Photometric selection and redshifts for quasars

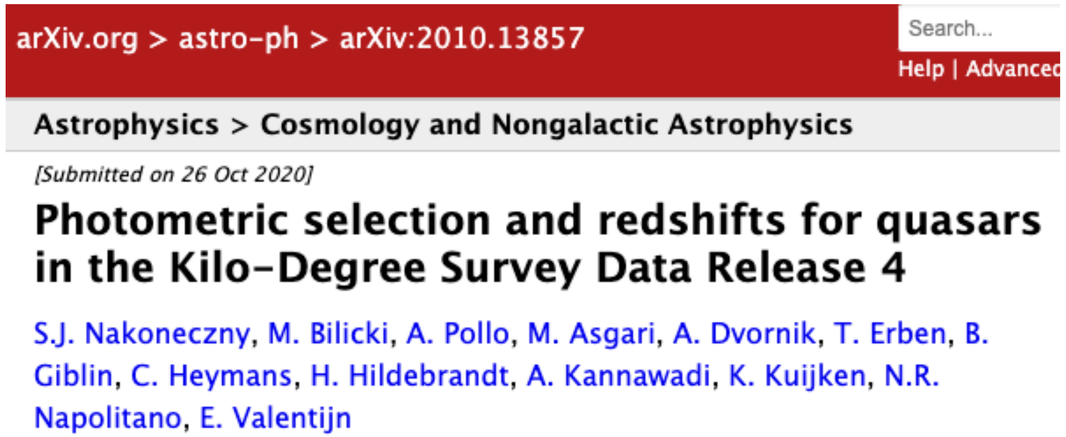
Kilo-Degree Survey Data Release 4

Szymon Nakoneczny - 17.10.2021

The catalogs

http://kids.strw.leidenuniv.nl/DR3/quasarcatalog.php

http://kids.strw.leidenuniv.nl/DR4/quasarcatalog.php



Changes to the DR4 release:

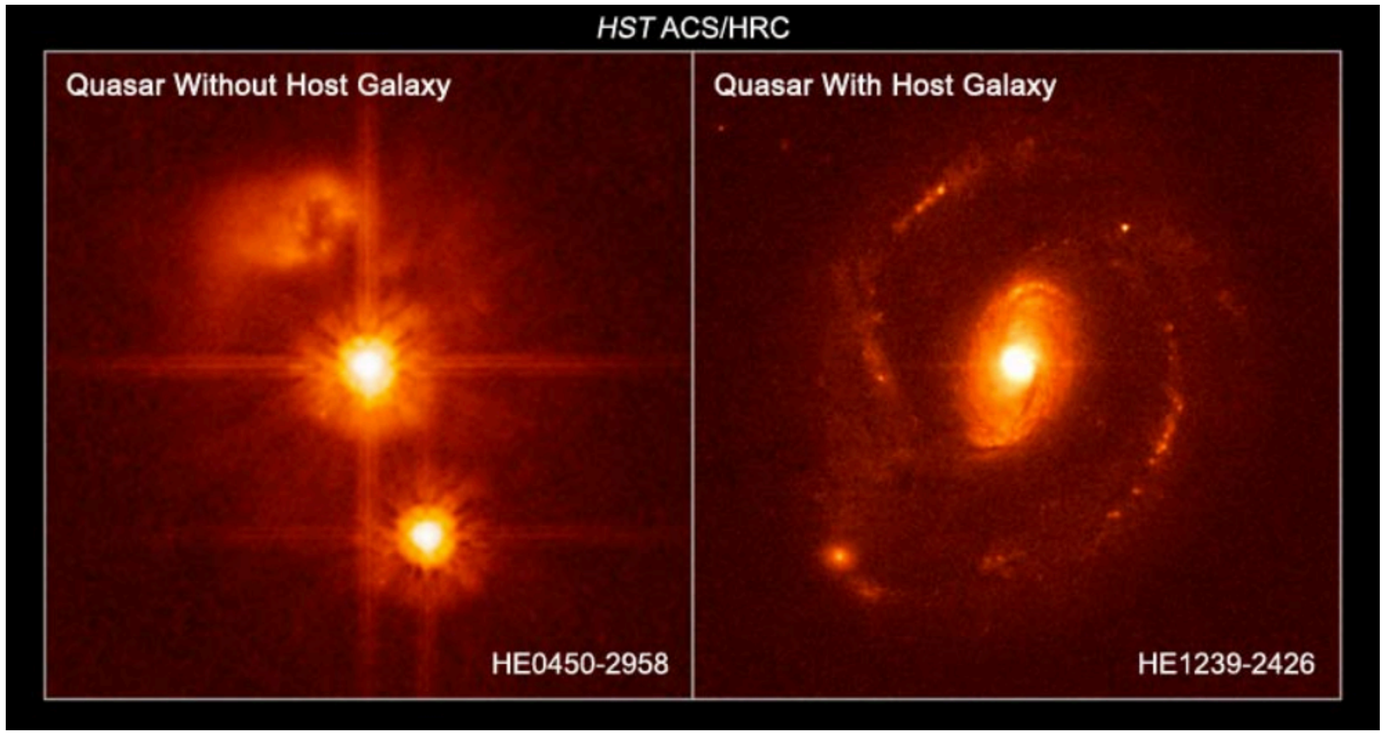
1. Added photo-zs and near-IR
2. Increased the magnitude coverage to $r < 23.5$

