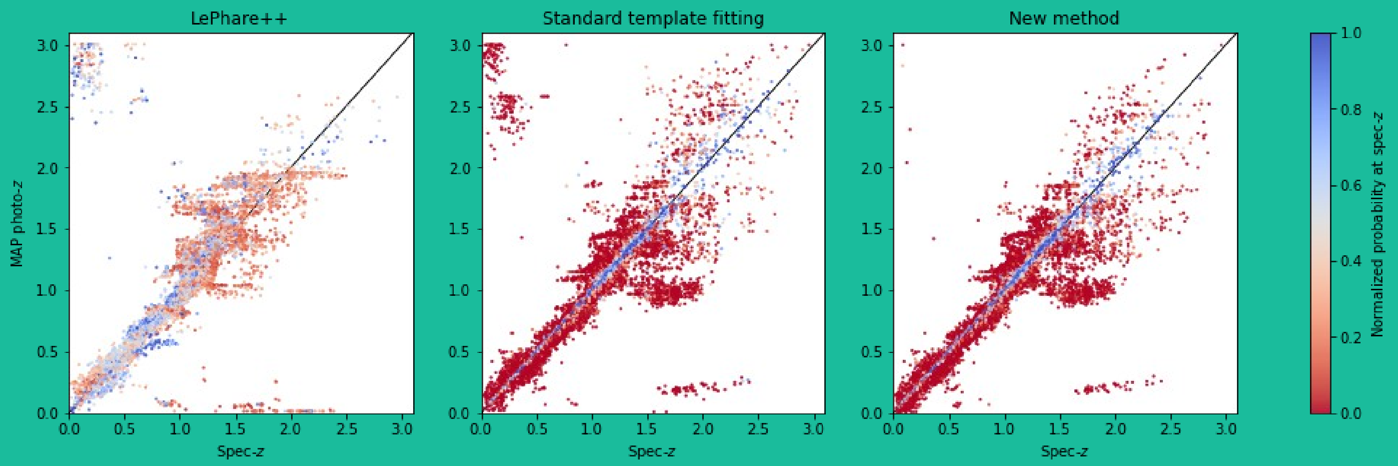


Study of a Hybrid Photo-Z estimator and potential improvements

J. Chevalier, Intern at IJClab, Orsay
LSST France days – May 18, 2022

Supervisor : Sylvie Dagoret-Campagne

Table of Contents

I. Introduction – context

i. Three types of photo-z estimators

- **LEPHARE++ [1] : a reference Template Fitting (TF) tool**

  Color-based ; very customizable and versatile

- **Delight [2] : hybrid TF & Bayesian predictions**

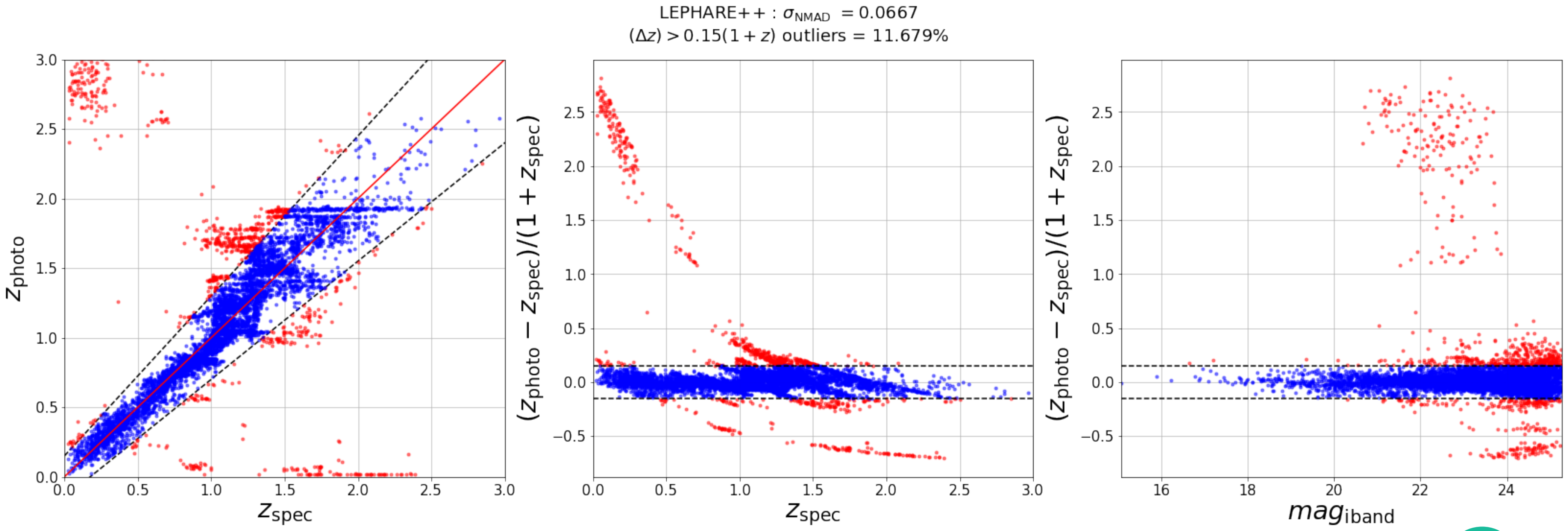
  Flux-redshift model ; Gaussian processes ; scale-free likelihood

