Using a Neural Network Classifier to Select Galaxies with the Most Accurate Photometric Redshifts

Adam Broussard
Advisor: Eric Gawiser

Website: http://www.adamcbroussard.com/
Email: adamcbroussard@physics.rutgers.edu

A New Method for Selecting Accurate Photo-z's

- Galaxy clustering (a.k.a. Large Scale Structure or LSS) is a powerful cosmological probe
- 3D clustering impossible without spec-z's
- Photo-z analysis: angular clustering in tomographic bins

- Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST) provides photometry for several billion galaxies
- Can greatly decrease sample size without hitting shot-noise limit
- We train a neural network to reject galaxies with inaccurate photo-z's to improve cosmological signal-to-noise ratio

Photometric Redshift Codes

Machine Learning-Based

Uses ML algorithms to regress or classify galaxies by redshift

Examples:

- Trees for Photo-z (TPZ; Carrasco-Kind & Brunner 2014, 2013)
- ANNz (Lahav & Collister 2012)
- NetZ (Schuldt et al. 2020)

Template-Based

Use empirical templates to calculate photometry in a grid of redshift to calculate minimum χ^2 relative to observations

Examples:

- EAZY (Brammer et al. 2008)
- Ilbert et al. (2006)
- Arnouts et al. (1999)

Pipeline Overview

1. Perform an initial redshift fit
    - We use Trees for Photo-Z (TPZ; Carrasco-Kind et al. 2013), a random forest algorithm
2. Using initial redshift fit results as additional features, classify galaxies that have accurate redshifts using a neural network classifier (NNC)
    - Outputs indicate the confidence of the given galaxy's photo-z being within $(z_{phot}-z_{spec})/(1+z_{spec}) < 0.10$

Training Features

Features used by TPZ & NNC

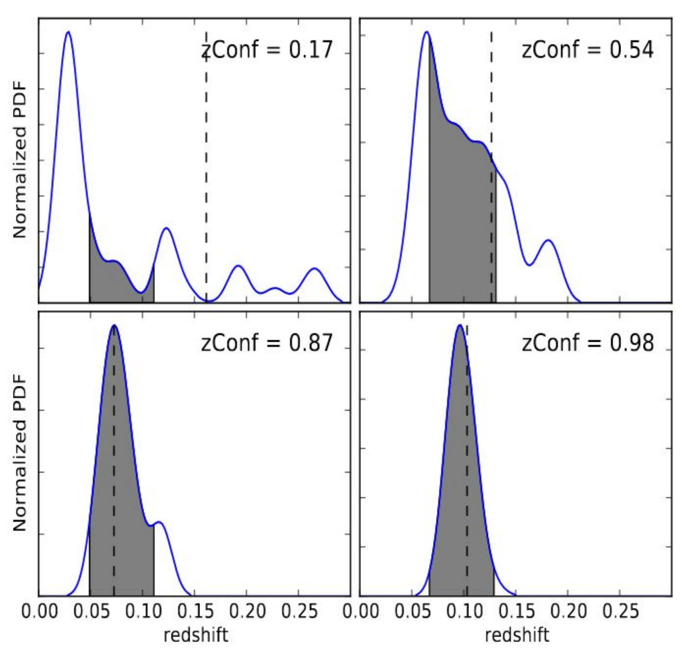
*g, r, **i**, z, y, (g-r), (r-i), (i-z), (z-y), [(g-r) - (r-i)], [(r-i) - (i-z)], [(i-z) - (z-y)]*

Magnitudes Colors Band Triplets

Features used by NNC

z_{TPZ}, σ_{TPZ}, zConf (TPZ Outputs)

zConf Explained



Carrasco Kind and Brunner (2013)

zConf is the PDF integrated over $z_{TPZ} \pm \sigma_{TPZ}(1+z_{TPZ})$

Neural Network Architecture

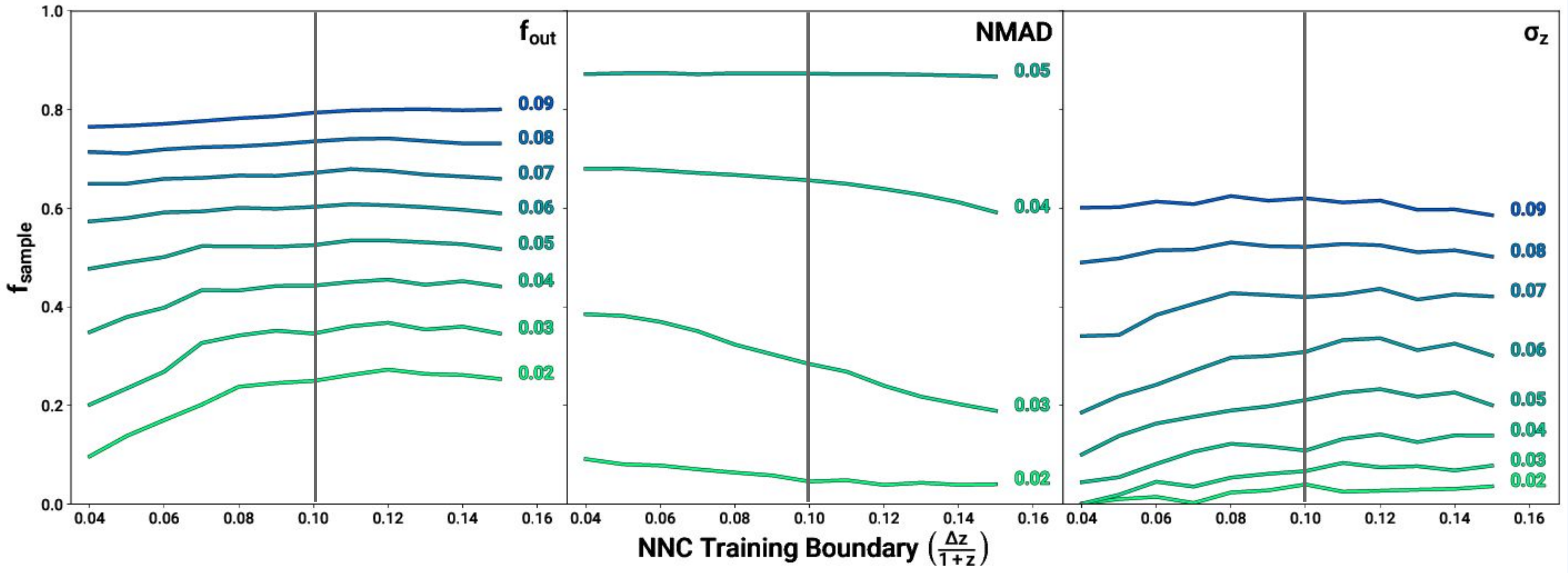
NNC

Input + 100 Neurons + 200 Neurons + 100 Neurons + 50 Neurons + 1 Output

Scaled Exponential Linear Unit (SELU) Activation*

Sigmoid Activation

Choosing a Training Boundary



Outlier fraction and σ_z are optimized at NNC training boundary of ~0.12, however the NNC increases slightly with increasing NNC training boundary, so we choose 0.10 an an optimal value

