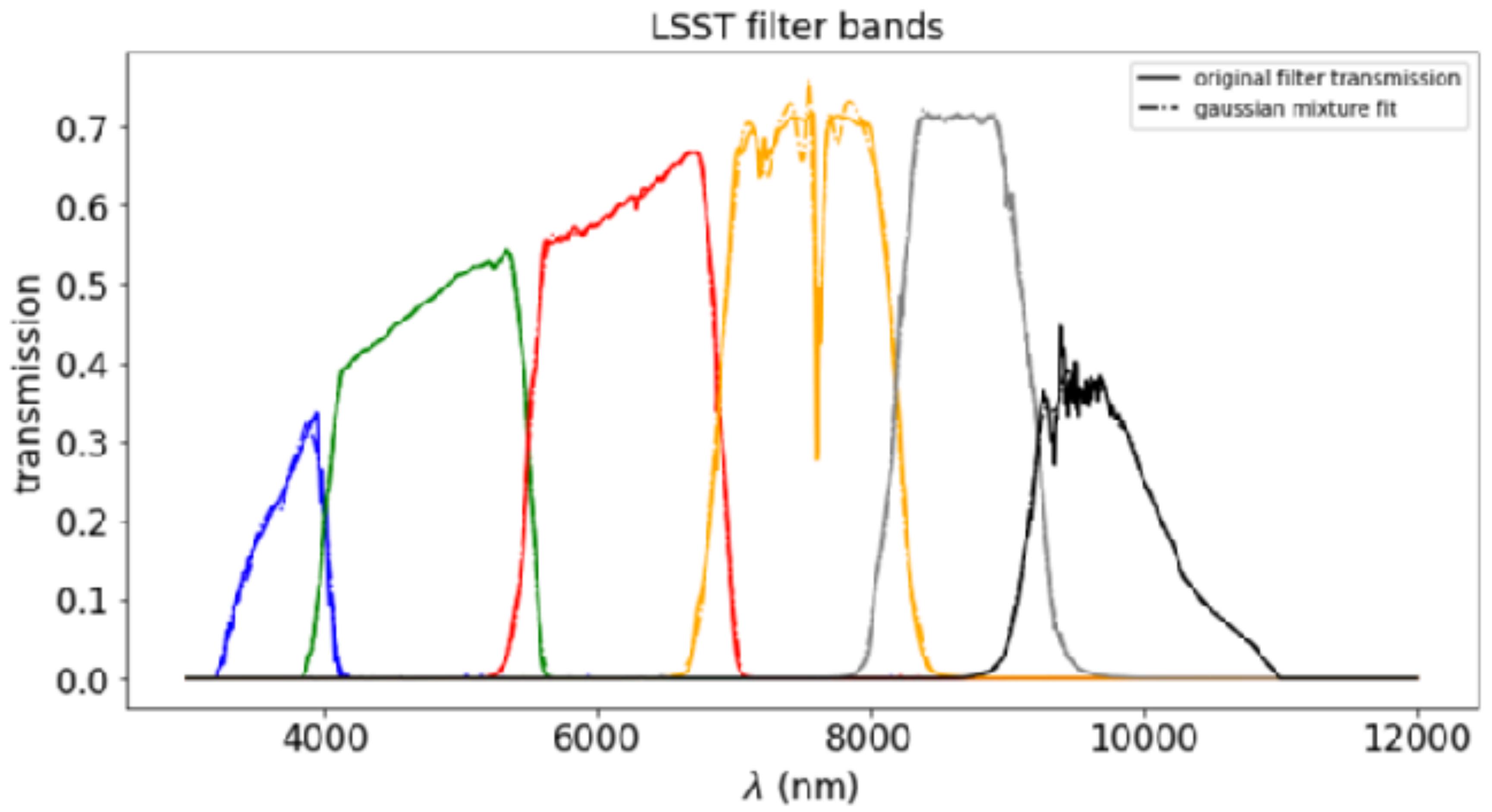
Used to build Fluxes - Redshift model for each template, for Template fitting and GP fitting

- Good redshift Templates for CWW SED

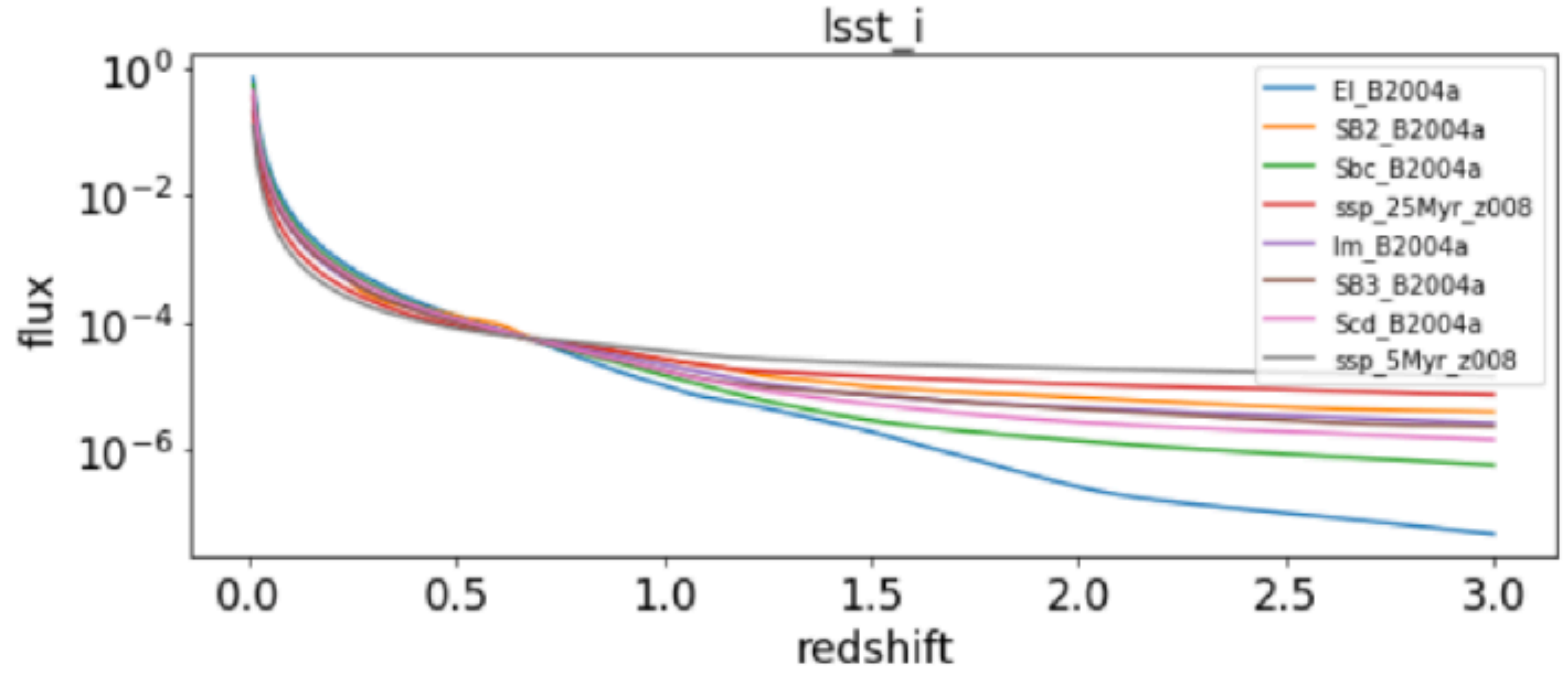


To be used later (see PAW)





Example of flux-redshift model in LSST band I for 8 $L_\nu(\lambda)$ template models



- Flux-Redshift model from Luminosity $L_\nu(\lambda_{em})$ and SED $f_\nu(\lambda_{obs}, z)$

$$f_\nu(\lambda_{obs}, z) = \frac{1+z}{4\pi D_L^2(z)} L_\nu\left(\frac{\lambda_{obs}}{1+z}\right)$$

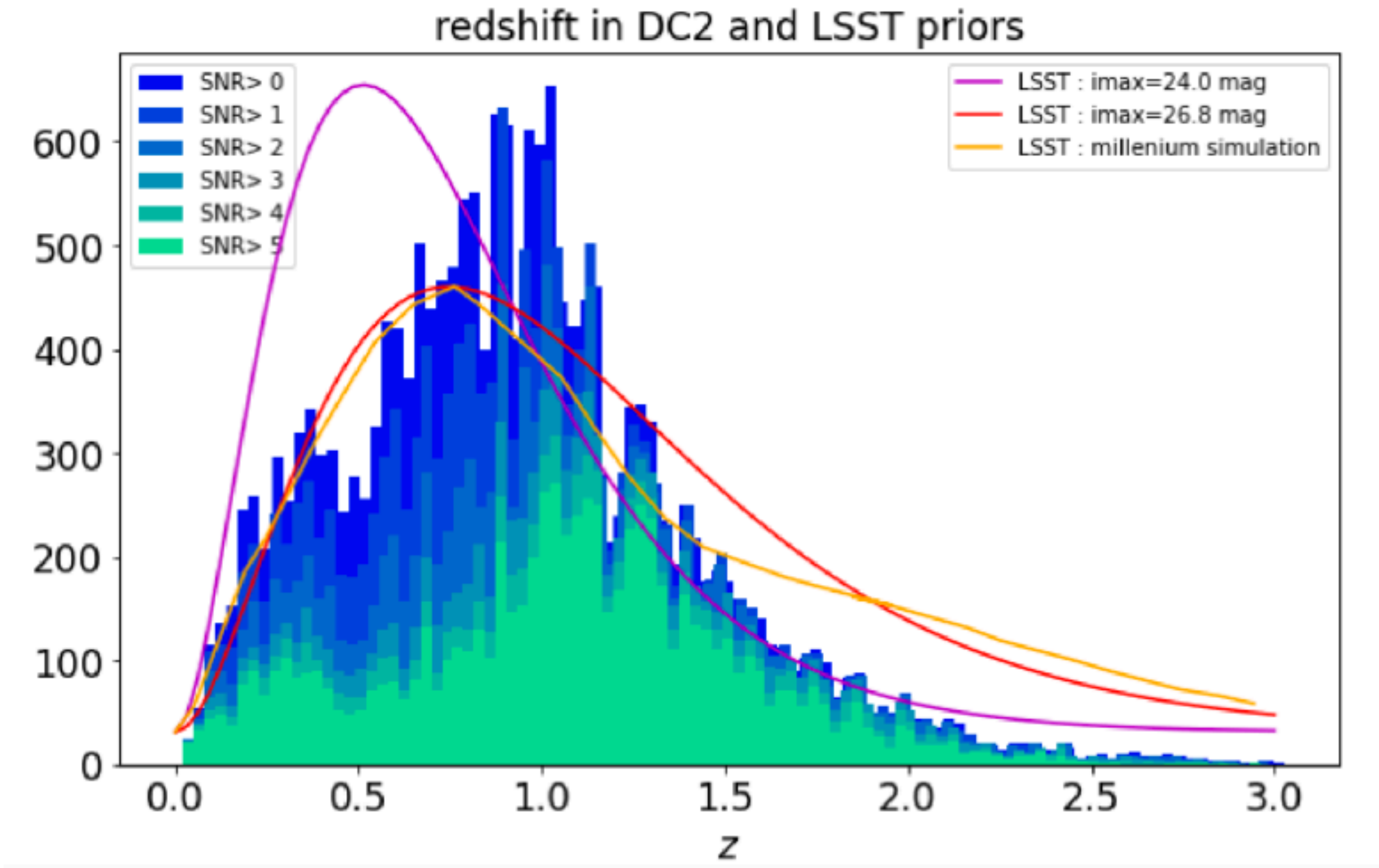
$$\boxed{F_b(z) = \frac{(1+z)^2}{4\pi D_L^2(z) g_{AB} C_b} \int_0^\infty L_\nu(\lambda_{em}, z) V_b\left(\lambda_{em}(1+z)\right) d\lambda_{em}}$$

