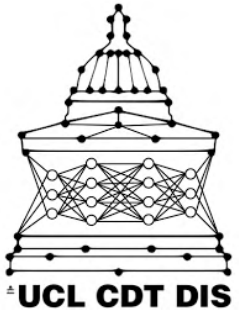


Joint redshift-stellar mass PDFs with Random Forest

Sunil Mucesh

Supervisors: Prof. Ofer Lahav & Dr. Will Hartley

Collaborators: Dr. Antonella Palmese & Dr. Lorne Whiteway

Motivation

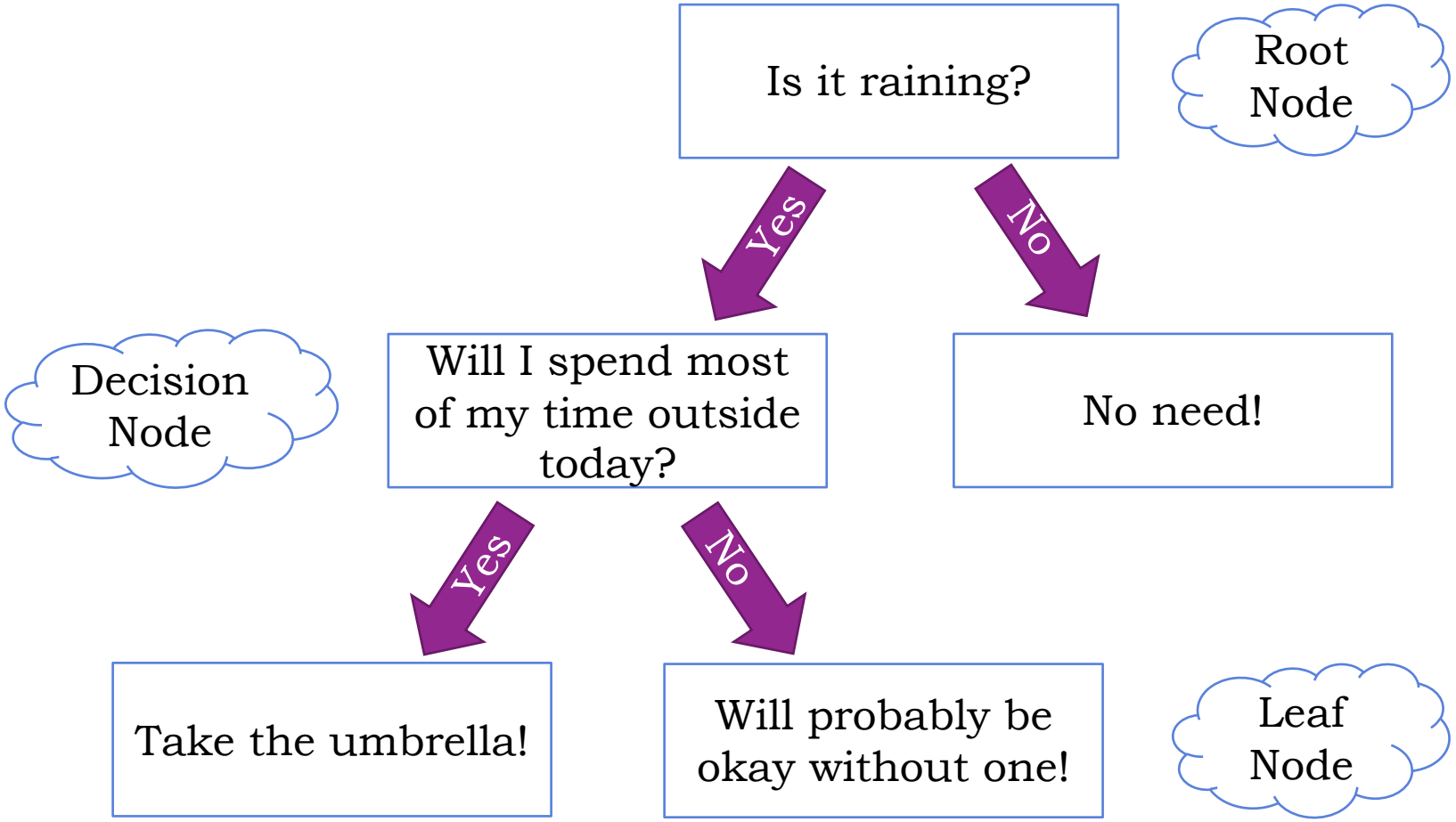
- Point estimates of galaxy properties determined with few photometric bands are imprecise.
- We require PDFs to fully characterise the uncertainty in the estimates.
- Much of the focus has been on generating redshift PDFs. Using redshift PDFs instead of point estimates has been shown to improve the accuracy of cosmological measurements.
- Statistically, a galaxy can be described by a multivariate PDF of redshift and physical properties.
- A new class of SED/template-fitting codes (BAGPIPES, BEAGLE, BAYESED etc…) use a Bayesian approach to derive these multivariate PDFs.
- However, they are not efficient for generating PDFs for a large number of galaxies.
- We use a ML-based approach to solve this problem. In particular, we focused on joint redshift - stellar mass pdfs.

Random Forest: Introduction

- Random forest is an ensemble supervised machine learning algorithm based on decision trees.
- Easy to implement and understand (i.e. not a black box).
- Generalises well, resistant to over-fitting.
- Can be used for regression and classification tasks.
- It has previously been used to predict redshifts, stellar masses and star formation rates of galaxies.

Credit: Flaticon

A simple decision tree: should I take my umbrella with me today?

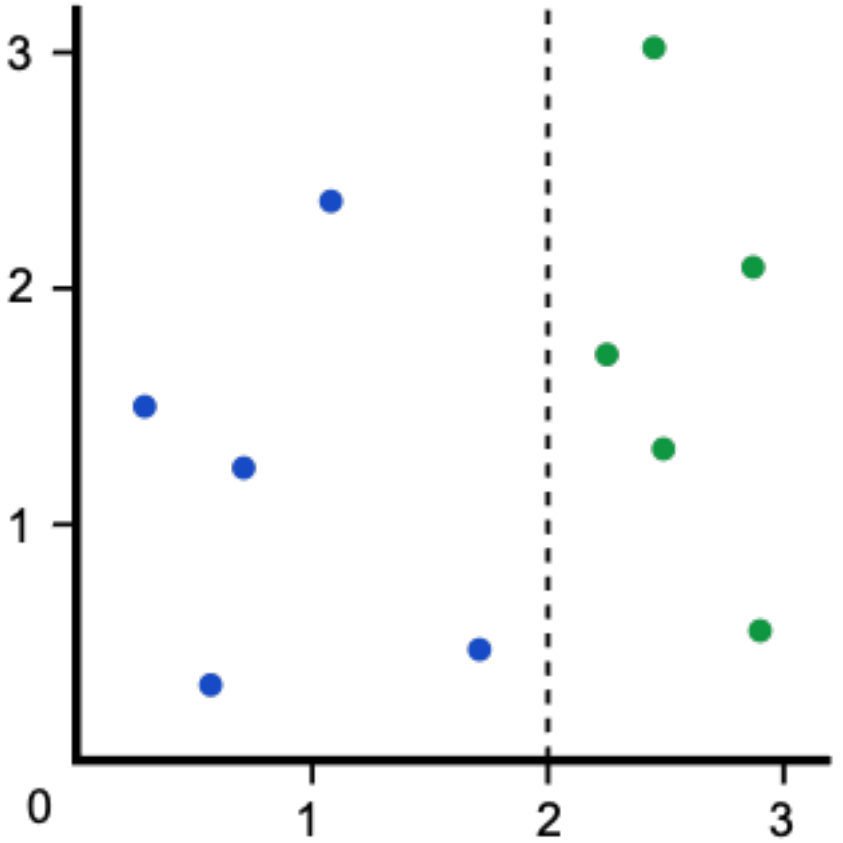


How to build a decision tree from data?

