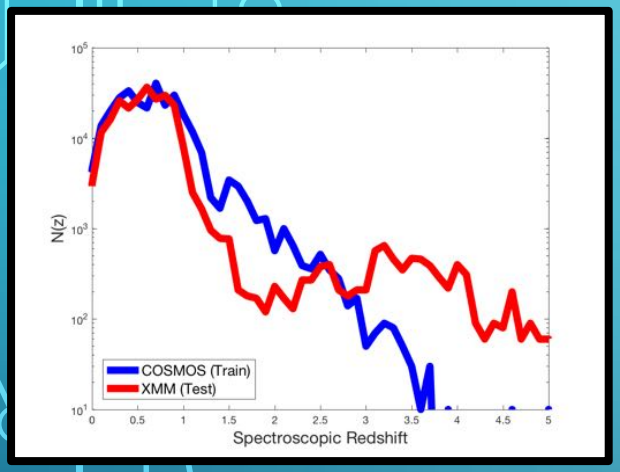
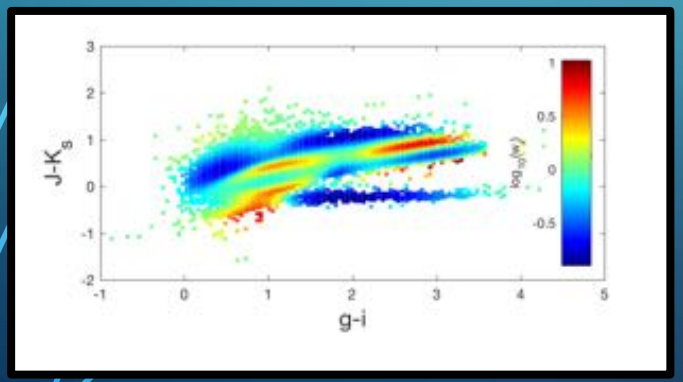
AUGMENTING MACHINE LEARNING PHOTOMETRIC REDSHIFTS WITH GAUSSIAN MIXTURE MODELS

PETER HATFIELD, MATT JARVIS, NATHAN ADAMS, REBECCA BOWER, IBRAHIM ALMOSALLAM, ZAHRA GOMES, STEPHEN ROBERTS, CORENTIN SCHREIBER (UNIVERSITY OF OXFORD)



Training (COSMOS) and test (XMM-LSS) data constructed to have different redshift (left) and colour (below) distributions

We used Gaussian Mixture Models to model the differences in colour-magnitude space

