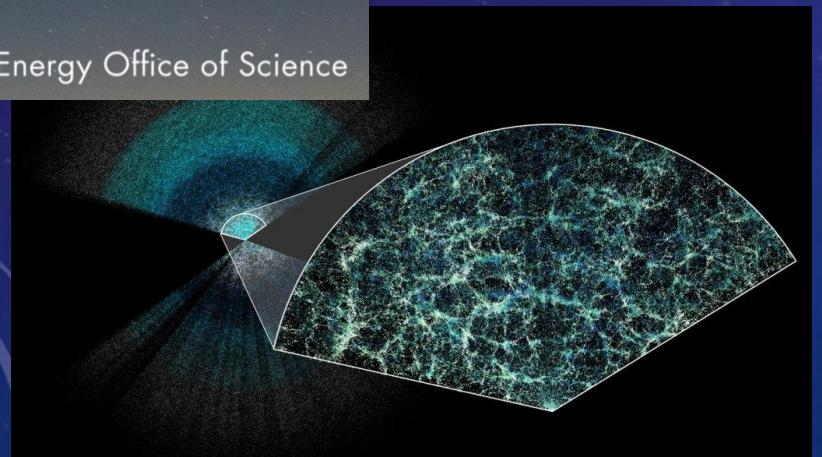
NSF Award # - 2020275

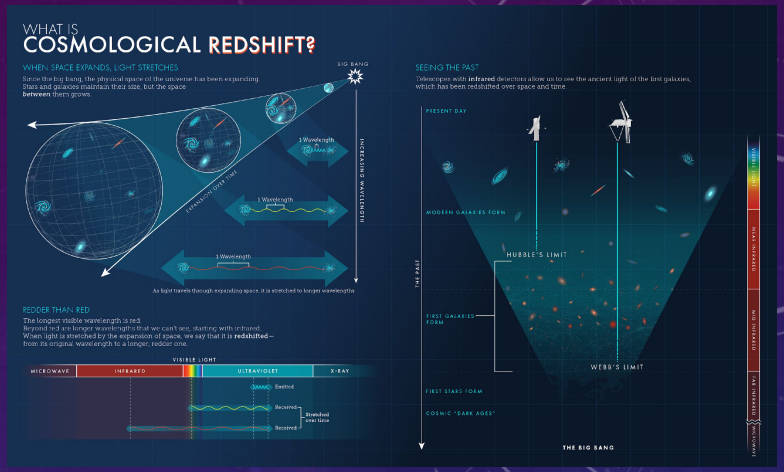
RANDOM FOREST PREDICTION OF PHOTOMETRIC REDSHIFTS FOR ELG, LRG, AND QSOS IN THE DARK ENERGY SPECTROSCOPIC INSTRUMENT

PRESENTED BY: QUENTIN LE NY

MENTOR: NOAH WEAVERDYCK



CONTEXT



Why does photometric redshift prediction matter?

- Spectroscopic redshifts are accurate but **resource-intensive** to obtain
- Large scale surveys (e.g., **LSST**, **DESI**) require fast, scalable alternatives
  - LSST: 5 million exposures of 40 billion objects – data file of ~50 Petabytes
- Precision photo-z enables key science: **galaxy clustering, cosmology, dark energy constraints**
  - ELGs, LRGs, and QSOs are **key tracers** of cosmic structure at different redshifts
  - Accurate redshifts for these tracers are essential for **galaxy clustering analysis** and **cosmological constraints**

CONTEXT

QSO
(Quasi-Stellar Object)

Crucial for constraining models of **dark energy and early structure formation**

ELG
(Emission Line Galaxy)

Useful for probing **growth of structure**, **bias evolution**, and **matter density**

LRG
(Luminous Red Galaxy)

Biased tracers of large-scale structure -- amplify clustering signal -- Ideal for **Baryon Acoustic Oscillation (BAO)** measurements

RESEARCH PROBLEM

- How accurately can we predict **redshift** from DESI photometric data using machine learning?
- **Goal:** build separate models per tracer to **maximize photo-z accuracy** without using spectroscopic inputs
- **Challenge:** high dimensionality and **feature redundancy** across magnitudes, colors, and observational depth
- Evaluate whether models generalize across tracers, and whether **feature selection** improves performance
- Key question: *Which features actually help, and how do their importances vary by galaxy type?*