Agenda

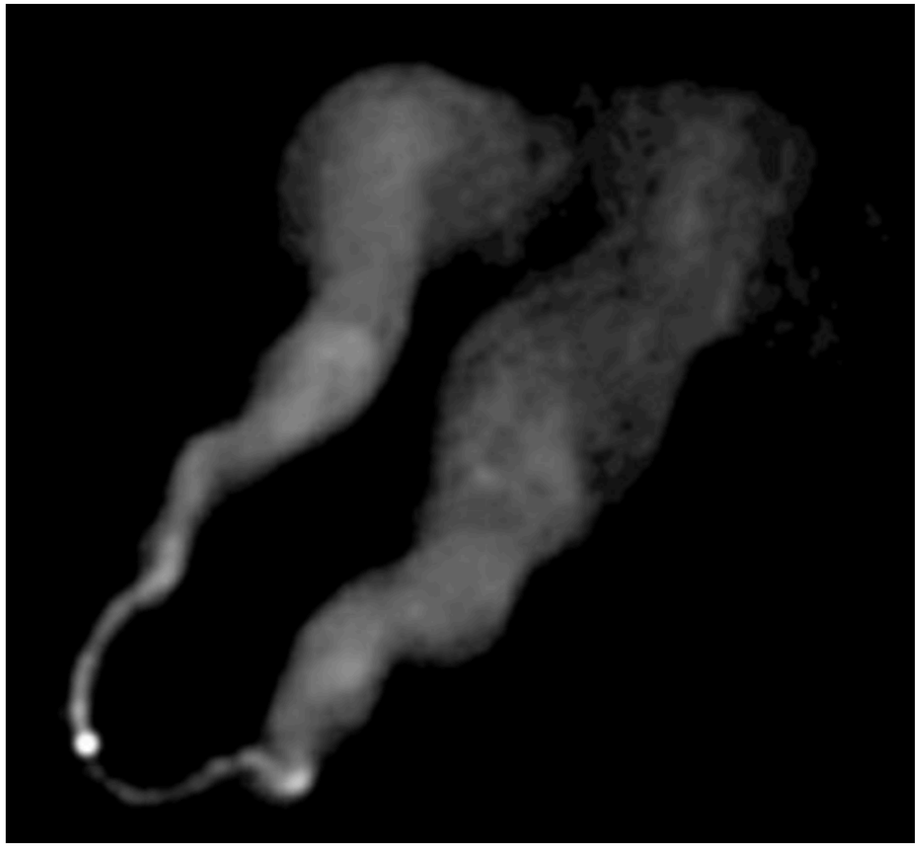
1. Introduction
    1. Quasars in photometry
2. Methodology
    1. Data (research and visualisation)
    2. Machine learning (validation)
3. Results
    1. Experiments
    2. Final catalog (testing)

