Spectroscopic Completeness and Photometric Redshift Performance in Astronomical Foundation Models

Andrew Engel
Department of Physics
The Ohio State University
Columbus, OH, 43210
engel.250@osu.edu

Hailey Widger
Department of Physics & Astronomy
Embry-Riddle Aeronautical University
Prescott, AZ 86301
widgerh@my.erau.edu

Annika Peter
Department of Physics
The Ohio State University
Columbus, OH, 43210
peter.33@osu.edu

Peter L. Taylor
Department of Physics
The Ohio State University
Columbus, OH, 43210
taylor.4264@osu.edu

Abstract

Using a combination of self-organized maps (SOM) to perform prototype learning and astronomical foundation models to provide embeddings, we present an analysis of a new photometric redshift model for the DESI Legacy Survey footprint. Using the groupings learned by our SOM, we investigate the role local training-sample density plays into performance. The SOM can be used to flag samples that users request which lie outside the distribution of training data, or those examples which are known to belong to volumes of input data-space where the model underperforms. The proposed flag could help scientists better understand the performance of our model on their specific sample to make educated decisions tailored to their downstream analysis.

1 Introduction

The redshift of galaxies is a fundamental observable used to probe the nature of dark matter and dark energy. Due to the expansion of the universe, modern cosmology allows us to relate the redshift of photons to the approximate distance an object was from Earth when the light was emitted (S. Dodelson, 2003). Measuring redshifts/distances is critical, because it allows us to understand the context of object's environments (R. J. Foley & K. Mandel, 2013), how far back in time the object is being observed (P. J. E. Peebles, 1993), and map out correlation in galactic locations (DESI Collaboration et al., 2024). While measuring redshift using spectroscopy is the gold-standard, it is also expensive and can not scale to the number of galaxies needed for next generation analyses (J. A. Newman et al., 2015).

Photometric redshift algorithms address this scale issue by estimating galactic redshifts using photometry (astronomical images or derived data products) rather than spectroscopy. Photometry is a much faster measurement than spectroscopy, but unfortunately does not contain the spectral resolution necessary to observe individual absorption/emission lines to measure redshift directly (M. Salvato et al., 2019). Instead, photometric redshift algorithms model the expected flux received in large pass-bands from a template library of spectra (S. Arnouts & O. Ilbert, 2011; R. Feldmann et al., 2006; M. Bolzonella et al., 2011) or simply regress the redshift given the labeled data from past spectroscopic surveys using machine learning (A. Collister & O. Lahav, 2003; R. Beck et al., 2016; M. Bilicki et al., 2018; A. H. Wright et al., 2019; J. Pasquet et al., 2019).

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

One open issue is that the available spectroscopic data is incomplete and biased towards volumes of galactic feature space that are targeted by various surveys for their individual science goals (J. A. Newman et al., 2015). For the end-users of these photometric redshift models, we need to characterize the available data distribution to understand where models can and can not be trusted.

In this work we evaluate the performance of a photometric redshift model trained from the embeddings of a new astronomical foundation model AIONS (L. Parker et al., 2025), evaluated in bins of color space from a self-organized map (SOM) (T. Kohonen, 1982). Our primary motivation is to develop a flagging algorithm to alert users that the requested galaxy is either not well-represented in the training data or known to be similar to under-performing galaxies to enable scientists to make informed decisions for their own analyses.

2 Methods

Data Preparation. We combine spectroscopic data from an extensive list of surveys including the Dark Energy Spectroscopic Instrument (DESI) (DESI Collaboration et al., 2025) and the Sloan Digital Sky Survey (SDSS) (A. Almeida et al., 2023). We make cuts to select for high quality redshift measurements from spectra of galaxies, following standards from previous work(R. Beck et al., 2021; A. Engel et al., 2025). See appendix C for visualization of the redshift distribution from each survey. The total number of spectroscopically identified galaxies is around 16 million.

Using the sky-coordinates given by the spectroscopic surveys, we query the DESI Legacy Survey (A. Dey et al., 2019) cutout service for 42"-square (equivalent to 160-pix) image cutouts for g,r,i, and z-band photometry. Samples which do not have overlap with the DESI Legacy Survey footprint are removed, leaving approximately 12 million viable datapoints. We then cross-match these sky-coordinates with a 2" circular aperture to find nearest-neighbor match within the DESI Legacy Survey catalog, recording the flux in the optical and infrared, shape measurements, galactic extinction (D. J. Schlegel et al., 1998), and photometric redshift estimates from R. Zhou et al. (2023a). We additionally query for these same photometric catalog features of 18 million random galaxies from the DESI Legacy Survey catalog, which form the representative sample to train our SOM upon. To summarize, we have three data types: redshifts which come from spectroscopy, image cutouts, and pre-calculated features summarizing those images from astronomical catalogs.

Spectroscopic Completeness and Self Organized Maps. This work seeks to understand the set of spectroscopically observed galaxies in the redshift region $z < 1.6$ in relation to a random, fiducial sample of galaxies. We utilize a self organized map (SOM), to perform prototype learning. See section 3.1 of (A. Campos et al., 2024) for a quick introduction to the algorithm; pertinent to our discussion here, SOMs learn a grid of prototypes while also maintaining a sense of similarity between neighboring cells on that grid. SOMs have been used to understand spectroscopic completeness for weak lensing samples (D. Masters et al., 2015), and to understand performance local to a set of relatively bright galaxies (E. R. Moran et al., 2025). We train our SOM on the catalog-level photometry for samples in the DR10-south (those including i-band photometry) rather than the features we could extract from the images using AIONS in this version of the work. At time of writing, we are working towards downloading and preparing a full-image dataset of random galaxies which could be used to train the SOM on the feature-vectors from AIONS directly, but was not yet complete. Part of our choice to use SOMs for this analysis is their familiarity to the target audience of astronomers; evaluating other choices for similarity or anomaly search is beyond our scope. Studying the relationship between sample-density and photometric redshift estimation performance is novel, and is our main focus.

Astronomical Foundation Models & Downstream Training. Computer vision astronomical data foundation models have become available in the last four years, embedding images to high dimensional latent spaces (M. Walmsley et al., 2024; M. J. Smith et al., 2024; M. A. Hayat et al., 2021). Such models are typically pretrained under self-supervision. In this work we evaluate the AIONS-small pretrained foundation model (300 million parameters). AIONS utilizes the 4M multi-modal architecture (D. Mizrahi et al., 2023), which allows it to selectively include whatever observables are available for any particular object. In the DESI Legacy Survey DR10, i-band images are not available for the Northern sky. This would typically mean one would need to train individual models to handle the change in available data (or deal with a large amount of missing data). The

flexibility in AIONS means that we can supply the same model with g, r, z band images for the northern sky, and g, r, i, z band images for the southern sky without a change of architecture or retraining.

While the AIONS research paper reports photometric redshift performance, that work did not report the performance using the standard metrics used in this field, making it difficult to understand the performance of that model in the context of well known models. As an additional contribution, we report the performance of the AIONS model on this task and compare against with the DESI Legacy Survey's photometric redshift model (R. Zhou et al., 2023a). The DESI Legacy Survey photometric redshift model is a Random Forest trained on catalog-level features derived from the images, so the relevance of this comparison is to answer whether the additional complication of having an image-to-redshift pipeline is worth the effort.

