Deblending galaxies with variational autoencoders

A joint multi-band, multi-instrument approach

Bastien Arcelin, Cyrille Doux, E. Aubourg and C. Roucelle

LSST Dark Energy Science Collaboration

arXiv:2005.12039, accepted in MNRAS

Cyrille Doux

University *of* Pennsylvania

Department of Physics and Astronomy
Warren Center for Data and Network Sciences

LIneA webinar

October 15th 2020

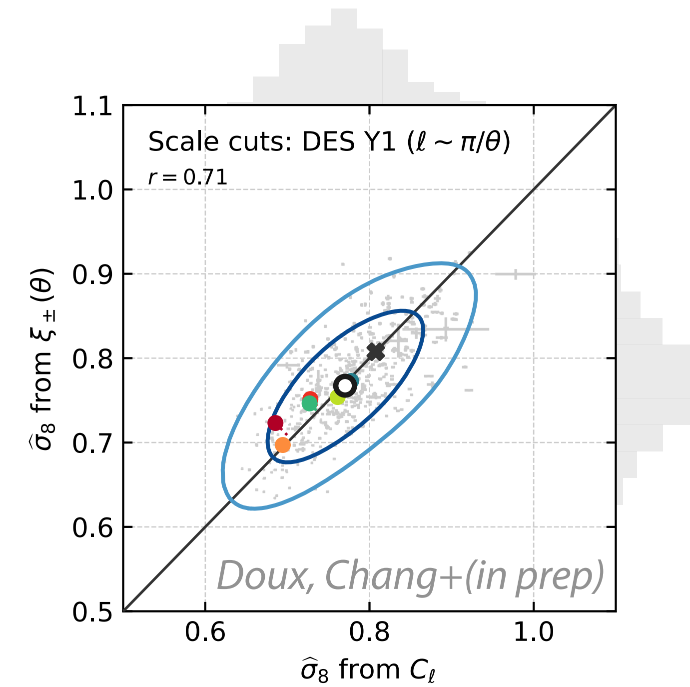
Penn
University *of* Pennsylvania

Other stuff I (try to) do

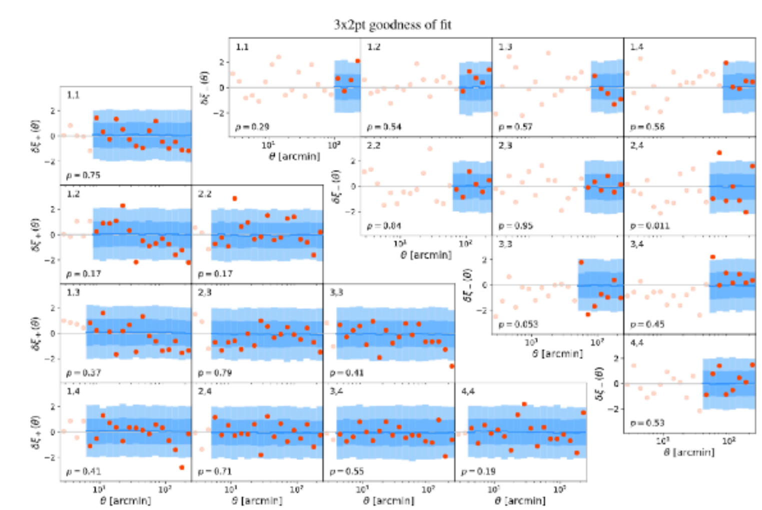
- **Research objective**: Studying dark energy through weak lensing and its combinations with other probes (incl. CMB) in a an era of multiple large astronomical surveys

- **Dark Energy Survey Year 3 analysis**
  - Weak lensing + clustering anaysis
    - Consistency tests with PPD (w/ E. Baxter, paper out soon)
  - Cosmic shear in harmonic space
    - Cosmic shear analysis + tests (B-mode, PSF)
    - Consistency with real space (w/ C. Chang, paper out soon)
  - DES x CMB
    - 6x2 analysis in ΛCDM/wCDM + extensions (eg $\sigma_8(z)$)
    - DES galaxy x ACT SZ-y for pressure profile
- **WFIRST**
  - Forecasts for 6x2 with SO on $w_0 w_a$CDM and extensions
- **Biology** (for fun)
  - Analysis of Morris Water Maze data for neuroscience.