- The goal is to cluster or group data with similar properties.
- We need a loss function, and generally for regression trees the variance is used.

$$S = \frac{1}{n_m} \sum_{m} \sum_{iem} (\tilde{y}_i - \overline{y_m})^2$$

- Choose a feature and location which minimises the variance.
- Repeat the process until some threshold.
- The decision tree can now be used to predict for new data. For classification the prediction is a class and for regression the outcome is a mean value.



Credit: https://victorzhou.com/blog/intro-to-random-forests/

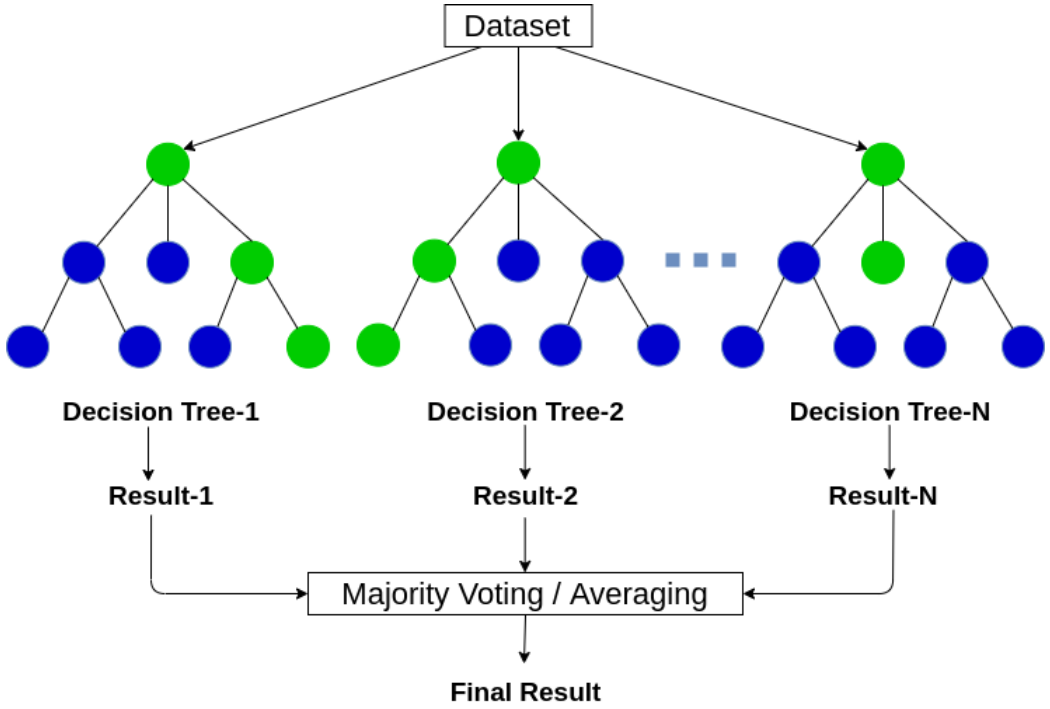
Random Forest: Algorithm

A random forest consists of many decision trees with a few tweaks.

1. Sample randomly from data with replacement.
2. Choose only a subset of input features.
3. Create a decision tree from the bootstrapped sample and the chosen features.
4. Repeat.

To make a prediction:

- Classification - Majority vote
- Regression - Average



Credit: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

Random Forest: Method

- Galaxies cluster together in n-dimensional space if they have similar values of input features (e.g. colours).
- RF aims to find these clusters by minimising a loss function (based on the variance), with redshift and stellar mass as the target variables.
- These clusters end up in the leaf nodes of the decision trees. The leaf nodes contain redshifts and stellar masses of similar galaxies.
- Once the random forest has been trained, we pass a 'new' galaxy down all the decision trees and it should end up in leaf nodes that are representative.
- To extract point estimates, we average the redshift and stellar mass values of training galaxies in the leaf nodes.
- To build marginal PDFs, we separately aggregate all the redshift and stellar mass values in the leaf nodes in all the decision trees.
- We combine the aggregated values to build joint PDFs.

Data: DES Y3 Deep Fields & COSMOS

- We use three different datasets: DES Y3 Deep Fields (DF), DES Y3 Balrog & COSMOS2015.
- Y3 DF catalogue contains deep and precise *griz* photometry of more than 1.7 million objects.
- We combine this with the COSMOS2015 catalogue, which has accurate redshifts and stellar masses to produce a 'baseline' DF dataset.
- However, our target is the main wide-field (WF) DES.
- We cannot use the DF dataset to train a RF model as the photometric errors present in the DF would not reflect those in the WF.
- Secondly, the COSMOS field does not overlap with the main survey area. The redshifts and stellar masses estimated using 4-band WF data would be imprecise, compared to those in the COSMO2015.
- In essence, we require a catalogue of DF galaxies which emulate galaxies in the WF.
- This leads us to Balrog.