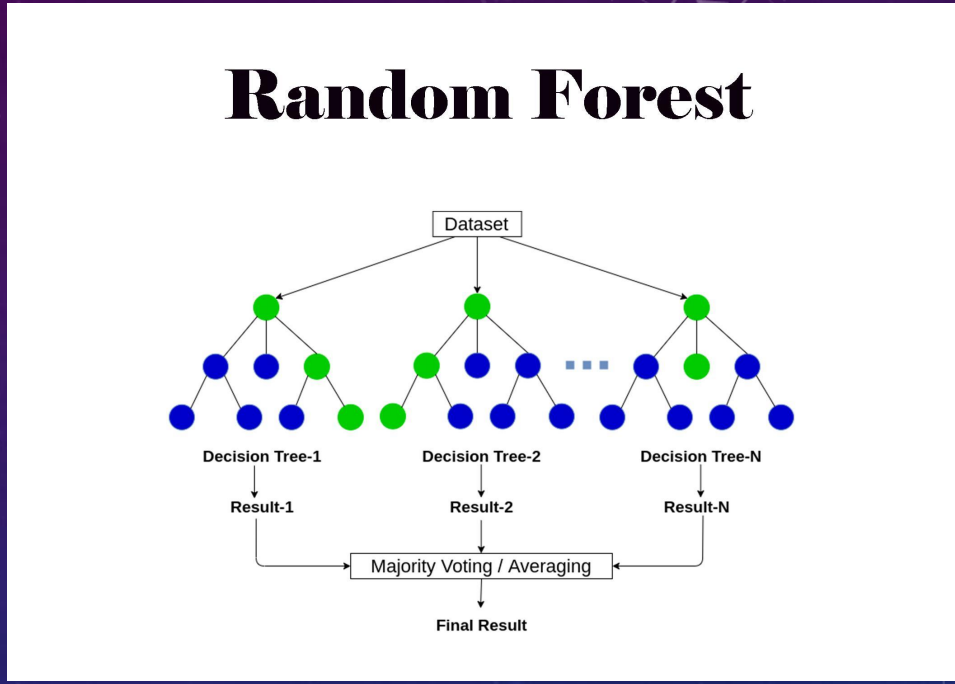
METHODOLOGY

- **Data**: DESI DR1* photometric catalogs for **ELG**, **LRG**, and **QSOs**

- **Preprocessing**:
- Filter out data points w/ high uncertainty
- Sample dataset to reduce runtime
- Engineered features: dereddened magnitudes, colors, fiber ratios

- **Model**:
- Random Forest Regressor
- Tuned w/ GridSearchCV to vary hyperparameters (# of estimators, depth, sample splits)

