Photometric Redshift Estimation with Convolutional Neural Networks and Galaxy Images: A Case Study of Resolving Biases in Data-Driven Methods

Qiufan Lin[1]

Advisors: Dr. Dominique Fouchez[1], Dr. Jérôme Pasquet[2]

[1]Aix Marseille Univ., CNRS/IN2P3, CPPM, Marseille, France

[2]UMR TETIS, Univ. Montpellier; AgroParisTech, Cirad, CNRS, Irstea, Montpellier, France

Spectroscopic redshift (spec-z) v.s. Photometric redshift (photo-z)

- **Spec-z by spectroscopy**

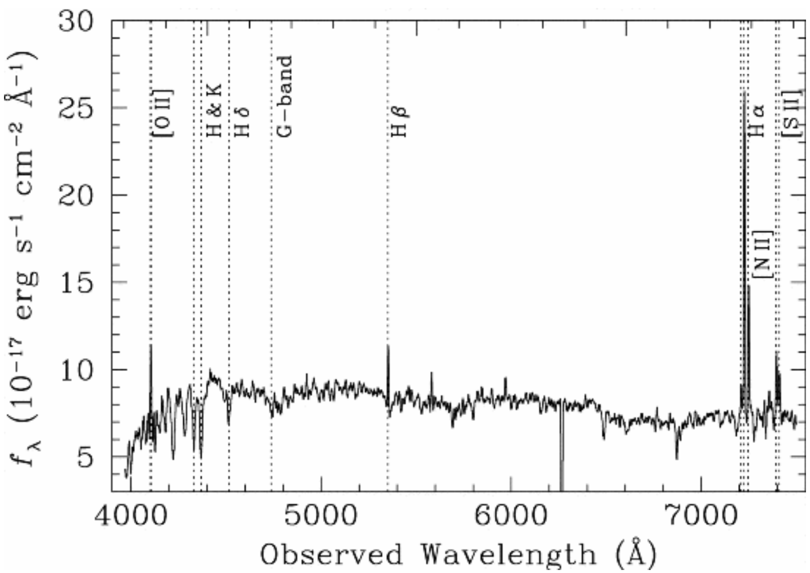


Image from: http://spiff.rit.edu/classes/phys301/lectures/doppler/doppler.html

- **Photo-z by broad-band photometry**

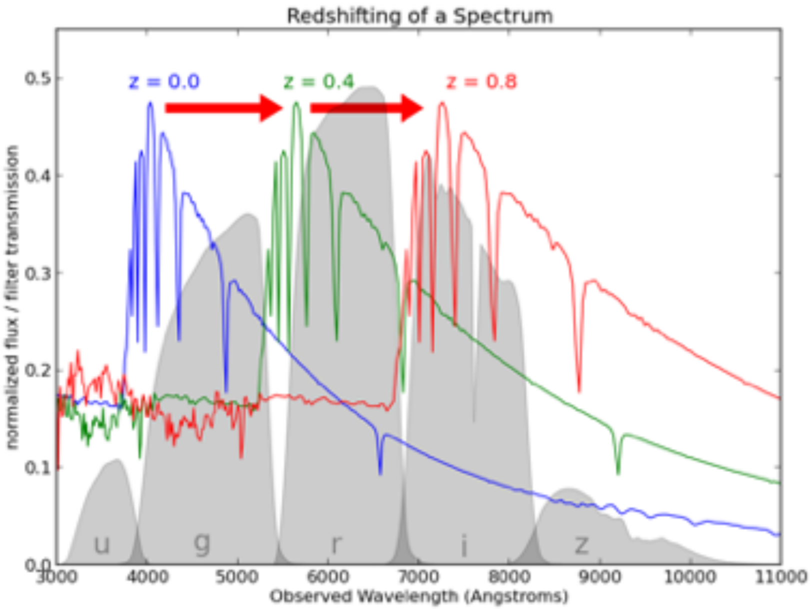
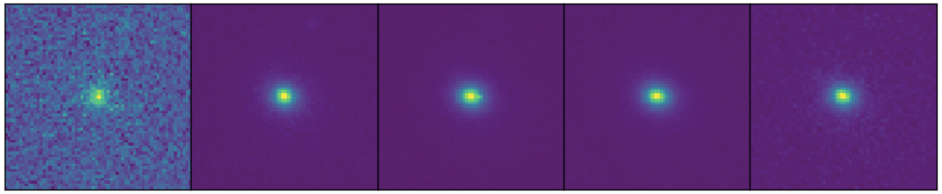


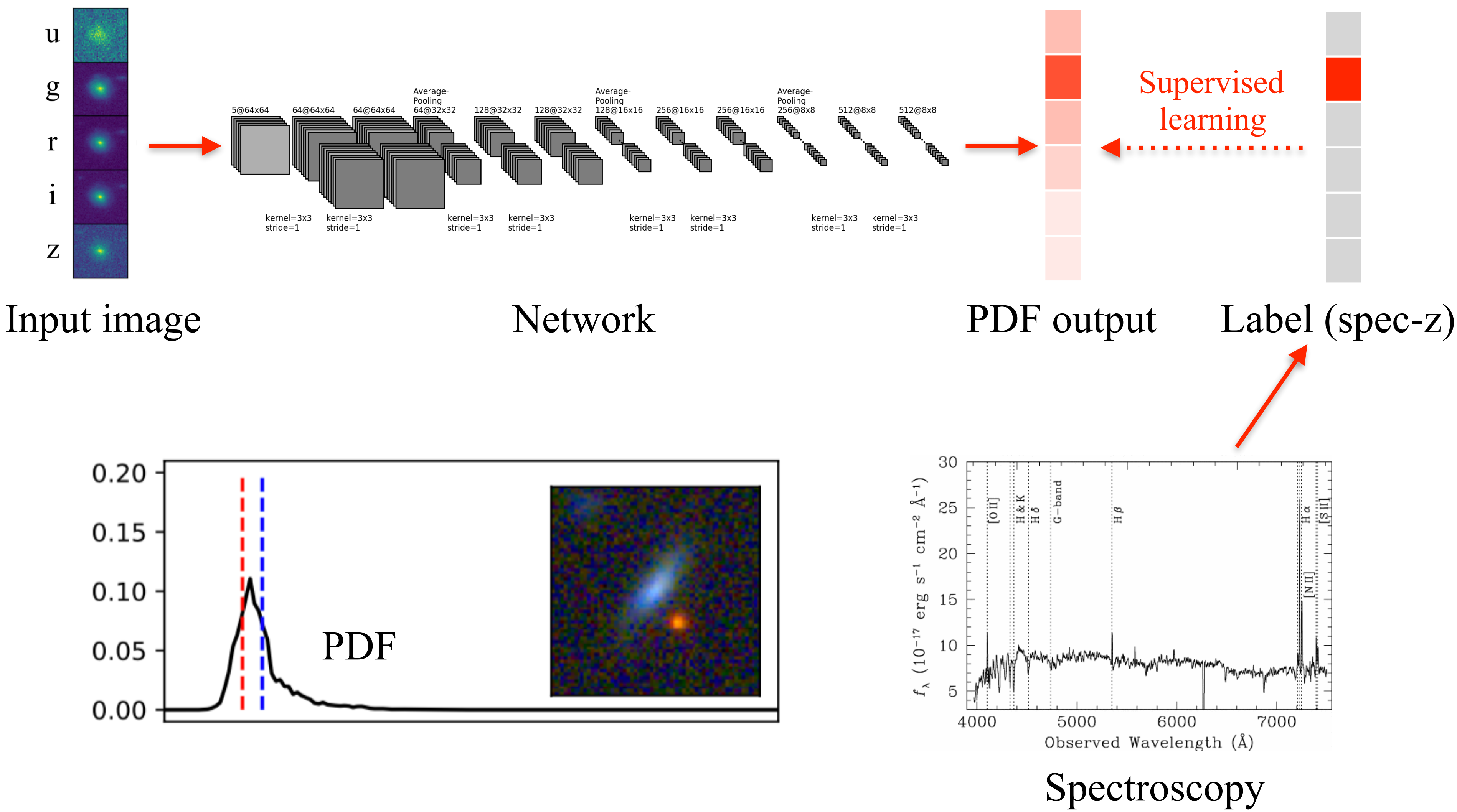
Image from: https://ogrisel.github.io/scikit-learn.org/sklearn-tutorial/tutorial/astronomy/regression.html

- **Photo-z by images**



u g r i z

Photometric redshift (photo-z) estimation as a classification problem supervised by spectroscopic redshift (spec-z)