Data: DES Y3 Balrog & COSMOS

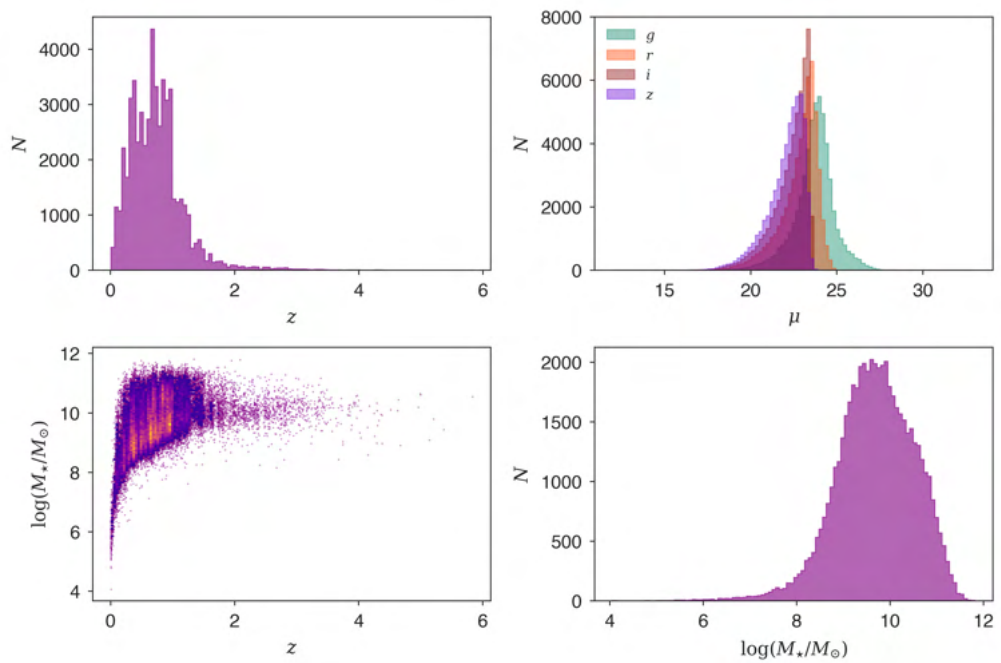
- Balrog is a Python package for measuring the transfer function of imaging surveys.
- We use the DES Y3 Balrog catalogue.
- Model fits of galaxies are drawn randomly from the Y3 DF catalogue and injected into DES-Y3 single-epoch images.
- The DES measurement pipeline is rerun on these injected images to produce the DES Y3 Balrog catalogue.
- The resulting catalogue is a Monte Carlo sampling of the DES transfer function and contains true and measured *griz* photometry.
- This catalogue provides us with ready-made emulated galaxies in our target wide-field dataset (DES Y3 Gold).
- We combine this catalogue with the COSMO2015 to produce our 'WF' dataset.

Pre-processing

- We apply some simple cuts to produce the DF and WF datasets.

1. $i < 23.5$
2. $0 < z < 9.99$ – to discard any galaxies with erroneous redshifts and stellar masses.
3. MEAS_CM_FLAG = 0 – to remove any galaxies with erroneous flux measurements.

- We convert fluxes into 'asinh' magnitudes or 'luptitudes' to avoid removing any galaxies with close to zero or negative fluxes.
- Finally, we perform an 80:20 split on each dataset for training and testing.
- DF and WF training : 42,792 & 314,196
- DF and WF testing: 10,699

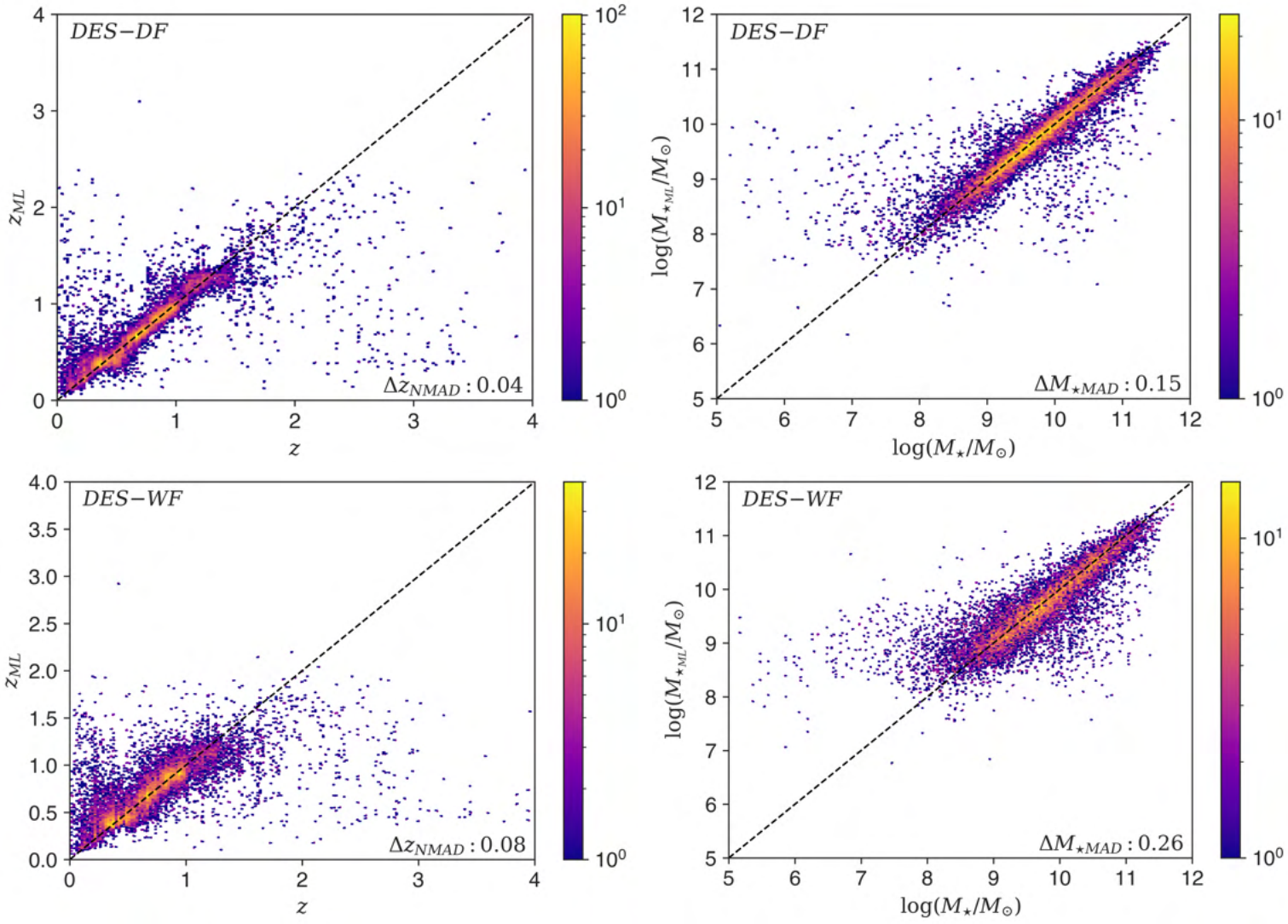


RF Models

- We build two RF models: DES-DF and DES-WF.
- DES-DF trained using the DF dataset and it allows us to establish the baseline performance.
- DES-WF trained using the WF dataset to produce joint PDFs for galaxies in our target dataset.
- We predict redshift and stellar mass simultaneously (multivariate target regression), with the following input features:

1. *griz luptitudes*
2. *g-r, r-i, i-z* lupticolours
3. luptitudes + lupticolour errors

Results: Point Estimates



DES-DF performing better than DES-WF. This is expected.

Taking into account the degraded photometry, DES-WF still performing well as most data points lie close to the diagonal.

Outliers at low and high redshift due to a lack of training data + degeneracies.

