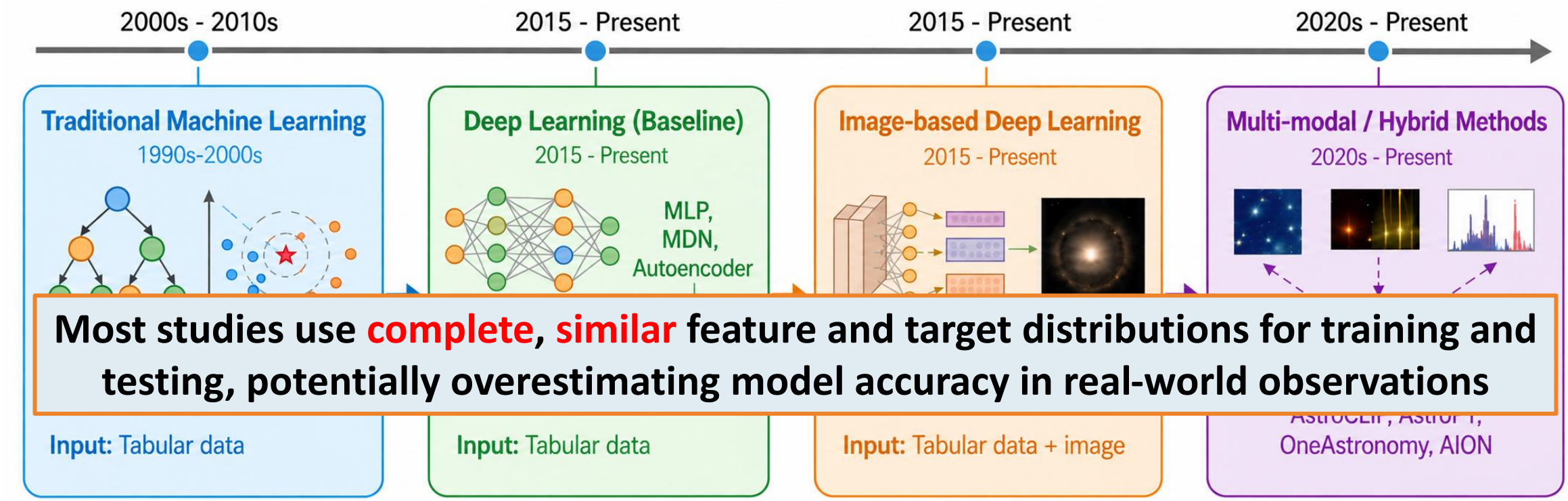
Rethinking the Validation for DL Photometric Redshifts: The Impact of Redshift Distribution Shift

ZHU CHEN (SHANGHAI NORMAL UNIVERSITY)

zhuchen@shnu.edu.cn

the 3rd International Workshop on AI + Astronomy: Models, Data, Discovery, Zhejiang Lab, Oct. 19-22, 2026

Evolution of ML-based Photo-z Estimation



Most studies use complete, similar feature and target distributions for training and testing, potentially overestimating model accuracy in real-world observations

Collister & Lahav, 2004 (ANNz)
Wadaekar, 2005 (SVM)
Ball et al. 2007 (kNN)
Carlilest et al. 2010 (RF)
Gerdes et al. 2010 (BDT)
Kind & Brunner 2013 (TPZ)
Lu, Luo, **Chen** et al. 2023 (RF)
…

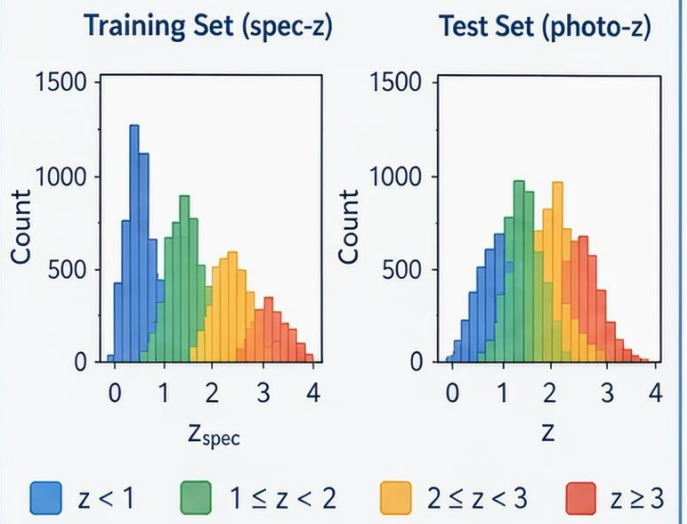
Firth et al. 2003 (MLP)
Eriksen et al. 2020 (Autoencoder+MDN)
Ansari et al. 2021 (MDN)
Zhou, …, **Chen** et al. 2022 (Bayesian NN)
Luo, …, **Chen** et al. 2024 (LSTM)
Teixeira et al. 2024, (RNN+MDN)
Ren et al. 2025 (Normalizing flow)
…

Hoyle, 2016 (CNN+MLP)
D'Isanto et al. 2018 (CNN+MDN)
Pasquet et al. 2019 (CNN)
Dey, 2022 (Capsule NN)
Zhou, …, **Chen** et al. 2021 (CNN+Bayesian NN)
Sen et al., 2025 (CNN+Transformer)
…

AstroCLIP (2024)
AstroPT (2024)
Euclid AstroPT Photo-z (2025)
OneAstronomy (2025)
AION-1 (2025)