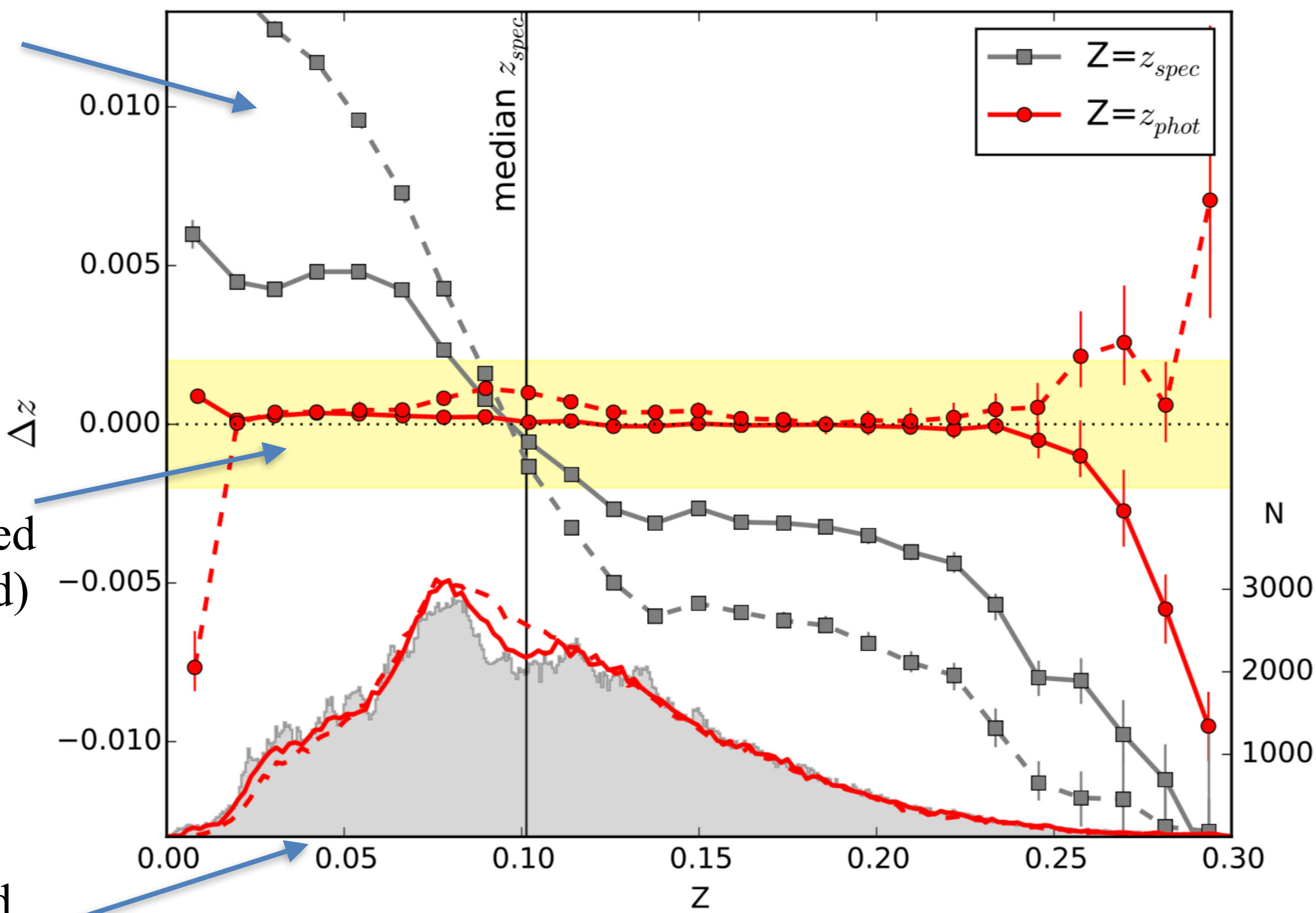
Bias 1: residuals as a function of spec-z or photo-z

Residual = (z_photo - z_spec) / (1 + z_spec)

w.r.t spec-z
(biased)

w.r.t photo-z
(not guaranteed
to be unbiased)

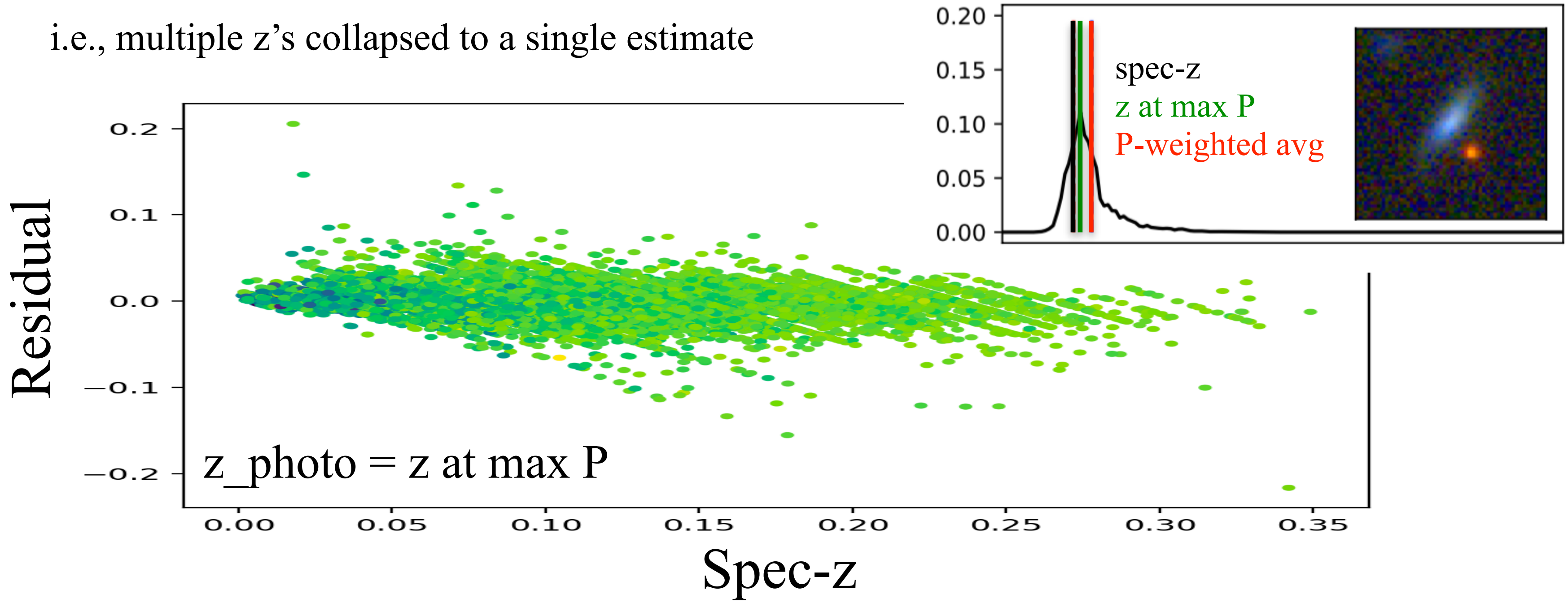
unbalanced
distribution

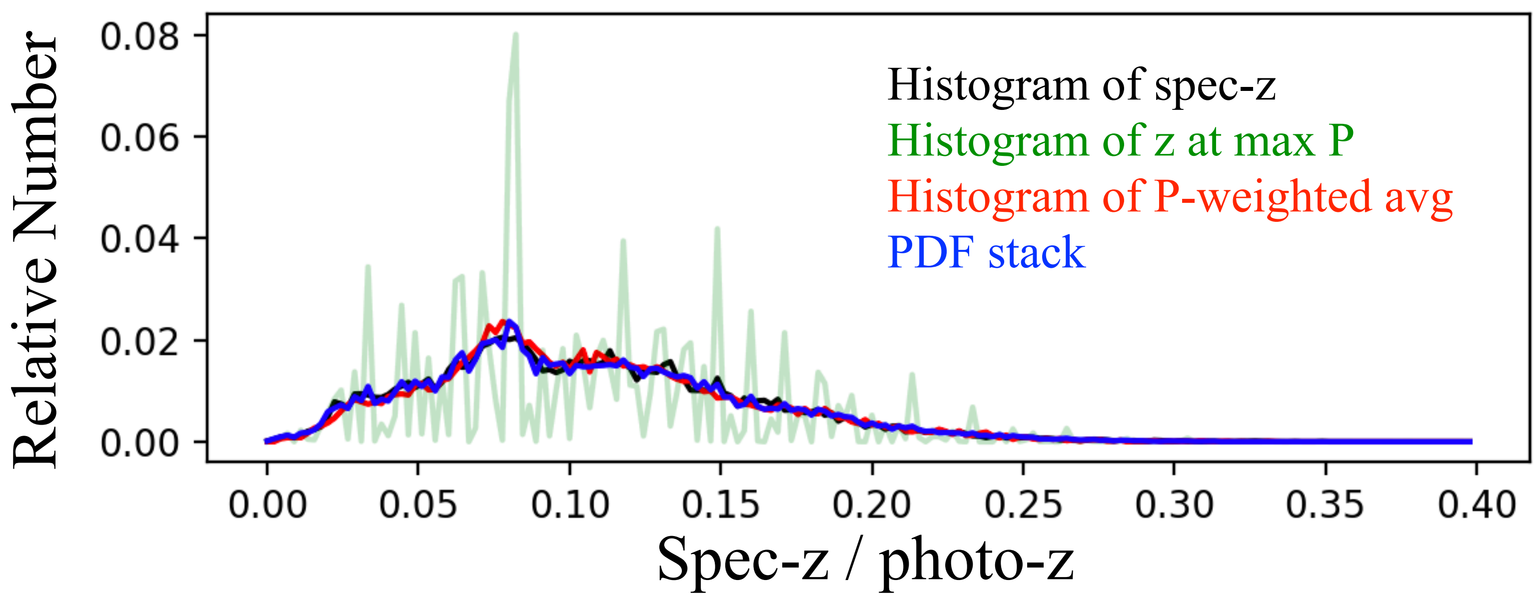


SDSS z<0.4
(Pasquet et al. 2019)