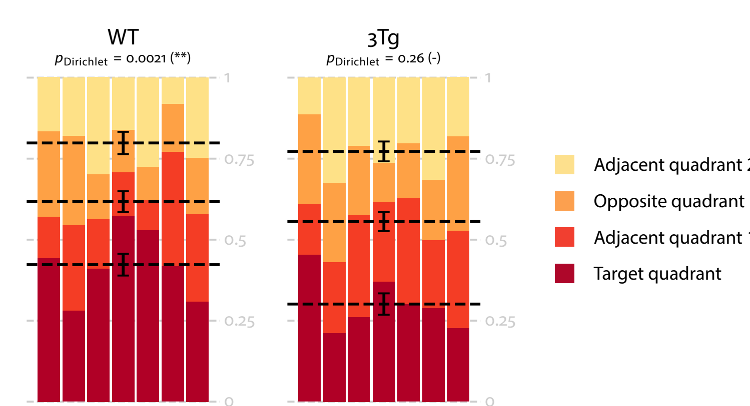
    Maugard, M., **Doux, C.** & Bonvento, G. *A new statistical method to analyze Morris Water Maze data using Dirichlet distribution*. F1000Research 2019 8:1601 8, 1601 (2019).



Doux, Chang+(in prep)



Doux, Baxter+(in prep)