Foundation models are used to create rich embeddings for downstream models. In this work, we train a 4-layer densely connected neural network to map from the space of embeddings to a 200 width output layer representing a coefficient expansion for a mixture of evenly spaced (and appropriately truncated) Gaussians with static standard deviation 0.0037. We interpret the resulting mixture of Gaussians as the network's estimate of the conditional density of redshift given the imaging data $P(\hat{z}|X)$. Our network is trained to minimize the continuous ranked probability score (J. E. Matheson & R. L. Winkler, 1976), which can be evaluated through the interpretation of our model's output as $P(\hat{z}|X)$. We holdout 10% of the data for hyperparameter tuning and an additional 10% for final performance measurement.

Metrics for Performance. The photometric redshift community has established guidelines for the evaluation of models that we adopt in this work (S. Schmidt et al., 2023). Due to the variety in analyses that rely on photometric redshifts, there is no one specific metric that determines what the "best" photometric redshift model is (S. J. Schmidt et al., 2020). Commonly used metrics quantify the scatter of the residuals using the median absolute deviation (MAD), the bias of the residuals, and the number of catastrophic outliers, η as the percentage of residuals with value greater than 0.15. We define $\hat{z}_i$ to be the estimate of redshift z_i for the i-th galaxy. Then the scaled residual is $\Delta z_i = \frac{\hat{z}_i - z_i}{1+z_i}$. It is common practice to use this scaled residual to account for the expected dependence of performance on redshift. The MAD is then computed as $1.4826 \times \mathrm{MED}(|\Delta z_i|)$, where MED denotes the median over galactic index i, and the scaling term makes MAD have value 1 in expectation for data sampled from the unit normal distribution. Bias is computed as $\langle \Delta z_i \rangle$, and catastrophic outliers $\eta = \frac{|\{\Delta z_i > 0.15\}|}{|\{\Delta z_i\}|}$.

Since our model outputs an estimate of the conditional density estimate $P(\hat{z}|X)$, we also quantify the number of probabilistic catastrophic outliers (PCOs) as those whose true redshift falls outside the inner 99% confidence region of $P(\hat{z}|X)$. For properly calibrated models, this value should of course be 1% of all datapoints. However, it is of interest to observe whether the probabilistic calibration of the model is dependent on the region of color-space.

3 Results

Performance of Foundation Models on Photometric Redshift Estimation Task. We trained a mixture density network using a continuous ranked probability score loss to learn the mapping from galaxy photometry to an estimate of the conditional probability density of redshift, $P(\hat{z}|X)$. These probability densities can be used quantify the uncertainty, can represent multi-modality, and can be used to estimate a photometric redshift point prediction, $\hat{z} = \langle P(z|X) \rangle$. We report the performance of our model in Table 3, and compare to the official DESI Legacy Survey DR10 photometric redshift model on a matched subset of our test-dataset. We further visualize the performance using predicted vs observed value plots, and visualize the calibration of our model using a probability integral transform plot. Both are available in appendix D. Our model is better calibrated (evaluated on the ensemble of datapoints), has a lower bias, less catastrophic outliers, but does not match the performance of the simple tabular model of R. Zhou et al. (2023a) in terms of scatter (MAD).

Is Local Density Predictive of Performance? Using the SOM we trained on a random selection of DESI LS DR10 galaxies, we evaluate whether the local density of available training data improves the performance of the model. In Figure 1 we demonstrate that scatter, η, and PCOs do trend downward

Model	MAD	Bias	η (%)	PCO (%)	R^2
AIONS MDN	0.029	-0.006	3.90%	1.40%	0.88
Zhou et. al, 2019	0.024	-0.013	4.91%	2.15%	0.84

Table 1: Point-performance measures of our mixture density network (ADN) trained on AIONS embeddings compared with the official DESI Legacy Survey photometric redshift model on a matched subset of our test dataset. Despite the increased complexity of the AIONS pipeline, e.g., its masked-pretraining and computer vision capabilities, we achieve only modest performance gains in Bias, catastrophic outlier rate, and in fact do worse in scatter (MAD) when evaluated globally. Our R^2 value is lower than what the AIONS authors report for their photometric redshift estimation performance; however, this is expected due to the change in population of galaxies we choose to include in this work.

with increased training samples. We also identify cells that have a statistically significant deviation from the performance measure evaluated over the entire dataset. A plot of the SOM cells evaluated on the test dataset for each performance measure is available in the appendix E, allowing one to visually see the connection between color and different aspects of the dataset. Higher count cells tend to be those with lower redshift. While we accounted for the expected heterogeneity of performance by scaling our residuals with $\frac{1}{1+z_i}$, our the model's performance and uncertainty remains dependent upon the redshift (i.e., remains heteroskedastic). We discuss the implications of this result in the following discussion.

4 Discussion

Flagging Under Performing cells to Users. Our primary goal was to determine whether SOMs could be used flag under-performing cells to users of photometric redshift algorithms. Given the statistically significant deviation from global performance measured in cells, we have determined there is meaningful information to be gained in such a flagging-system. We visualize examples from 16 different cells that have statistically significant and large in magnitude bias away from the global average value of bias in our residuals in appendix 2, demonstrating how these flags could be used to identify under-performing sub-spaces. Continuing work would study how users could set thresholds for performance depending on the needs of their own analysis.

5 Acknowledgments

A.W.E. acknowledges support from the Ohio State University Dean's Distinguished Fellowship Award. H.W. participation in this research was supported by the National Science Foundation under award #2447734 and The Ohio State University. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or The Ohio State University. The authors thanks the three anonymous reviewers who helped provide constructive criticism on the draft. We thank the L. Parker, F. Lanusse, and the remaining developers of AIONS for releasing their model early to our team. A continued acknowledgment for the source of our data and facilities follows in the appendix.