Data Sets

Names

`HSC Spec` - All Hyper Suprime Cam PDR2 (Aihara et al. 2018) Wide Field galaxies with spectroscopic redshifts and photometry

`HSC Phot` - HSC PDR2 Wide Field galaxies with photo-z's

`COSMOS2015` - All `HSC Phot` galaxies matched to a COSMOS2015 (Laigle et al. 2016) galaxy with a well-measured 30-band photo-z

Data

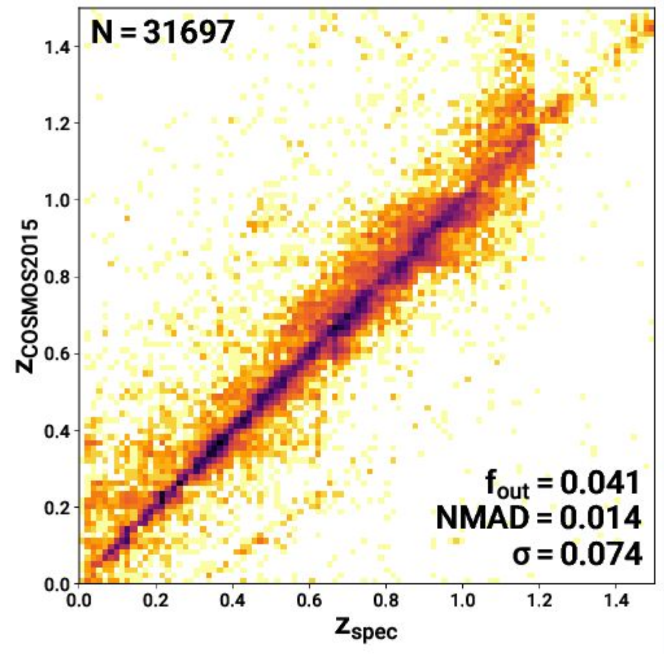
HSC Wide galaxies with photo-z's are analogous to LSST Wide Fast Deep

HSC Wide galaxies with spec-z's are analogous to LSST Deep Drilling Fields

Ideally train on HSC galaxies with spec-z's and fit galaxies with HSC photo-z's

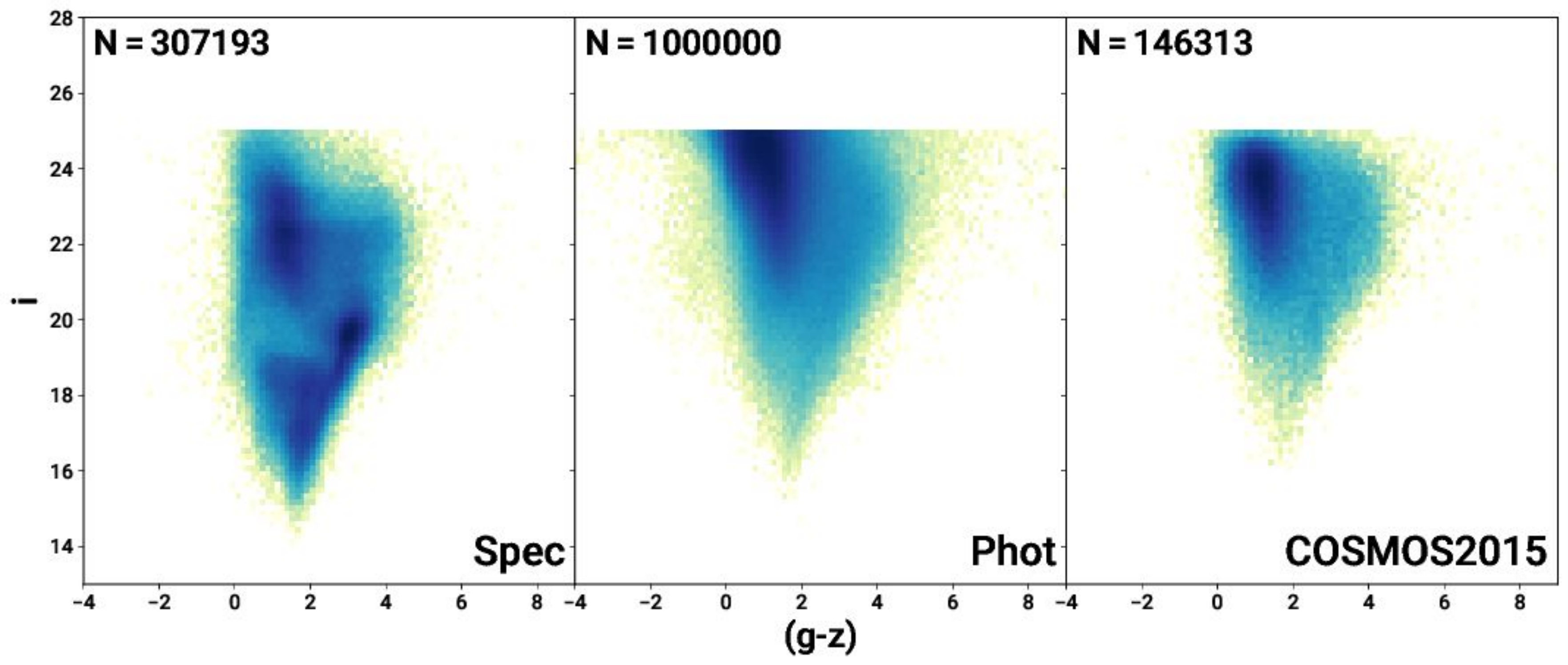
Instead, we use COSMOS2015 (Laigle et al. 2016) photo-z's to construct the test set to evaluate performance

COSMOS2015 Photo-z's



While not as good as spec-z's the COSMOS2015 30-band photo-z's add only modest scatter to the sample if treated as "truth".