Reshift prior choice : Extension of redshift range from [0,1] to [0,3]

- Template fitting use this prior for each SED template
- Gaussian process fitting use training data redshift distribution as prior
- Redshift in mock data are generated according a uniform distribution in [0.,3]

LSST Science Book

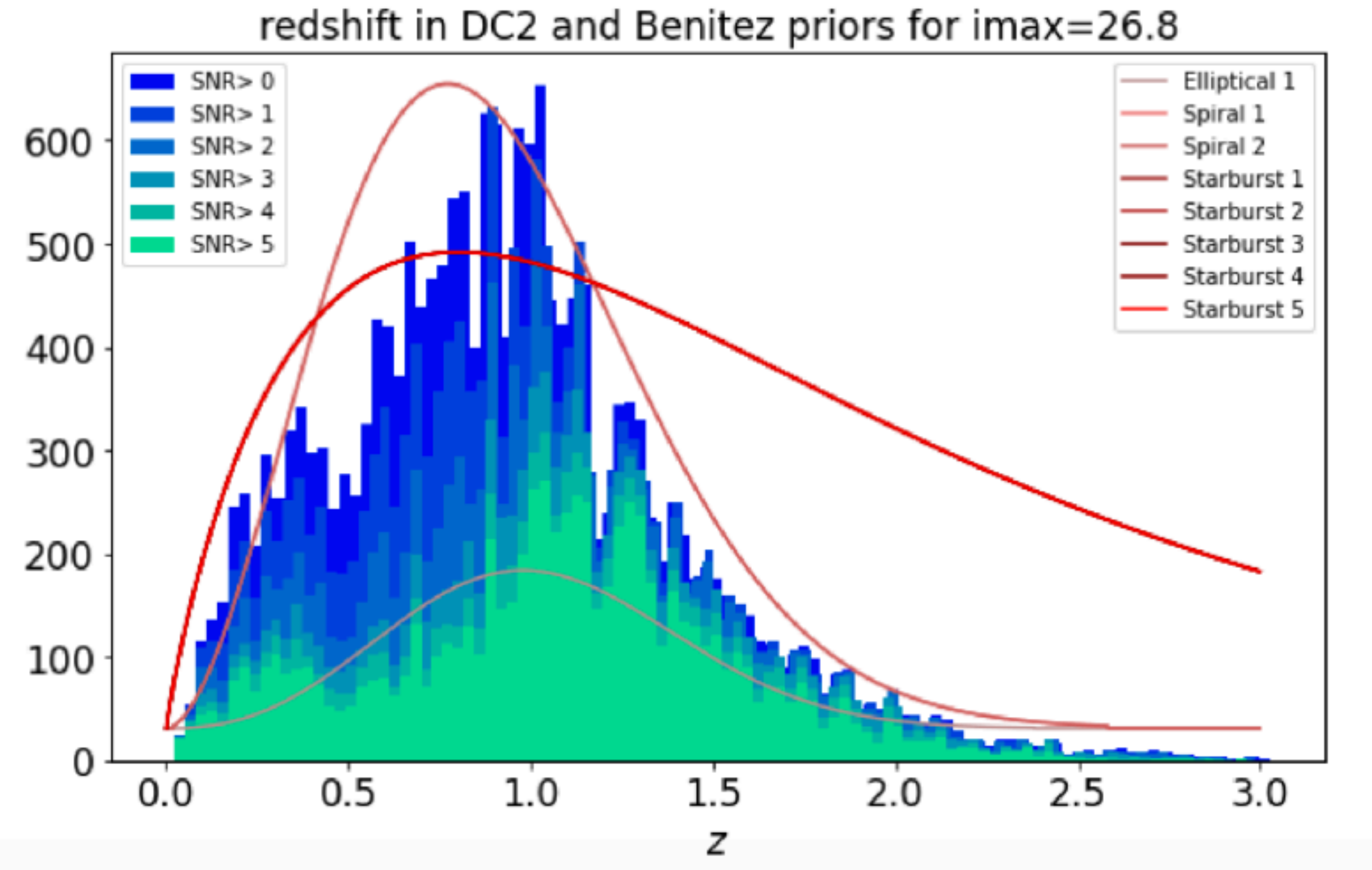
$$p(z) = \frac{1}{2z_0}\left(\frac{z}{z_0}\right)^2 \exp(-z/z_0) \qquad z_0 = 0.0417 * m_{i\ max} - 0.744$$



Benitez 2000 priors

$$p(z) = b_{in}\frac{z}{\beta^2}\exp(-z^2/(2\beta^2))$$

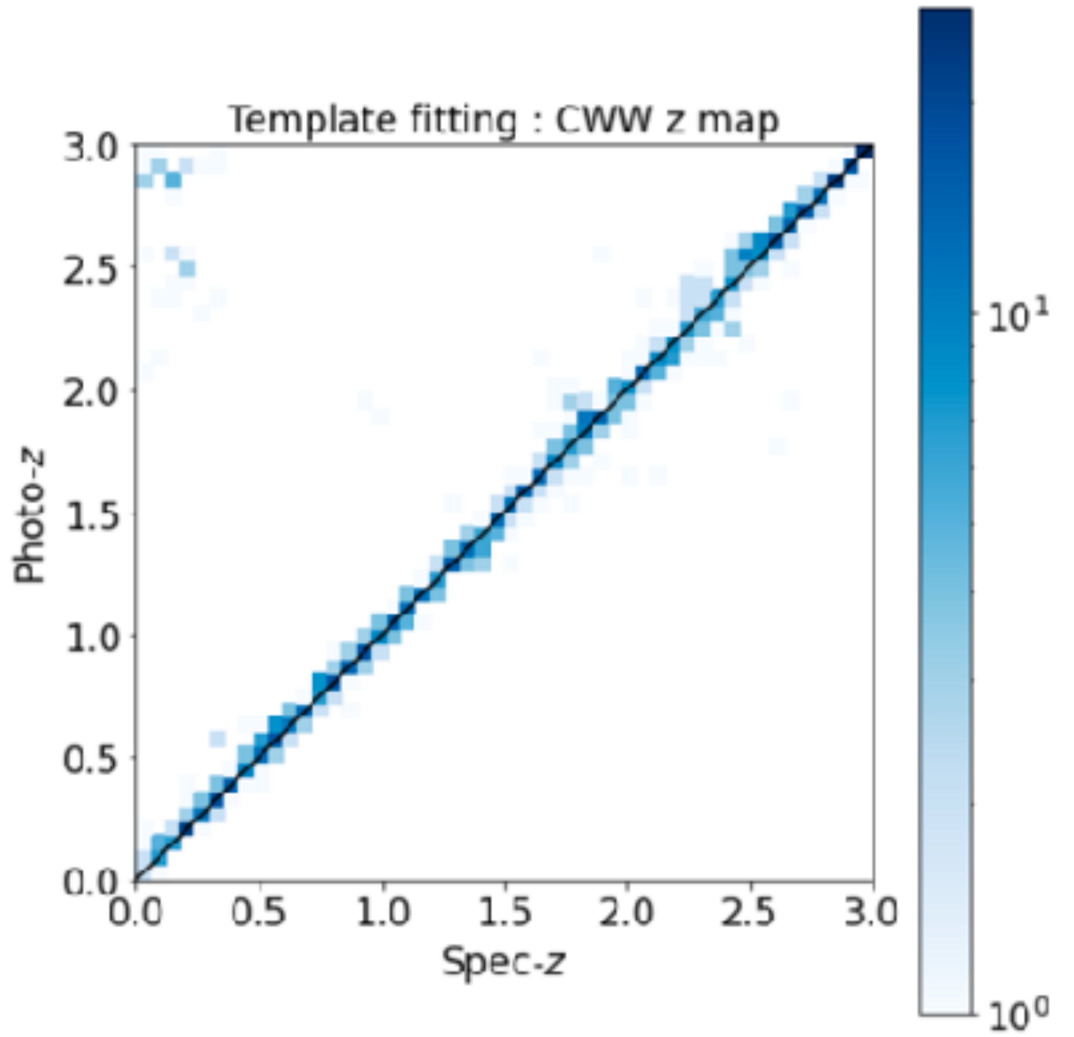
Z pdf width increases with maximum magnitude



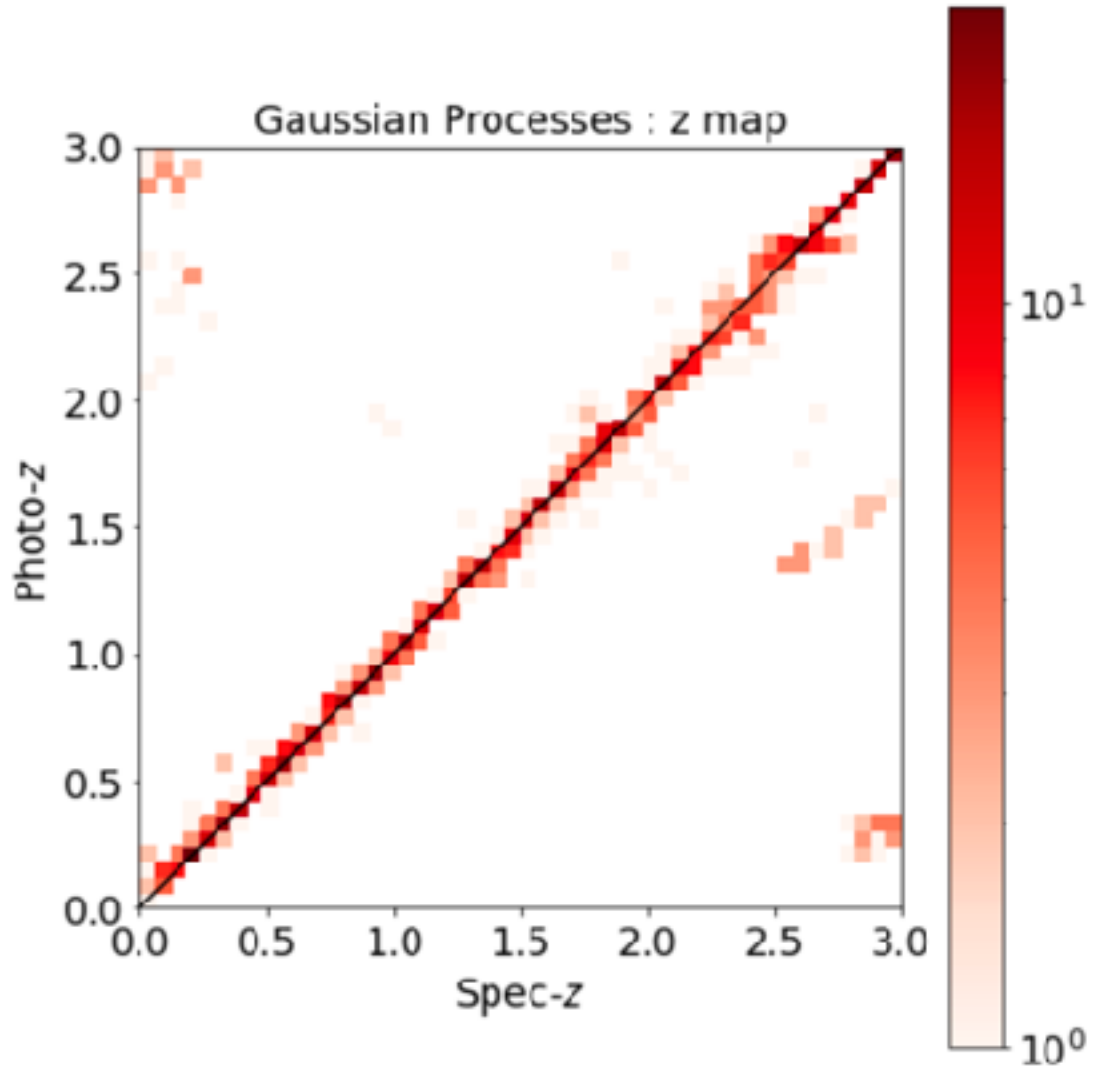
Benitez priors can be calibrated on Data