Introduction

Left: quasar (in the center) and a star



Right: quasar and a spiral host galaxy

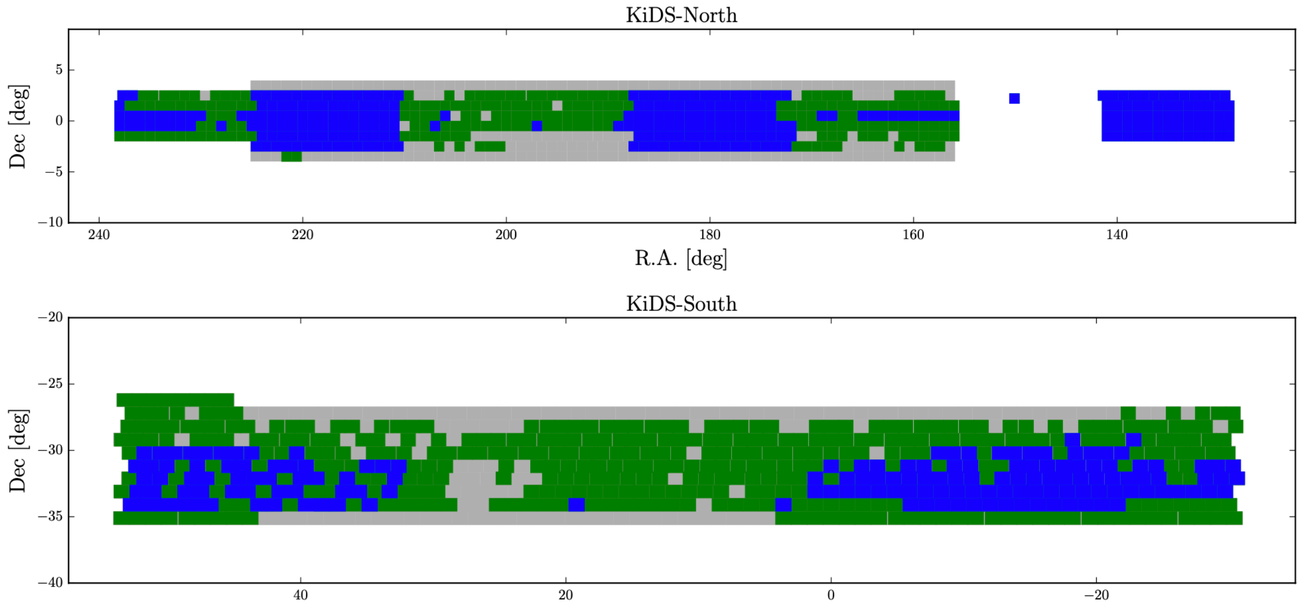


Active galactic nucleus (AGN) with jets

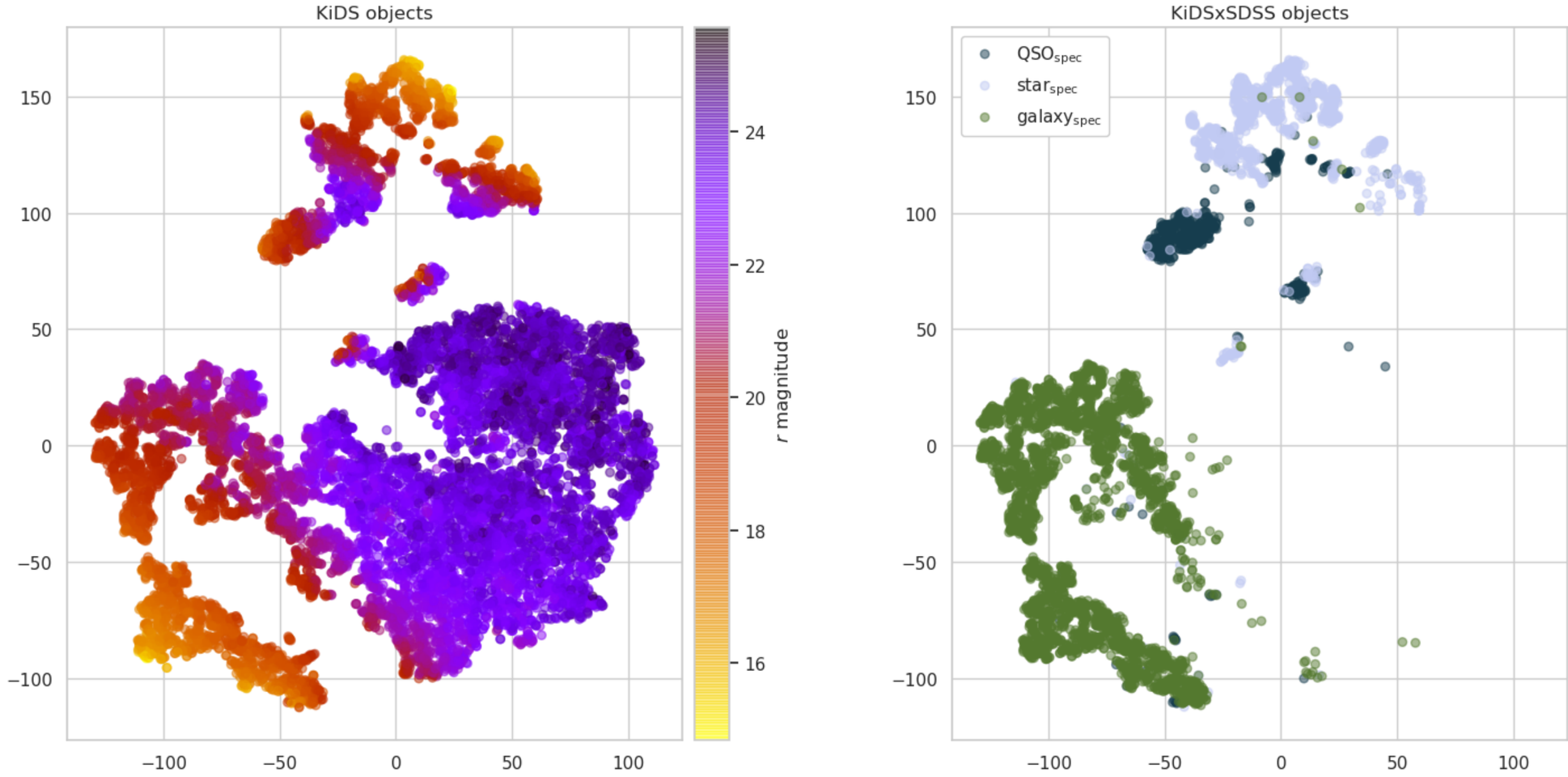
Methodology

Data

- Kilo-Degree Survey (KiDS): optical *ugri*
- VISTA Kilo-degree Infrared Galaxy Survey (VIKING): near-infrared *ZYJHKs*
- Depth: $r < 25$
- Data release 4: 1006 square-degrees (1350 in the final DR5), ~100M objects
- 9 band detections yield 45M objects

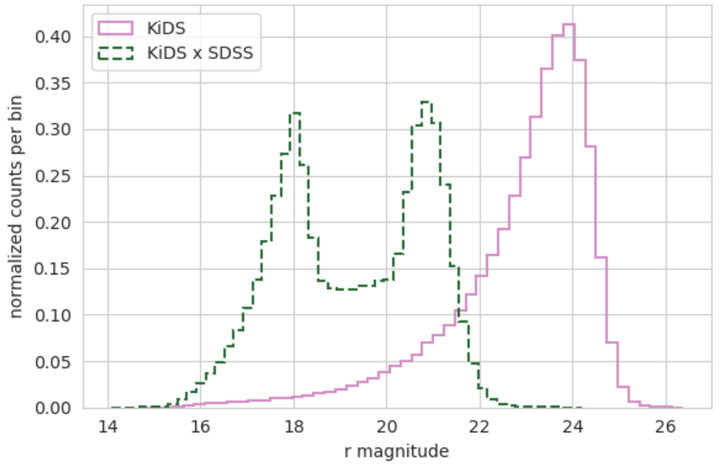


Extrapolation problem

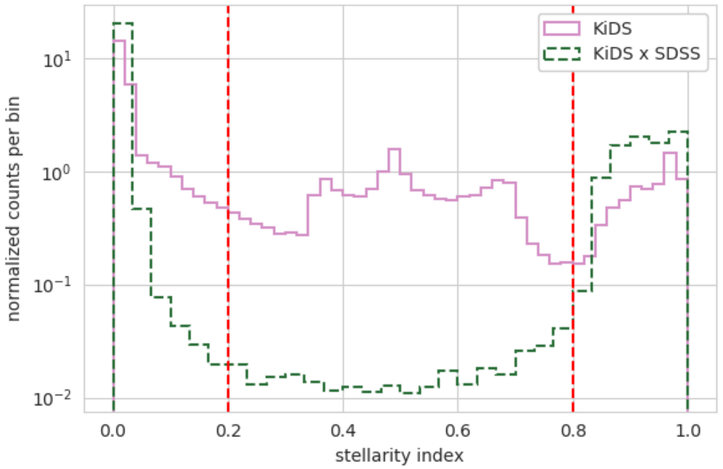


t-SNE projection of **KiDSxSDSS data**. *Left*: *r* magnitude, *right*: SDSS spectroscopic classification.

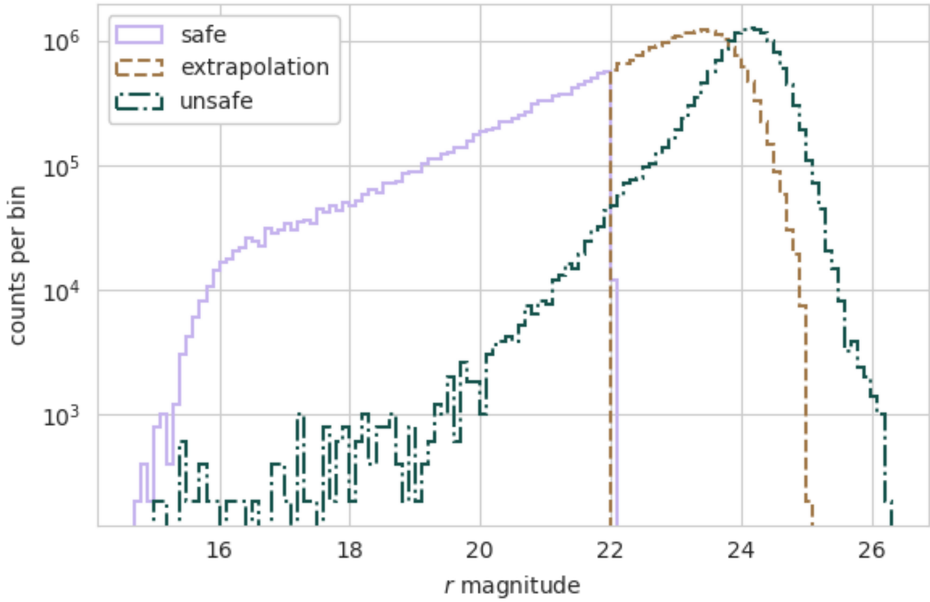
Inference subsets



Normalized histograms of the *r* magnitude for the KiDS inference data and the KiDS × SDSS training set.