Color–Magnitude Comparison



Training and Test Set Combination Overview

Case 1:

Train
HSC Spec (80%)

↓

Test
HSC Spec (20%)

- Idealistic
- Shows best-case

Case 2:

Train
HSC Spec

↓

Test
COSMOS2015

- Approximates Train/Test Mismatch

Case 3:

Train
COSMOS2015

↓

Test
COSMOS2015

- Idealistic
- Shows best-case using 30-band photo-z's

Case 4:

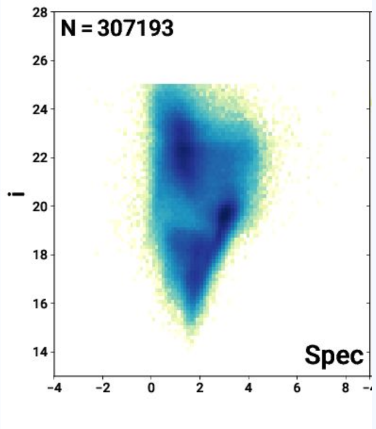
Train
Match

↓

Test
COSMOS2015

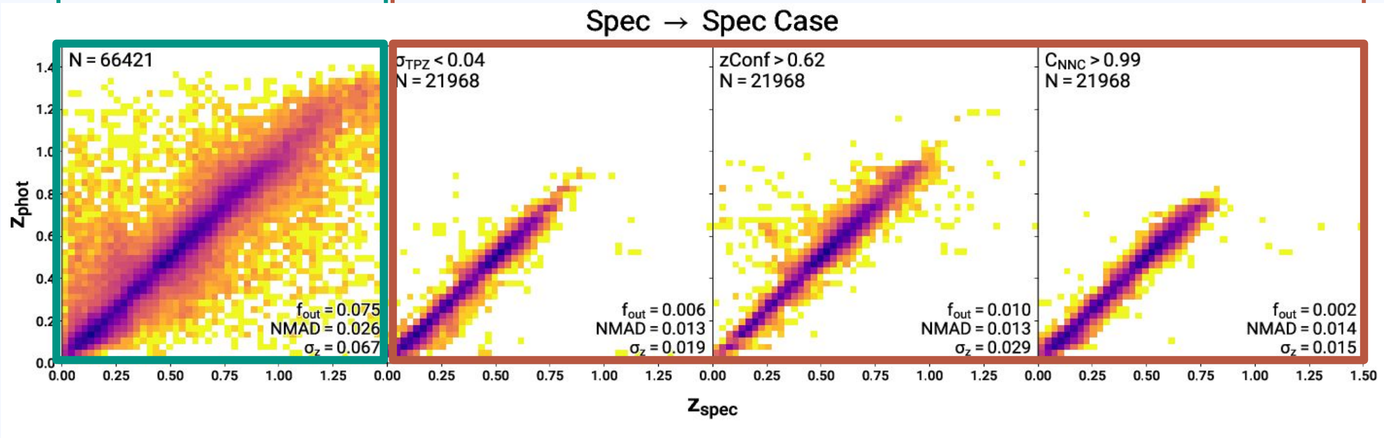
- Constructs training set to match test set
- Most sophisticated approach

Case 1 Results: Spec → Spec



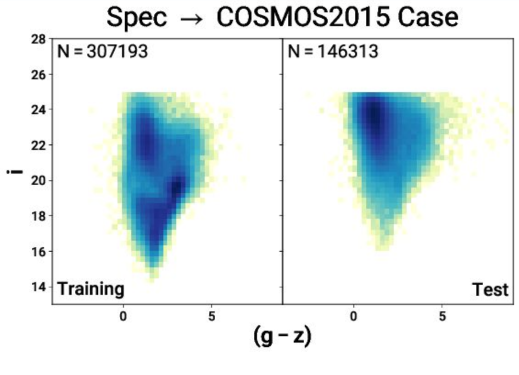
Standard Approach

Selection Methods



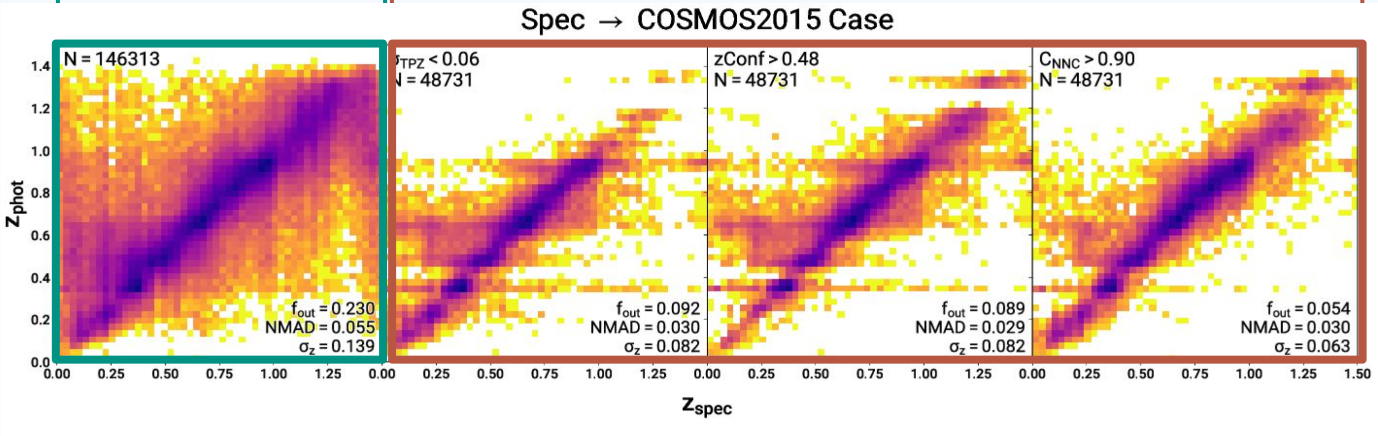
Results are idealistic, and demonstrate the best-case scenario