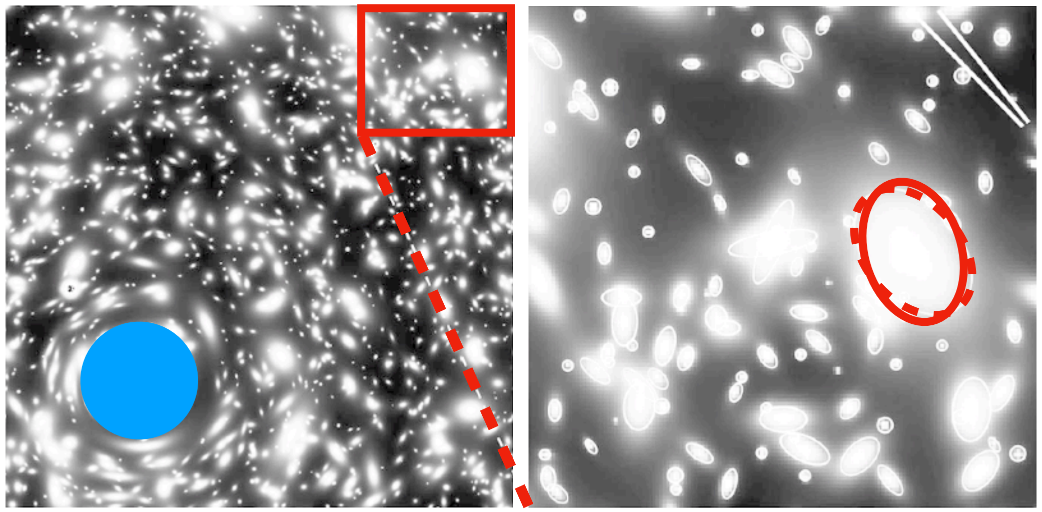


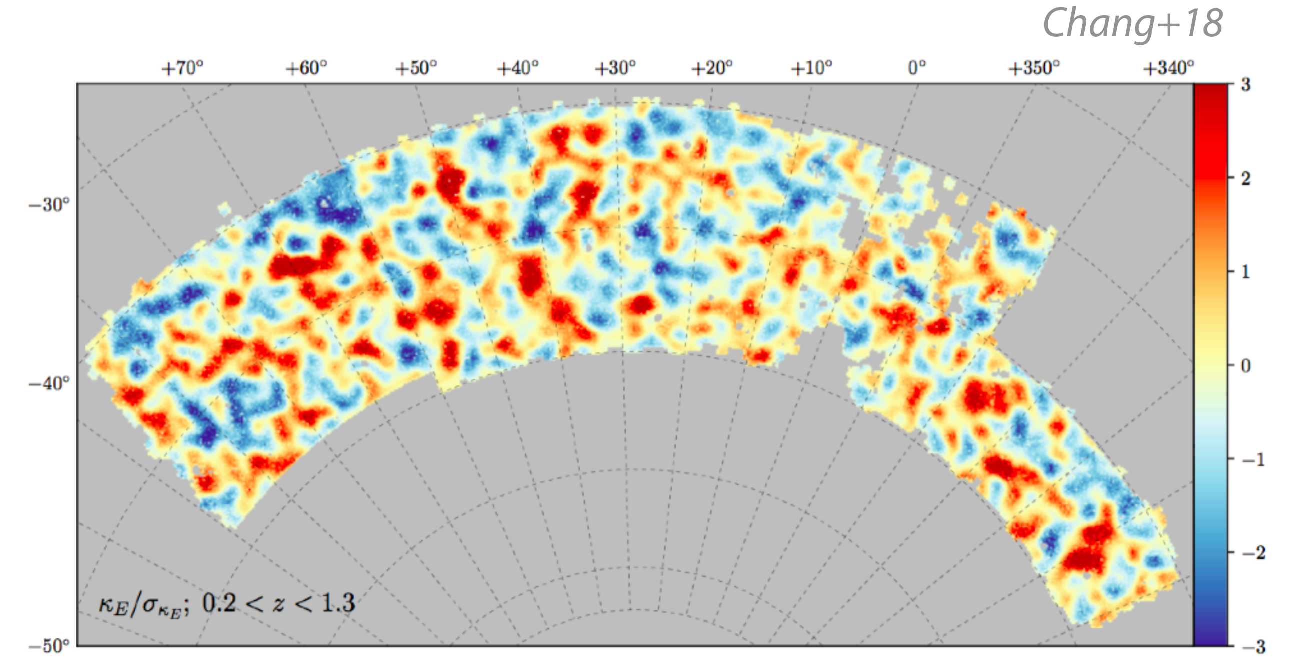
Maugard, Doux+2019

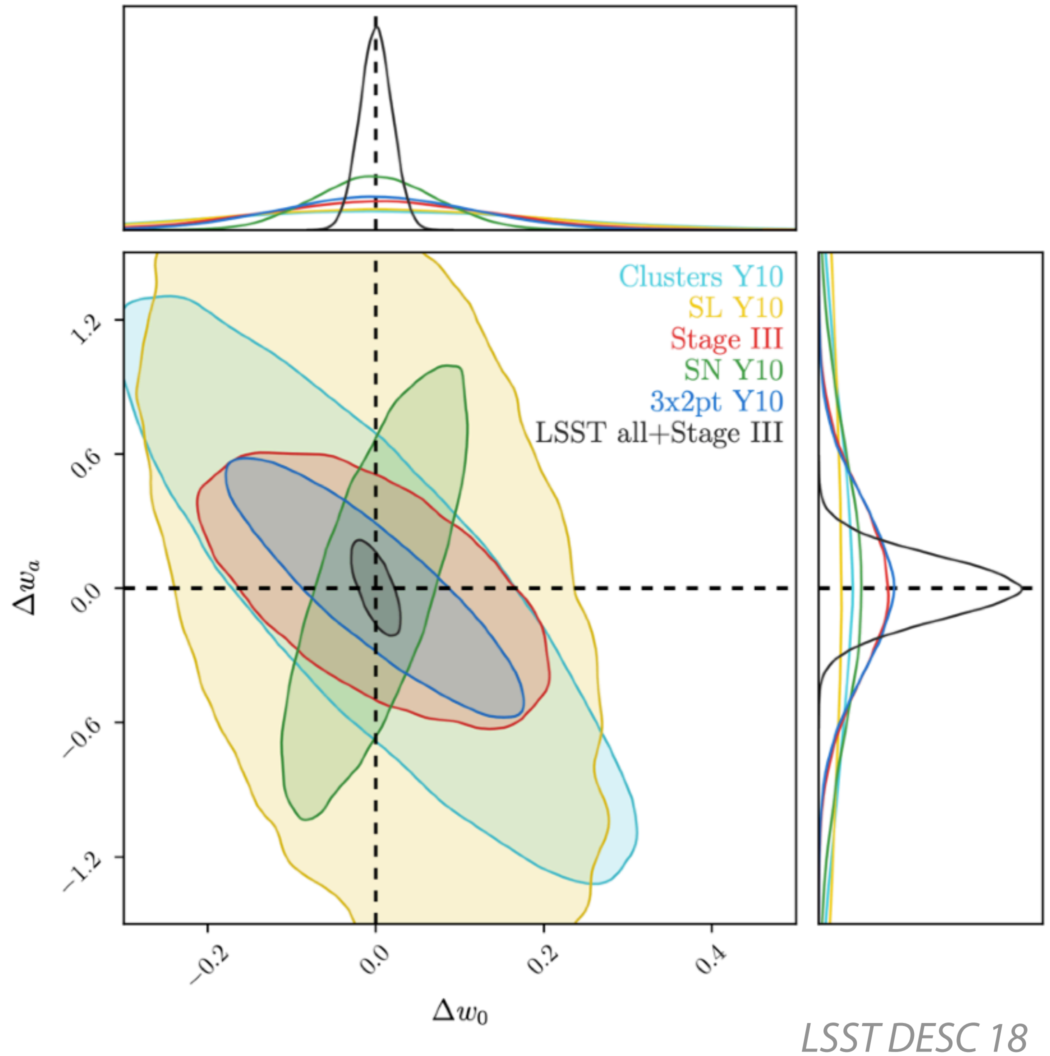
Outline

- Context and motivations
- The *blending* problem
- Method — variational autoencoders and simulated images
- Results — deblending performances
- Discussion and perspectives