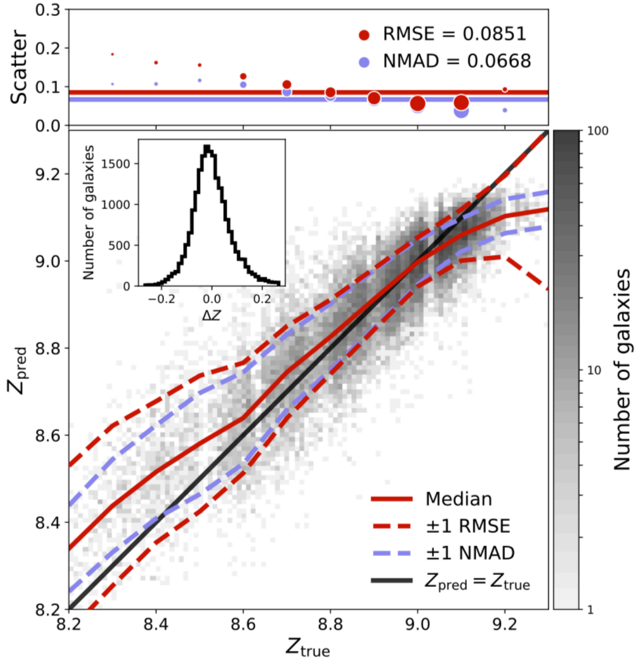
Bias 2: mode collapse

i.e., multiple z's collapsed to a single estimate

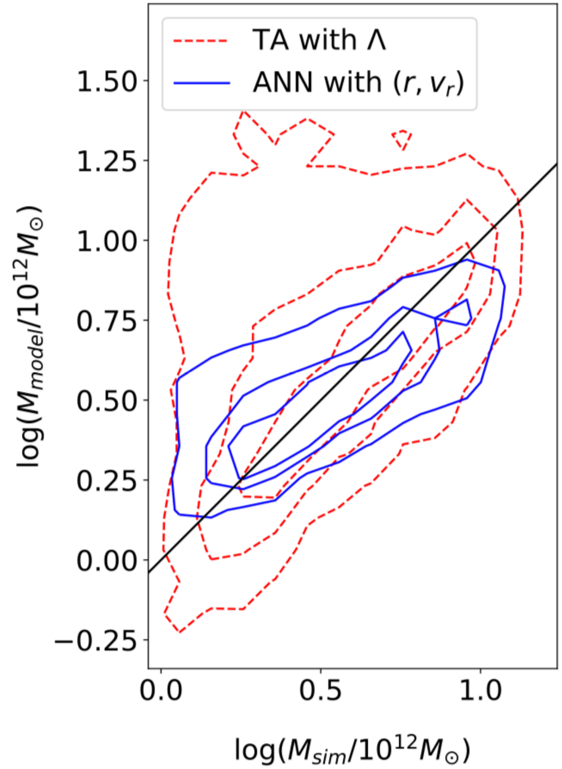




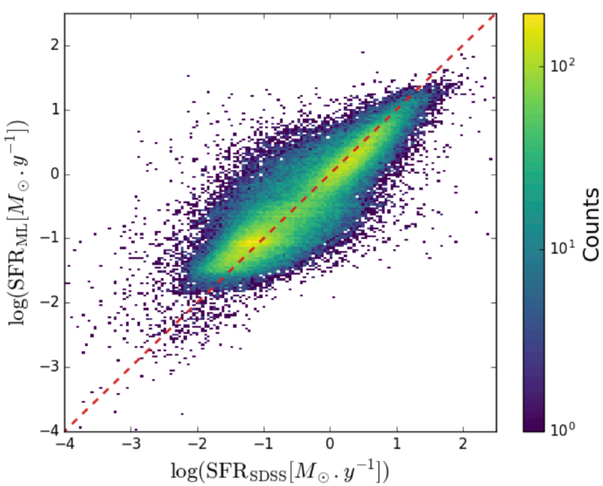
Biases exist in various classification & regression applications



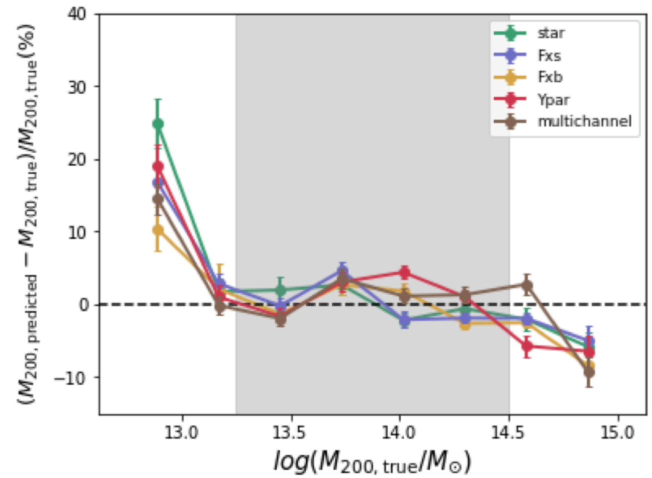
Metallicity (Wu et al. 2019)



Mass of the Local Group (McLeod et al. 2017)



Star Formation Rate (Bonjean et al. 2019)



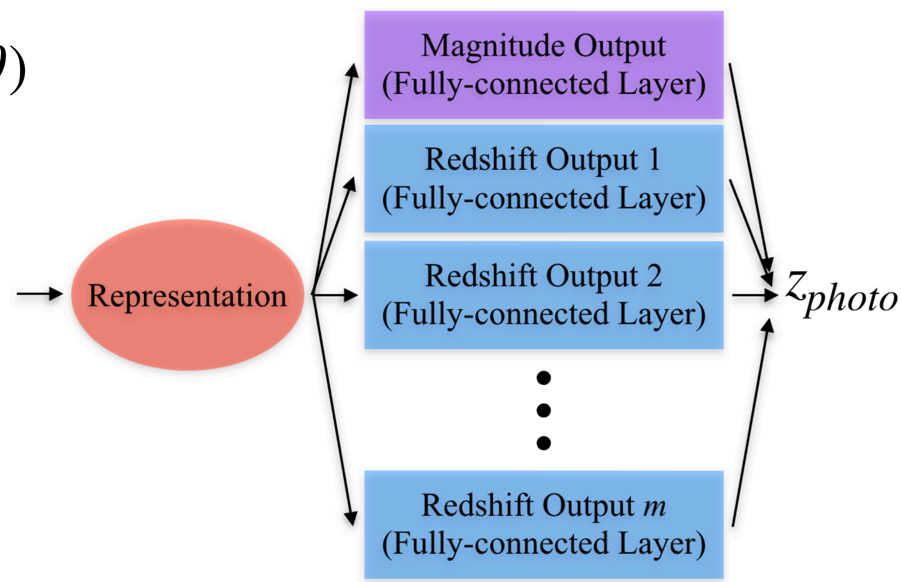
Cluster Mass (Yan et al. 2020)

Splitting the learning of representation and classification

Representation Learning (all data) **Classification (a near-balanced subset)**

Multi-channel outputs (redshift & r-band magnitude)

$$p(z\,|\,\theta) = \sum_{m} p(m\,|\,\theta)\,p(z\,|\,m,\theta)$$



Bias correction procedure

Causes of biases due to data, model, etc.

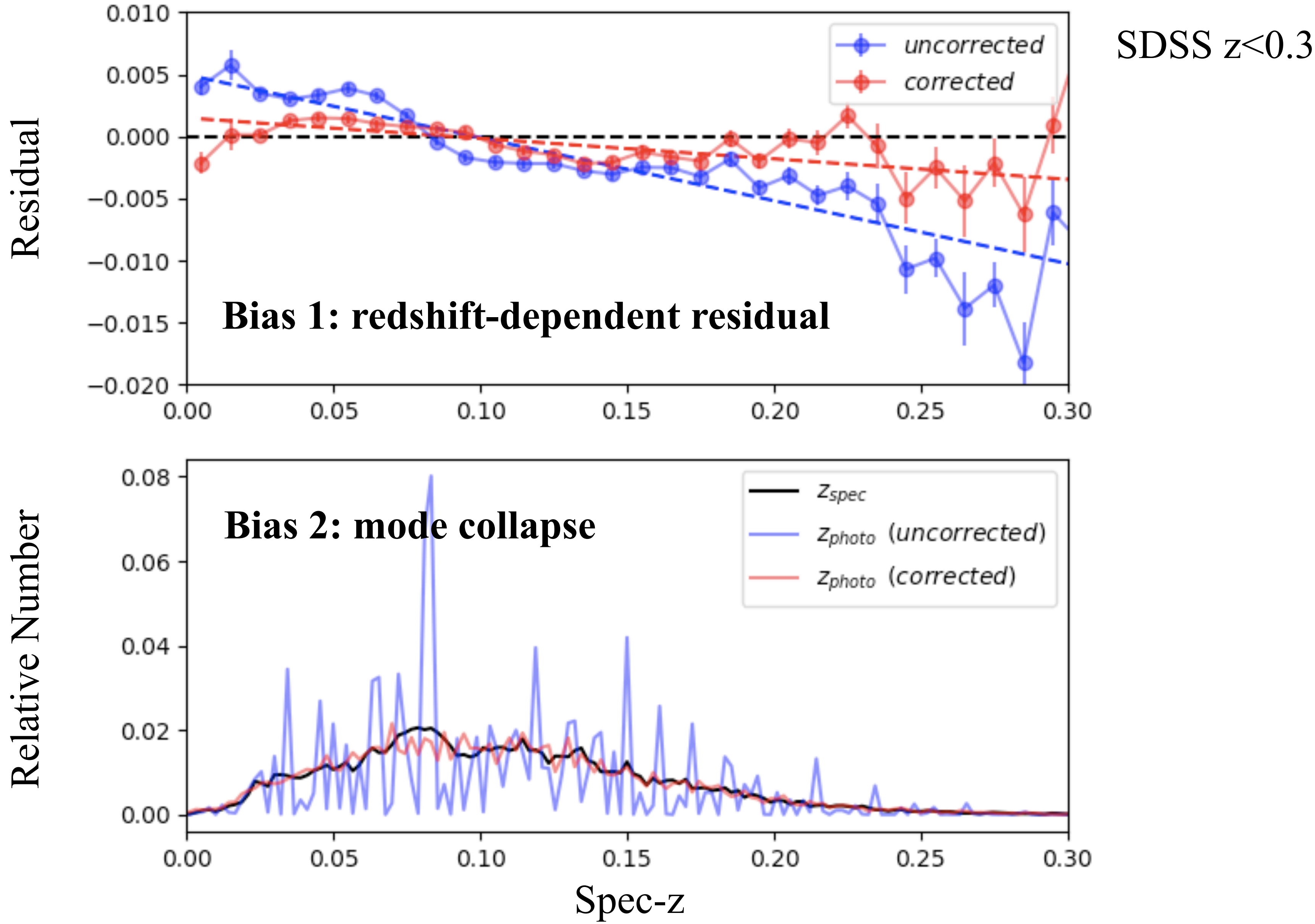
$$p'(z_{photo}\,|\,D) \sim \int q(z_{photo}\,|\,z_{spec}, D)\, p(z_{spec}\,|\,D)\, dz_{spec}$$