Case 2 Results: Spec → COSMOS2015



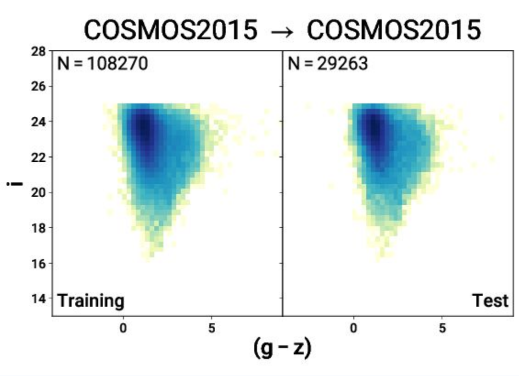
Standard Approach

Selection Methods



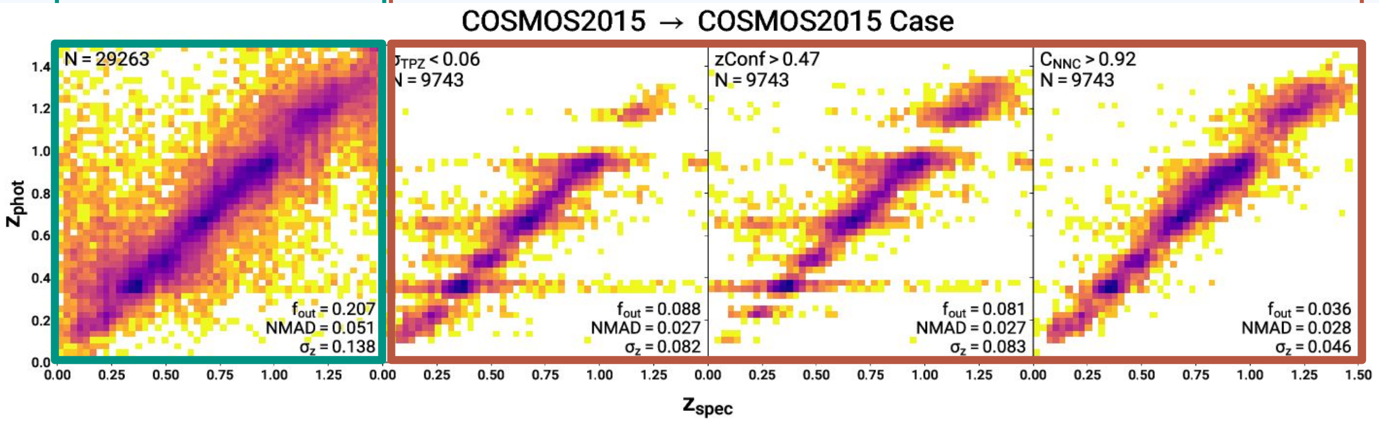
Shows a realistic mismatch between training and test set data if applied to HSC Phot

Case 3 Results: COSMOS2015 → COSMOS2015



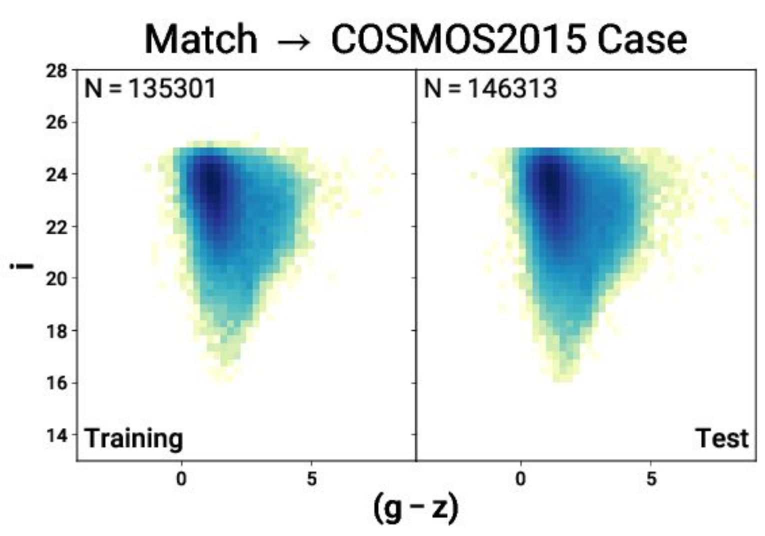
Standard Approach

Selection Methods



Idealistic similar to Case 1, but brackets the best case scenario when applying our method to COSMOS2015 photo-z's as true redshifts.

Using Data Augmentation to Construct a Training Set



Relies on assumption of scalable galaxy SED's

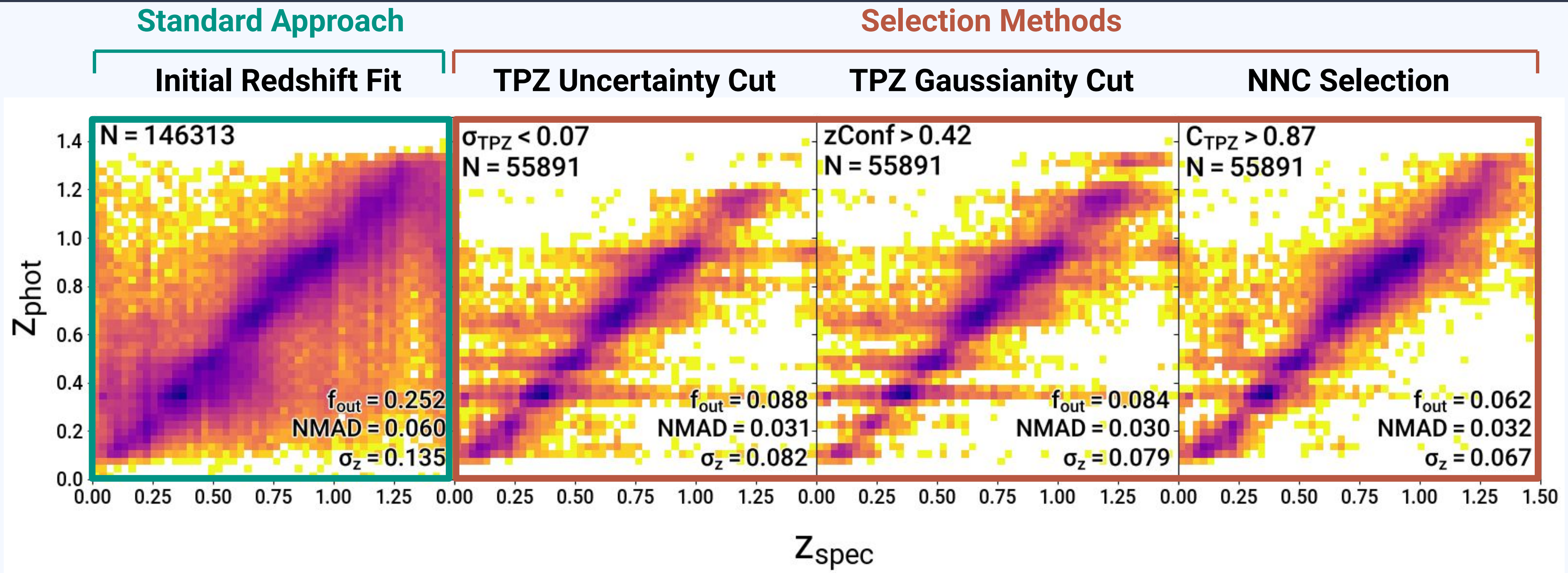
For each `COSMOS2015` galaxy:

- Treat galaxy SED's as 5D vectors
- Find the most similar SED shape in `HSC Spec` using vector dot product of normed vectors
- Renormalize `HSC Spec` SED to `COSMOS2015` object's vector norm
- Add scatter to photometry if needed
- Retain spec-z from matched `HSC Spec` object

Color magnitude distribution of constructed sample nearly identical to test sample

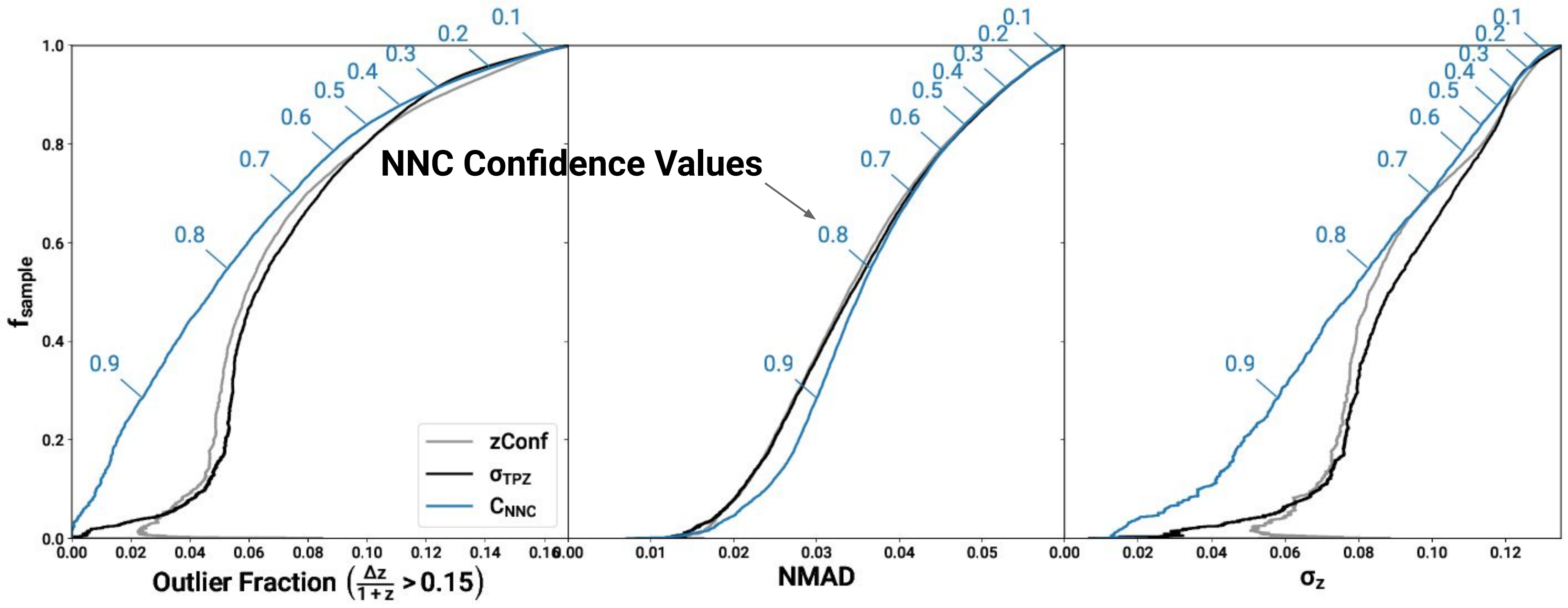
We benefit from dim spec-z's, but this is promising for analyses with a larger mismatch in sample brightness.

Case 4 Results: Match → COSMOS2015



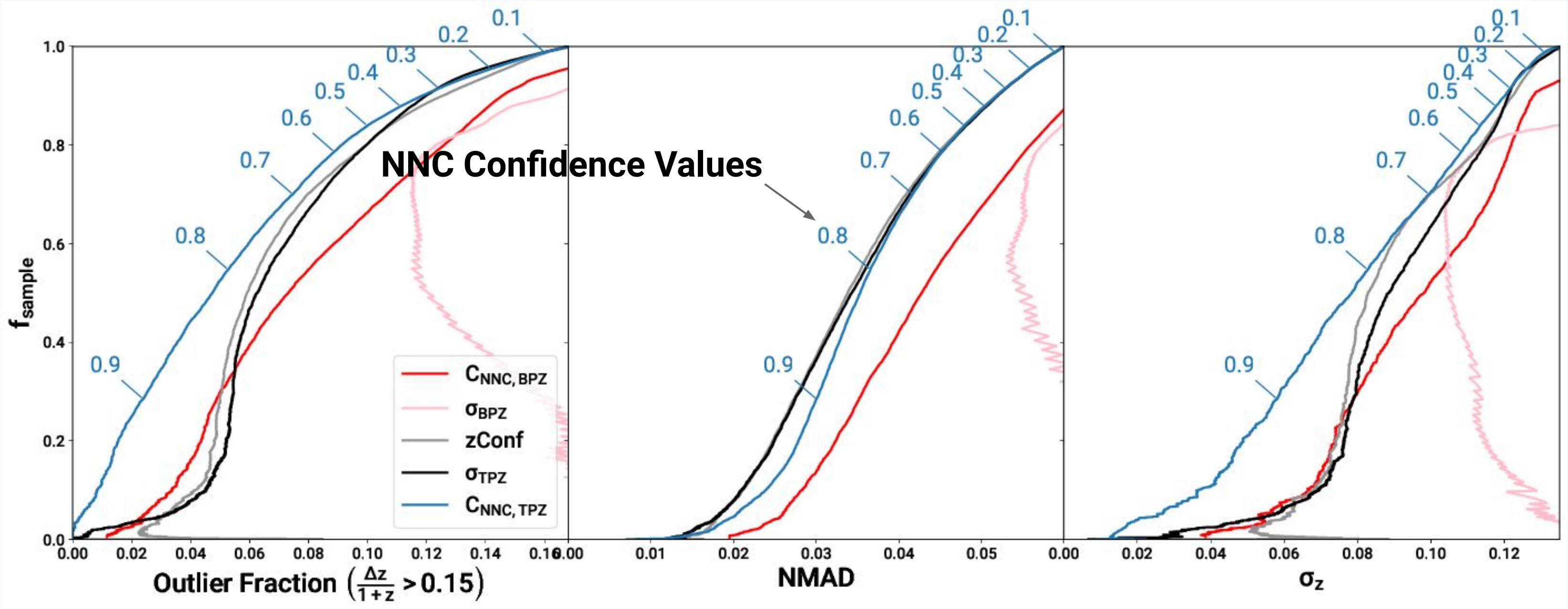
The rightmost panel shows significant improvements in f_{out} and σ_z while maintaining the same normalized median absolute deviation (NMAD) when performing a fiducial selection of ⅓ of the original sample.