- **Machine Learning (ML) [3] : random forest of decision trees**

  Easy to use ; adaptable and readily available ML algorithm (scikit-learn) ; robust

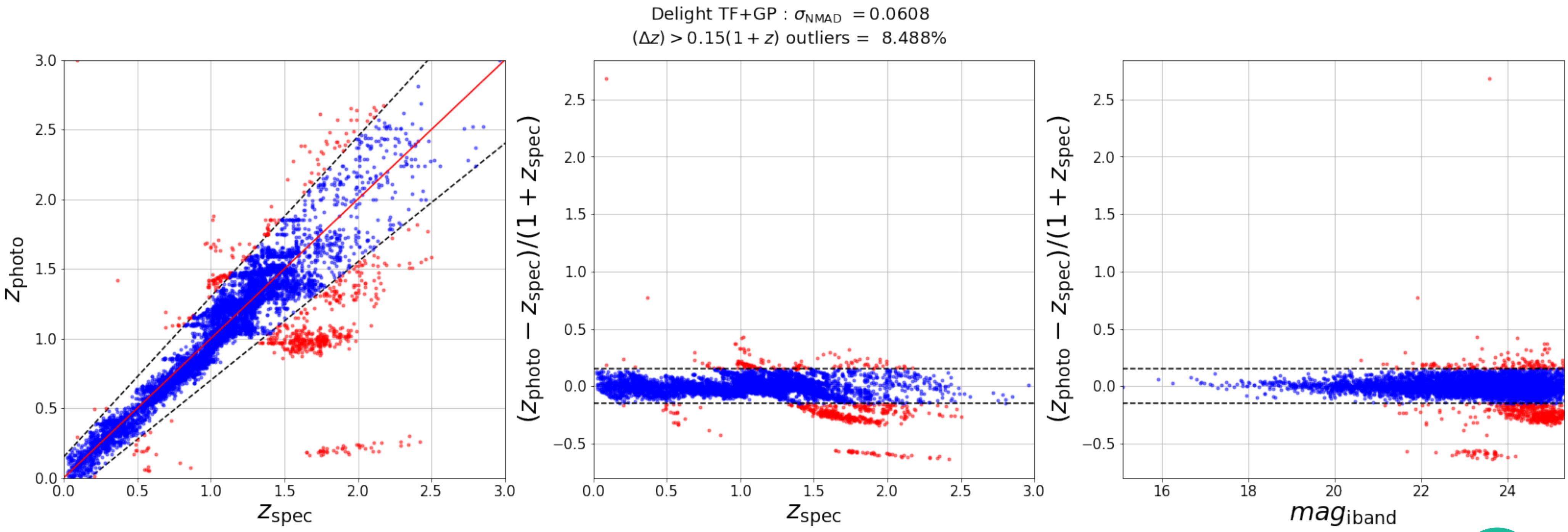
I. Introduction – context (cont.)

