Machine Learning Applied to Millions of DESI Spectra

Geometric Spectral Modeling

Andy Morgan & Stéphanie Juneau

NSF NOIRLab & NSF-Simons CosmicAI Institute

Discovering Our Universe Together

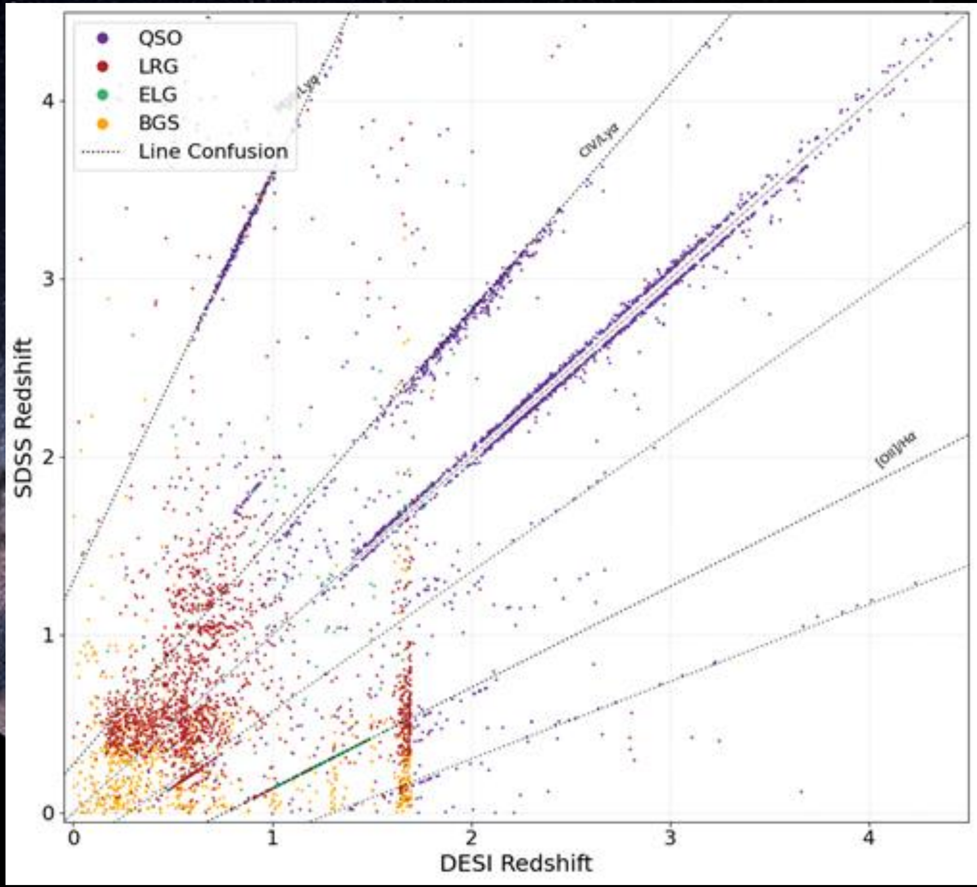
What is DESI?

The Dark Energy Spectroscopic Instrument (DESI) is a highly parallel fiber-fed spectrograph that studies the evolution of dark energy by collecting and analyzing extragalactic redshifts (distances)

Our redshift measurements are not always accurate

Relative failure rate of ~0.5% in the nearby and bright regime for confident objects → hundreds of thousands of failures

Want to flag and fix them in an automated way when possible. This is a conditional probability problem: p(redshift | spectrum)



Andy Morgan andy.morgan@noirlab.edu

Beyond Redshifts; Necessity is the Mother of Invention

Goal: Flag and fix redshifts → Need: p(redshift | spectrum)

While we're at it: Want p(spectrum) ← Outlier detection, population statistics, sampling, etc

Which means: We need a general purpose spectroscopic foundation model for DESI spectra to generate a suitable representation for modeling p(spectrum) in the first place

Andy Morgan andy.morgan@noirlab.edu

Not Another Autoencoder

The latent space of an autoencoder is:

1. Uninterpretable: Requires significant post-training probing to decipher. Science demands interpretability
2. Ungrounded/Agnostic: Latent representations of autoencoders lack a ground-truth. They are a function of downstream architecture and initialization
3. Non-metric: Latent geometry is uncorrelated with the data geometry (distances, densities, angles). Not suitable for evaluating p(spectrum)

