Optimizing Photometric Redshift Training Sets with UMAP

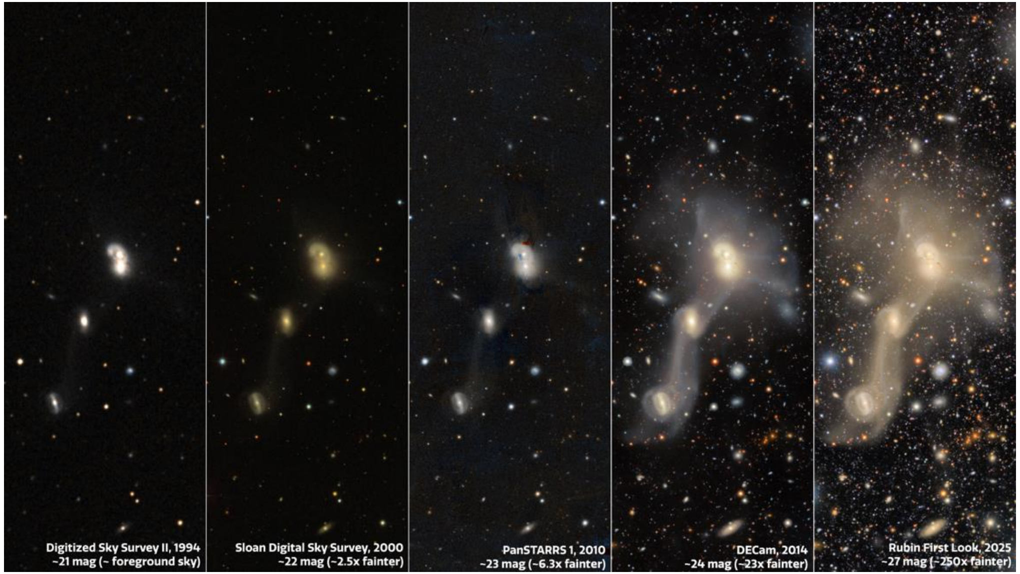
Finian Ashmead
University of Pittsburgh
finianashmead@pitt.edu
2nd Annual Cosmic Horizons Conference 7/16/26

supervisors Jeff Newman and Brett Andrews
this work supported by WFS grant (PI: Newman)

arxiv: 2512.09032

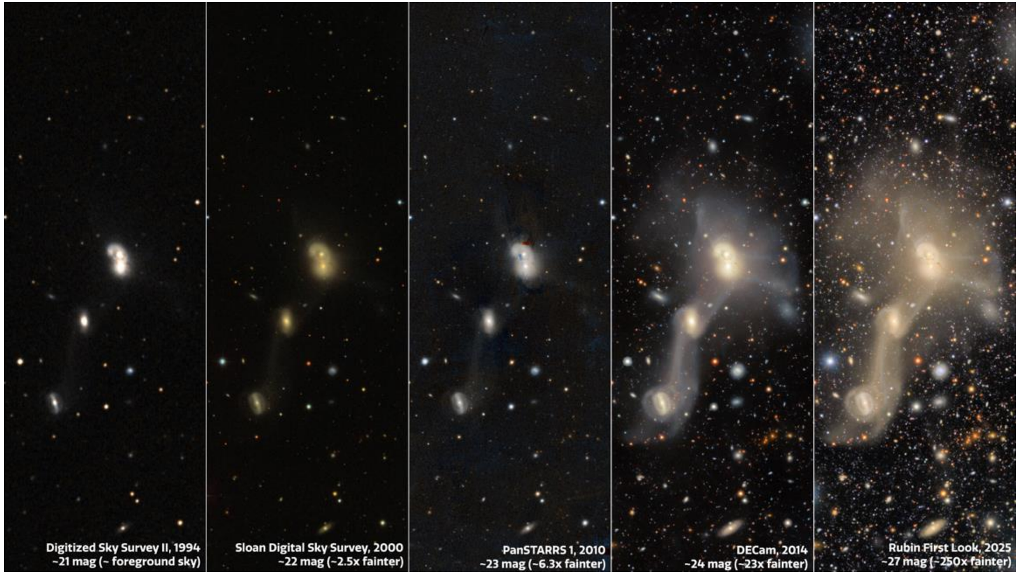
Why photometric redshifts (photo-*z*’s)?

- with large imaging surveys we have a **massive amount of photometric data,** including **very faint objects**



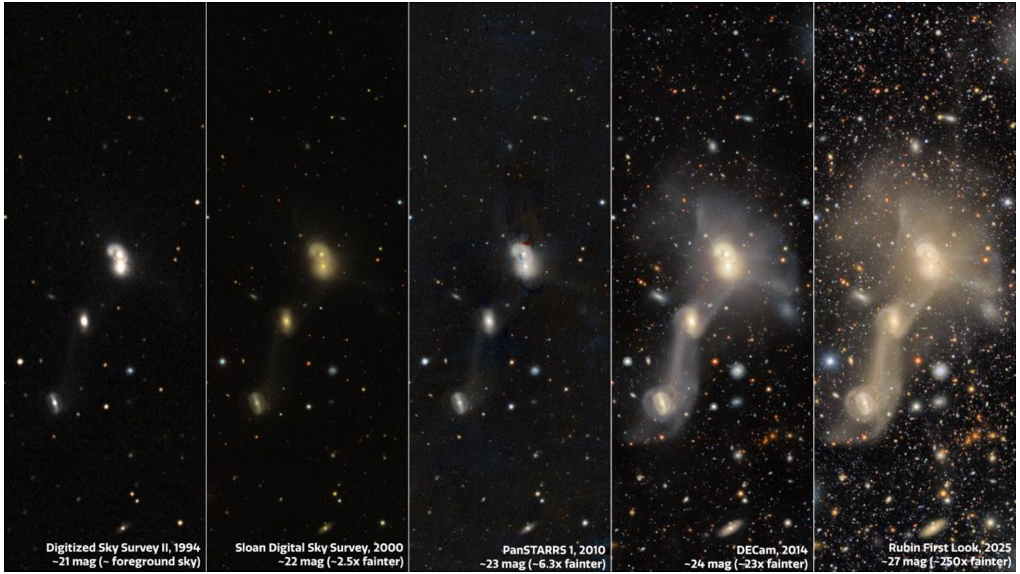
Why photometric redshifts (photo-*z*'s)?

- with large imaging surveys we have a **massive amount of photometric data,** including **very faint objects**
- need redshifts for cosmology + galaxy evolution

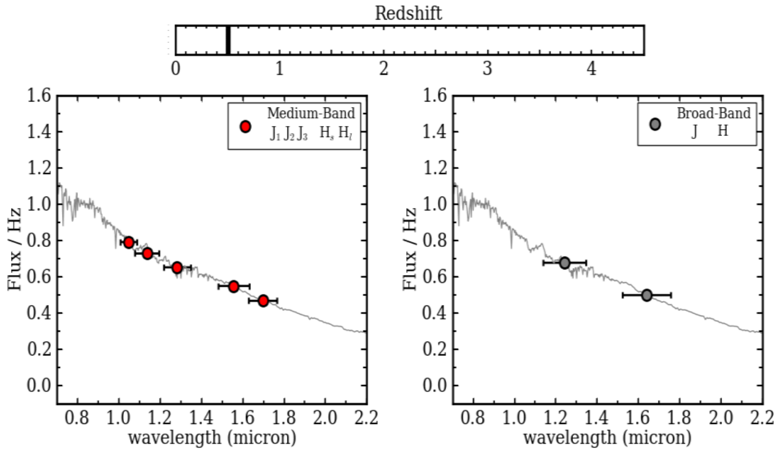


Why photometric redshifts (photo-*z*'s)?

- with large imaging surveys we have a **massive amount of photometric data,** including **very faint objects**
- need redshifts for cosmology + galaxy evolution
- spectroscopic redshifts (spec-*z*'s) are **completely infeasible**
- → we need to get **redshifts from photometry**



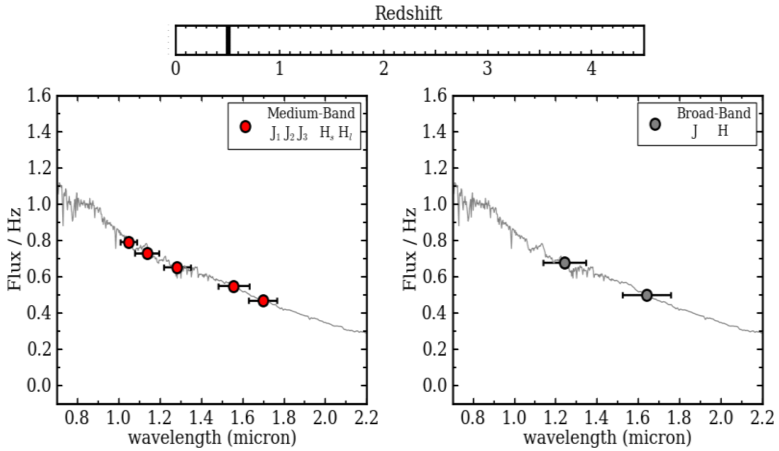
Why do photo-*z*'s work?

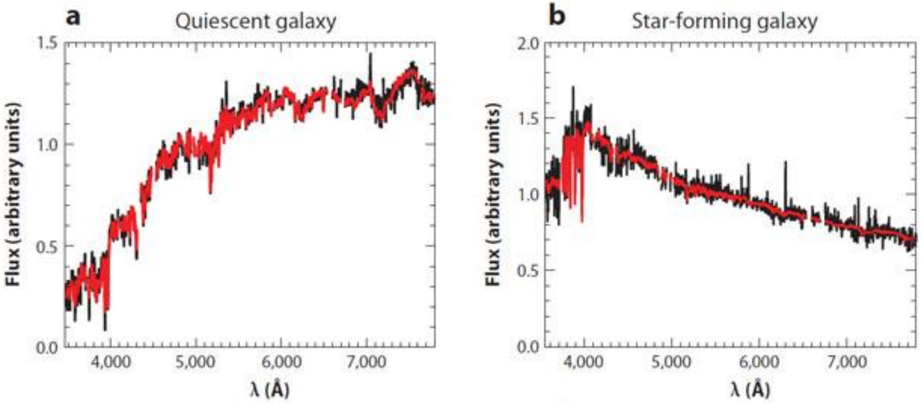


- using the galaxies we have spectra for as training data, learn to estimate redshifts from the photometry

A. Tomczak gif

Why do photo-z's work?



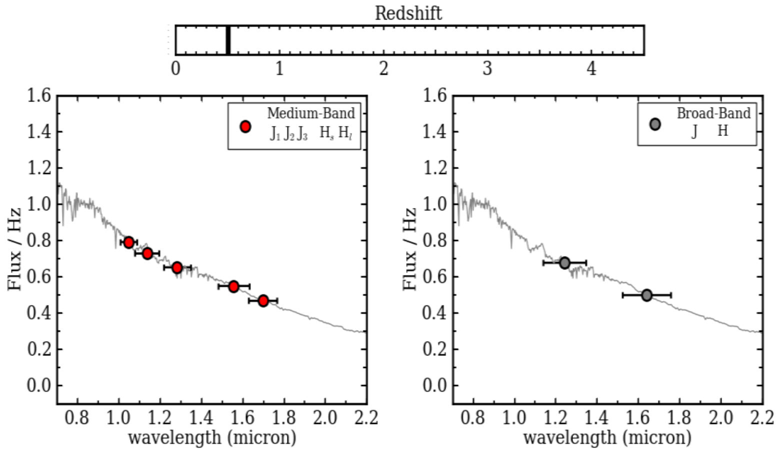


- using the galaxies we have spectra for as training data, learn to estimate redshifts from the photometry

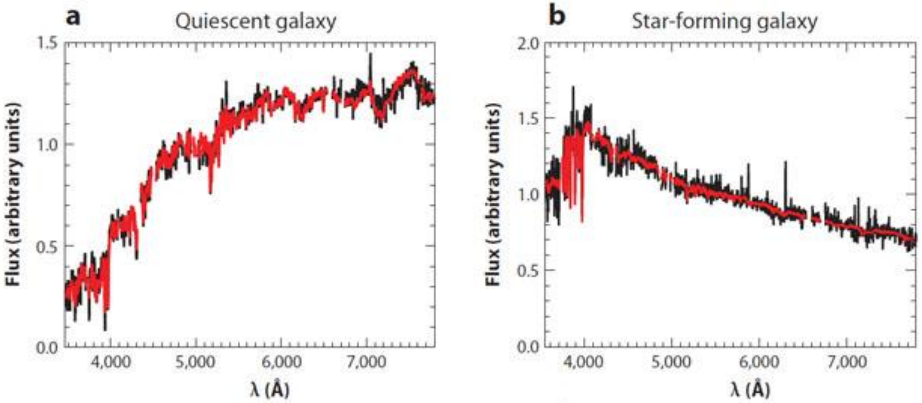
- observed galaxy colors are highly correlated and mainly determined by two physical parameters: redshift and specific star formation rate (sSFR)

[Tojeiro et al. 2009](#) [A. Tomczak gif](#)

Why do photo-z's work?



