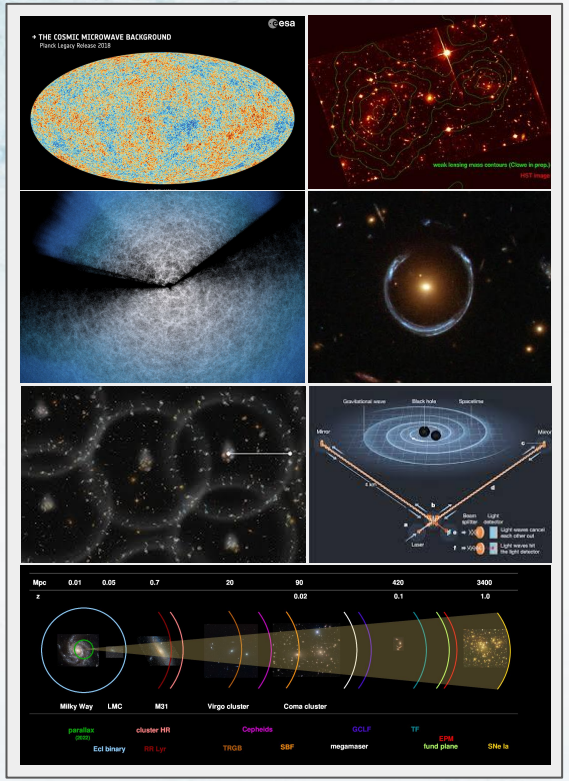
From spectroscopic redshift errors to AI-Assisted Diagnostics

– Shengyu HE

2026 AI+Astronomy

Observational cosmology in a nutshell

Experiments and Probes



Theory basis

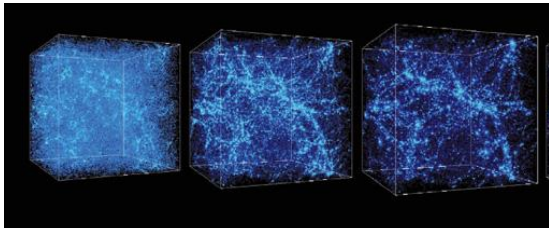
GR, perturbation …

$$G_{\mu\nu} = 8\pi G \sum_s T^{(s)}_{\mu\nu} \qquad s \in \{\gamma, \nu, b, c, \Lambda\}.$$

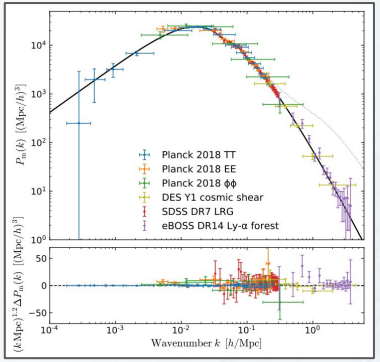
$$\nabla_\mu \sum_s T^{\mu\nu}_{(s)} = 0 \qquad g_{\mu\nu} = \bar{g}_{\mu\nu} + \delta g_{\mu\nu}$$

$$\frac{\partial f_s}{\partial t} + \frac{dx^i}{dt}\frac{\partial f_s}{\partial x^i} + \frac{dp_i}{dt}\frac{\partial f_s}{\partial p_i} = C_s[f]$$