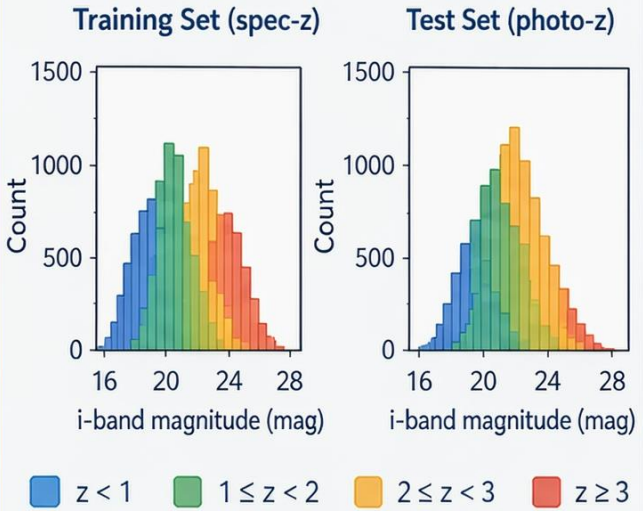
Challenges of Real-World Training Data

1. Different Redshift Distributions



- **Training set:** more low-z sources, fewer high-z sources.
- **Test set:** less skewed redshift distribution.

2. Different Brightness Distributions (Selection Effects)



- Selection effects cause the training set to be **biased toward brighter objects**, while the test set contains more fainter sources.

3. Missing Data Patterns

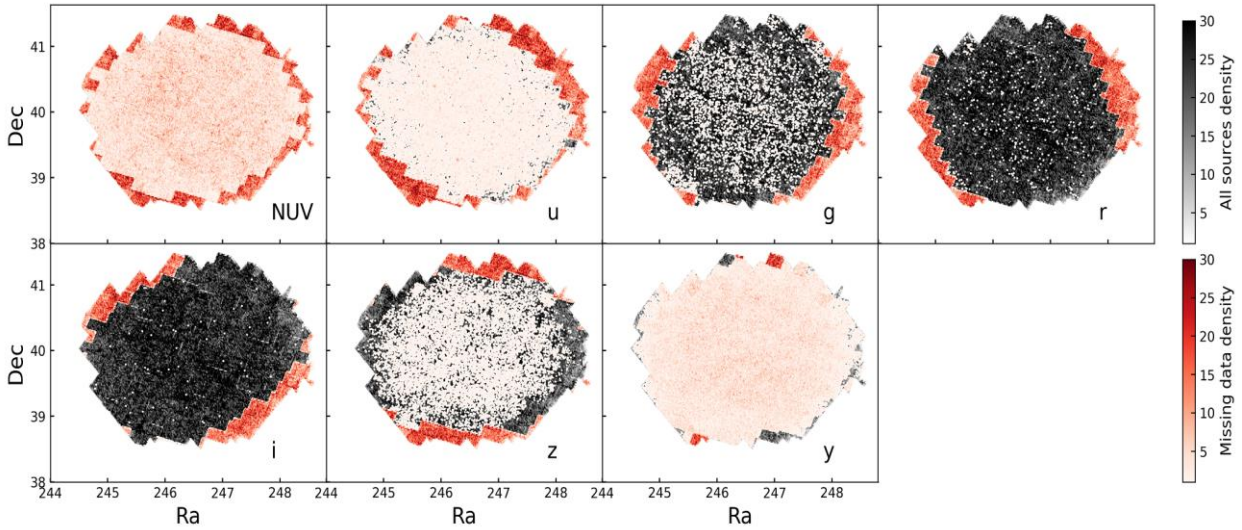


- **Training set:** fewer missing bands, more complete data.
- **Test set:** more missing bands (e.g., u, z, Y), more incomplete samples.

How to construct representative training/test samples?