$$p''(z'_{photo}\,|\,D) \sim \int \tilde{q}(z'_{photo}\,|\,z_{photo}, D)\, p'(z_{photo}\,|\,D)\, dz_{photo}$$

Corrections according to pre-estimated redshifts

Bias	Cause	Correction
Over-population-induced residuals	Over-density	Construct near-balanced subset
Under-population-induced residuals	Under-density	Shift labels
Mode collapse	Local over-confidence	Use soft labels

Biases w.r.t spec-z are reduced by our method



Correcting biases w.r.t photo-z

- Biases w.r.t photo-z not compatible with biases w.r.t spec-z
- First perform correction for spec-z then perform calibration for photo-z

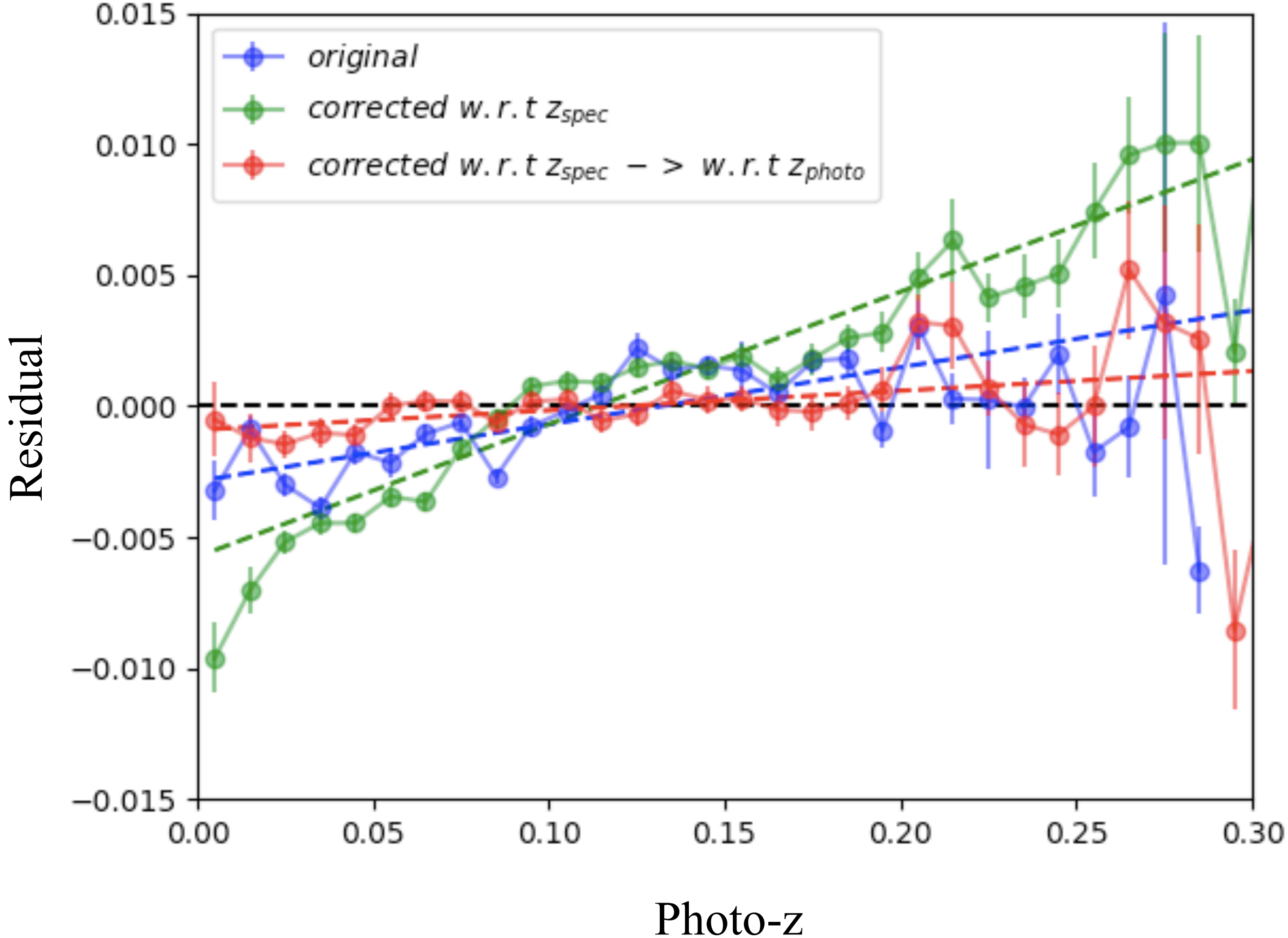
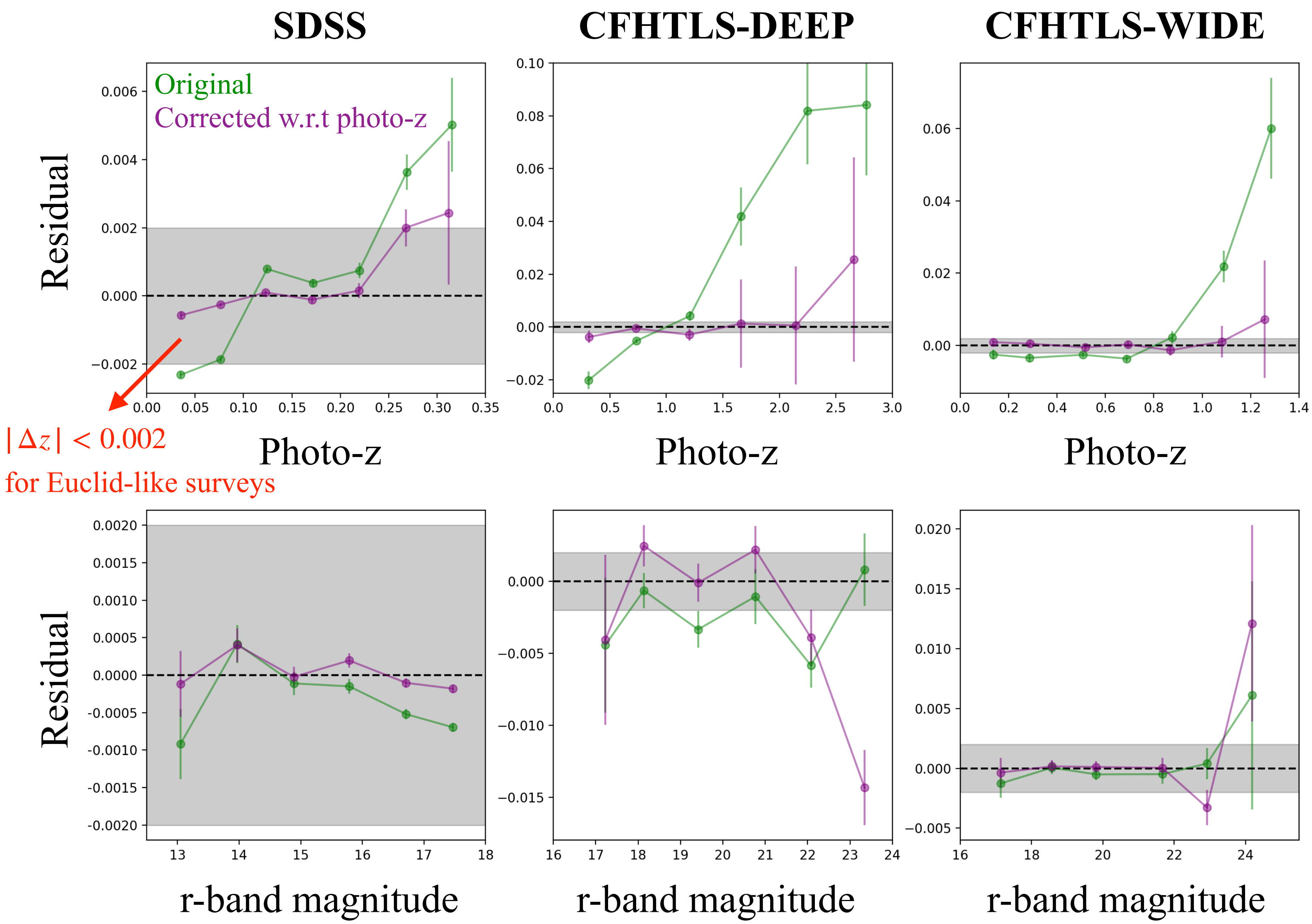


