Probabilistic analysis methods for cosmology using uncertainty-dominated photometric data

Alex Malz *advised by David Hogg*

3 May 2019

A dissertation defended in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Department of Physics, New York University

Overview

Cosmology with large photometric galaxy surveys

Stress-testing probabilistic redshift estimates

Practicalities of probabilistic data

Hierarchical inference of the cosmological parameters

Summary and future directions

Overview

Cosmology with large photometric galaxy surveys

Stress-testing probabilistic redshift estimates

Practicalities of probabilistic data

Hierarchical inference of the cosmological parameters

Summary and future directions

Dark energy accelerates the expansion of the universe.

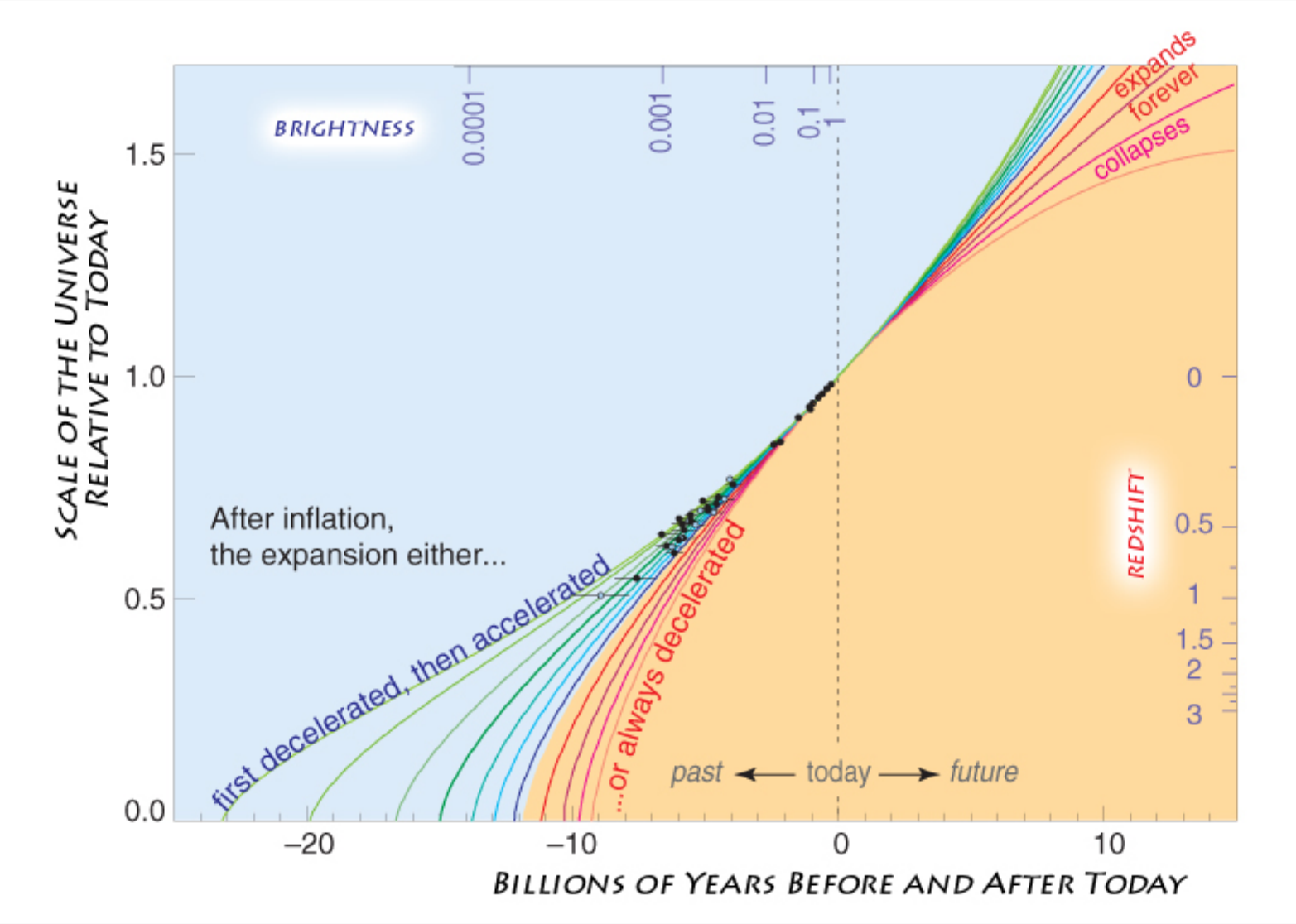
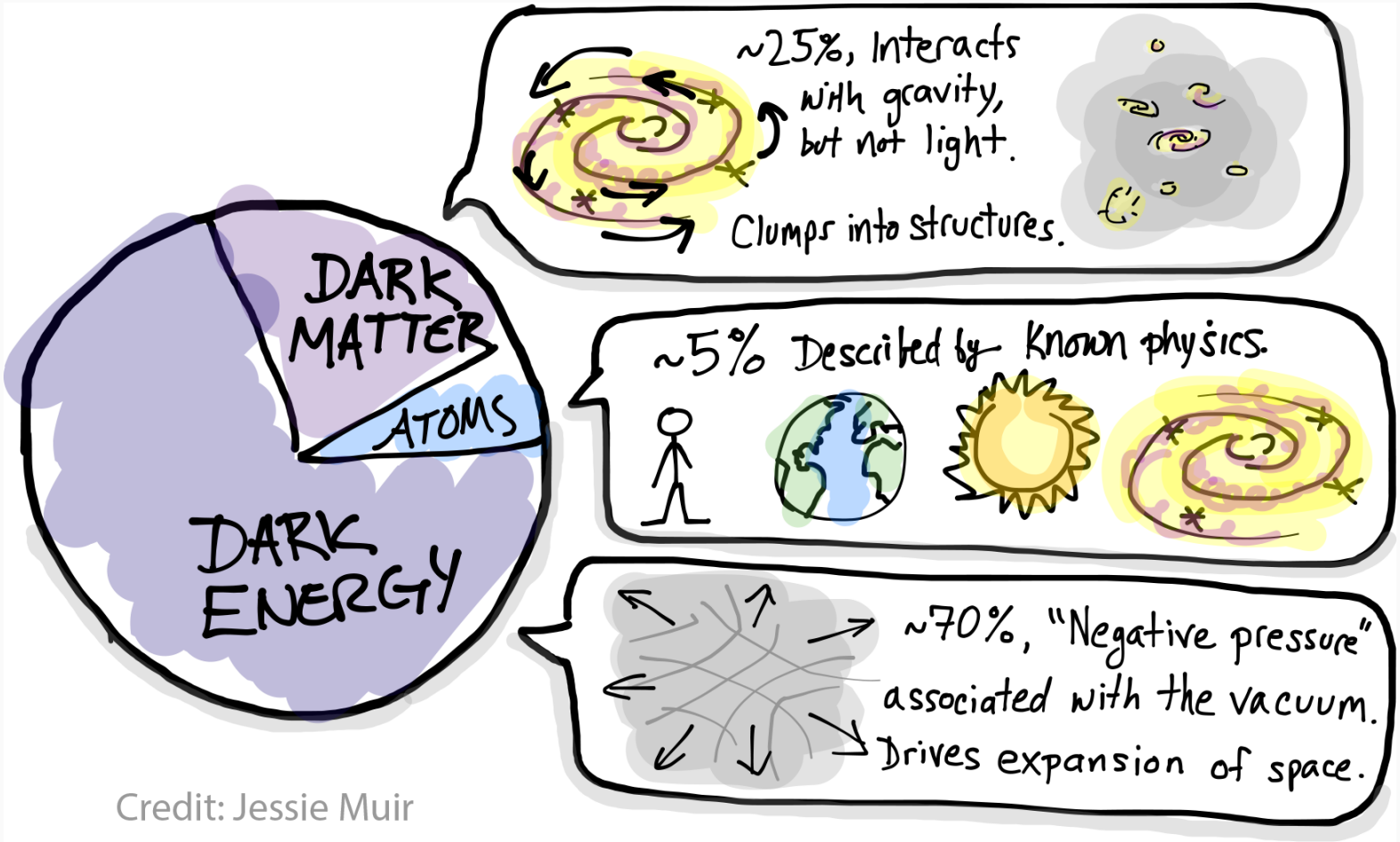


image source: *Saul Perlmutter*

Dark energy and dark matter dominate the universe.



Credit: Jessie Muir

DE & DM models predict the distribution of matter.

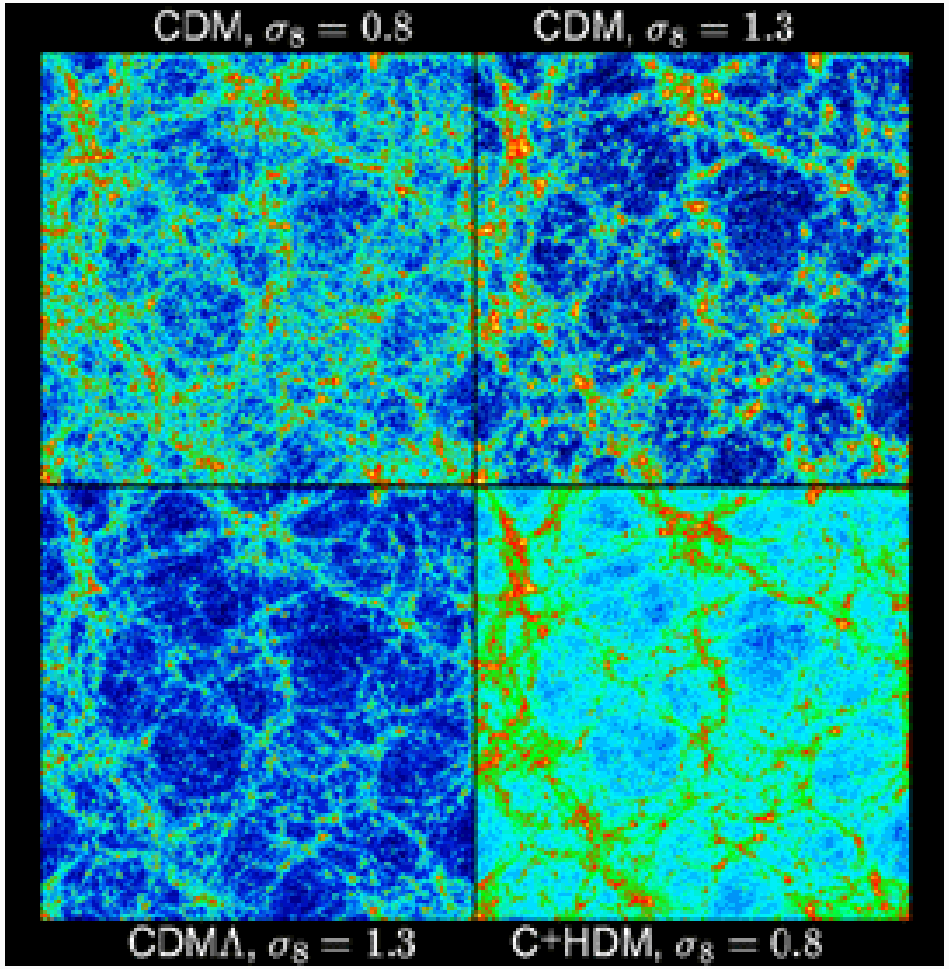


image source: *Swinburne Centre for Astrophysics and Supercomputing*

Galaxy positions trace the large-scale structure of dark matter.

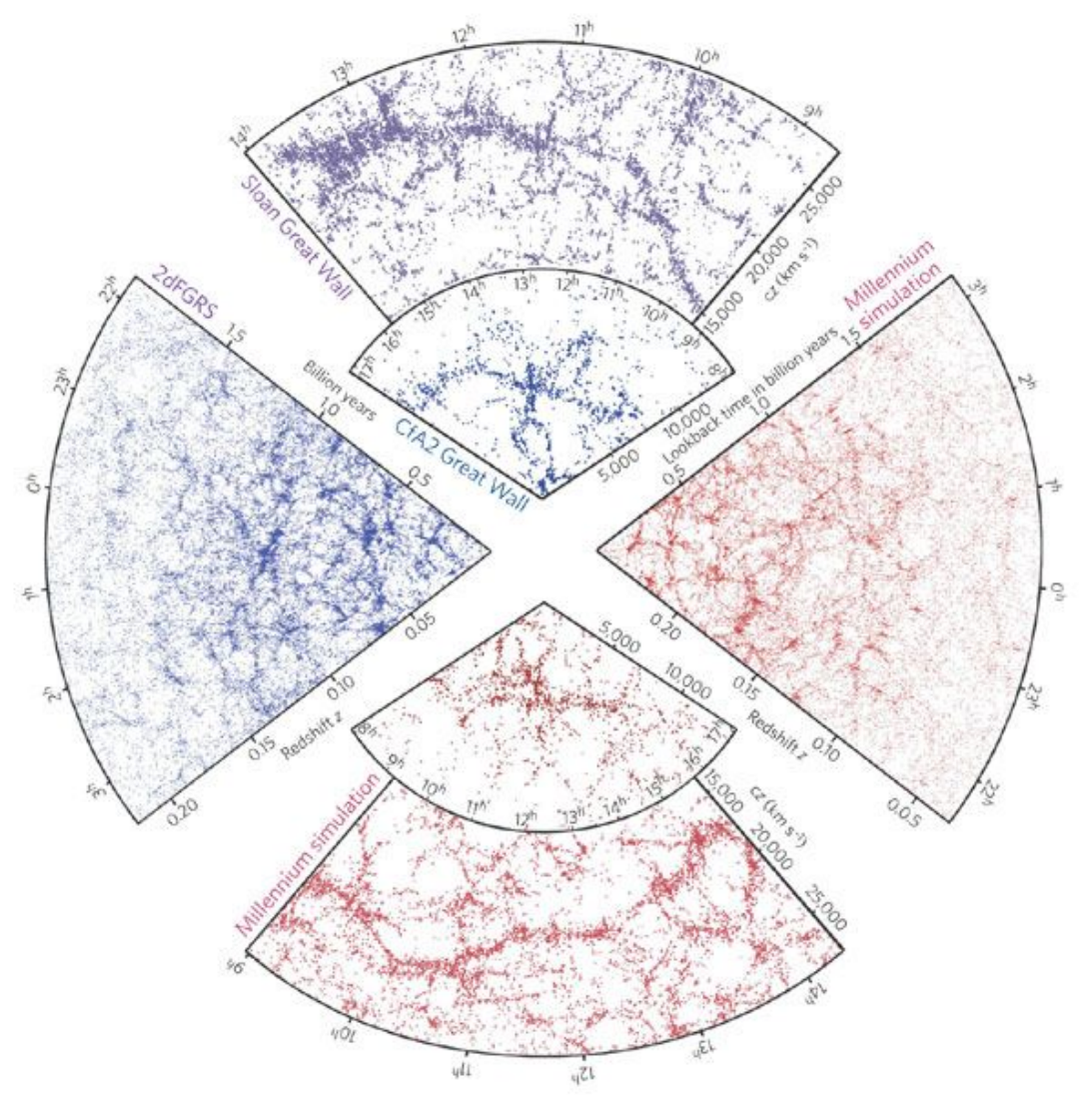


image source: *Springel+06*

Galaxy redshifts are related to galaxy distances.

Galaxy redshifts are related to galaxy distances.

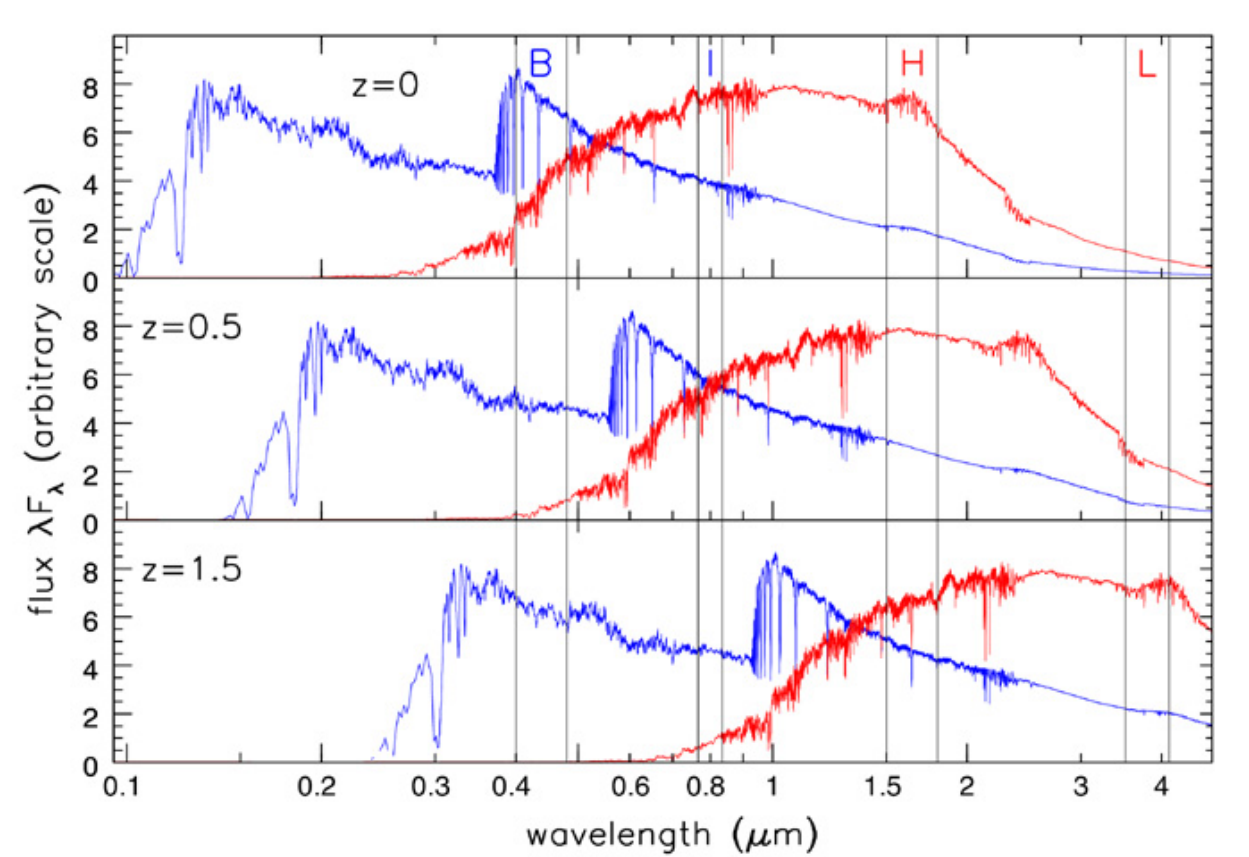


Fig 8.12 (S. Charlot) 'Galaxies in the Universe' Sparke/Gallagher CUP 2007

LSST will observe billions of galaxies without spectroscopy.

image source: BNL

Photo-zs approximate redshifts in lieu of spectra.