Trading Sample Size and Accuracy



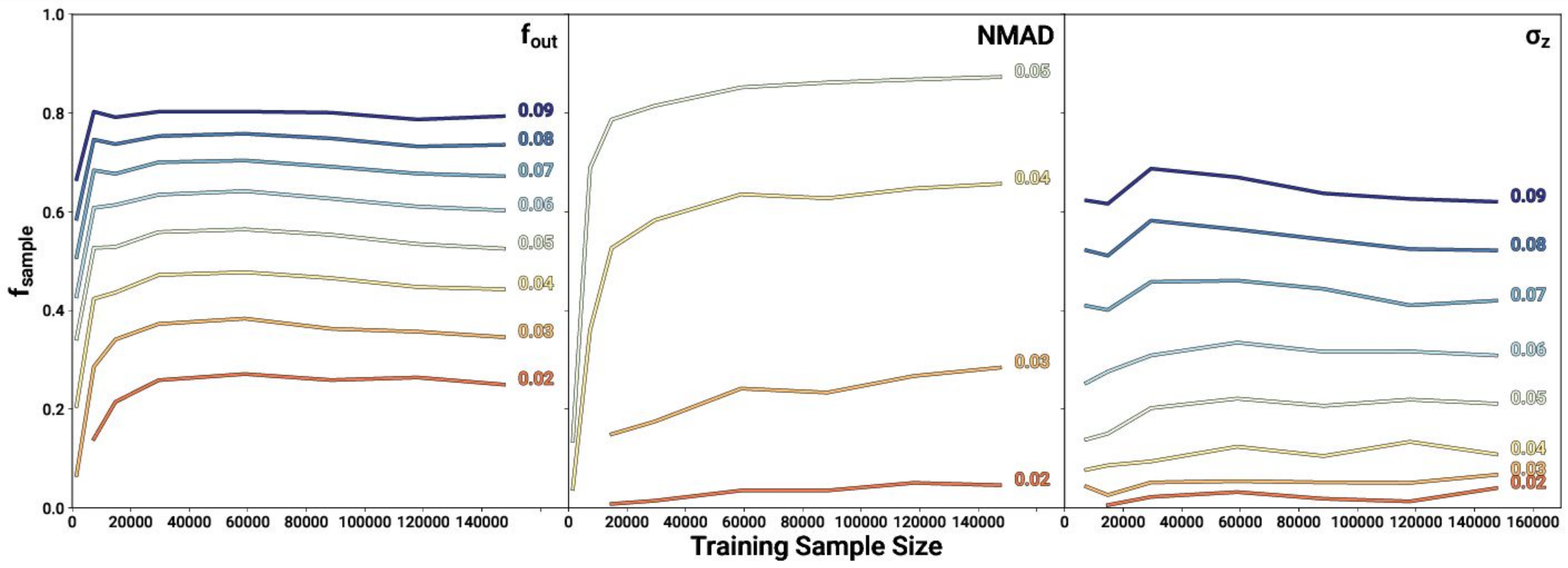
Selections made with the NNC are able to provide greatly improved f_{out} and σ_z over selections made with reported uncertainty or Gaussianity for any selected sample fraction with only a small increase in NMAD.

What about other photo-z codes?



BPZ yields similar results, but decreased photo-z quality yields a smaller selected sample for a given degree of accuracy.