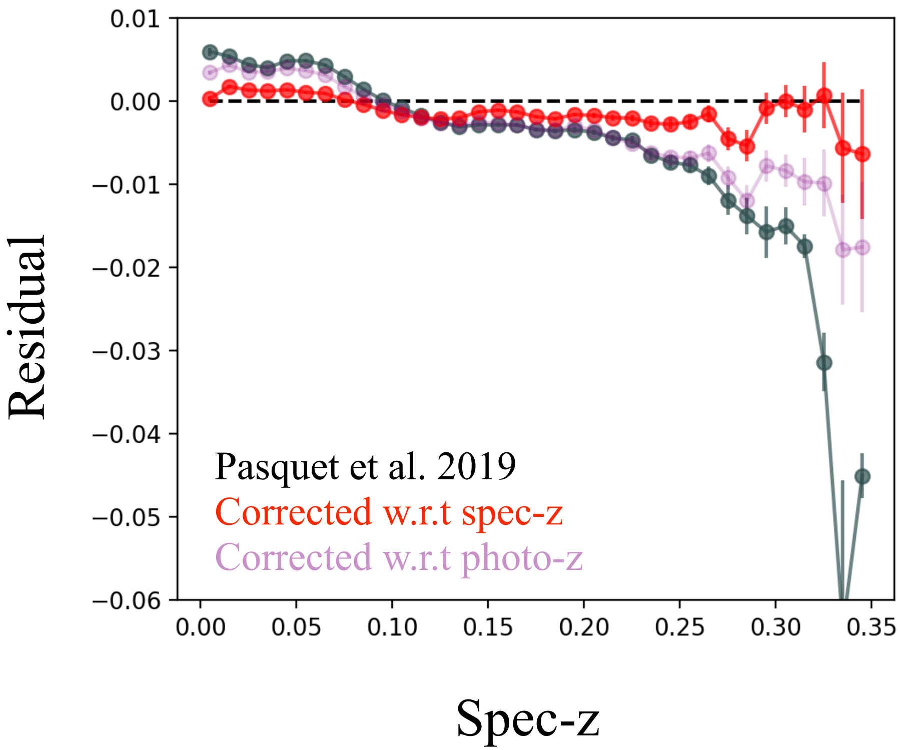
Photo-z calibration for cosmological analysis



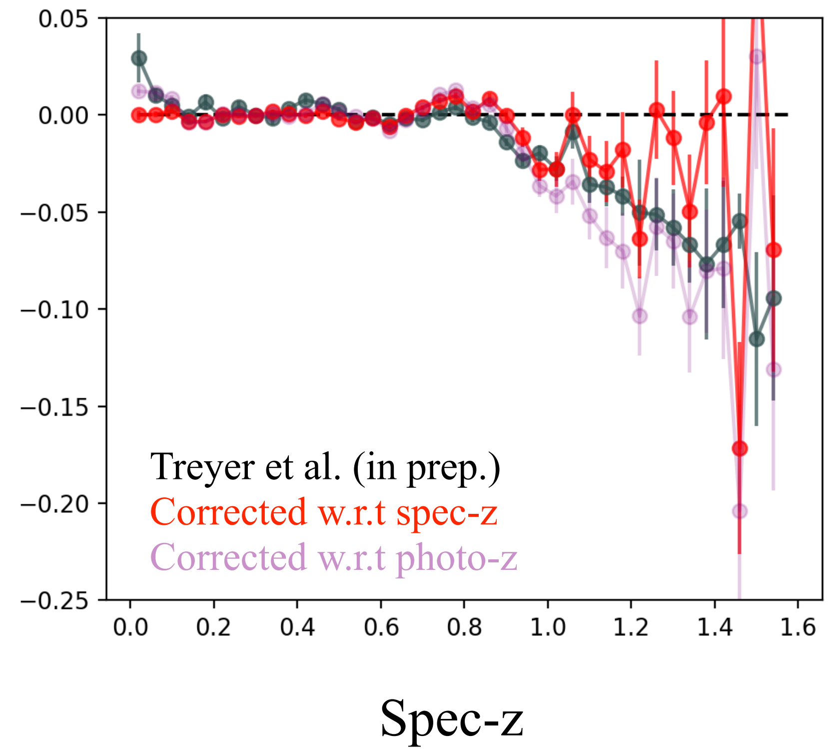
Comparison with state-of-the-art results

- Bias correction w.r.t spec-z

SDSS

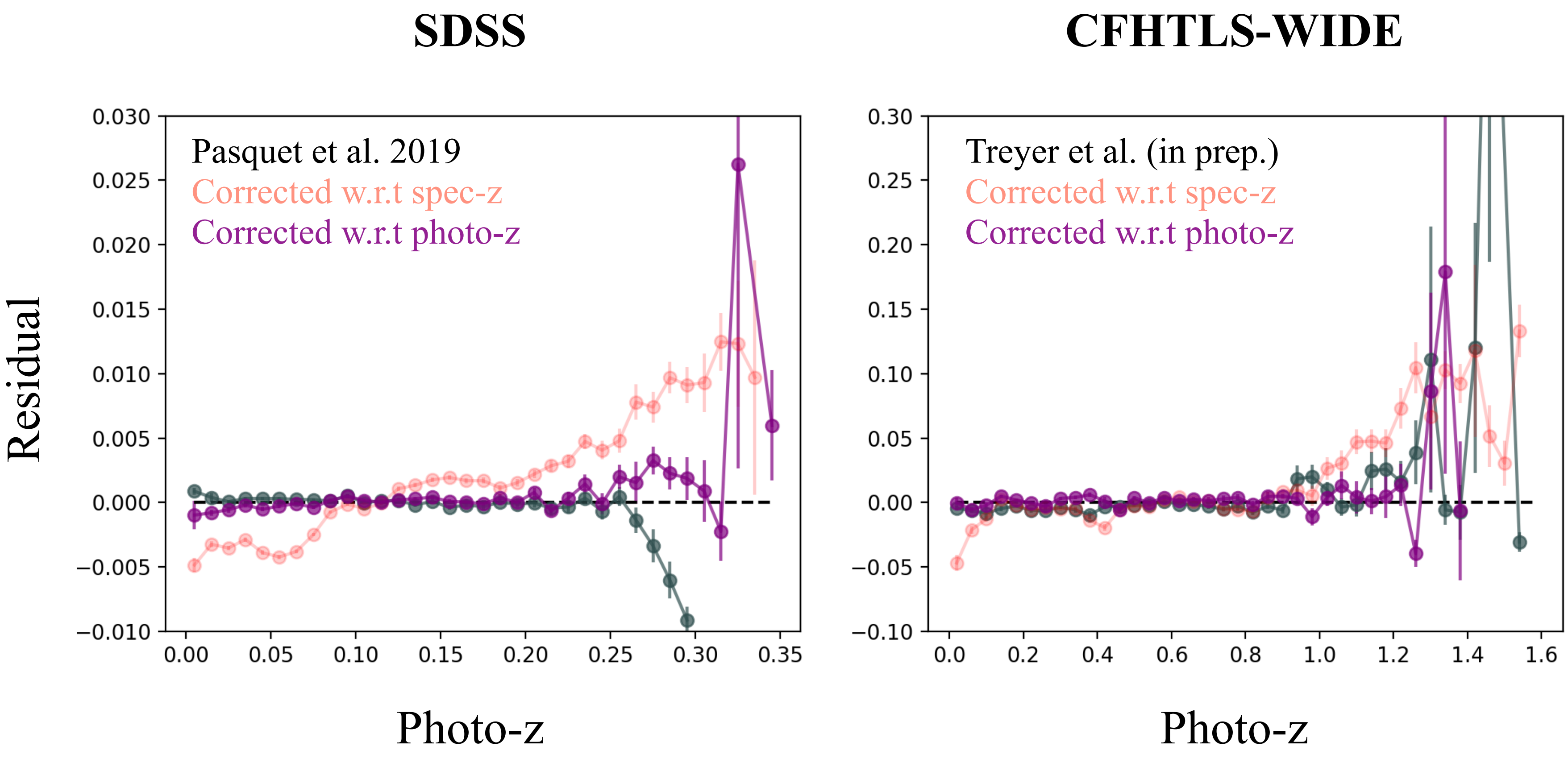


CFHTLS-WIDE



Comparison with state-of-the-art results (continued)

- Bias correction w.r.t photo-z



Conclusion

- Two forms of biases.

- Redshift-dependent residuals
  (Over-population-induced & under-population-induced)
- mode collapse

- Key 1: split the learning of representation and classification.