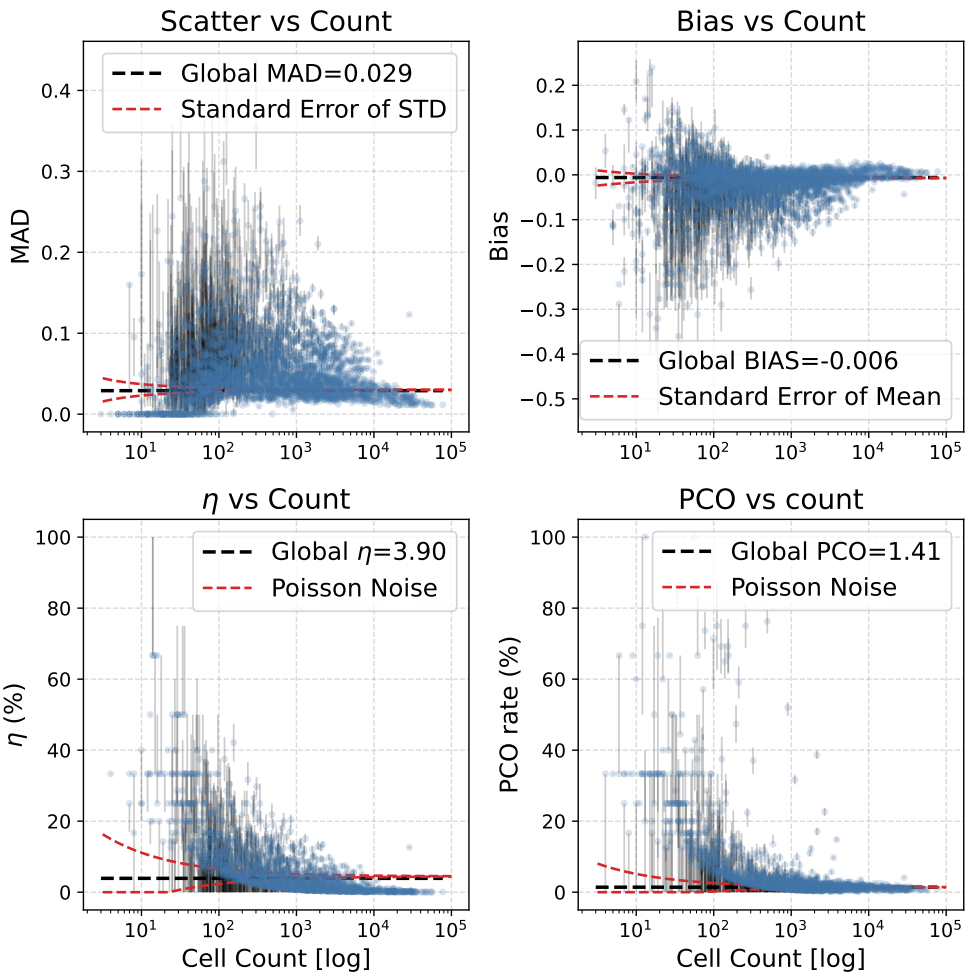

Figure 1: We plot performance measures against the cell's occupancy of training data (count). We estimate error bars from our sample using a bootstrap and plot the horizontal line representing the performance measure estimated from the entire test set. The red dashed lines are the 1-σ confidence interval representing the statistical noise expected from calculating the global-value of the metric from a sample of size equal to the count. The MAD, η and PCOs slightly trend to better scores with increasing training size, but there remain outliers of cells with statistically significant under-performance even at high training data count. This demonstrates that performance does generally trend with increased training density, albeit slowly. In this image we have masked out cells with no outliers for clarity of the general trend, see Figure 8 for a version with these points included.

References

Aihara, H., Allende Prieto, C., An, D., et al. 2011, The Eighth Data Release of the Sloan Digital Sky Survey: First Data from SDSS-III, The Astrophysical Journal Supplement, 193, 29, doi: 10.1088/0067-0049/193/2/29

Almeida, A., Anderson, S. F., Argudo-Fernández, M., et al. 2023, The Eighteenth Data Release of the Sloan Digital Sky Surveys: Targeting and First Spectra from SDSS-V, The Astrophysical Journal Supplement, 267, 44, doi: 10.3847/1538-4365/acda98

Arnouts, S., & Ilbert, O. 2011, LePHARE: Photometric Analysis for Redshift Estimate,, Astrophysics Source Code Library, record ascl:1108.009

Beck, R., Dobos, L., Budavári, T., Szalay, A. S., & Csabai, I. 2016, Photometric redshifts for the SDSS Data Release 12, MNRAS, 460, 1371, doi: 10.1093/mnras/stw1009

Beck, R., Szapudi, I., Flewelling, H., et al. 2021, PS1-STRM: neural network source classification and photometric redshift catalogue for PS1 3π DR1, Monthly Notices of the Royal Astronomical Society, 500, 1633, doi: 10.1093/mnras/staa2587

Bilicki, M., Hoekstra, H., Brown, M. J. I., et al. 2018, Photometric redshifts for the Kilo-Degree Survey. Machine-learning analysis with artificial neural networks, Astronomy and Astrophysics, 616, A69, doi: 10.1051/0004-6361/201731942

Bolzonella, M., Miralles, J.-M., & Pelló, R. 2011, Hyperz: Photometric Redshift Code,, Astrophysics Source Code Library, record ascl:1108.010

Campos, A., Yin, B., Dodelson, S., et al. 2024, Enhancing weak lensing redshift distribution characterization by optimizing the Dark Energy Survey Self-Organizing Map Photo-z method, arXiv e-prints, arXiv:2408.00922, doi: 10.48550/arXiv.2408.00922

Carlsten, S. G., Greene, J. E., Beaton, R. L., Danieli, S., & Greco, J. P. 2022, The Exploration of Local VolumE Satellites (ELVES) Survey: A Nearly Volume-limited Sample of Nearby Dwarf Satellite Systems, The Astrophysical Journal, 933, 47, doi: 10.3847/1538-4357/ac6fd7

Coil, A. L., Blanton, M. R., Burles, S. M., et al. 2011, The PRIsm MUlti-object Survey (PRIMUS). I. Survey Overview and Characteristics, The Astrophysical Journal, 741, 8, doi: 10.1088/0004-637X/741/1/8

Colless, M., Peterson, B. A., Jackson, C., et al. 2003, The 2dF Galaxy Redshift Survey: Final Data Release, arXiv e-prints, astro, doi: 10.48550/arXiv.astro-ph/0306581

Collister, A., & Lahav, O. 2003, ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks, Publications of the Astronomical Society of the Pacific, 116, 345 . https://api.semanticscholar.org/CorpusID:119089041

Cooper, M. C., Aird, J. A., Coil, A. L., et al. 2011, The DEEP3 Galaxy Redshift Survey: Keck/DEIMOS Spectroscopy in the GOODS-N Field, Astrophysical Journal Supplement Series, 193, 14, doi: 10.1088/0067-0049/193/1/14

DESI Collaboration, Adame, A. G., Aguilar, J., et al. 2024, DESI 2024 V: Full-Shape Galaxy Clustering from Galaxies and Quasars, arXiv e-prints, arXiv:2411.12021, doi: 10.48550/arXiv.2411.12021

DESI Collaboration, Abdul-Karim, M., Adame, A. G., et al. 2025, Data Release 1 of the Dark Energy Spectroscopic Instrument, arXiv e-prints, arXiv:2503.14745, doi: 10.48550/arXiv.2503.14745

Dey, A., Schlegel, D. J., Lang, D., et al. 2019, Overview of the DESI Legacy Imaging Surveys, The Astronomical Journal, 157, 168, doi: 10.3847/1538-3881/ab089d

Dodelson, S. 2003, Modern Cosmology (Academic Press), 33–37

Engel, A., Byler, N., Tsou, A., et al. 2025, Mantis Shrimp: Exploring Photometric Band Utilization in Computer Vision Networks for Photometric Redshift Estimation, arXiv e-prints, arXiv:2501.09112, doi: 10.48550/arXiv.2501.09112

Euclid Collaboration, Desprez, G., Paltani, S., et al. 2020, Euclid preparation - X. The Euclid photometric-redshift challenge, A&A, 644, A31, doi: 10.1051/0004-6361/202039403

Feldmann, R., Carollo, C. M., Porciani, C., et al. 2006, The Zurich Extragalactic Bayesian Redshift Analyzer and its first application: COSMOS, Monthly Notices of the Royal Astronomical Society, 372, 565, doi: 10.1111/j.1365-2966.2006.10930.x