LEPHARE results on DESC-DC2 simulated data :



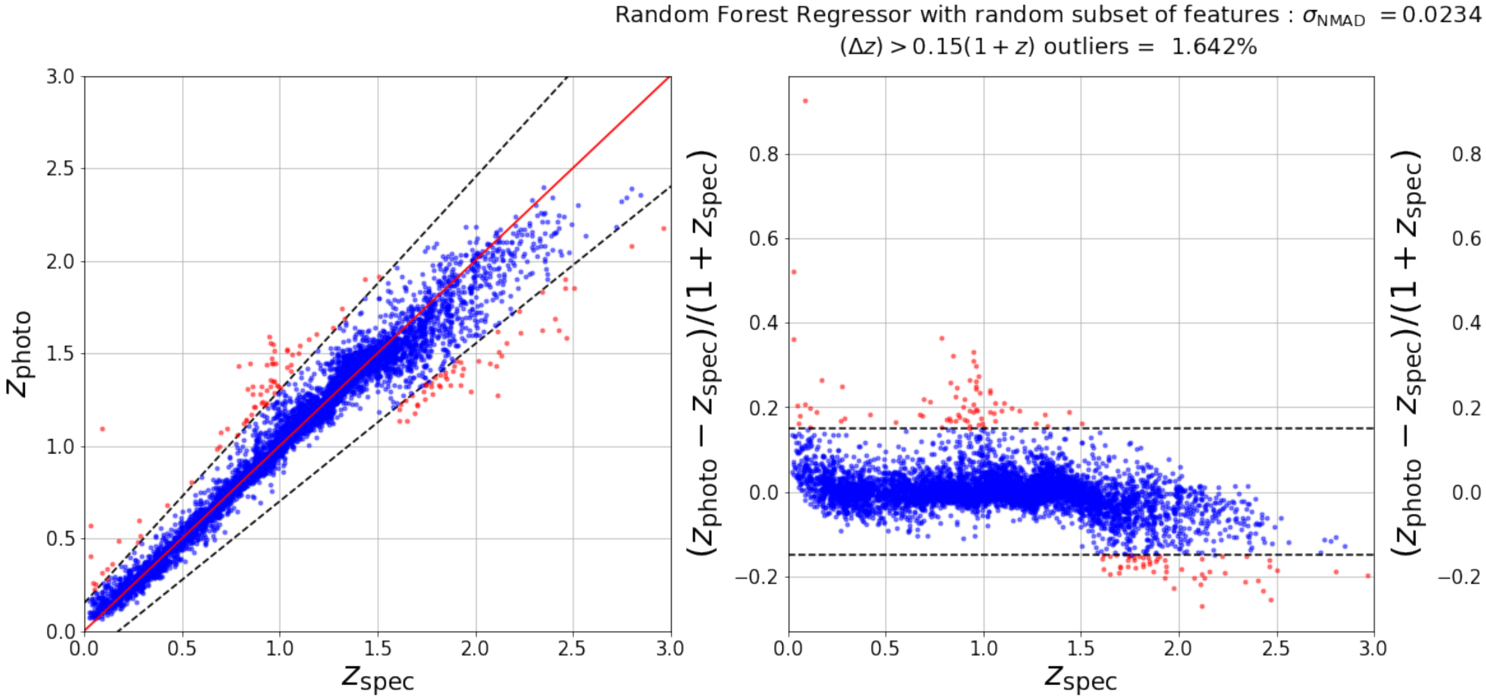
I. Introduction – context (cont.)

Delight results on DESC-DC2 simulated data :



I. Introduction – context (cont.)

Random forest results on DESC-DC2 simulated data :



I. Introduction – context (cont.)

Results on DESC simulated data :

	LEPHARE (TF)	Delight (TF/GP)	Random Forest (ML)
Std deviation	0.33	0.10	0.05
σ_{NMAD}	0.07	0.06	0.02
Outliers	11.7%	8.5%	1.6%
R2 score	0.31	0.72	0.92

- (Optimized) Random forest : best performances... in an ideal case
- LEPHARE : room for improvement, far from ML in an ideal scenario
- Delight : comparable to TF but does not require much knowledge of the population studied to be improved. Fewer outliers than TF.