Constraining dark energy with weak lensing



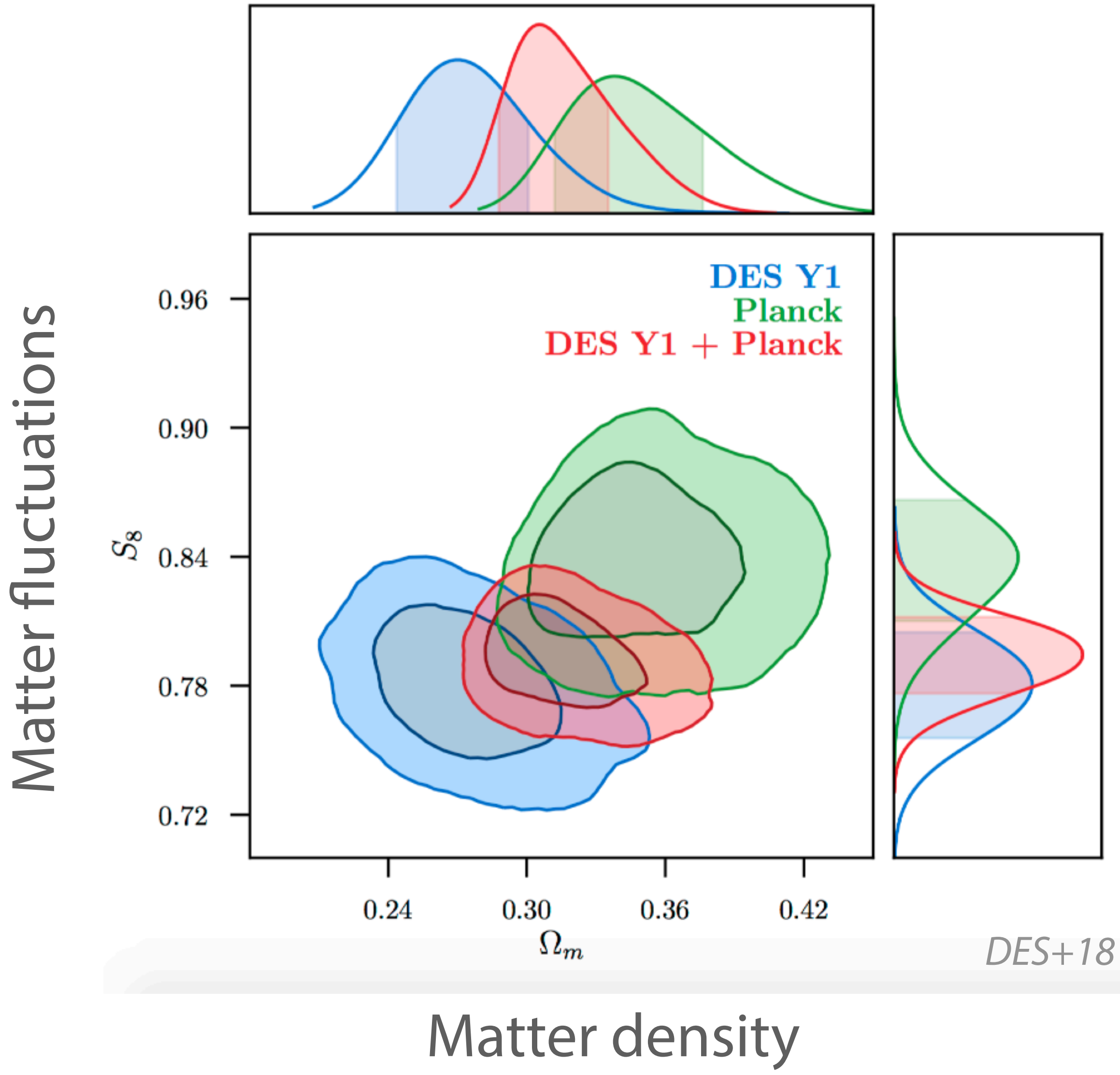




- Weak lensing by large-scale structure imprints **coherent distorsion** (~1%) on *galaxy shapes*
- Direct **mapping of matter distribution** (tomography) → measurement of power spectrum/2pt-functions
- Powerful probe of geometry+growth over wide redshift range → constraints **dark energy**