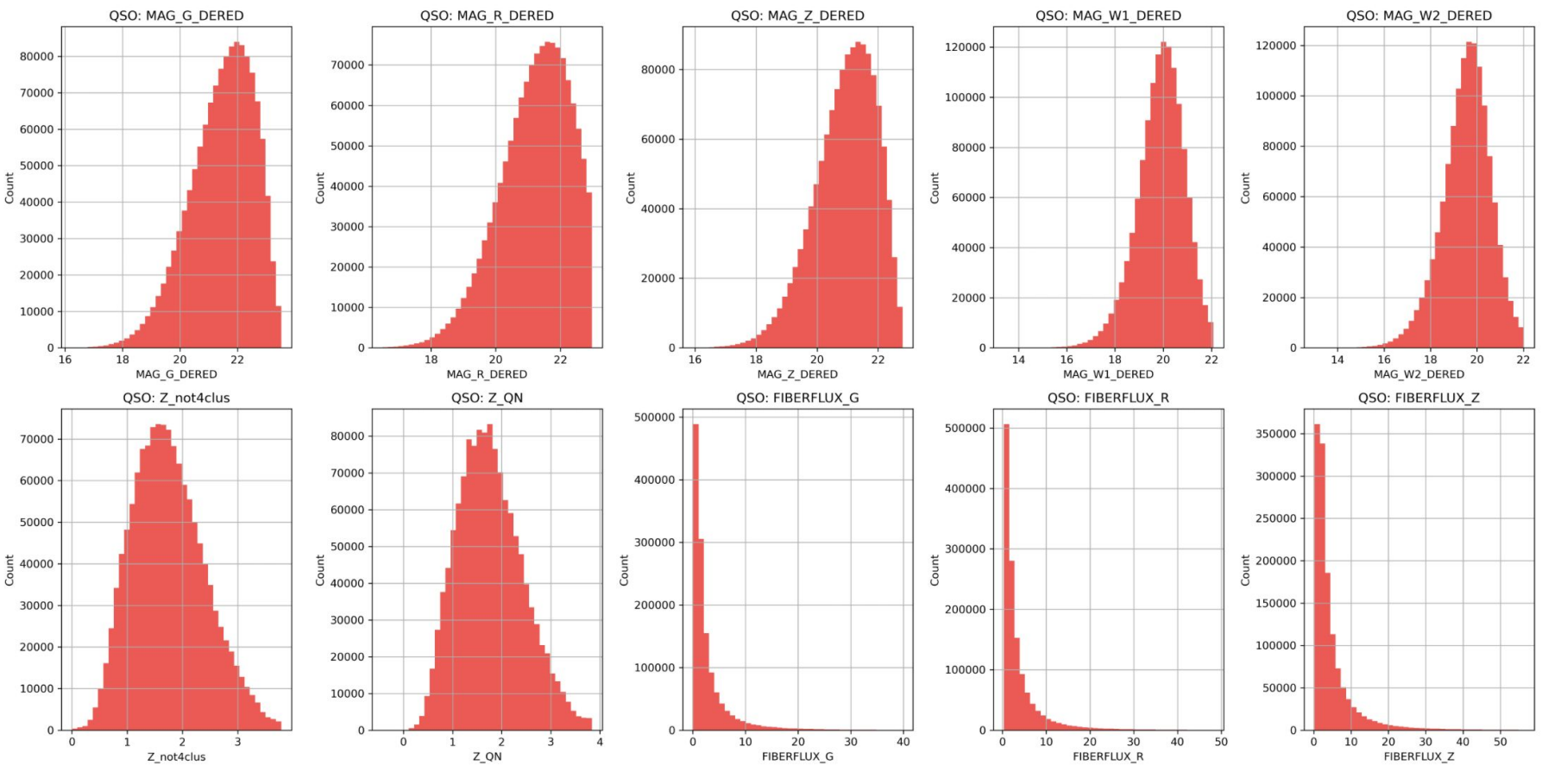
- **Evaluation Metrics**:
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R^2 score