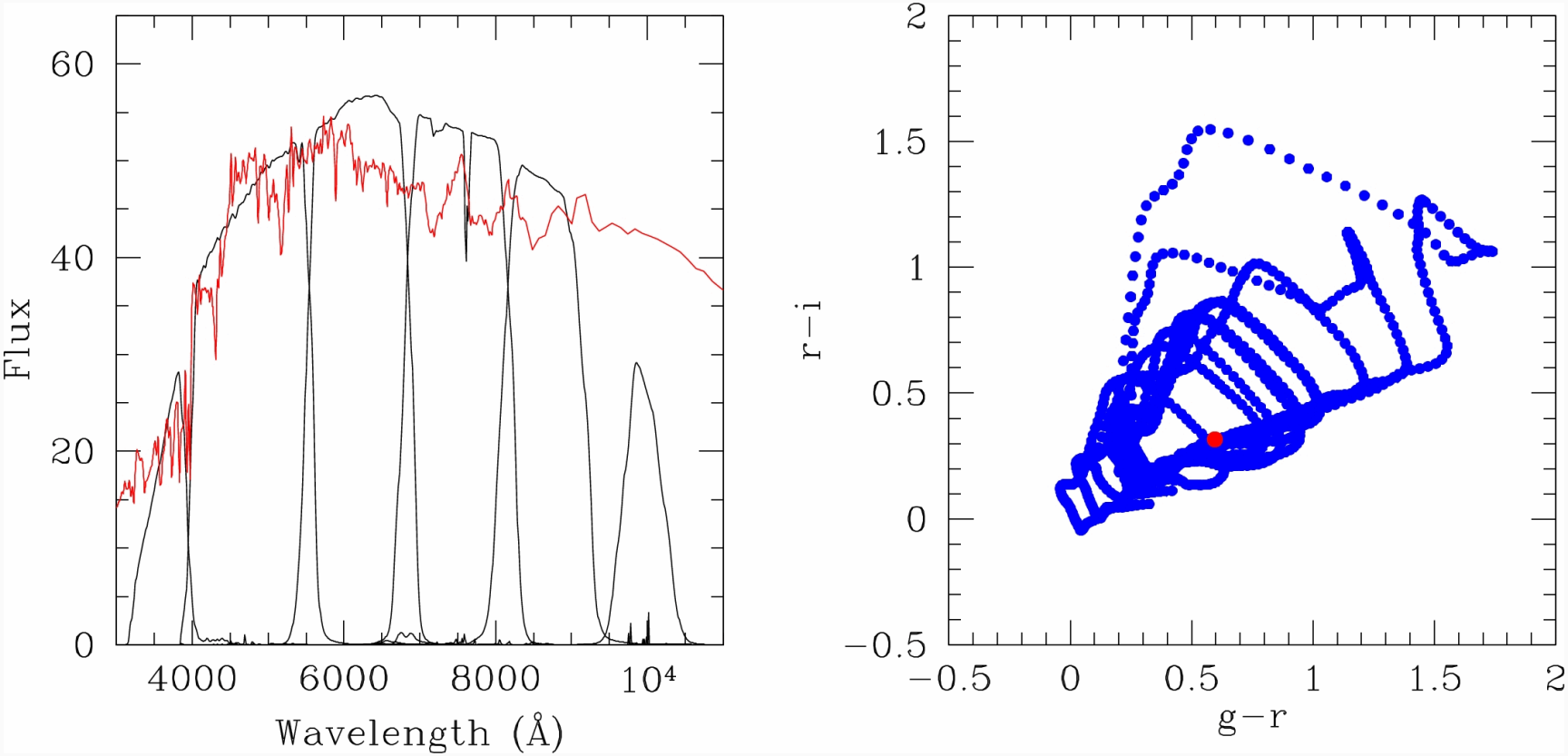


image source: LSST

Photo-zs approximate redshifts in lieu of spectra.

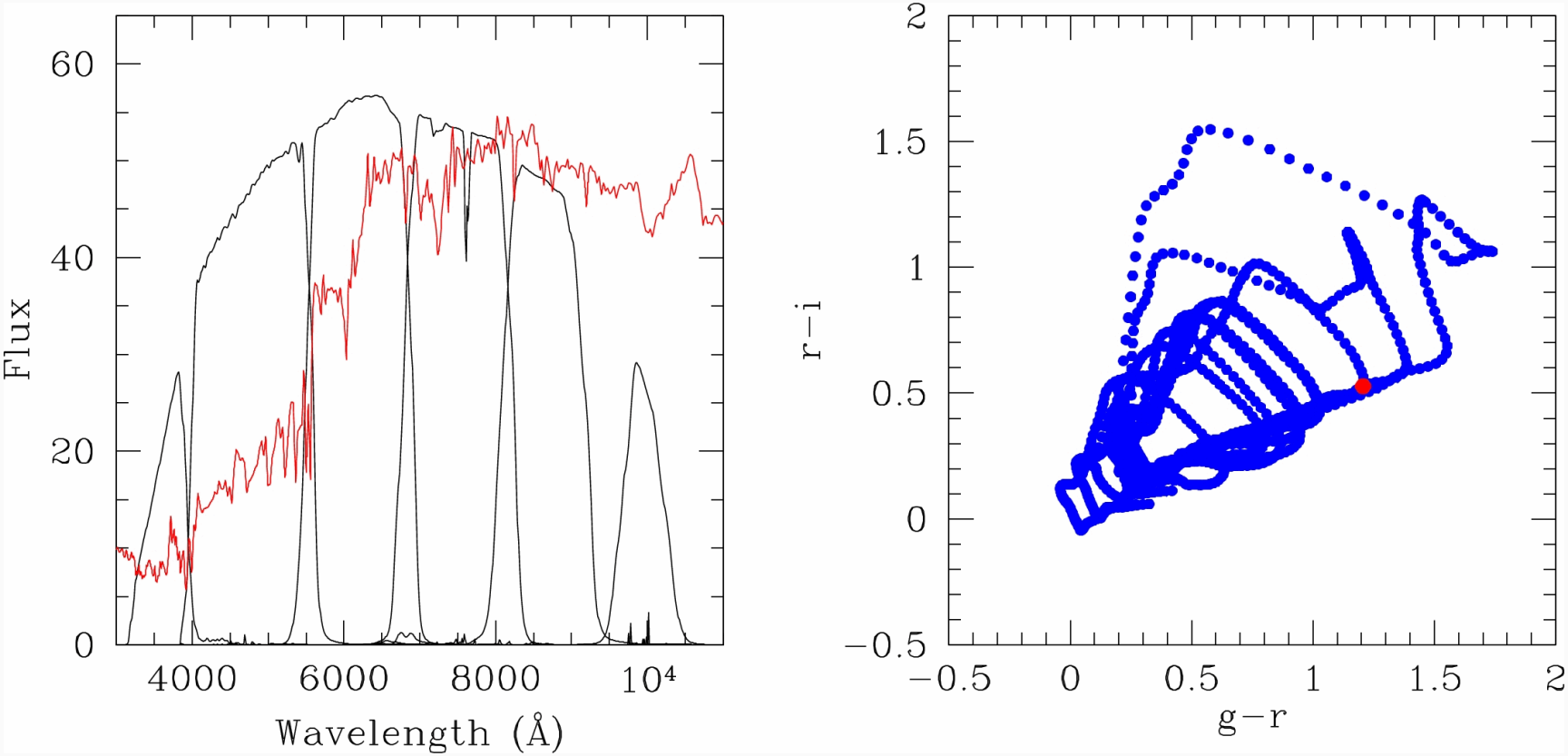


image source: LSST

Photo-zs approximate redshifts in lieu of spectra.

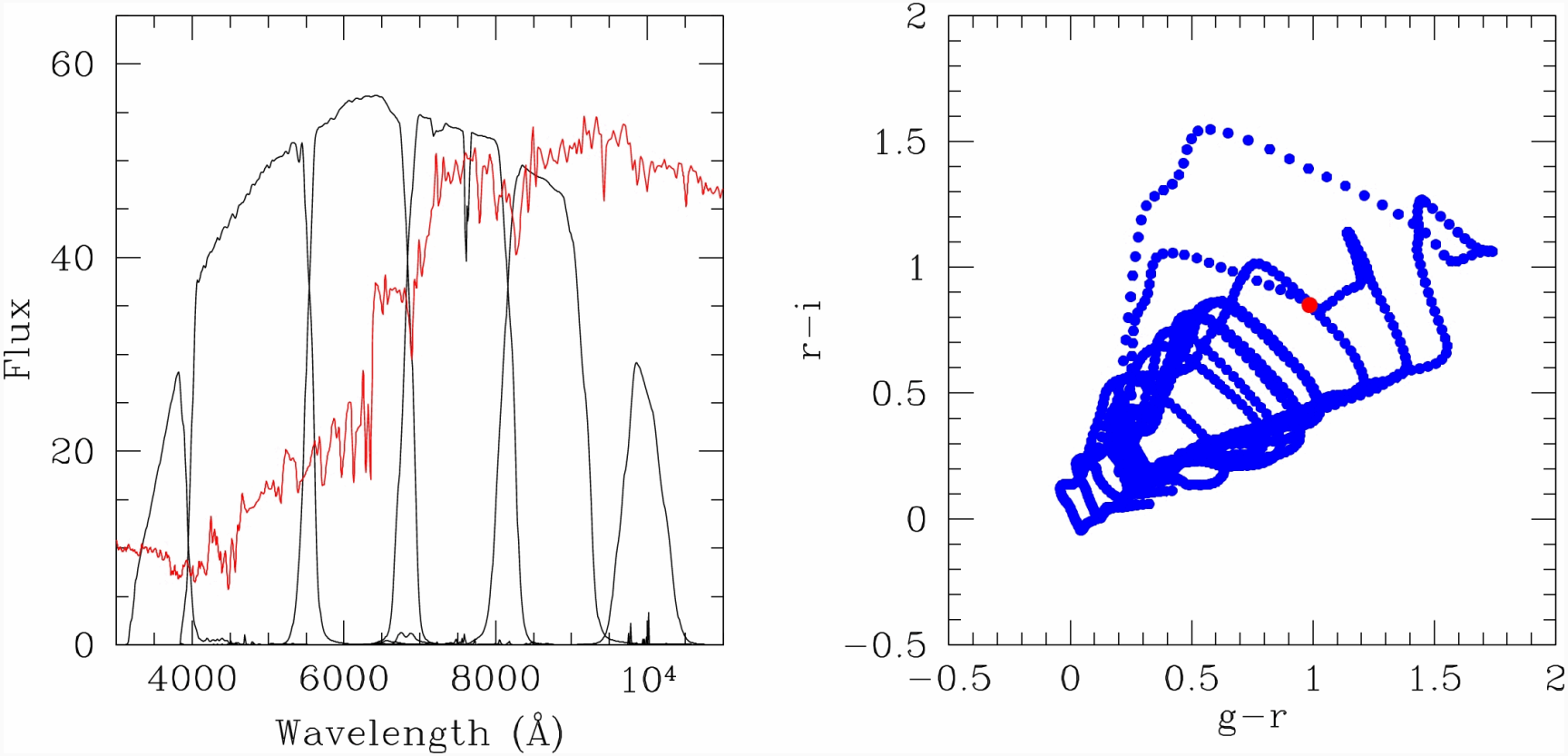


image source: LSST

Photo-zs suffer from uncertainty.

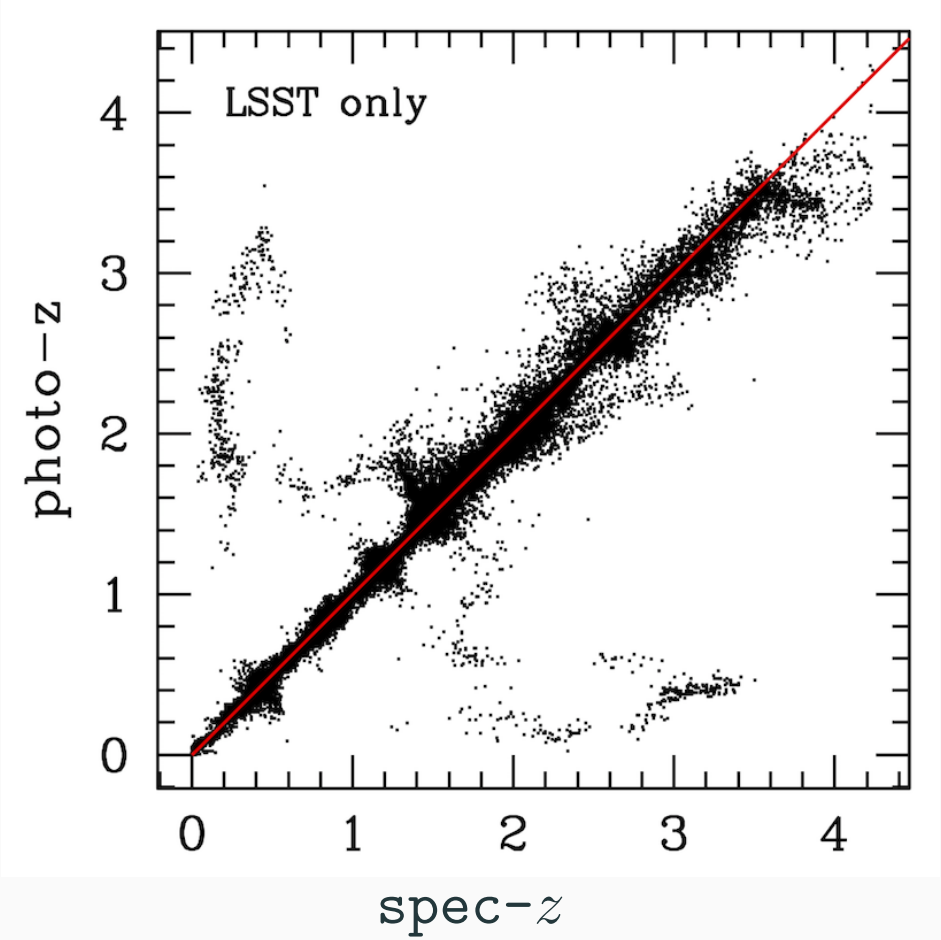


image source: Jain+15

Overview

Cosmology with large photometric galaxy surveys

Stress-testing probabilistic redshift estimates

Practicalities of probabilistic data

Hierarchical inference of the cosmological parameters

Summary and future directions

Photo-z PDFs convey nontrivial uncertainties.