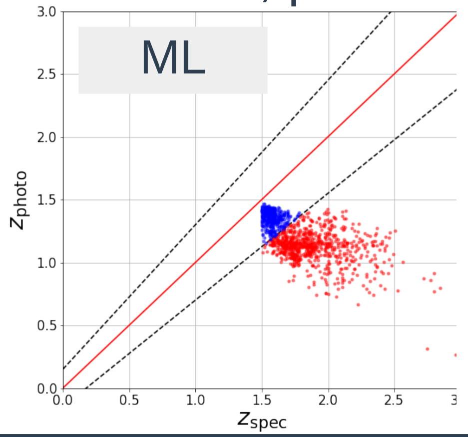
Internship work : looking for improvements

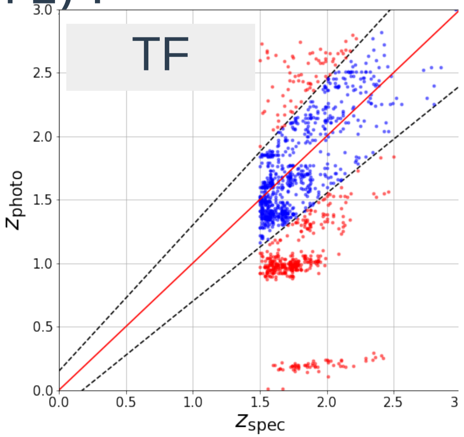
- More SED templates (CWW [5] vs. Brown [6])
- Improve LEPHARE++ settings
- Optimize Delight hyperparameters
- Pre-classification
- **Detection bias in models**

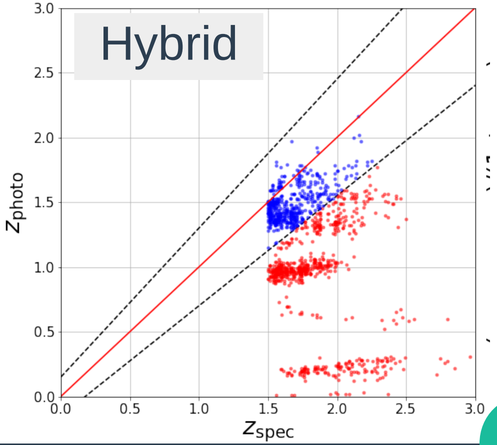
I. Introduction – context (cont.)

ii. Drawbacks

- Template fitting : optimized through good knowledge of the population studied. For instance in LEPHARE++ : extinction law, prior on Z, emission lines, magnitude limits, etc.
- Machine Learning requires appropriate training, *e.g.* for extrapolation (training at low z, prediction at high z) :



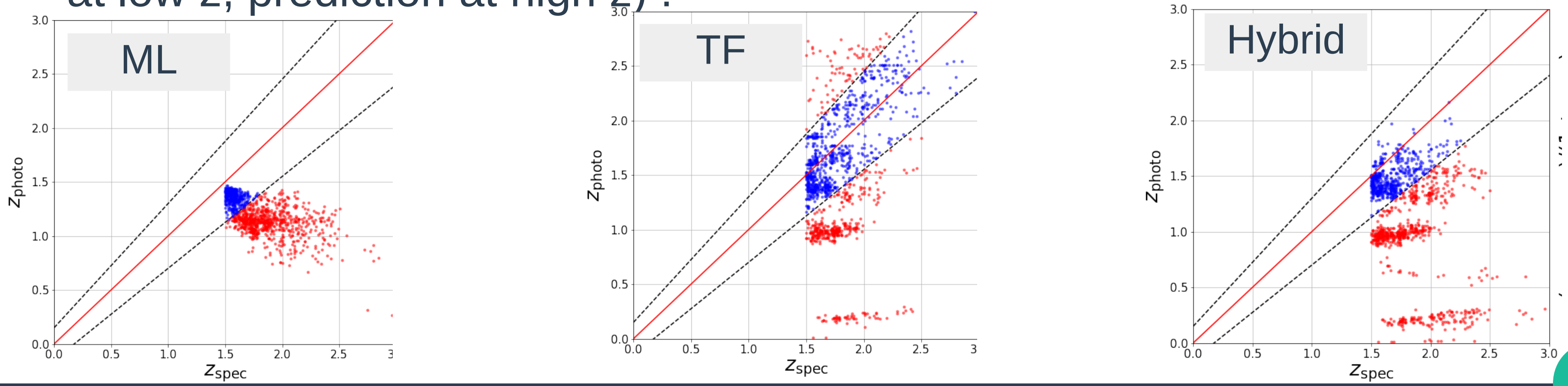




ii. Drawbacks

- Template fitting : optimized through good knowledge of the population studied. For instance in LEPHARE++ : extinction law, prior on Z, emission lines, magnitude limits, etc.