CSST Mock data

Handling Missing data-- imputation

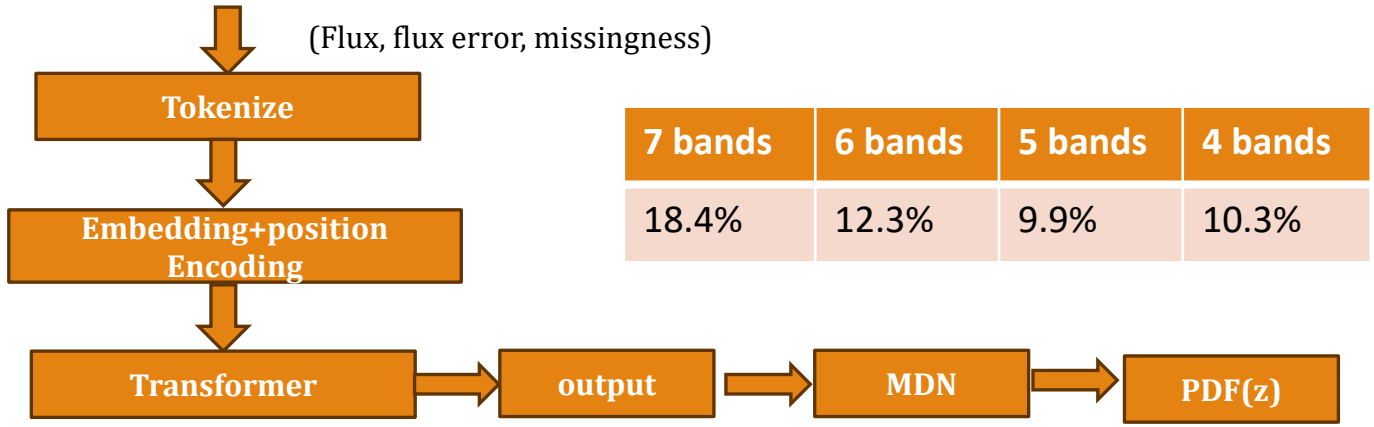


Model	Drop 1 band		Drop 2 bands		Drop 3 bands	
	$f_{out}(\%)/\sigma_{NMAD}$/bais	mae/rmse/mre	$f_{out}(\%)/\sigma_{NMAD}$/bais	mae/rmse/mre	$f_{out}(\%)/\sigma_{NMAD}$/bais	mae/rmse/mre
Test (non-imputed)	3.79/0.061/0.0022	...	11.33/0.086/-0.0112	...	26.34/0.131/-0.0375	...
KNN	1.41/0.049/0.0100	0.028/0.062/0.036	1.73/0.051/0.0099	0.033/0.073/0.041	2.48/0.053/0.0095	0.040/0.088/0.050
RF	1.45/0.049/0.0100	0.027/0.060/0.034	7.38/0.059/0.0050	0.124/0.291/0.158	19.22/0.080/0.0038	0.236/0.443/0.299
CatBoost	1.73/0.051/0.0094	0.048/0.099/0.061	18.17/0.077/0.0148	0.258/0.383/0.329	34.63/0.129/0.0289	0.364/0.527/0.463
US-GAN	4.09/0.055/0.0087	0.182/0.252/0.231	18.88/0.080/0.0308	0.325/0.420/0.410	21.02/0.094/0.0062	0.277/0.397/0.351
GP-VAE	7.97/0.064/0.0152	0.231/0.304/0.293	22.78/0.093/0.0243	0.309/0.393/0.391	32.26/0.123/0.0032	0.390/0.517/0.494
M-RNN	32.79/0.107/0.0285	0.785/1.013/0.995	44.93/0.175/0.0411	0.787/1.004/0.995	51.22/0.215/0.0543	0.785/1.005/0.994
BRITS	4.04/0.054/0.0087	0.163/0.289/0.207	8.94/0.065/0.0147	0.210/0.344/0.256	20.02/0.094/0.0229	0.257/0.373/0.326
Transformer	1.94/0.052/0.0104	0.071/0.119/0.090	2.58/0.053/0.0093	0.058/0.114/0.074	5.72/0.063/0.0069	0.091/0.151/0.115
iTransformer	1.67/0.052/0.0095	0.051/0.103/0.064	2.40/0.054/0.0100	0.053/0.109/0.067	4.37/0.060/0.0074	0.075/0.139/0.095
SAITS	1.41/0.048/0.0100	0.024/0.055/0.030	2.10/0.052/0.0106	0.043/0.087/0.054	3.27/0.056/0.0080	0.051/0.107/0.065

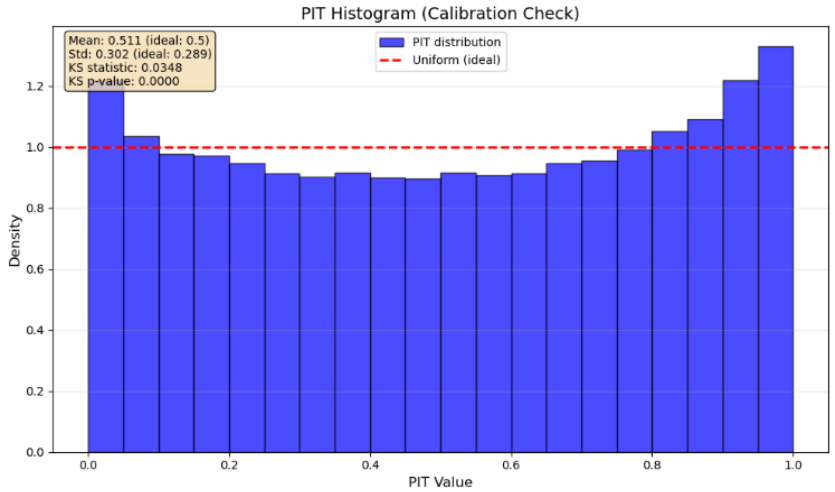
Wang, Chen, et al. 2026, A&A

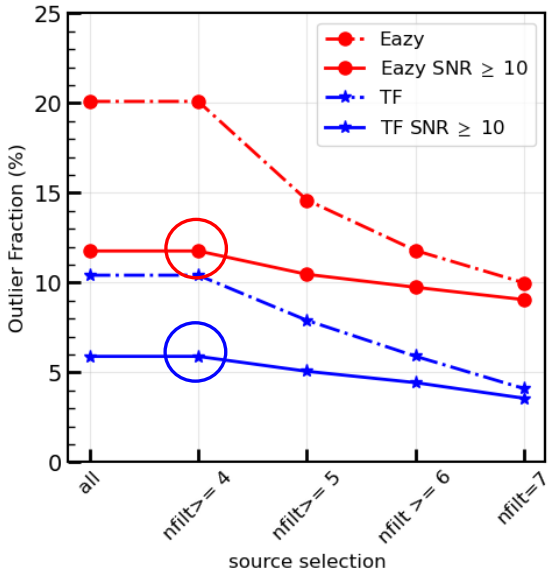
- KNN and SAITS excel under complete training data, even with small training sets.
- SAITS remains reliable with highly incomplete training data, while KNN degrades significantly.
- SAITS leads under realistic CSST conditions, highlighting the importance of matching training and test distributions.
- Imputation is **ineffective** for **non-random missingness** and may even degrade photometric redshift prediction accuracy.