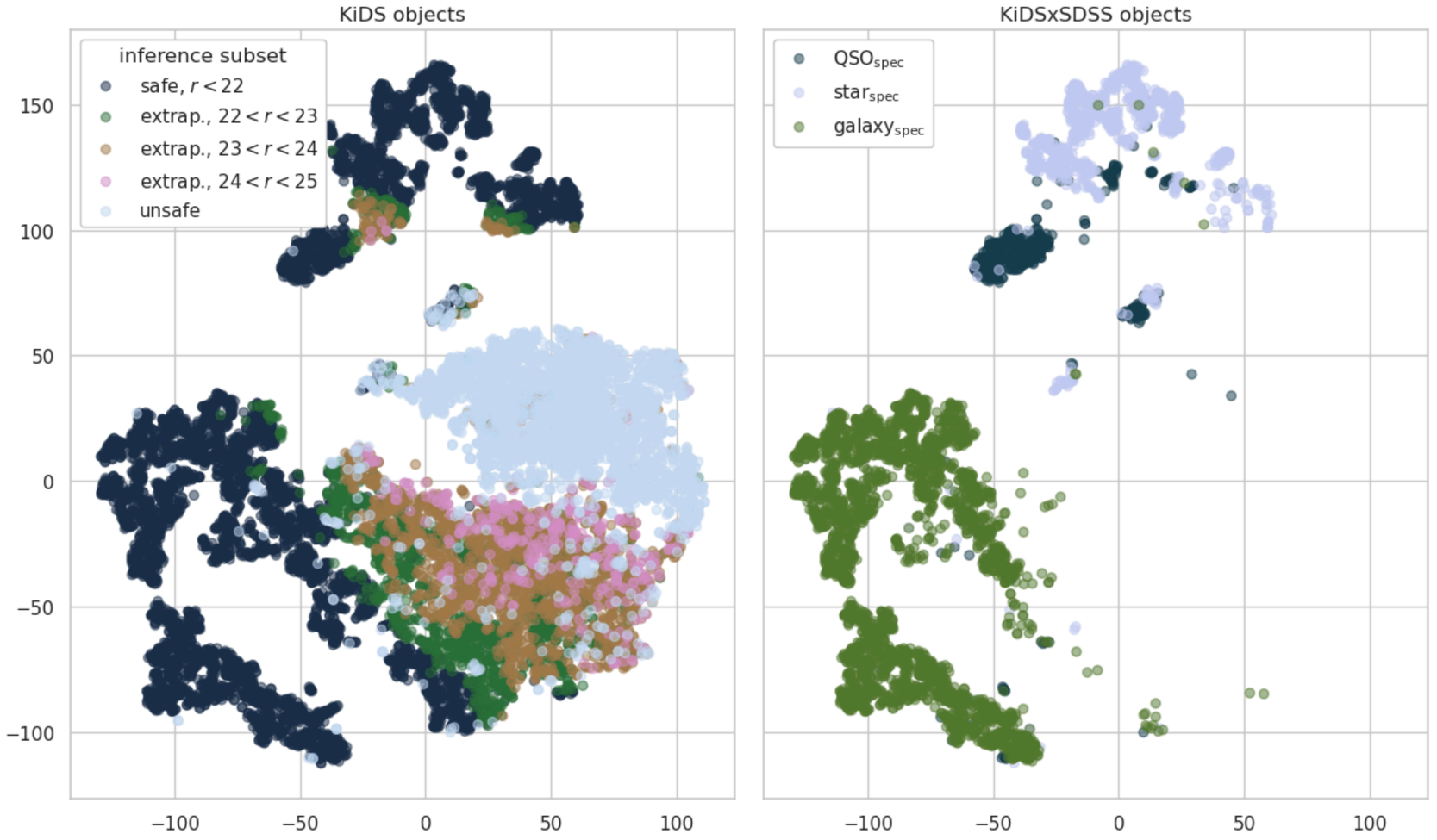
Normalized histograms of Sextractor's stellarity index for training and inference datasets.



Distribution of inference subsets over the *r* magnitude.

subset name	r magnitude	stellarity index	size
safe	r < 22	not in (0.2, 0.8)	21%
extrapolation	22 < r < 25	not in (0.2, 0.8)	45%
unsafe	r > 25	in (0.2, 0.8)	34%

Inference subsets

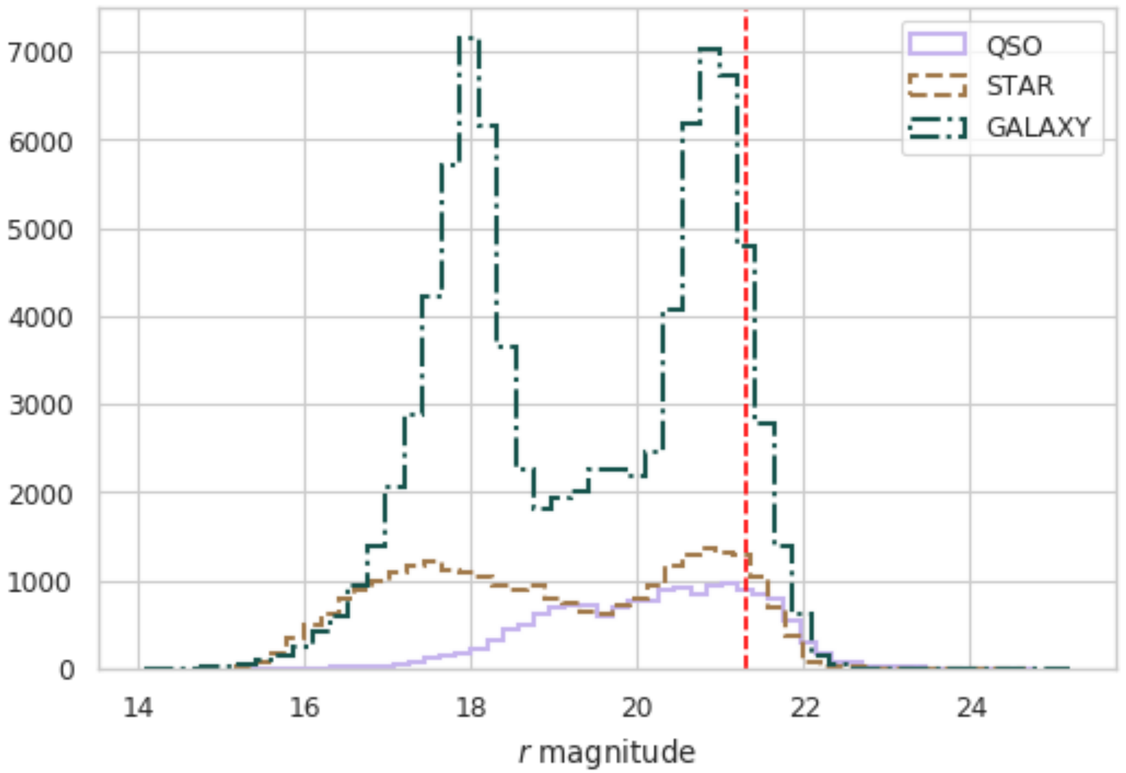


t-SNE projection of KiDSxSDSS data. *Left*: **inference subsets**, *right*: SDSS spectroscopic classification.