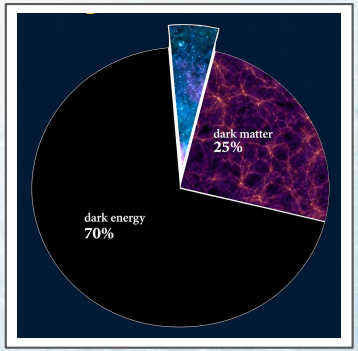
Cosmological Simulation



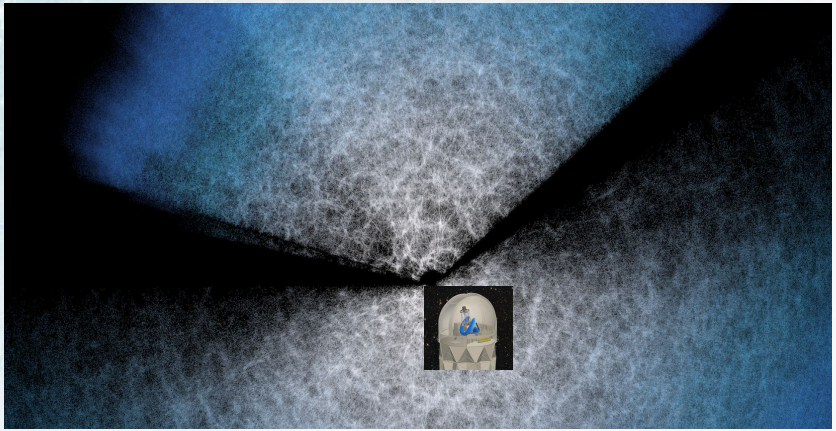
Summary statistics

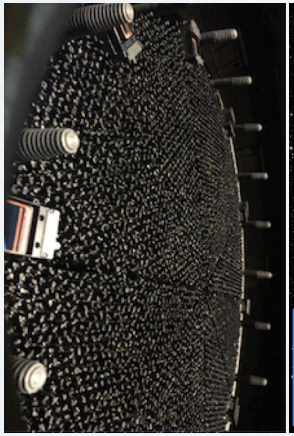


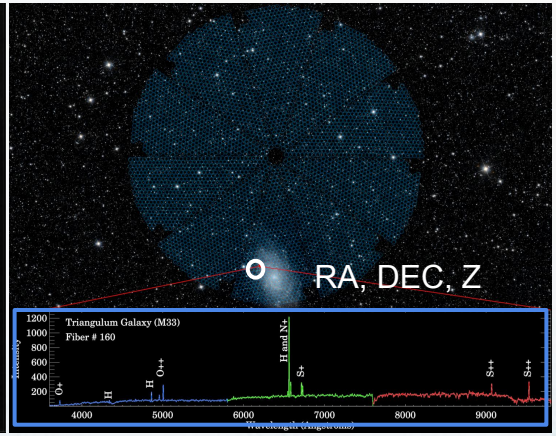
Cosmological inference



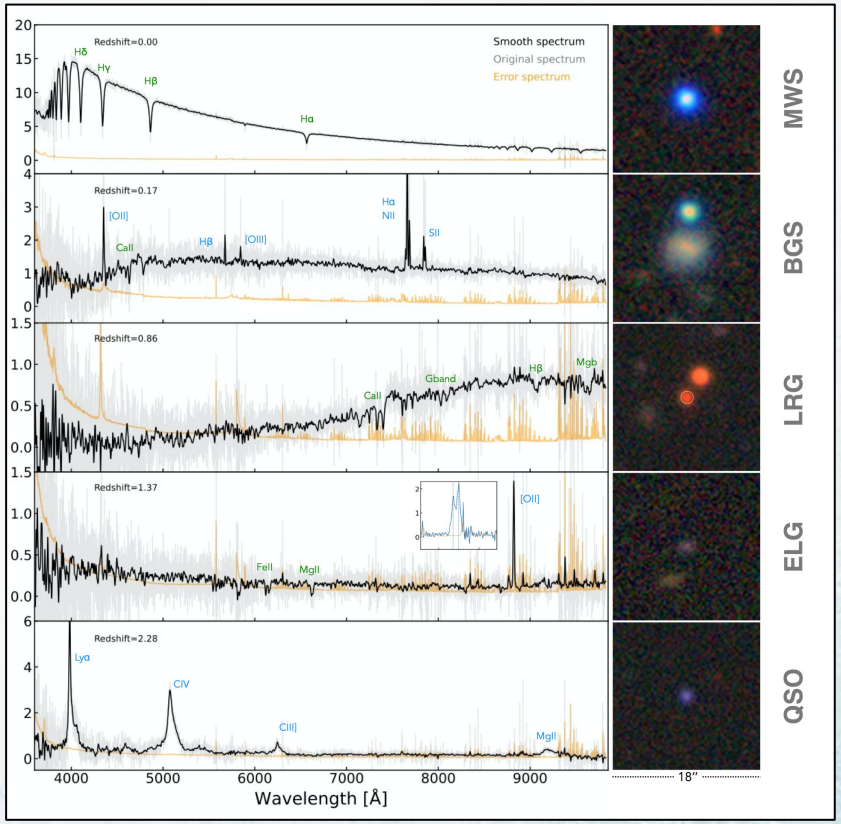
Spectroscopic survey







DESI spectrum

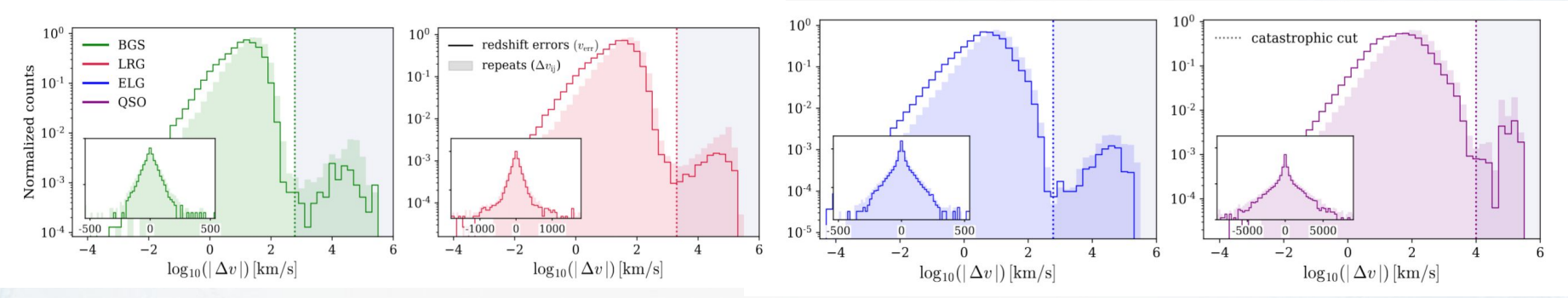


Use repeat observations to probe redshift errors

$$z_{\rm err} = \frac{z_{\rm meas} - z_{\rm true}}{1 + z_{\rm true}}$$

$$\Delta z_{ij} \equiv z_{{\rm err},i} - z_{{\rm err},j} = \frac{z_{{\rm meas},i} - z_{{\rm meas},j}}{1 + z_{\rm true}}.$$

He et al. 2026 in preparation, Assessing Spectroscopic redshift errors for DESI DR2