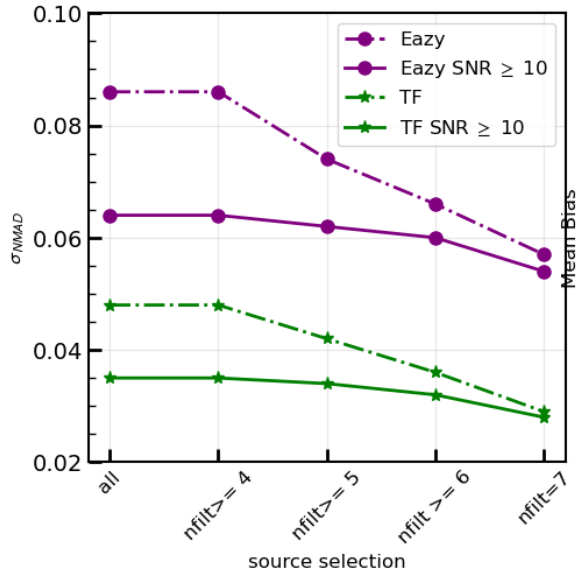
Handling Missing data – photo-z regression

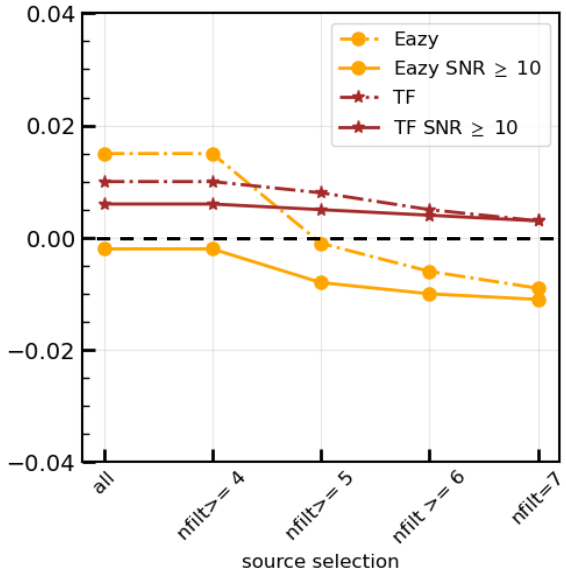


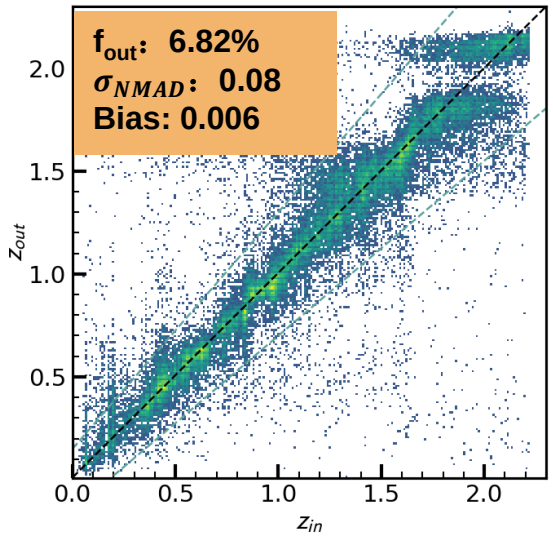
7 bands	6 bands	5 bands	4 bands
18.4%	12.3%	9.9%	10.3%