- using the galaxies we have spectra for as training data, learn to estimate redshifts from the photometry



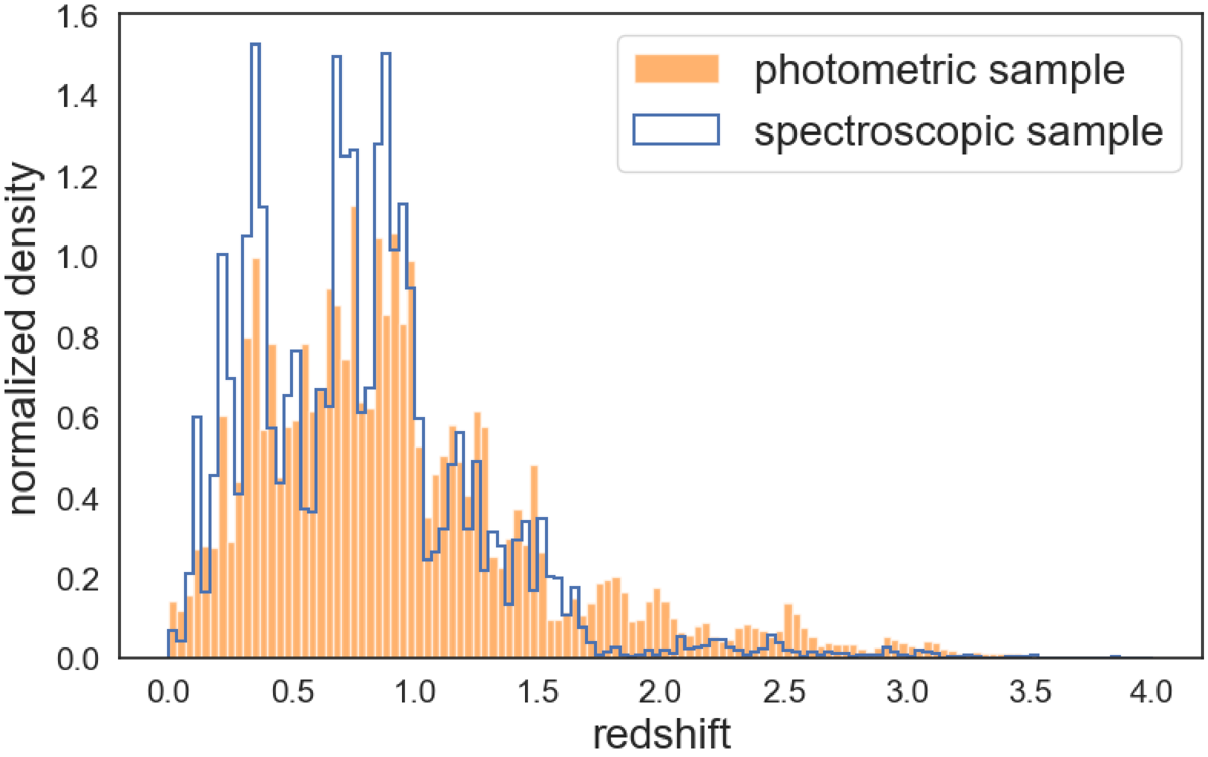
- observed galaxy colors are highly correlated and mainly determined by two physical parameters: redshift and specific star formation rate (sSFR)

→ use **dimensionality reduction** to recover the low-dimensional manifold underlying observed galaxy colors

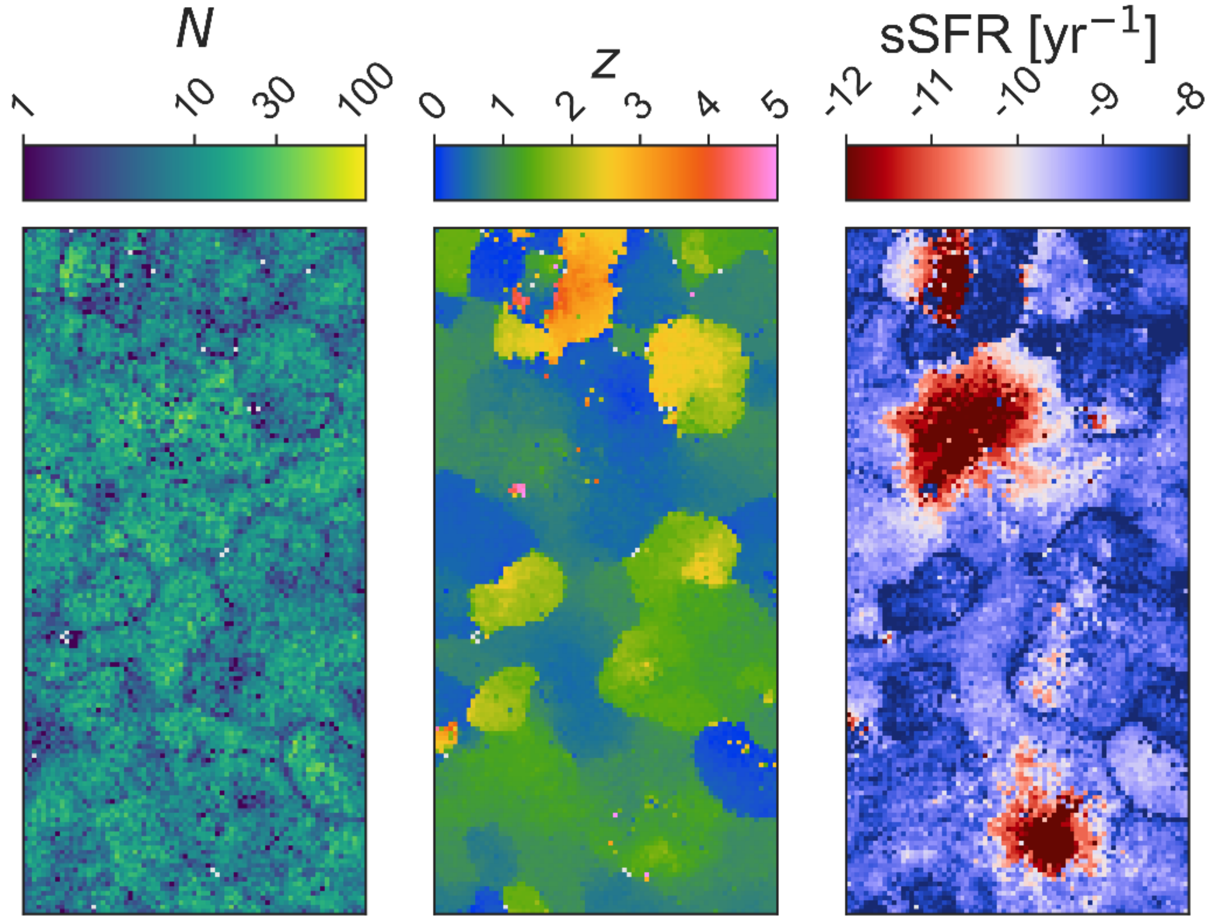
[Tojeiro et al. 2009](#) [A. Tomczak gif](#)

Data

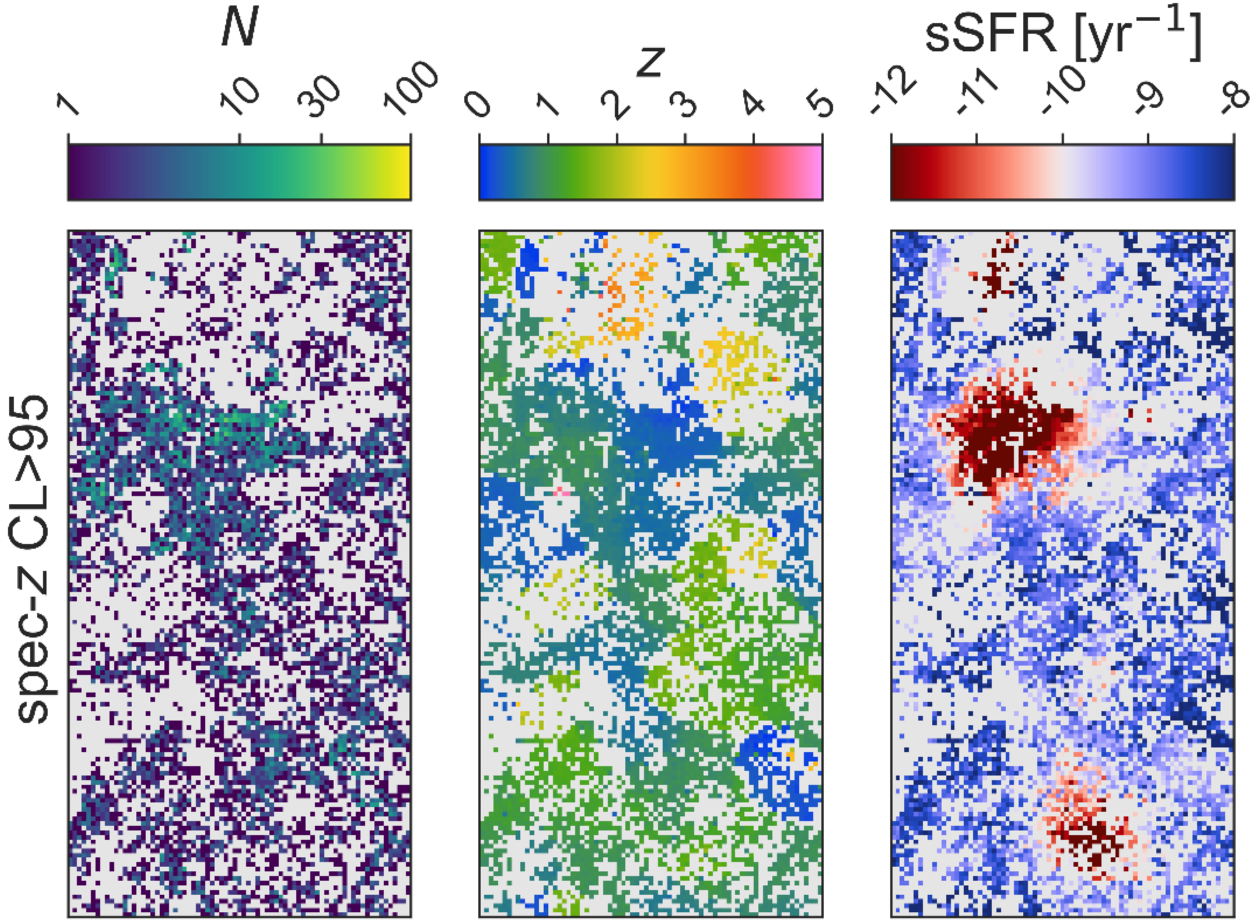
- ~110k galaxies in COSMOS2020 with *ugrizyJH* colors, and LePHARE photo-*z*'s and sSFR's
- ~14k with spec-*z*'s
    - spec-*z*'s have **biased coverage** of the redshift distribution and color space