- The representation potentially contains all required information
  (though biased).
- Re-train the classification part for resolving biases or other needs.
- Balance the number density of training data.
- Adjust the target output (= shift & soften labels).

- Key 2: Correct biases separately for spec-z and photo-z.

- First correct the biases w.r.t spec-z.
- Calibrate for photo-z if needed.

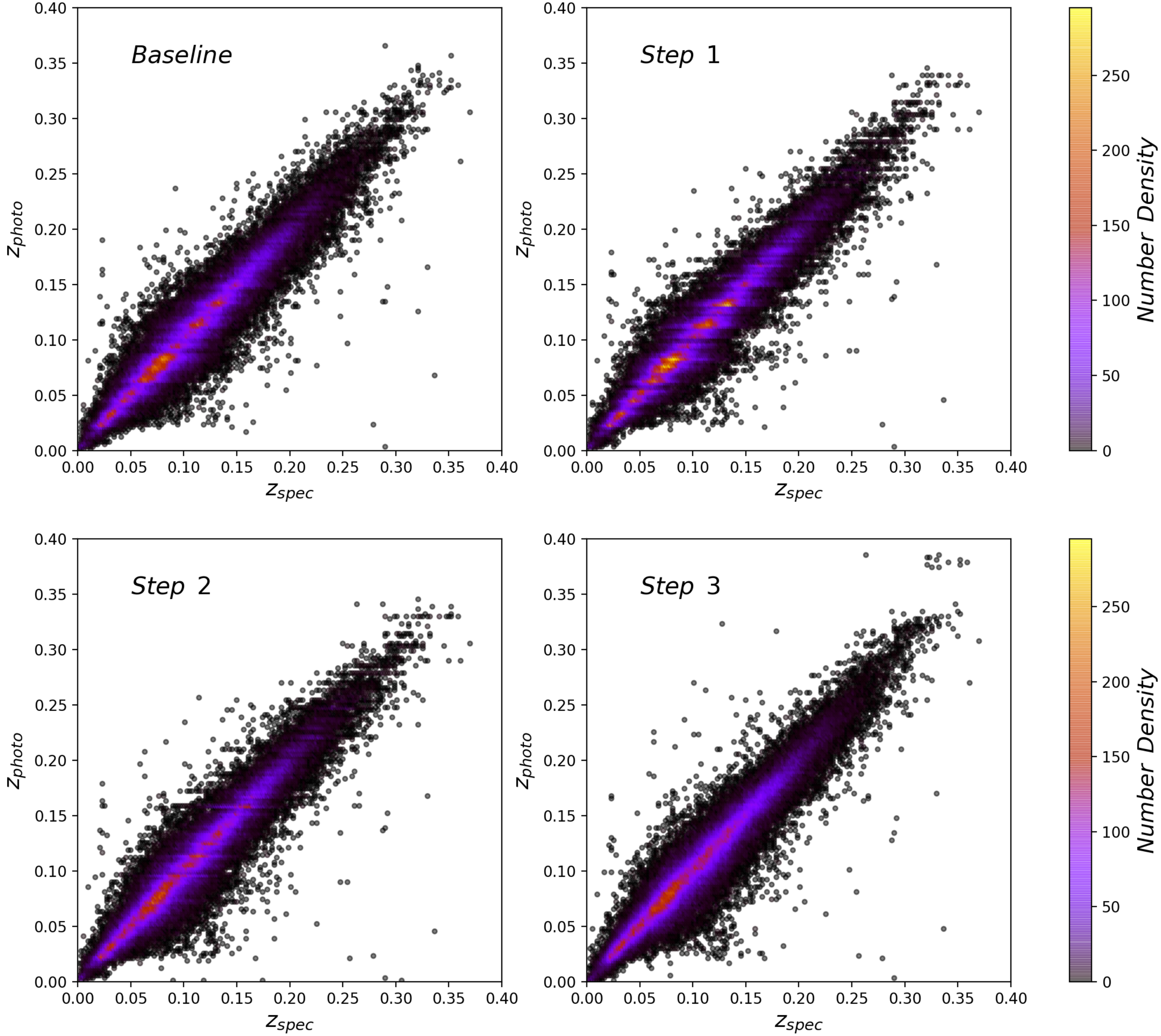
- Prospect

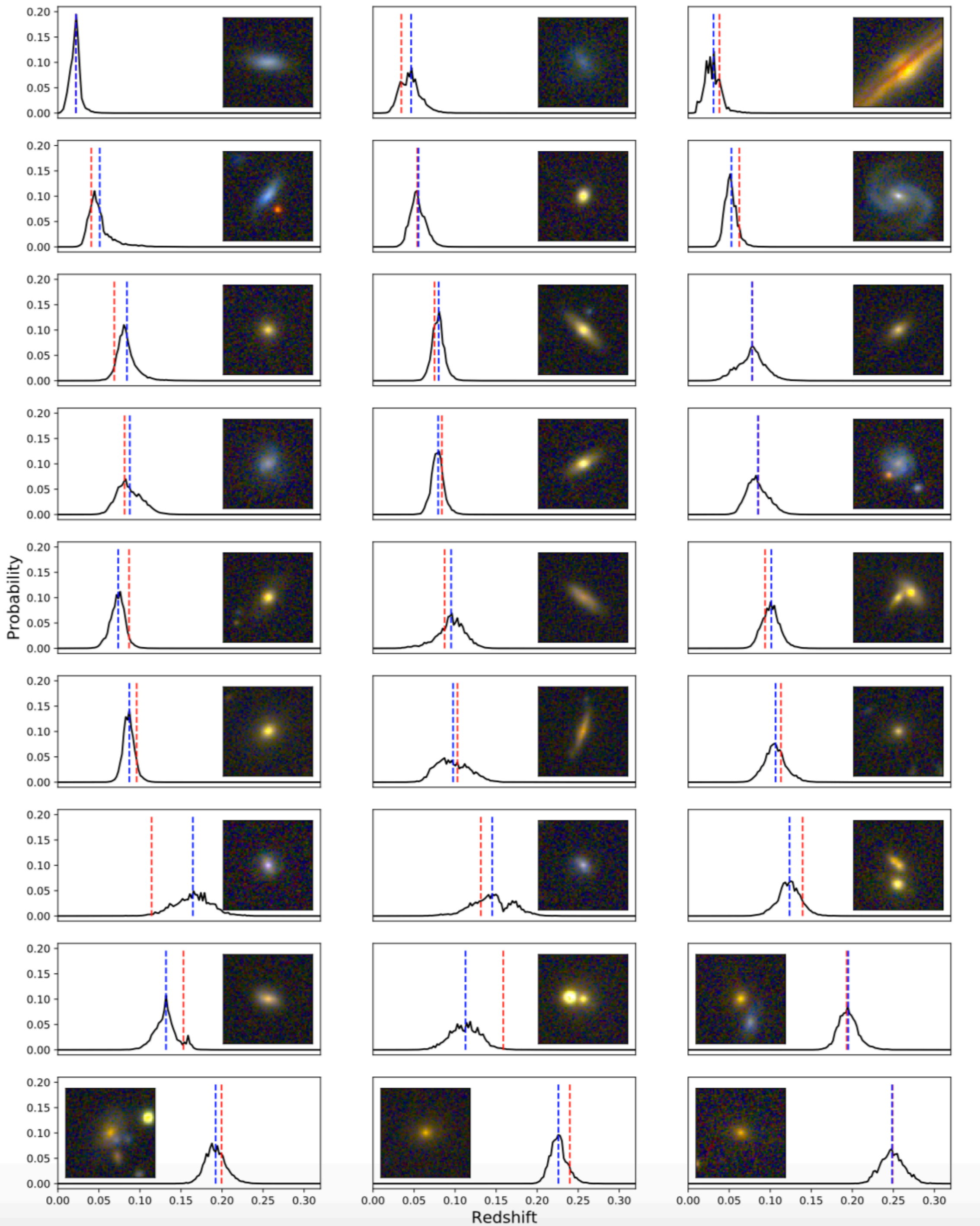
- Combined with other photo-z methods.
- Generalized to regression problems and used in other applications.