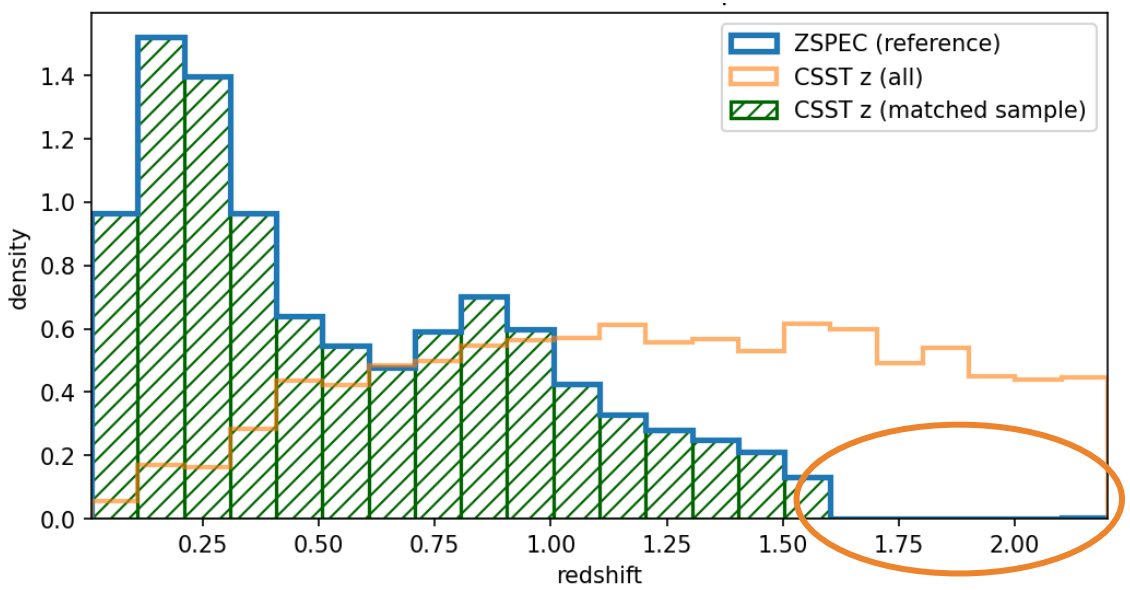


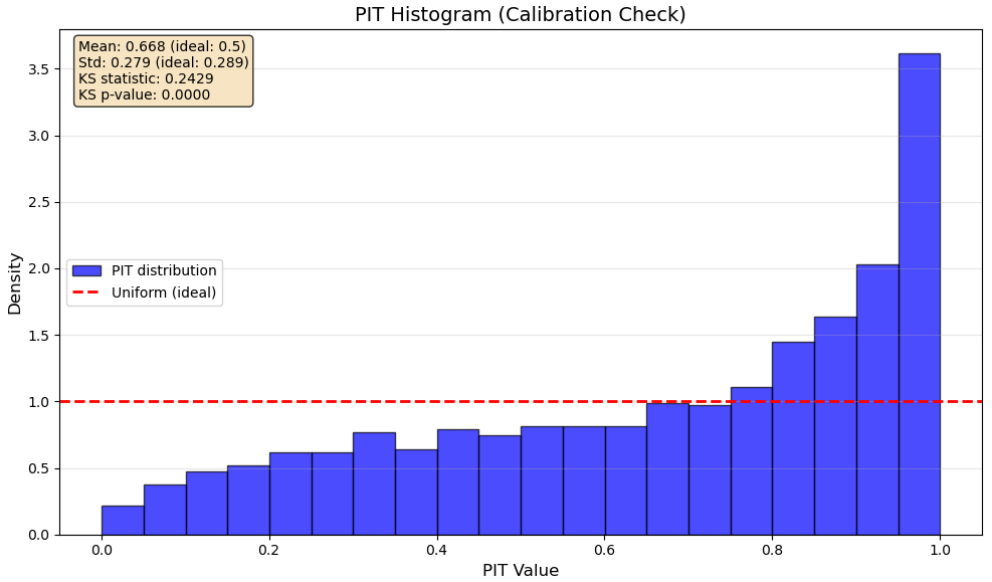


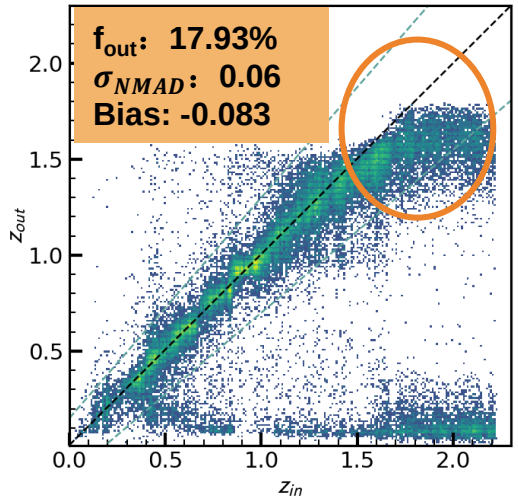


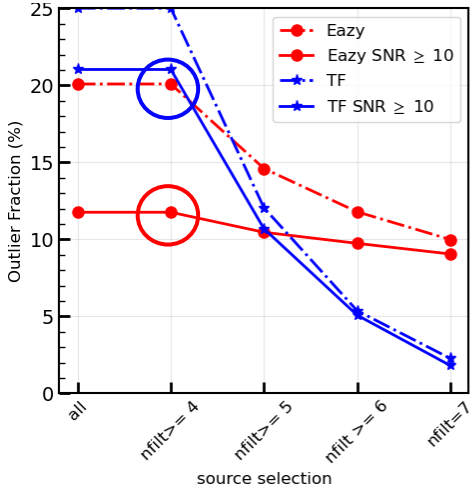


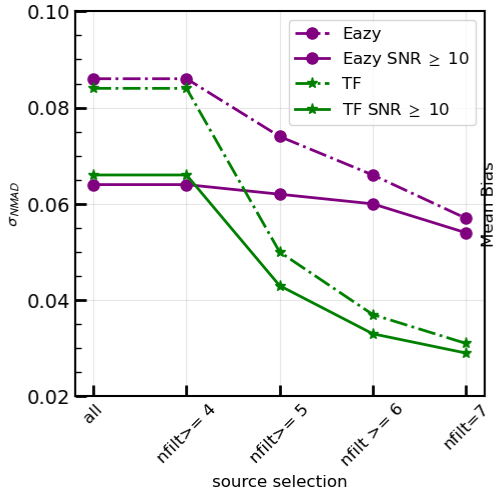
Redshift distribution shift effect

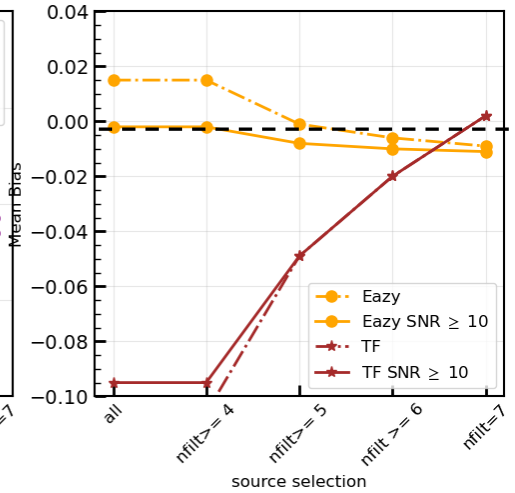




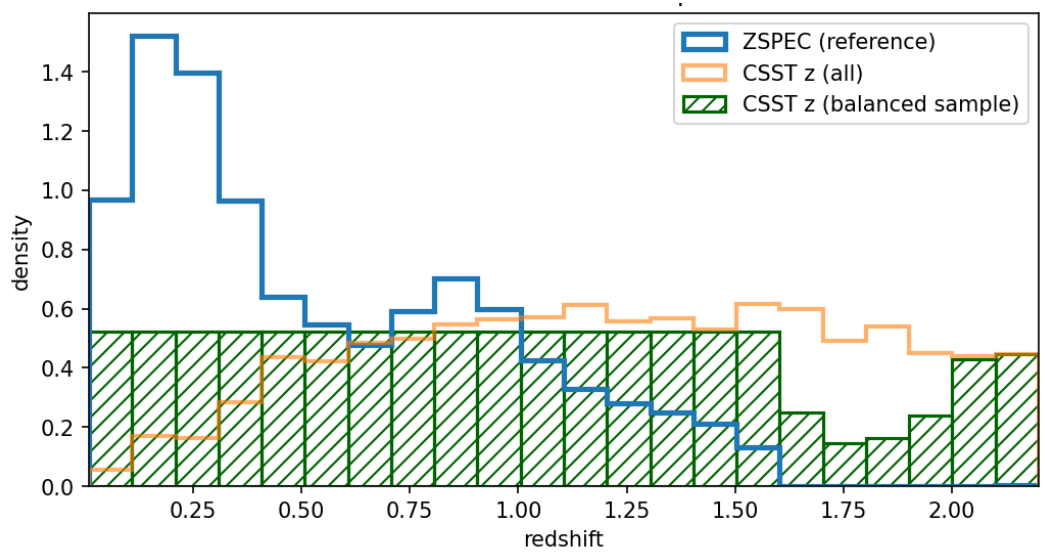