Testing

Using internal data structure

- Model training is based on random test
- Parameters tuning is based on random and faint extrapolation tests
- New models are trained for inference with parameters derived from experiments (except number of epochs/trees)

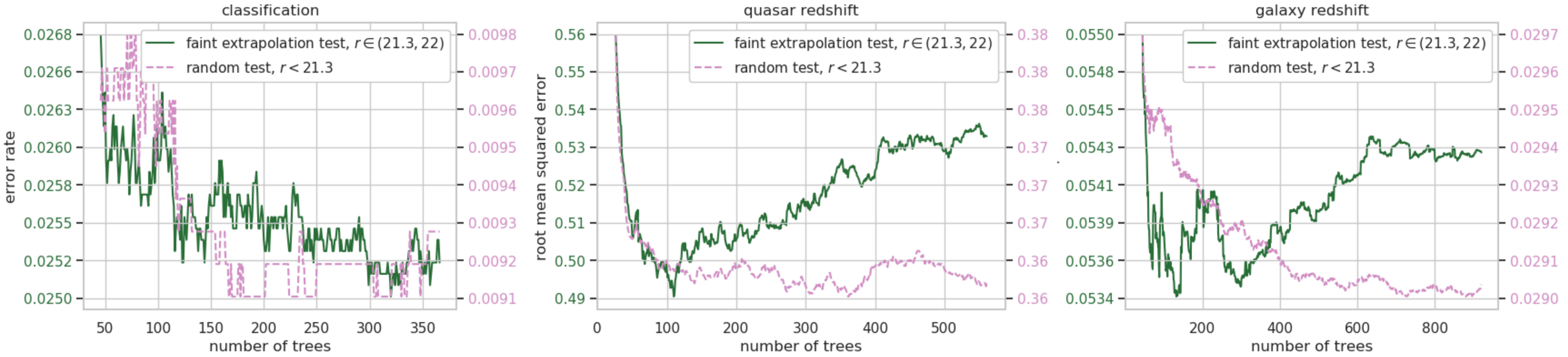


Distribution of spectroscopic classes in the training dataset. The red dashed line at $r = 21.3$ separates data used for the faint extrapolation test.

Experiment results

Testing

Method comparison



Learning histories for XGBoost. *Left:* classification, *center:* QSO redshift, *right:* galaxy redshift.

Model selection

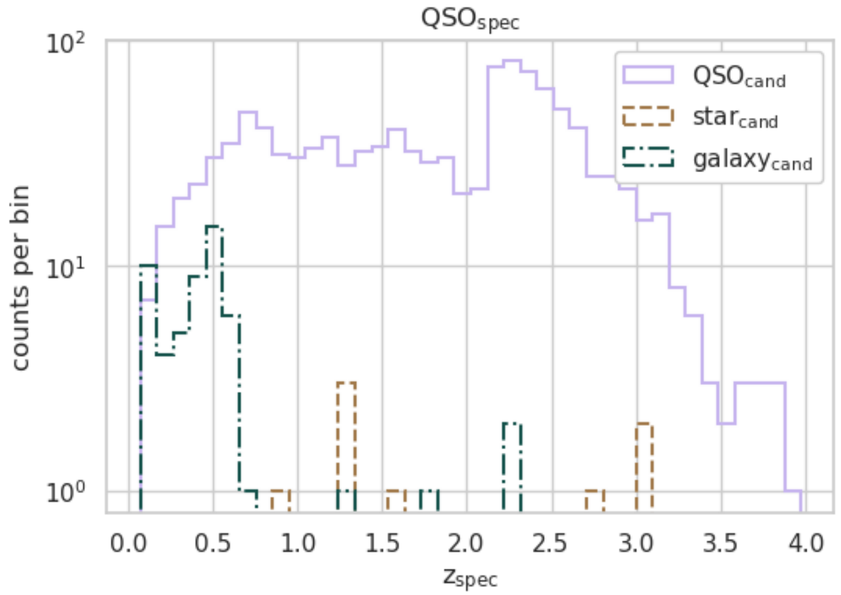
Experimental comparison

test	model	classification accuracy	classification purity	classification recall	redshift in QSO_{spec} MSE	redshift in QSO_{spec} R2	redshift in QSO_{spec} z error	redshift in QSO_{photo} MSE	redshift in QSO_{photo} R2	redshift in QSO_{photo} z error
random	RF	99.00%	97.44%	94.31%	0.12	85%	0.018 ± 0.14	0.12	84%	0.032 ± 0.21
	XGB	**99.09%**	**97.85%**	**94.75%**	0.13	84%	0.017 ± 0.15	0.13	83%	0.030 ± 0.21
	ANN	98.98%	96.93%	94.67%	**0.10**	**88%**	**0.009 ± 0.12**	0.11	85%	0.023 ± 0.22
	clf XGB, z ANN	**99.09%**	**97.85%**	**94.75%**	**0.10**	**88%**	**0.009 ± 0.12**	**0.10**	**87%**	**0.020 ± 0.19**
faint extrap.	RF	**97.44%**	96.12%	**92.37%**	0.31	31%	0.019 ± 0.25	0.33	31%	0.046 ± 0.38
	XGB	**97.44%**	96.48%	92.12%	0.27	39%	0.036 ± 0.23	0.34	29%	0.077 ± 0.41
	ANN	97.27%	**96.52%**	90.89%	**0.22**	**51%**	**-0.0004 ± 0.19**	**0.28**	**39%**	**0.042 ± 0.37**
	clf XGB, z ANN	**97.44%**	96.48%	92.12%	**0.22**	**51%**	**-0.0004 ± 0.19**	0.31	35%	0.050 ± 0.40