Preliminary PZ results for unoptimized Delight

mock data



Template Fitting

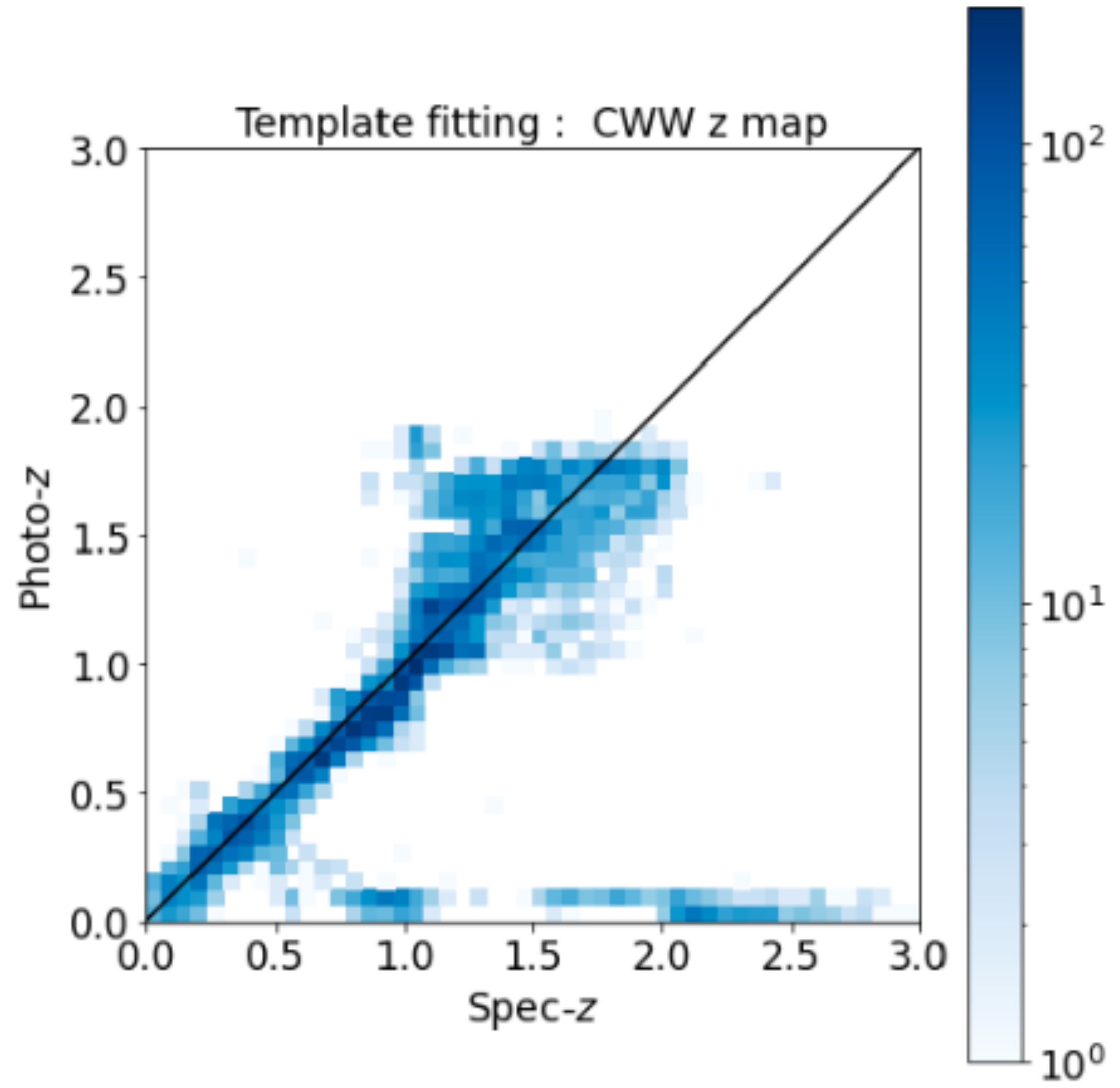


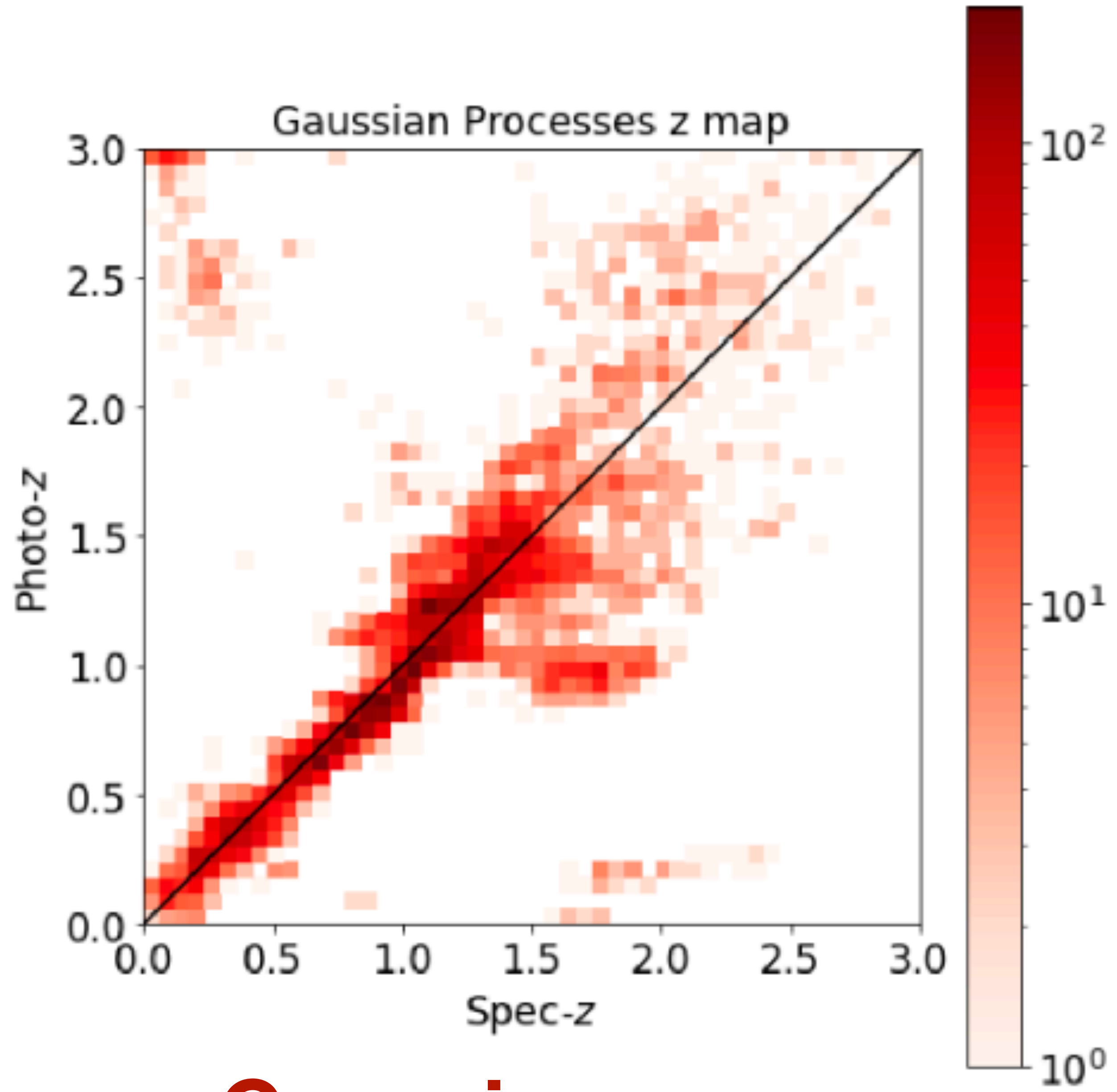
Gaussian Process

Perfect redshift estimation with mock data
For both:

- Template fitting
- Gaussian process

DESC-DC2 flux-redshift data





Template fitting

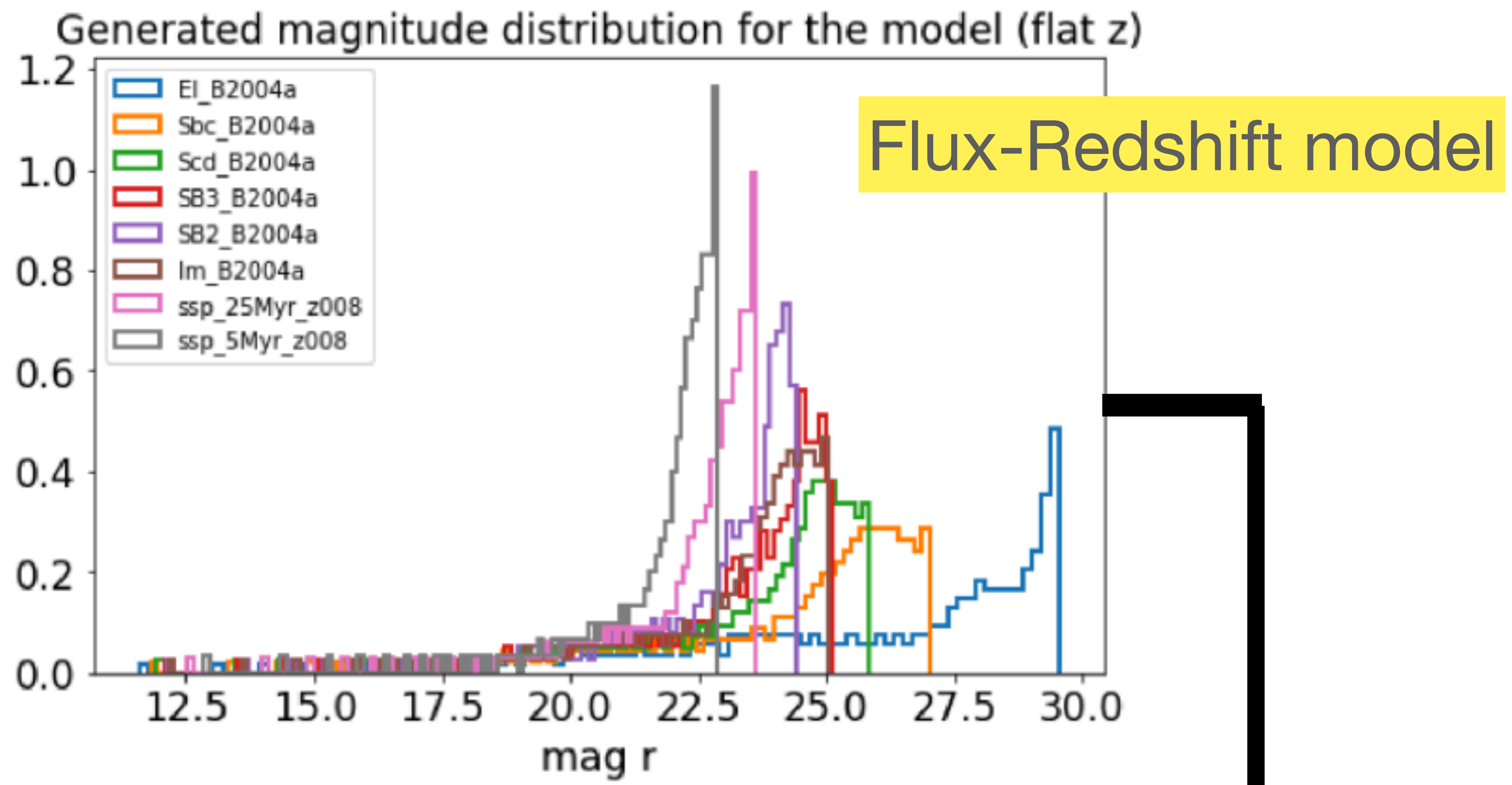
- Good for z< 1
- Not good for 1<z<2
- Fails for z> 2