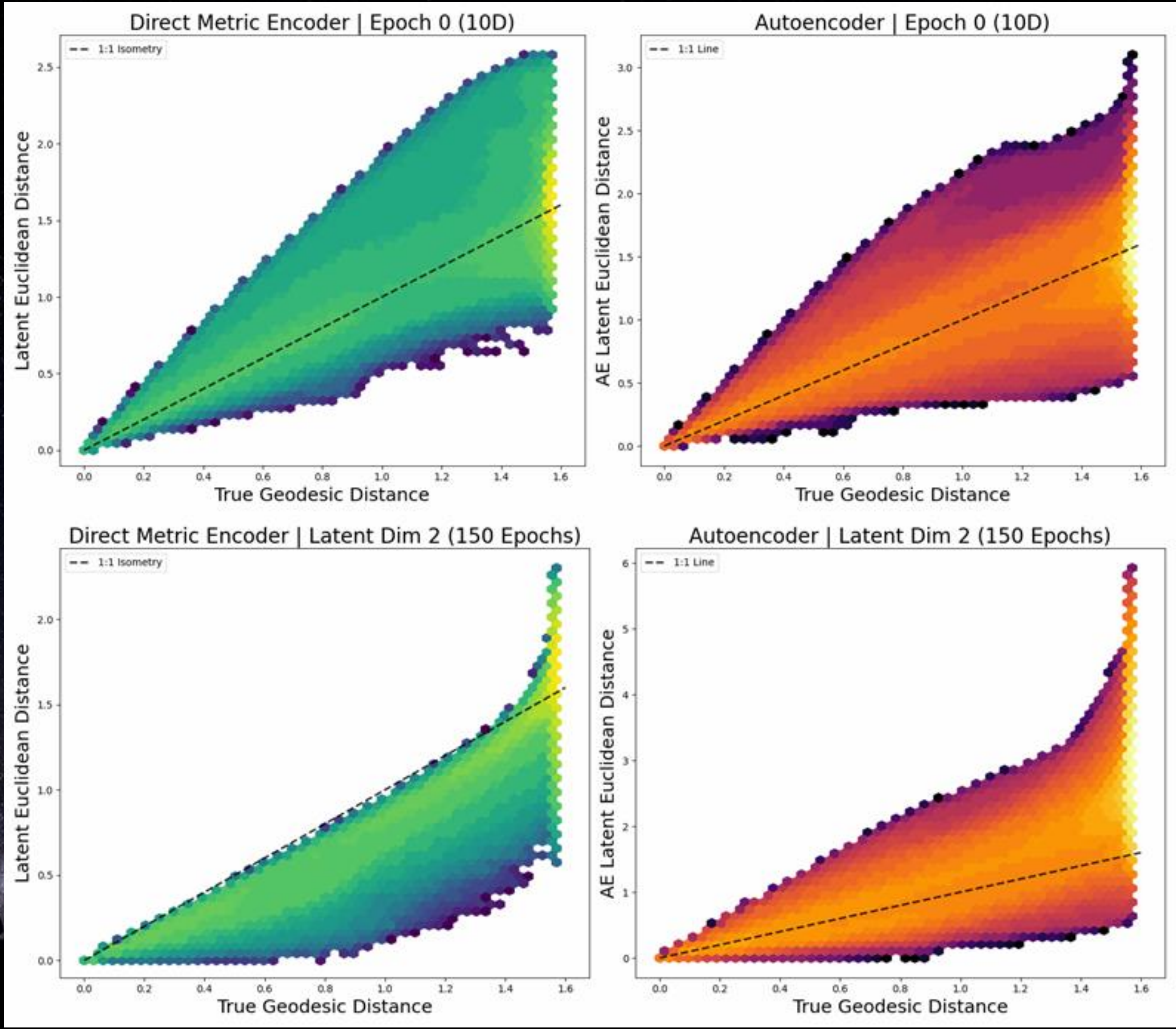
Andy Morgan andy.morgan@noirlab.edu

Proof

Synthetic hyperspherical data

Identical encoders trained in parallel: left trained via metric regression, right trained with an autoencoding objective

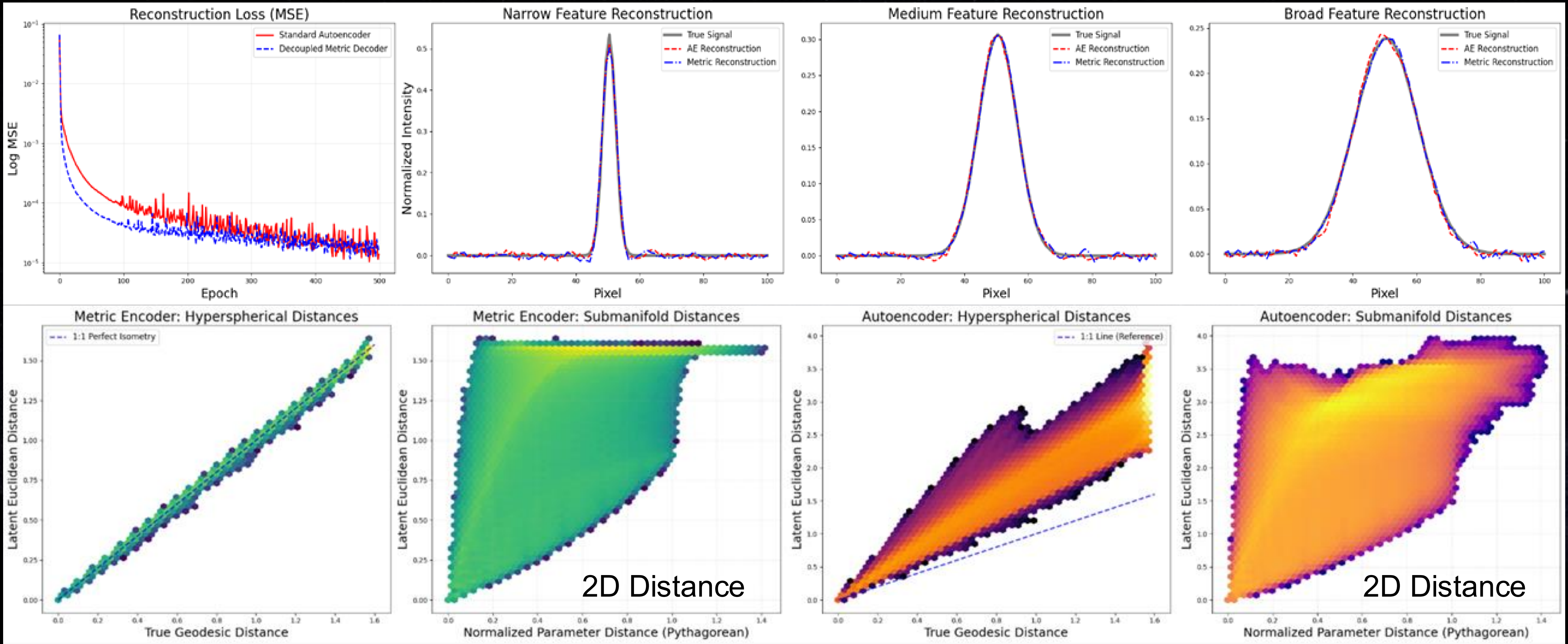
Plotting latent Euclidean distance (y) against true geodesic distance (x)