- Machine Learning requires appropriate training, *e.g.* for extrapolation (training at low z, prediction at high z) :



II. One improvement attempt

i. Model fitting (WIP)

- In LSST Science Book [4] : $\sigma^2_{\rm rand} = (0.04 - \gamma)x + \gamma\, x^2 \;\; ({\rm mag}^2),$ $x \equiv 10^{0.4(m - m_5)}$

→ first estimate : at m = m_5, σ = 0.2

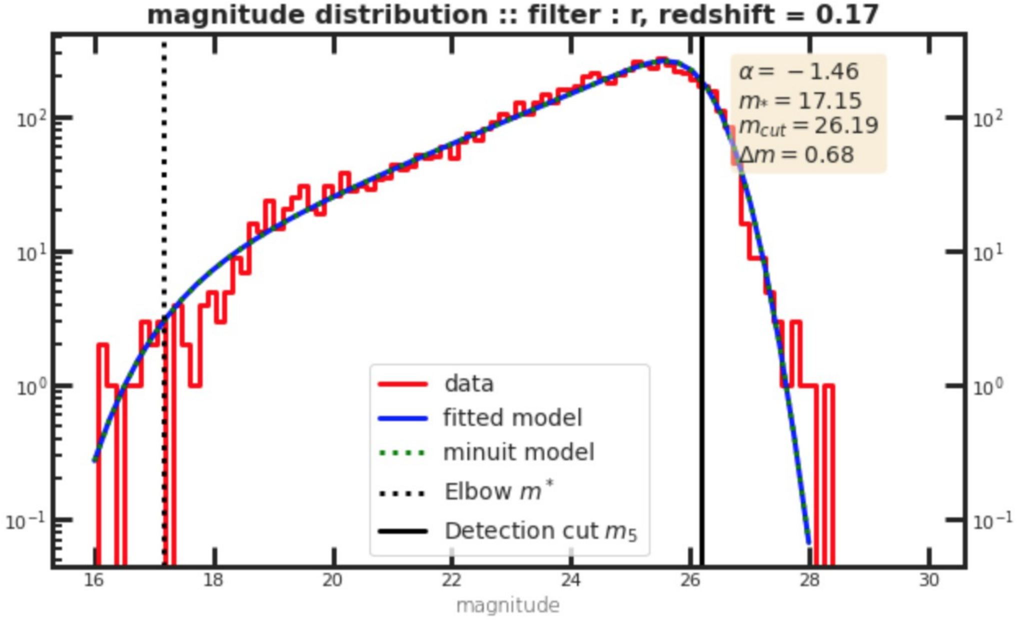
- Galaxies distribution model : Schechter equation [6], incomplete gamma function

$$\phi(L)\,\mathrm{d}L = \phi^*(L/L^*)^\alpha \; \exp(-L/L^*)\,\mathrm{d}(L/L^*) \qquad (1)$$

with the normalization parameter ϕ^*, or equivalently in magnitudes

$$\phi(M)\,\mathrm{d}M = (0.4\ln 10)\phi^* 10^{0.4(M^*-M)(1+\alpha)} \times \exp[-10^{0.4(M^*-M)}]\,\mathrm{d}M.$$

$$(1) \rightarrow N(L' \geq L_{min}) = \int_{L_{min}}^{\infty} \Phi(L)\,dL = \Phi_0\, \gamma\!\left(\alpha+1, \frac{L_{min}}{L_0}\right)$$