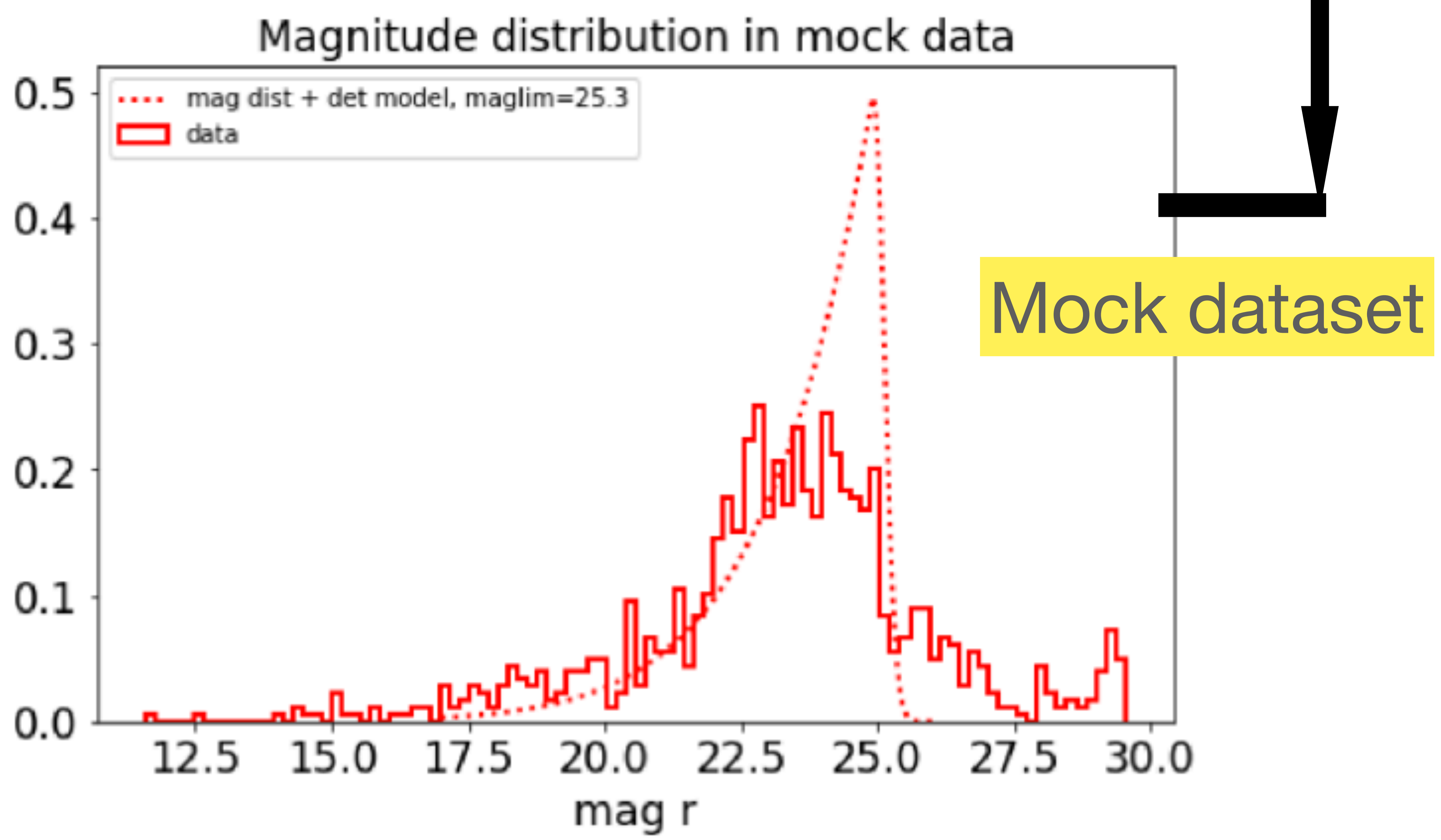
Gaussian process

- Good for z< 1
- Not good for z>1
-

Interpretation of those results in the following slides

Magnitude distribution



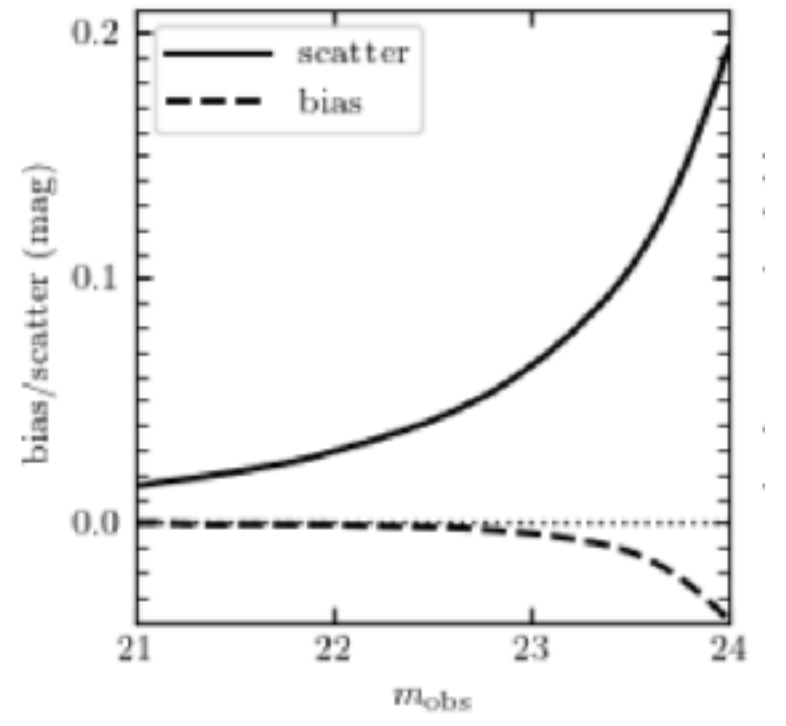


Mock Data:

- Magnitude range : Δm = 18

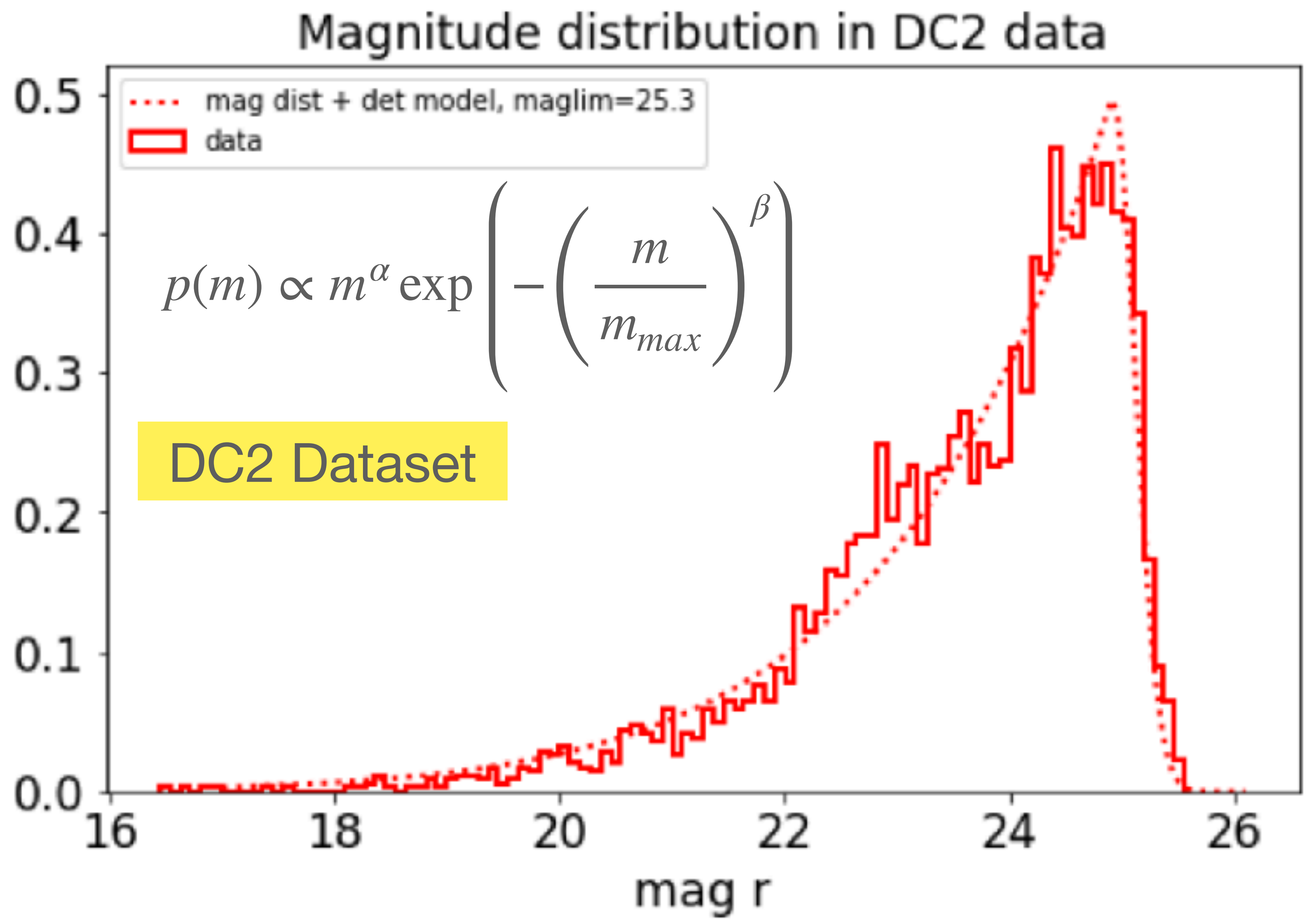
DC2 Data:

- Magnitude range : Δm = 9



From astroml book Statistics, data mining And ML in astronomy

- Malmquist-Eddington bias:
  ➡ Photometric flux errors induces bias toward low magnitudes
- <u>Base Flux-Redshift Model does not describe well DC2 data</u>