Acknowledgement

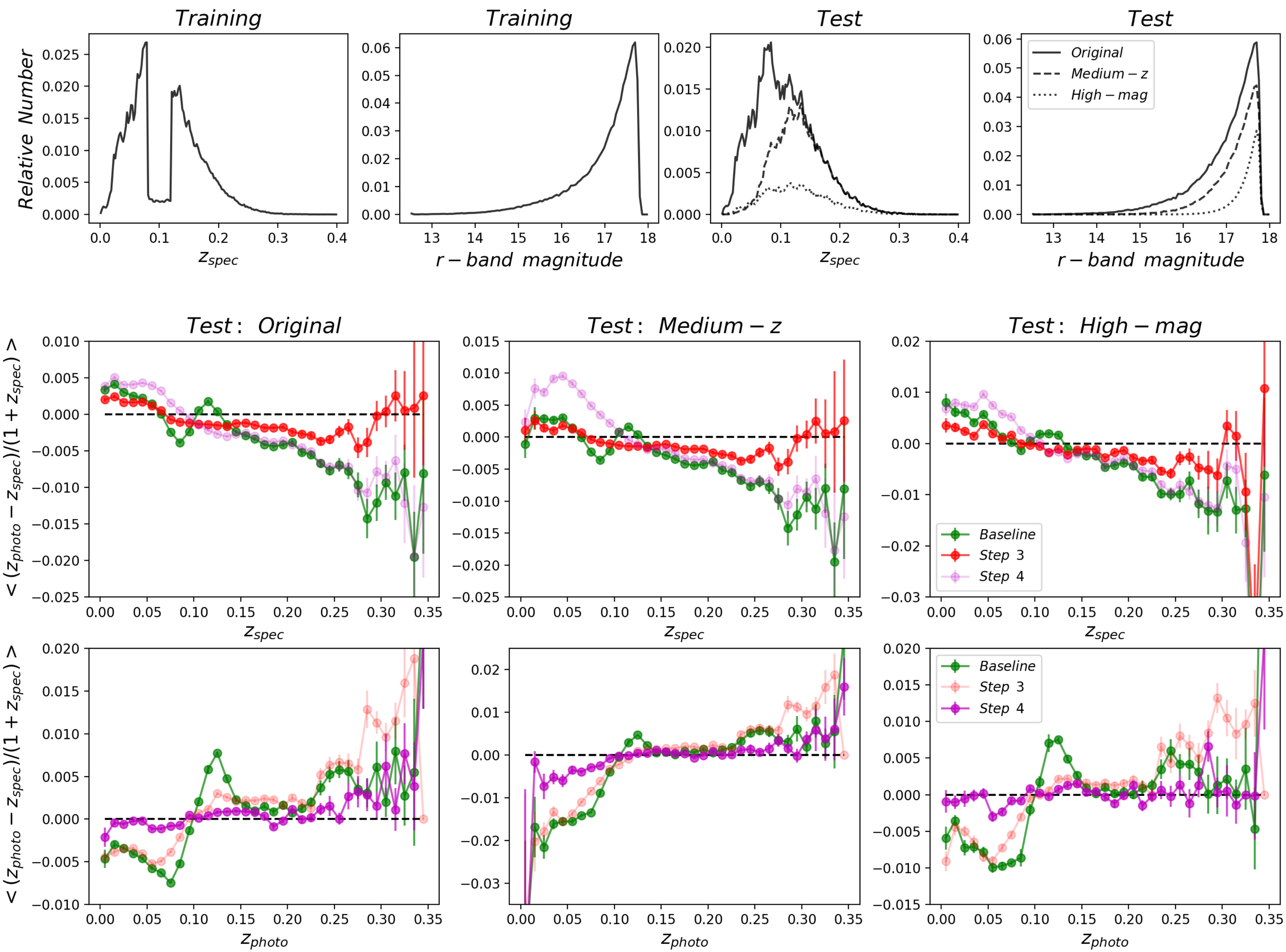
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No713750. Also, it has been carried out with the financial support of the Regional Council of Provence-Alpes-Côte d’Azur and with the financial support of the A*MIDEX (n° ANR-11-IDEX-0001-02), funded by the Investissements d'Avenir project funded by the French Government, managed by the French National Research Agency (ANR).

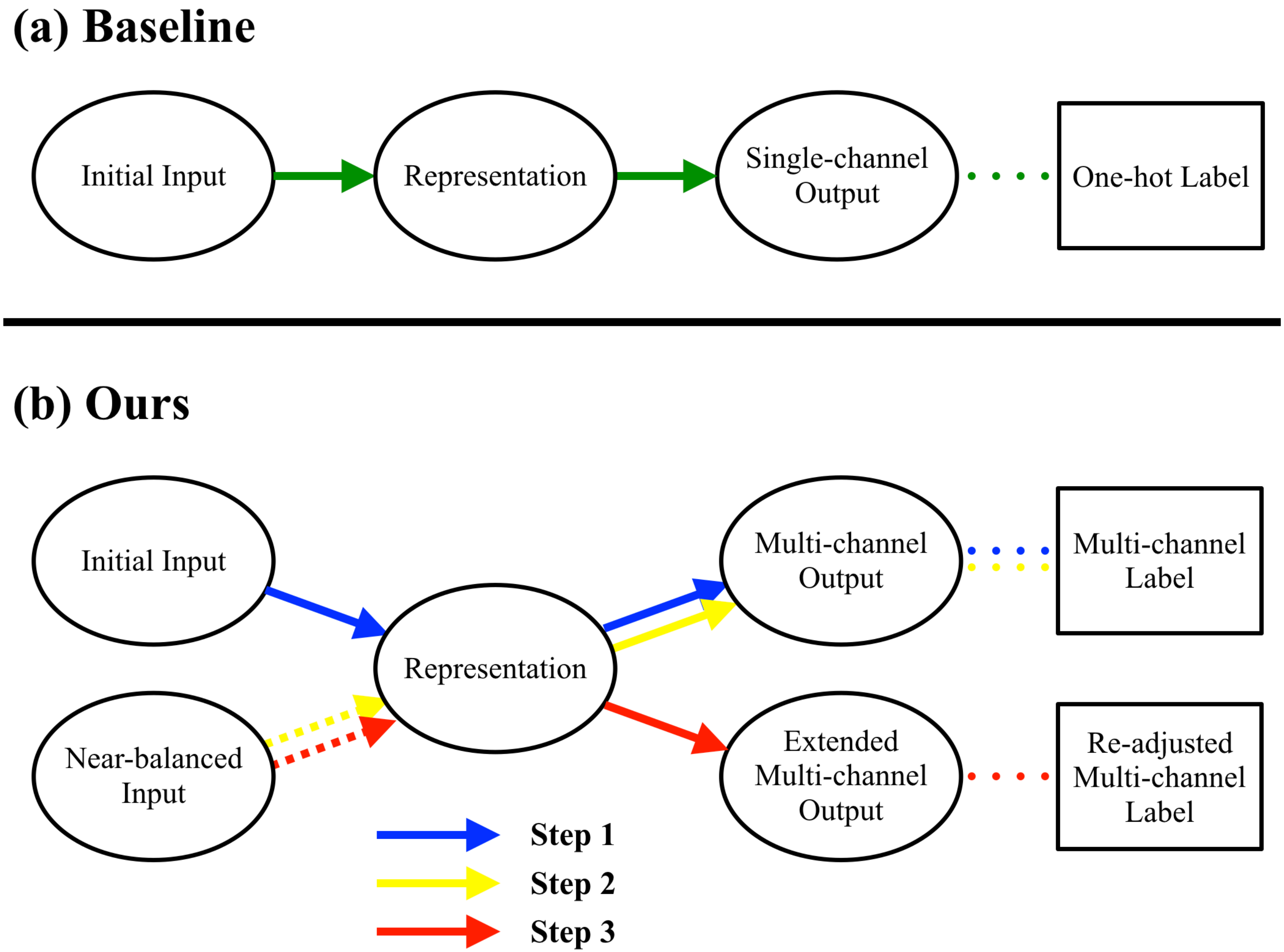
Back-up slides


Baseline
Step 1
Step 2
Step 3
z_{photo}
z_{spec}
Number Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
50
100
150
200
250