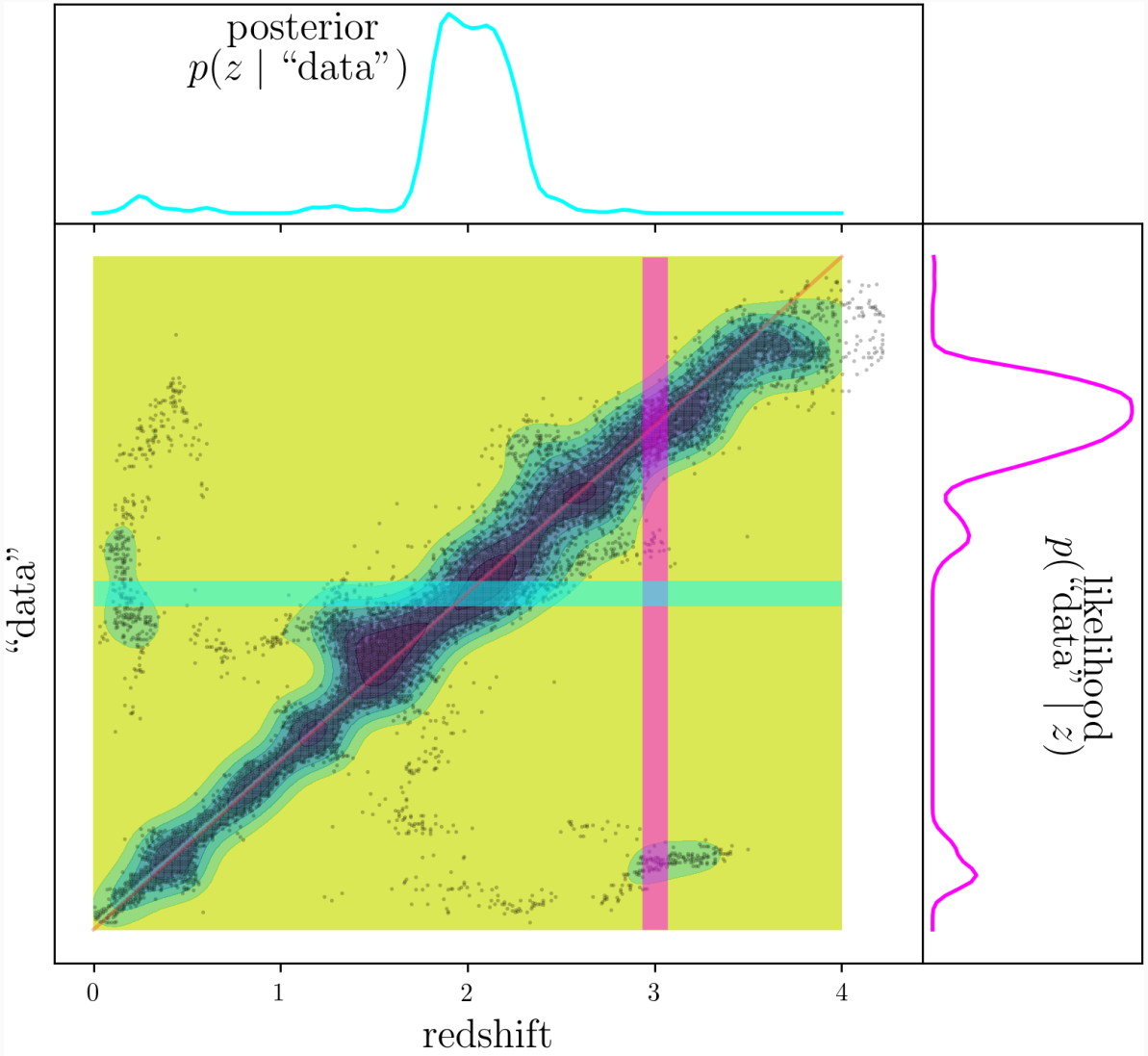


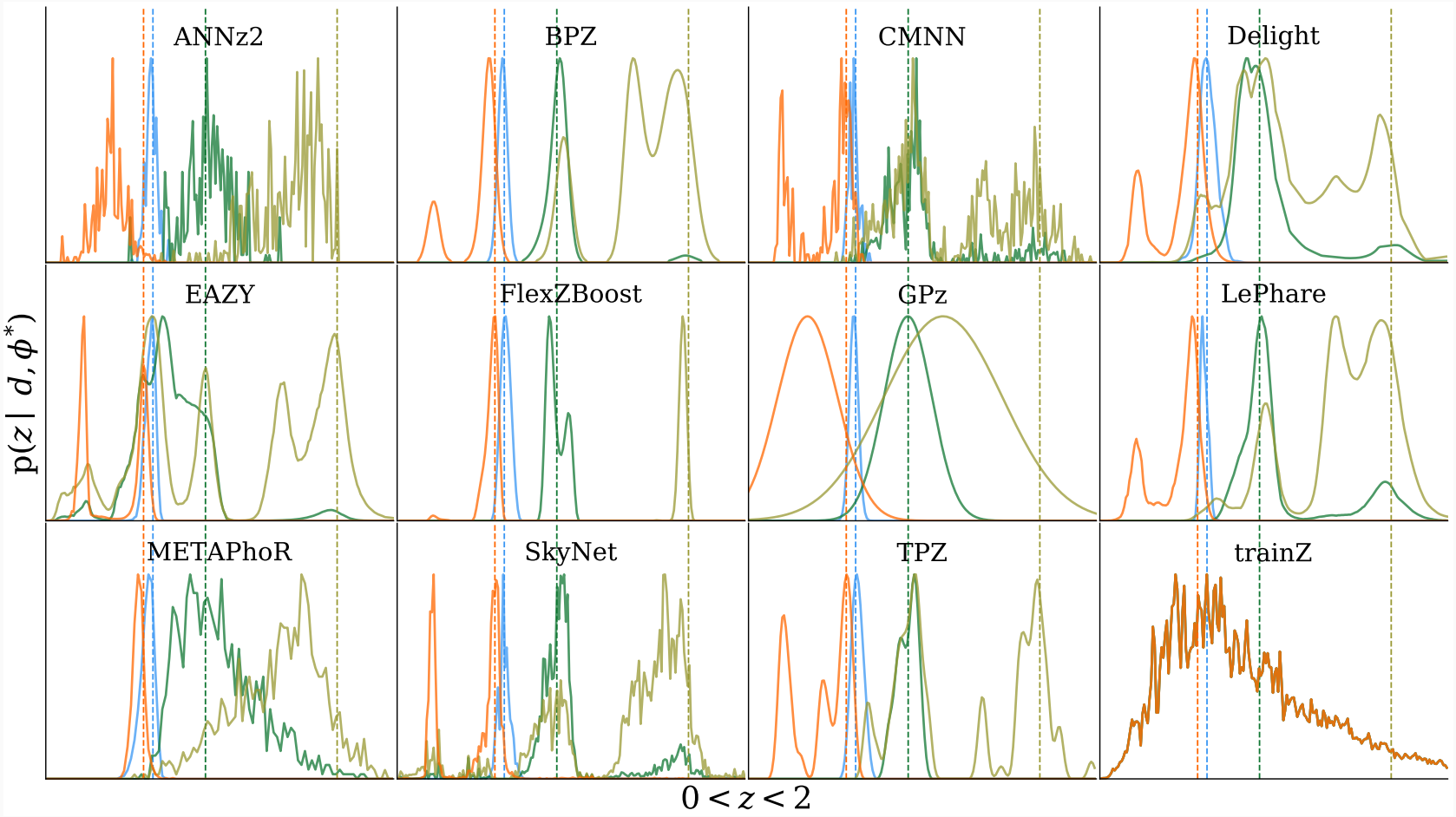
Photo-z PDFs come with their own ~~problems~~ opportunities.

Photo-z PDFs come with their own ~~problems~~ opportunities.

LSST-DESC PZ WG DC1 experimental design

- Goal: benchmark many competing photo-z PDF codes
- Data: redshifts & photometry catalog painted on N-body simulation
- Control: idealized prior information
- Test conditions: true template library for model-fitting codes & representative training set for data-driven codes
- Analysis criteria: TBD

Photo-z PDF codes are distinguished by their implicit priors.



Schmidt&Malz+(in review)

The state-of-the-art trainZ algorithm

The state-of-the-art trainZ algorithm

1. Make histogram of training set galaxy redshifts.

The state-of-the-art trainZ algorithm

1. Make histogram of training set galaxy redshifts.
2. Read in data file of photometry.

The state-of-the-art trainZ algorithm

1. Make histogram of training set galaxy redshifts.
2. Read in data file of photometry.
3. Count N lines of photometry.

The state-of-the-art trainZ algorithm

1. Make histogram of training set galaxy redshifts.
2. Read in data file of photometry.
3. Count N lines of photometry.
4. Discard the data.

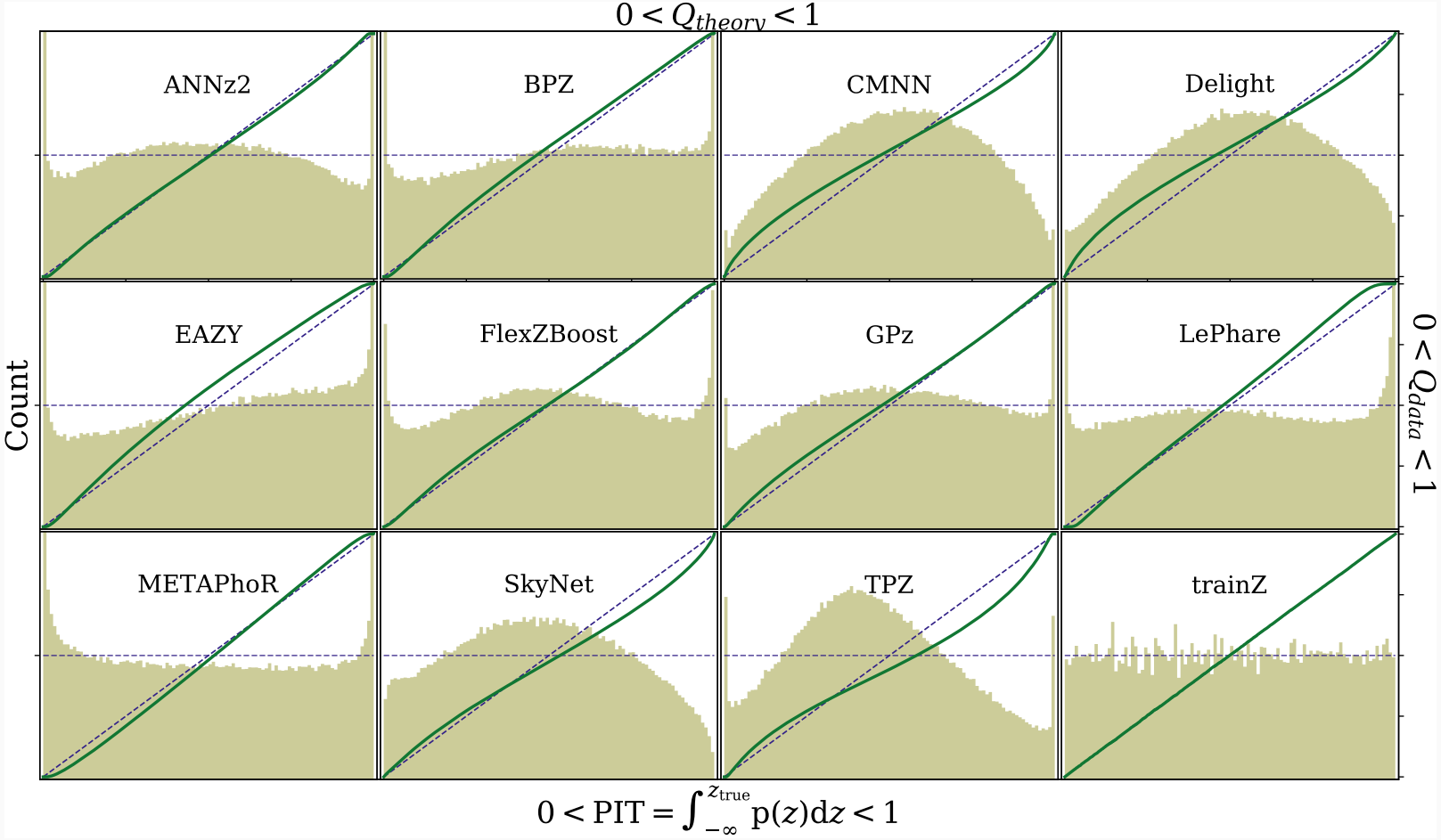
The state-of-the-art trainZ algorithm

1. Make histogram of training set galaxy redshifts.
2. Read in data file of photometry.
3. Count N lines of photometry.
4. Discard the data.
5. Print the histogram of training set galaxy redshifts N times.

The state-of-the-art trainZ algorithm

1. Make histogram of training set galaxy redshifts.
2. Read in data file of photometry.
3. Count N lines of photometry.
4. Discard the data.
5. Print the histogram of training set galaxy redshifts N times.
6. . . . Profit!

Established photo-z PDF comparison metrics are inadequate.



Schmidt&Malz+(in review)

LSST-DESC PZ DC1 summary

Next steps for vetting photo-z PDF algorithms

Use mock data with true photo-z PDFs to recover.

Test sensitivity to realistically degraded prior information.

Isolate or minimize implicit priors.

Stumbling blocks on the road ahead

Conclusions depend on the choice of comparison metric.

Inappropriate metrics can be gamed.

Without true photo-z PDFs, few principled metrics are available.

Check out the full results at

`github.com/LSSTDESC/PZDC1_Paper`.

Overview

LSST will provide photo-z PDFs to astronomers.

LSST will provide photo-z PDFs to astronomers.

Why must photo-z PDFs be approximated?

- hedging bets in absence of obviously superior method
- optimizing limited storage resources for big data

LSST will provide photo-z PDFs to astronomers.

Why must photo-z PDFs be approximated?

- hedging bets in absence of obviously superior method
- optimizing limited storage resources for big data

How can photo-z PDFs be stored?

LSST will provide photo-z PDFs to astronomers.

Why must photo-z PDFs be approximated?

- hedging bets in absence of obviously superior method
- optimizing limited storage resources for big data

How can photo-z PDFs be stored?

- histogram

LSST will provide photo-z PDFs to astronomers.

Why must photo-z PDFs be approximated?

- hedging bets in absence of obviously superior method
- optimizing limited storage resources for big data

How can photo-z PDFs be stored?

- histogram
- samples

LSST will provide photo-z PDFs to astronomers.

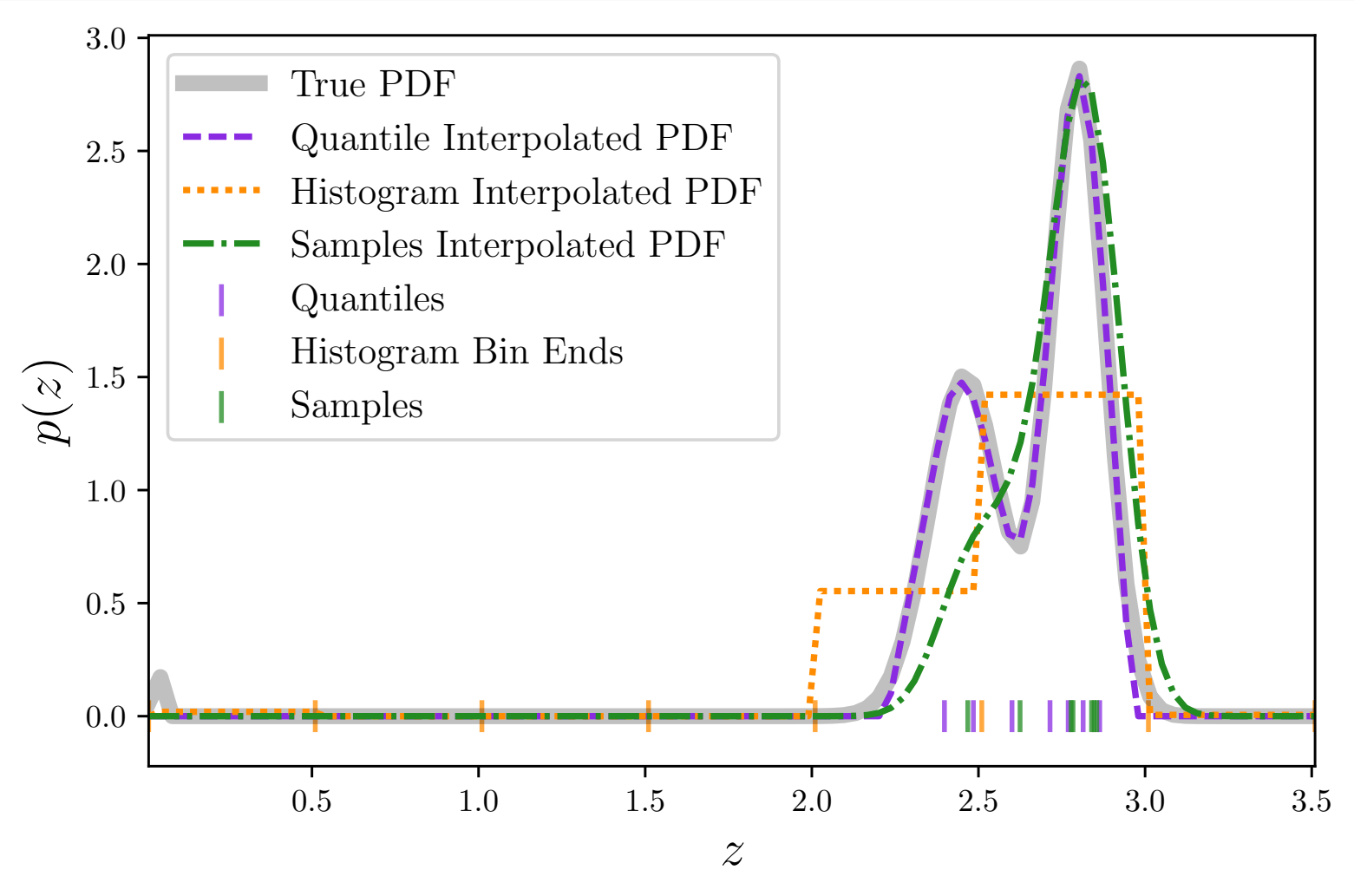
Why must photo-z PDFs be approximated?

- hedging bets in absence of obviously superior method
- optimizing limited storage resources for big data

How can photo-z PDFs be stored?

- histogram
- samples
- quantiles

qp is a tool for approximating photo-z PDFs.