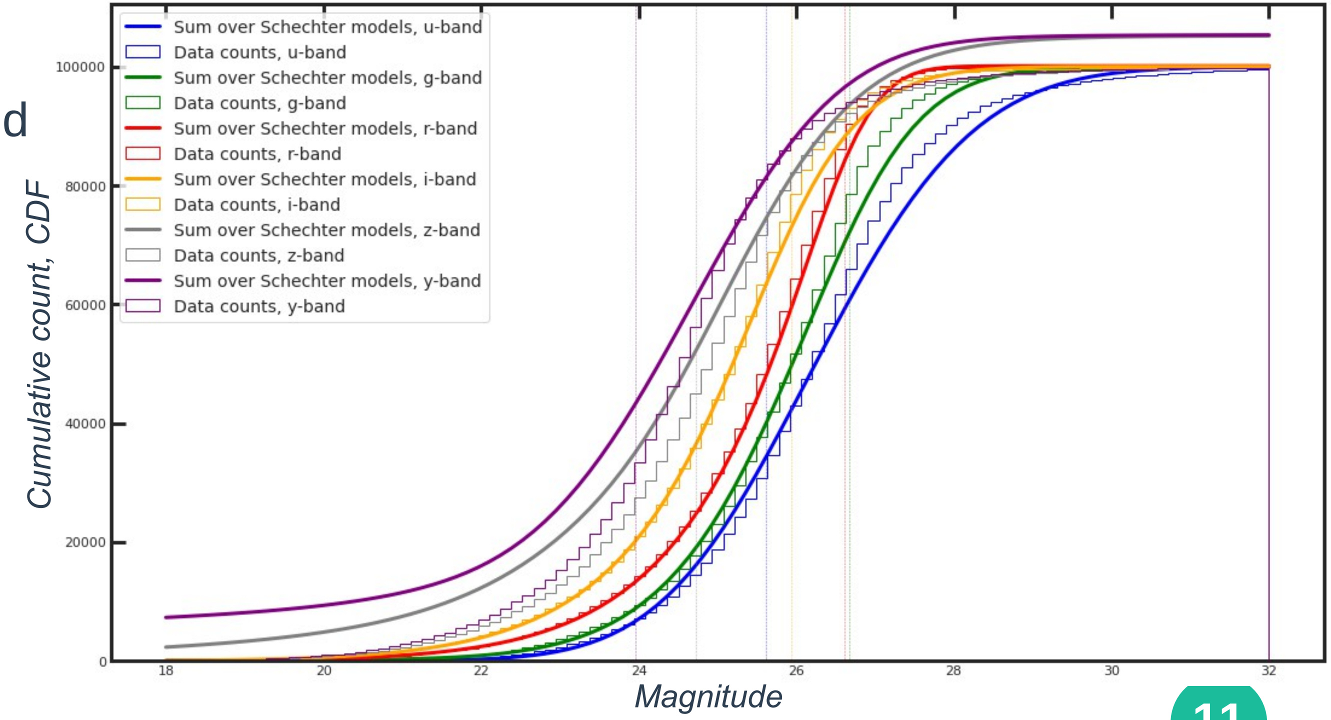
Not accounted for in templates

This leads to a simple model of detection bias relying on few parameters : Φ^*, L_{min}, L^*, α

II. One improvement attempt (cont.)

Use Minuit minimizer [8] to f nd the model parameters on DP0 data :

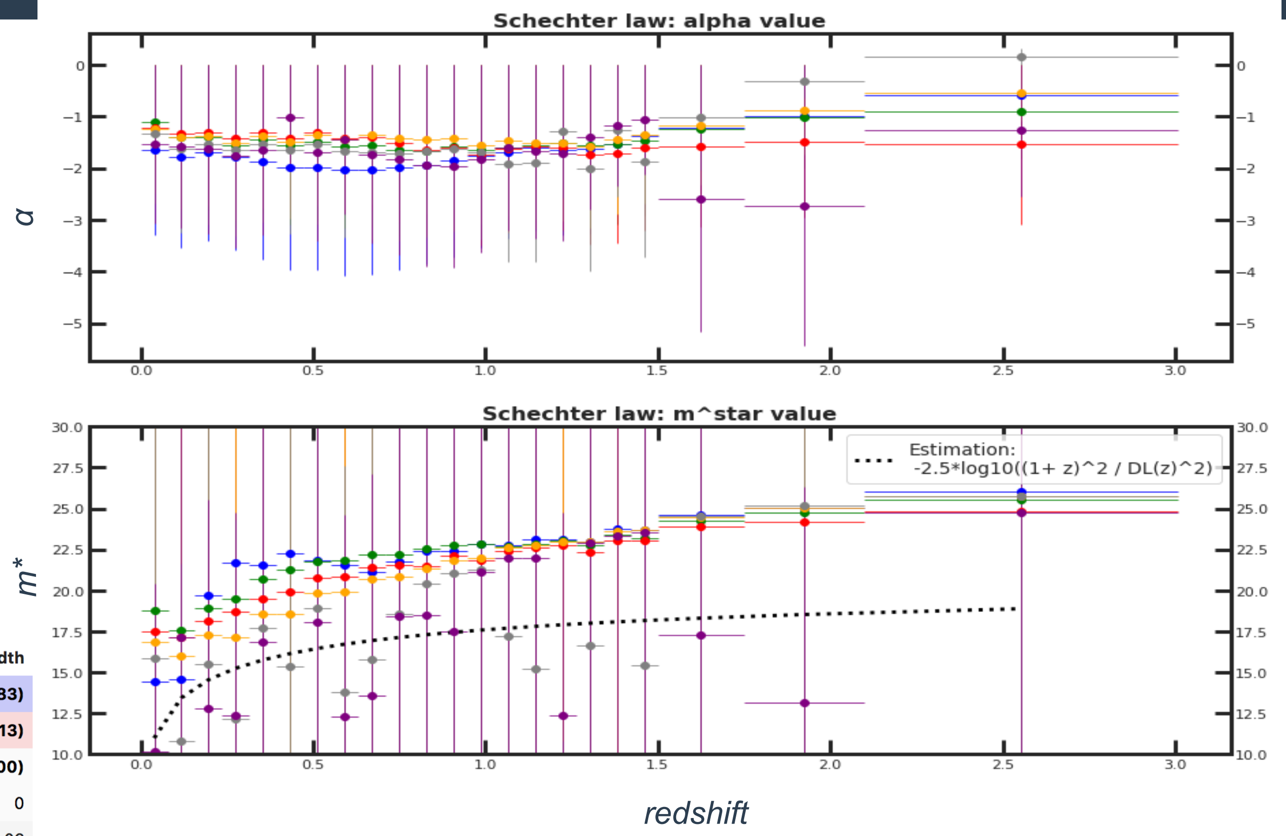
- Normalisation parameter Φ^* and reference magnitude m*
- α value
- Effective 5σ detection threshold m_5