Andy Morgan andy.morgan@noirlab.edu

Not a Convergence Problem

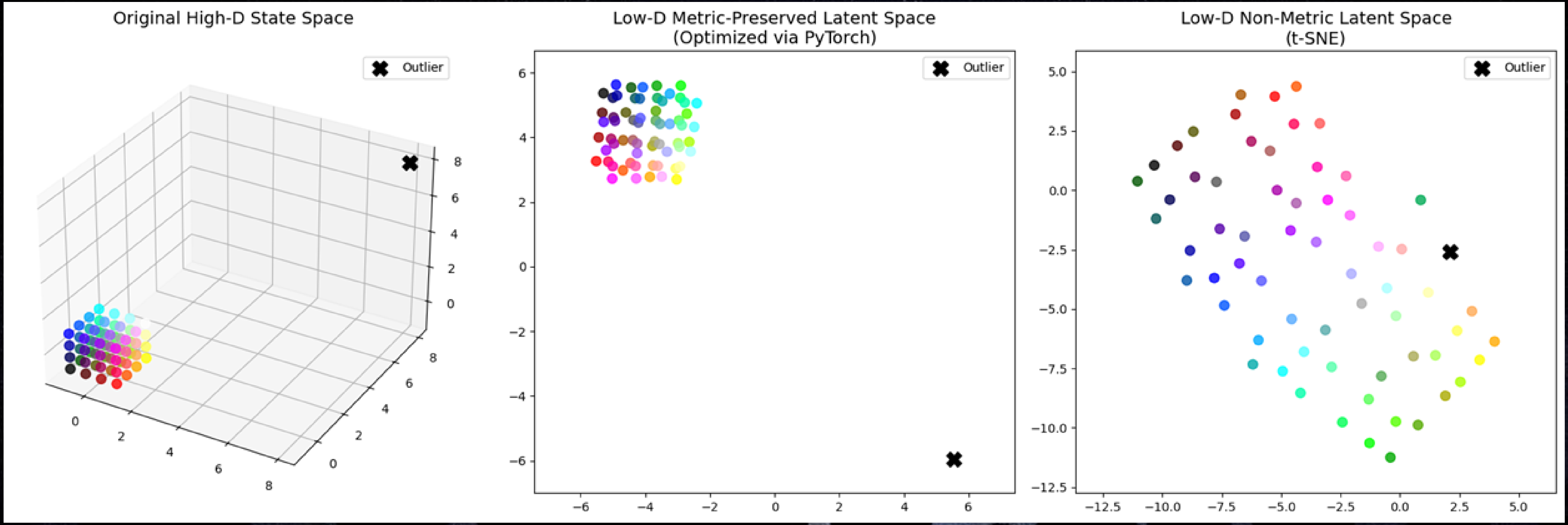
15D, 500 Epochs; Same synthetic data, now with 2D submanifold distances



Andy Morgan andy.morgan@noirlab.edu

Implications

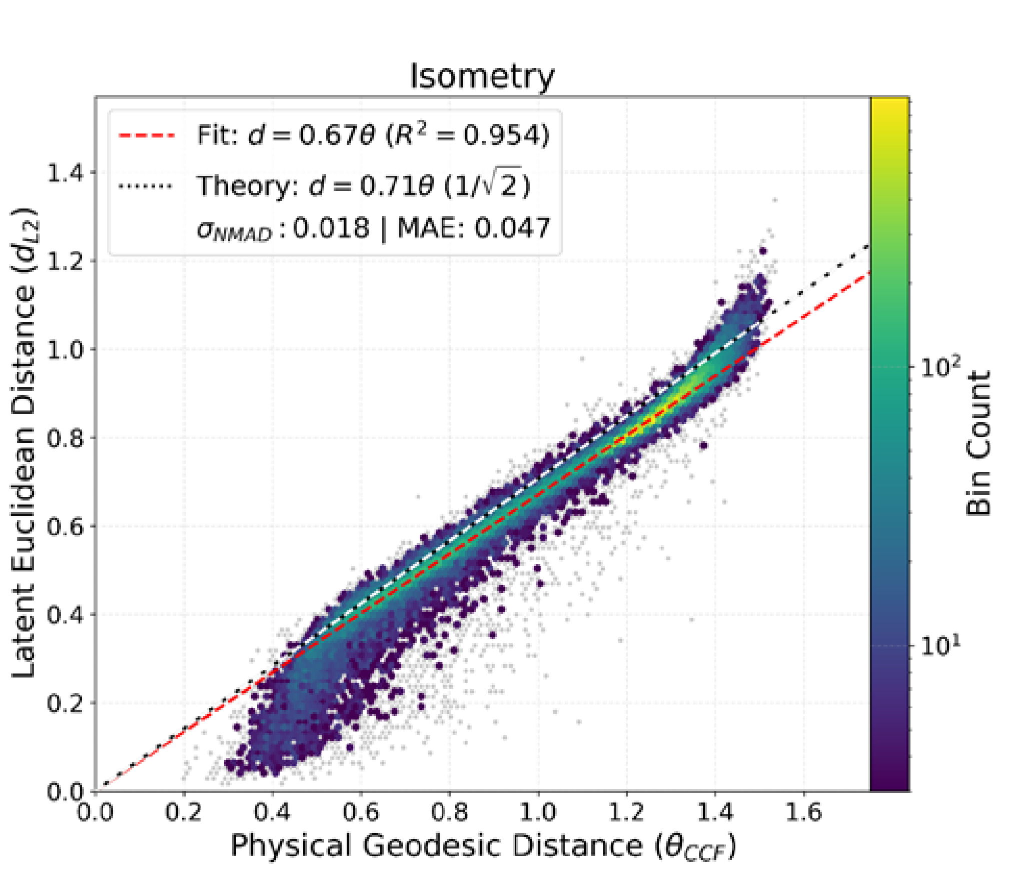
We can rigorously do: density modeling→outlier detection, clustering, vector algebra, etc



Andy Morgan andy.morgan@noirlab.edu

Determine a Metric

Standardized spectra live on a hypersphere; a valid metric space
Enforce latent Euclidean distance to be proportional to the physical hyperspherical distance



$$D^2_{latent} = 1 - \max_{\tau} \left(\frac{f_a \cdot f_{b,\tau}}{||f_a|| \; ||f_b||} \right)$$