Malz&Marshall+1806.00014

qp metrics

qp metrics: moment percent errors of $\{p_i(z)\}$

moment percent error

$$\delta_i(m) \equiv 100 \times \frac{\left| \int z^m \hat{p}_i(z)\, dz - \int z^m p_i(z)\, dz \right|}{\int z^m p_i(z)\, dz}$$

$m = 1$ mean

$m = 2$ variance

$m = 3$ skewness

qp metrics: $\hat{n}(z)$ as a science metric

the stacked estimator of the redshift distribution

$$\hat{n}(z) \equiv \frac{1}{N} \sum_{i=1}^{N} p_i(z)$$

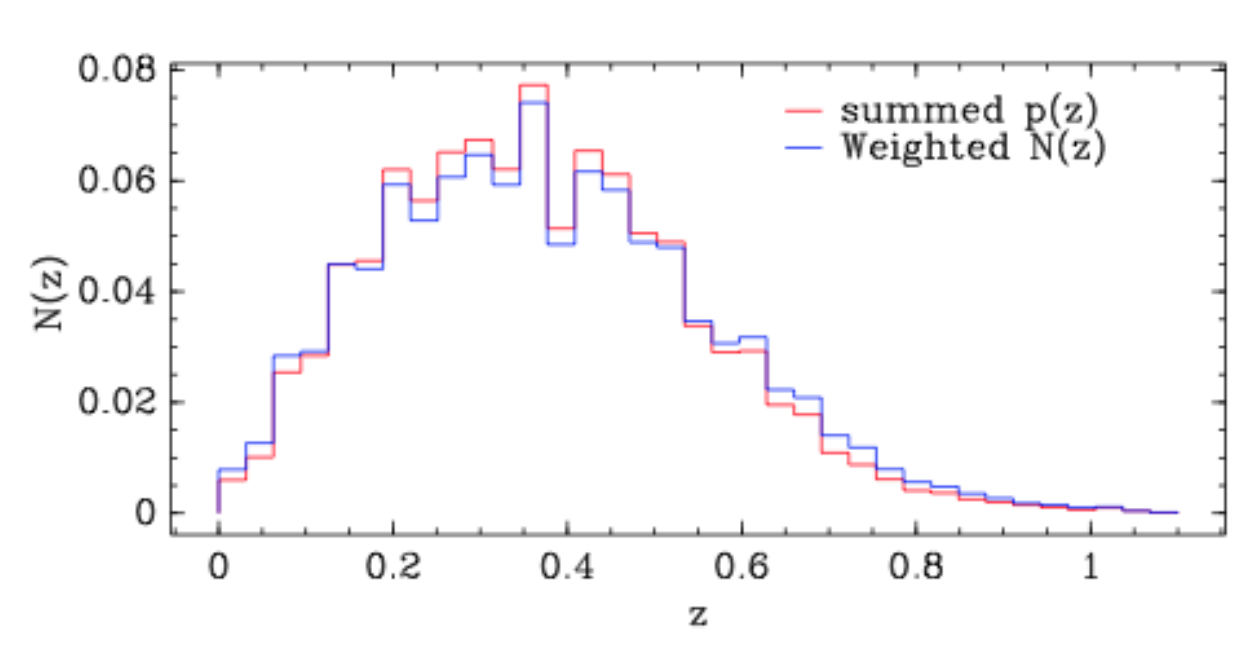


image source: Sheldon+12

qp metrics: KLD of $\hat{n}(z)$

Kullback-Leibler Divergence

$$\mathrm{KLD}[p;\hat{p}] \;=\; \int_{-\infty}^{\infty} \; p(z) \;\; \ln\left[\frac{p(z)}{\hat{p}(z)}\right] \;\; dz$$

qp metrics: KLD of $\hat{n}(z)$

Kullback-Leibler Divergence

$$\mathrm{KLD}[p;\hat{p}] \;=\; \int_{-\infty}^{\infty} \; p(z) \;\; \ln\left[\frac{p(z)}{\hat{p}(z)}\right] \;\; dz$$

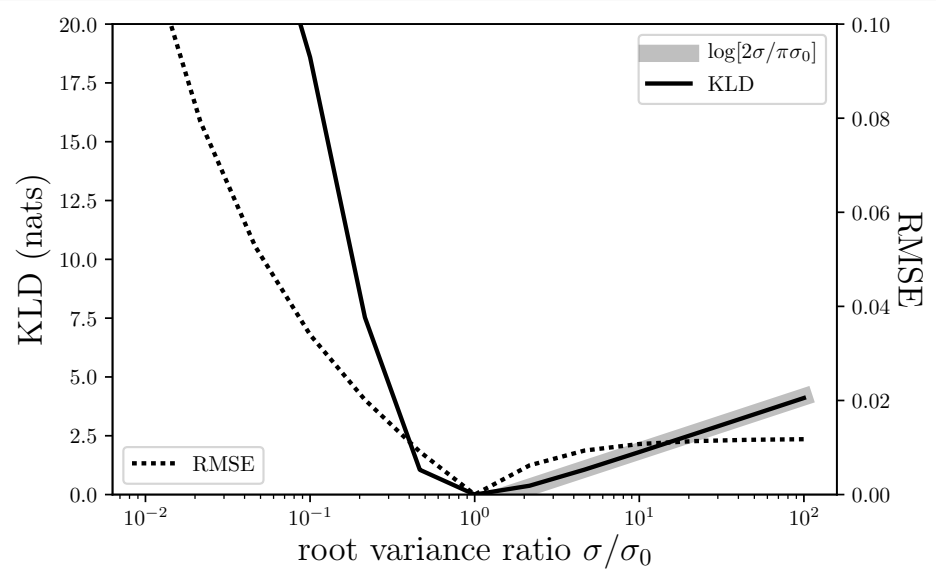
Gaussian example:

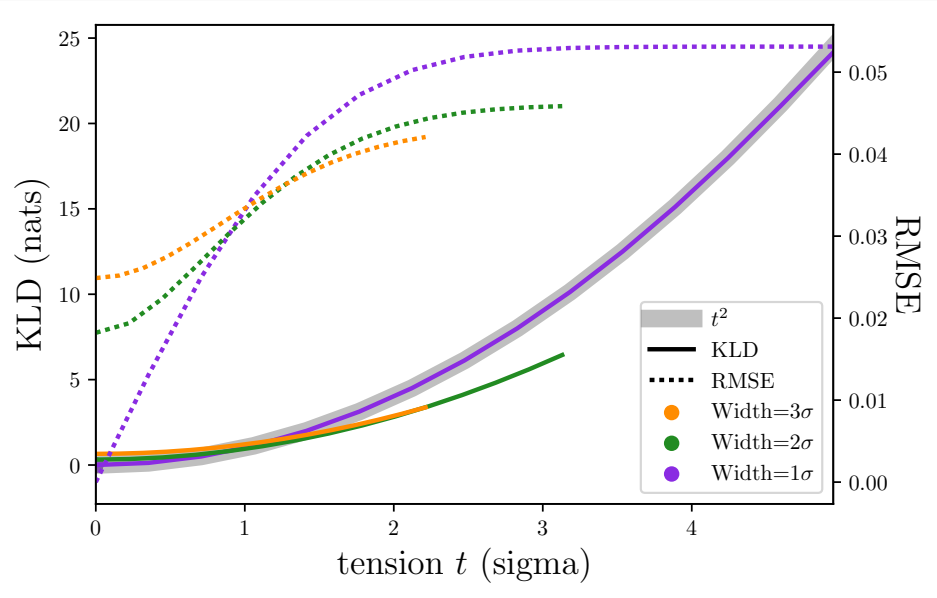
qp metrics: KLD of $\hat{n}(z)$

Kullback-Leibler Divergence

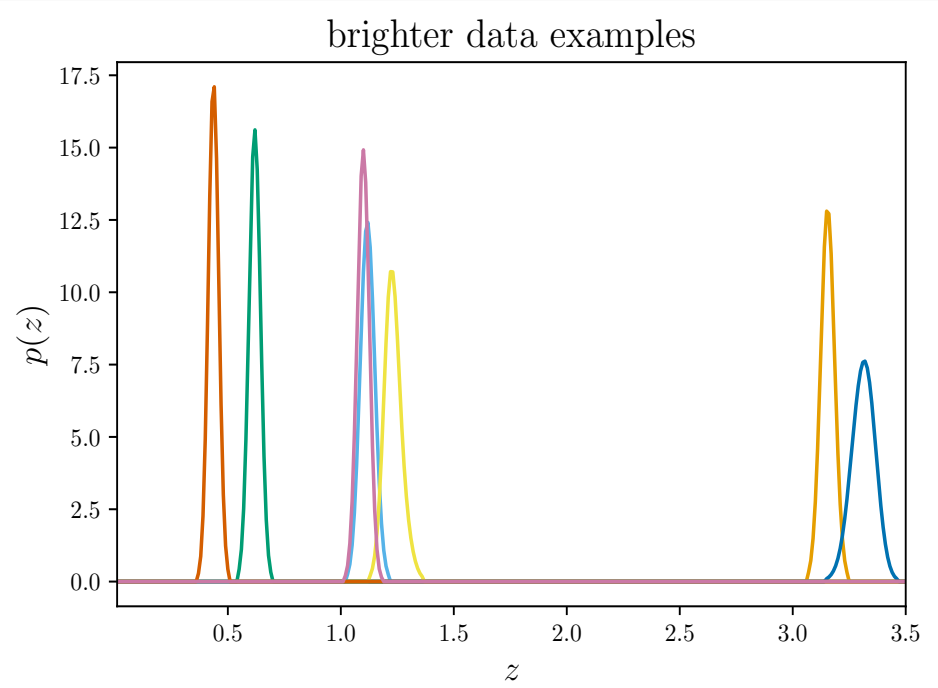
$$\mathrm{KLD}[p;\hat{p}] \;=\; \int_{-\infty}^{\infty} p(z)\;\ln\left[\frac{p(z)}{\hat{p}(z)}\right]\;dz$$

Gaussian example:



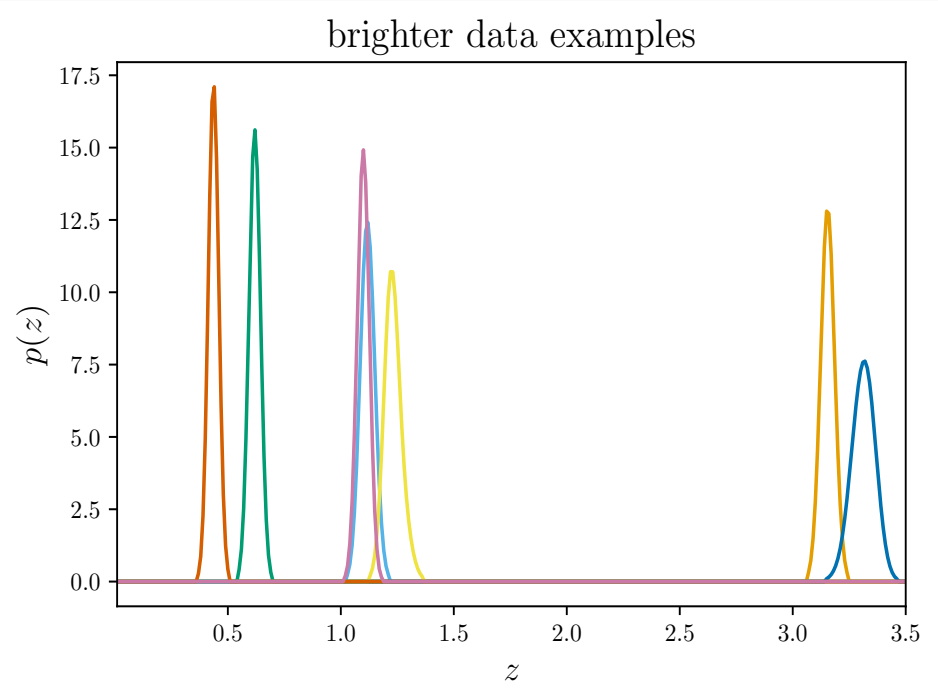


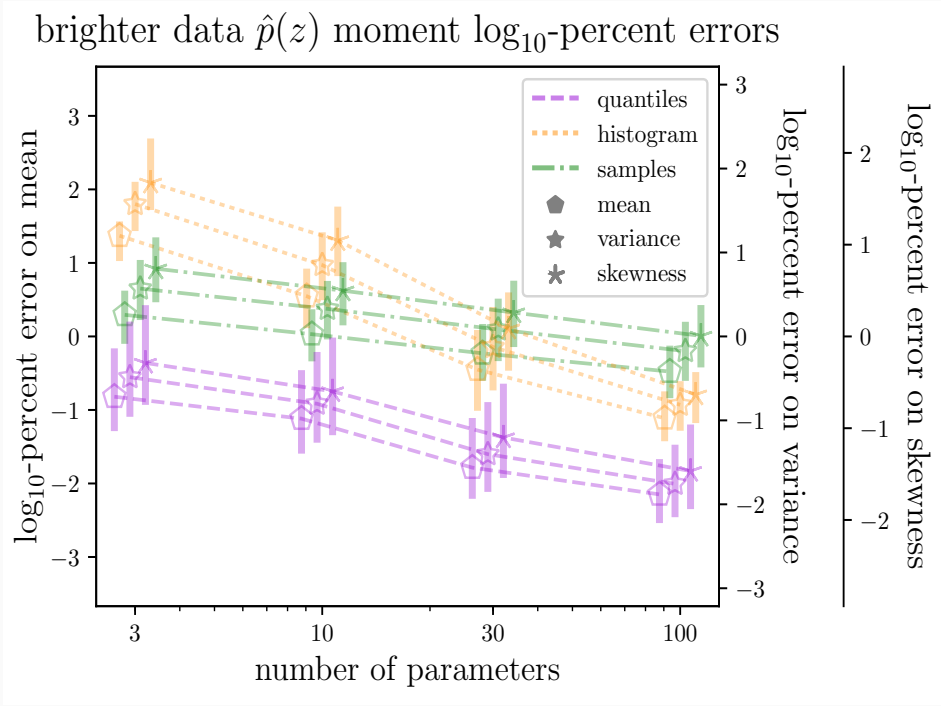
LSST “platinum sample”-like mock photometry



Malz&Marshall+1806.00014

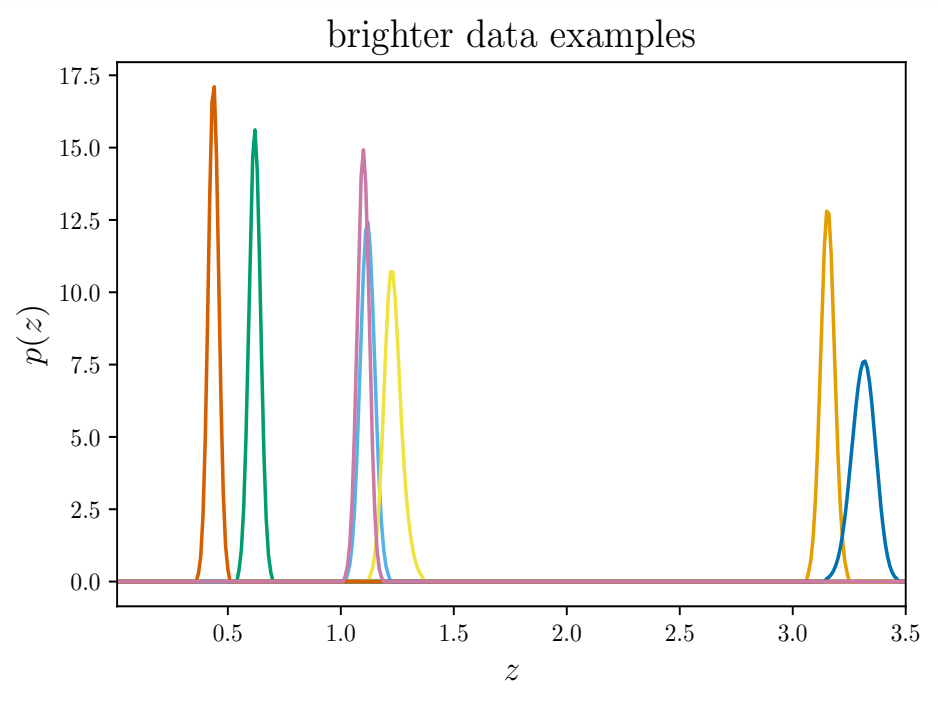
LSST “platinum sample”-like mock photometry