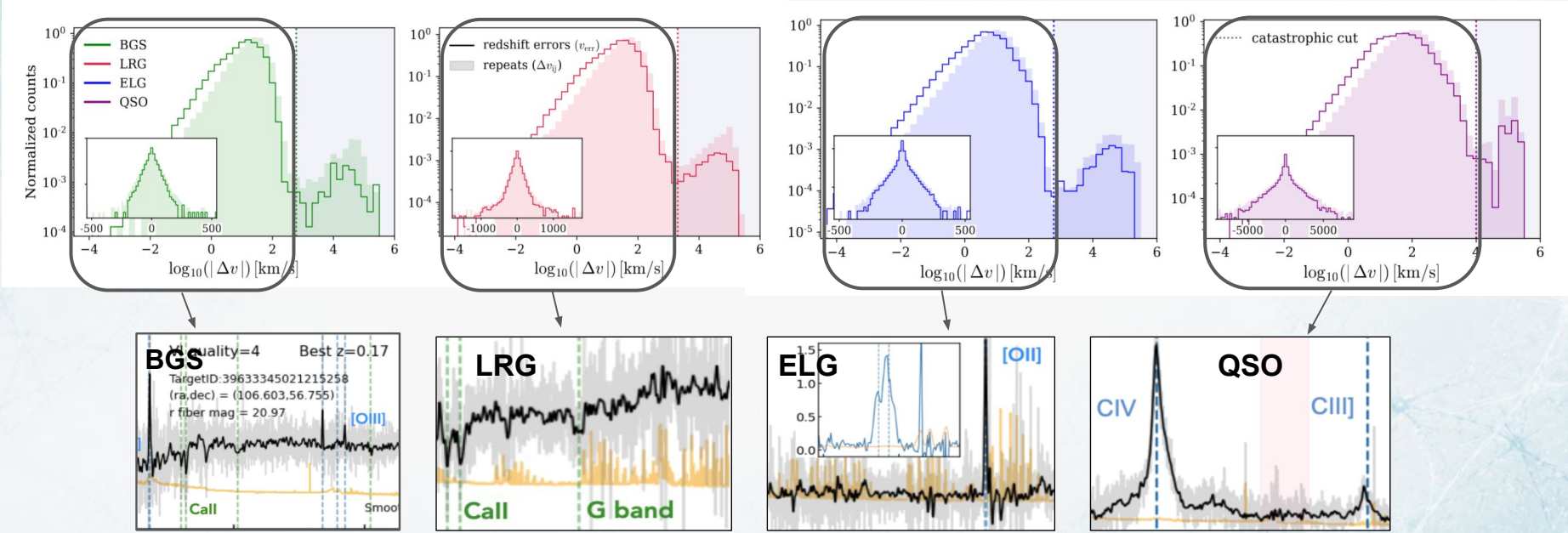


Use repeat observations to probe redshift errors

$$z_{\rm err} = \frac{z_{\rm meas} - z_{\rm true}}{1 + z_{\rm true}}$$

$$\Delta z_{ij} \equiv z_{{\rm err},i} - z_{{\rm err},j} = \frac{z_{{\rm meas},i} - z_{{\rm meas},j}}{1 + z_{\rm true}}.$$

He et al. 2026 in preparation, Assessing Spectroscopic redshift errors for DESI DR2



(Lan et al. 2023, Alexander et al. 2023; DESI visual inspection)

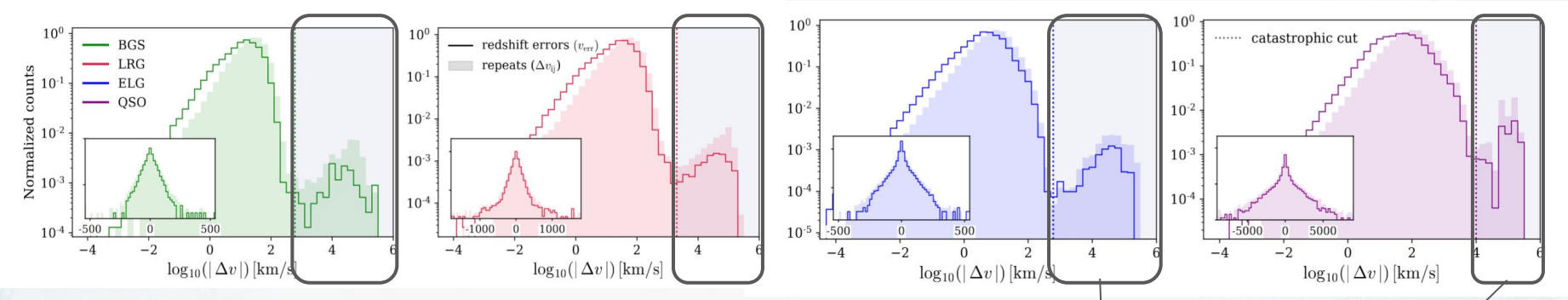
Redshift uncertainties: minor differences in measuring redshifts

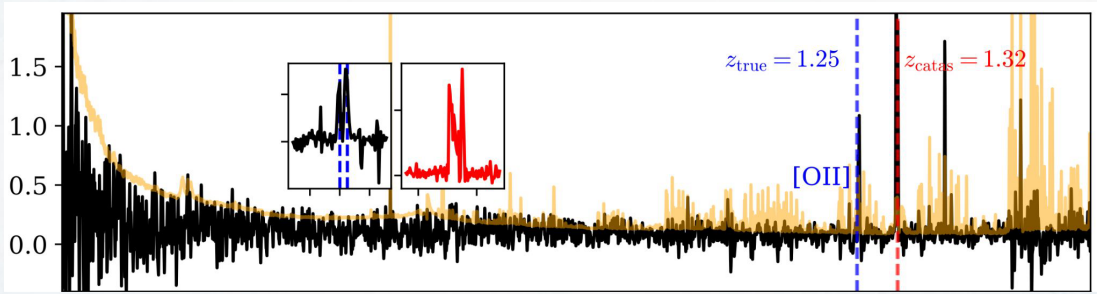
Use repeat observations to probe redshift errors

$$z_{\rm err} = \frac{z_{\rm meas} - z_{\rm true}}{1 + z_{\rm true}}$$

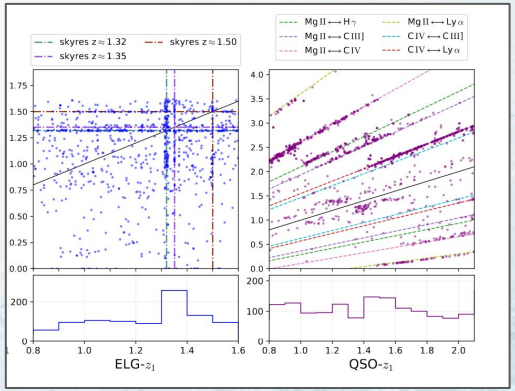
$$\Delta z_{ij} \equiv z_{{\rm err},i} - z_{{\rm err},j} = \frac{z_{{\rm meas},i} - z_{{\rm meas},j}}{1 + z_{\rm true}}.$$

He et al. 2026 in preparation, Assessing Spectroscopic redshift errors for DESI DR2