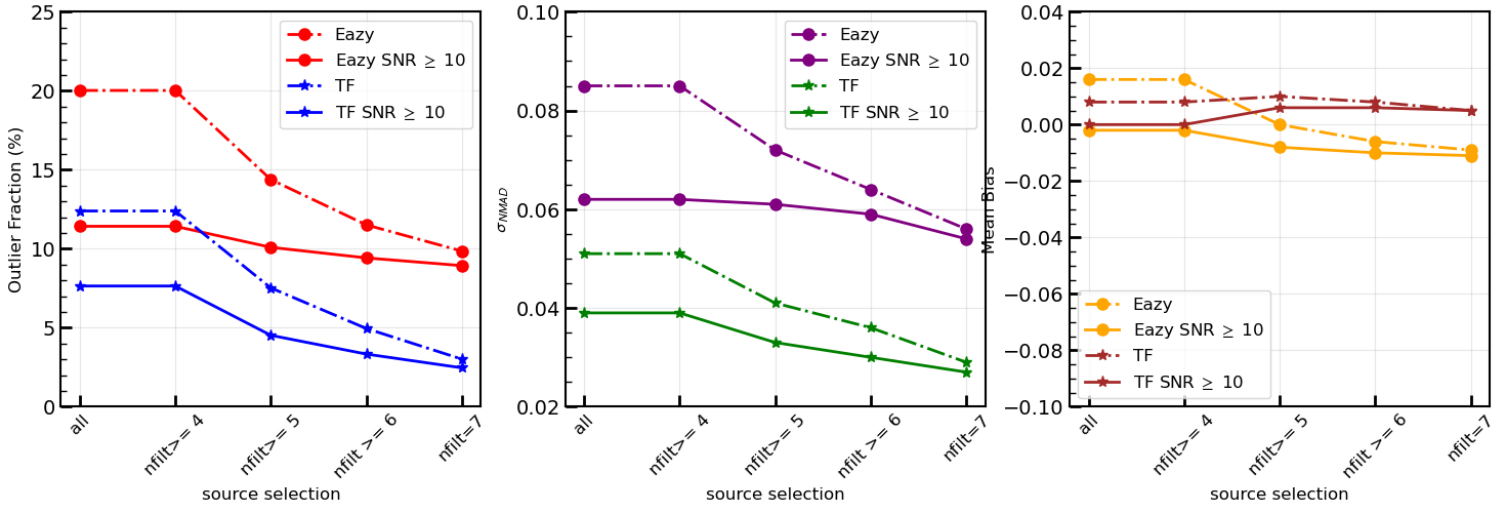


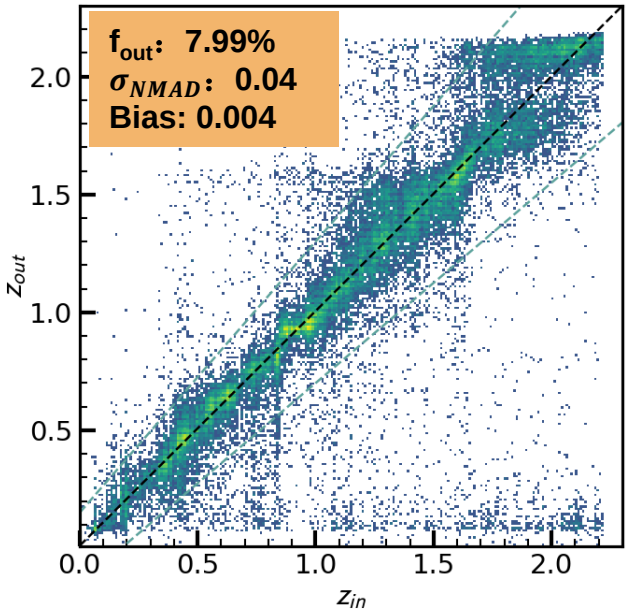


Resampling strategy



- **Simple balanced sampling:** retain rare high-redshift samples while downsampling the abundant low-redshift samples.
- **Epoch-wise balanced sampling**: perform balanced sampling at each training epoch.