Validation: Marginal PDFs

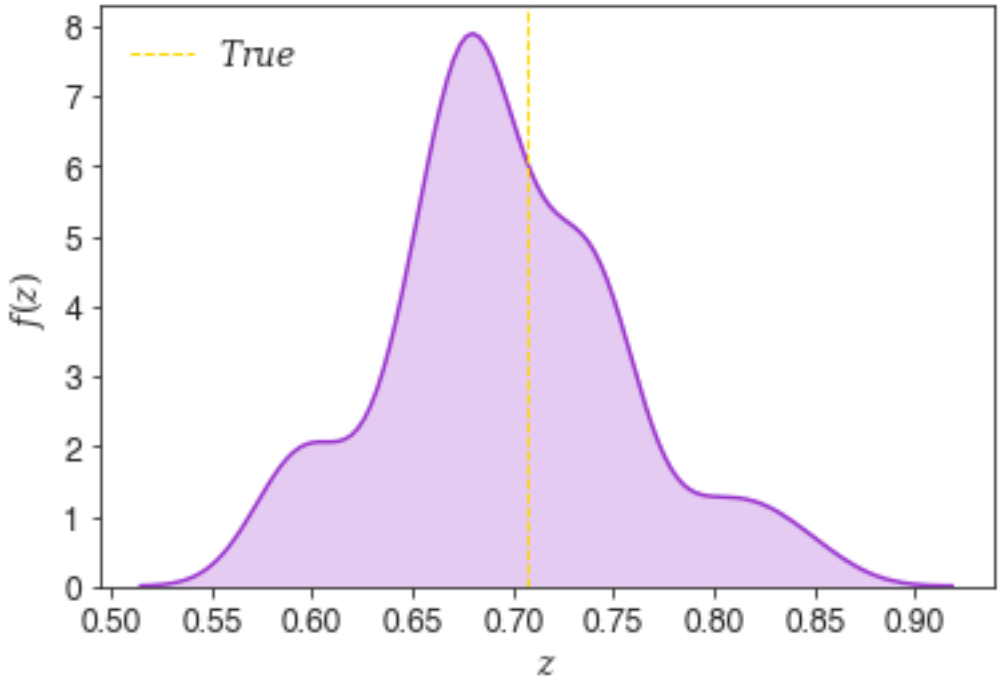
- Unlike point estimates, the ‘true’ PDFs are not available for comparison.
- We cannot validate individual PDFs, but we can determine their overall validity.
- We use two different modes of calibration: probabilistic and marginal calibration.

Probabilistic calibration

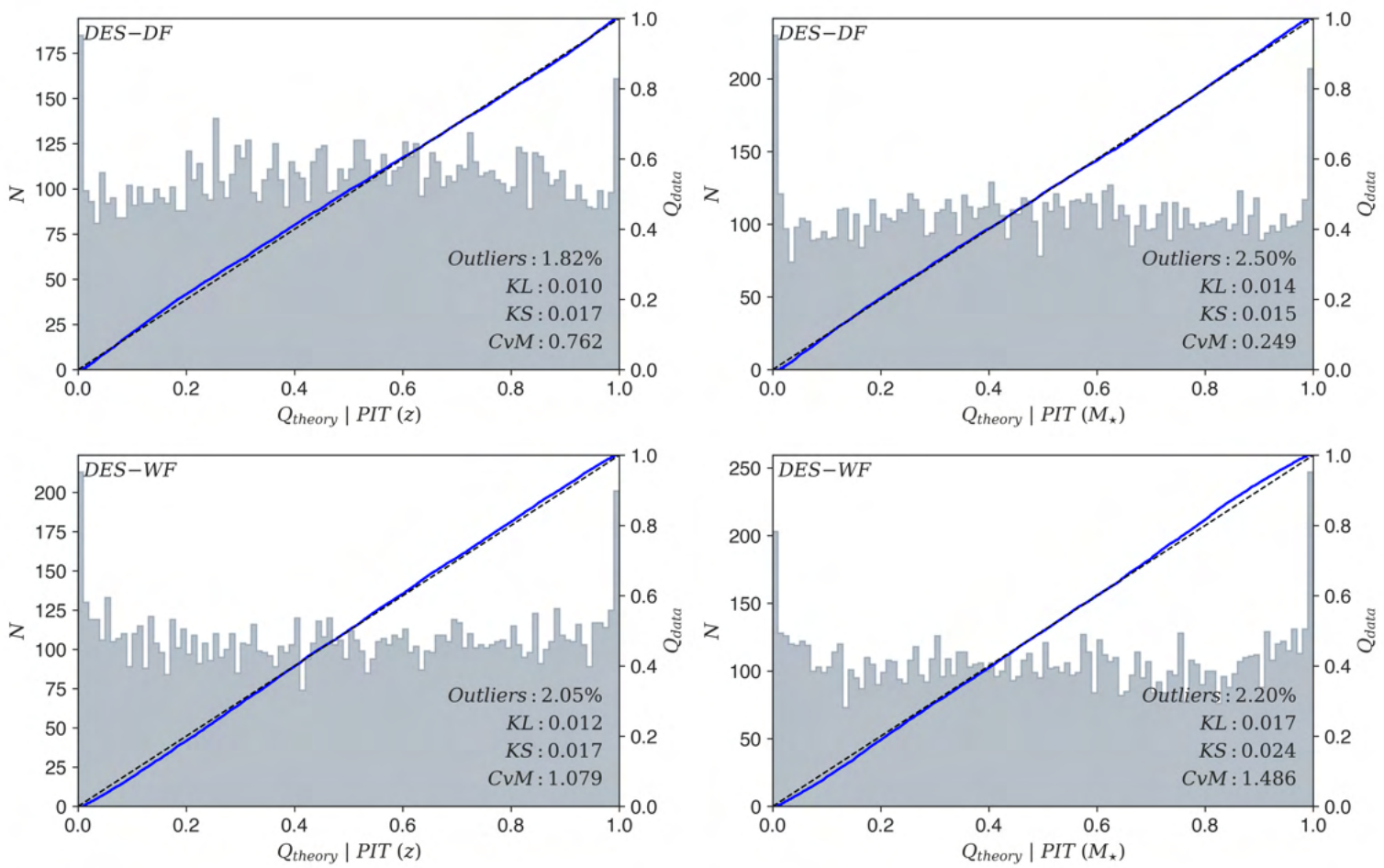
- True values of redshift and stellar mass should be random draws from their respective marginal PDFs.
- We can determine probabilistic calibration using the probability integral transform (PIT).

$$PIT = \int_{-\infty}^{\tilde{y}} f(y)\, dy$$

- If the values are random draws, then the PIT will be a random number between 0 and 1.
- As a result, for an ensemble of galaxies, the distribution of PIT values should follow the standard uniform distribution.