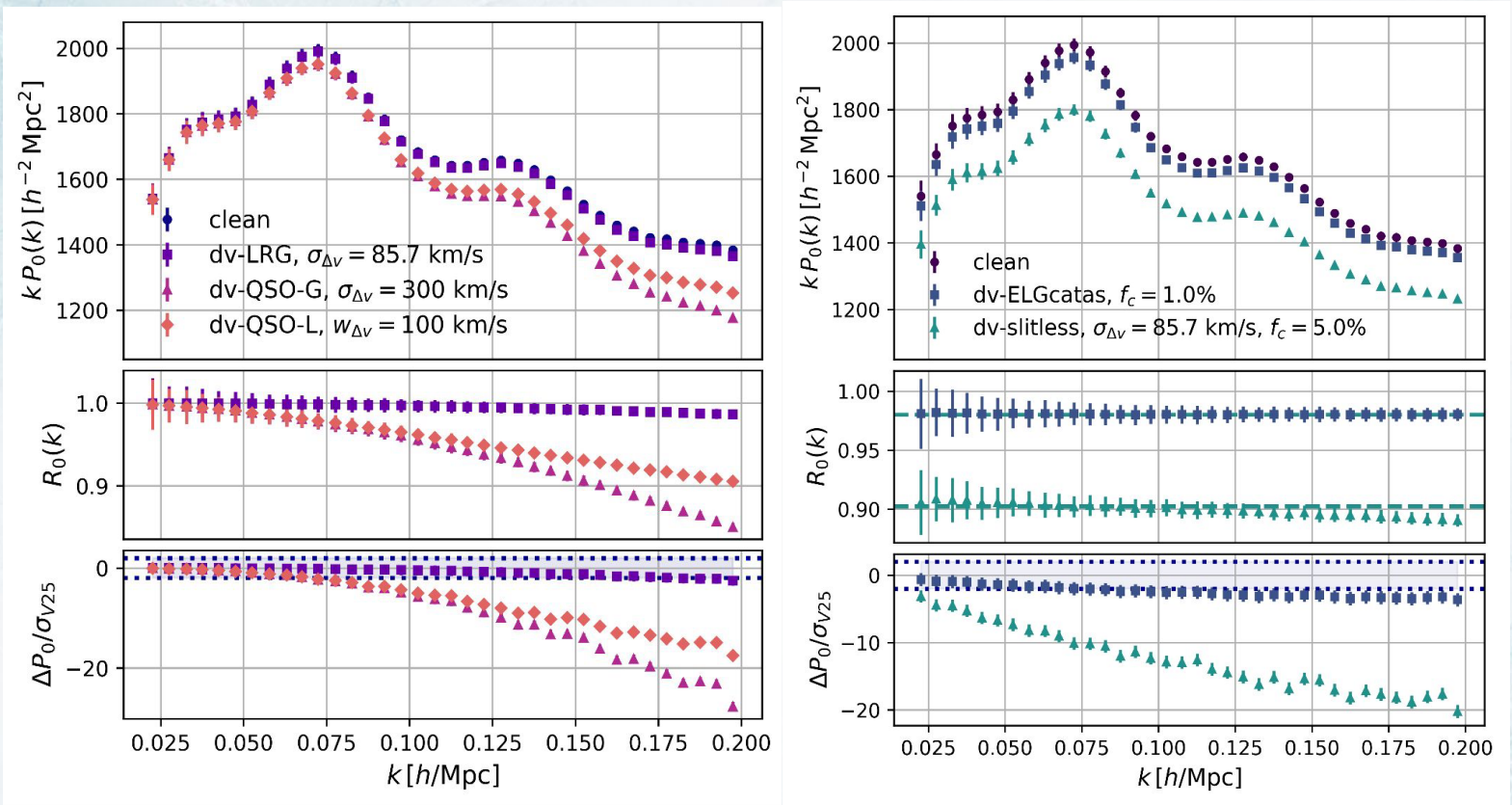


(Yu et al. 2025, ELG spectroscopic systematics)



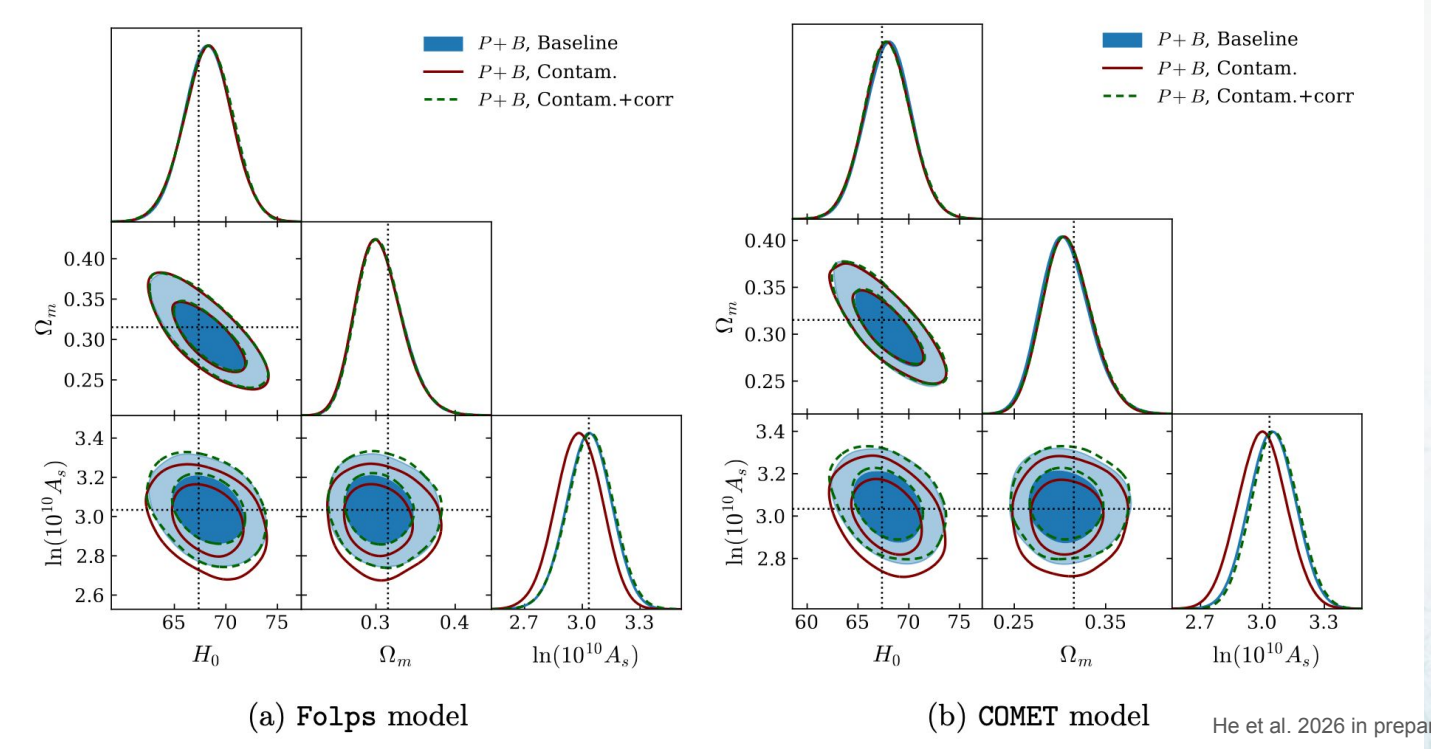
Redshift catastrophic failures: significant difference from its true value