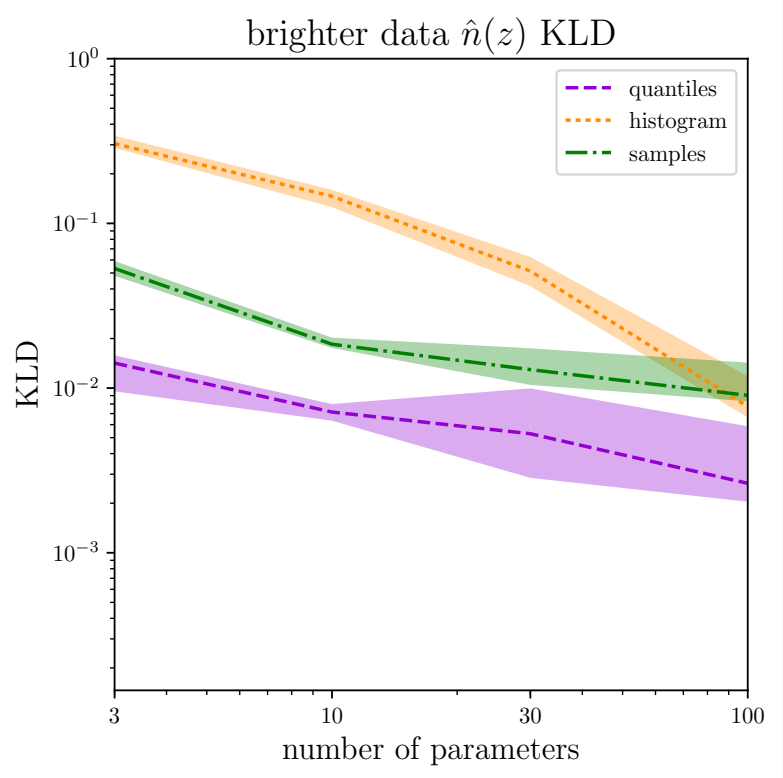




Malz&Marshall+1806.00014

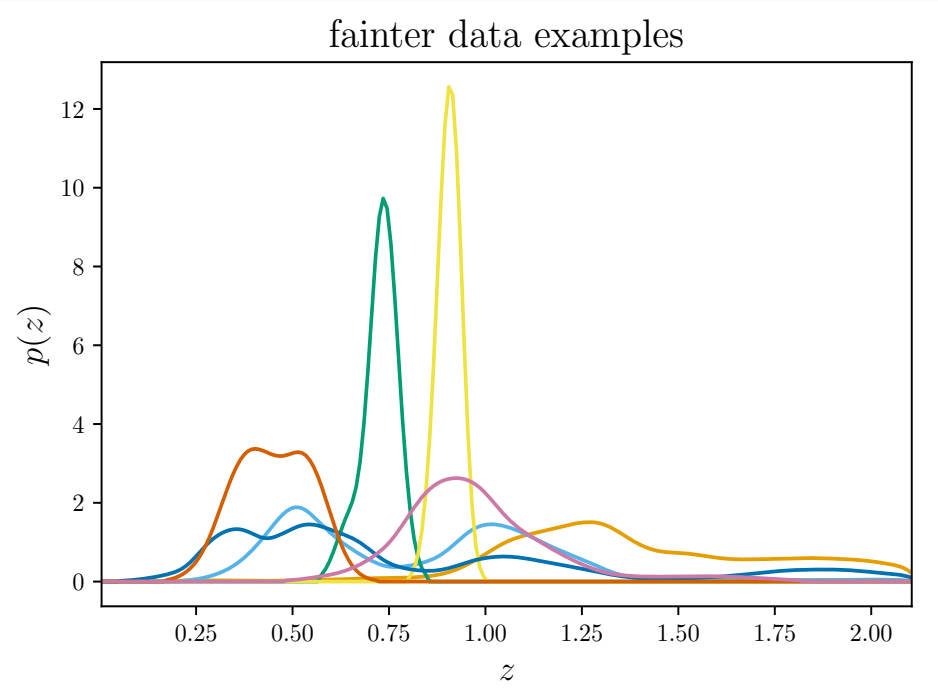
LSST "platinum sample"-like mock photometry





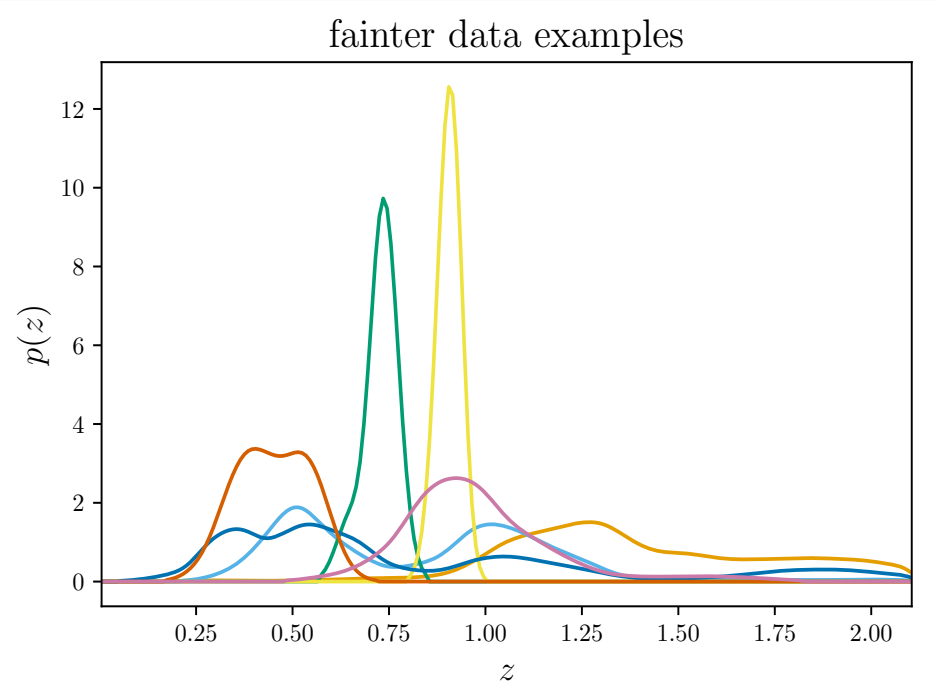
Malz&Marshall+1806.00014

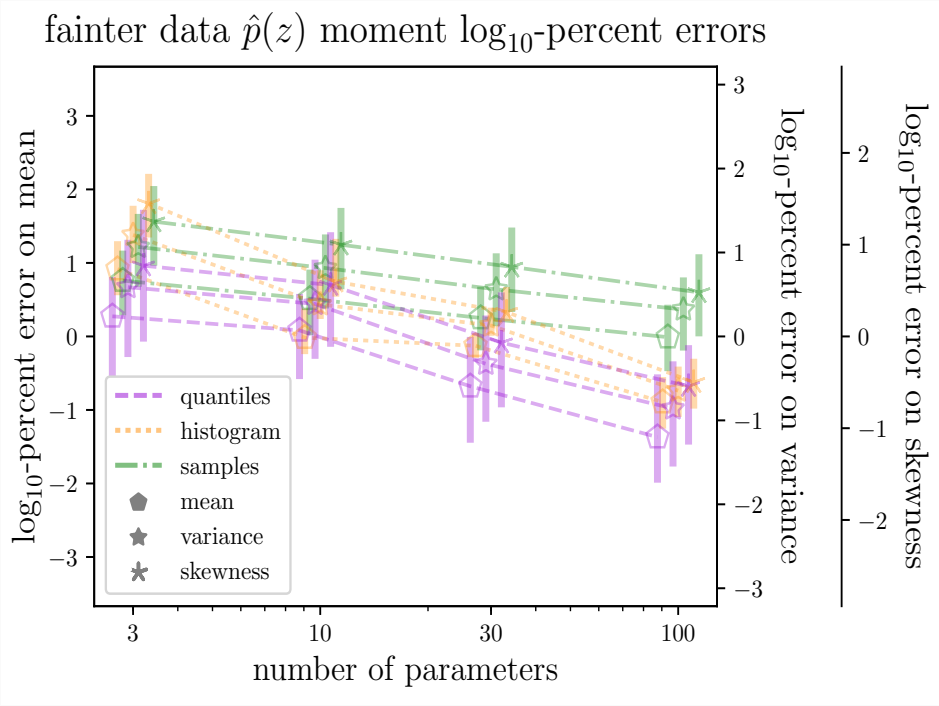
LSST “gold sample”-like mock photometry



Malz&Marshall+1806.00014

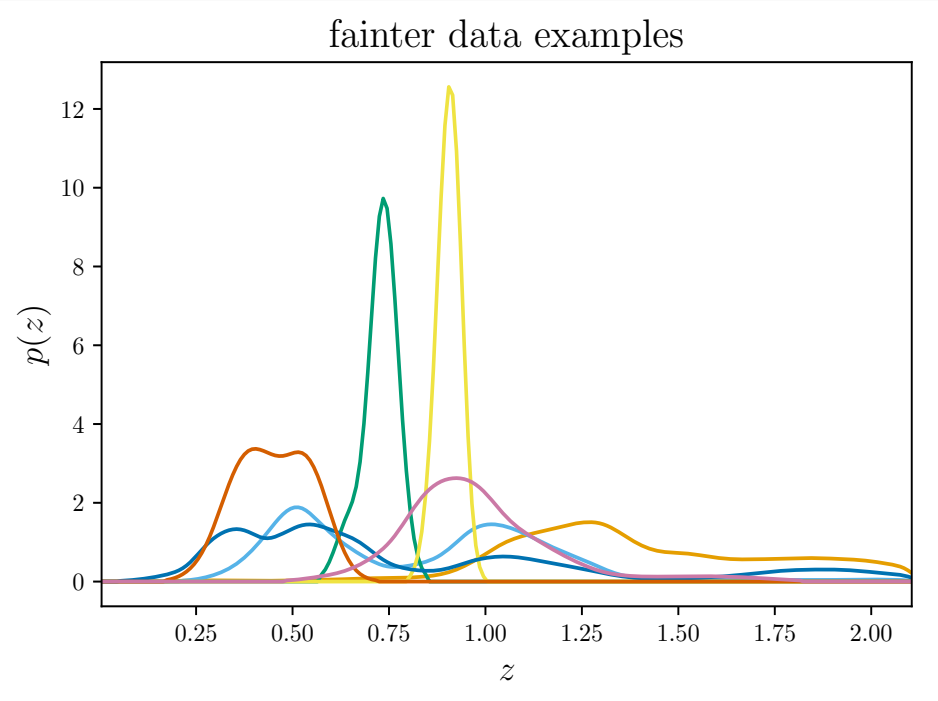
LSST "gold sample"-like mock photometry

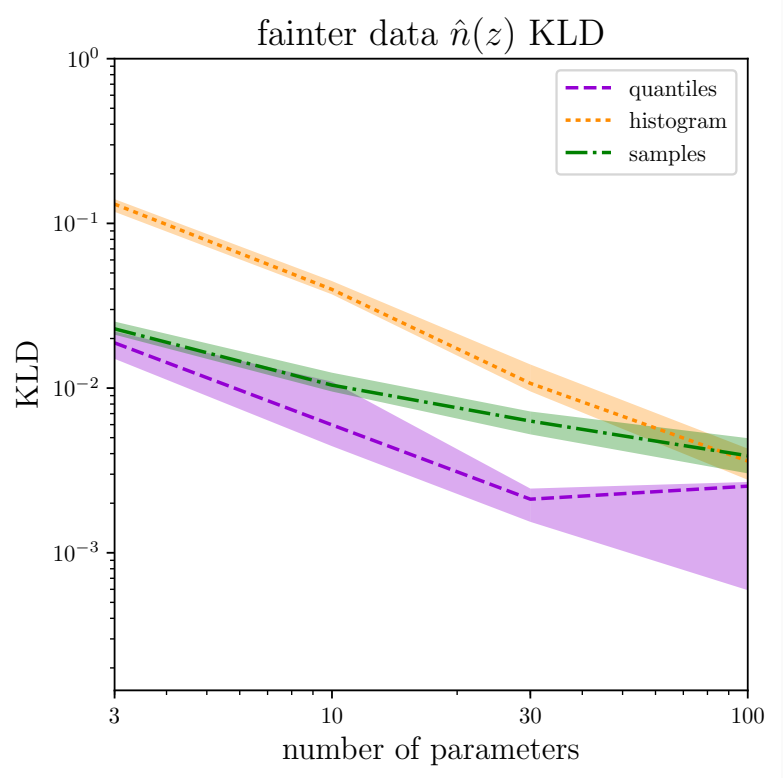




Malz&Marshall+1806.00014

LSST "gold sample"-like mock photometry





Malz&Marshall+1806.00014

qp summary

Significance

Surveys can only store limited photo-z PDF information.

Storage parametrization impacts photo-z PDF accuracy.

qp enables optimizing the photo-z PDF storage parametrization.

Findings

Science applications must motivate parameterization metrics.

Samples and quantiles can outperform the histogram format.

(Samples and quantile PDF reconstruction may yet improve.)

Get qp from `github.com/aimalz/qp`.

Overview

Cosmology with large photometric galaxy surveys

Stress-testing probabilistic redshift estimates

Practicalities of probabilistic data

Hierarchical inference of the cosmological parameters

Summary and future directions

How not to use photo-z PDFs: stacking

How not to use photo-z PDFs: stacking

the stacked estimator of the redshift distribution

$$\hat{n}(z) \equiv \frac{1}{N} \sum_{i=1}^{N} p_i(z)$$

How not to use photo-z PDFs: stacking

the stacked estimator of the redshift distribution

$$\hat{n}(z) \equiv \frac{1}{N} \sum_{i=1}^{N} p_i(z)$$

and statistical independence

$$p(A \cap B) = p(A) \times p(B) \neq p(A) + p(B)$$

are incompatible!

What is a photo-z PDF?

A photo-z PDF is a

posterior.

$$\begin{aligned} \text{“}p(z)\text{”} &\equiv p(z \mid \mathrm{data}) \\ &\equiv p(z \mid \vec{d}) \end{aligned}$$

What is a photo-z PDF?

A photo-z PDF is an

implicit posterior.

$$\begin{aligned} \text{“}p(z)\text{”} &\equiv p(z \mid \text{data},\ \text{assumptions}) \\ &\equiv p(z \mid \vec{d},\ \vec{\phi}^{*}) \end{aligned}$$

Why does it matter?

the stacked estimator of the redshift distribution

$$\hat{n}(z) \equiv \frac{1}{N} \sum_{i=1}^{N} p_i(z \mid \vec{d}_i, \vec{\phi}^*)$$

and statistical independence

$$p(A \cap B \mid C) = p(A \mid C) \times p(B \mid C) \neq p(A \mid C) + p(B \mid C)$$

are incompatible!

How to use photo-z PDFs: The CHIPPR model

Cosmological

Hierarchical

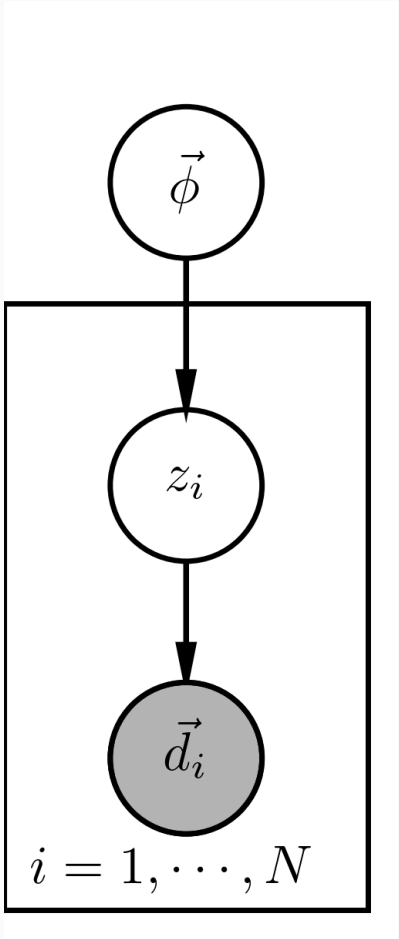
Inference with

Probabilistic

Photometric

Redshifts

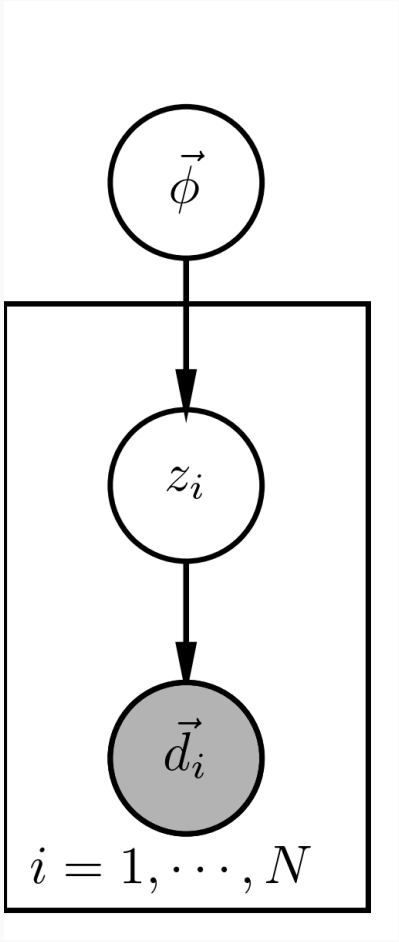
How to use photo-z PDFs: The CHIPPR model



Cosmological
Hierarchical
Inference with
Probabilistic
Photometric
Redshifts

Malz&Hogg(in prep)

How to use photo-z PDFs: The CHIPPR model



Malz&Hogg(in prep)