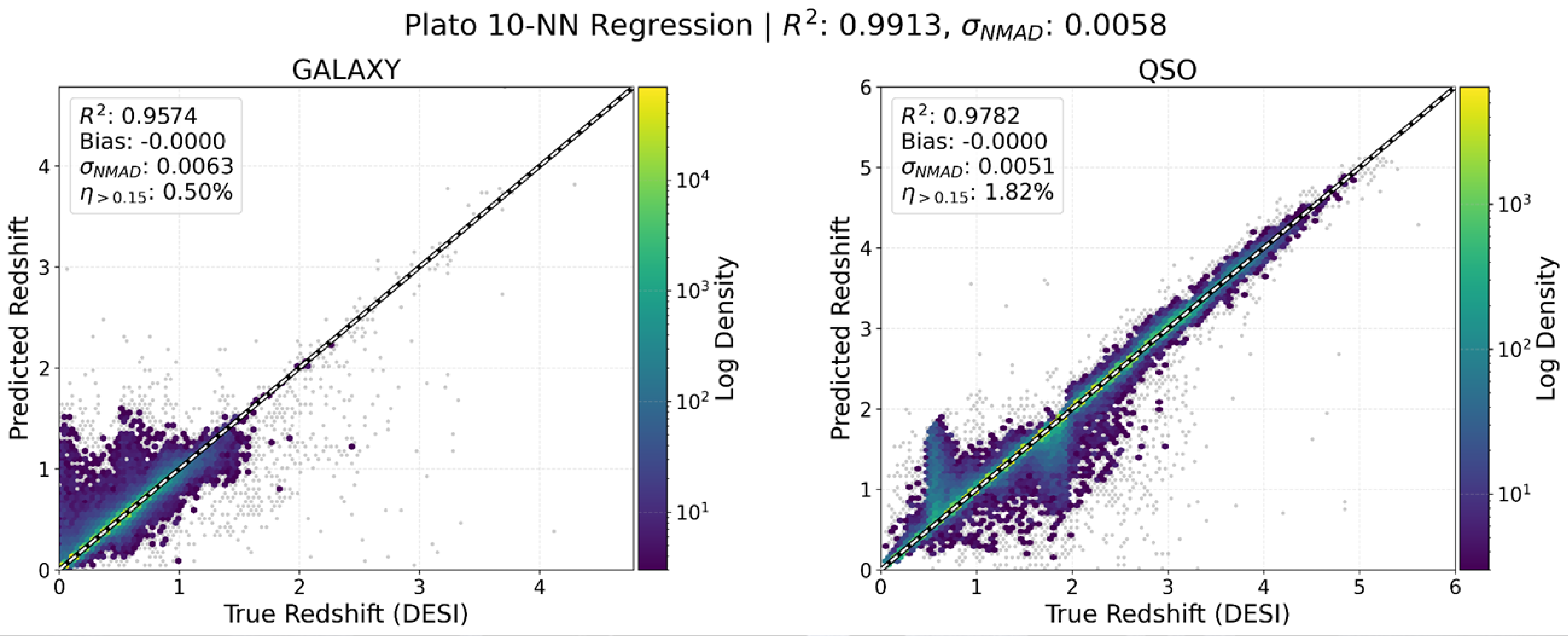
f = flux vector
τ = lag/shift in a zero-padded field