Cosmological impact of redshift errors



Cosmological impact of redshift errors

Case 1: DESI DR2 QSO redshift errors

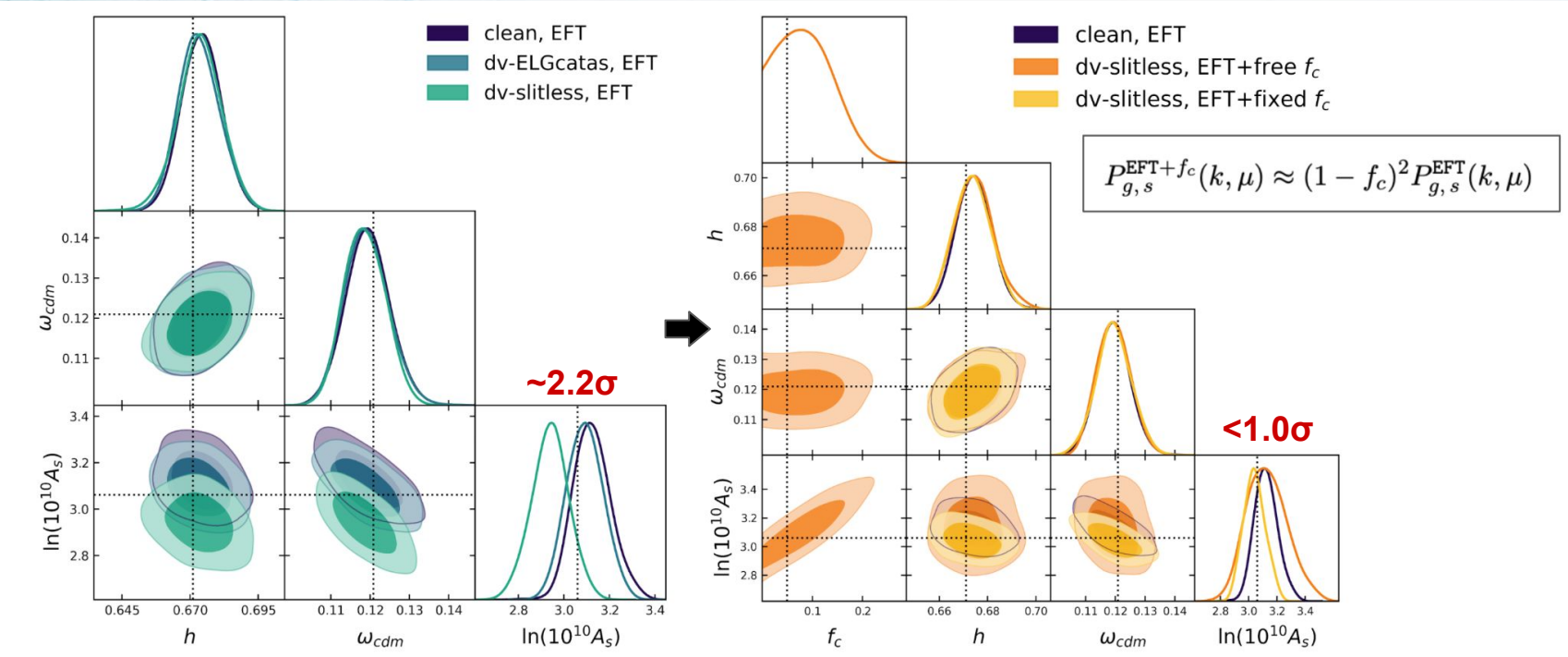


He et al. 2026 in preparation

With redshift errors correction, reduce logA bias from **~0.5σ** to **<0.1σ**

Cosmological impact of redshift errors

Case 2: Redshift Catastrophic failures fc=5% ⇒ slitless-like survey such as *Euclid* and CSST



He et al. 2025 – The impact of spectroscopic redshift errors on cosmological measurements

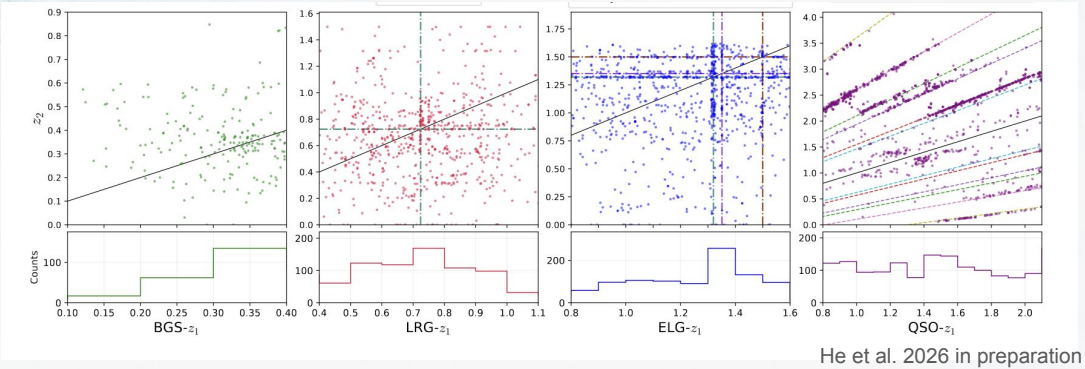
With catastrophic correction, reduce logA bias from **~2.2σ** to **<1.0σ**

What cause catastrophic redshift failures?

1. Catastrophic pairs in DESI DR2 :

BGS: 388 LRG: 1002

ELG: 1112 QSO: 3108



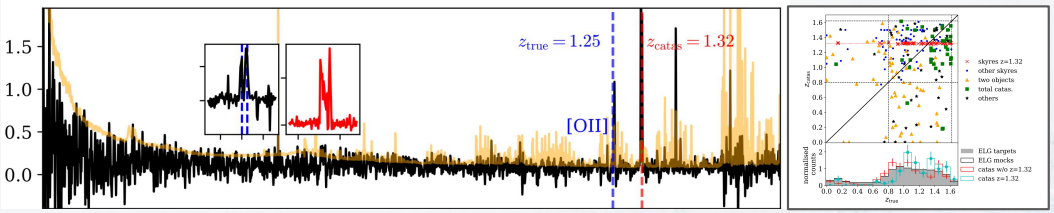
He et al. 2026 in preparation

2. Forward-modeling of DESI spectrum:

Peng et al. 2026, in preparation – Redshift Misidentification in DESI DR2 Emission-Line Galaxies