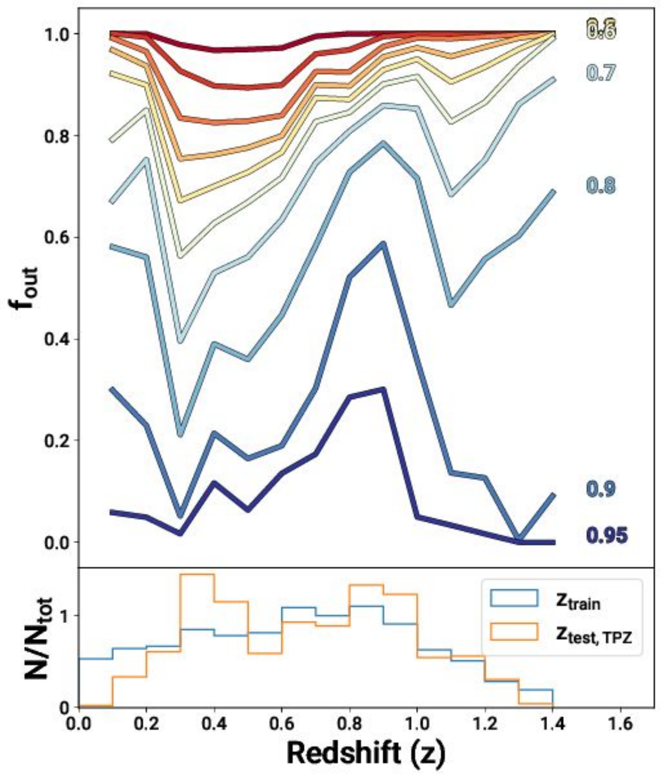
Model Validation



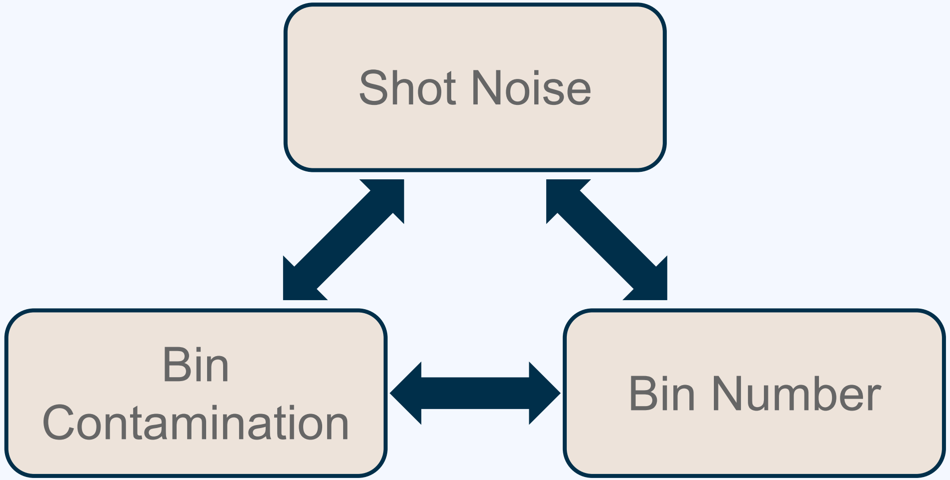
Outlier fraction and σ_z are optimized at ~40,000 objects in the training sample while the NMAD continues to improve with increasing sample size

What does the selected redshift distribution look like?

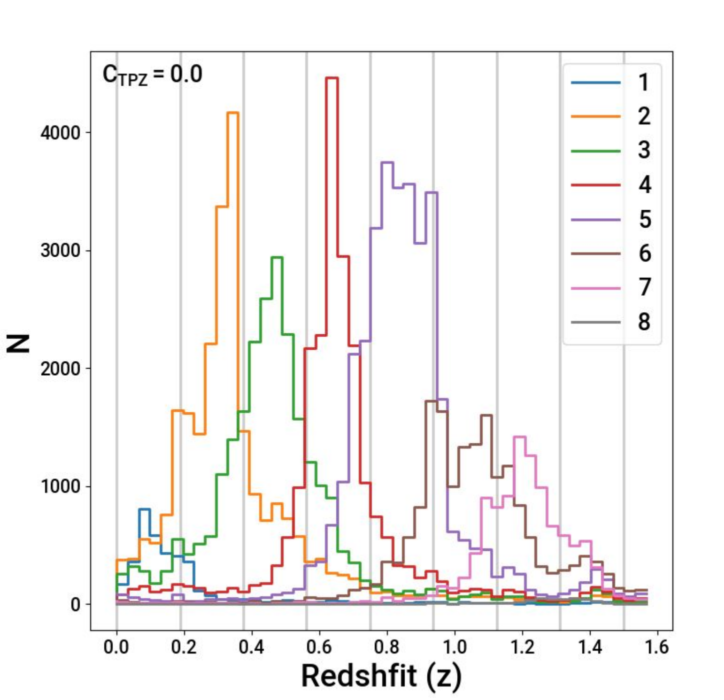
Selection is not uniform in redshift, but contains a large enough sample fraction to enable tomographic binning



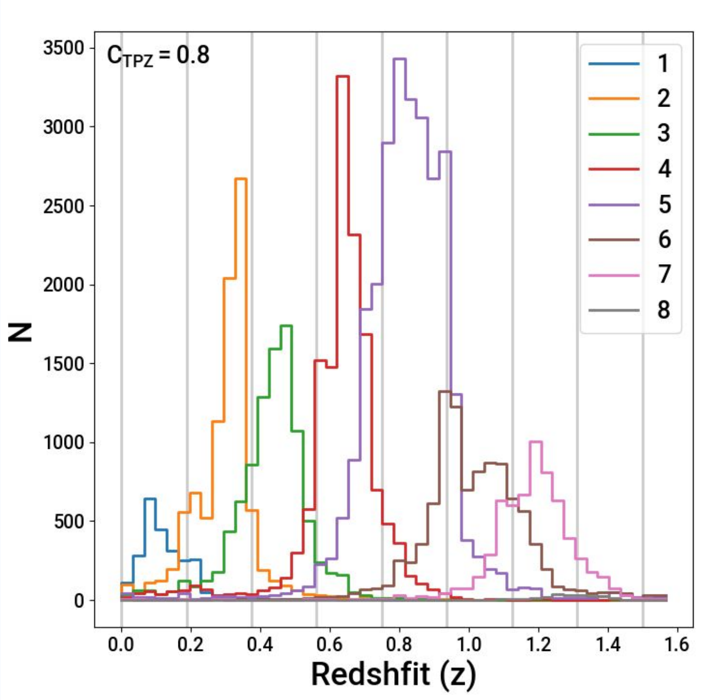
What about the cosmology?



Improvements to N(z)