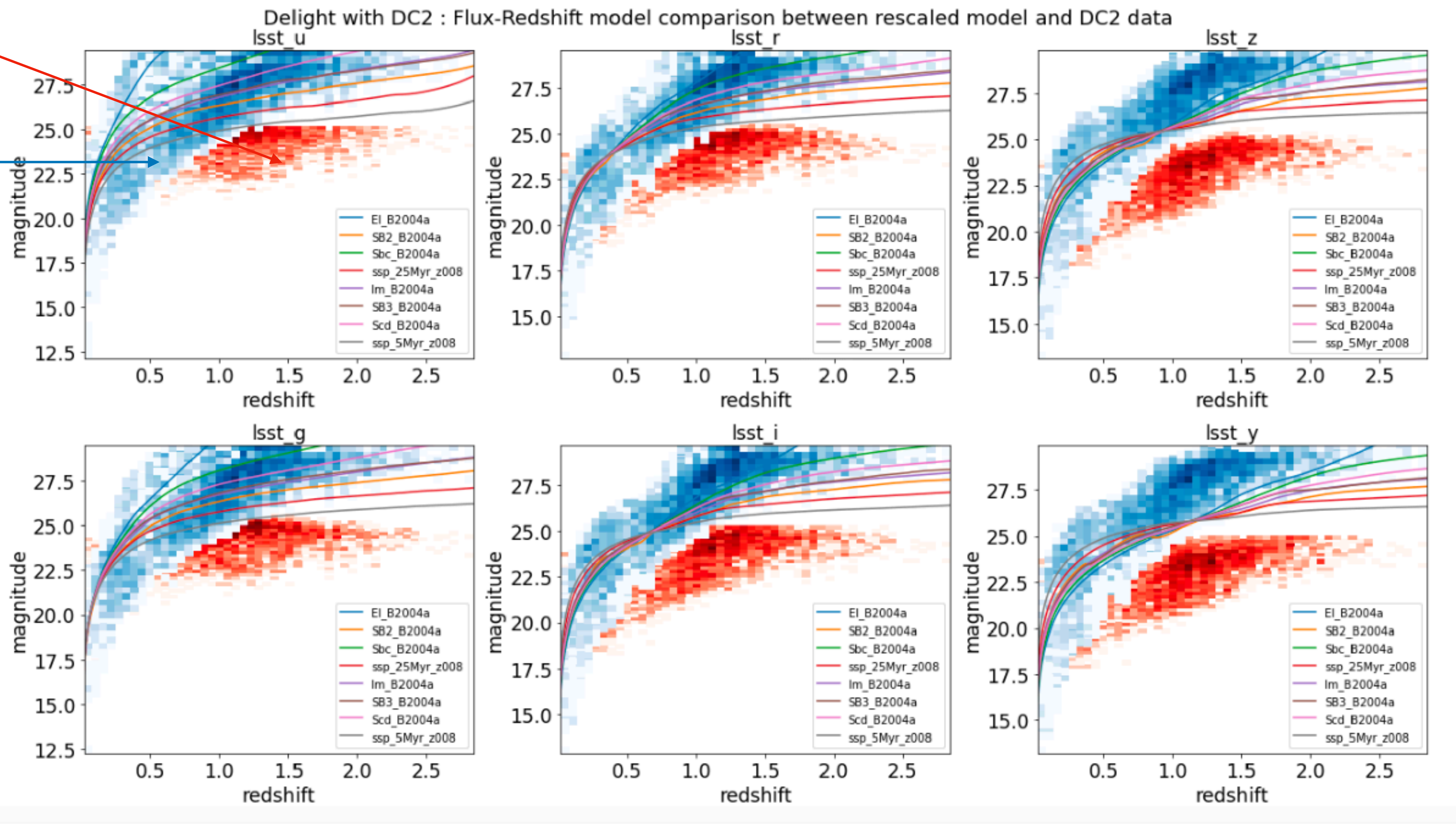


Flux biases in DC2 Data

In DC2 training dataset

(At input of Gaussian Process Learning)

- Flux- Redshift DC2 data
- For each training sample $(z_i, \hat{F}_i^B)$, prediction of fluxes at any other z for each template t_i
- The models are roughly rescaled



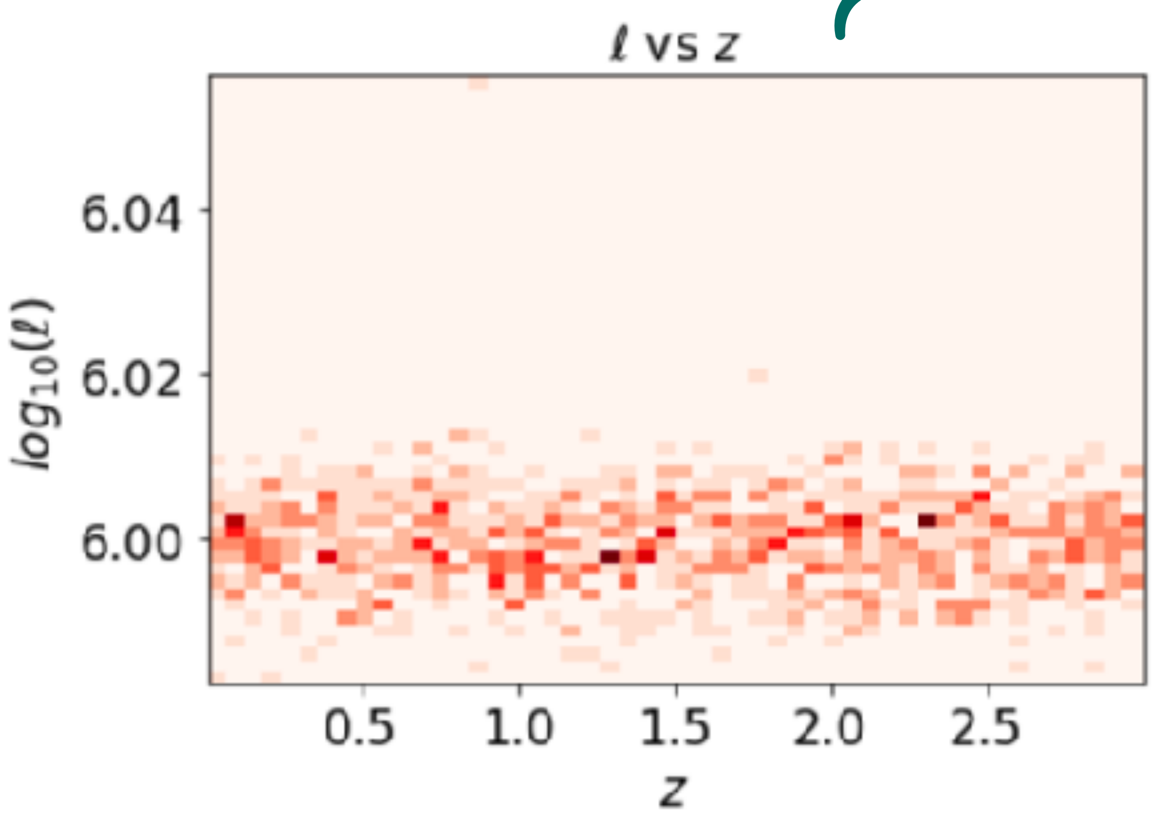
Correction for Flux bias and luminosity evolution wrt z in Gaussian Process learning

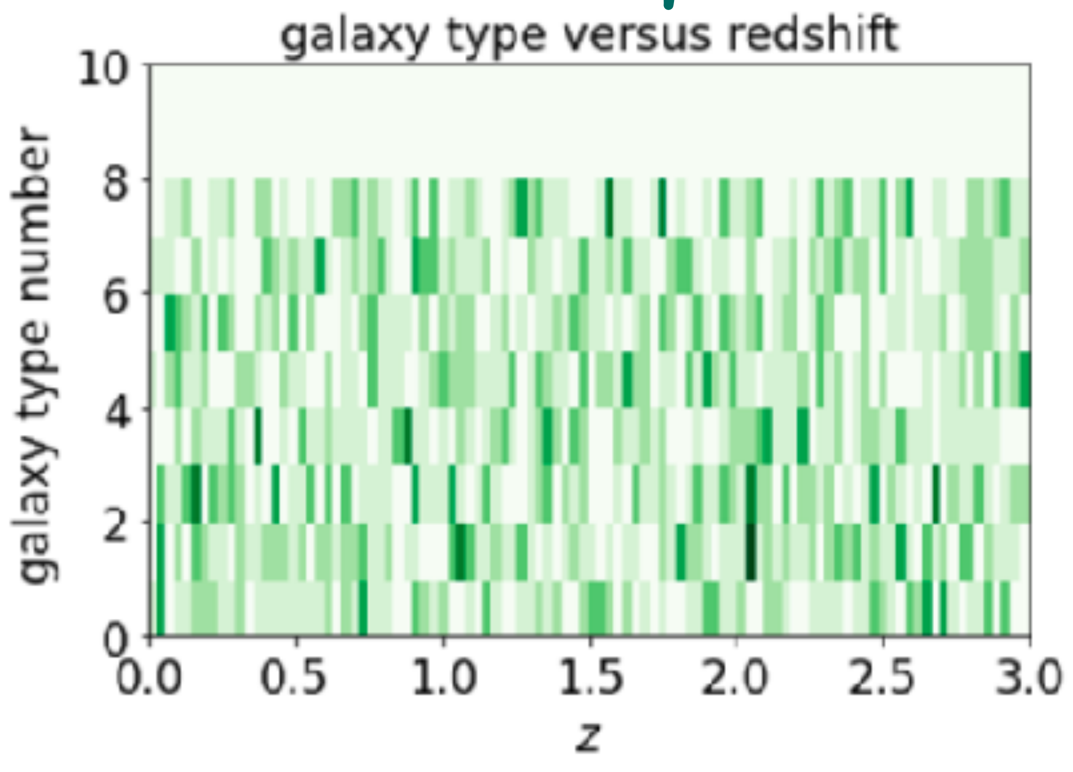
Luminosity (nuisance) parameter $\ell(z)$ inside likelihood

$$p(\hat{\boldsymbol{F}}/z, z_i, \hat{F}_i) = \int d\ell \, \mathcal{N}\left(\hat{\boldsymbol{F}} - \ell(z)\boldsymbol{F}^*(z); \boldsymbol{\Sigma}_{\hat{\boldsymbol{F}}} + \ell^2(z)\boldsymbol{\Sigma}^*_{\boldsymbol{F}}(z)\right) p(\ell) \qquad p(\ell) = \mathcal{N}(\hat{\ell} - \ell, \sigma^2_\ell)$$

- For each training sample at z, find the best rescaling factor ℓ of data wrt model
- Use the best template type t_i (minimum χ^2 for each training sample) —> « latent SED »