Foley, R. J., & Mandel, K. 2013, Classifying Supernovae Using Only Galaxy Data, The Astrophysical Journal, 778, 167, doi: 10.1088/0004-637X/778/2/167

Hahn, C., Wilson, M. J., Ruiz-Macias, O., et al. 2023, The DESI Bright Galaxy Survey: Final Target Selection, Design, and Validation, The Astronomical Journal, 165, 253, doi: 10.3847/1538-3881/accff8

Hayat, M. A., Stein, G., Harrington, P., Lukić, Z., & Mustafa, M. 2021, Self-supervised Representation Learning for Astronomical Images, The Astrophysical Journall, 911, L33, doi: 10.3847/2041-8213/abf2c7

Jones, D. H., Read, M. A., Saunders, W., et al. 2009, The 6dF Galaxy Survey: final redshift release (DR3) and southern large-scale structures, Monthly Notices of the Royal Astronomical Society, 399, 683, doi: 10.1111/j.1365-2966.2009.15338.x

Kohonen, T. 1982, Self-Organized Formation of Topologically Correct Feature Maps, Biological Cybernetics, 43, 59

Le Fèvre, O., Cassata, P., Cucciati, O., et al. 2013, The VIMOS VLT Deep Survey final data release: a spectroscopic sample of 35,016 galaxies and AGN out to z 6.7 selected with 17.5 <= i <= 24.75, Astronomy and Astrophysics, 559, A14, doi: 10.1051/0004-6361/201322179

Lidman, C., Tucker, B. E., Davis, T. M., et al. 2020, OzDES multi-object fibre spectroscopy for the Dark Energy Survey: results and second data release, Monthly Notices of the Royal Astronomical Society, 496, 19, doi: 10.1093/mnras/staa1341

Liske, J., Baldry, I. K., Driver, S. P., et al. 2015, Galaxy And Mass Assembly (GAMA): end of survey report and data release 2, Monthly Notices of the Royal Astronomical Society, 452, 2087, doi: 10.1093/mnras/stv1436

Mao, Y.-Y., Geha, M., Wechsler, R. H., et al. 2024, The SAGA Survey. III. A Census of 101 Satellite Systems around Milky Way–mass Galaxies, The Astrophysical Journal, 976, 117, doi: 10.3847/1538-4357/ad64c4

Masters, D., Capak, P., Stern, D., et al. 2015, MAPPING THE GALAXY COLOR–REDSHIFT RELATION: OPTIMAL PHOTOMETRIC REDSHIFT CALIBRATION STRATEGIES FOR COSMOLOGY SURVEYS, The Astrophysical Journal, 813, 53, doi: 10.1088/0004-637X/813/1/53

Matheson, J. E., & Winkler, R. L. 1976, Scoring rules for continuous probability distributions, Manage. Sci., 22, 1087

Mentuch Cooper, E., Gebhardt, K., Davis, D., et al. 2023, HETDEX Public Source Catalog 1: 220 K Sources Including Over 50 K Lyα Emitters from an Untargeted Wide-area Spectroscopic Survey, The Astrophysical Journal, 943, 177, doi: 10.3847/1538-4357/aca962

Mizrahi, D., Bachmann, R., Kar, O. F., et al. 2023, 4M: Massively Multimodal Masked Modeling, arXiv e-prints, arXiv:2312.06647, doi: 10.48550/arXiv.2312.06647

Moran, E. R., Andrews, B. H., Newman, J. A., & Dey, B. 2025, Deep Learning Improves Photometric Redshifts in All Regions of Color Space, arXiv e-prints, arXiv:2507.06299, doi: 10.48550/arXiv.2507.06299

Newman, J. A., Cooper, M. C., Davis, M., et al. 2013, THE DEEP2 GALAXY REDSHIFT SURVEY: DESIGN, OBSERVATIONS, DATA REDUCTION, AND REDSHIFTS*, The Astrophysical Journal Supplement Series, 208, 5, doi: 10.1088/0067-0049/208/1/5

Newman, J. A., Abate, A., Abdalla, F. B., et al. 2015, Spectroscopic needs for imaging dark energy experiments, Astroparticle Physics, 63, 81, doi: 10.1016/j.astropartphys.2014.06.007

Parker, L., Lanusse, F., Shen, J., et al. 2025, AION-1: Omnimodal Foundation Model for Astronomical Sciences, arXiv e-prints, arXiv:2510.17960, doi: 10.48550/arXiv.2510.17960

Parkinson, D., Riemer-Sørensen, S., Blake, C., et al. 2012, The WiggleZ Dark Energy Survey: Final data release and cosmological results, Physics Review Letters D, 86, 103518, doi: 10.1103/PhysRevD.86.103518

Pasquet, J., Bertin, E., Treyer, M., Arnouts, S., & Fouchez, D. 2019, Photometric redshifts from SDSS images using a convolutional neural network, Astr & Astro, 621, A26, doi: 10.1051/0004-6361/201833617

Peebles, P. J. E. 1993, Principles of Physical Cosmology (Princeton University Press), doi: 10.1515/9780691206721

Raichoor, A., Moustakas, J., Newman, J. A., et al. 2023, Target Selection and Validation of DESI Emission Line Galaxies, The Astronomical Journal, 165, 126, doi: 10.3847/1538-3881/acb213

Reid, B., Ho, S., Padmanabhan, N., et al. 2016, SDSS-III Baryon Oscillation Spectroscopic Survey Data Release 12: galaxy target selection and large-scale structure catalogues, Monthly Notices of the Royal Astronomical Society, 455, 1553, doi: 10.1093/mnras/stv2382

Salvato, M., Ilbert, O., & Hoyle, B. 2019, The many flavours of photometric redshifts, Nature Astronomy, 3, 212, doi: 10.1038/s41550-018-0478-0

Schlegel, D. J., Finkbeiner, D. P., & Davis, M. 1998, Maps of Dust Infrared Emission for Use in Estimation of Reddening and Cosmic Microwave Background Radiation Foregrounds, The Astrophysical Journal, 500, 525, doi: 10.1086/305772

Schmidt, S., Gschwend, J., Crenshaw, J. F., et al. 2023, LSSTDESC/RAIL: v0.98.5, v0.98.5 Zenodo, doi: 10.5281/zenodo.7927358

Schmidt, S. J., Malz, A. I., Malz, A. I., et al. 2020, Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST), Monthly Notices of the Royal Astronomical Society. https://api.semanticscholar.org/CorpusID:224909363

Scodeggio, M., Guzzo, L., Garilli, B., et al. 2018, The VIMOS Public Extragalactic Redshift Survey (VIPERS). Full spectroscopic data and auxiliary information release (PDR-2), Astronomy and Astrophysics, 609, A84, doi: 10.1051/0004-6361/201630114

Smith, M. J., Roberts, R. J., Angeloudi, E., & Huertas-Company, M. 2024, AstroPT: Scaling Large Observation Models for Astronomy, arXiv e-prints, arXiv:2405.14930, doi: 10.48550/arXiv.2405.14930