Probabilistic calibration



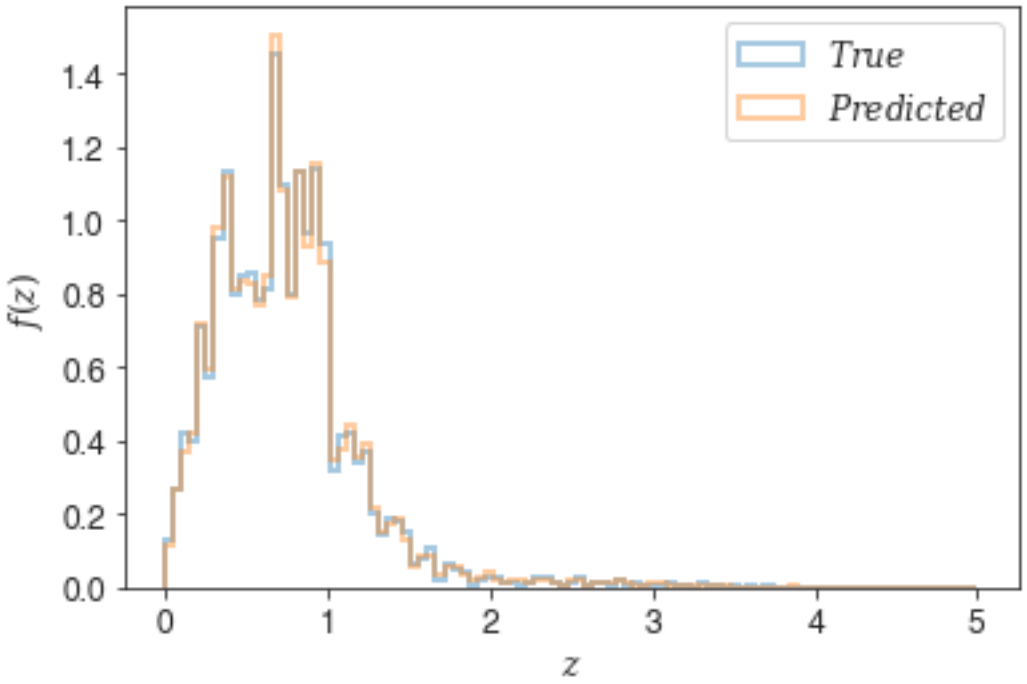
Uniform PIT histograms, catastrophic outliers approx. 2%.

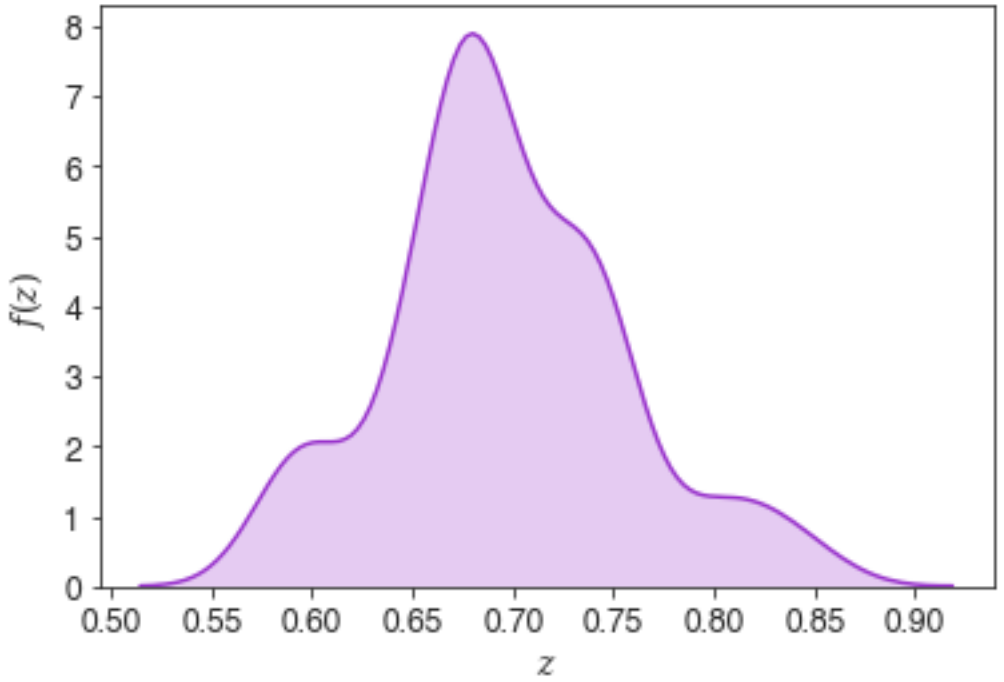
DES-DF performing marginally better than DES-WF in terms of probabilistic calibration.

Marginal calibration

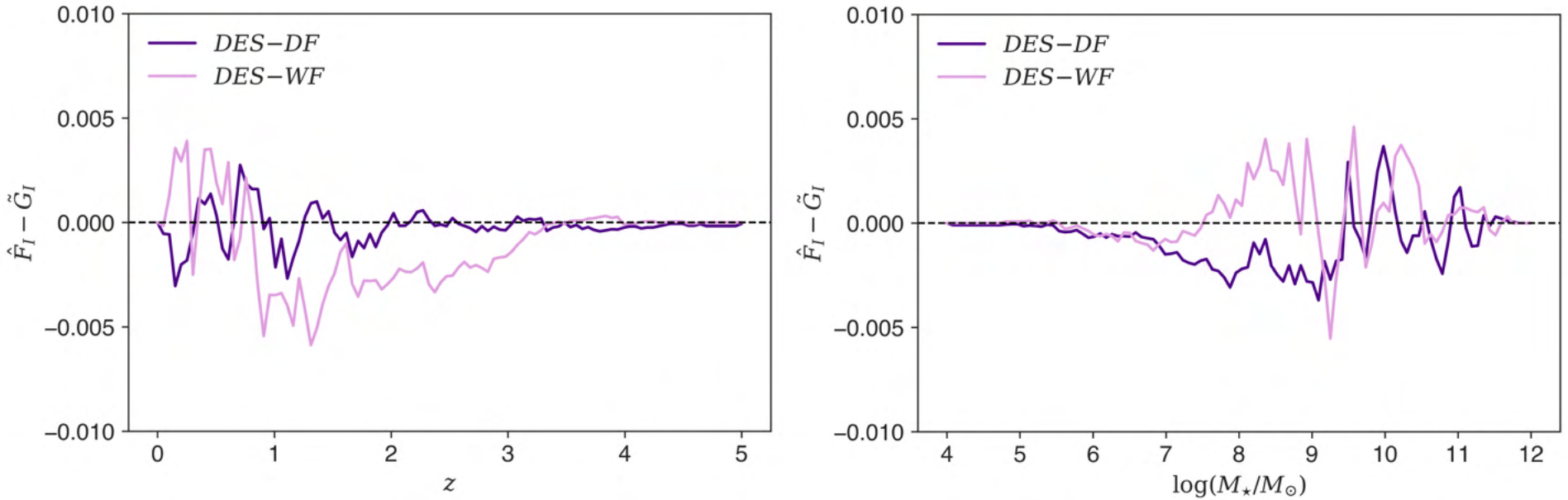
- Average predictive CDF should match the 'true' empirical CDF.