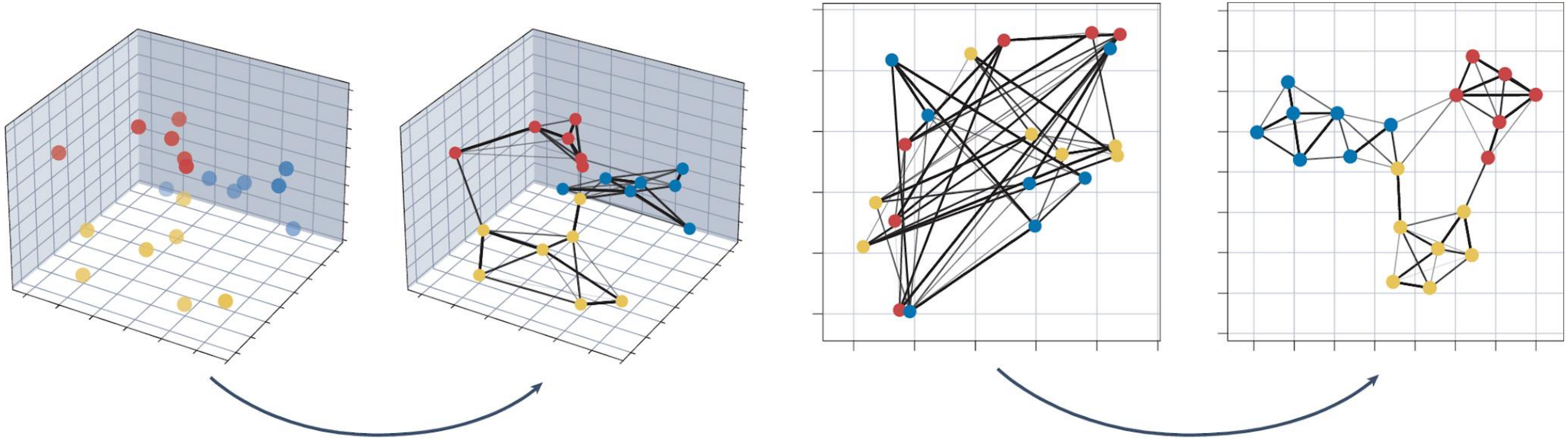
SOM: full imaging sample w/ photo-z's



SOM: spectroscopic sample of ~14k



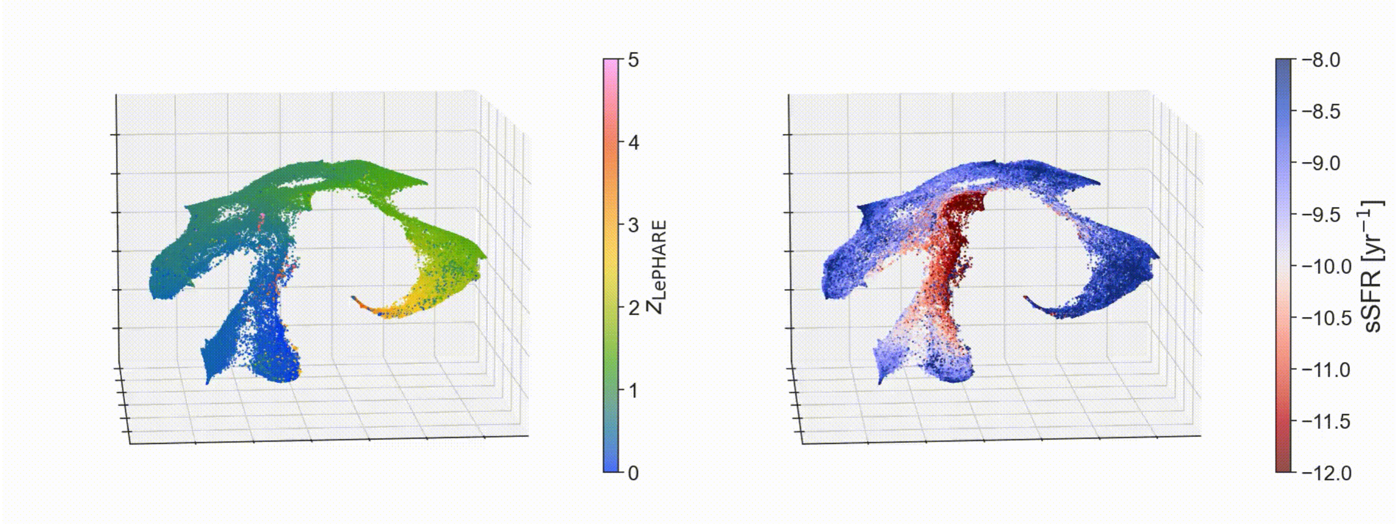
UMAP learning



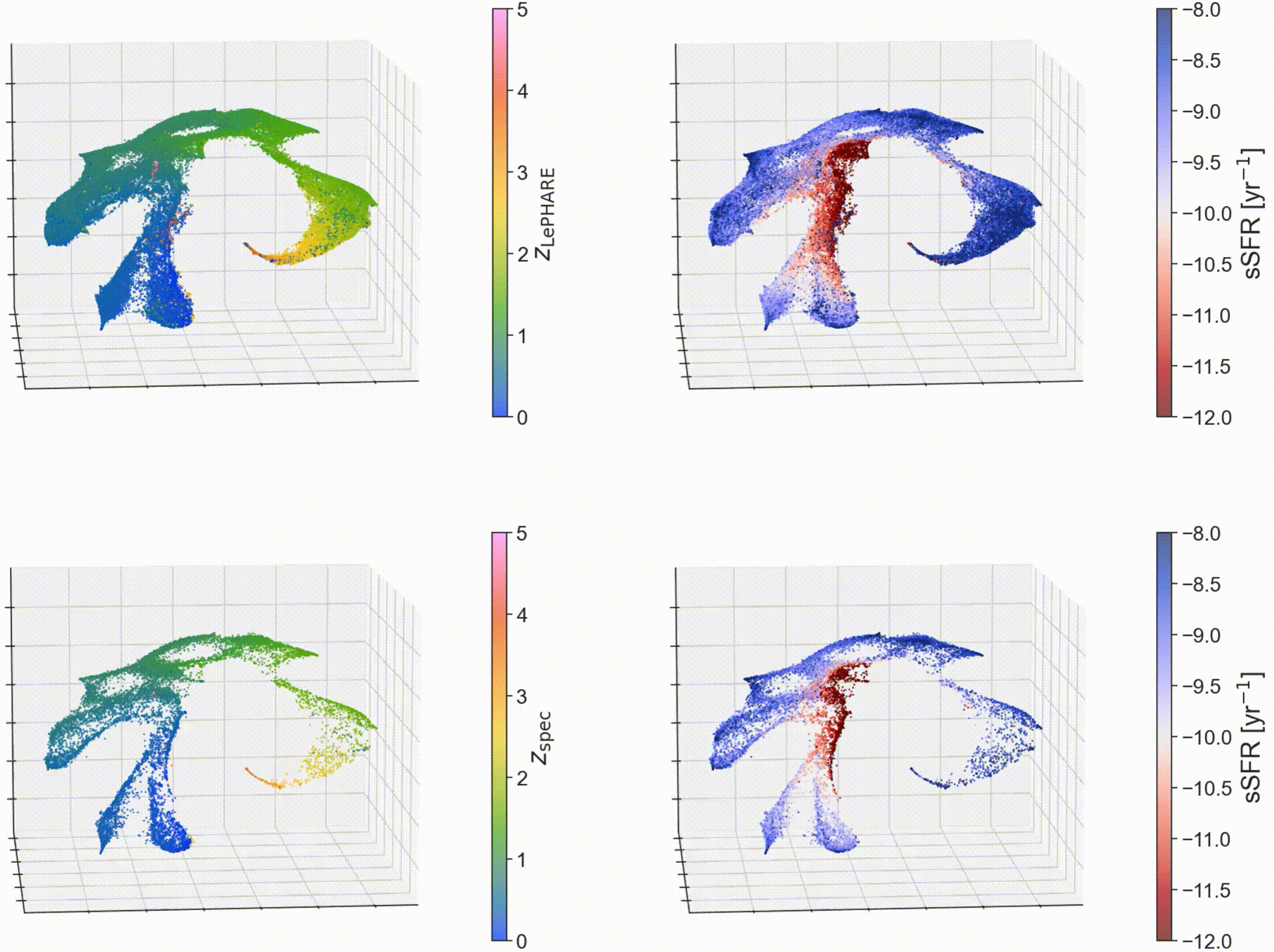
Build a topological representation of
the high-dimensional data

Optimize a low-dimensional embedding
to have similar topological representation

[source]

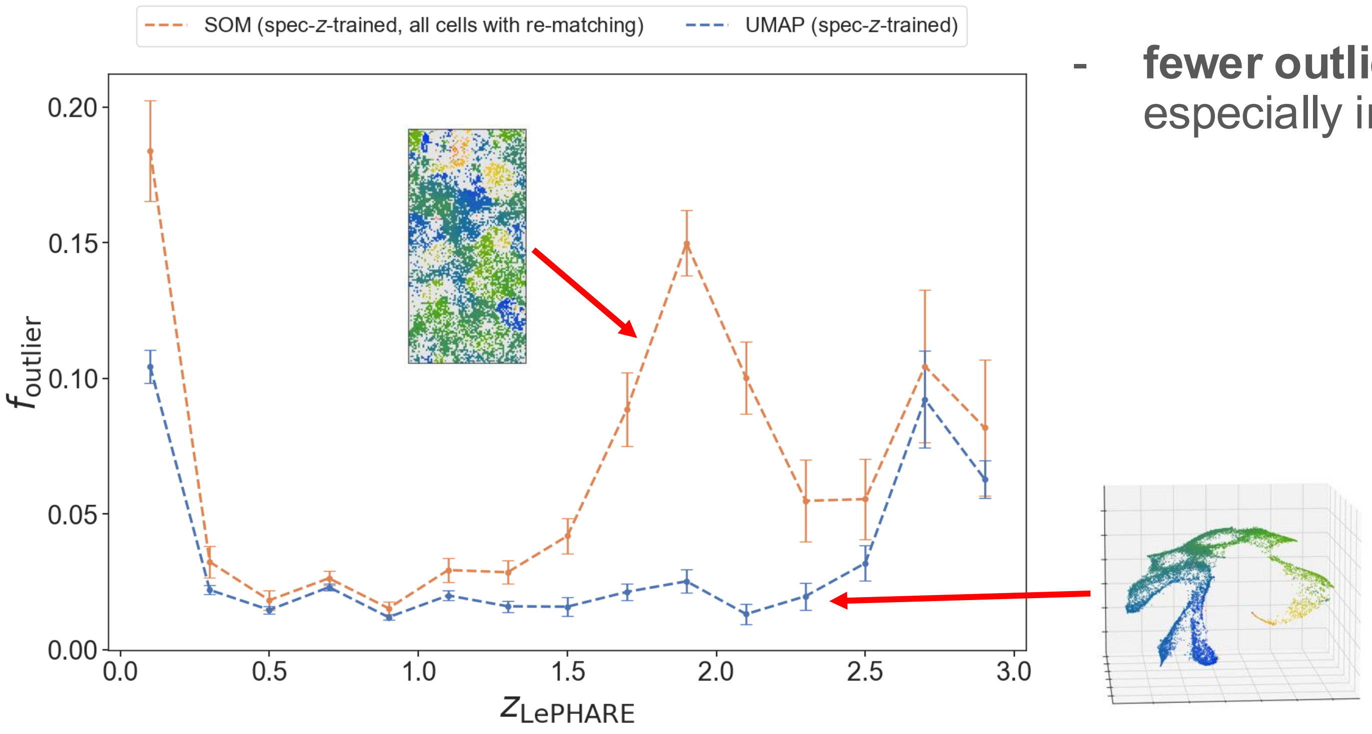


*mp4 animation and interactive versions are available [here](#) on [my website](#)
finianashmead.github.io



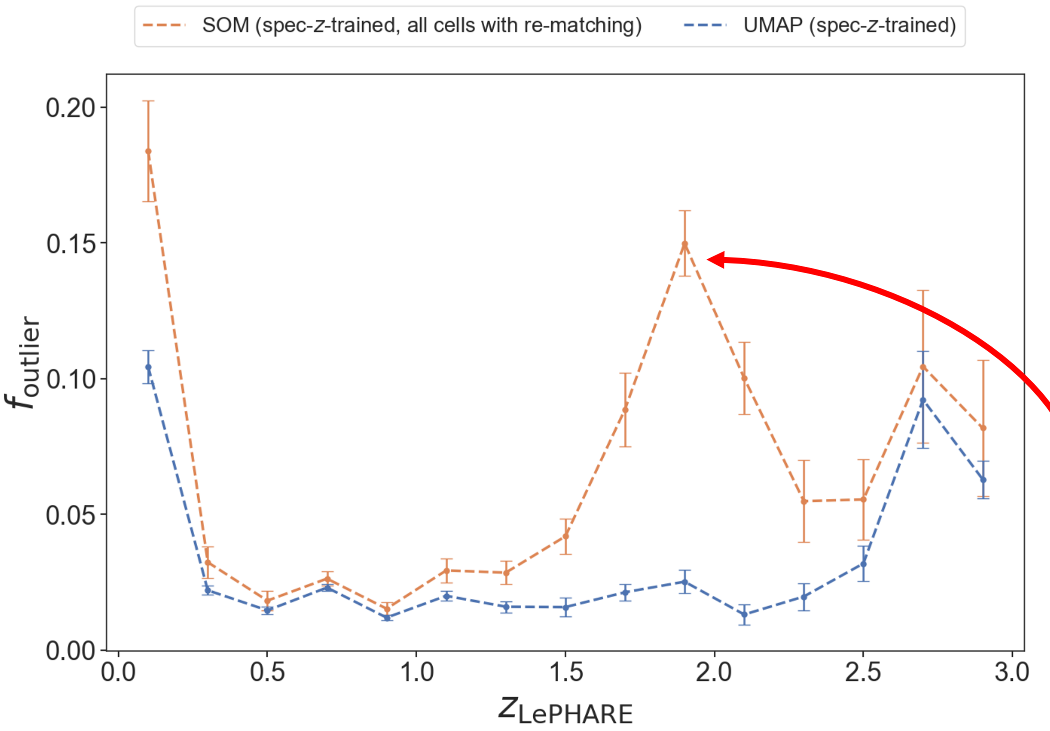
*mp4 animation and interactive versions are available [here](#) on [my website](#) finianashmead.github.io