Sohn, J., Geller, M. J., Hwang, H. S., et al. 2023, HectoMAP: The Complete Redshift Survey (Data Release 2), The Astrophysical Journal, 945, 94, doi: 10.3847/1538-4357/acb925

Walmsley, M., Bowles, M., Scaife, A. M. M., et al. 2024, Scaling Laws for Galaxy Images, arXiv e-prints, arXiv:2404.02973, doi: 10.48550/arXiv.2404.02973

Wright, A. H., Hildebrandt, H., Kuijken, K., et al. 2019, KiDS+VIKING-450: A new combined optical and near-infrared dataset for cosmology and astrophysics, Astronomy and Astrophysics, 632, A34, doi: 10.1051/0004-6361/201834879

Zaritsky, D., Donnerstein, R., Dey, A., et al. 2019, Systematically Measuring Ultra-diffuse Galaxies (SMUDGes). I. Survey Description and First Results in the Coma Galaxy Cluster and Environs, The Astrophysical Journal Supplement, 240, 1, doi: 10.3847/1538-4365/aaefe9

Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, LAMOST spectral survey — An overview, Research in Astronomy and Astrophysics, 12, 723, doi: 10.1088/1674-4527/12/7/002

Zhou, R., Ferraro, S., White, M., et al. 2023a, DESI luminous red galaxy samples for cross-correlations, Journal of Cosmology and Astroparticle Physics, 2023, 097, doi: 10.1088/1475-7516/2023/11/097

Zhou, R., Dey, B., Newman, J. A., et al. 2023b, Target Selection and Validation of DESI Luminous Red Galaxies, The Astronomical Journal, 165, 58, doi: 10.3847/1538-3881/aca5fb

A Data and Facility Acknowledgments

The Legacy Surveys consist of three individual and complementary projects: the Dark Energy Camera Legacy Survey (DECaLS; Proposal ID #2014B-0404; PIs: David Schlegel and Arjun Dey), the Beijing-Arizona Sky Survey (BASS; NOAO Prop. ID #2015A-0801; PIs: Zhou Xu and Xiaohui Fan), and the Mayall z-band Legacy Survey (MzLS; Prop. ID #2016A-0453; PI: Arjun Dey). DECaLS, BASS and MzLS together include data obtained, respectively, at the Blanco telescope, Cerro Tololo Inter-American Observatory, NSF's NOIRLab; the Bok telescope, Steward Observatory, University of Arizona; and the Mayall telescope, Kitt Peak National Observatory, NOIRLab. The Legacy Surveys project is honored to be permitted to conduct astronomical research on Iolkam Du'ag (Kitt Peak), a mountain with particular significance to the Tohono O'odham Nation. NOIRLab is operated by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation. This project used data obtained with the Dark Energy Camera (DECam), which was constructed by the Dark Energy Survey (DES) collaboration. Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo, Financiadora de Estudos e Projetos, Fundacao Carlos Chagas Filho de Amparo a Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Cientifico e Tecnologico and the Ministerio da Ciencia, Tecnologia e Inovacao, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey. The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energeticas, Medioambientales y Tecnologicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenossische Technische Hochschule (ETH) Zurich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciencies de l'Espai (IEEC/CSIC), the Institut de Fisica d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig Maximilians Universitat Munchen and the associated Excellence Cluster Universe, the University of Michigan, NSF's NOIRLab, the University of Nottingham, the Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, and Texas A&M University. BASS is a key project of the Telescope Access Program (TAP), which has been funded by the National Astronomical Observatories of China, the Chinese Academy of Sciences (the Strategic Priority Research Program "The Emergence of Cosmological Structures" Grant # XDB09000000), and the Special Fund for Astronomy from the Ministry of Finance. The BASS is also supported by the External Cooperation Program of Chinese Academy of Sciences (Grant # 114A11KYSB20160057), and Chinese National Natural Science Foundation (Grant # 11433005). The Legacy Survey team makes use of data products from the Near-Earth Object Wide-field Infrared Survey Explorer (NEOWISE), which is a project of the Jet Propulsion Laboratory/California Institute of Technology. NEOWISE is funded by the National Aeronautics and Space Administration. The Legacy Surveys imaging of the DESI footprint is supported by the Director, Office of Science, Office of High Energy Physics of the U.S. Department of Energy under Contract No. DE-AC02-05CH1123, by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility under the same contract; and by the U.S. National Science Foundation, Division of Astronomical Sciences under Contract No. AST-0950945 to NOAO.url: http://www.lamost.org/public/?locale=en Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is http://www.sdss.org/. The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the

Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington. Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/. SDSS-III is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS-III Collaboration including the University of Arizona, the Brazilian Participation Group, Brookhaven National Laboratory, University of Cambridge, Carnegie Mellon University, University of Florida, the French Participation Group, the German Participation Group, Harvard University, the Instituto de Astrofisica de Canarias, the Michigan State/Notre Dame/JINA Participation Group, Johns Hopkins University, Lawrence Berkeley National Laboratory, Max Planck Institute for Astrophysics, Max Planck Institute for Extraterrestrial Physics, New Mexico State University, New York University, Ohio State University, Pennsylvania State University, University of Portsmouth, Princeton University, the Spanish Participation Group, University of Tokyo, University of Utah, Vanderbilt University, University of Virginia, University of Washington, and Yale University. Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic Telescope LAMOST) is a National Major Scientific Project built by the Chinese Academy of Sciences. Funding for the project has been provided by the National Development and Reform Commission. LAMOST is operated and managed by the National Astronomical Observatories, Chinese Academy of Sciences. Funding for the DEEP2 Galaxy Redshift Survey has been provided by NSF grants AST-95-09298, AST-0071048, AST-0507428, and AST-0507483 as well as NASA LTSA grant NNG04GC89G. This publication also makes use of data products from NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology, funded by the Planetary Science Division of the National Aeronautics and Space Administration.

B Flagging Samples for Photometric Redshift Estimation.

In this appendix we include Figure 2, where we visualize samples from cells that our SOM method would flag.

C Spectroscopic Data Distribution

In Figure 3 we plot the distribution and provide a total count from each of the 20 different spectroscopic surveys used to assemble our training targets, including 2dF (M. Colless et al., 2003), 6dF (D. H. Jones et al., 2009), DESI DR1 (DESI Collaboration et al., 2025), ELVES (S. G. Carlsten et al., 2022), GAMA (J. Liske et al., 2015), HectoMAP (J. Sohn et al., 2023), HetDEX (E. Mentuch Cooper et al., 2023), LAMOST (G. Zhao et al., 2012), OzDES (C. Lidman et al., 2020), PRIMUS (A. L. Coil et al., 2011), SAGA (Y.-Y. Mao et al., 2024), SDSS (A. Almeida et al., 2023), SMUDGes (D. Zaritsky et al., 2019), VIPERS (M. Scodeggio et al., 2018), WiggleZ (D. Parkinson et al., 2012), Deep2/3 (J. A. Newman et al., 2013; M. C. Cooper et al., 2011), and VVDS (Le Fèvre, O. et al., 2013). We note that the total distribution is dominated by DESI and SDSS observations, which include components that are magnitude-limited down to $r < 19.5$ and $r < 17.77$, respectfully (H. Aihara et al., 2011; C. Hahn et al., 2023). Beyond these limits, the surveys become color-selected towards large red galaxies (B. Reid et al., 2016; R. Zhou et al., 2023b) and emission line galaxies (A. Raichoor et al., 2023).