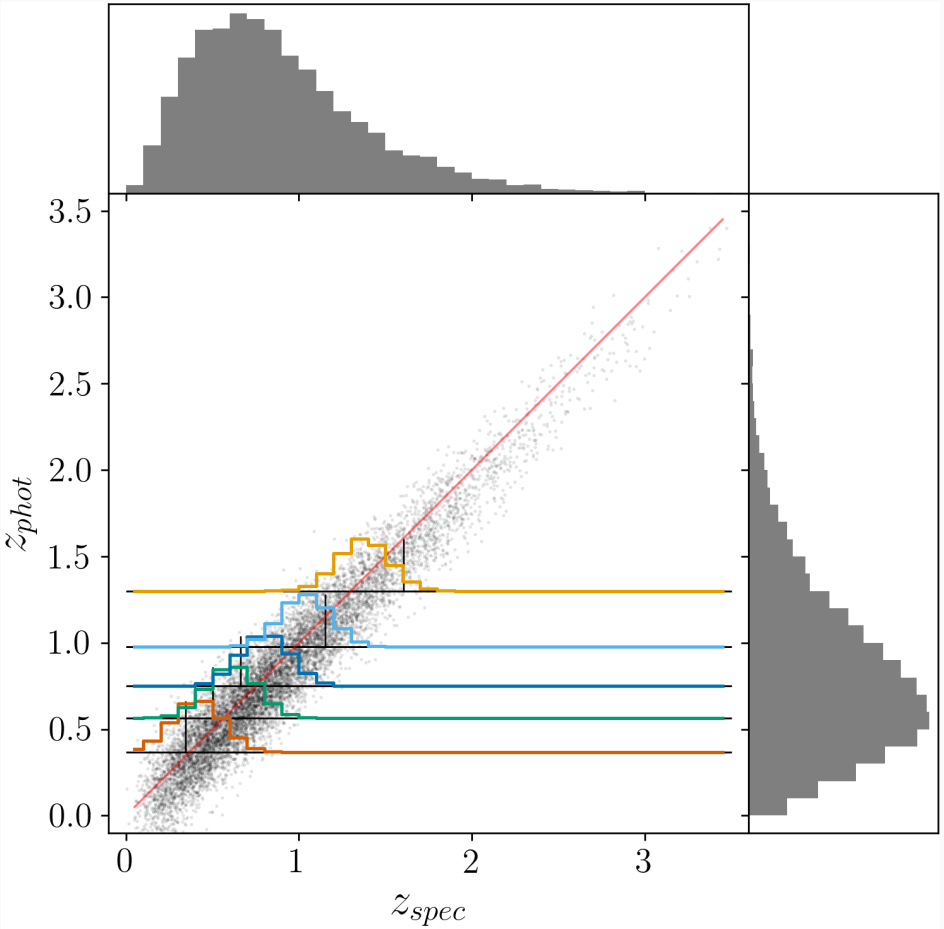
Cosmological
Hierarchical
Inference with
Probabilistic
Photometric
Redshifts

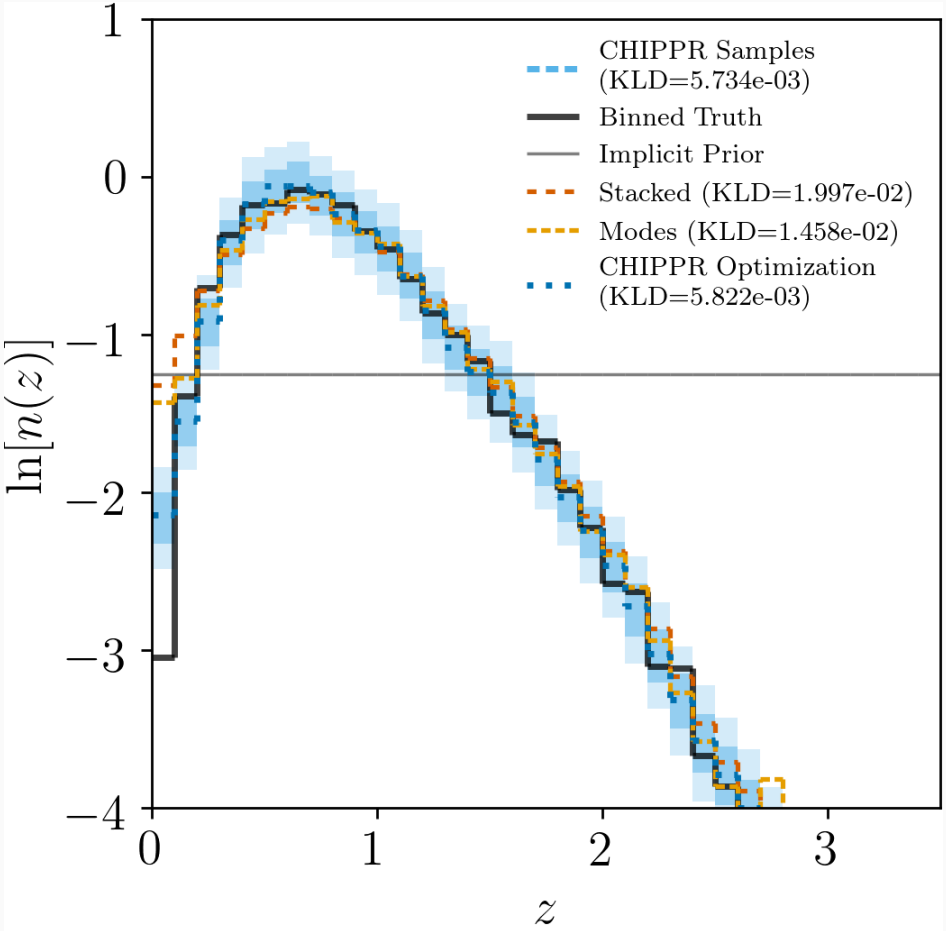
$$p(\vec{\phi} \mid \{\vec{d}_i\}) \propto p(\vec{\phi}) \prod_i^N \int p(z \mid \vec{d}_i, \vec{\phi}^*) \frac{p(z \mid \vec{\phi})}{p(z \mid \vec{\phi}^*)} dz$$

CHIPPR outperforms stacking on realistically complex data.

CHIPPR outperforms stacking on realistically complex data.

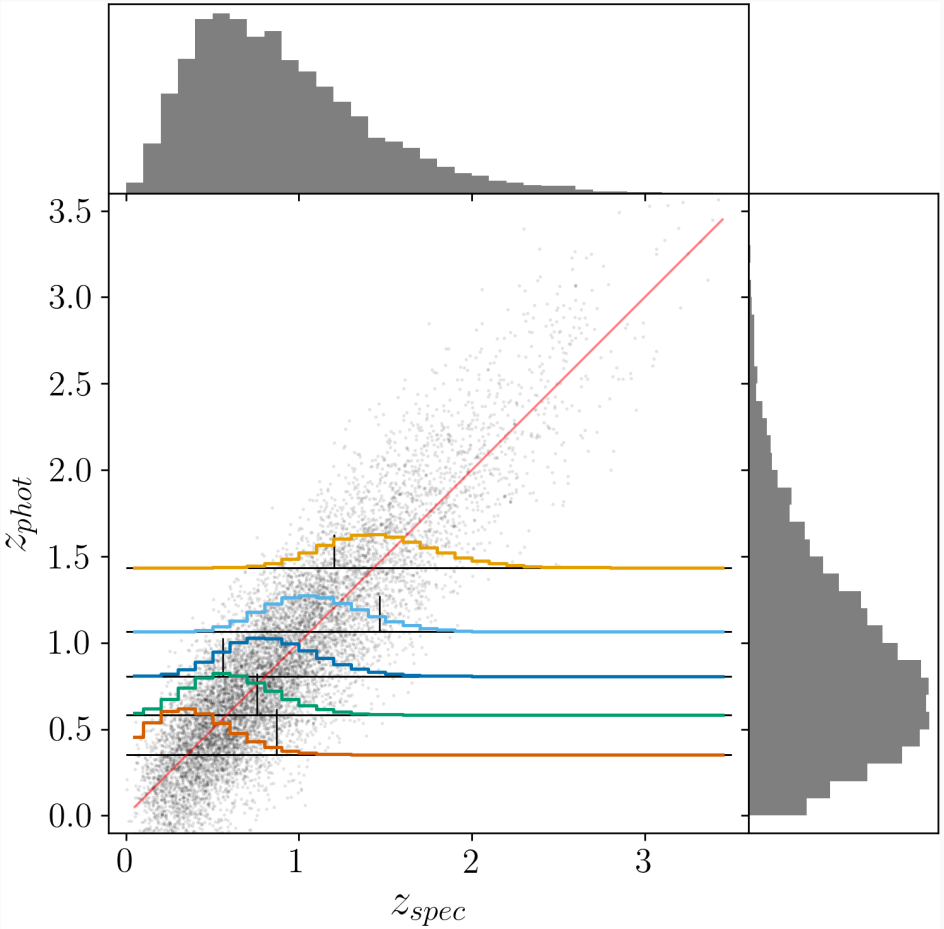
bias $\Delta z = 0.03$ ($10\times$ LSST requirements)

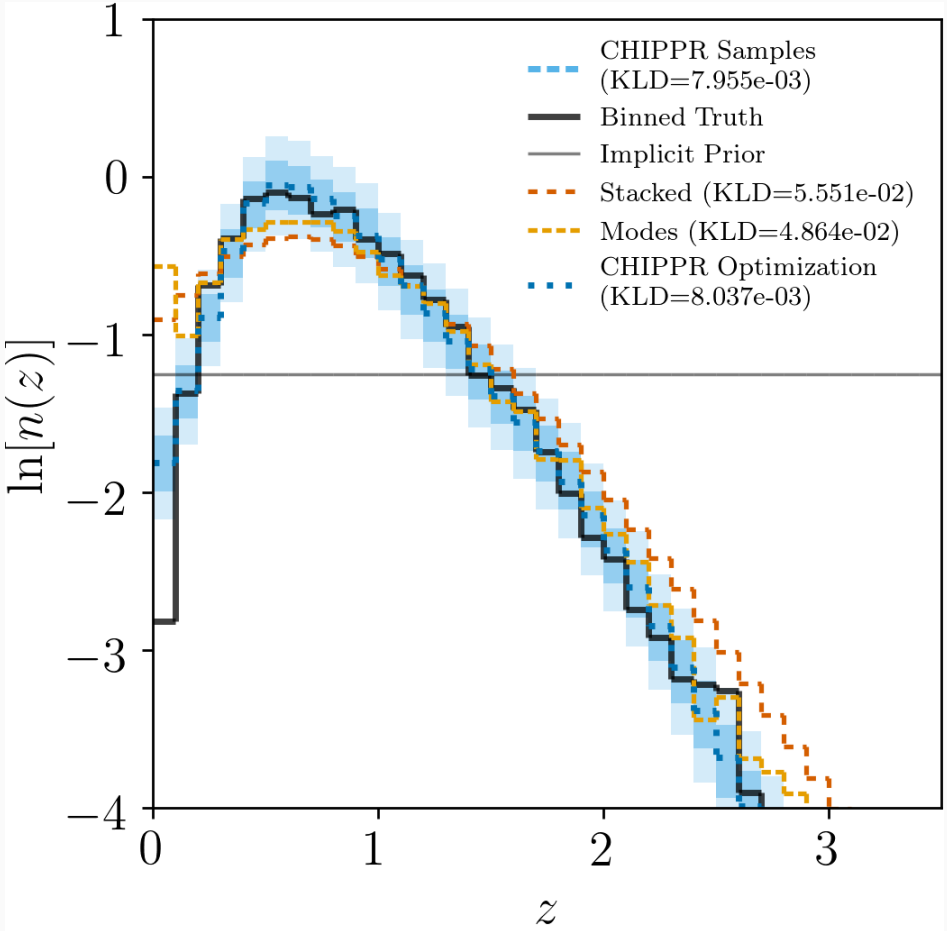




CHIPPR outperforms stacking on realistically complex data.

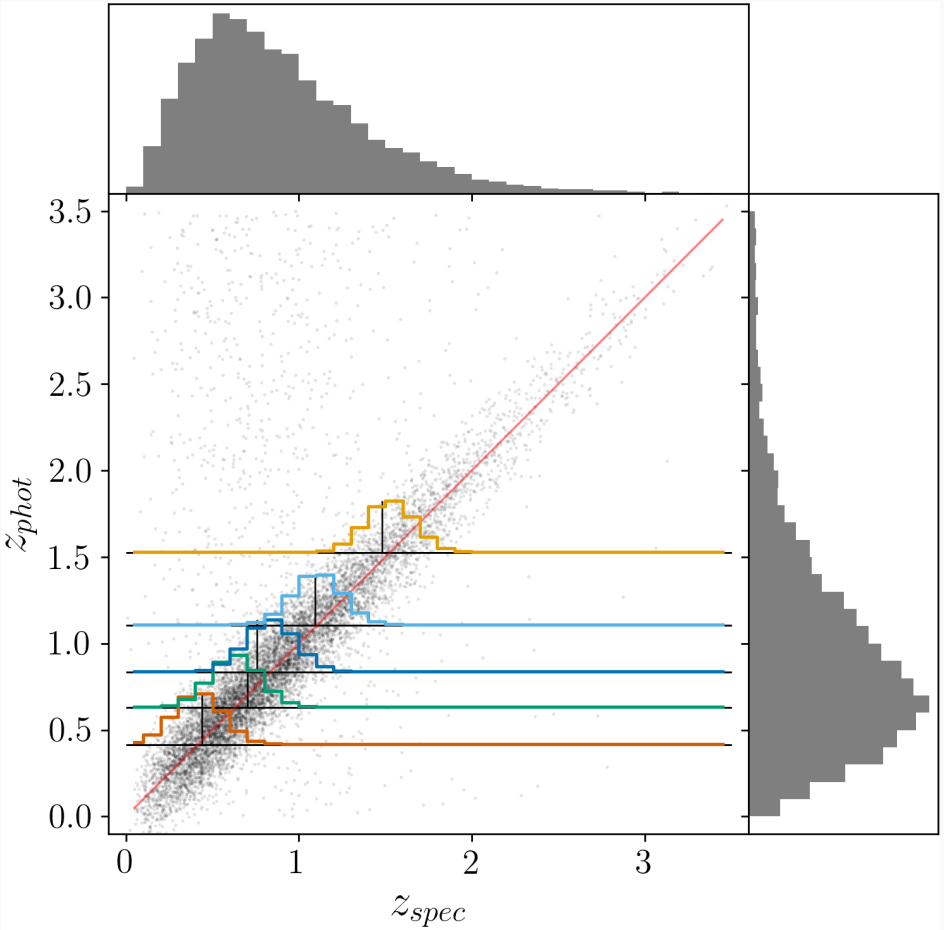
4% intrinsic scatter (2× LSST requirements)

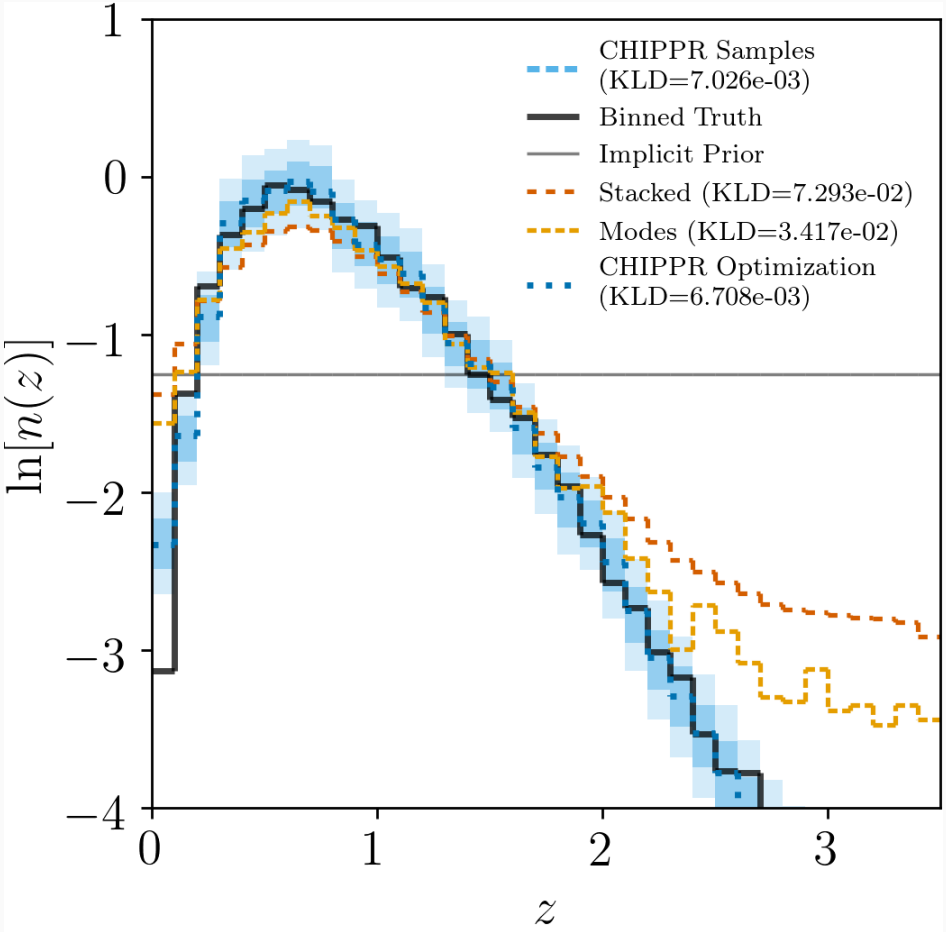




CHIPPR outperforms stacking on realistically complex data.