Term on right is just maximum of the normalized CCF over all lags

Valid metric on the quotient space of translation orbits and scale-invariance

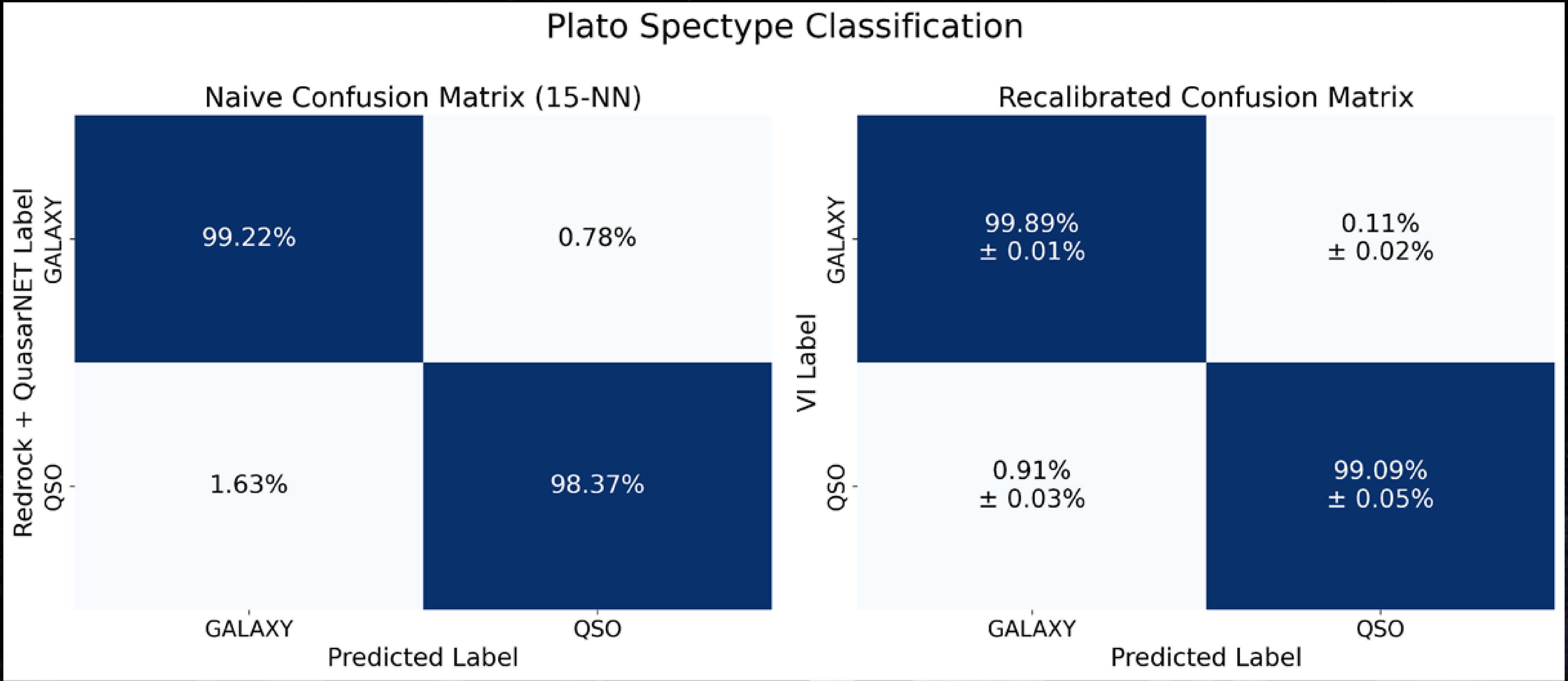
Andy Morgan andy.morgan@noirlab.edu

Initial Results: kNN Redshifts



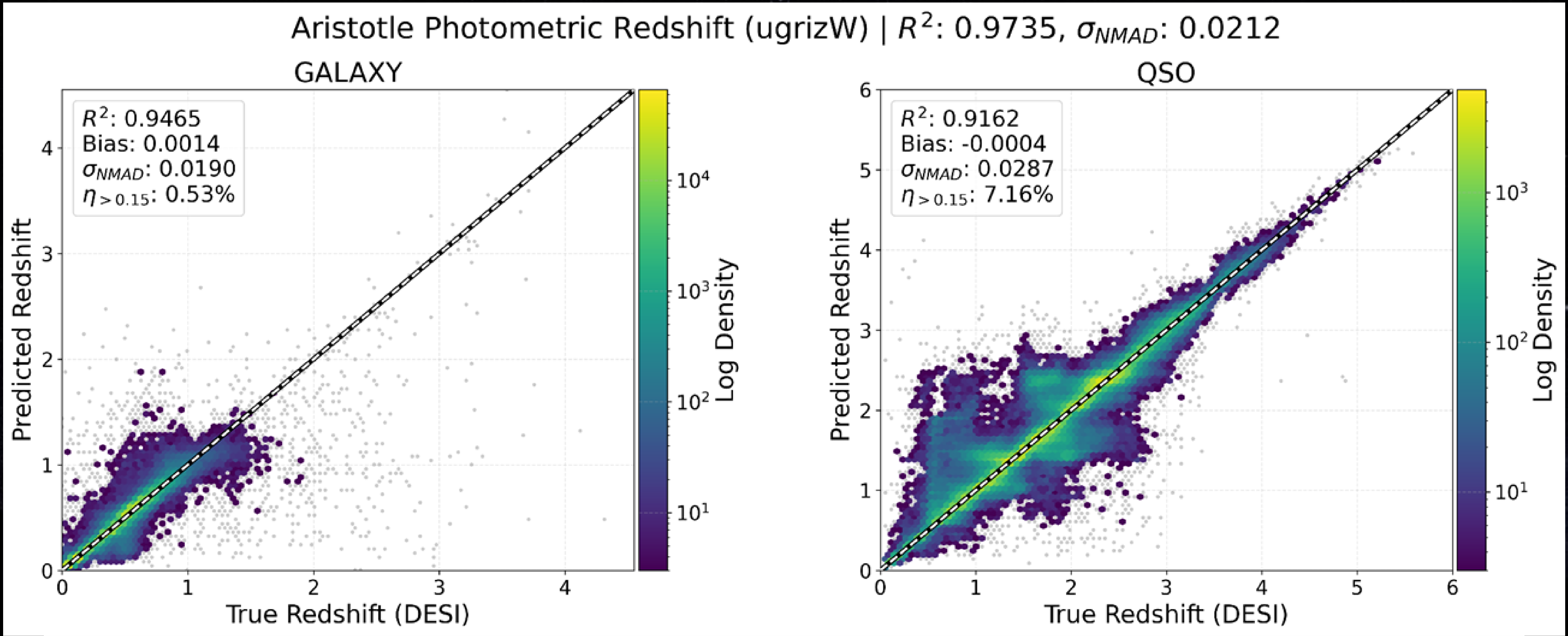
Andy Morgan andy.morgan@noirlab.edu

Initial Results: kNN Classification



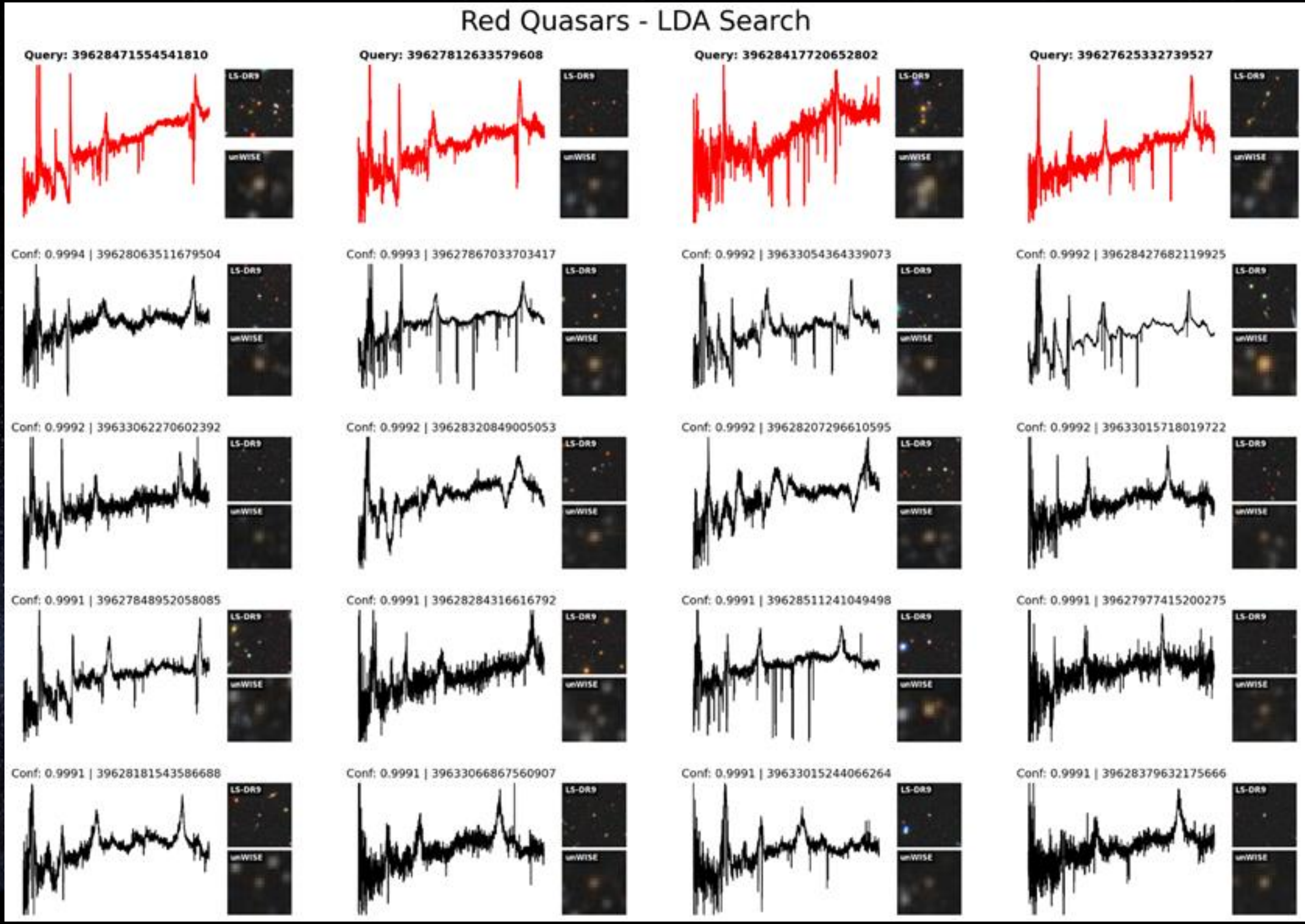
Andy Morgan andy.morgan@noirlab.edu

Initial Results: Photo-z



Andy Morgan andy.morgan@noirlab.edu

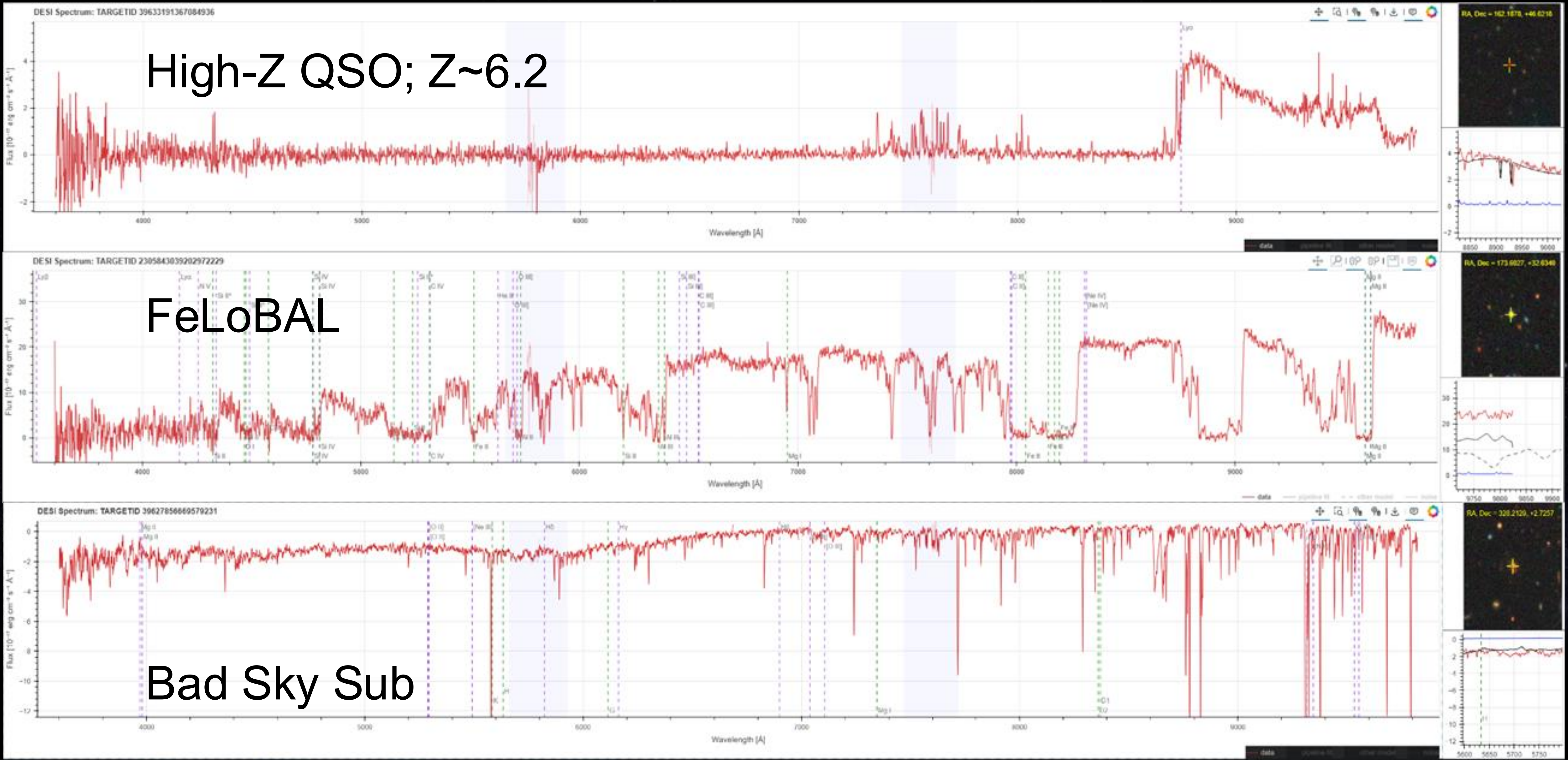
Initial Results: Similarity Search