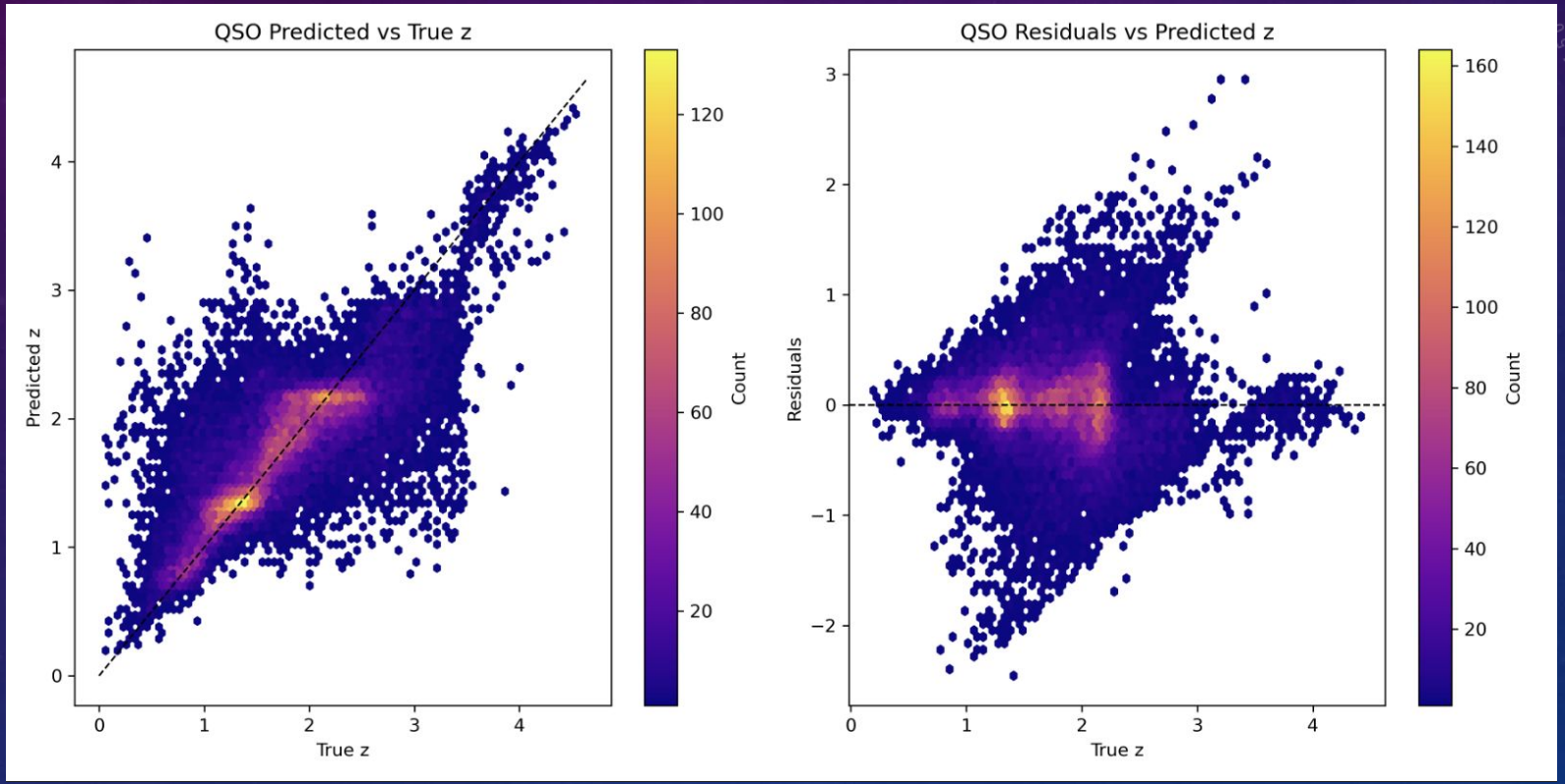
Goal: Identify optimal features and model parameters for accurate redshift prediction across galaxy types