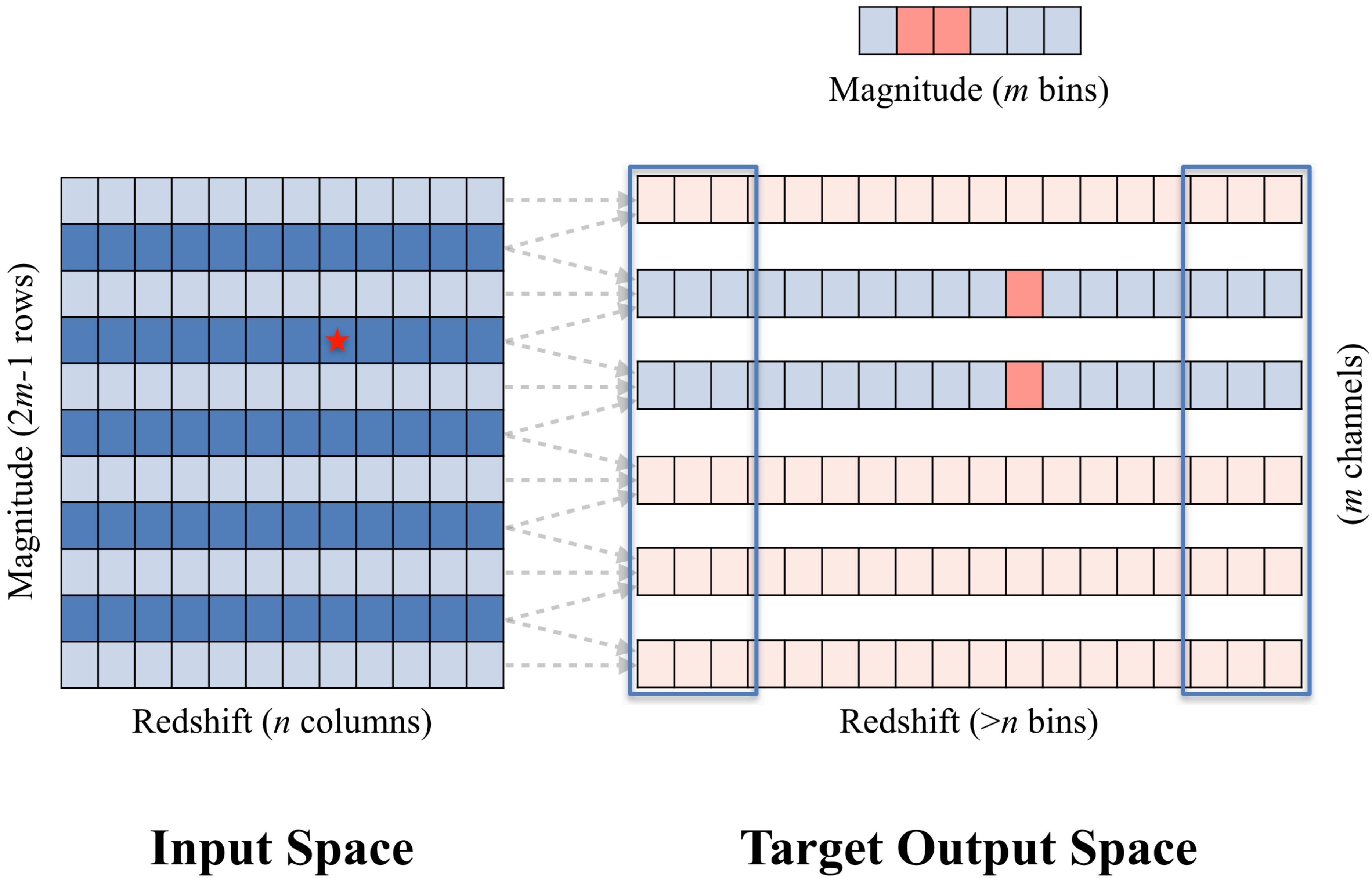


(Pasquet et al. 2019)


Training
Training
Test
Test
Relative Number
z_{spec}
$r-band\ \ magnitude$
Original
Medium − z
High − mag
Test: Original
Test: Medium − z
Test: High − mag
$<(z_{photo}-z_{spec})/(1+z_{spec})>$
z_{photo}
Baseline
Step 3
Step 4

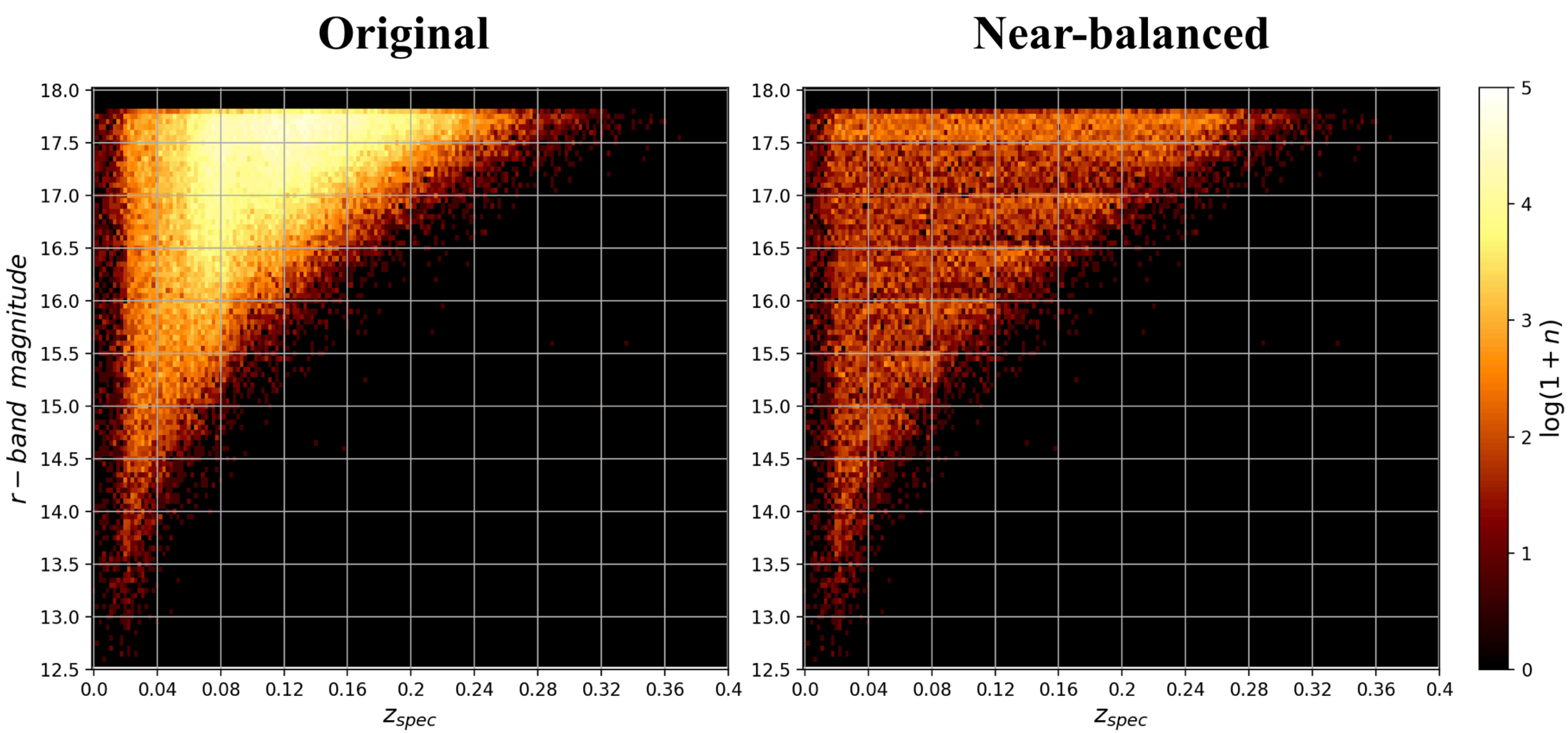

(a) Baseline
Initial Input
Representation
Single-channel Output
One-hot Label
(b) Ours
Initial Input
Near-balanced Input
Representation
Multi-channel Output
Multi-channel Label
Extended Multi-channel Output
Re-adjusted Multi-channel Label
Step 1
Step 2
Step 3

Correspondence between input space and output channels



Correction of over-population-induced residuals: construct a near-balanced subset

- Divide the input space into two-dimensional (z, mag) cells.
- Near-balanced subset: randomly select N instances in each cell (N <= Nth).

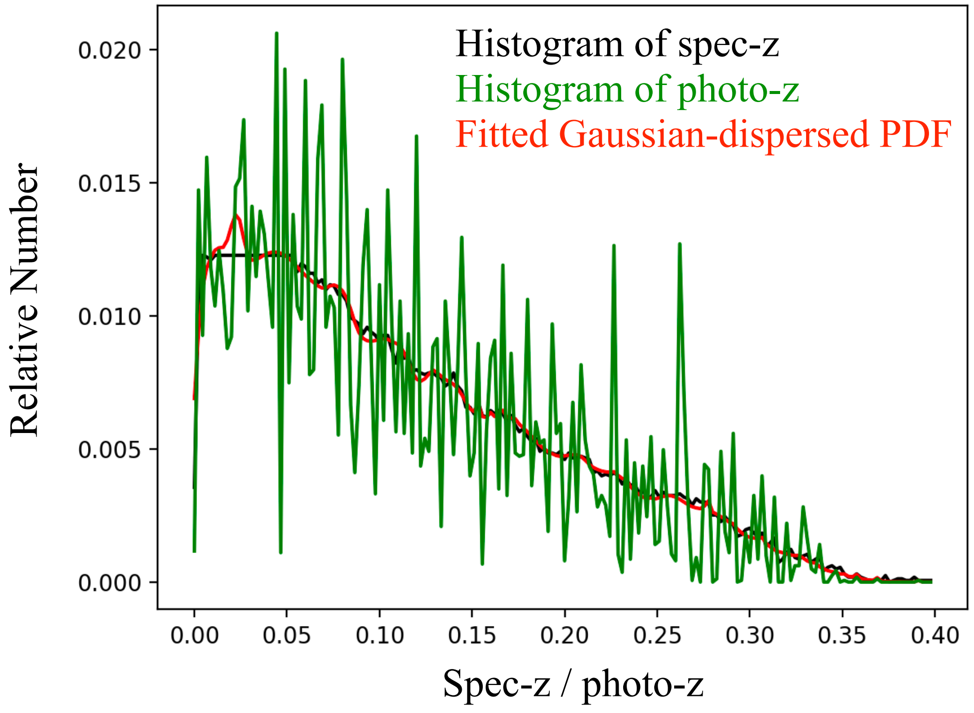


Correction of mode collapse: introduce dispersion to labels

- Model $\tilde{q}(z'|z,D)$ as Gaussians. Labels are given by $\int \tilde{q}(z'|z,D)\,\delta(z|D)\,dz$
- Fit with the histogram of spec-z and the histogram of pre-estimated photo-z.

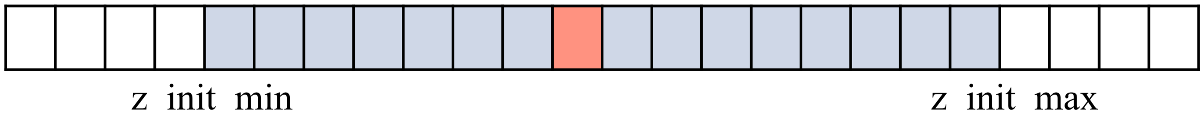
$$\min_{\sigma}\{-p_{spec}\log(p_{photo} * N(0,\sigma))\}$$

Approximation: same labeling dispersion along z for each r-band magnitude.



Correction of under-population-induced residuals

- Extend the range.



- Relocate the (soft) label according to the center-of-mass of the modified distribution.