II. One improvement attempt (cont.)

	Name	Value	Hesse Error	Fixed
0	phiStar	7.59e3	0.21e3	
1	alpha	-1.27	0.04	
2	m_star	24.757	0.027	
3	m5cut	23.95	0.24	yes
4	m5width	2.56	0.04	

	phiStar	**alpha**	**m_star**	**m5cut**	**m5width**
phiStar	5.93e+03	0.269 **(0.266)**	0.315 **(0.290)**	0	-4.64e+04 **(-0.383)**
alpha	0.269 **(0.266)**	0.000172	0.000147 **(0.794)**	0	4.39 **(0.213)**
m_star	0.315 **(0.290)**	0.000147 **(0.794)**	0.000199	0	0.00624 **(0.000)**
m5cut	0	0	0	0	0
m5width	-4.64e+04 **(-0.383)**	4.39 **(0.213)**	0.00624 **(0.000)**	0	2.47e+06



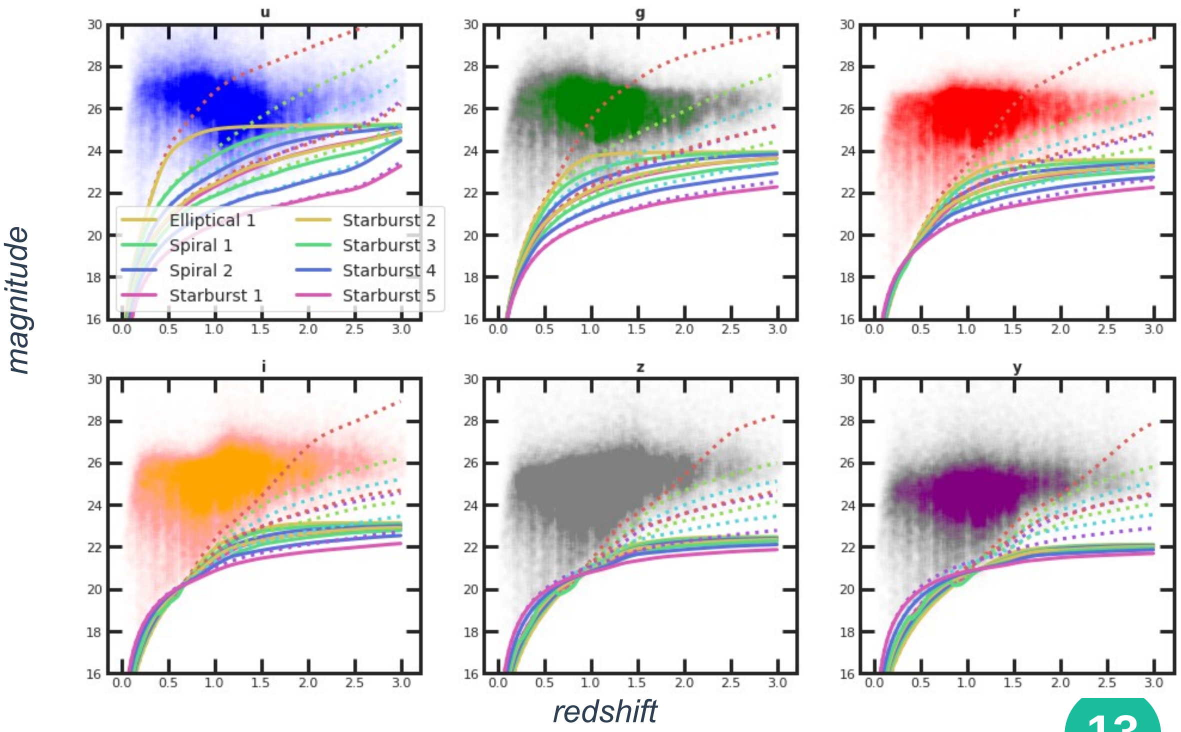
II. One improvement attempt (cont.)