*DESI DR1 LSS Iron Catalog v1.5 — public release

METHODOLOGY

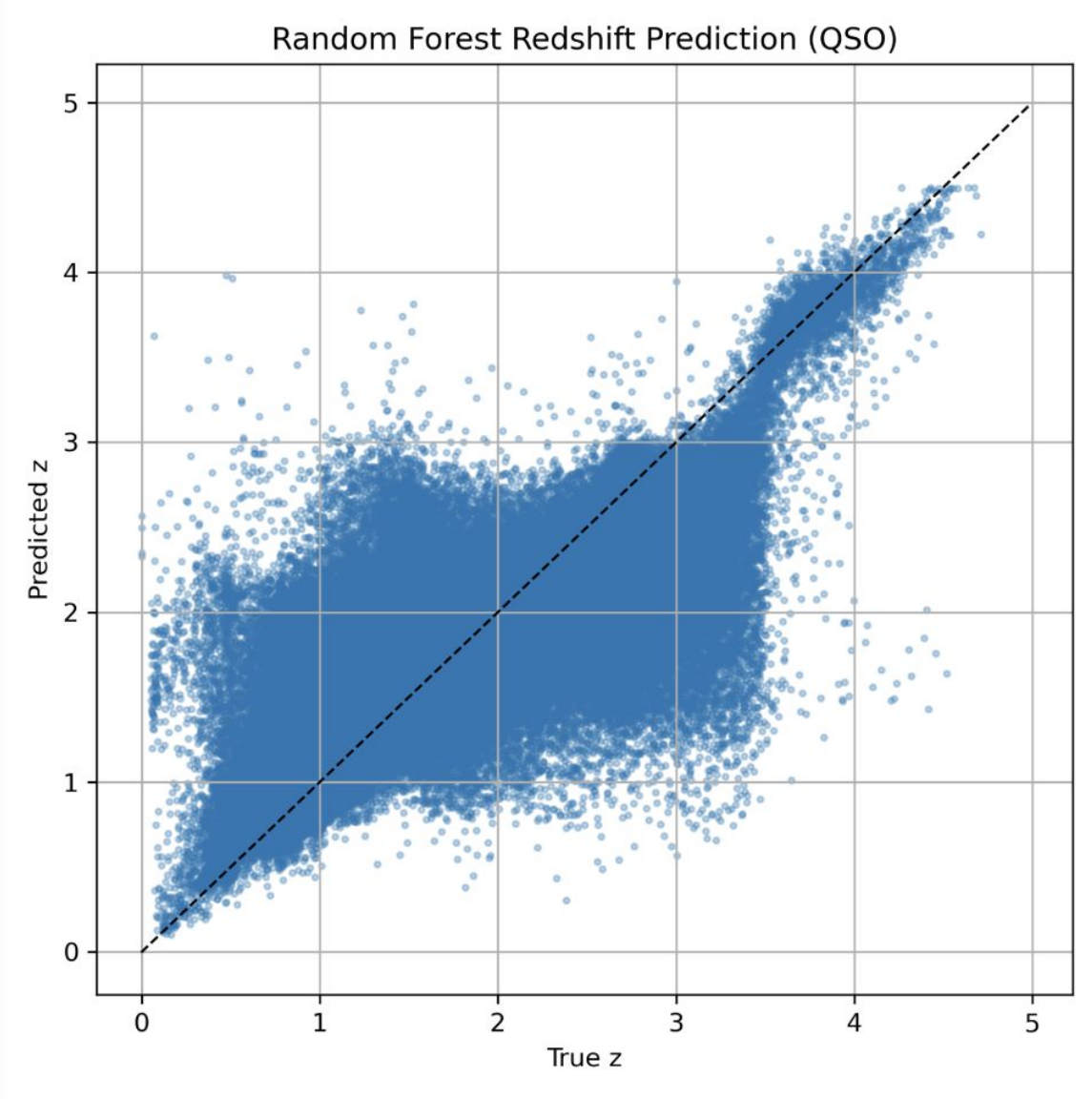


RESULTS

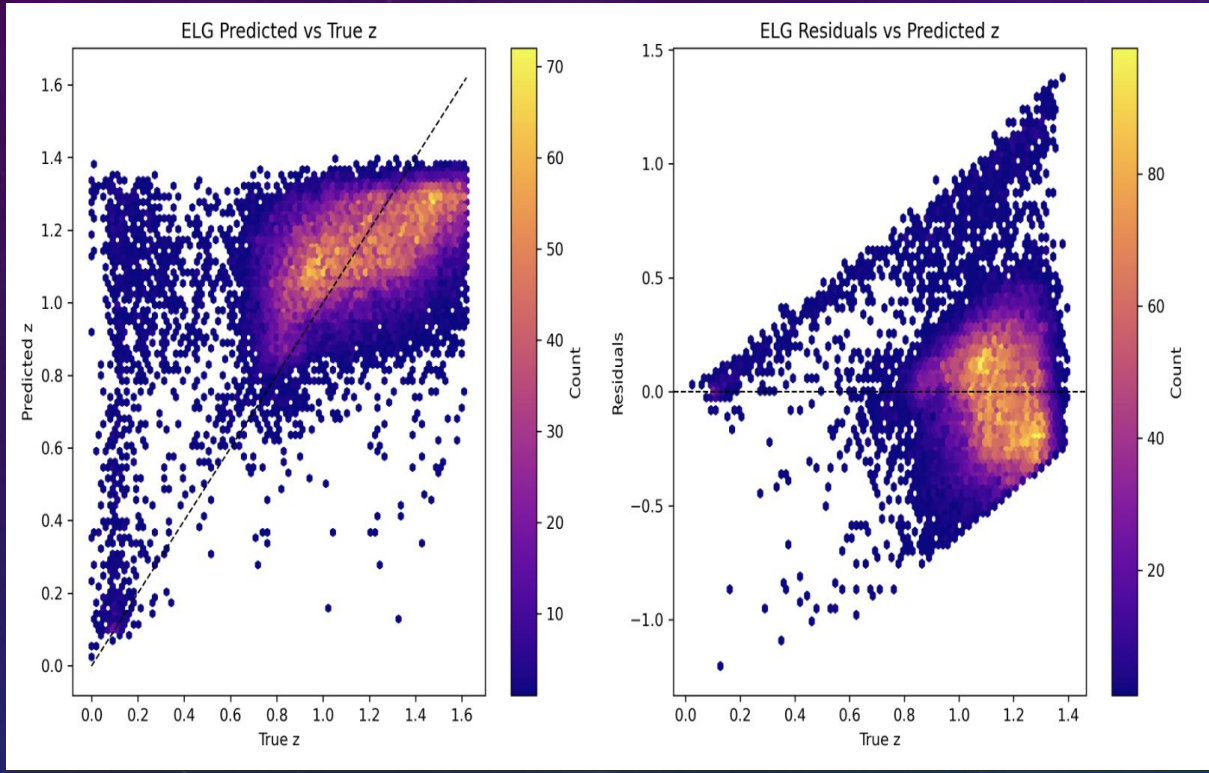


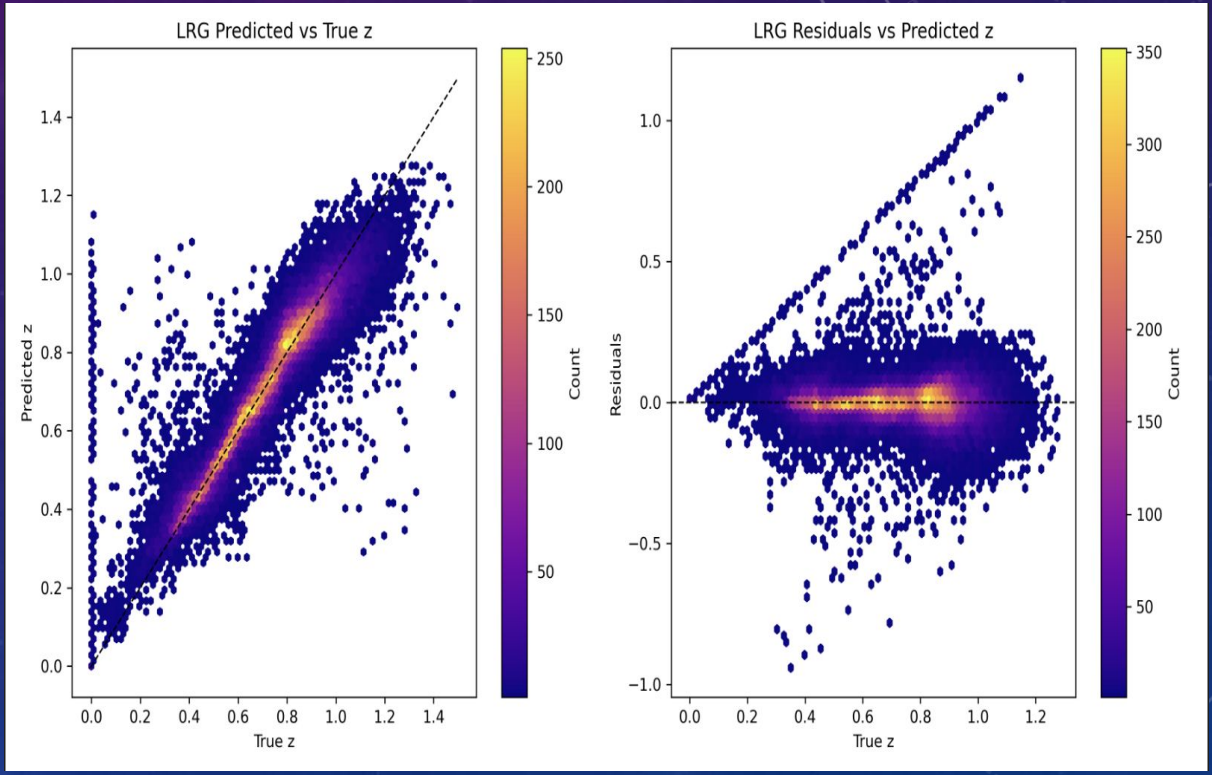
MAE: 0.28 **MSE: 0.17** **R^2: 0.62**

RESULTS


Random Forest Redshift Prediction (QSO)
Predicted z
True z
0
1
2
3
4
5

NEXT STEPS





MAE: 0.19 **MSE: 0.06** **R^2: 0.29**

MAE: 0.045 **MSE: 0.005** **R^2: 0.885**

NEXT STEPS

- Continue **feature pruning** (e.g., remove low-importance magnitudes, magnitude uncertainties, any columns with repeatedly low feature importances)
- Explore **dimensionality reduction** (e.g., PCA) to eliminate redundancy across photometric features
- Compare Random Forests to other ML models — **Neural Networks, XGBoost, or ensemble methods**
- Incorporate addtl. uncertainties or quality metrics into model predictions

CONCLUSION

- Developed Random Forest models to estimate photometric redshifts for ELG, LRG, and QSO samples
- **Strong performance** for LRGs ($R^2 \approx 0.88$); **moderate** for QSOs, and **limited** for ELGs
- Feature importances varied by tracer, offering insight into photometric predictors of redshift
- Despite extensive tuning, model performance plateaued
- Highlights the need for feature reduction and exploration of alternate model architectures