D Expanded Visualizations of Photometric Redshift Performance

We include an expanded series of visualization to explore the performance of our photometric redshift model. In Figure 4, we visualize the $\hat{z}$ vs z plot using a Kernel Density Estimate to visualize the distribution of samples. Using our estimate of the conditional density estimate $P(\hat{z}|X)$ we plot a random selection of point estimates and their 90% confidence regions as error bars. In Figure 5, we plot the same KDE but overlay points that whose estimates 99% confidence interval exclude the true value, to get a sense of where these probabilistic catastrophic outliers (PCOs) lie in target space. Finally, in Figure 6, we plot the probability integral transform, which is simply a histogram of the cumulative density function evaluated to the true value across all of our conditional density estimates. If the model was perfectly calibrated, the PIT histogram would fall along the dashed horizontal line. Our model tends to be overconfident with large overabundance at the outermost

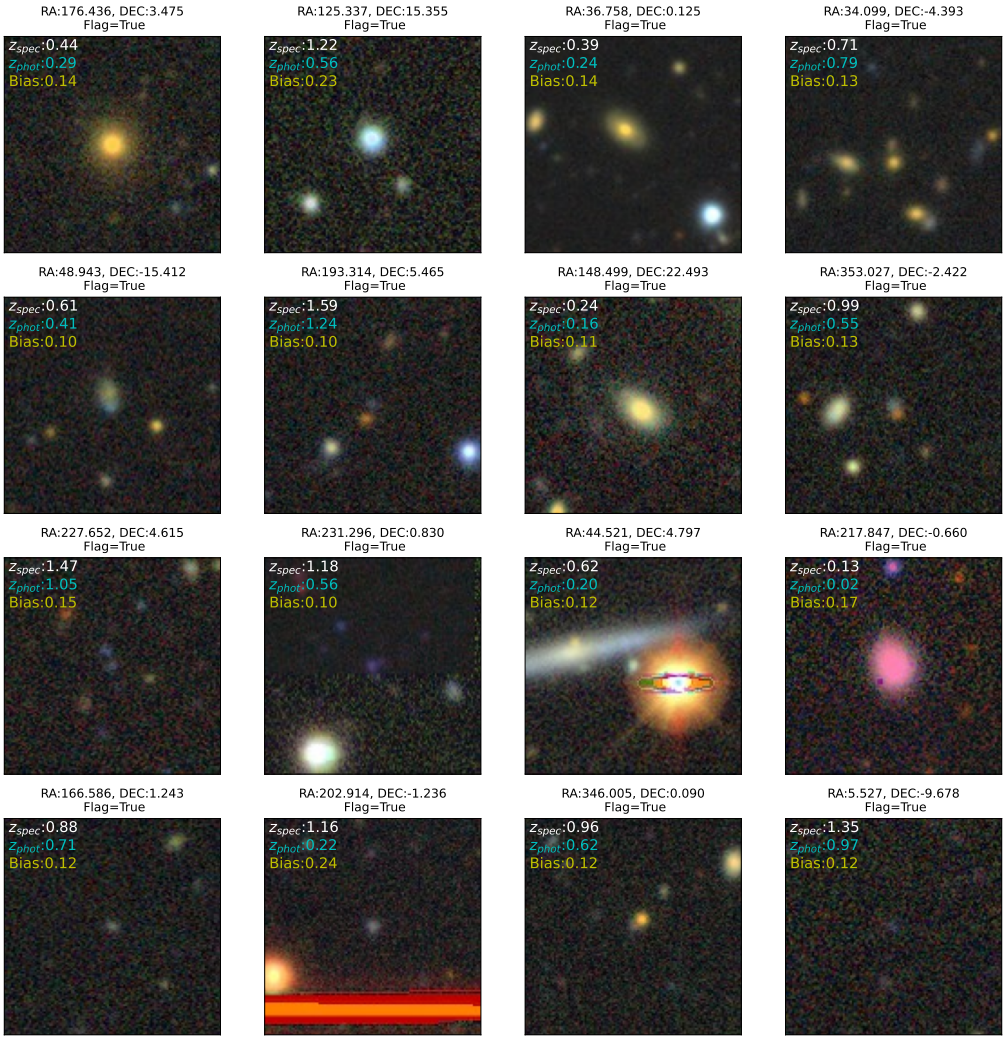

Figure 2: Each subplot shows a unique galaxy from our labeled dataset from a unique cell that has statistically significant deviation from the global bias value across all samples. We argue that feature space local to these cells is under performing the model's average, therefore, when a user requests a photometric redshift on a galaxy falling into one of these cells, that user should be alerted to the possibility of bad performance. If we take these examples to be indicative of the kinds of objects in their cells, we see that many cells may be astronomical blends which could explain their bad performance. However, some cells also seem to be of relatively bright and nearby objects, where one might expect the model to perform well. We leave these investigations to future work.

histogram bins, representing PCOs. This is common of many photometric redshift algorithms, see Euclid Collaboration et al. (2020).

E SOM visualization

In Figure 7 we visualize the SOM cells overlaid with the performance measures as well as the spectroscopic training data in each cell and the average redshift in the cell.

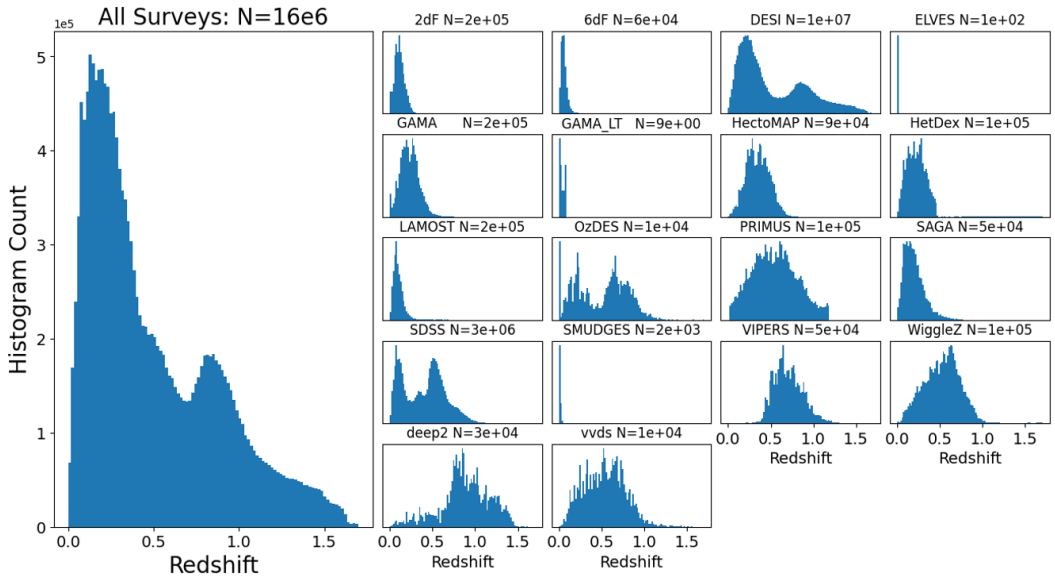

Figure 3: We combine spectroscopy from a broad range of surveys, visualizing the normalized distribution from each component survey in the right half of the plot, and the total combined distribution on the left side. In each subtitle, we list the survey name and N, the number of datapoints used from that survey. Actual usage in this work depends upon the availability of DESI LS DR10 data, which cuts our sample size down to 12e6 samples. Many of these surveys seek to characterize the distribution of galaxies to perform cosmology, so become color-selected to maximize science return. To some degree, these surveys and auxillary observations may be complimentary, providing spectra for different parts of the galactic feature space. The goal of this paper could be said to determine whether local to those sub-populations, can we trust photometric redshift models, despite the fact that the entire distribution is knowably biased towards specific color spaces.