3. Visual Inspection:

~300 paris catastrophic spectra for ELG

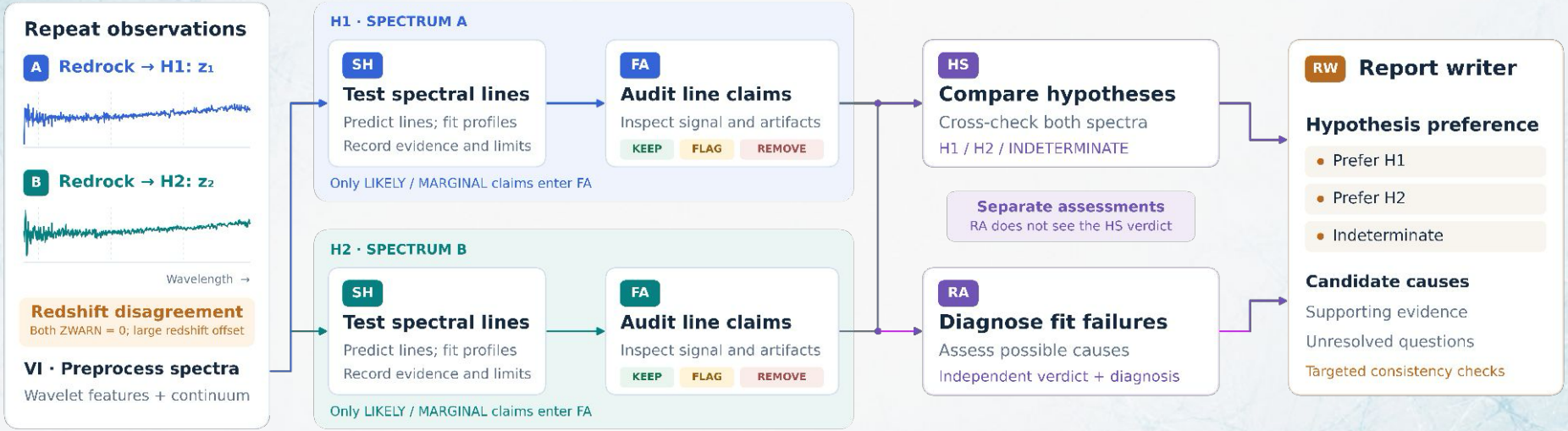


Yu et al. 2025 – DESI DR1 ELG Spectroscopic systematics

Can an agent investigate these disagreements and document the evidence?

CatasAgent (on-going)

Investigating redshift disagreements with spectral evidence



SH: Single Hypothesis **FA: Feature Auditor** **HS: Hypothesis Synthesis** **RA: Result Auditor** **RW: Report Writer**

HS and RA assess shared evidence independently
Real DESI spectra · binned for display

FORMA: Haosong's talk next session

Agent report

CatasAgent report

Report guide

Report essentials · What the main sections provide

Input facts (calculated) Observation metadata, Redrock fits, object types, and fit diagnostics.

Exposure	Tile	Night	Petal	Redrock z	SPECTYPE	ZWARN	DELTACHI2
A	X	X	X	X	X	X	X
B	X	X	X	X	X	X	X

Overview

How H1 and H2 fit the evidence, which redshift is preferred, and why.

Proposed catastrophe source

RA's proposed failure mechanism, supporting reasoning, and alternative explanations.

Decisive evidence

The strongest spectral features and measurements that support or reject each redshift.

Assessment

HS and RA verdicts, their agreement or disagreement, and RA's confidence.

Evidence conflicts and limits

Remaining disagreements and limits on the redshift, object type, and proposed cause.

Follow-up

Recommended next checks or observations, and which uncertainties they would resolve.

Core report sections · Full reports also include detailed stage and audit history

Example:

```
## Input facts (calculated)

| Exposure | Tile | Night | Petal | Redrock z | SPECTYPE | ZWARN | DELTACHI2 | Wavelength range (Å) |
|---|---|---|---|---|---|---|---|---|
| A | 7712 | 20240114 | 6 | 0.226403088734664 | GALAXY | 0 | 253.69595907256007 | 3600.0–9824.0 |
| B | 9263 | 20240403 | 2 | 2.9521562050680608 | QSO | 0 | 2610.1402104124427 | 3600.0–9824.0 |

**|Δz| = 2.726**

## Overview

Both synthesis stages prefer H2: the object is most plausibly a broad-line QSO near z ≈ 2.95, with the exact systemic redshift unconfirmed. RA attributes the A/B redshift catastrophe to `noise_overfit` — Redrock's GALAXY fit at z = 0.2264 in the noise-dominated exposure A (median S/N 2.0) locking onto a spurious χ² minimum — an inferred mechanism, not a reproduced test. The decisive evidence is a broad (~150 Å) emission hump at ~4805 Å, present in both exposures, matching H2's Lyα prediction at 4805.8 Å and unexplained by any H1 line; H1's only two FA-eligible claims were both removed as noise. The most consequential uncertainty is that H2 rests on this single feature: C IV was removed by FA, no Lyα forest or DLA is detected, and alternative identifications of the hump (C IV at z ≈ 2.10, Mg II at z ≈ 0.72) were not excluded.

## Proposed catastrophe source

RA's saved `catastrophe_type` is `noise_overfit`: in exposure A's noise-dominated spectrum (median S/N 2.0, DELTACHI2 = 254), Redrock's GALAXY template is assessed to have settled on a spurious local χ² minimum, while the QSO template in exposure B anchored correctly on the broad hump as Lyα. This mechanism is inferred from the absence of any surviving H1 feature and the weak DELTACHI2, not demonstrated by a recorded template re-fit; per the pipeline's own caution, lack of support for H1 does not by itself prove noise overfitting, and RA's MEDIUM confidence applies to its overall assessment. Separately from the mechanism, there is a genuine object-type disagreement (GALAXY vs QSO) between the two Redrock runs; the spectral evidence — one very broad emission line plus a blue continuum — disfavors A's GALAXY label, but with only one confirmed line the classification remains a candidate, not established. Remaining alternatives RA could not exclude: shared pipeline systematics (sky subtraction, flux calibration, templates) common to both coadds, line-identification ambiguity for the single hump, and fiber/source confusion.

## Decisive evidence
```

Set up and computational cost

CSCS – Swiss National Supercomputing Centers

DESI spectra

A Repeat observation 1



B Repeat observation 2



Wavelength

Stored at CSCS

CatasAgent

SH · FA · HS · RA · RW

Same model · role-specific prompts

Scientific tools

Line fitting · spectrum reads · physics KB

Runs on CSCS

API

Requests →

← Responses

LLM inference service

GLM-5.3

zai-org/GLM-5.3

Reasoning effort: high

Hosted and operated by CSCS

One Catastrophic failure analysis:

≈ 25 min total / 55k tokens

Prepare spectra	SH	FA	HS + RA	RW
≈ 6 s	≈ 8 min	≈ 5.2 min	≈ 10 min	≈ 2 min
Load · CWT · continuum	H1 + H2 in parallel	H1 + H2 in parallel	HS ≈ 5 min · RA ≈ 10 min	Write report

Averages over 16 completed targets · two DESI spectra per target · parallel stages overlap

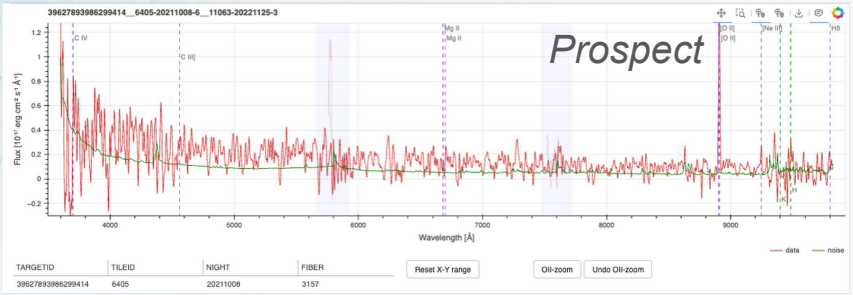
Total and HS/RA timed directly; other stages estimated from output timestamps. Queue time excluded.

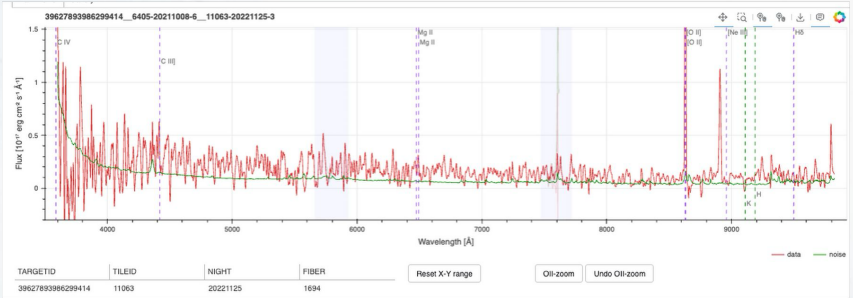
Preliminary tests with visual inspection

Visual inspection of 16 catastrophic cases

2 BGS, 7 LRG, 4 ELG, 3 QSO

Example: ELG / sky_1p315 · TARGETID 39627893986299414





Catastrophic sources proposed by RA

Sky confusion: 2 · Line confusion: 2 · Artifact: 2 · Noise overfit / unclassified: 10

Original 16-pair VI sample · Counts describe proposed causes, not validated accuracy

CatasAgent–VI agreement

GLM-5.3 · high

16 test pairs · Preliminary performance comparison

Overlap in failure diagnosis and redshift choice

Cause comparison: 9 pairs · Redshift preference: 10 pairs*

6 **Same catastrophic source**
Noise overfit (2) · artifact (1) · sky (2) · line confusion (1)

3 **Same issue, different main label**
Artifact · sky residuals · wrong object type

7 **Same redshift preference**
5 both agents agree · 2 synthesis agrees; auditor undecided

6 pairs agree on both the issue and a redshift preference.

Improving VI and agent diagnoses

VI can miss details; the agent can flag overlooked issues.

Future work: complete VI and assess diagnostic accuracy.

Counts overlap; this preliminary comparison is not an accuracy estimate.
* Includes 2 tentative cause notes and 5 tentative/inferred redshift preferences.

Improvements and future applications

CatasAgent: next improvements

PROPOSED NEXT STEPS

From describing spectra to explaining catastrophic redshift failures

1 Investigate the cause

Go beyond “low S/N”

Identify each fit’s anchor.

Check sky, artifacts, line IDs, and template/type errors.

Cause + evidence

Keep unresolved causes explicit.

2 Test both redshifts

2 spectra × 2 hypotheses

	z_1	z_2
Spectrum A	Tested	Test
Spectrum B	Test	Tested

Reuse preparation to limit cost.

Extra tests increase token use.

3 Consistent evidence

Shared measurements

Feature strength + uncertainty

Noise + masks + repeatability

Specific diagnosis

Primary cause + contributors

Independent agent judgments

4 Clarify the report

Clear conclusions first

Preferred redshift + why

Proposed cause

Decisive evidence + limits

Short, focused, little repetition.

Future validation

Complete VI · compare cause and redshift agreement · measure tokens and runtime

VI can miss details; agent findings can help refine the reference labels.

Improvements and future applications

CatasAgent: future work & framework extensions

FUTURE DIRECTIONS

Systematic failure analysis and pipeline validation, with broader uses of the framework

CATASAGENT WORK

1 DESI DR2 catastrophe analysis

Build a catalogue of causes and evidence.
Study trends with target type, redshift, S/N, and observing conditions.

A systematic analysis of catastrophic failures

2 Quality control & pipeline validation

Find recurring instrumental or reduction problems.
Compare pipeline versions on the same targets.
Check which failure mechanisms are resolved.

Validate improvements in spectral analysis

FRAMEWORK EXTENSIONS · BEYOND CATASAGENT

3 DESI × 4MOST

Compare matched spectra to explain survey differences.
Inform 4MOST exposure needs and survey strategy.

4 Search for unusual spectra

Flag blends, unusual features, or changes between epochs.
Check artifacts before pursuing astrophysical explanations.

Compare matched targets and successful controls; account for depth, resolution, coverage, and selection.

Take away messages

1. **Redshift errors matter for cosmology**

   Repeat observations reveal uncertainties and catastrophic failures.
   Modelling their impact can reduce biases in cosmological inference.

2. **CatasAgent investigates why redshifts disagree**

   Initial tests identify issues also noted during visual inspection.
   Reports link proposed failure causes to measured spectral evidence.