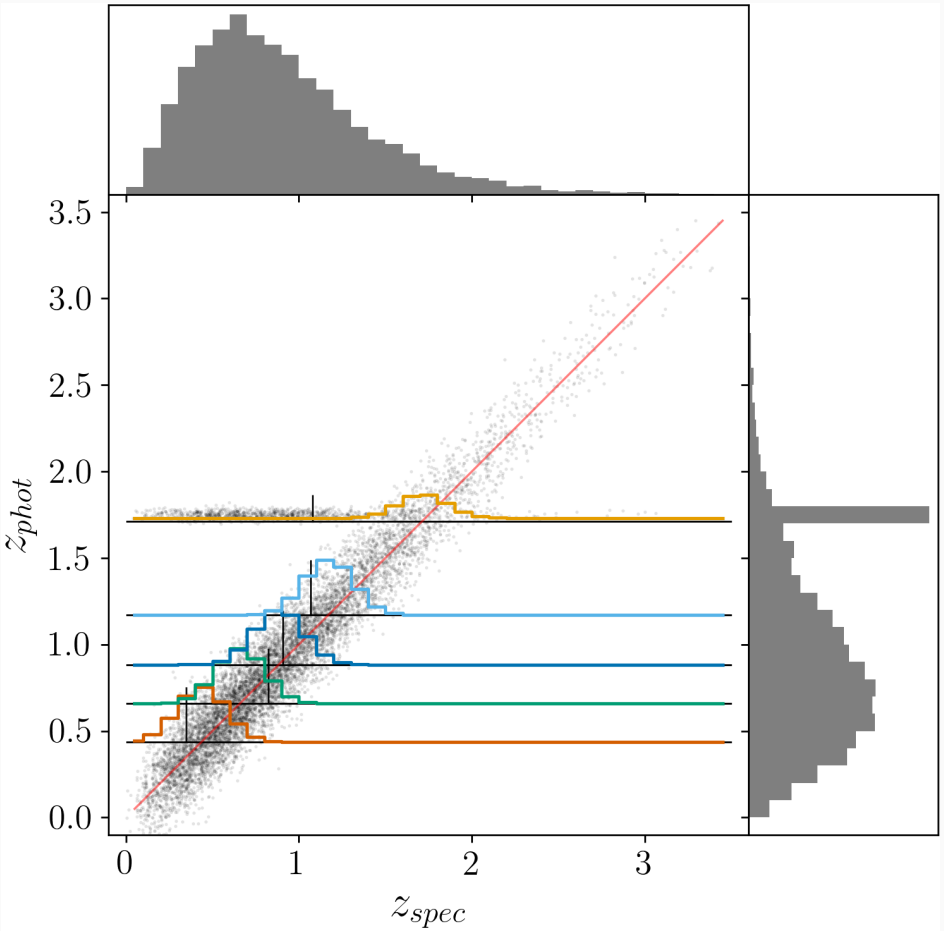
10% uniformly distributed catastrophic outliers

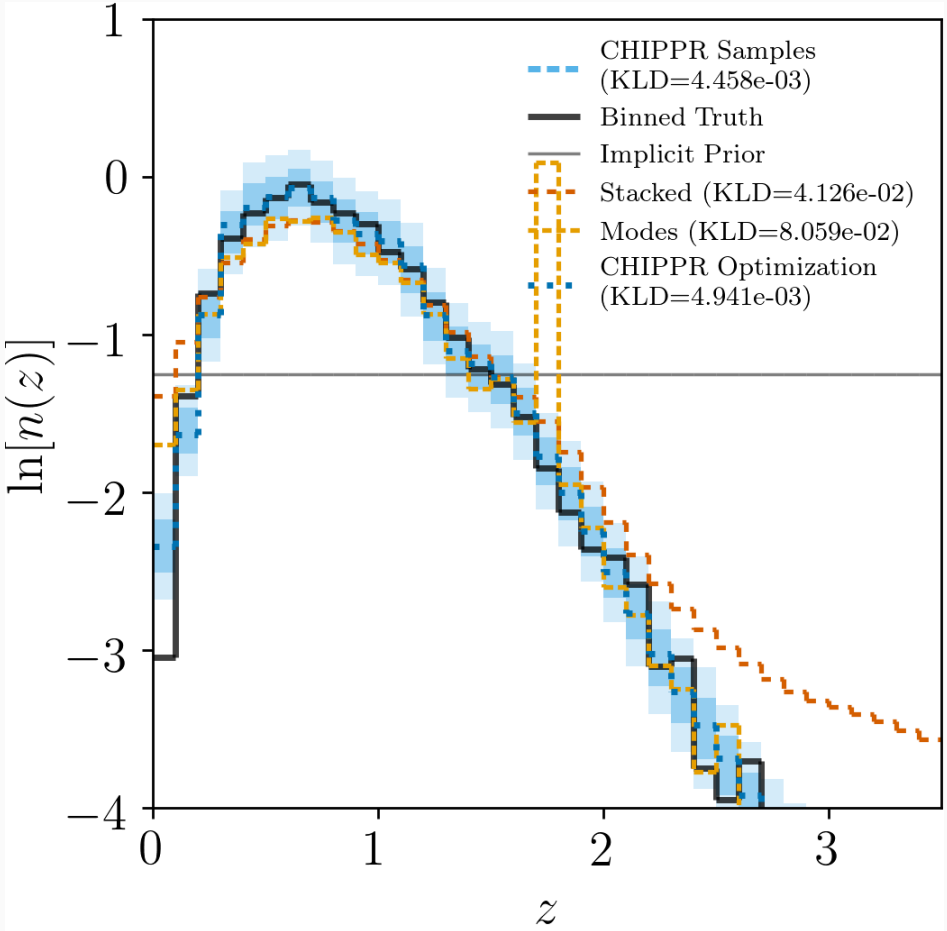




CHIPPR outperforms stacking on realistically complex data.

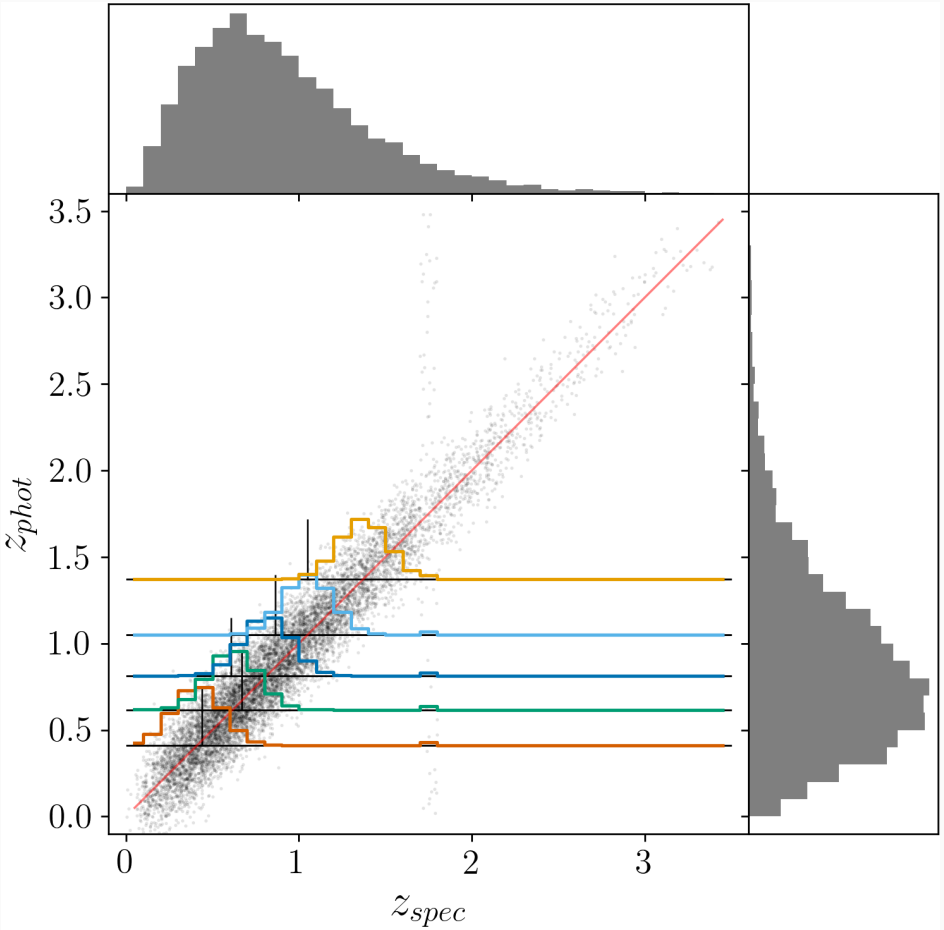
$\lesssim 10\%$ catastrophic outliers a la model-fitting codes

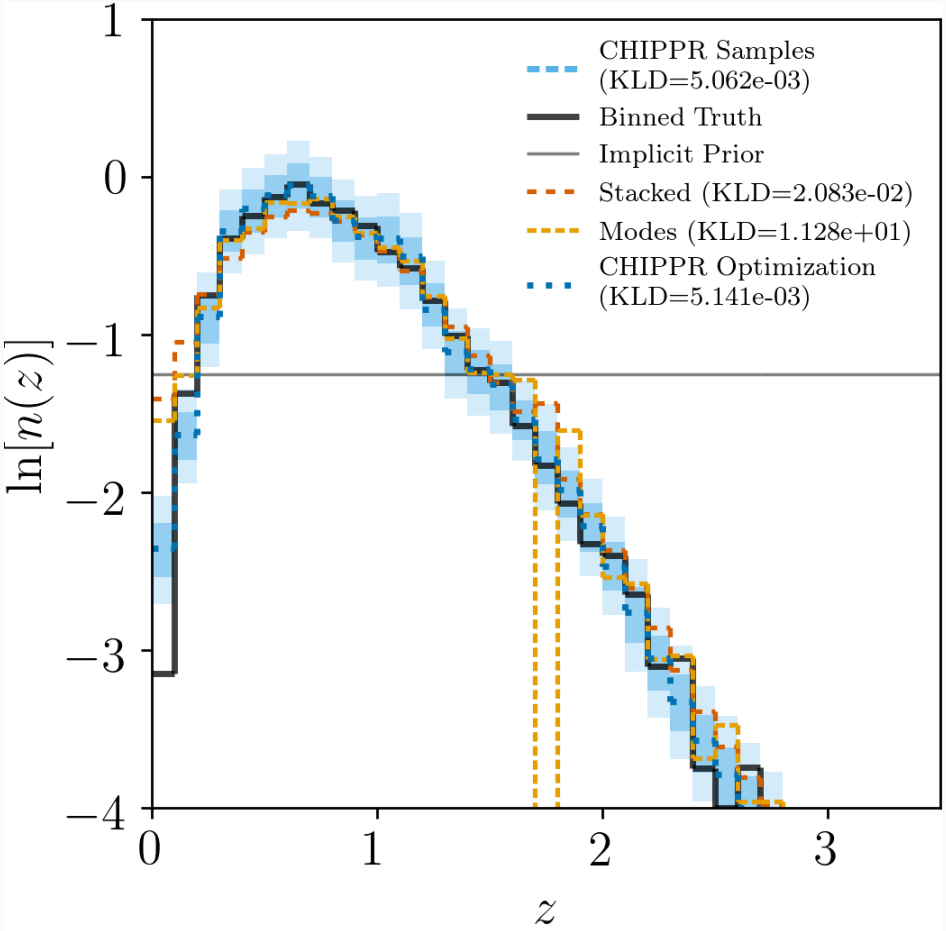




CHIPPR outperforms stacking on realistically complex data.

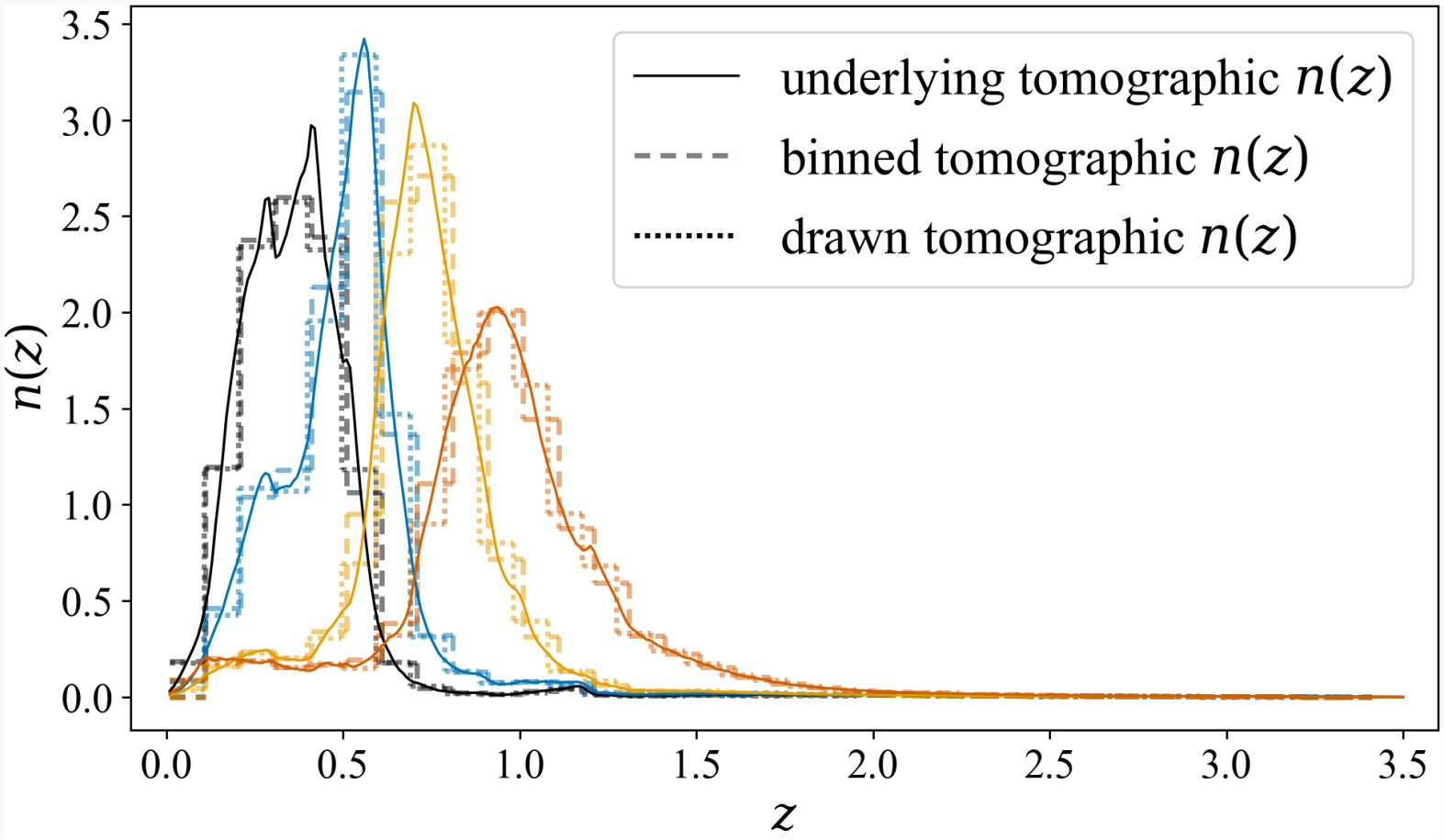
$\lesssim 10\%$ catastrophic outliers a la data-driven codes





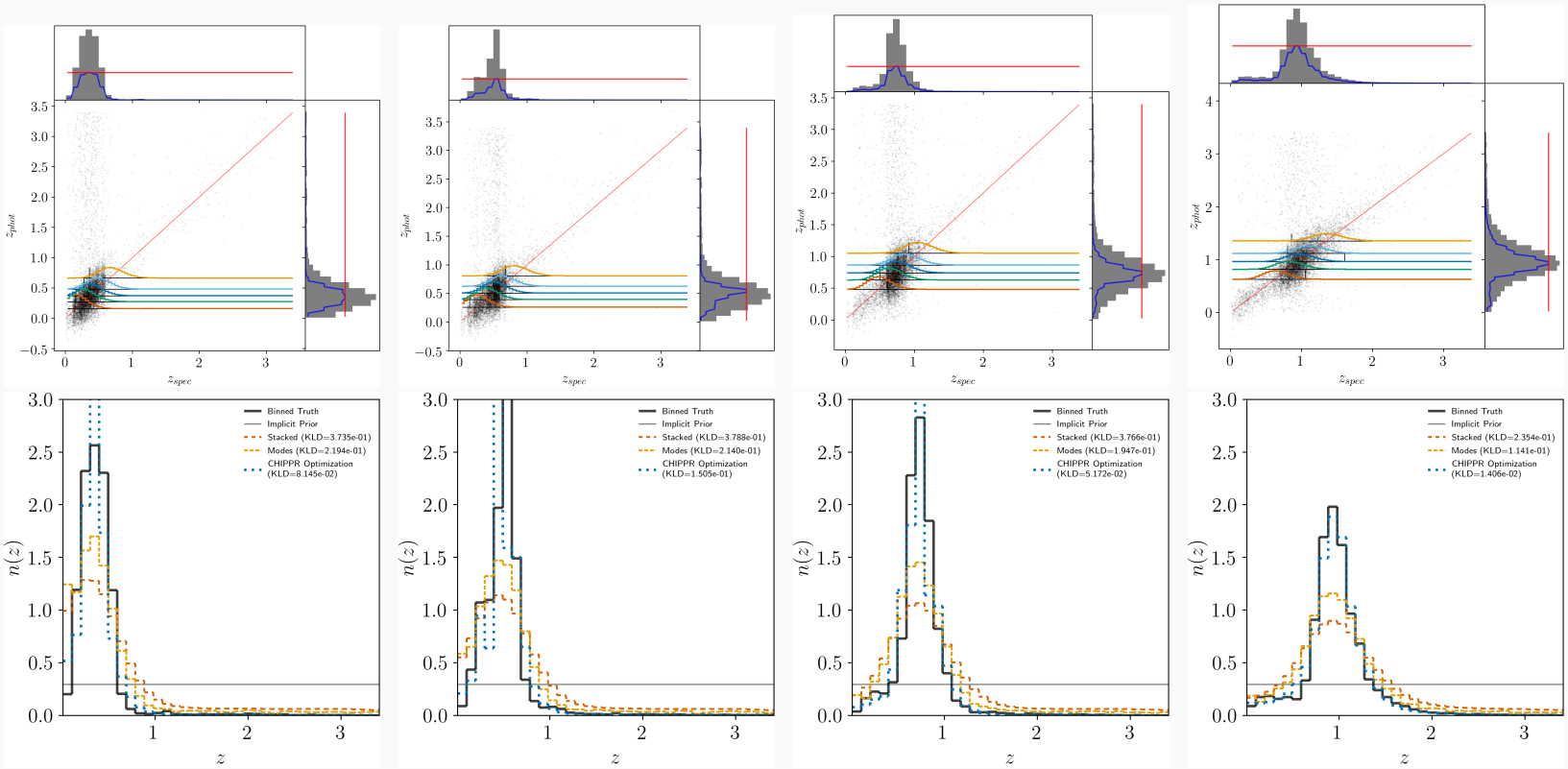
CHIPPR outperforms stacking on LSST-like data.