F Performance Measures Figure with Zero Outlier Cells

This appendix contains Figure 8, which shows the main figure of the paper including datapoints excluded to simplify the visualization.

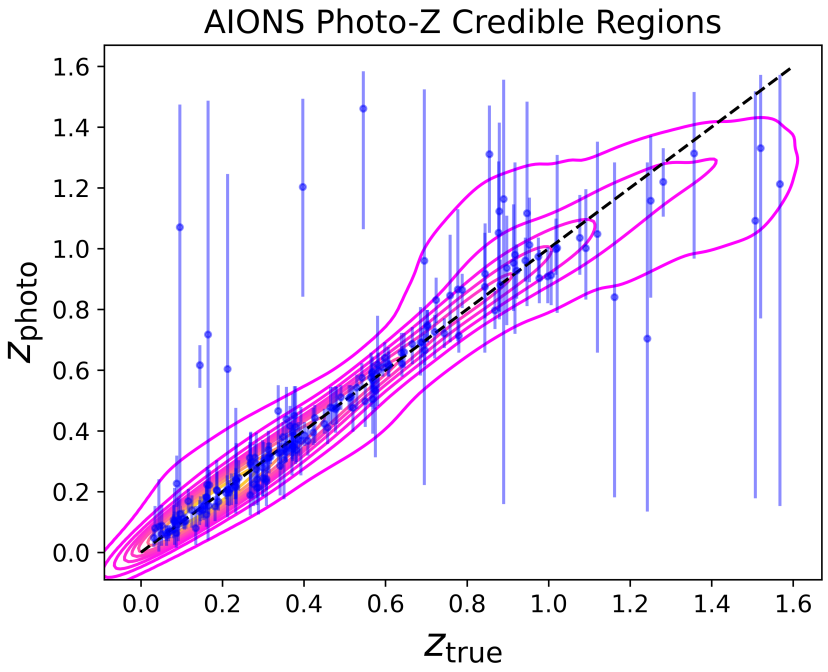

Figure 4: AIONS estimation vs observation plot using a gaussian kernel density to visualize the distribution of points. The y-error bars are sampled randomly and we visualize them as the inner 90% confidence interval from our conditional probability estimate.

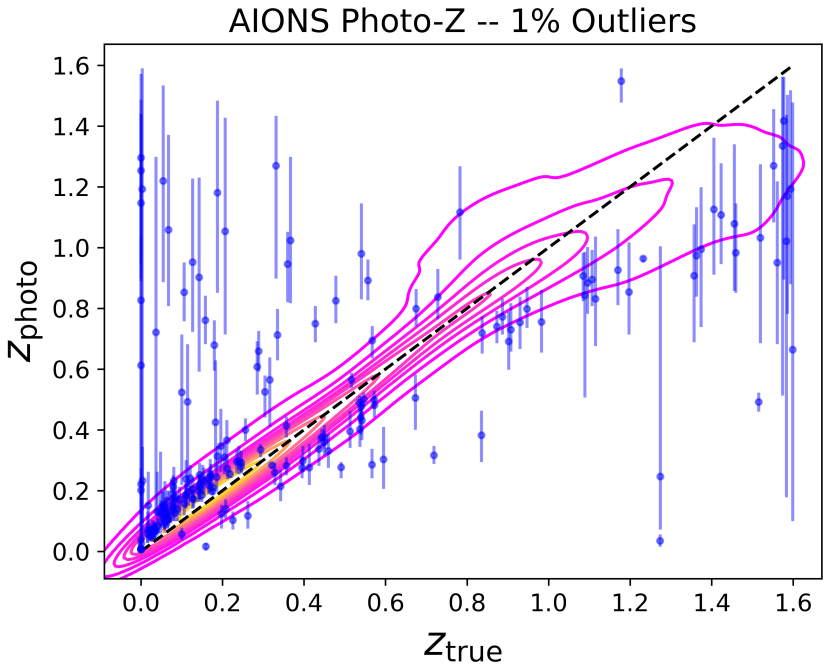

Figure 5: AIONS estimation vs observation plot using the same gaussian kernel density as in Figure 4, but now sampling points which are evaluated to be >99% outliers based on our conditional probability estimate. Notice that the distribution of points has shifted to nearby objects which appear far away (possibly low surface brightness objects) and high redshift objects that the network is unable to constrain.

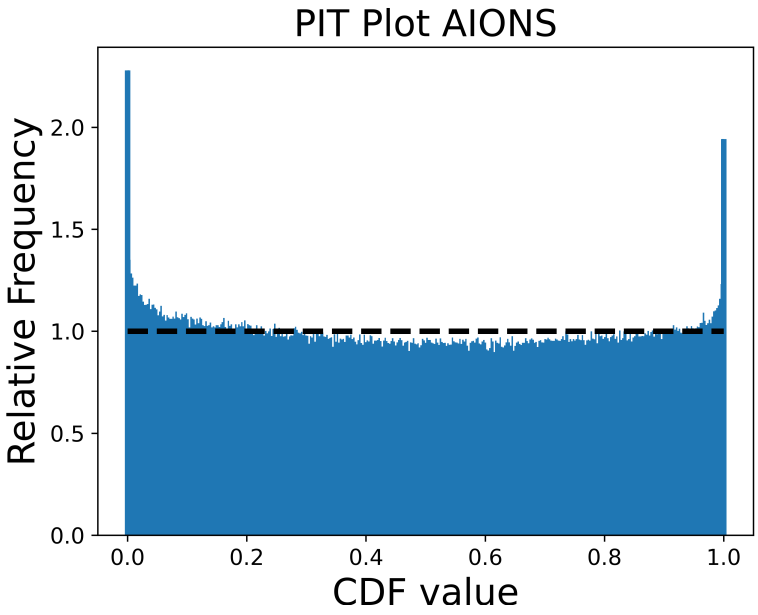

Figure 6: AIONS Probability Integral Transform. The outermost histogram bars have been widened for readability. The PIT is a commonly used visual metric in the photometric redshift community to visualize the probabilistic calibration of the model; it can be thought of as a kind of QQ plot. The PIT plot indicates our model has an over abundance of outliers that are more likely to be over-estimated than under-estimated.