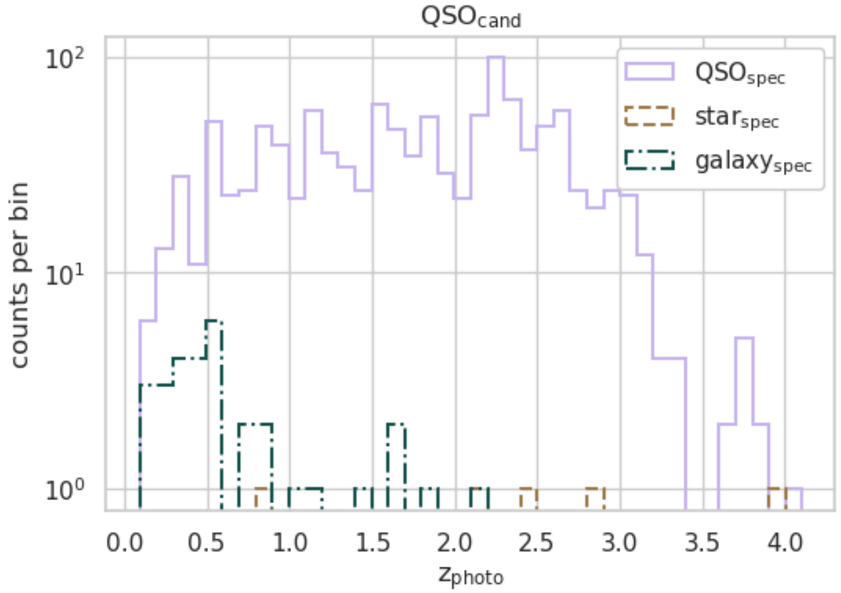
Final method: classification network
+ QSO redshift network with Gaussian uncertainty

QSO classification

	3-class accuracy	purity	completeness
randomly selected (r < 21.3)	99.0%	96.9%	94.7%
faint extrapolation (21.3 < r < 22)	97.3%	96.5%	90.9%



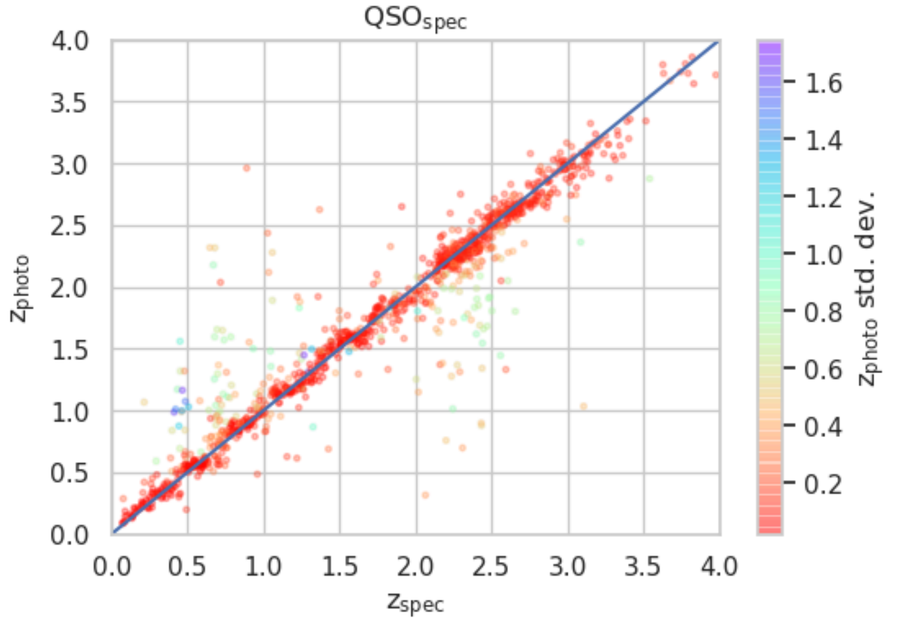
Completeness test

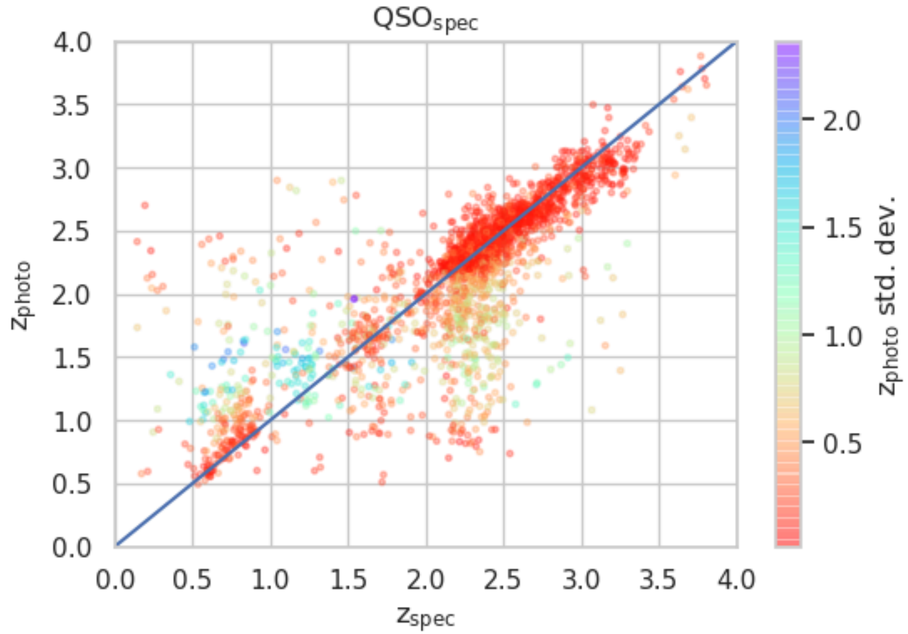


Purity test

QSO photo-zs

	$\delta z = \frac{(z_{photo} - z_{spec})}{(1 + z_{spec})}$
randomly selected (r < 21.3)	0.009 +/- 0.12
faint extrapolation (21.3 < r < 22)	-0.0004 +/- 0.19