Quantitative Test: Results

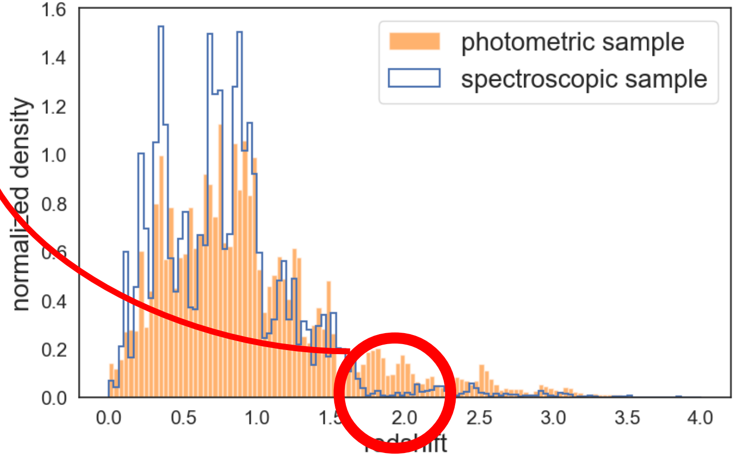


- **fewer outliers with UMAP!**
  especially in difficult regimes

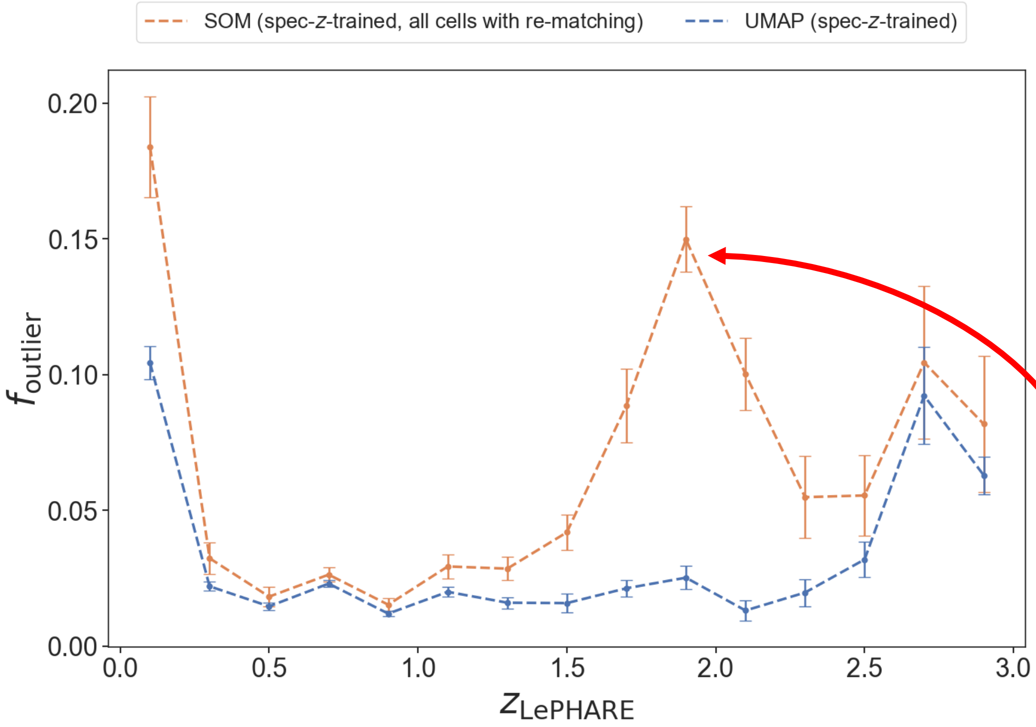
Quantitative Test: Results



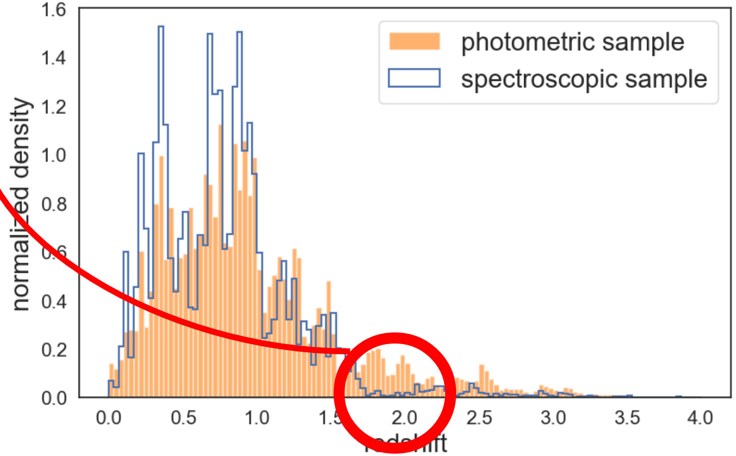
- **fewer outliers with UMAP!** especially in difficult regimes
- SOM outlier rate spikes where training data is sparse



Quantitative Test: Results



- **fewer outliers with UMAP!** especially in difficult regimes
- SOM outlier rate spikes where training data is sparse
  - we largely rely on ground based instruments, OII leaves optical at z~>1.5

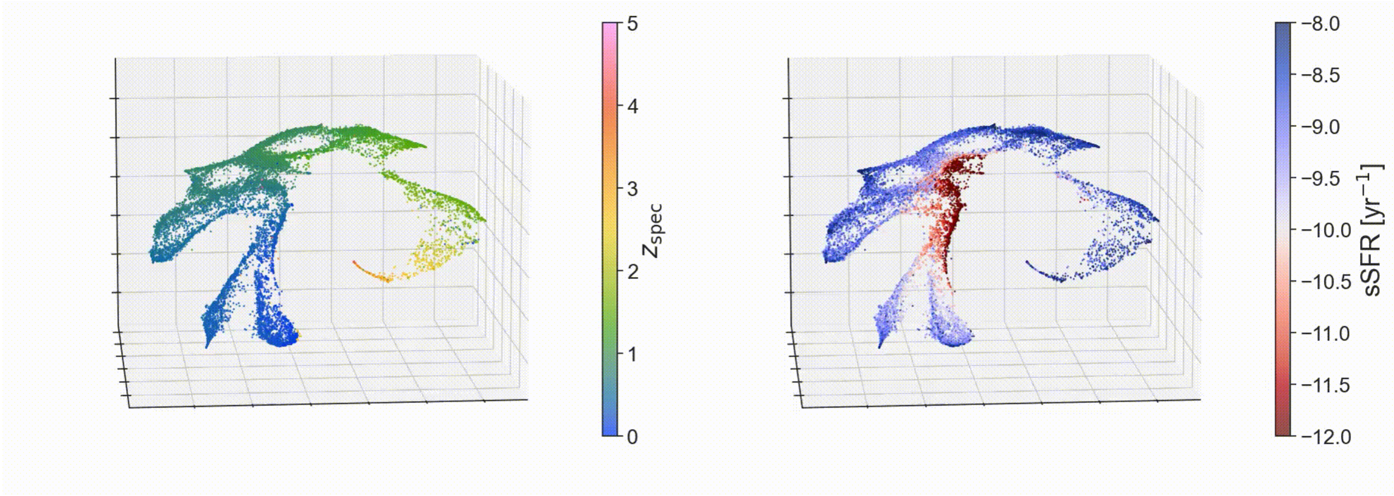


Summary:

- UMAP→ work in a space densely sampled by the data that maps simply and reliably to redshift
- we can make balanced spec-z training sets as input for existing frameworks
- going forward:
    - regression and p(z) estimation with techniques like GPR and normalizing flows

*animated/interactive plots, github repo @ finianashmead.github.io

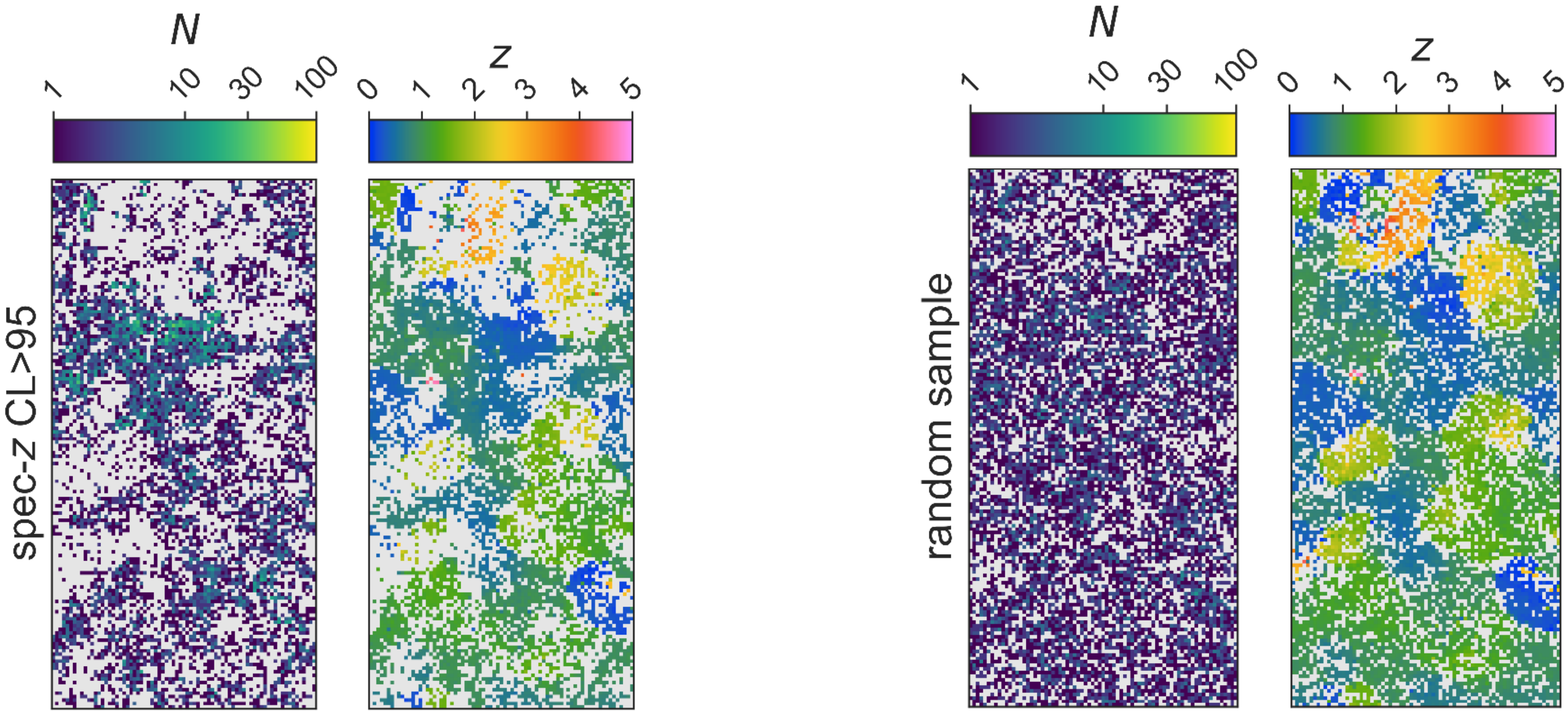
arxiv paper : 2512.09032



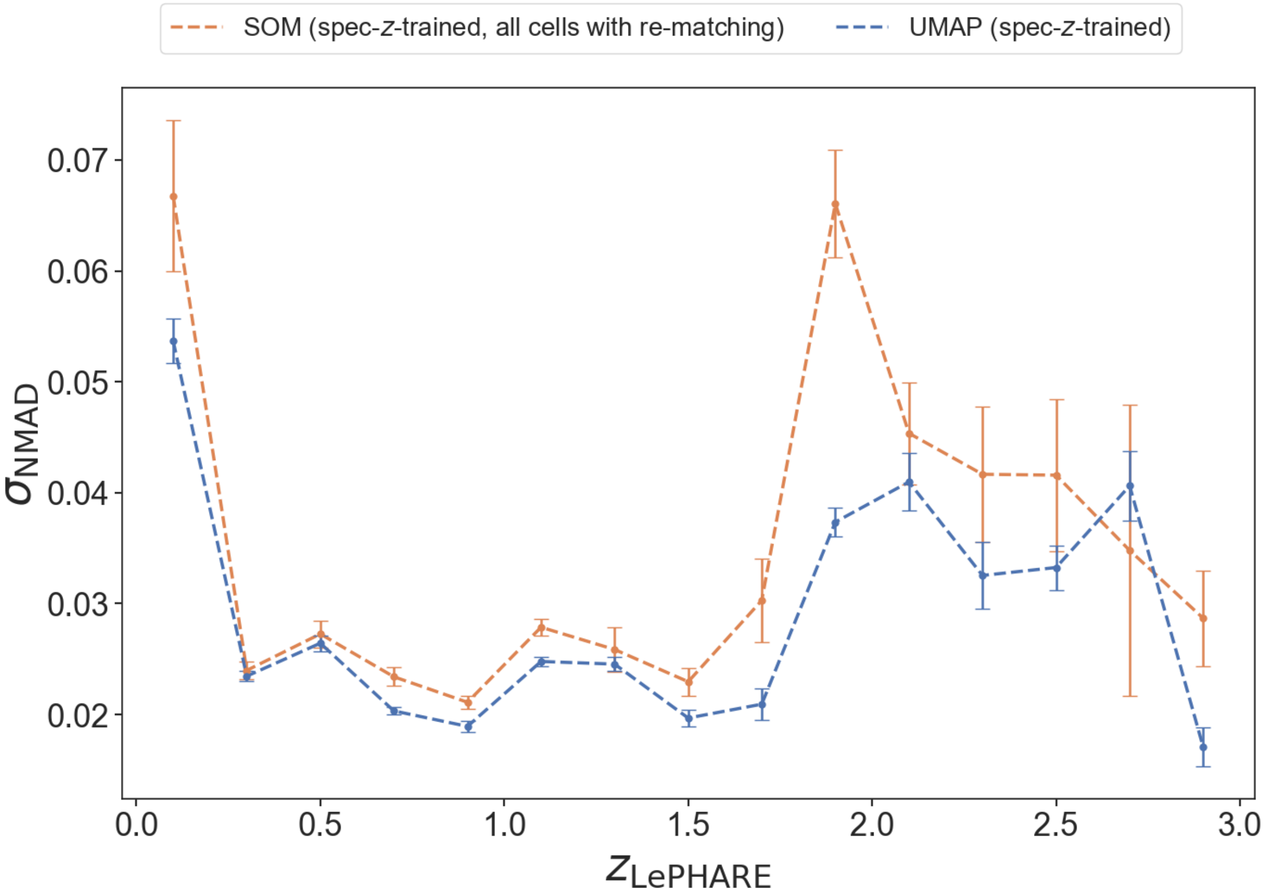
BONUS SLIDES!