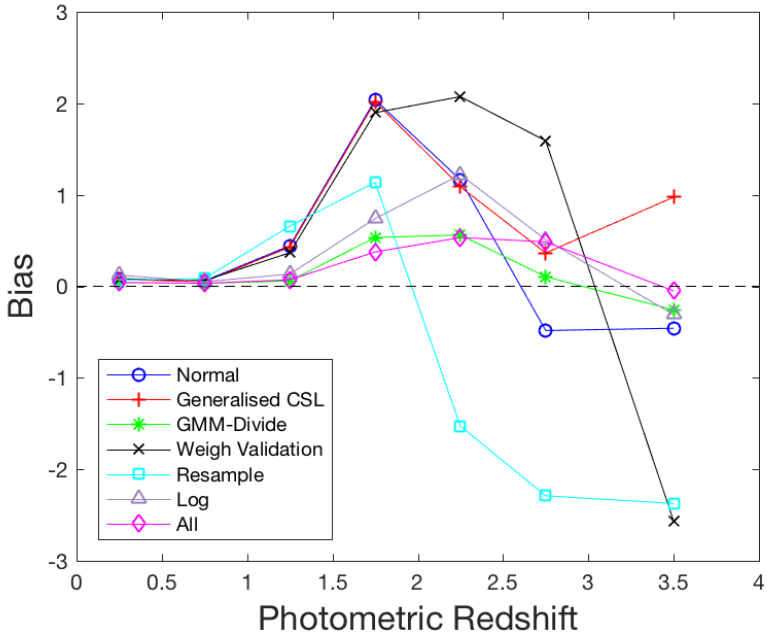


- Measuring the redshifts of large numbers (10^5-10^9) of galaxies is essential for many problems in galaxy physics and cosmology. Spectroscopic redshifts can unfortunately only be obtained for a small number of these galaxies; photometry must be used
- Photometric redshifts (photo-z) can be calculated in two main ways: a) fitting model galaxy spectra templates, or b) using machine learning (ML), trained on galaxies for which we do know the redshift
- GPz is a machine learning code for photo-z developed in Almosallam+2016, applied in Gomes+2018 , Duncan+2018
- We seek to develop ways to account for training and test data having different distributions in parameter space - traditionally challenging for ML methods
- In COSMOS and XMM-LSS uGRIZYJHK data we find bias on predictions can be moderately reduced with no additional data if the differences in the colour distributions are taken into account, Hatfield+2020

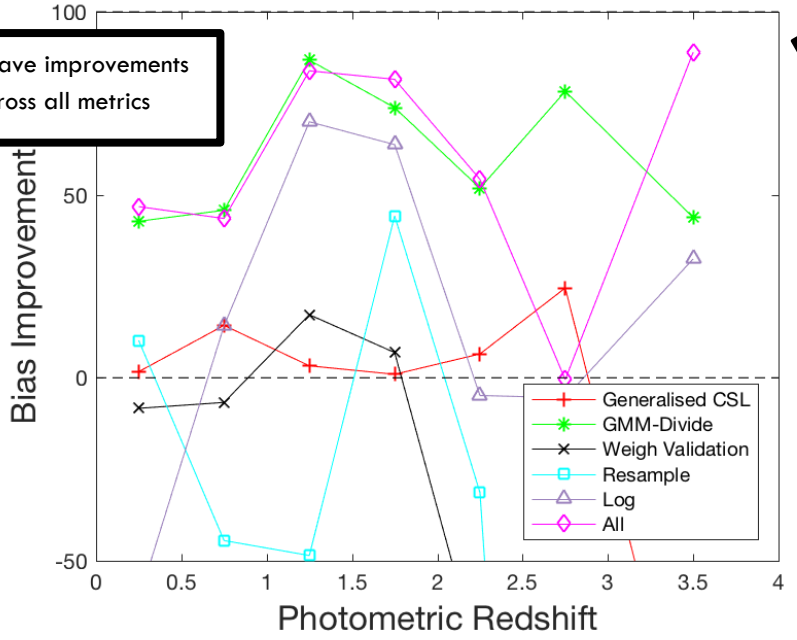
METHODS CONSIDERED & RESULTING IMPACT ON PHOTO-Z METRICS

Bias on predictions (mean prediction minus true redshift): zero means predictions are unbiased)

Improvement in bias (100%=complete removal of bias)