CHIPPR outperforms stacking on LSST-like data.



Malz&Hogg (in prep)

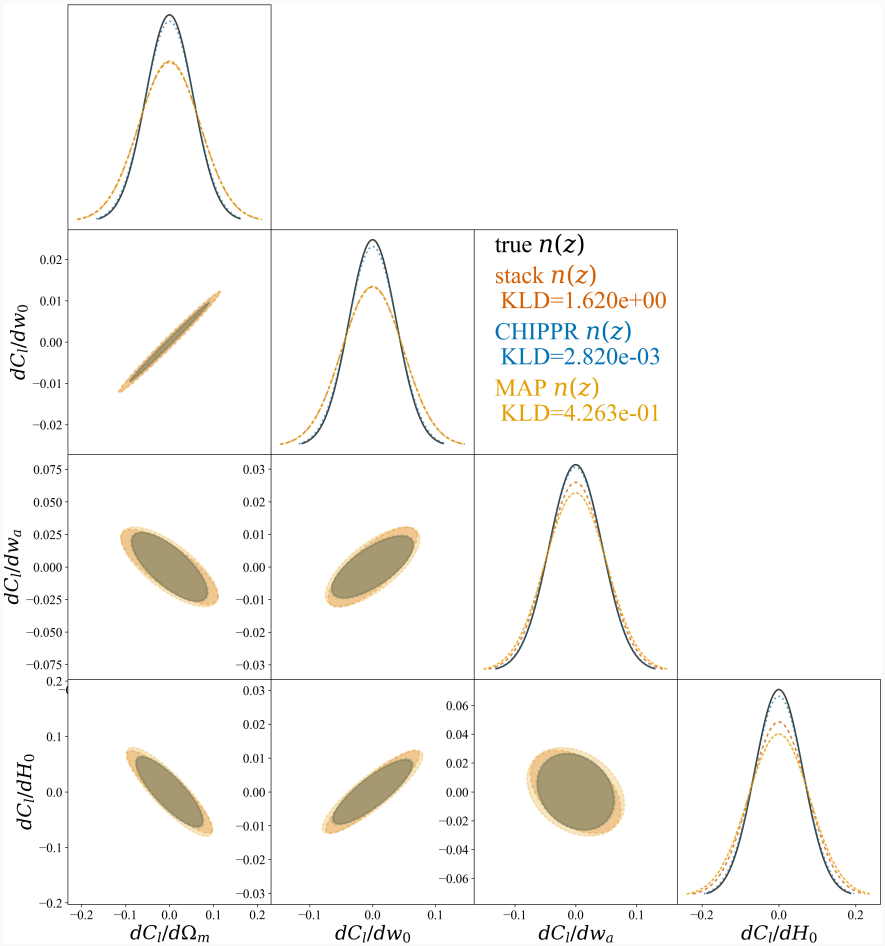
CHIPPR outperforms stacking on LSST-like data.



Malz&Hogg (in prep)

CHIPPR outperforms stacking on LSST-like data.

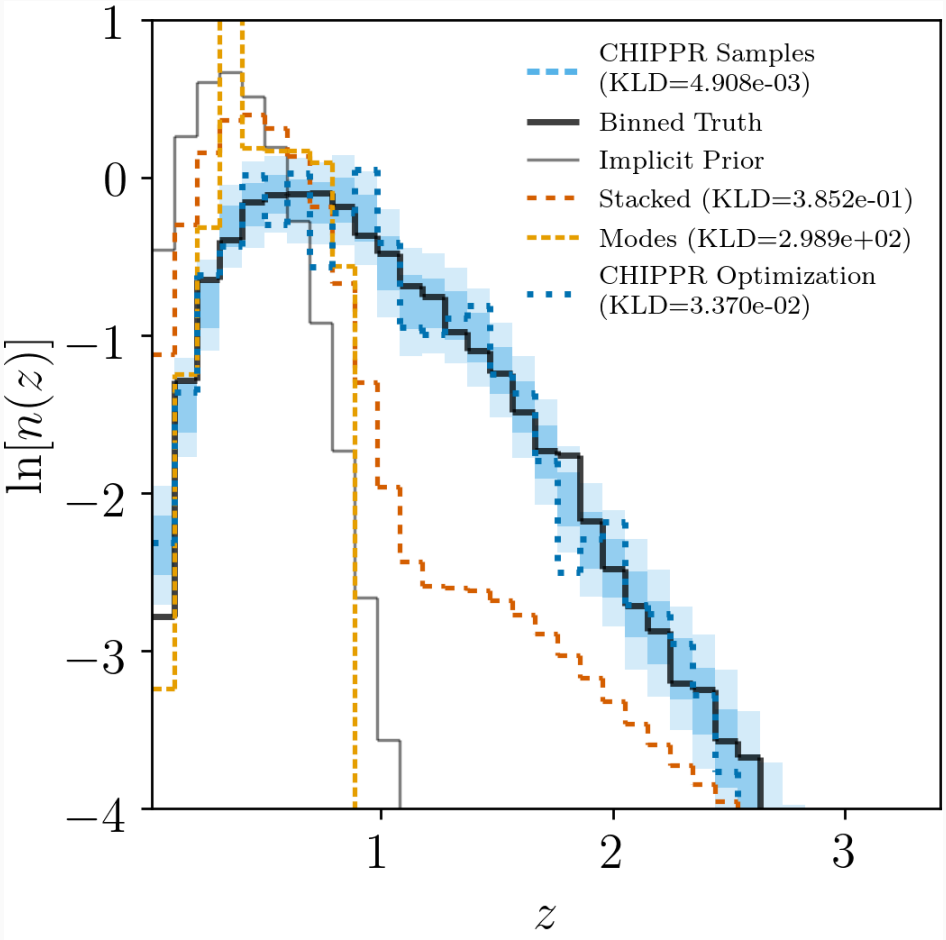
CHIPPR outperforms stacking on LSST-like data.



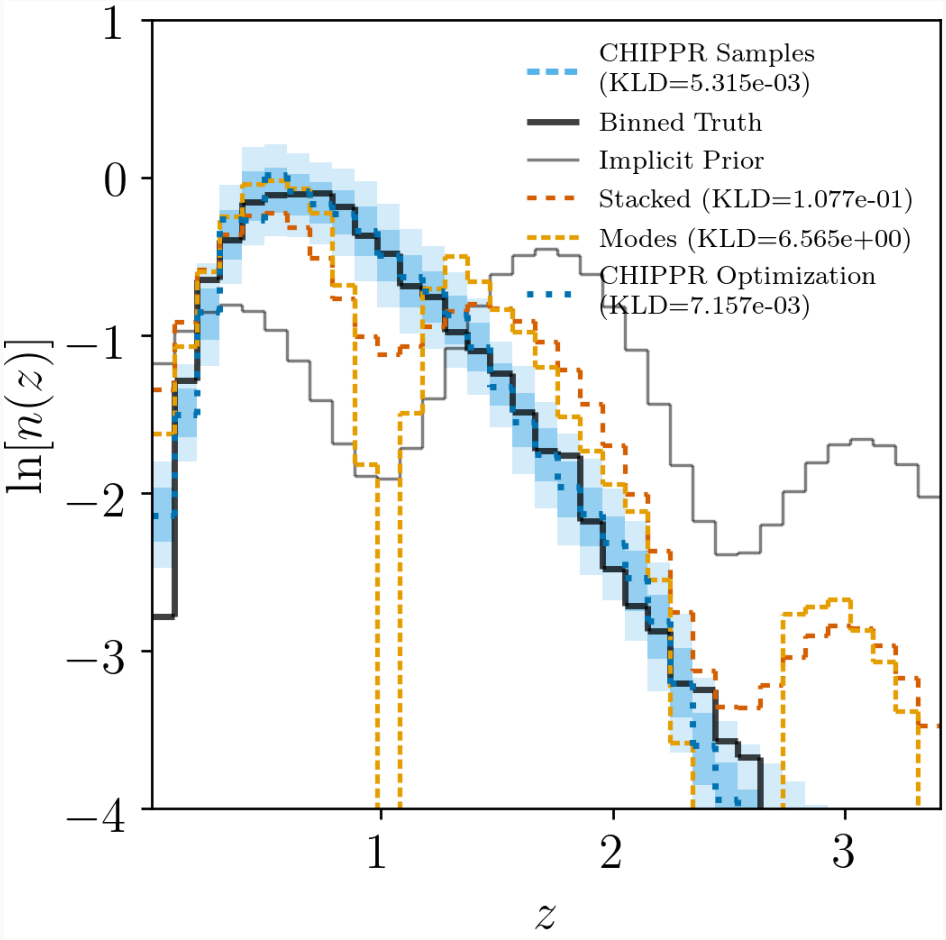
Malz&Hogg (in prep)

CHIPPR is robust to a nontrivial implicit prior.

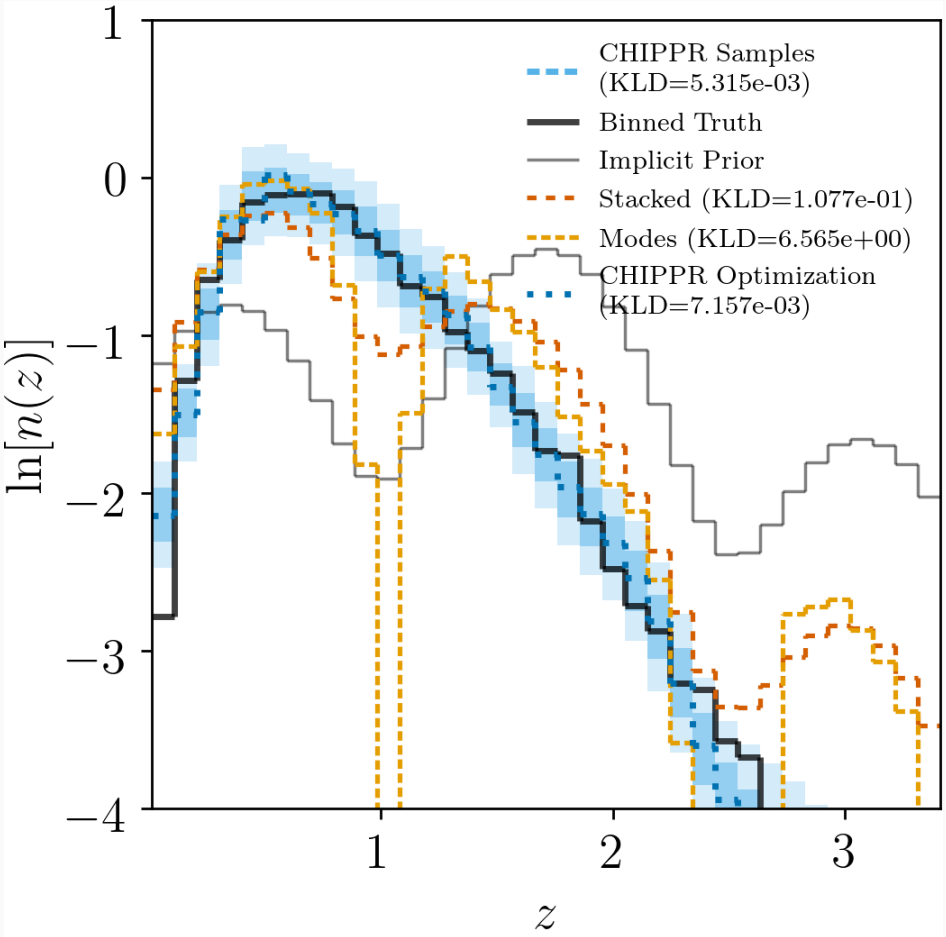
CHIPPR is robust to a nontrivial implicit prior.

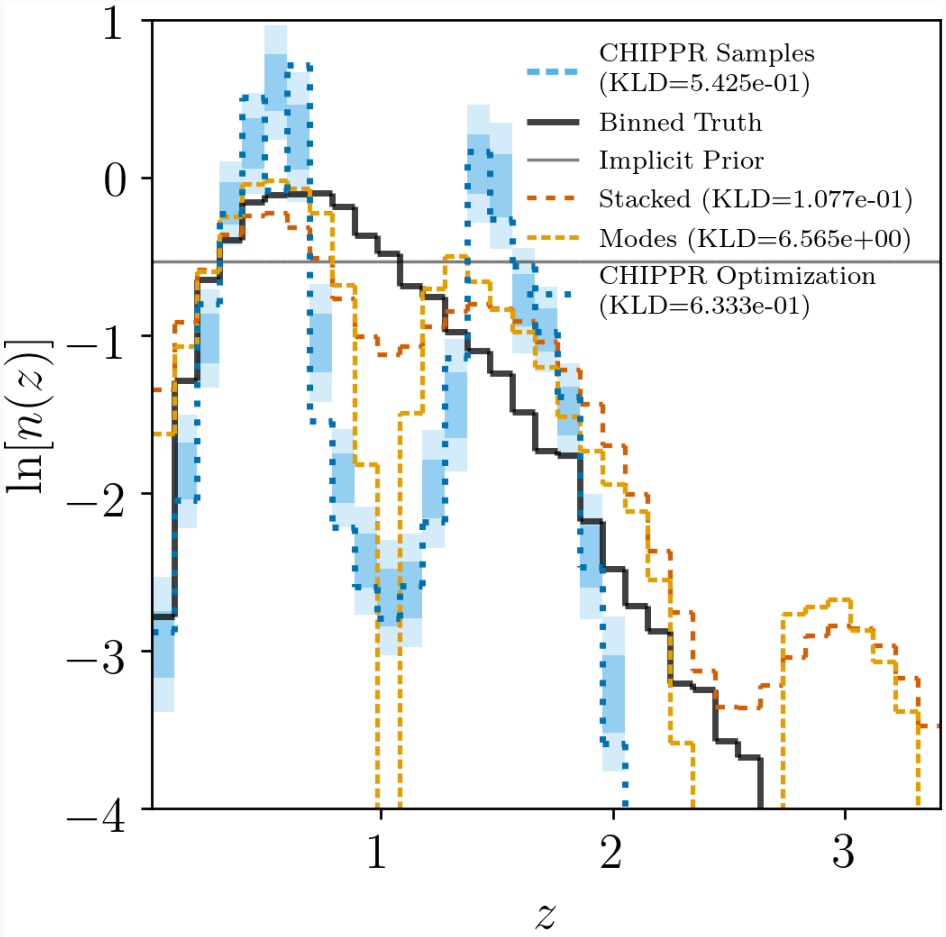


CHIPPR is robust to a nontrivial implicit prior.



Even CHIPPR fails under prior misspecification.





chippr summary

Context

Stacking photo-z PDFs to get $\hat{n}(z)$ is mathematically incorrect.

CHIPPR is a provably valid model for $n(z)$.

chippr samples $n(z)$ given any photo-z PDF catalog.

Findings

Stacking yields a biased estimate of $n(z)$ for LSST-like data.

CHIPPR recovers the true $n(z)$ when the implicit prior is known.

chippr yields the posterior probability distribution of $n(z)$.

Get chippr from github.com/aimalz/chippr.

Overview

Probabilistic data products and the road ahead

"Take a good, long look at this. ... We don't know what it is, but it's the only part of the buffalo we don't use."

Probabilistic data products and the road ahead

Making probabilistic data products

Probabilistic data products best preserve physical information.

Catalog-level decisions can cause science-level biases.

Experimental design determines the questions we can answer.

Using probabilistic data products for science

A probabilistic data product mandates a probabilistic approach.

Implicit assumptions impact our science results.

Principled science metrics must guide decisionmaking.

Probabilistic data products can enhance our science, but only if we are careful to not misuse them.

We can learn cosmology with uncertainty-dominated data!

A probabilistic data product necessitates a probabilistic approach.

Catalog-level decisions propagate to science-level biases.

The choice of performance metric must be principled.

Valid science hinges on documenting implicit assumptions.

Thank you to David Hogg & Phil Marshall, LSST-DESC,
my thesis committee, and all the friends who have supported me!