Quantitative Test: Training Samples

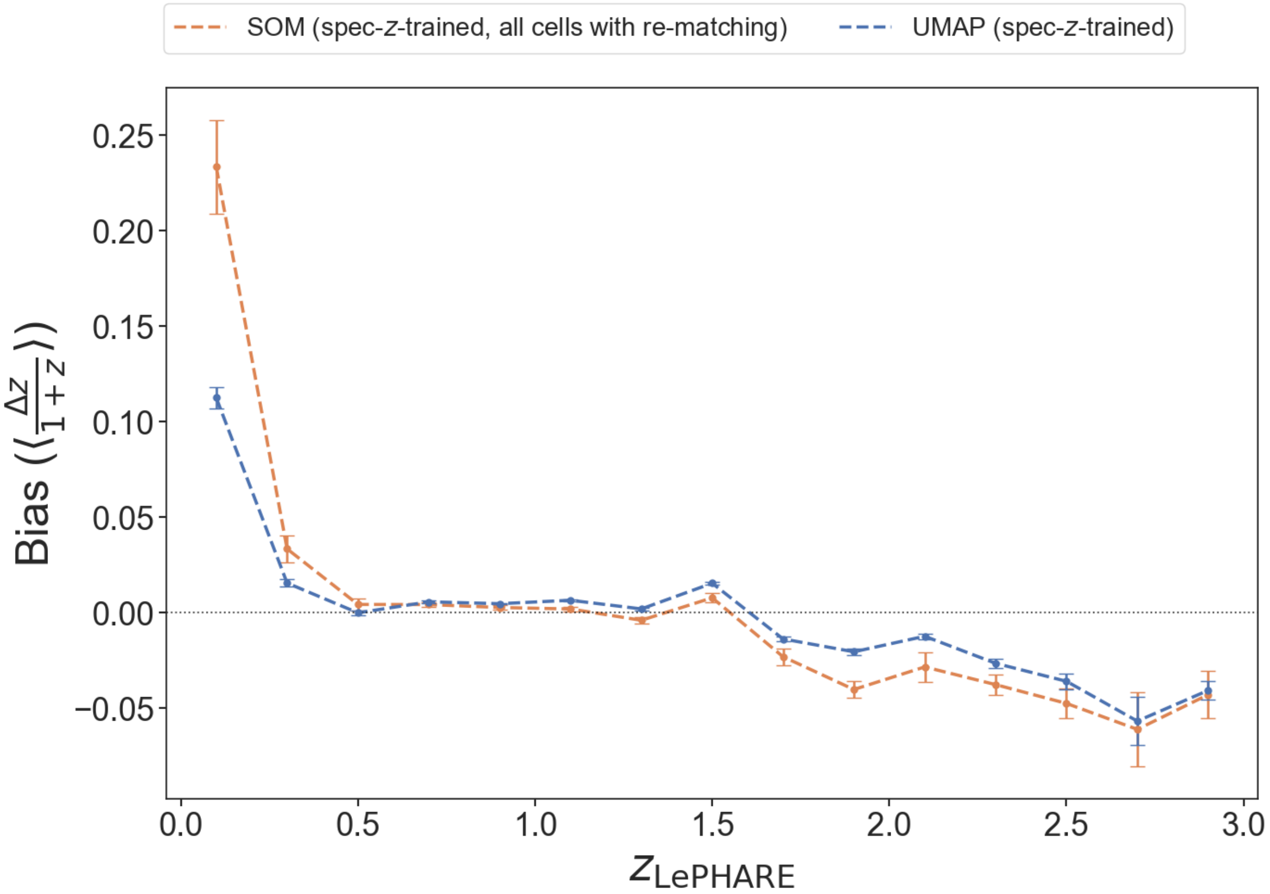
- 14.5k/110k have high-confidence spec-z's (CSRC, Khostovan et al. 2026)
- random 14.5k with LePHARE-z's