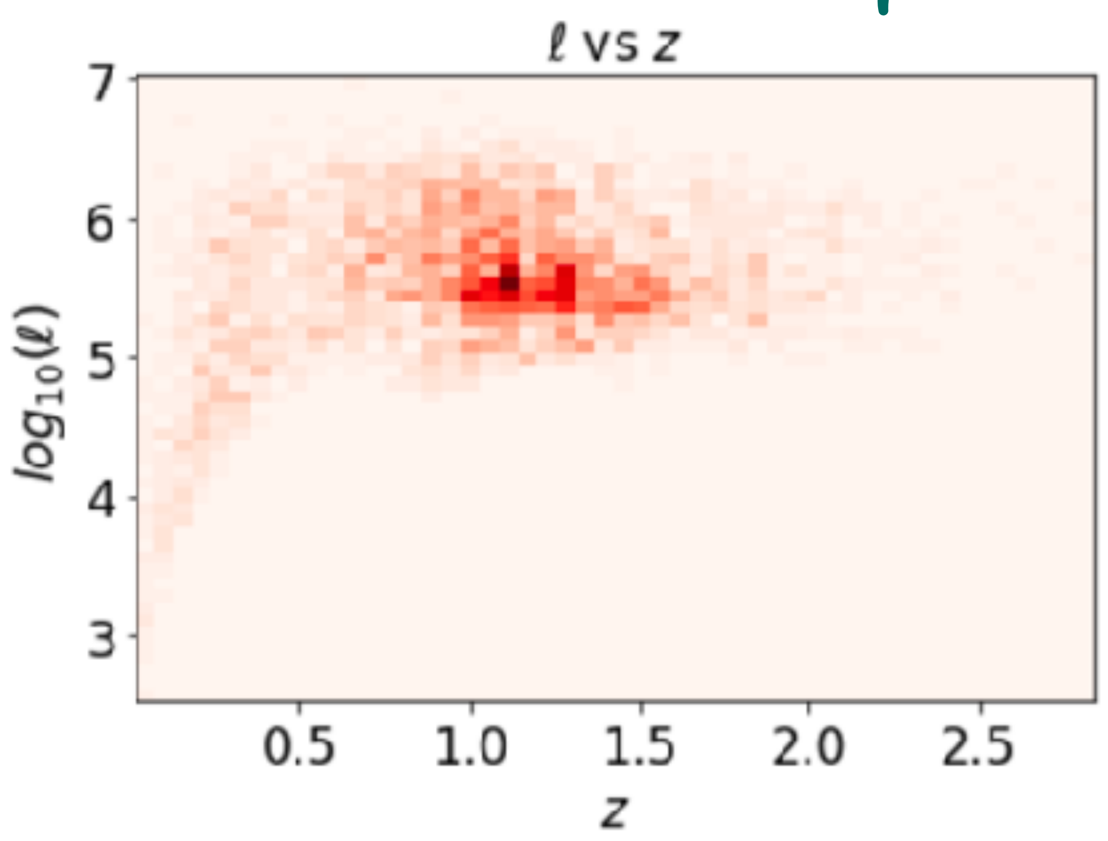
- Mock dataset :

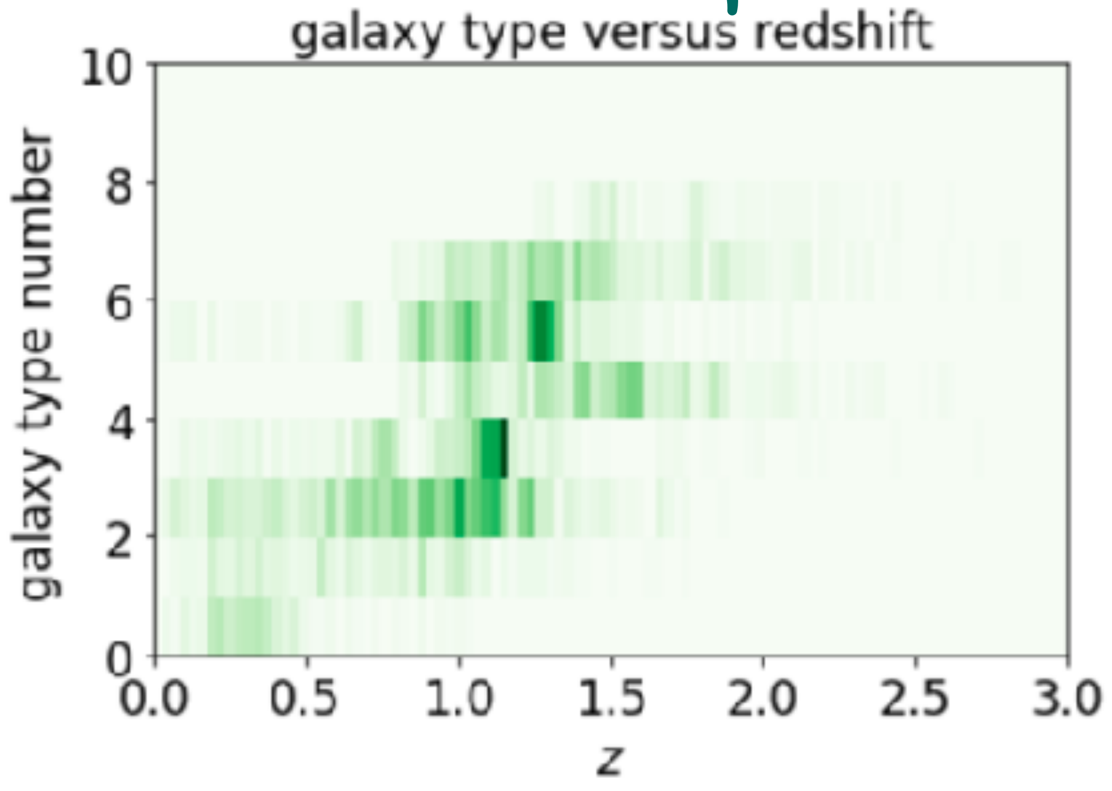




Mock data were drawn with uniform pdf in z and a fixed luminosity factor $\ell = 10^6$

- DC2 data :





Selected Galaxy type during GP learning

Correction of flux bias and luminosity evolution in Template Fitting

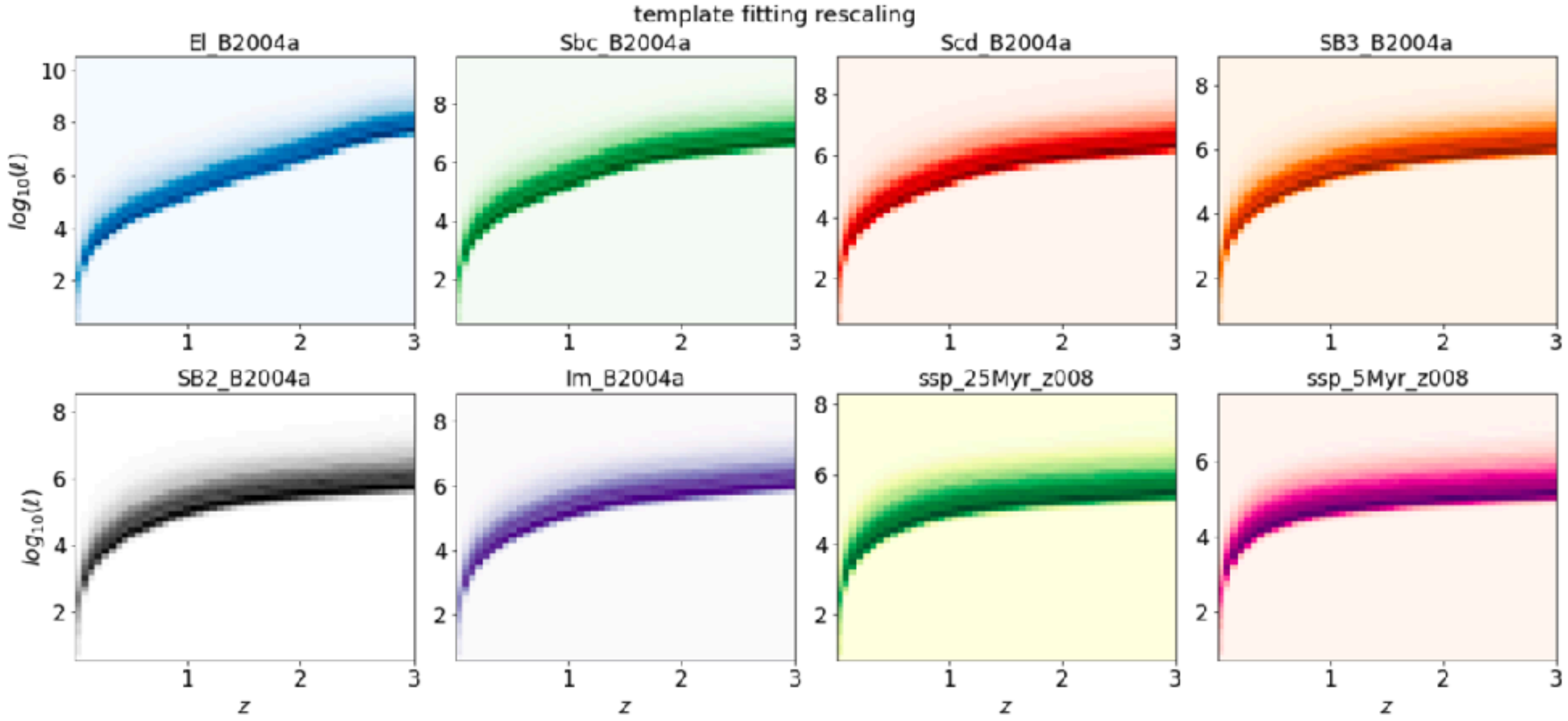
Luminosity (nuisance) parameter ℓ and magnitude vs redshift

- One curve $\ell(z)$
- One $F_b(z)$

per SED template

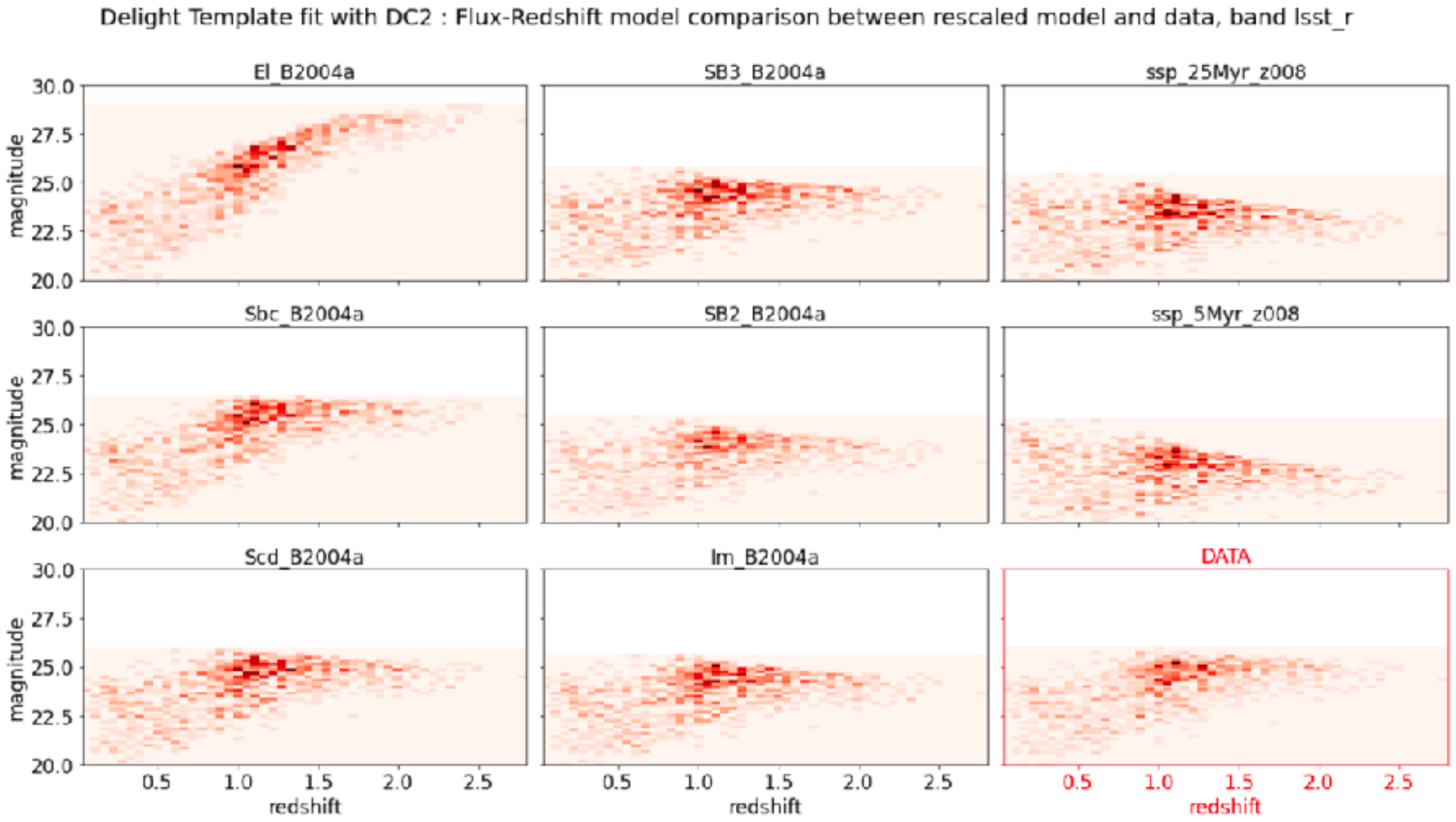
Shown for DC2 Data (similar for mock data)