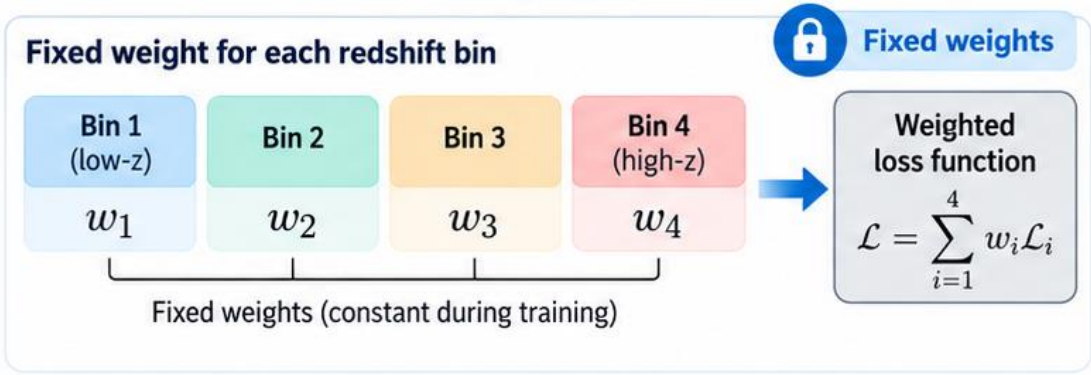
Loss weighting strategy

- **Static Loss Weighting**: Assign fixed sample weights based on redshift bins.

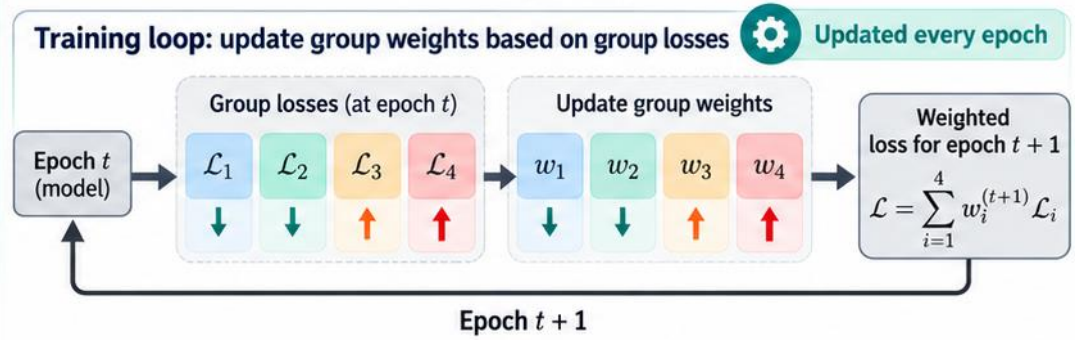
$$p_{target}(z) = const \qquad w(z) \propto \frac{1}{p_{train}(z)}$$

To account for the long-tailed redshift distribution of the training set, we adopt soft-uniform weighting and weight clipping.

$$w(z) = \left[\frac{1}{\hat{p}_{train} + \epsilon}\right]^{\alpha} \qquad w_i \leftarrow clip(w_i, w_{min}, w_{max},) \qquad L = \frac{\sum w_i L_i}{\sum L_i}$$



- **Dynamic Loss Weighting:** Divide training samples into redshift groups. Dynamically adjust group weights based on their losses at each epoch, encouraging the model to focus on poorly predicted redshift ranges.