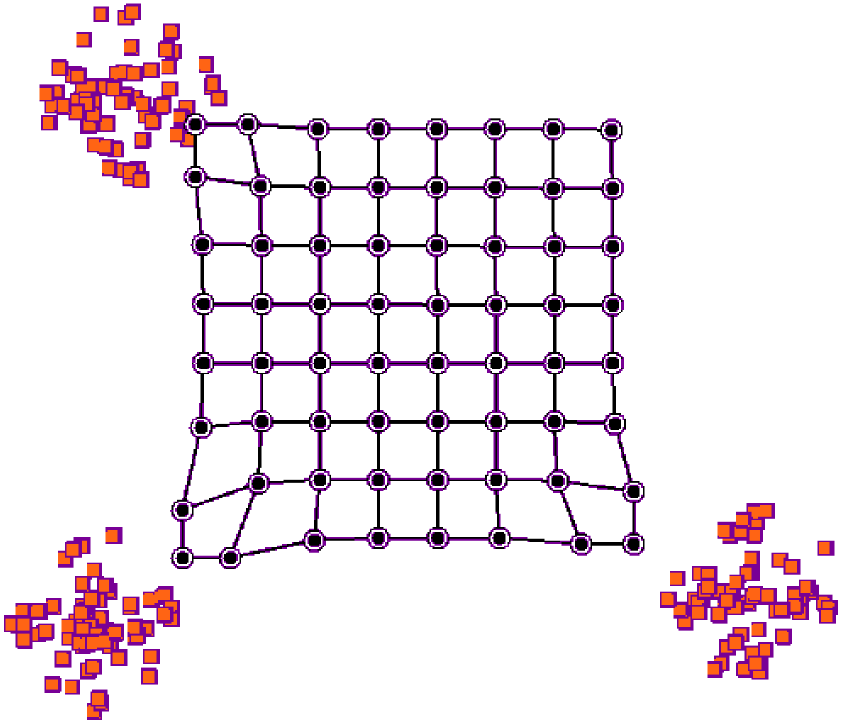
Quantitative Test: Results



Quantitative Test: Results

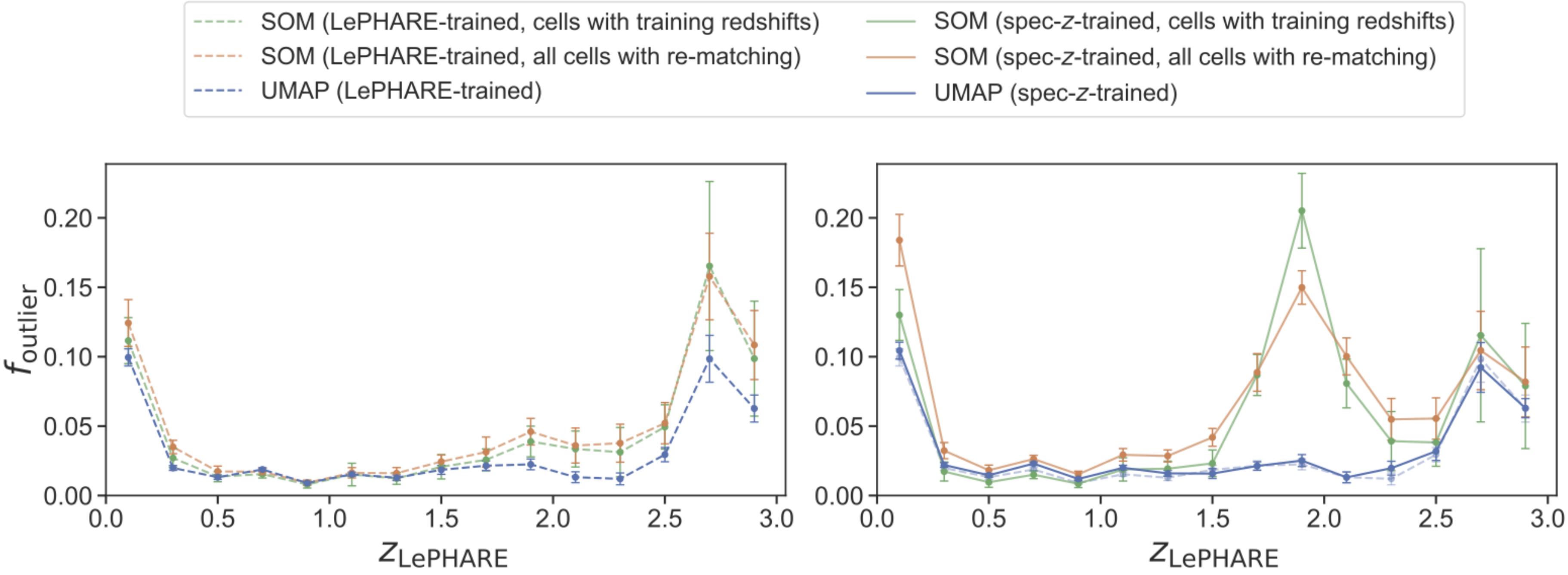


SOM learning

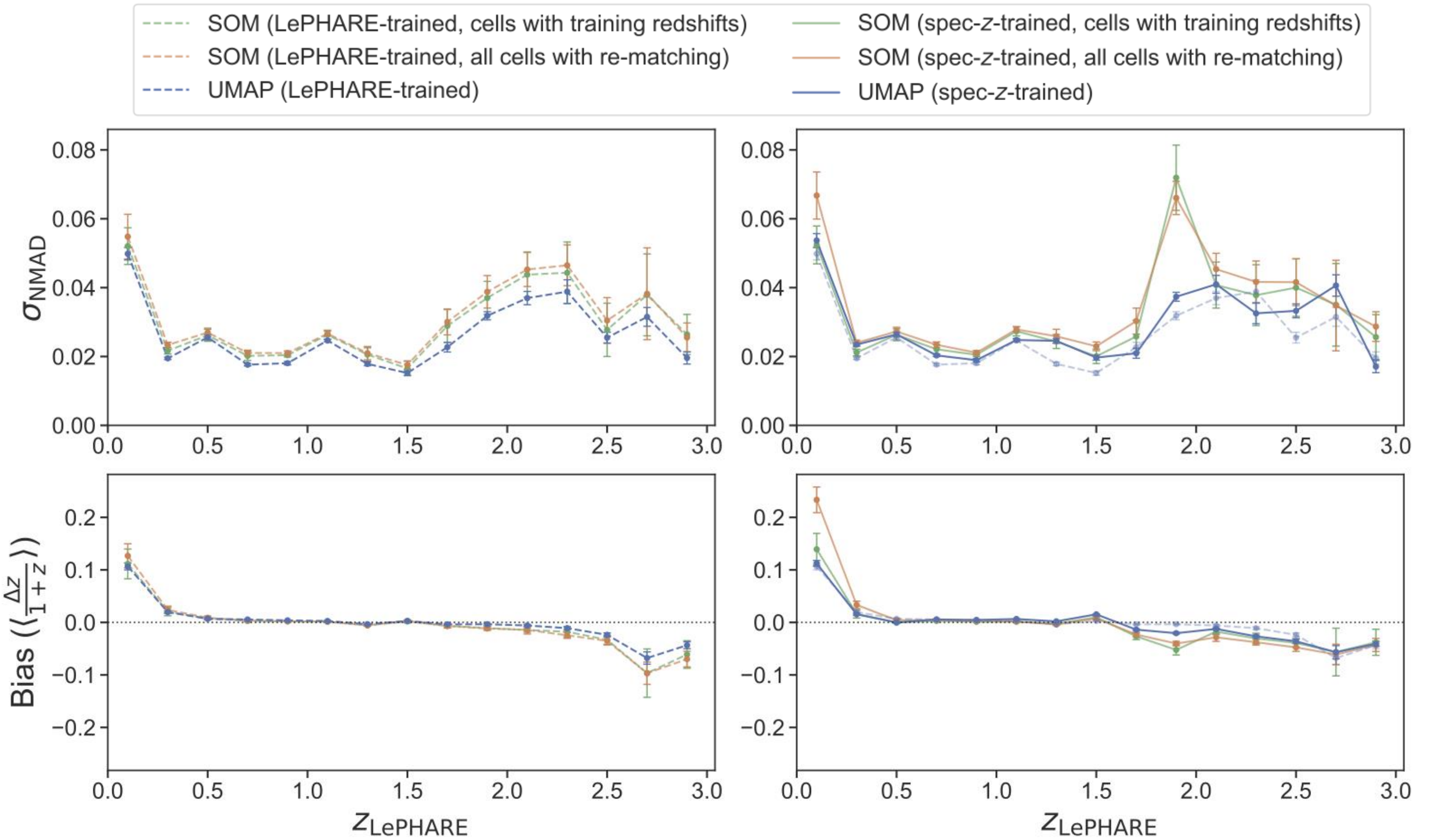


source

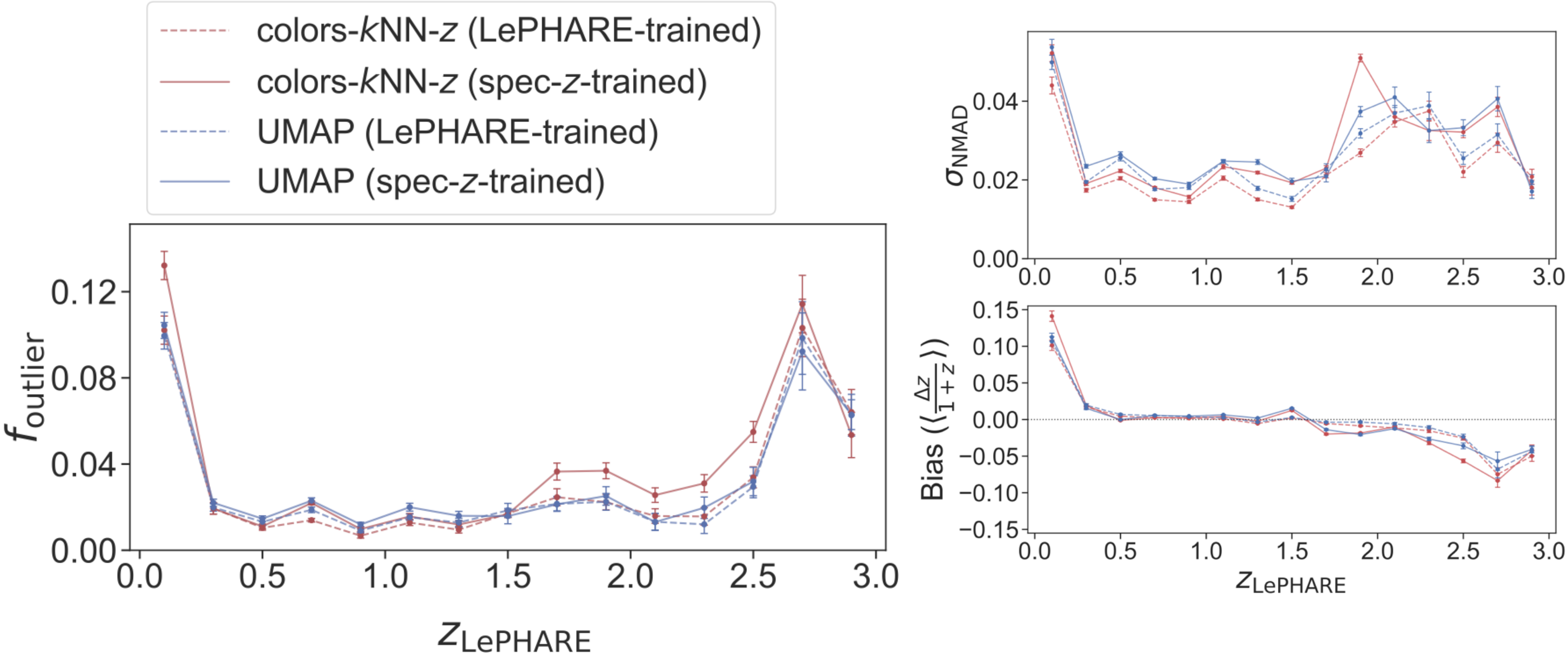
Quantitative Test: Results



Quantitative Test: Results



Quantitative Test: Comparison to *ugrizyJH* color space



	All Redshifts			$z_{\rm LePHARE}>1.5$		
	$f_{\rm outlier}$	$\sigma_{\rm NMAD}$	Bias	$f_{\rm outlier}$	$\sigma_{\rm NMAD}$	Bias
			`LePHARE`-trained			
colors-kNN-z	**0.0198 ± 0.0005**	**0.0182 ± 0.0002**	**0.0039 ± 0.0004**	**0.0377 ± 0.0017**	**0.0256 ± 0.0005**	-0.0251 ± 0.0010
SOM-z [a]	0.0194 ± 0.0026 (0.2666 ±0.0035) [b]	0.0234 ± 0.0006	0.0048 ± 0.0013	0.0368 ± 0.0040 (0.3191 ±0.0080) [c]	0.0315 ± 0.0009	-0.0187 ± 0.0018 [d]
nearest-z	0.0279 ± 0.0017	0.0247 ± 0.0005	0.0064 ± 0.0016	0.0521 ± 0.0032	0.0334 ± 0.0007	-0.0268 ± 0.0021
UMAP-kNN-z	0.0218 ± 0.0005	0.0215 ± 0.0001	0.0064 ± 0.0005	**0.0377 ± 0.0017**	0.0280 ± 0.0005	**-0.0230 ± 0.0008**
			spec-z-trained			
colors-kNN-z	0.0257 ± 0.0006	**0.0223 ± 0.0002**	**0.0037 ± 0.0004**	0.0509 ± 0.0013	0.0342 ± 0.0004	–0.0362 ± 0.0008
SOM-z [e]	0.0215 ± 0.0026 (0.4468 ±0.0034) [f]	0.0236 ± 0.0006	0.0031 ± 0.0012 [g]	0.0725 ± 0.0071 (0.6787 ±0.0079)	0.0339 ± 0.0013	-0.0274 ± 0.0027 [h]
nearest-z	0.0430 ± 0.0018	0.0276 ± 0.0004	0.0088 ± 0.0017	0.0972 ± 0.0043	0.0418 ± 0.0008	-0.0395 ± 0.0021
UMAP-kNN-z	**0.0246 ± 0.0007**	0.0250 ± 0.0001	0.0047 ± 0.0004	**0.0370 ± 0.0016**	**0.0328 ± 0.0004**	**-0.0293 ± 0.0009**

[a] These statistics reflect cells with training data only.

[b] $f_{\rm outlier}$ (all redshifts, `LePHARE`-trained, values only for objects in cells used for SOM-z): 0.0122 ± 0.0004 (colors-kNN-z) and 0.0134 ± 0.0006 (UMAP-kNN-z)

[c] $f_{\rm outlier}$ ($z_{\rm LePHARE}>1.5$ `LePHARE`-trained, values only for objects in cells used for SOM-z), 0.0230 ± 0.0014 (colors-kNN-z) and 0.0233 ± 0.0016 (UMAP-kNN-z)

[d] Bias ($z_{\rm LePHARE}>1.5$ `LePHARE`-trained, values only for objects in cells used for SOM-z): -0.0170 ± 0.0010 (colors-kNN-z) and -0.0155 ± 0.0012 (UMAP-kNN-z)

[e] These statistics reflect cells with training data only.

[f] $f_{\rm outlier}$(all redshifts, spec-z-trained, values only for objects in cells used for SOM-z): 0.0113 ± 0.0005 (colors-kNN-z) and 0.0106 ± 0.0006 (UMAP-kNN-z)

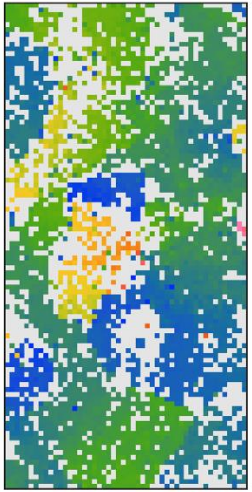
[g] Bias (all redshifts spec-z-trained, values only for objects in cells used for SOM-z): 0.0019 ± 0.0005 (colors-kNN-z) and 0.0020 ± 0.0005 (UMAP-kNN-z)

[h] Bias ($z_{\rm LePHARE}>1.5$ spec-z-trained, values only for objects in cells used for SOM-z): -0.0235 ± 0.0016 (colors-kNN-z) and -0.0208 ± 0.0019 (UMAP-kNN-z)

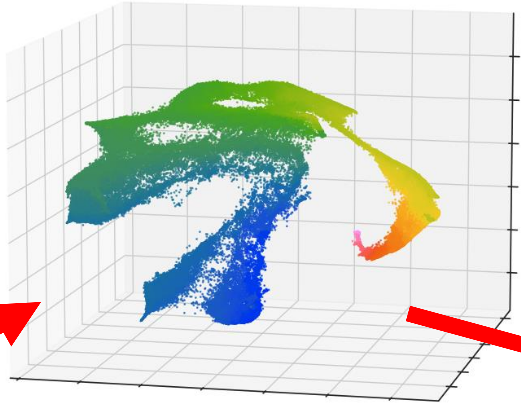
Table 1. This table summarizes the performance of the various algorithms for the two training cases, across all redshifts (left) and for test objects with redshifts $z_{\rm LePHARE}>1.5$ only (right). The parenthetical values in the SOM-z $f_{\rm outlier}$ columns correspond to the values if all test objects in cells with no training data are considered outliers. The bolded values indicate the best algorithm based on the corresponding performance metric for that training case and redshift regime. In several cases, the lowest values of the performance metrics were obtained with SOM-z. We have bolded the next best values instead in these cases, as the SOM-z statistics ignore objects that were placed in cells without training data, making for a lopsided comparison. In all such cases, both colors-kNN-z and UMAP-kNN-z yield superior performance to SOM-z when using identical datasets (i.e., ignoring the objects in SOM cells without training redshifts for colors-kNN-z and UMAP-kNN-z as well). These cases and the corresponding results are detailed in the footnotes. The values for nearest-z ignoring objects in cells without training redshifts are by definition the SOM-z values, so are not reported again.

Ok great, but how does this help optimize photo-*z* training sets in general?

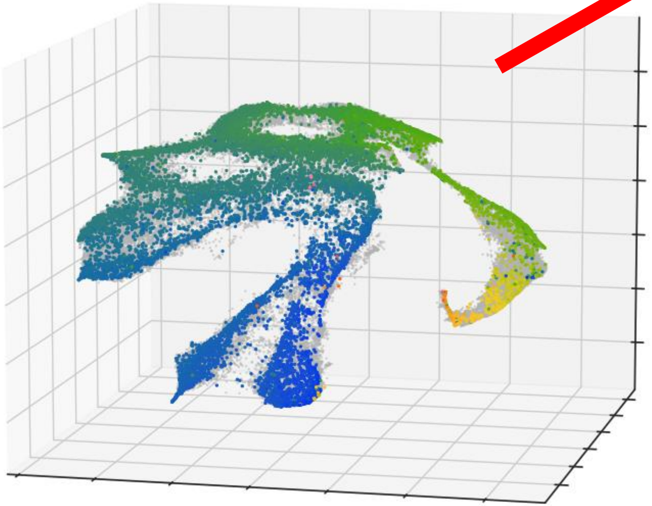
Simple Example: Easy Adoption in SOM-based Frameworks



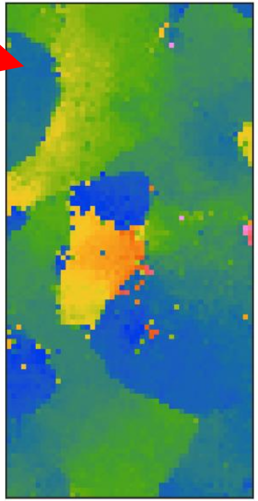
(1) spec-z sources sample only a subset of the color space occupied by galaxies