Luminosity scaling factor vs target galaxy redshift
For each SED template model



AB Magnitude galaxy vs target galaxy redshift in LSST red filter
Comparison of SED template models / DC2 Data

Used to compute the likelihood $p(\hat{\boldsymbol{F}}/z, t_i)$

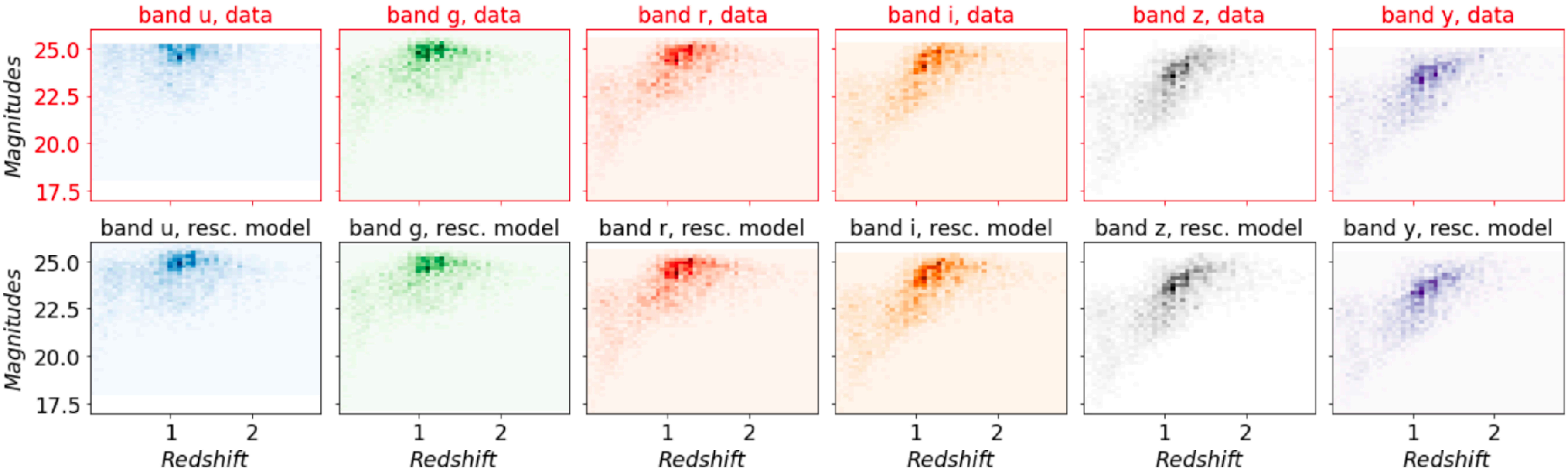


Redshift Posterior $$p(z/\hat{\boldsymbol{F}}) \simeq \sum_i p(\hat{\boldsymbol{F}}/z, t_i)p(z/t_i)p(t_i)$$ Where i is the SED template index

Correction for bias in Gaussian Process learning

- Comparison of magnitudes/redshift between data and rescaled model

Shown for DC2 Data



Similar for mock data

Gaussian Process redshift estimation

Compute flux Likelihood on Target galaxy

$$p(\underbrace{\hat{\boldsymbol{F}}/z}_{target}, t_i) = p(\hat{\boldsymbol{F}}/z, \underbrace{z_i, \hat{F}_i}_{training}) = \int d\boldsymbol{F}\, p(\hat{\boldsymbol{F}}/\boldsymbol{F})\, p(\boldsymbol{F}/z, z_i, \hat{F}_i)$$

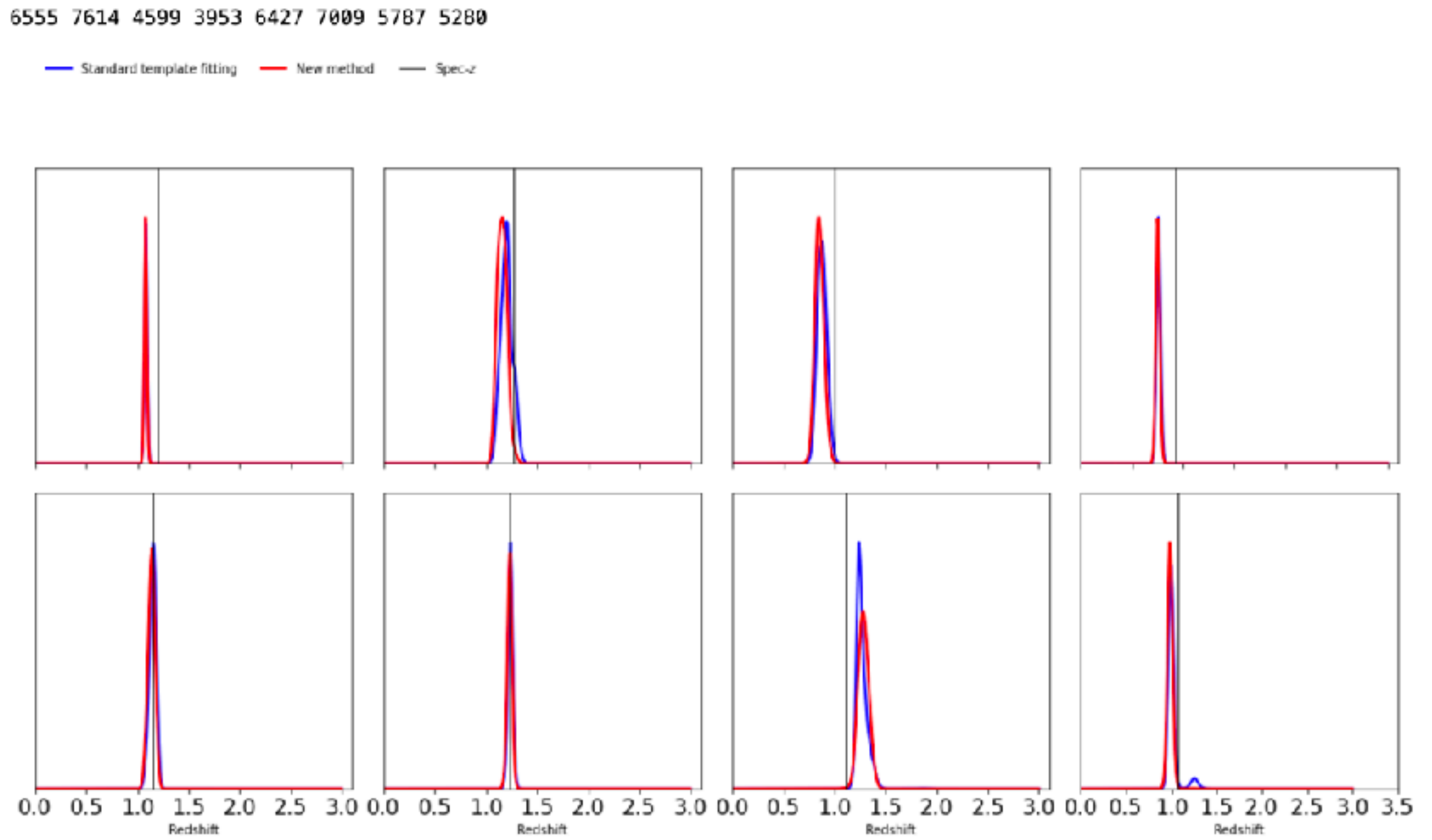
Using GP prediction

$$p(\boldsymbol{F}/z, z_i, \hat{F}_i) = \mathcal{N}(\boldsymbol{F} - \boldsymbol{F}(z)^*; \boldsymbol{\Sigma}^*_{\boldsymbol{F}}(z))$$

GP posterior on target Galaxy

$$p(z/\hat{\boldsymbol{F}}) = \sum_i p(\hat{\boldsymbol{F}}/z, \underbrace{z_i, \hat{F}_i}_{training})\, \overbrace{\mathcal{N}(z_i, \sigma_z)}^{redshift\ prior}$$

Distribution of the highest evidence over target galaxies



Examples of pdf

Standard template fitting — New method

Reminder on what is Gaussian Process and definition of notations

After past introduction of GP by François Leget

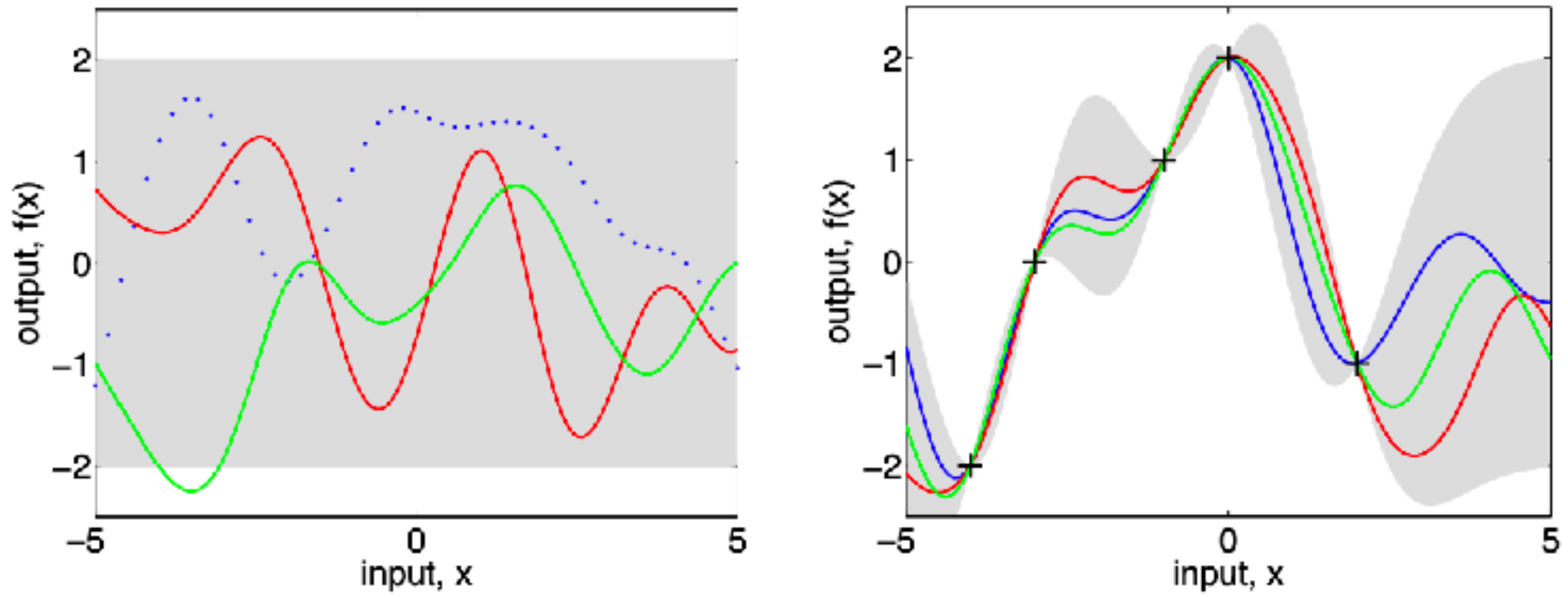
Find the prediction of the function $y = f(x)$