3. **Next: a systematic analysis of DESI DR2**

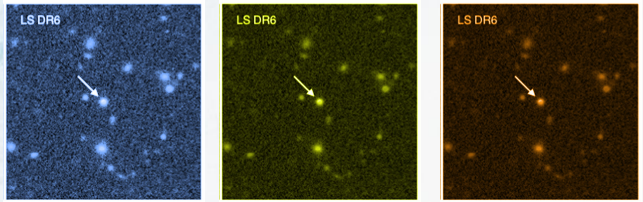
   Validate failure causes and redshift choices; build a failure catalogue
   to support quality control and pipeline improvements.

Goal: automated, evidence-based diagnosis at survey scale.

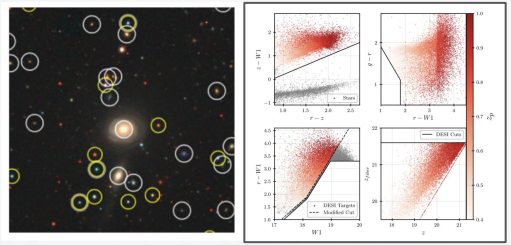
Backup

Galaxy redshift survey

Imaging survey

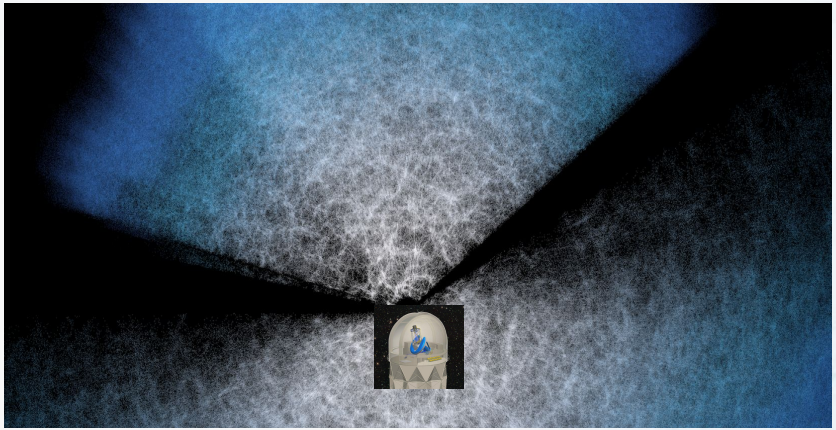


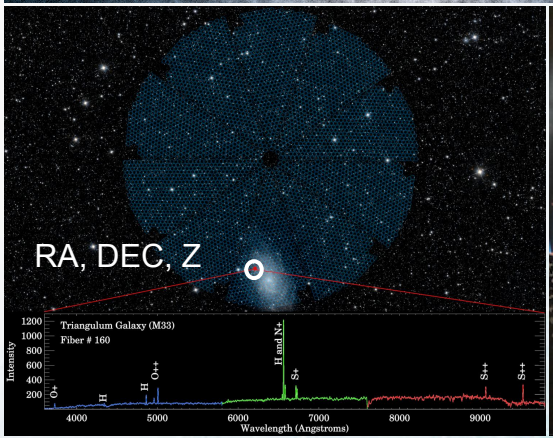
Target Selection

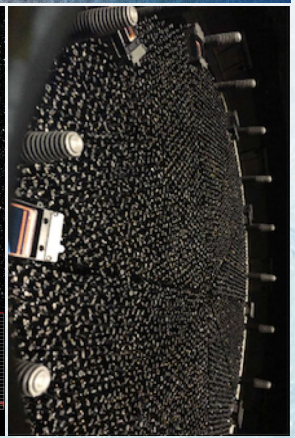


Verdier et al. 2026, 4MOST target selection

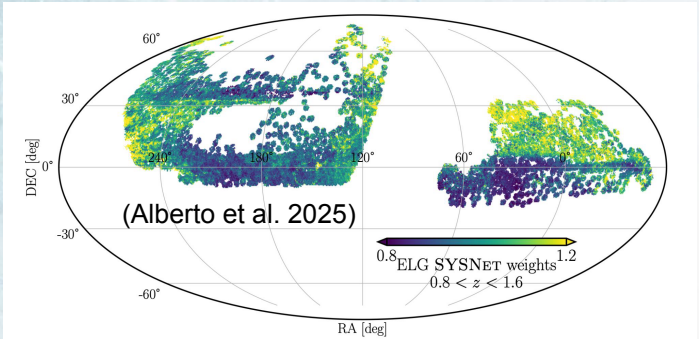
Spectra and redshifts





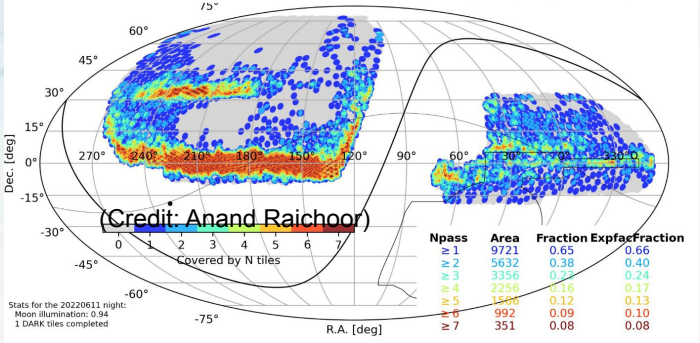


Observational systematics in DESI



Imaging Systematics

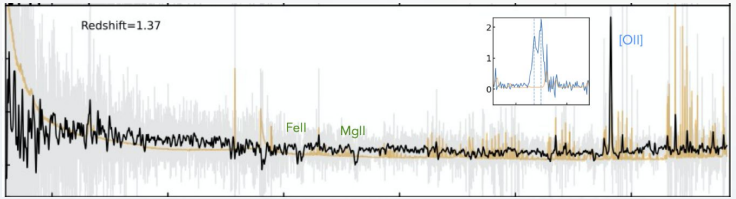
Weight Map



Fiber Incompleteness

Θ-cut /AUW

(Pinon et al. 2025)

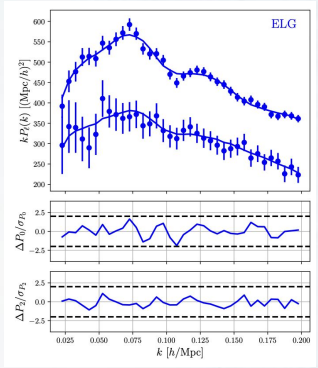


Spectroscopic Redshift errors

Theory Correction

	RA	DEC	redshift	weights
Galaxy 1				
...				
Galaxy N				

Validation with Model Galaxies Mock catalogs



Unbiased cosmological results