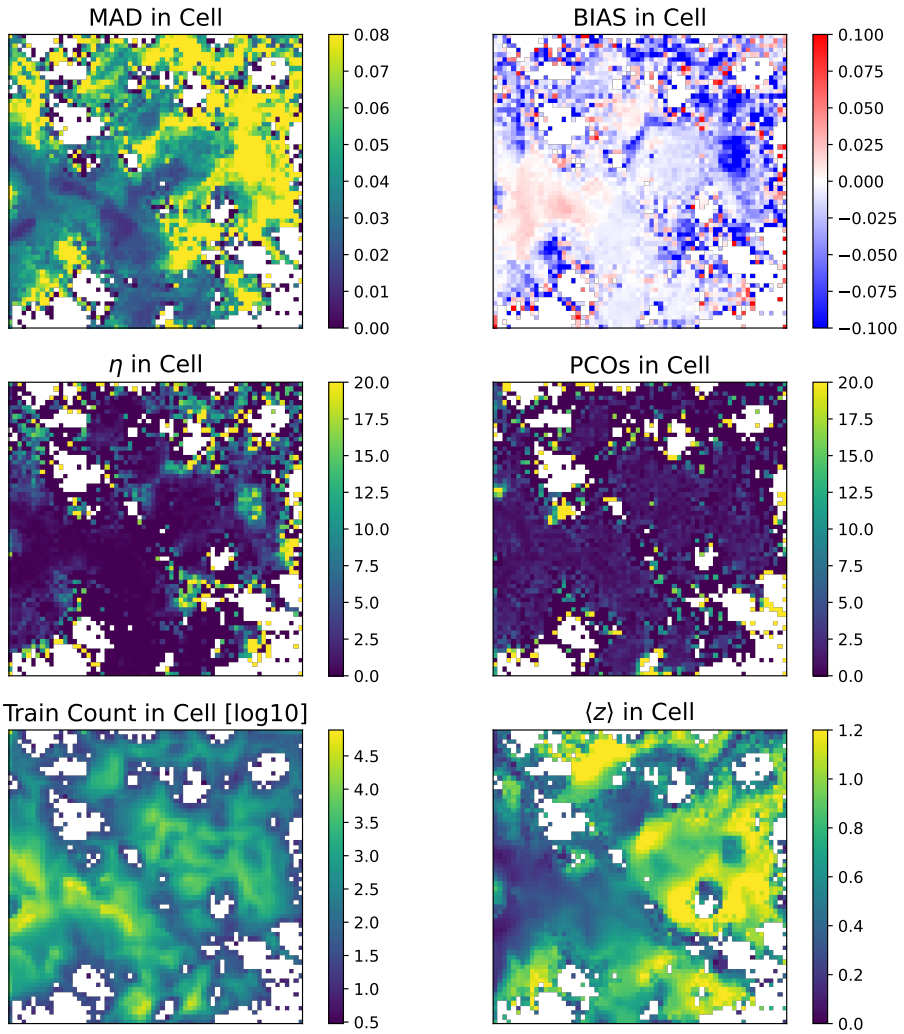

Figure 7: We visualize the SOM cells overlaid with MAD (upper left), bias (upper right), η (middle left), probabilistic catastrophic outliers (middle right), the log of spectroscopic training data in each cell (bottom left) and the average redshift of the galaxies within the cell (bottom right). These visualizations highlight the locality of populations of galaxies with different redshifts across colors. It highlights the general correlation of superior performance with lower redshift. While this is to be expected as high signal-to-noise ratio galaxies might be expected to be expected to be the most nearby sample, is it unfortunate, as the residuals remain heteroskedastic with the target variable, complicating analyses using our model. The values of cells are clipped so as to keep a sensible dynamic range for the color maps.

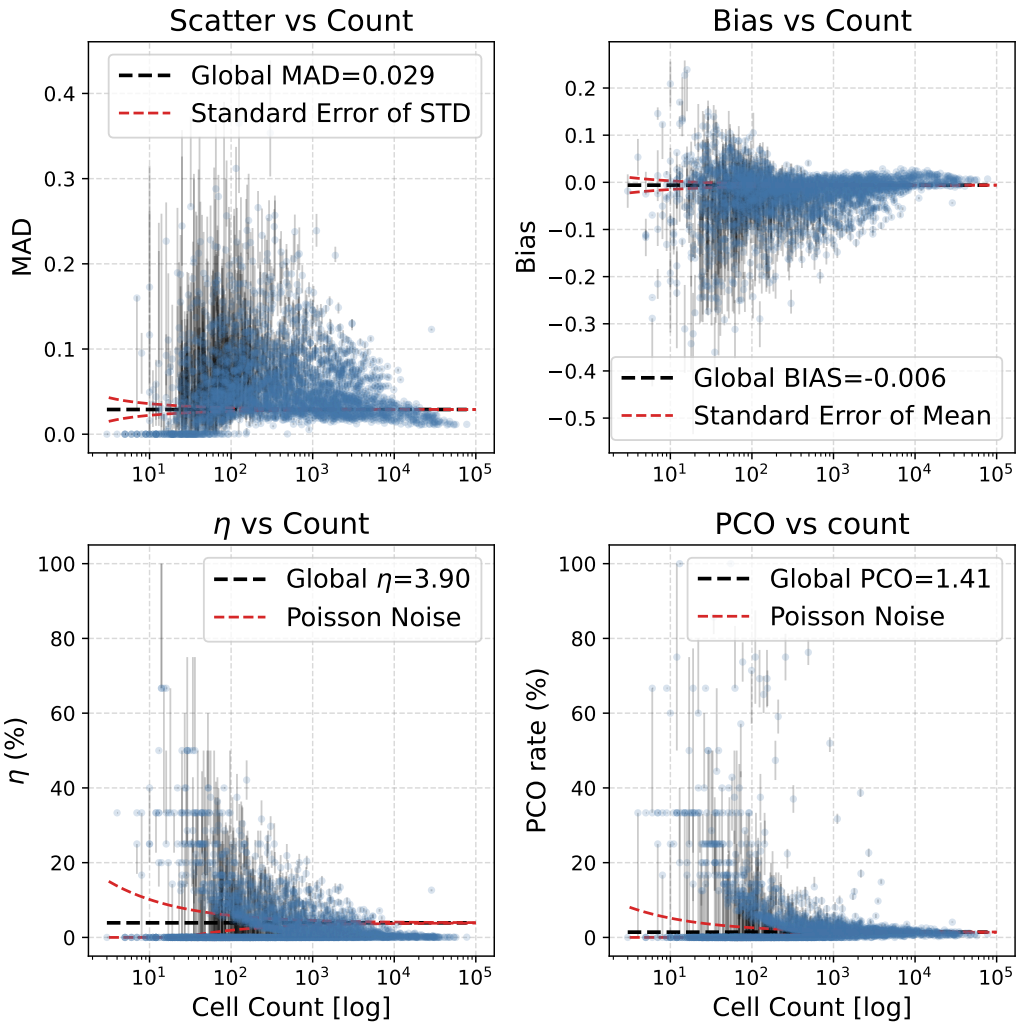

Figure 8: We plot the same data as in Figure 1 but do not mask out the cells which contain no outliers. For low cell count, these miraculous cells can be explained by simply, randomly, having no outlier in the cell. There are also cells which seemingly constrain very bright galaxies that should be easy for the algorithm to recognize as nearby. Additional work is needed to completely characterize these cells and understand whether these are the same cells which have near 0 MAD.