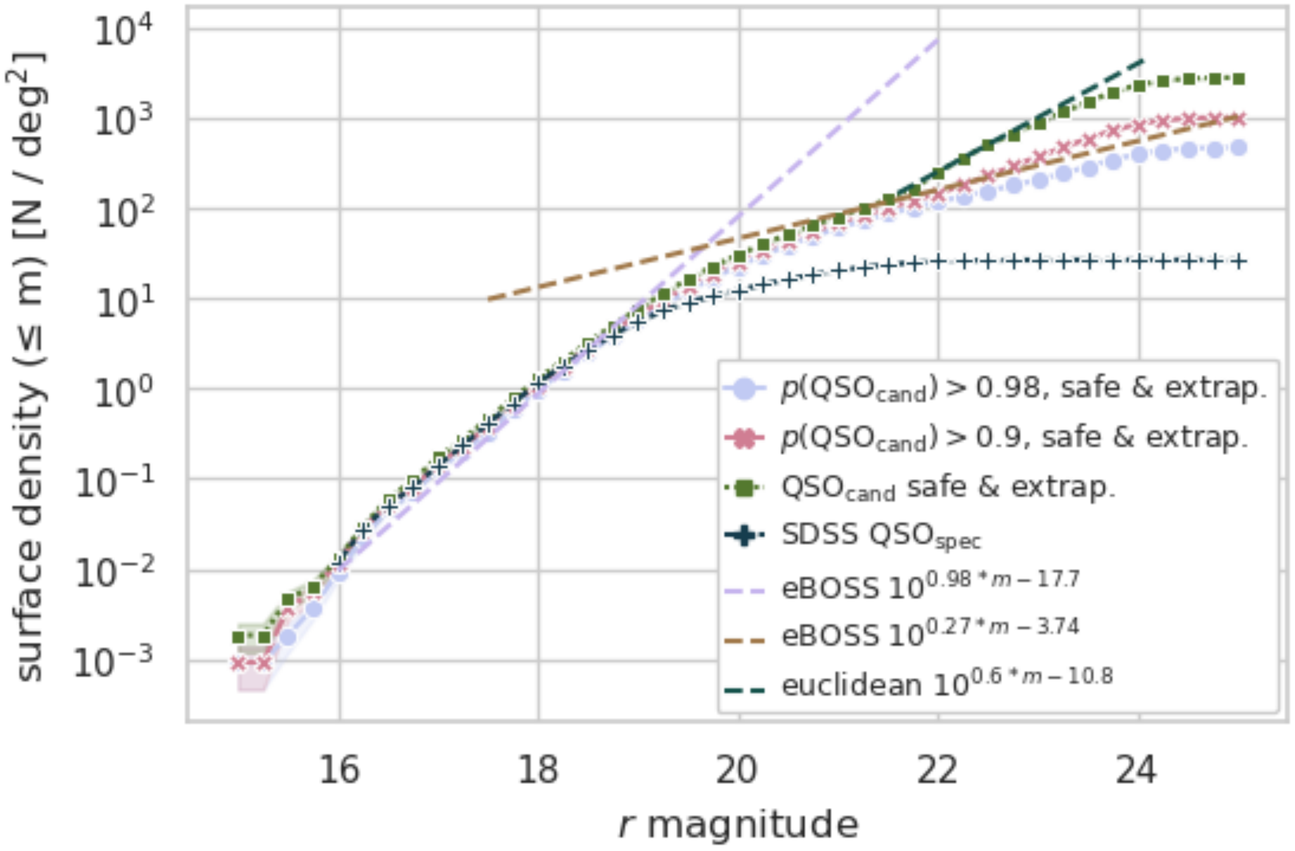


Comparison of the spectroscopic and photometric redshifts on SDSS quasars. *Left:* random test ($r < 21.3$), *right:* faint extrapolation test ($21.3 < r < 22$)

Final catalog properties

Number counts

Completeness of the catalog

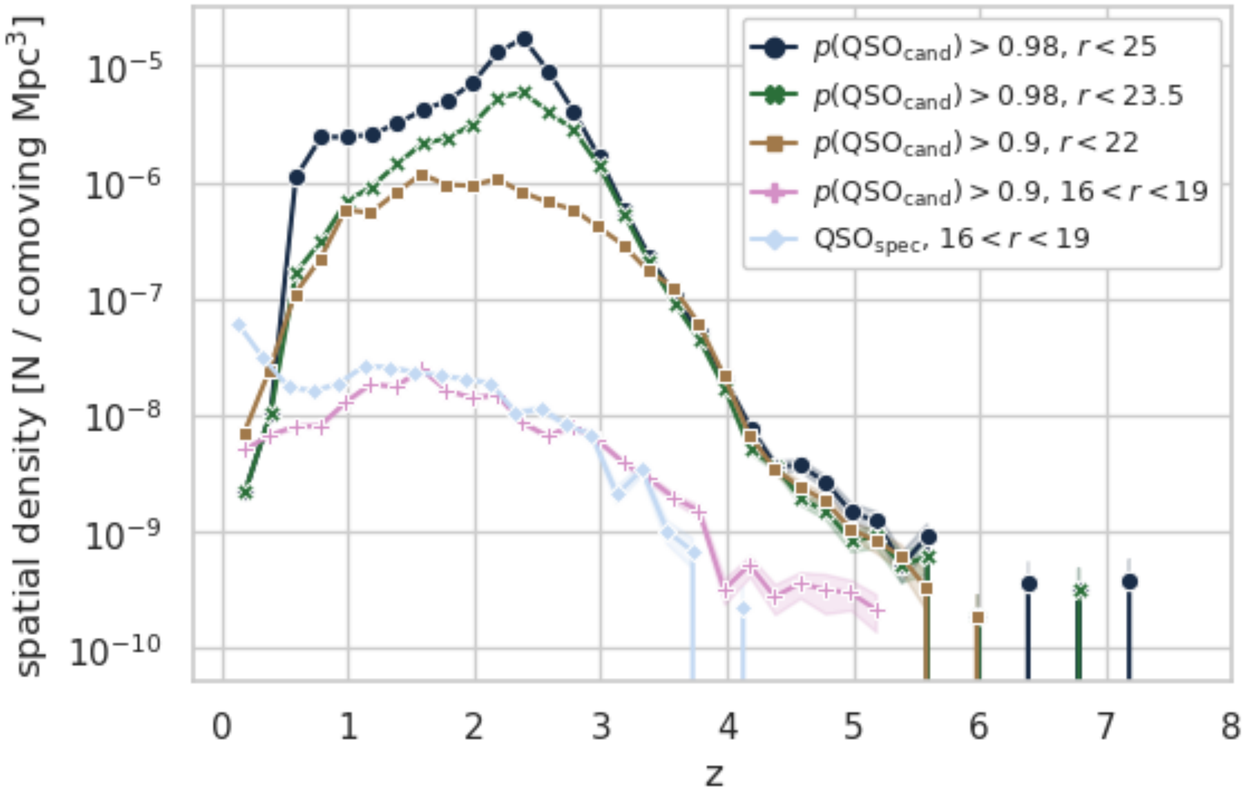


KiDS photometric QSO catalog at progressing probability cuts

- Straight dashed lines: quasar model from eBOSS
- Expected number counts at $p(\mathrm{QSO_{cand}}) > 0.98$
- Limited by the extrapolation subset which stops growing at $r > 23.5$