➡ for a new value y_* at x_* (n targets)

➡ from previously m observed training samples (X, y)

$X \simeq (m \times k)$ matrix $\quad y \simeq (m \times 1)$ vector

$$p(\boldsymbol{f}_* \,|\, X_*, X, \boldsymbol{y}) = \mathcal{N}(\bar{\boldsymbol{f}}_*, cov(\boldsymbol{f}_*))$$



(a), prior $\qquad$ (b), posterior

From Rasmussen and Williams book « Gaussian Processes for Machine Learning »

Standard formula of Gaussian Process for noisy data points (on $\boldsymbol{y}$)

Average on predicted $\boldsymbol{y}_*$

$$\underbrace{\bar{\boldsymbol{f}}_*}_{\text{vector } (n \times 1)} = E[\boldsymbol{f}_* \,|\, X, \boldsymbol{y}, X_*] = \underbrace{K(X_*, X)}_{\substack{\text{Prediction phase} \\ (n \times m)}} [\overbrace{K(X, X)}^{\substack{\text{Learning phase} \\ (m \times m)}} + \underbrace{\sigma_n^2 I}_{\substack{\text{Noise on the } m \text{ training} \\ \text{Data points}}}]^{-1} \boldsymbol{y}$$

The Kernel $K(X_1, X_2)$ chosen according an assumption (or a prior)
Ex: on the expected regularity of the function
Ex: the RBF (Radial Basis Function)

$$k(x_1, x_2) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}(x_1^2 - x_2^2)\right)$$

Covariance on predicted $\boldsymbol{y}_*$

$$cov(\boldsymbol{f}_*) = \underbrace{K(X_*, X_*)}_{\substack{\text{Prediction phase} \\ (n \times n)}} - \underbrace{K(X_*, X)}_{\substack{\text{Prediction phase} \\ (n \times m)}} [\underbrace{K(X, X)}_{\substack{\text{Learning phase} \\ (m \times m)}} + \sigma_n^2 I]^{-1} \underbrace{K(X, X_*)}_{\substack{\text{Prediction phase} \\ (m \times n)}}$$

Estimation of target redshift with Gaussian Process in Delight

Find a non parametric function $y = f(x)$, where

- **Y is the vector of LSST fluxes $F(b,z)$ in the 6 LSST filters,**
- **X is a complicated vector of a band index b_j, redshift z and luminosity scaling factor ℓ <u>for each training & target galaxy.</u>**

$X_j = (b_j, z, \hat{l})$, size $(B \times 3)$

$X_k^* = (b_k^*, z^*, l^*)$, size $(B^* \times 3)$

Training noisy fluxes: $\hat{\boldsymbol{F}} = (\hat{F}_1, \ldots \hat{F}_b, \ldots \hat{F}_{N_b})$, size $(B \times 1)$ and covariance matrix $\Sigma_{\hat{F}}$

Predicted noiseless fluxes : $\boldsymbol{F}^* = (F_1^*, \ldots F_b^*, \ldots F_{N_b^*}^*)$, size $(B^* \times 1)$ and covariance matrix Σ_{F^*}

Prior on noiseless model fluxes: $p(\boldsymbol{F}/\boldsymbol{X}) = \mathcal{N}(\mu^F(\boldsymbol{X}), k^F(\boldsymbol{X}, \boldsymbol{X}))$

μ^F average , k^F kernel

ℓ is a nuisance parameter which is to be marginalized in Flux likelihood

Standard formula of Gaussian processes For prediction

Average

$$\boxed{\boldsymbol{F}^* = \underbrace{\mu^F(\boldsymbol{X}^*)}_{\text{The only term available for template fitting}} + \underbrace{k^F(\boldsymbol{X}^*, \boldsymbol{X})[k^F(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\Sigma}_{\hat{\boldsymbol{F}}}]^{-1} \times \left(\hat{\boldsymbol{F}} - \mu^F(\boldsymbol{X})\right)}_{\text{Additional term for GP}}}$$

Covariance

$$\boxed{\boldsymbol{\Sigma}_{\boldsymbol{F}}^* = k^F(\boldsymbol{X}^*, \boldsymbol{X}^*) - k^F(\boldsymbol{X}^*, \boldsymbol{X})[k^F(\boldsymbol{X}, \boldsymbol{X}) + \boldsymbol{\Sigma}_{\hat{\boldsymbol{F}}}]^{-1} k^F(\boldsymbol{X}, \boldsymbol{X}^*)}$$

<u>But what are the chosen expression for μ^F and K^F ?</u>

From which cosmological concepts μ^F and k^F are derived ?

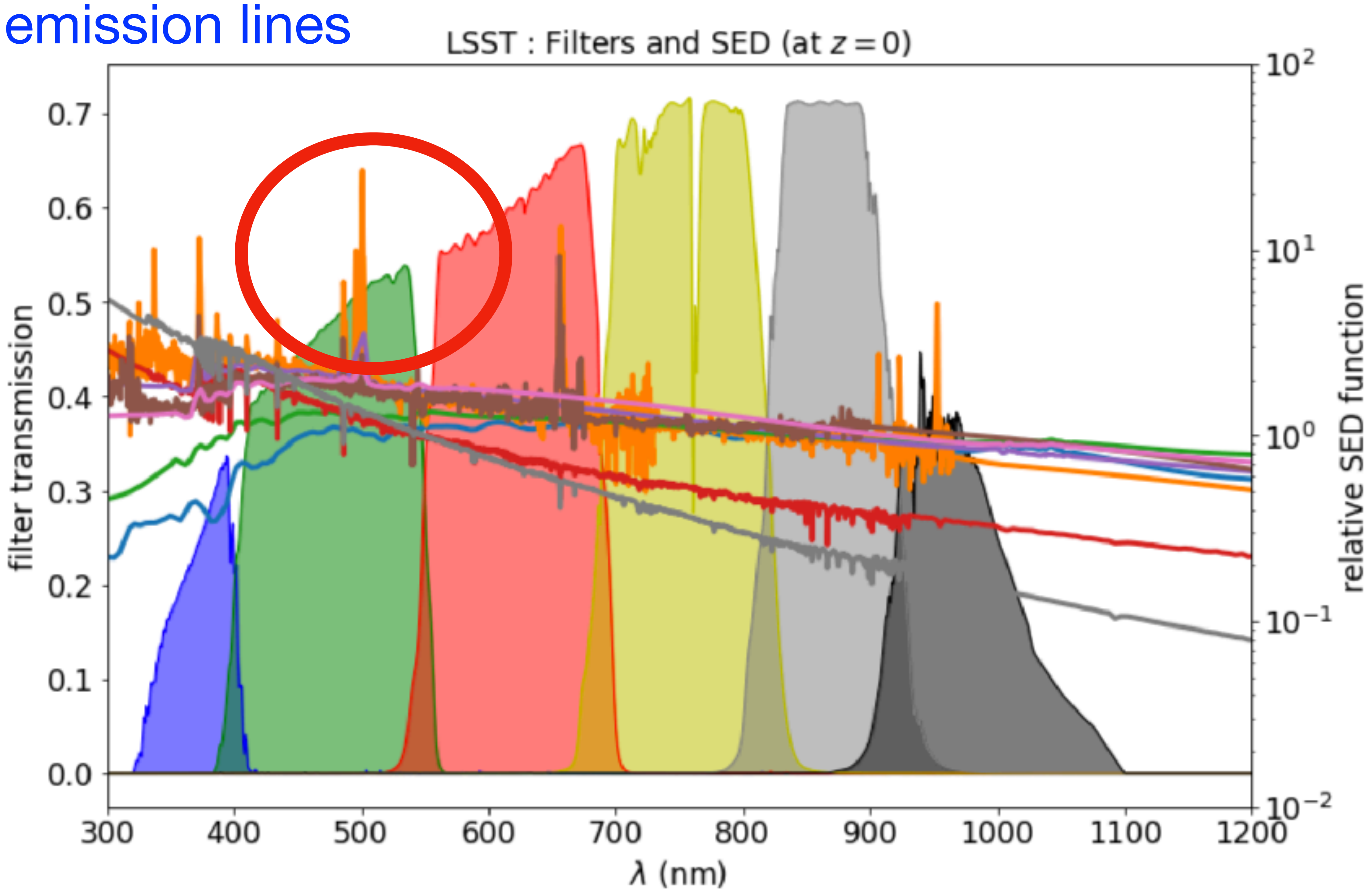
Luminosity is a linear combination of Template + adding eventual emission lines

Luminosity : $L_\nu(\lambda, \boldsymbol{\alpha}, l) = \ell \underbrace{\sum_t^{N_T} \alpha_t T_\nu^t(\lambda)}_{SED\ templates} + \ell \underbrace{R_\nu(\lambda)}_{residuals}$

Residuals: $R_\nu \sim \mathcal{GP}(0, k^\lambda(\lambda, \lambda'))$ $\qquad k^\lambda(\lambda, \lambda')$ chosen to be a RBF

$$L_\nu(\lambda, \boldsymbol{\alpha}, l) \sim \mathcal{GP}\left(\ell \sum_t^{N_T} \alpha_t T_\nu^t(\lambda), \ell \ell' k^\lambda(\lambda, \lambda')\right)$$

Flux : $F_b(z, \boldsymbol{\alpha}, \ell) \sim \mathcal{GP}(\mu^F(b, z, \boldsymbol{\alpha}), k^F(b, b', z, z', \ell, \ell'))$



$$\begin{cases} \boxed{\mu^F(b, z, \ell, \boldsymbol{\alpha}) = \dfrac{\ell(1+z)^2}{4\pi D_L^2(z) g_{AB} C_b} \displaystyle\sum_t^{N_t} \int_0^\infty T_\nu^t(\lambda_{em}, z) V_b\left(\lambda_{em}(1+z)\right) d\lambda_{em} = \ell \sum_t^{N_t} \alpha_t F_b^t(z)} \\ \boxed{k^F(b, b', z, z', \ell, \ell') = \left(\dfrac{(1+z)(1+z')}{4\pi D_L(z) D_L(z') g_{AB}}\right)^2 \dfrac{\ell\ell'}{C_b C_{b'}} \displaystyle\int_0^\infty V_b((1+z)\lambda) V_{b'}((1+z')\lambda') k^\lambda(\lambda, \lambda') d\lambda d\lambda'} \end{cases}$$

Conclusion on this work

- **Delight provides a new way for PZ estimation** based on GP in the context of Bayesian statistics.

  ➡ Compromise between ultra flexible ML without priors on physics requiring a very representative training set and rigid Template fitting with «hard » coded physics model in it,

  ➡ **Extended physical hierarchical model with a moderate number of hyperparameters** (understandable physically) requiring a limited training dataset not necessarily fully representative

- Delight standard configuration (for SDSS) has been **extended for LSST**

  ➡ Redshift priors extended to redshift [0-3] (used for Template Fitting only)

- Delight works well (Template fit & GP) with mock data (no luminosity evolution and flux bias)

- Delight works not that well for DC2 fluxes by now

  ➡ Was expected for Template Fitting,

  ➡**Results for GP are better than Template fitting but far from optimal however encouraging,**

  ★Namely No optimization has been performed

Conclusion / Next steps

- **Optimize GP hyper parameters over DC2** data using CWW SED latent SED
- Extend **SED CWW set to Brown SED** and try to optimize again.
- Many path to explore ways **to refine the GP model**
  - ➡ Add more emission lines,
  - ➡ Find Other features
  - ➡ Many new idea for models see Leistedt, Boris & Hogg(2019) not implemented in Delight