Andy Morgan andy.morgan@noirlab.edu

Andy Morgan andy.morgan@noirlab.edu

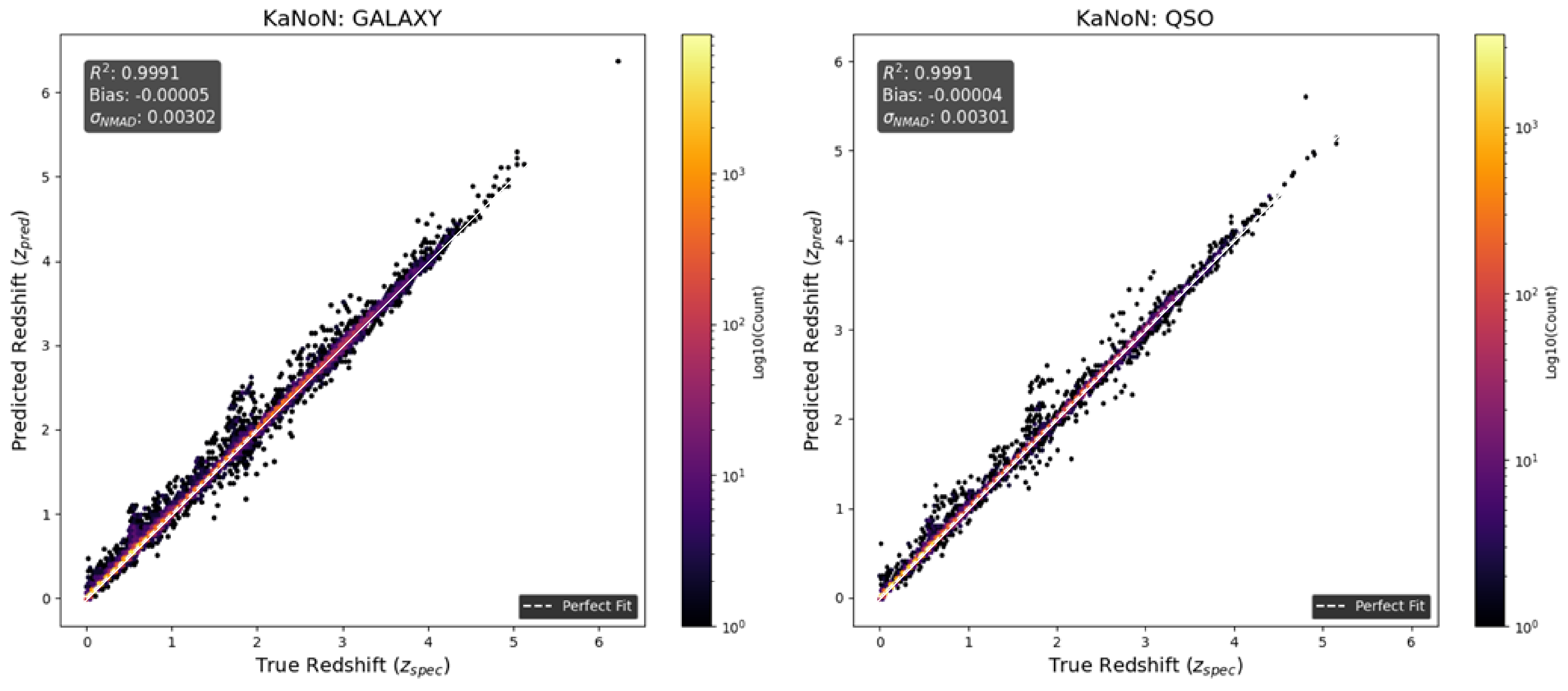
Density Modeling



Andy Morgan andy.morgan@noirlab.edu

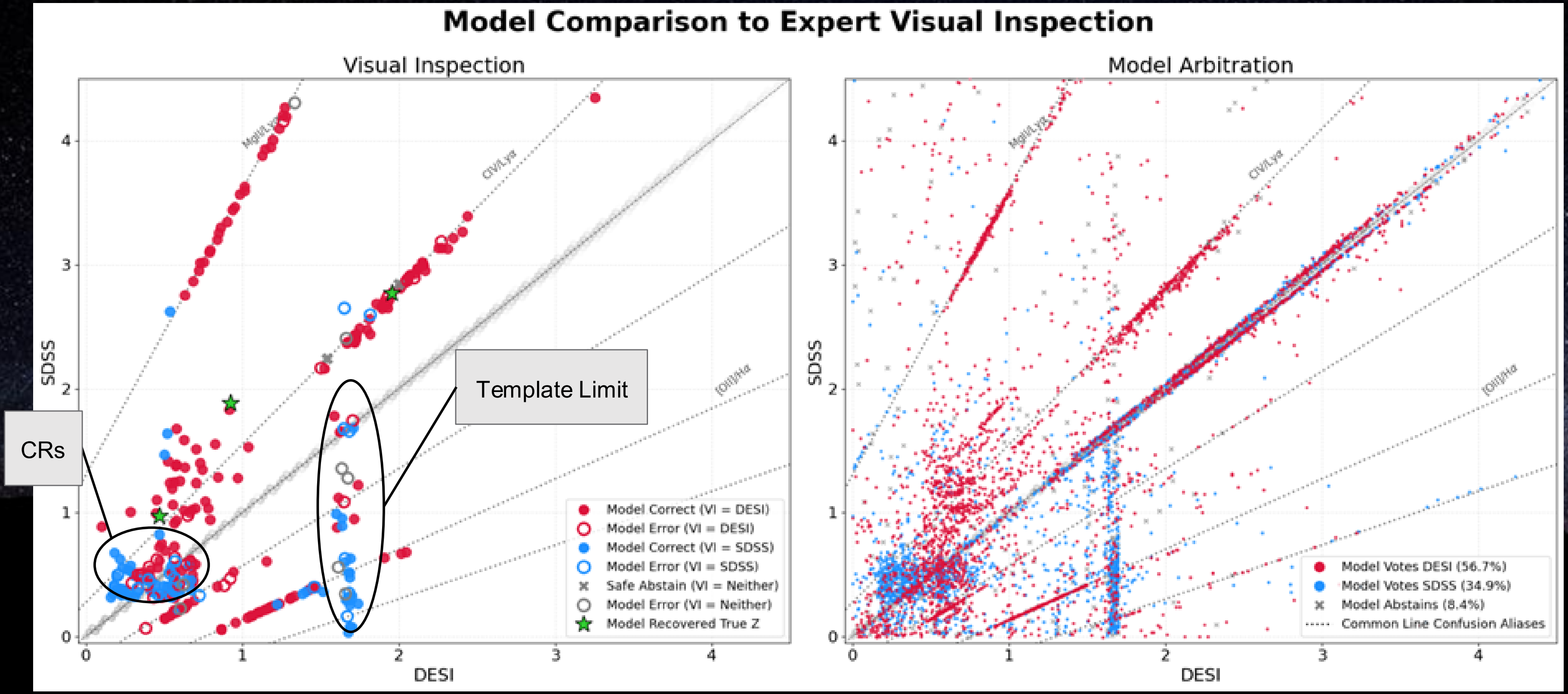
Density Modeling

Model p(redshift | spectrum) → redshift estimate is the argmax



Probabilistic Redshift Arbitration

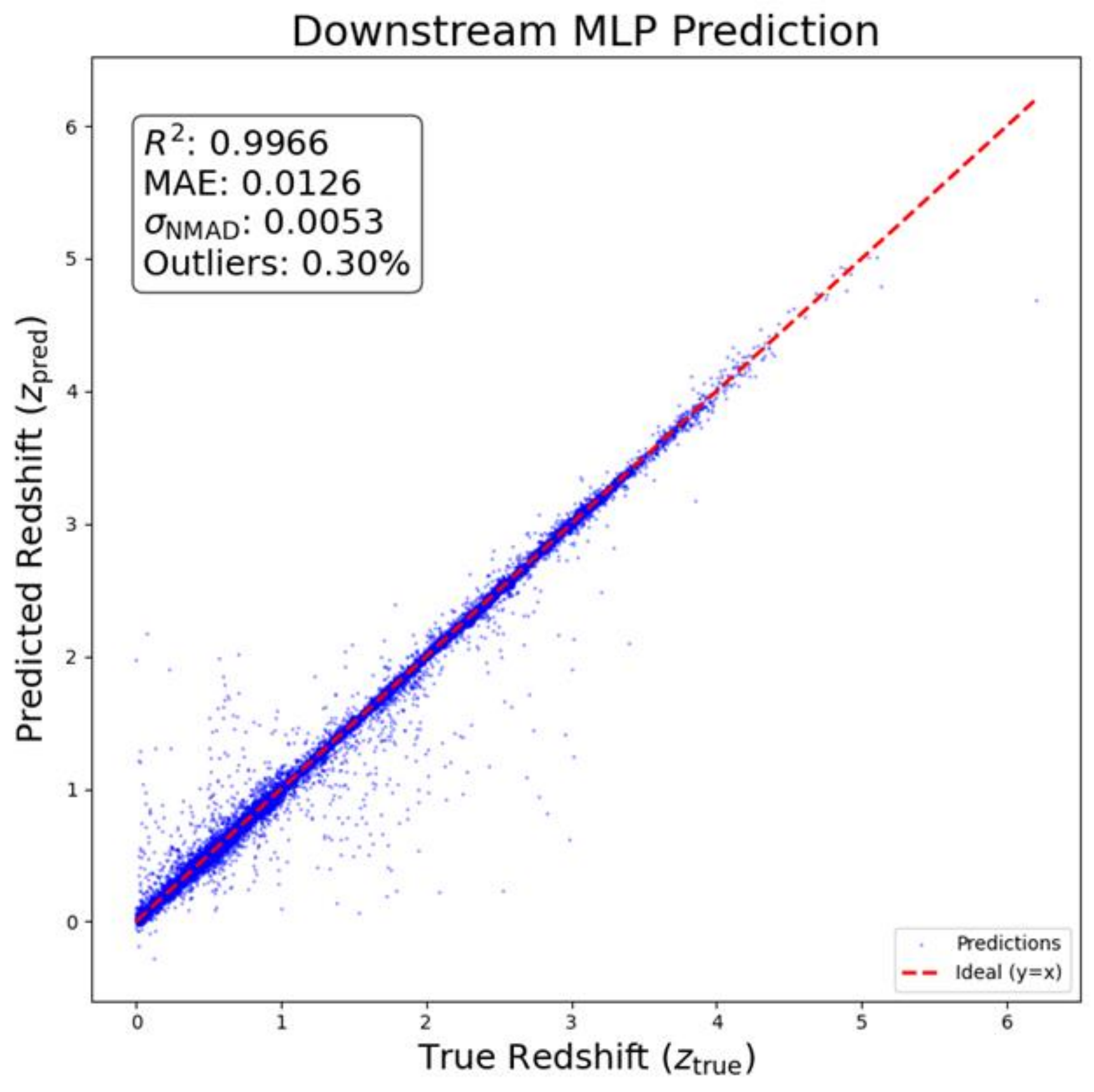
Initial tests have shown success in flagging ~60% of Redrock's errors with a ~2% False Positive Rate → ~98% reduction in manual visual inspection for a ~60% reduction in pipeline failures



Credit to Yufeng Luo, Adam Tedeschi, Dixon Reid (U. Wyoming), and Arjun Dey (NOIRLab) for contributing to this VI effort

Andy Morgan andy.morgan@noirlab.edu

The Metric is Self-Diagnosing



Worked really well, but…

- **<u>Metric stress is ~0</u>:** Not a model capacity issue.
- **<u>Errors diagnose the geometry</u>:** Failures point directly to misrepresented physics, probing the data instead of the model.
- **<u>Iterative discovery</u>:** Update the metric to actively converge on the data's true geometry.

Andy Morgan andy.morgan@noirlab.edu

Conclusion

Key Takeaways

- **<u>Generative to Geometric</u>:** We're stepping away from the prevailing generative paradigm towards a latent representation that acts as a self-diagnosing surrogate for the target data via metric-learning
- **<u>Foundation Level Performance</u>:** We demonstrate competitive and SOTA performance across a number of downstream tasks at a fraction of the parameter count of contemporary generative models, even with simple non-parametric methods and a "false" metric

Future Developments

- **Metric Refinement:** Iterating upon the diagnostic nature of the approach
- **Full Survey Integration:** Scaling to the complete extragalactic DESI DR2

Andy Morgan andy.morgan@noirlab.edu