Spatial number density

Redshift reliability

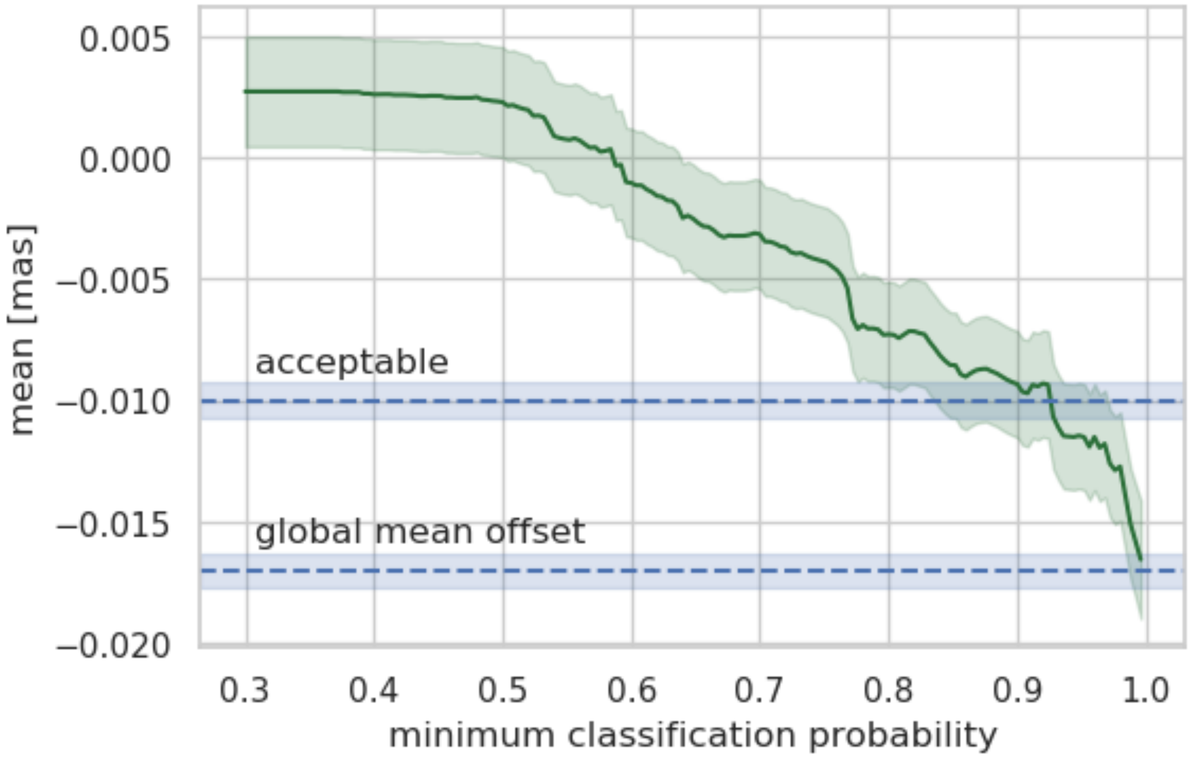


Spatial number density for KiDS photometric QSO at progressing magnitudes and corresponding probability cuts.

- Good result at the SDSS completeness limit r < 19
- Expected peek at z > 2 for p($\mathrm{QSO_{cand}}$) > 0.98 and r < 23.5
- Contamination with low-z galaxies at r > 23.5

Parallax test

Purity of the catalog



Mean parallax value for photometric quasars as a function of minimum classification probability.

The final catalog

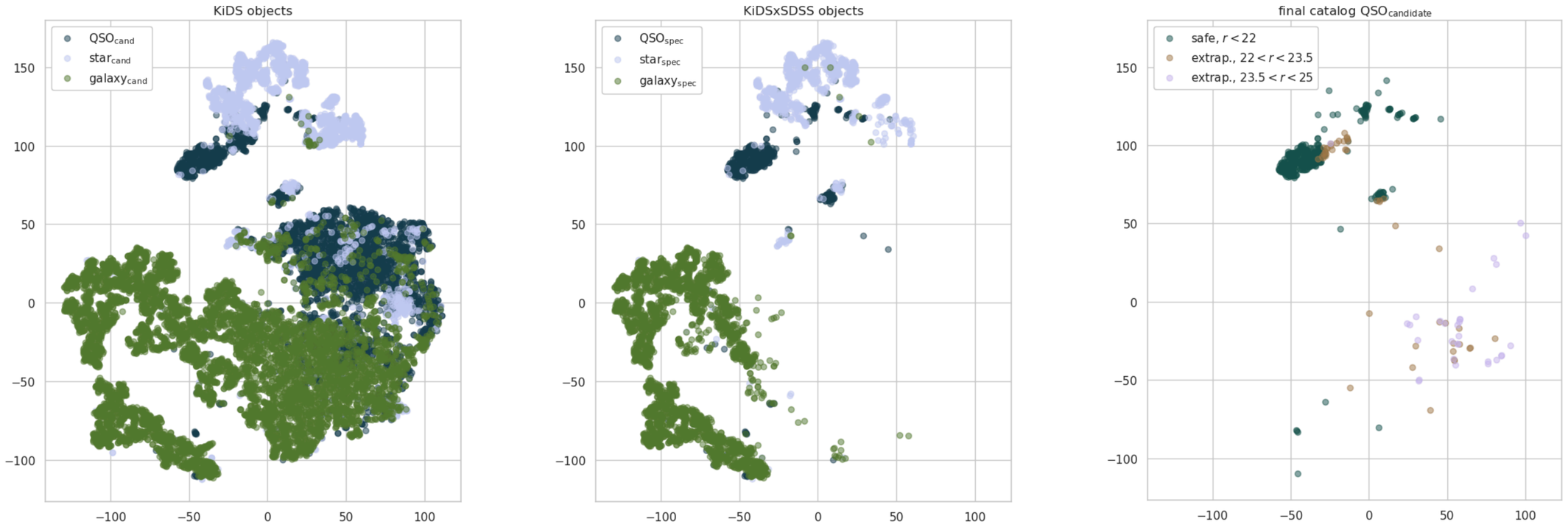
Suggested probability cuts

	safe $r < 22$	safe & extrap. $r < 23.5$	safe & extrap. $r < 25$
$\mathrm{QSO_{cand}}$	266k (100%)	1.6M (100%)	3M (100%)
$p(\mathrm{QSO_{cand}}) > 0.90$	**158k (59%)**	637k (39%)	1.1M (36%)
$p(\mathrm{QSO_{cand}}) > 0.98$	127k (48%)	**311k (19%)**	507k (17%)

Suggested probability cuts at progressing magnitudes

Classification

Visualisation



t-SNE projection of KiDS + KiDSxSDSS data. *Left*: photometric classification, *center*: SDSS spectroscopic classification, *right*: final photometric QSO catalog.

Applications

Catalog ready for:

- tomographic study of LSS
- AGN studies (e.g. SED fitting, possible test on photo-zs)

Main takeaways

Key points to successful ML application:

1. Data research and visualisations
   t-SNE visualisations, feature understanding, inference subsets

2. Tough model validation method
   Faint extrapolation test to fit bias vs variance trade-off

3. Physically-driven testing approach
   Number counts and GAIA parallaxes to fit purity vs completeness

Thank you!