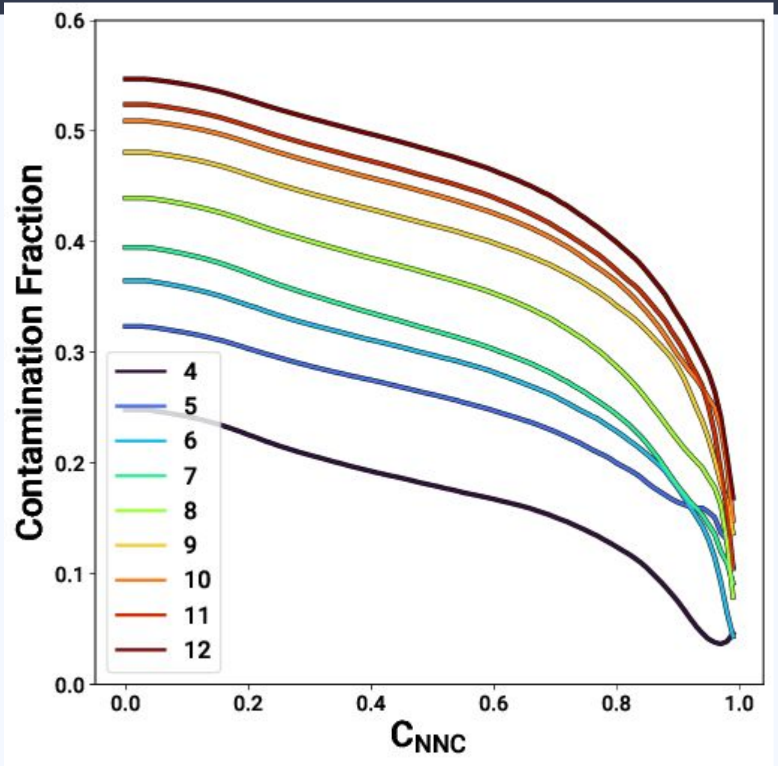


Initial Tomographic Bin N(z)'s



Improved Tomographic Bin N(z)'s

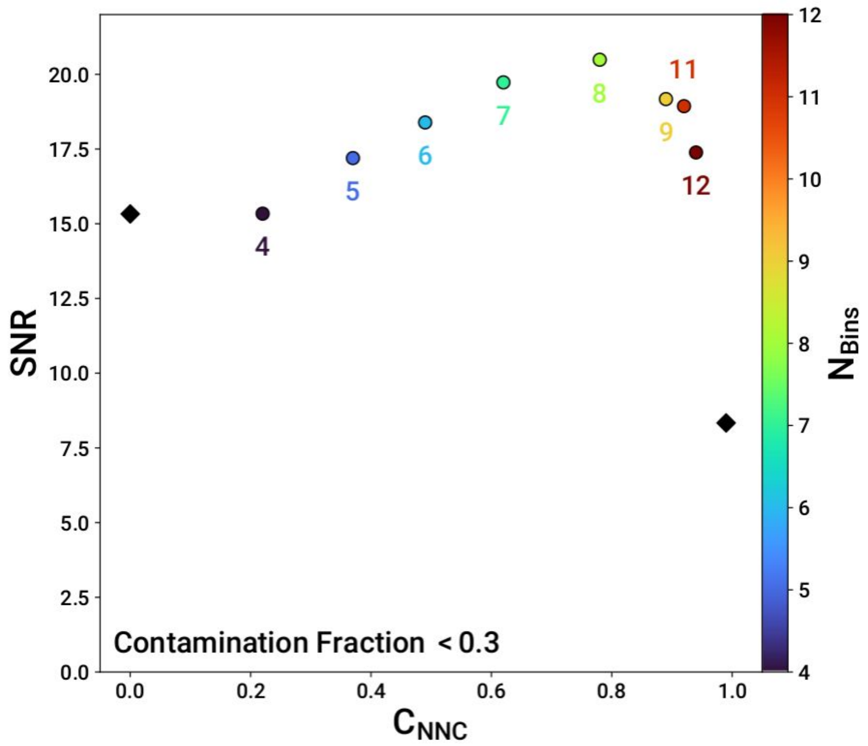
Total Contamination Fraction vs. C_{NNC}



The NNC reduces total contamination fraction by a similar amount regardless of tomographic bin number

Total Signal-to-Noise vs. C_{NNC}

The NNC optimizes the angular power spectrum SNR at 8 tomographic bins (for this example)



Computed using the Core Cosmology Library (CCL; Chisari et al. 2019)

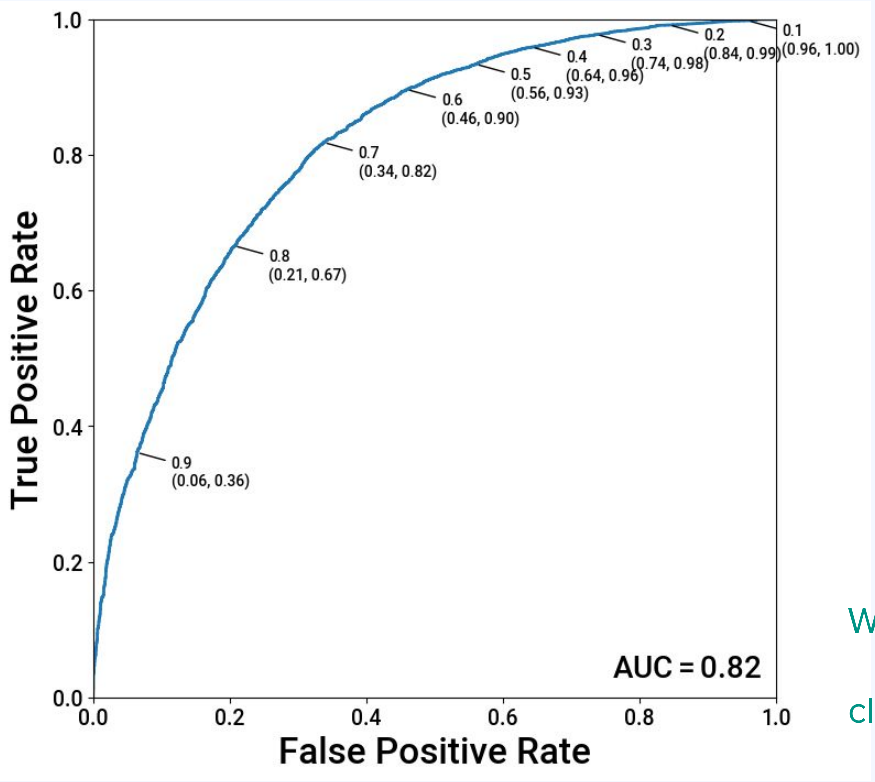
Conclusions and Future DESC Work

- We present a **new method of photo-z selection** using a neural network classifier that uses photometry and photo-z statistics as features
- This method **outperforms selections made using photo-z statistics alone**, and can provide substantial improvements in outlier fraction and σ_z when selecting $\lesssim$ 50% of the initial sample
- We find **significant improvements in galaxy angular clustering SNR** using this method
- Accepted by ApJ **(Broussard & Gawiser 2021)**; being applied to the DESC Tomographic Challenge by Irene Moskowitz
- **Future plans:** 3x2pt DESC analyses of DC2 and Hyper Suprime-Cam PDR2 catalogs to measure improvements in cosmological constraints

Backup Slides

Receiver-Operator Characteristic Curve

What fraction of well-fit galaxies get correctly classified as well-fit?



What fraction of the poorly-fit galaxies get incorrectly classified as well-fit galaxies?