$$\widehat{F}_I(y) = \frac{1}{n}\sum_{i=1}^{n} F_i(y) \qquad \widetilde{G}_I(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{\widetilde{y_i} \leq y\}$$



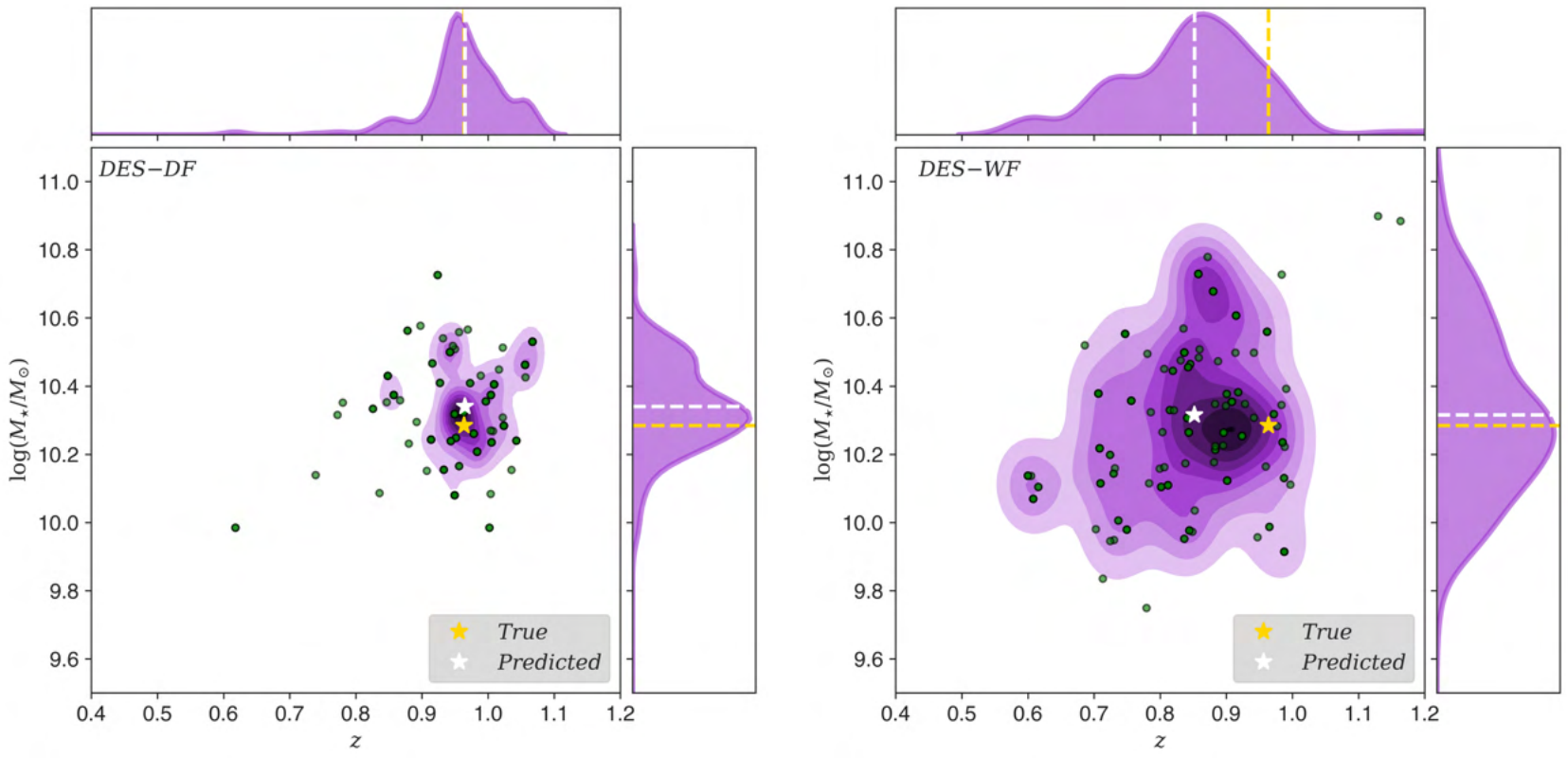


Marginal calibration



Small fluctuations about the zero line indicate DES-DF and DES-WF are performing well.

Joint PDFs



Joint PDFs of the same test galaxy occupy similar regions of the redshift-stellar mass space.

DES-DF produces more compact PDFs, reflecting the precise photometry used.

Validation: Joint PDFs

- The methods we have used so far cannot be used to validate multivariate PDFs.
- For example, the PIT distribution is no longer uniform.
- We use the multivariate extensions of probabilistic and marginal calibration to validate our joint PDFs. These are probabilistic copula calibration and Kendall calibration.
- These modes of calibration can be interpreted in the same manner as their univariate counterparts.

Probabilistic copula calibration

- Probabilistic copula calibration can be assessed by using the copula probability integral transform (copPIT):

$$copPIT = \mathcal{K}_H\big(H(\tilde{y})\big)$$

- The Kendall distribution function is defined as:

$$\mathcal{K}_H\big(H(\tilde{y})\big) = P(H(y) \leq H(\tilde{y}))$$

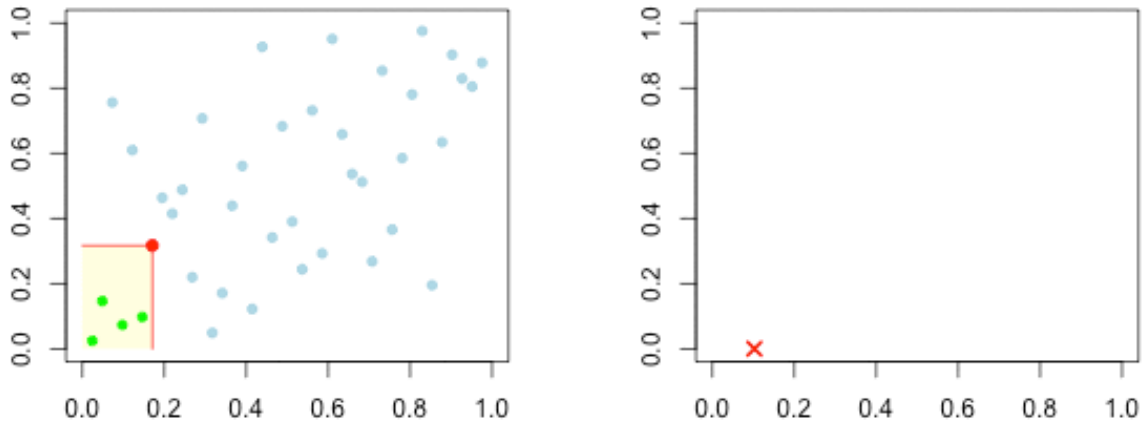
Probabilistic copula calibration

- Multivariate analogue of the PIT distribution.

1. Evaluate predicted joint CDF H(y) at each point prediction.
2. Evaluate the CDF at the 'true' redshift and stellar mass.
3. Compute copPIT.