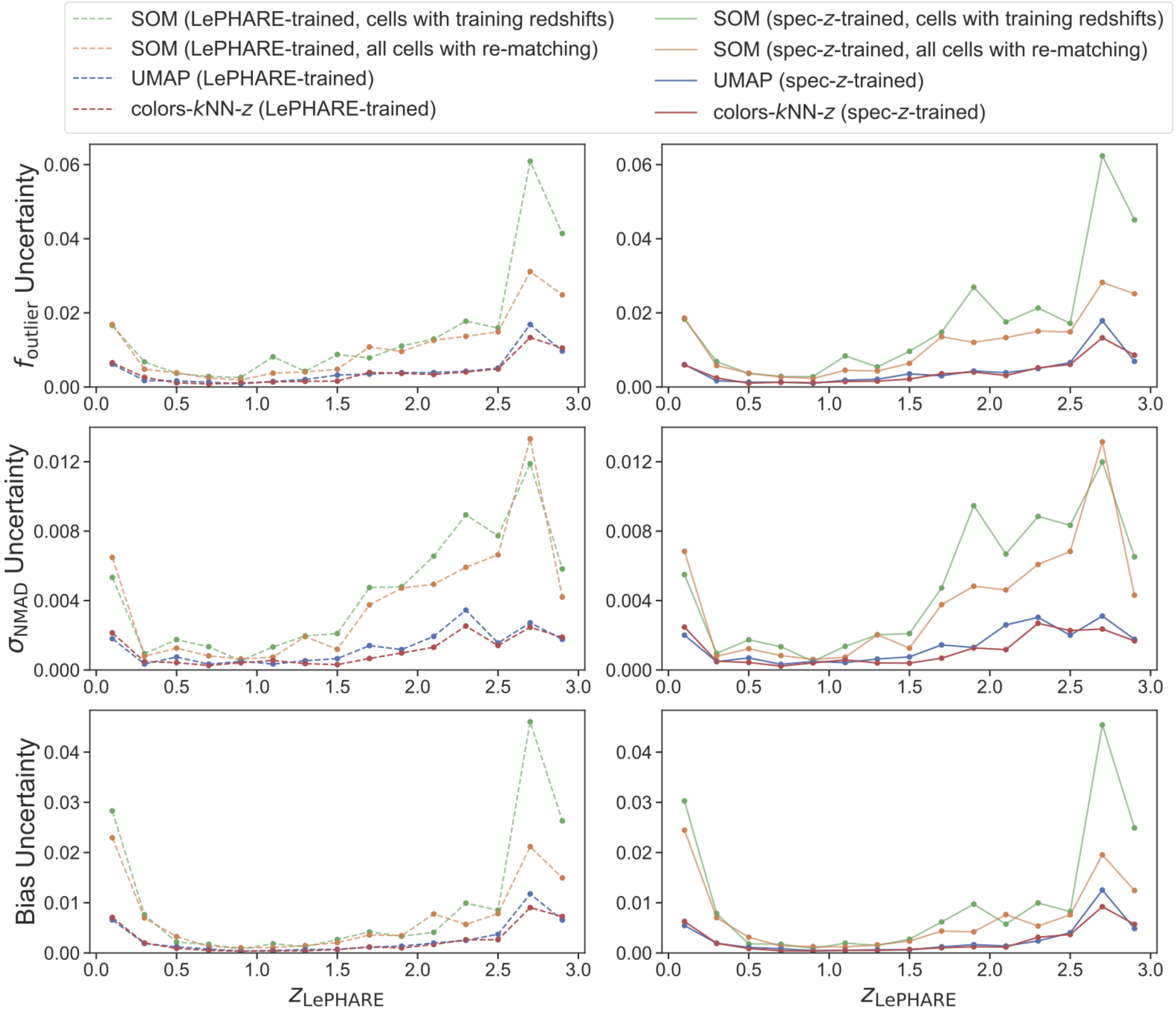


(3) using UMAP-*z*'s to fill in the SOM, this approach can easily be adopted as a preprocessing step in SOM-based frameworks



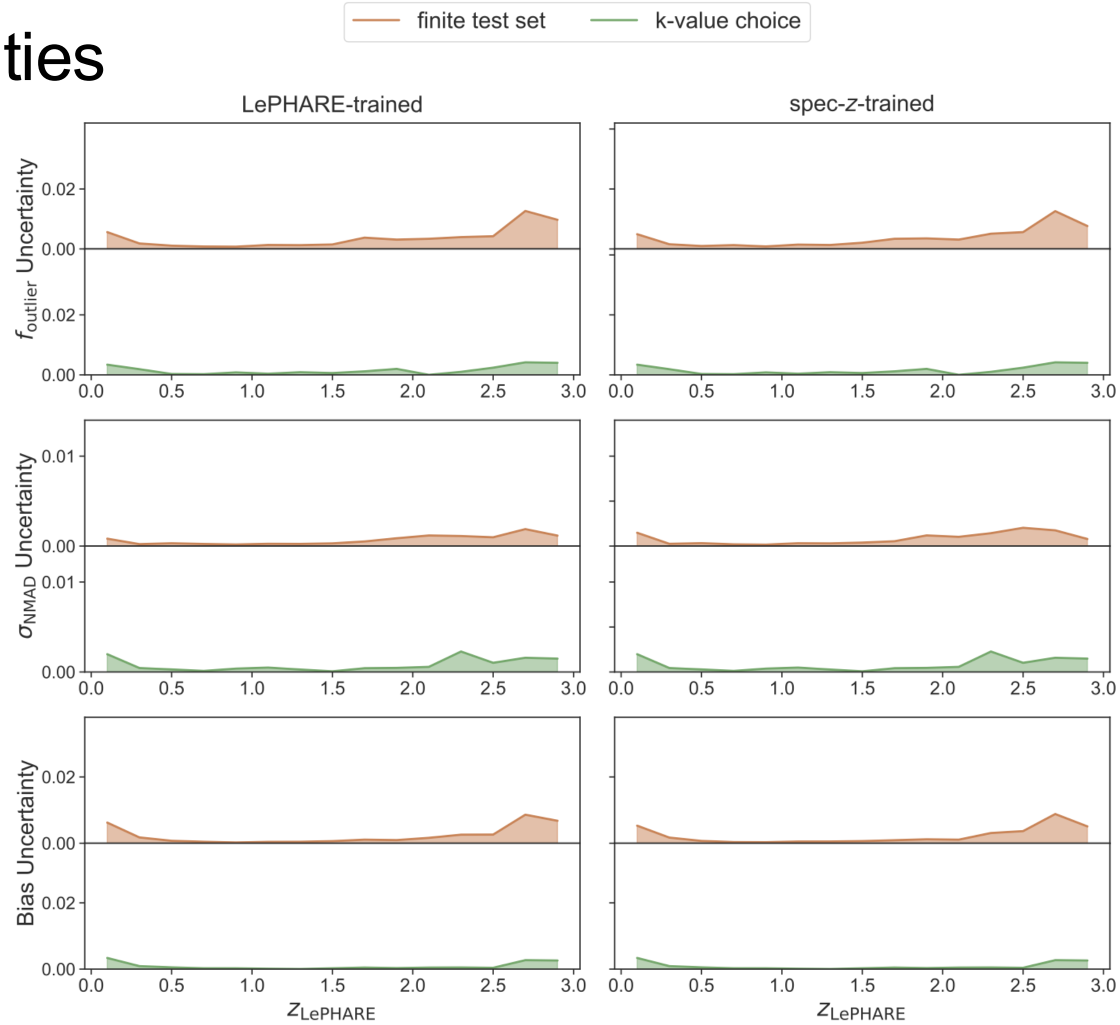
(2) leverage the UMAP's coherent mapping to *z* to fill in the missing redshifts



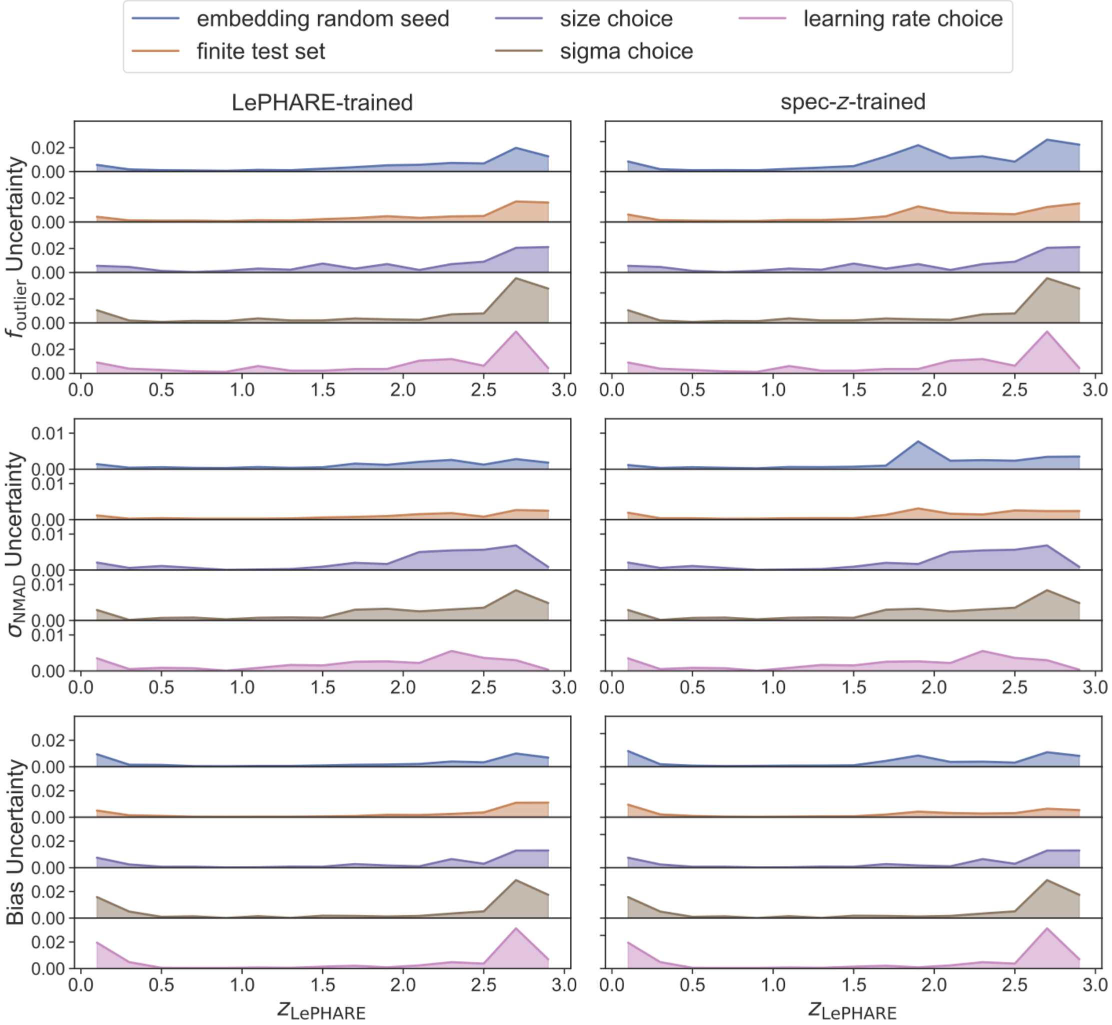


SOM (LePHARE-trained, cells with training redshifts)
SOM (LePHARE-trained, all cells with re-matching)
UMAP (LePHARE-trained)
colors-kNN-z (LePHARE-trained)
SOM (spec-z-trained, cells with training redshifts)
SOM (spec-z-trained, all cells with re-matching)
UMAP (spec-z-trained)
colors-kNN-z (spec-z-trained)
f_{outlier} Uncertainty
σ_{NMAD} Uncertainty
Bias Uncertainty
z_{LePHARE}
z_{LePHARE}