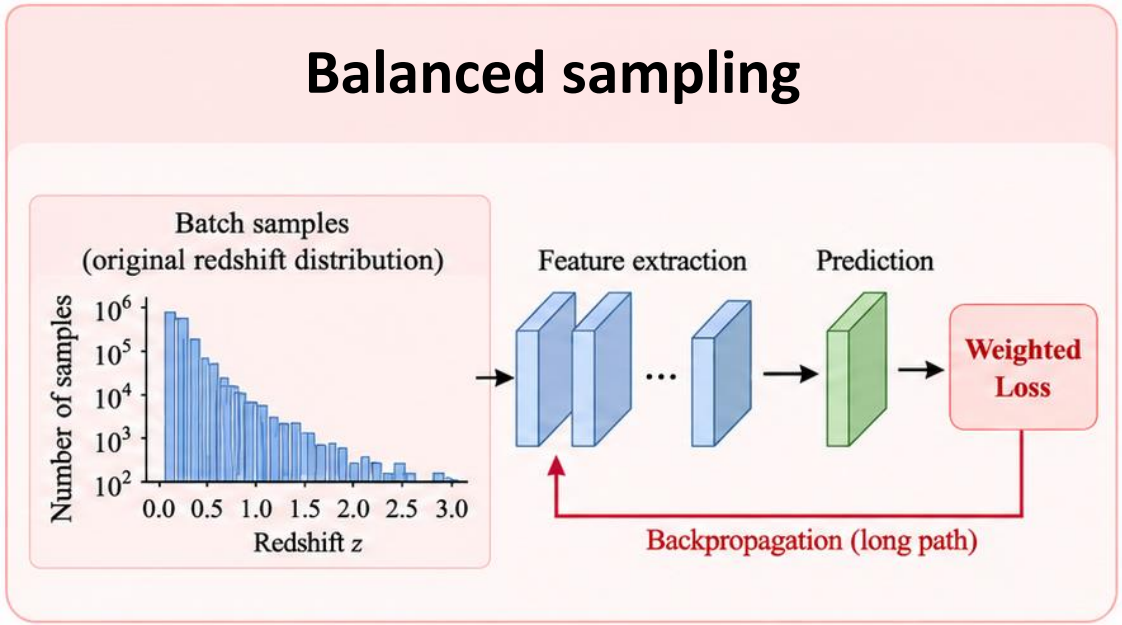


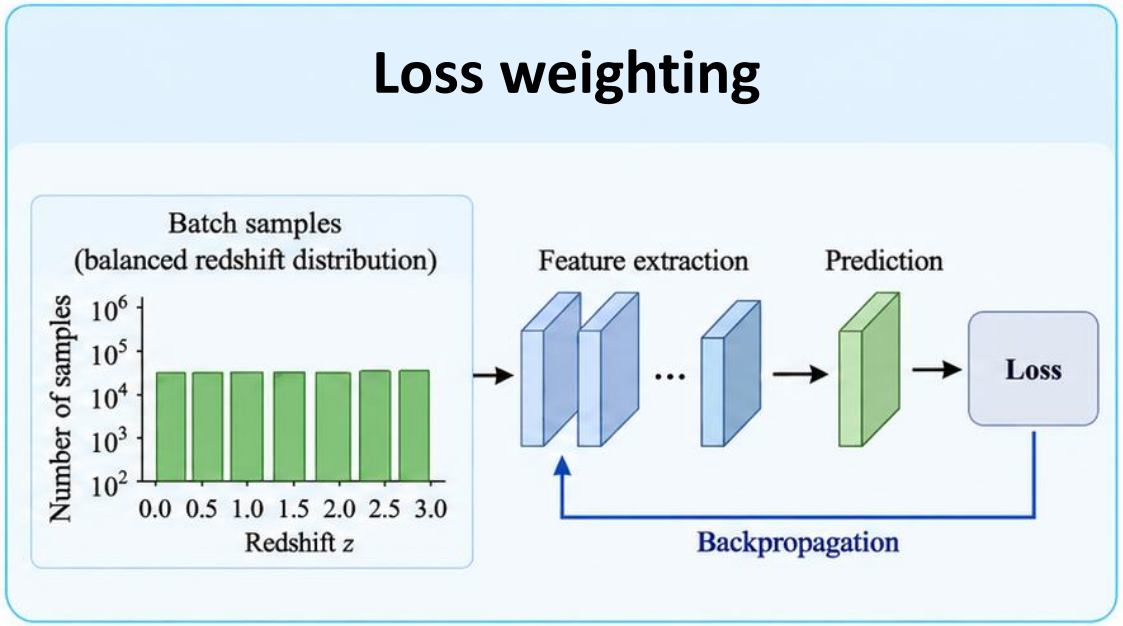
Some hyperparameter tuning results

	Outlier fraction	Sigma_NMAD	Bias_mean	Bias_median
Simple balanced sampling	2.46%	0.027	0.005	0.004
Epoch-wise sampler	2.90%	0.033	0.000	-0.002
$\alpha = 1.0, \epsilon = 0.001, \mathrm{w_{min}} = 0.2, \mathrm{w_{max}} = 5.0$	3.19%	0.033	-0.006	-0.002
$\alpha = 0.5, \epsilon = 0.001, \mathrm{w_{min}} = 0.2, \mathrm{w_{max}} = 5.0$	3.31%	0.031	-0.002	0.000
$\alpha = 0.25, \epsilon = 0.001, \mathrm{w_{min}} = 0.2, \mathrm{w_{max}} = 5.0$	3.27%	0.035	0.007	0.003
$\alpha = 1.0, \epsilon = 0.0, \mathrm{w_{min}} = 0.2, \mathrm{w_{max}} = 5.0$	4.43%	0.033	-0.014	-0.007
$\alpha = 0.25, \epsilon = 0.001, \mathrm{w_{min}} = 0.2, \mathrm{w_{max}} = 5.0$ +Epoch sampler	3.45%	0.045	0.009	0.006

Balanced Sampling Outperform Loss Weighting

Balanced sampling v.s. Loss weighting





- Unchanged redshift distribution → Feature learning dominated by low-z galaxies.
- Loss reweighting provides limited correction to early feature extraction.

- Balanced redshift distribution in every batch → Less biased feature learning.
- Promotes representative feature extraction across all redshifts, improving generalization.

Conclusions & Future Perspectives

- Bridging the gap between idealized training conditions and real-world observations is essential for developing robust and reliable photometric redshift models.
- Further advances in deep learning require validation and optimization using more realistic observational training samples.
- Realistic missing-data patterns should be explicitly considered in the design of deep learning frameworks.
- Resampling strategies can effectively mitigate redshift distribution shifts and improve prediction accuracy.
- Other distribution shifts, such as magnitude distribution shifts, require further systematic investigation and targeted algorithmic improvements.
- Further validation is needed to identify which types of galaxies would yield the greatest improvement in photo-z accuracy when included in the spectroscopic training sample.

Thank you!