`All' gave improvements across all metrics



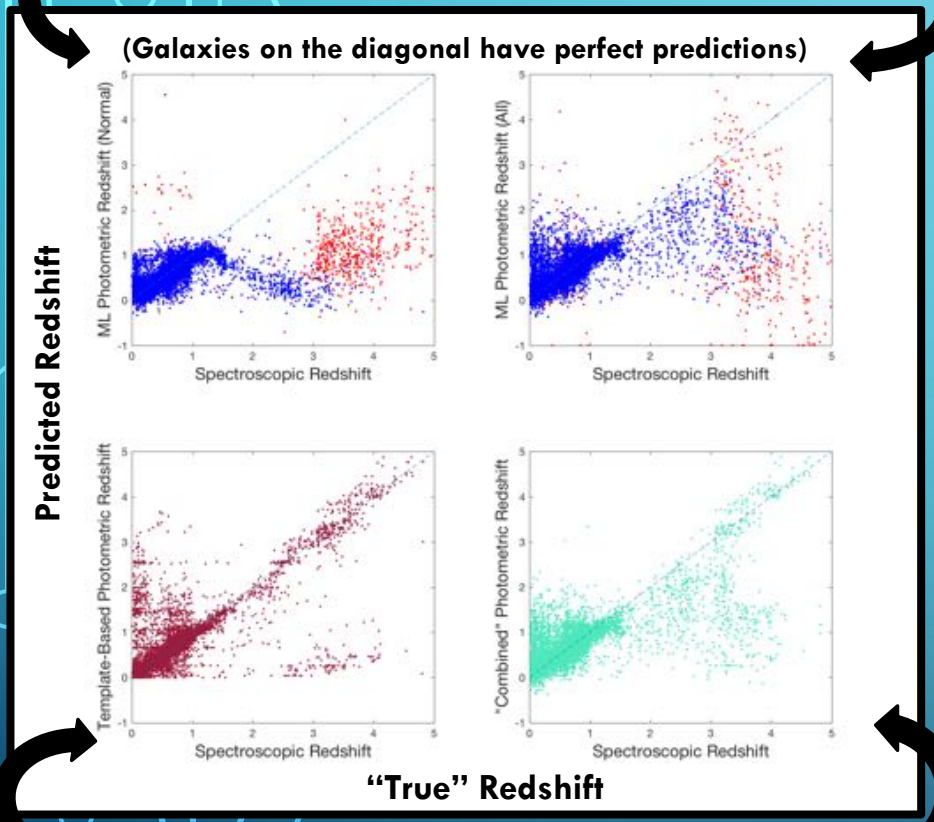
Methods Considered

- **Normal:** base use of GPz
- **GCSL:** Upweighting parts of colour space common in the test data but rare in the training data
- **GMM-Divide:** Using a GMM to divide parameter space into smaller segments in an unsupervised way, and then training on them separately
- **Weigh Validation:** Making the validation data look more like the test data
- **Resample:** Retrain the algorithm multiple times, each time resampling new photometry values based on the photometry uncertainty
- **Log:** Modelling log(z) rather than z
- **All:** Using Weigh Validation, Resample and GMM-Divide simultaneously

COMBINING ML AND TEMPLATE FITTING

Conventional Application of ML

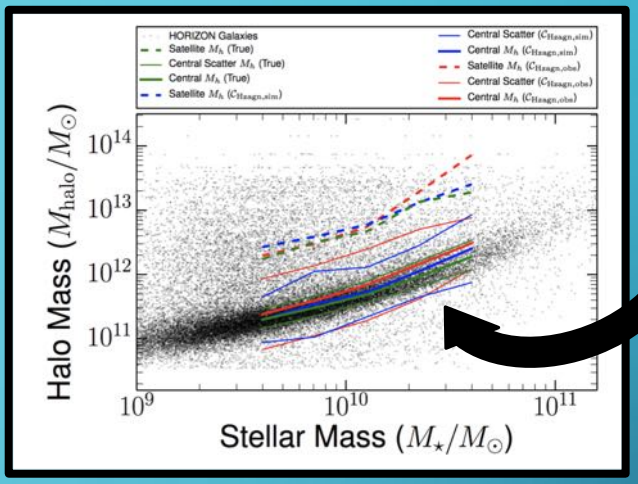
ML with the modifications described in this poster



Template based photo-z (from Adams+2020)

Hybrid combination of ML and Template methods – for overall better predictions

RELATED WORK



TRUE PHYSICS
ESTIMATE FROM CLUSTERING
ESTIMATE BASED ON PHOTO-Z

In Hatfield+2019 we looked at how measurements of the relationship between galaxies and their host dark matter halo are biased if photo-z are used rather than the true redshifts, for the hydrodynamical cosmological simulation Horizon-AGN

Have applied methods here to the Rubin DESC Tomography Challenge