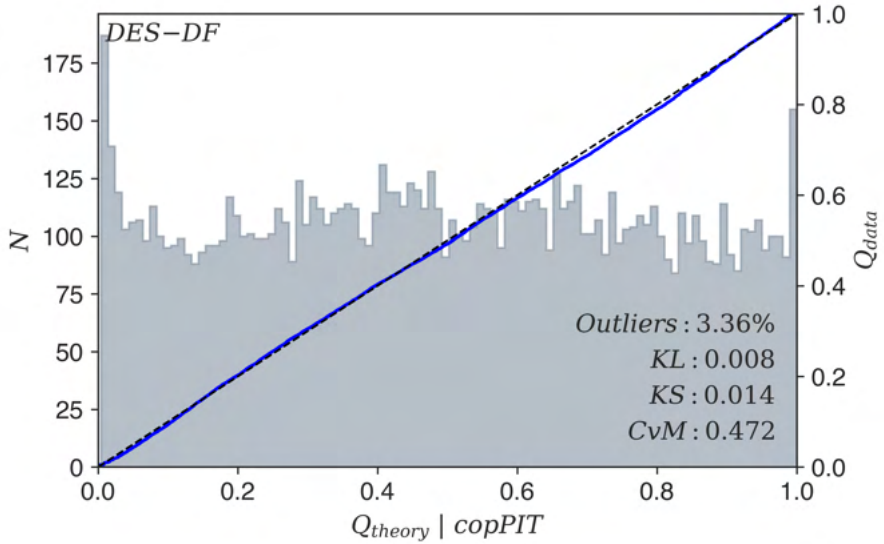
$$copPIT = P(H(y) \leq H(\tilde{y}))$$

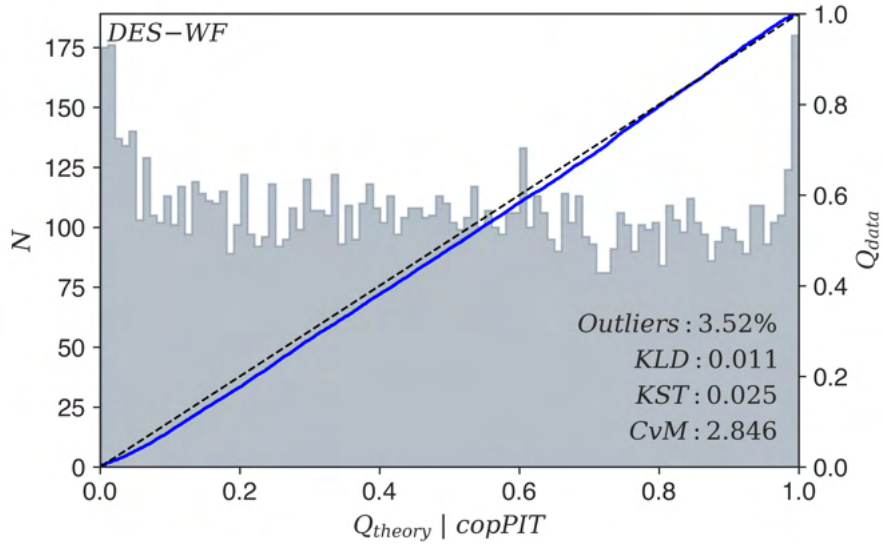
- copPIT distribution uniform if joint PDFs are copula calibrated.



Credit: https://www.freakonometrics.hypotheses.org/1126

Probabilistic copula calibration





Uniform copPIT histogram, with DES-DF performing slightly better than DES-WF.

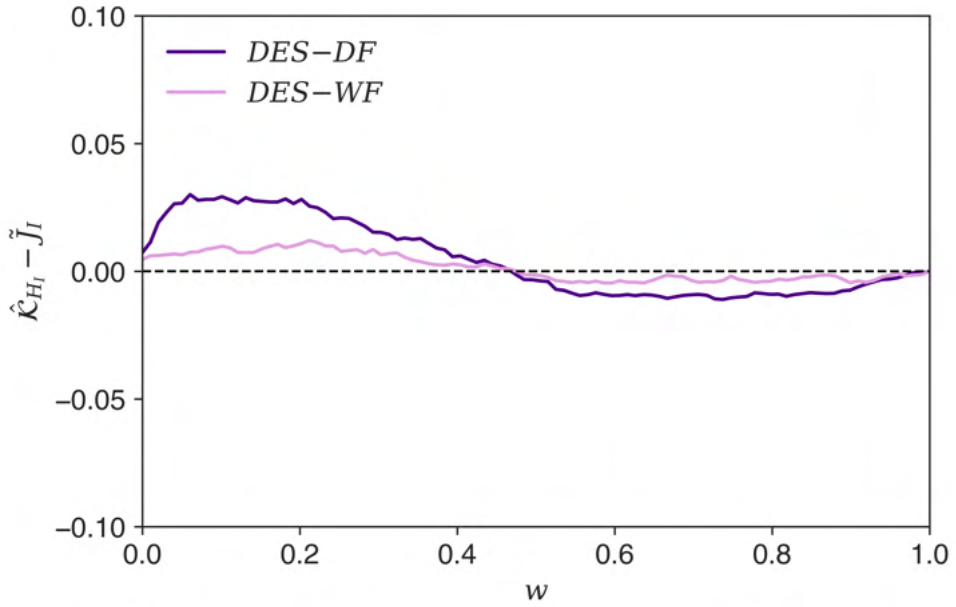
Kendall calibration

- Kendall calibration can be assessed by comparing the 'average Kendall distribution function', $\widehat{\mathcal{K}_{H_I}}$, to the empirical CDF of the predicted joint CDFs evaluated at the true redshifts and stellar mass, $\widetilde{J_I}$:

$$\widehat{\mathcal{K}_{H_I}}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{H_i}(w) \qquad \widetilde{J_I}(w) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{H_i(\widetilde{y_i}) \leq w\}$$

- Kendall calibration probes how well the dependence structure between redshift and stellar mass is predicted on average.