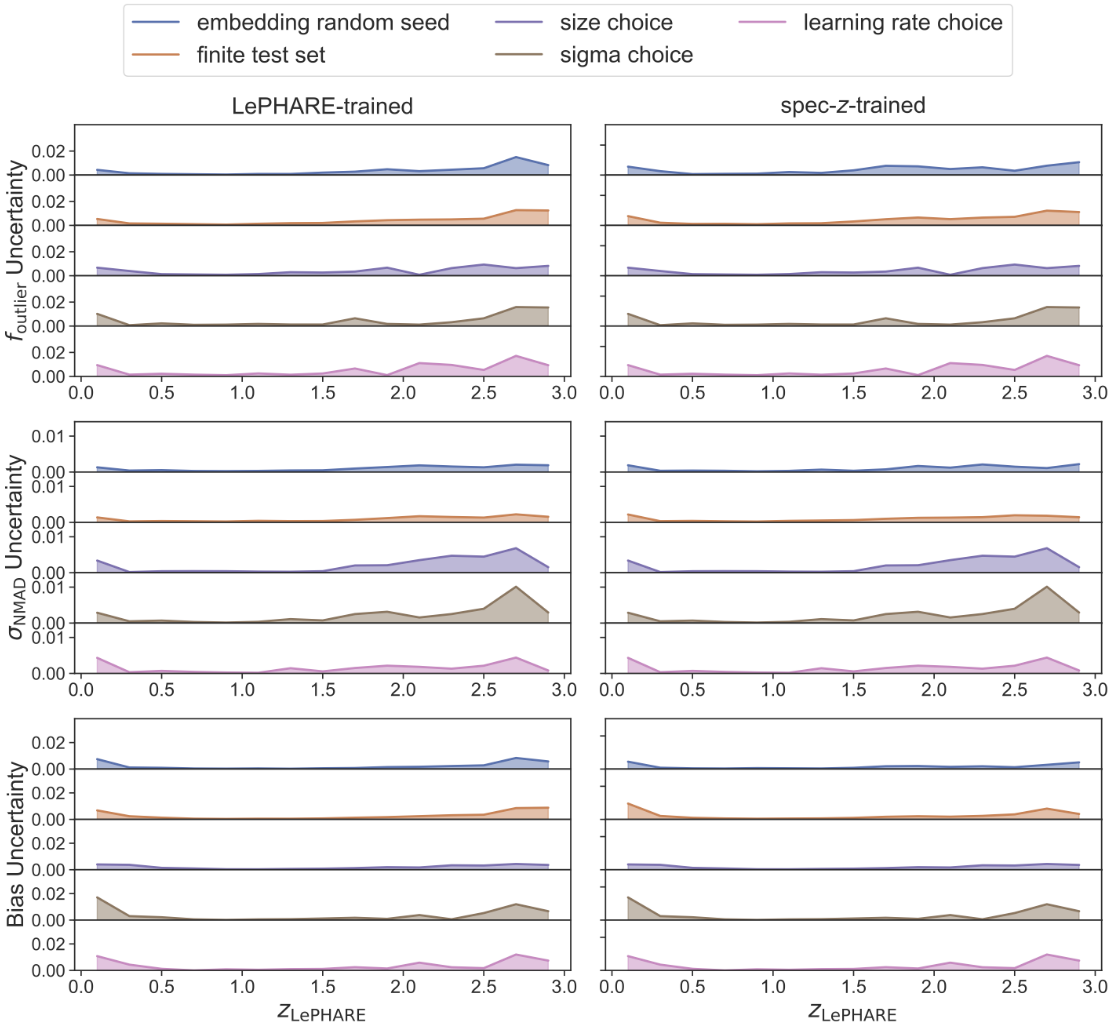
colors-*k*NN-*z* uncertainties



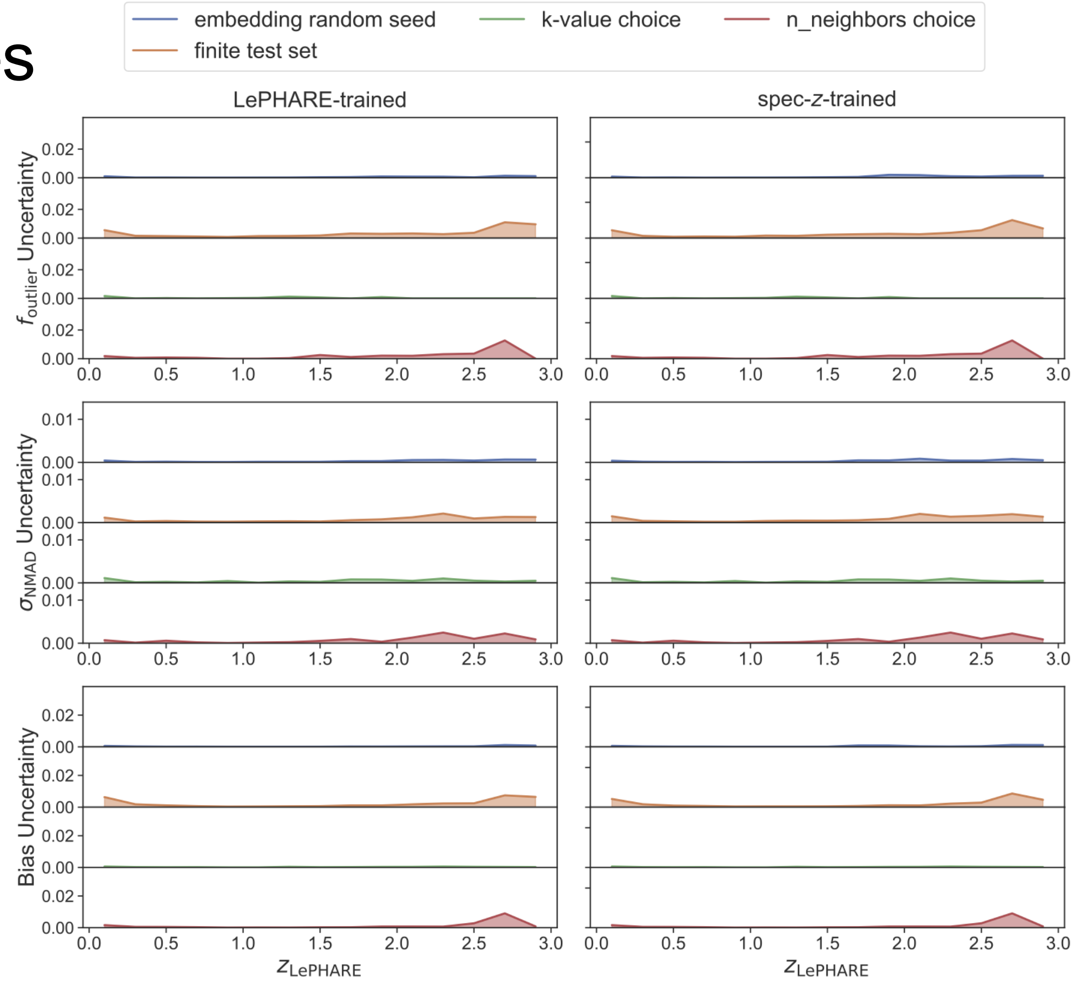
SOM-z uncertainties



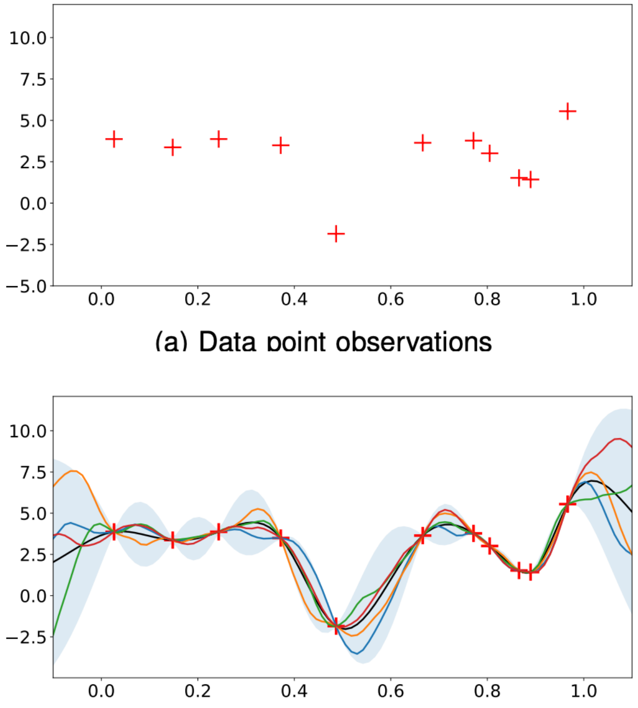
nearest-z uncertainties



UMAP-kNN-z uncertainties



Currently: extending this research



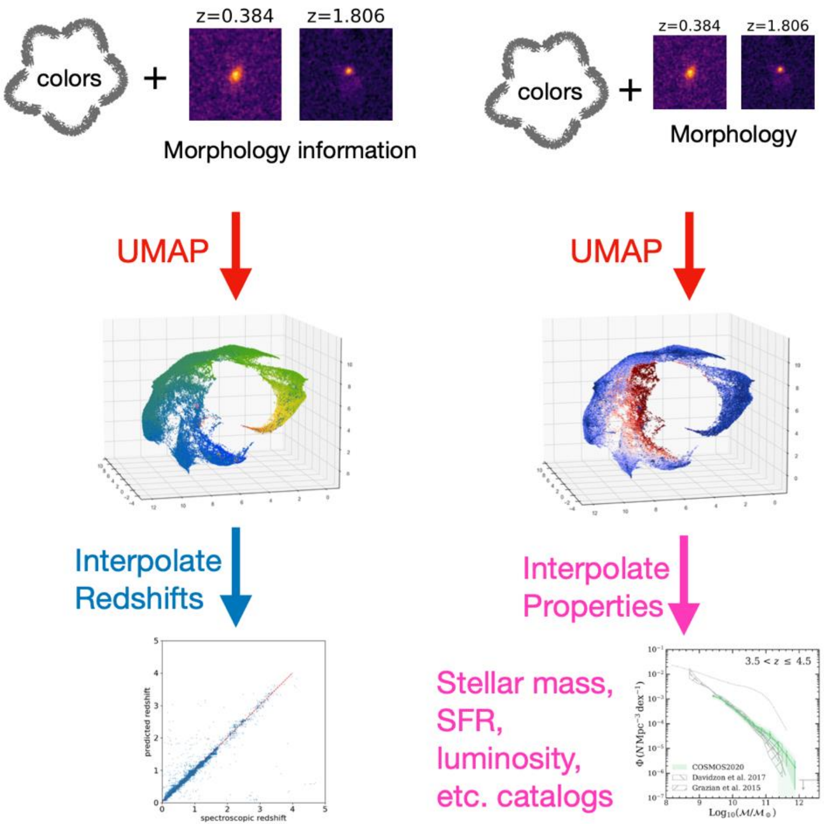
(a) Data point observations

(b) Five possible functions by GPR

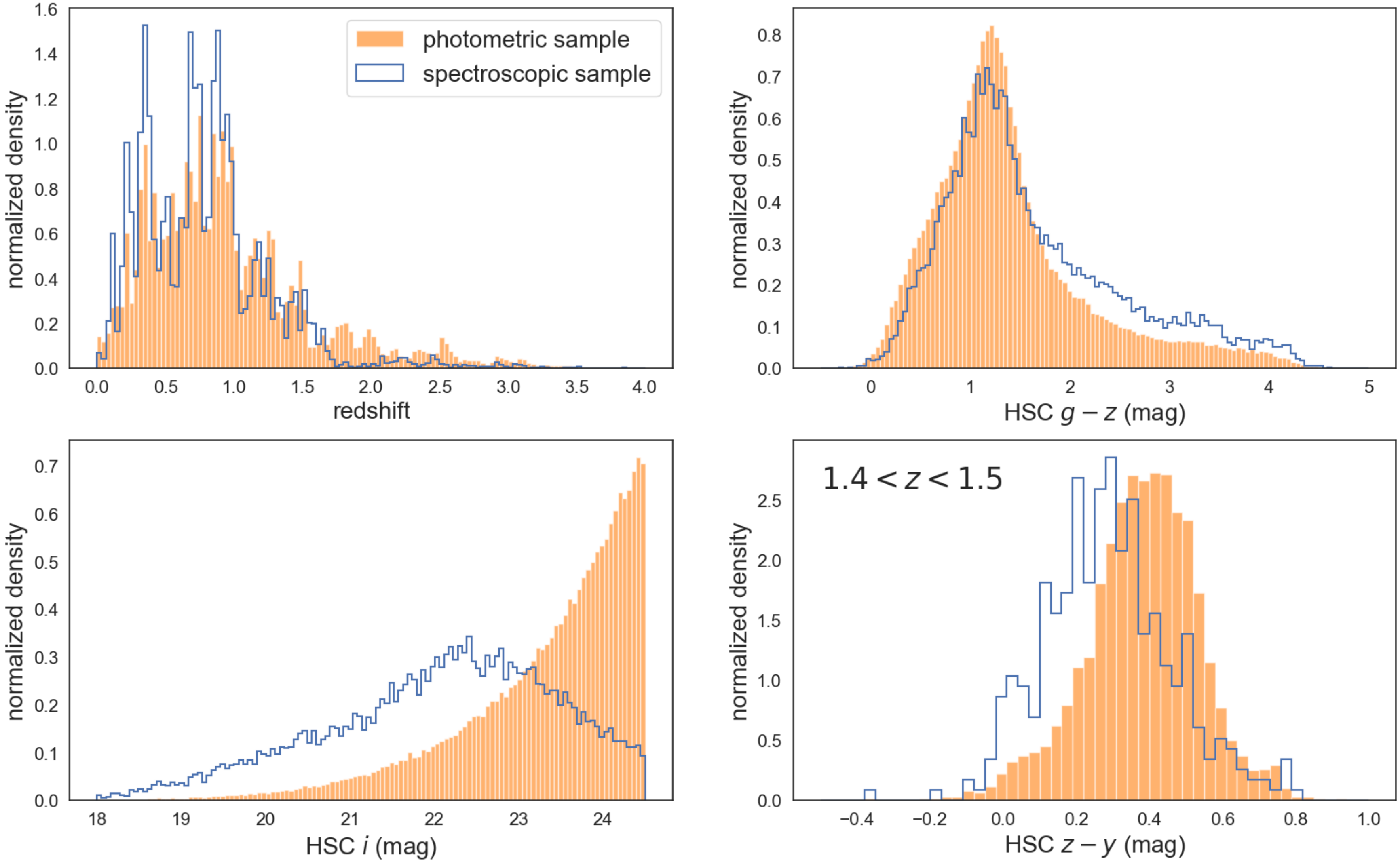
FIGURE 1: A regression example: (a) Observed data points, (b) Five sample functions fitting the observed data points.

- the first paper used *k*NN interpolation in the UMAP space → this is far from the most sophisticated approach!
- robust GPR will allow us to get better predictions + redshift PDFs throughout the UMAP manifold!

Third Paper and Beyond



- catalog-level morphology information (axis ratio, Sersic index, surface brightness, etc.) can be useful for redshift prediction
  - potentially break degeneracies in optical color spaces this way, reduce reliance on infrared photometry → big for LSST!
- leverage the nice mapping to sSFR shown in this work to predict properties other than redshift using similar UMAP-based techniques



photometric sample
spectroscopic sample
normalized density
redshift
HSC $g-z$ (mag)
HSC i (mag)
$1.4 < z < 1.5$
HSC $z-y$ (mag)

Metrics:

- **the fraction of outliers** (f_{outlier}), defined as the fraction of predicted photo-z's for which $|\frac{\Delta z}{1+z}| > 0.15$;

- **the normalized median absolute deviation** (σ_{NMAD}), defined as $1.4826 \times \text{Median}(|\frac{\Delta z}{1+z} - \text{Median}(\frac{\Delta z}{1+z})|)$; and

- **the prediction bias**, defined as $\langle \frac{\Delta z}{1+z} \rangle$.