ii. Detection bias in templates (TBD)

- Apply the bias to templates:
  - Issue for $\alpha \leqslant -1$
- Optimize the coverage of data points with templates :
  - Value : normalisation of flux-redshift model
  - Slope : detection bias

→ Automated pre-fit within the templates creation process to improve predictions at fainter magnitudes

Relative Biased Flux (AB-mag) - DP0-DC2 (No m5 cut) Redshift Model



III. Prospects

- Optimize the use of TF methods,
  - More & more representative SED templates
  - Use all the software's capabilities
- Improve templates generation : bias model, maybe other effects ?
- Build good, appropriate training data for ML
- Hybrid methods : TF+GP (Delight) with optimized hyperparameters, pre-classification, etc.

→ ***combine the reliability of TF and the performances of ML***

Thank you for your attention

References

[1] LEPHARE++ : Arnoults S. *et al.*, 1999, MNRAS, 310, 450 ; Ilbert O. *et al.,* 2006, A&A, 457, 841

[2] Delight : Data-driven, Interpretable Photometric Redshifts Trained on Heterogeneous and Unrepresentative Data, Boris Leistedt B. and Hogg D. W., 2017 *ApJ* 838 5

[3] Scikit-learn : Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[4] LSST Science Book : LSST Science Collaborations and LSST Project 2009, LSST Science Book,Version 2.0, arXiv:0912.0201

[5] CWW, KIN SEDs : Kinney, A. L., Calzetti, D., Bohlin, R. C., McQuade, K., Storchi-Bergmann, T., & Schmitt, H. R.1996, ApJ, 467, 38

Calzetti, D., Kinney, A. L., & Storchi-Bergmann, T. 1994, ApJ, 429, 582

[6] Brown SEDs : Brown, M. J. I., et al. 2014, ApJS, 212, 18

[7] Schechter luminosity function : Luminosity Function of Galaxies, Jerjen H., 2006, Encyclopedia of Astronomy & Astrophysics (P. Murdin)

[8] Minuit : Dembinski H., *et al.*, odidev. (2022). scikit-hep/iminuit: v2.11.2 (v2.11.2). Zenodo. https://doi.org/10.5281/zenodo.6389982

I. Introduction – context (cont.)

LSST requirements [4] :

Photometric redshifts for LSST will be applied and calibrated over the redshift range $0 < z < 4$ for galaxies to r › 27.5. For the majority of science cases, such as weak lensing and BAO, a subset of galaxies with $i < 25.3$ will be used. For this high S/N gold standard subset over the redshift interval, $0 < z < 3$, the photometric redshift requirements are:

- The root-mean-square scatter in photometric redshifts, $\sigma_z/(1+z)$, must be smaller than 0.05, with a goal of 0.02.
- The fraction of 3σ outliers at all redshifts must be below 10%.
- The bias in ez = (zphoto −zspec)/(1+zspec) must be below 0.003 (or 0.01 for combined analyses of weak lensing and baryon acoustic oscillations); the uncertainty in $\sigma_z/(1+z)$ must also be known to similar accuracy.