Kendall calibration

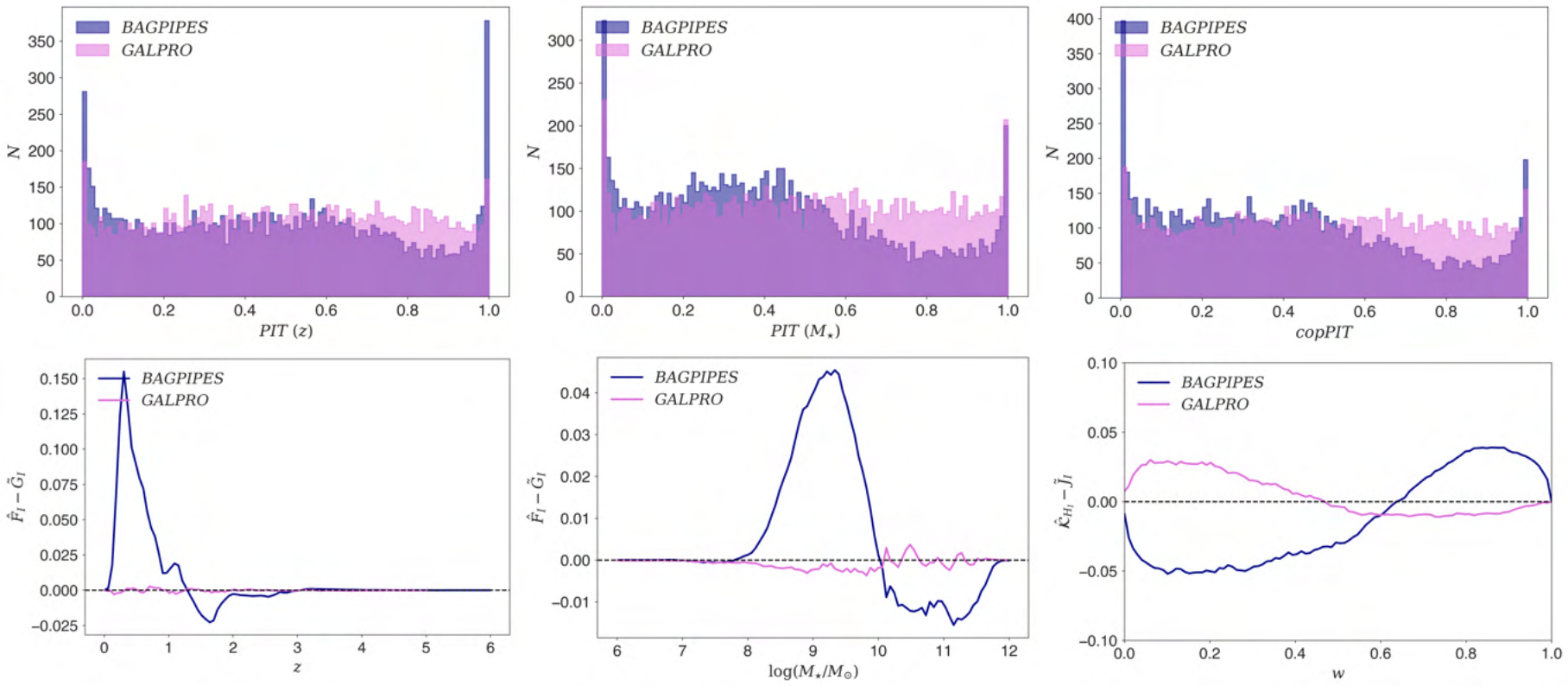


DES-WF performing better than DES-DF in this metric due to incorporation of photometric errors as it is trained on multiple scattered copies of DF galaxies.

Comparison: Template-fitting

- The diagnostic plots and the metrics we have utilised are difficult to fully appreciate without familiar context.
- As a result, we compare our results against those achieved by the SED-fitting method Bayesian Analysis of Galaxies for Physical Inference and Parameter Estimation, or BAGPIPES.
- It generates complex model galaxy spectra and fits these models to spectroscopic and/or photometric data to infer galaxy properties.
- BAGPIPES uses MultiNest nested sampling algorithm to generate multivariate posterior PDFs of redshift and physical properties of galaxies.
- We run BAGPIPES on test galaxies in the DF dataset, inputting photometry in Subaru V, r, i+ and z++ bands.
- To validate the PDFs, we run BAGPIPES again, but this time with 22 COSMOS bands (including the four mentioned above).

Comparison: Results



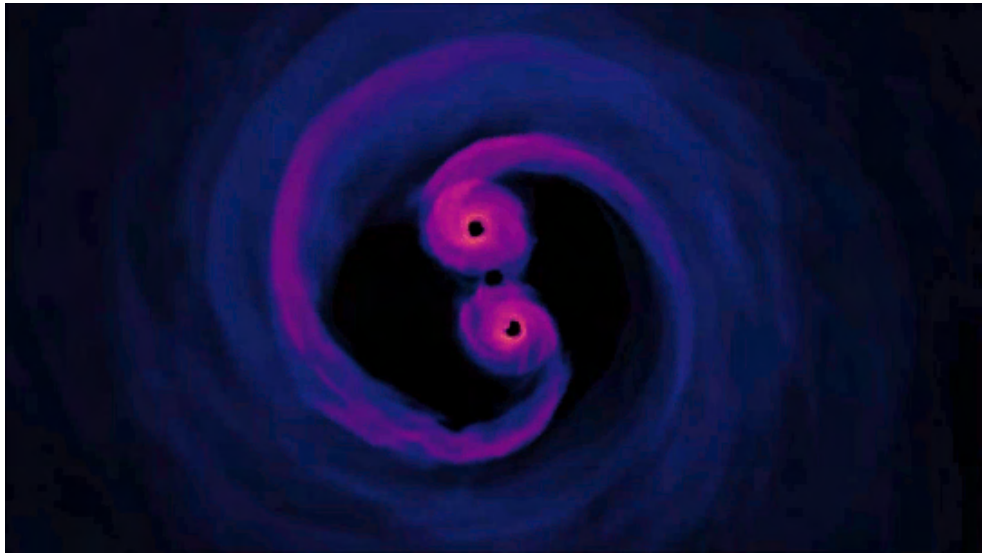
Our ML-based method performs better than BAGPIPES in all the metrics we have considered in our analysis. The redshift PIT distribution for BAGPIPES is still competitive with other template-fitting codes.

GALPRO

- GALPRO is a highly intuitive and efficient Python package based on the random forest algorithm to generate multivariate PDFs of galaxy properties on-the-fly.
- Code: https://github.com/smucesh/galpro
- Documentation: https://galpro.readthedocs.io/en/latest/

Applications

- Joint redshift-stellar mass PDFs have many applications.
- For example, to study the evolution of the stellar mass function.
- An interesting application of GALPRO could be to generate joint redshift-luminosity PDFs for measurement of the Hubble constant from dark standard sirens.
- Using full redshift PDFs has been shown to improve measurements, and joint redshift-luminosity PDFs allows one to define the selection function of a galaxy sample.



Credit: https://www.nasa.gov/feature/goddard/2018/new-simulation-sheds-light-on-spiraling-supermassive-black-holes

Summary & Outlook

- Successfully extracted point estimates, marginal and joint PDFs of redshift and stellar mass using the random forest algorithm.
- Performed validation checks for both the marginal and joints PDFs using different metrics.
- Compared our results to those achieved by the template-fitting code BAGPIPES.
- We find that our method is producing highly accurate joint PDFs, with only small calibration errors.
- We have developed GALPRO, a Python package which can generate n-dimensional PDFs on-the-fly, thus removing the problem of storage (https://galpro.readthedocs.io/en/latest/).
- In terms of speed, GALPRO is extremely fast, potentially able to generate joint PDFs for a million galaxies in just under 6 minutes with consumer computer hardware.
- Future applications to LSST and Euclid data.