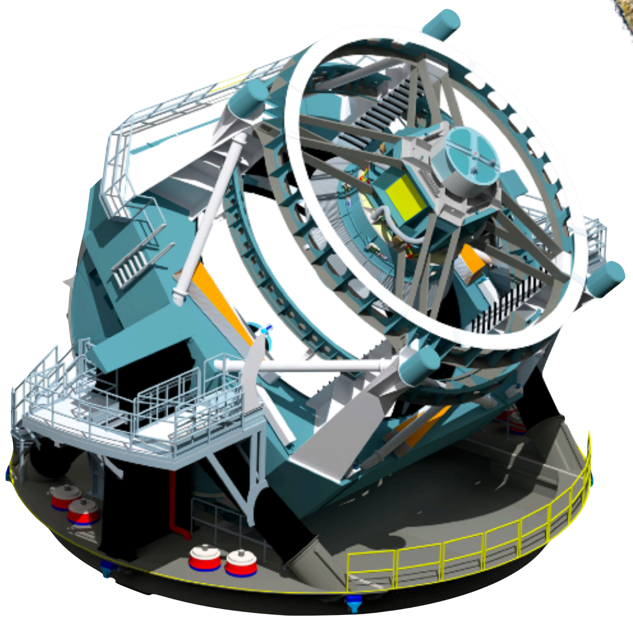
Ongoing surveys

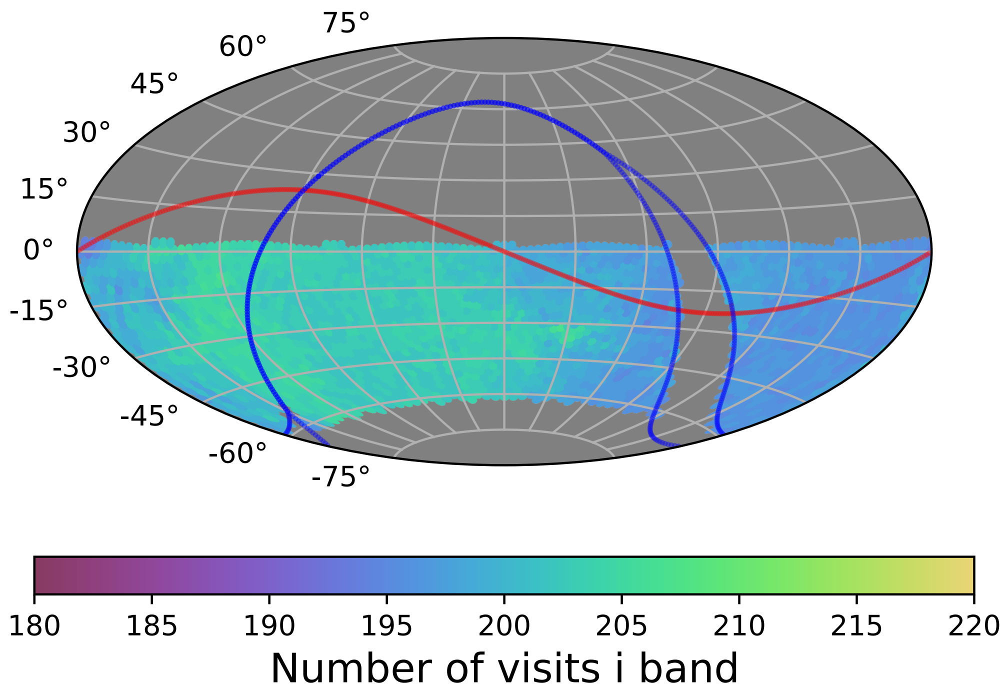
- **Dark Energy Survey (DES)**
  - 5000 deg^2 in Southern sky in *griz* at i<24
  - Upcoming Y3 analysis with 100M galaxies, stay tuned!
- **Hyper Suprime Cam (HSC)**
  - 1400 deg^2 in *grizy*, r<26 (much deeper)
  - Y1 analyzed in 2019-2020
- **Kilo Degree Survey (KiDS)**
  - 1300 deg^2 in *ugri* + IR bands
  - Recent release of KiDs-1000 with BOSS

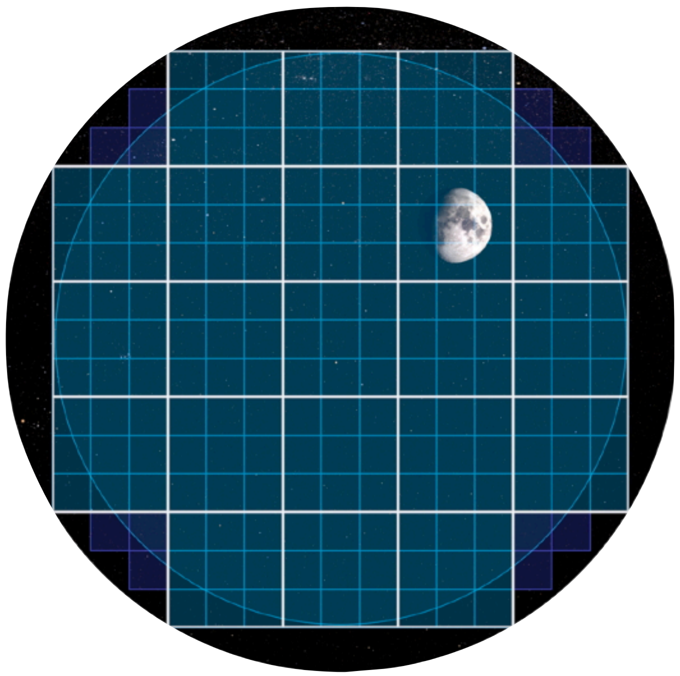


Next generation: LSST

- **Vera Rubin Observatory**
  - 8.4m wide-field telescope at Cerro Pachón
  - 3200 megapixel camera with *ugrizy* filters
- **LSST fast-wide-deep survey**
  - 10 years 2022-2032
  - Depth *r*<27.5, 10G galaxies
  - Raw data 10Tb/night



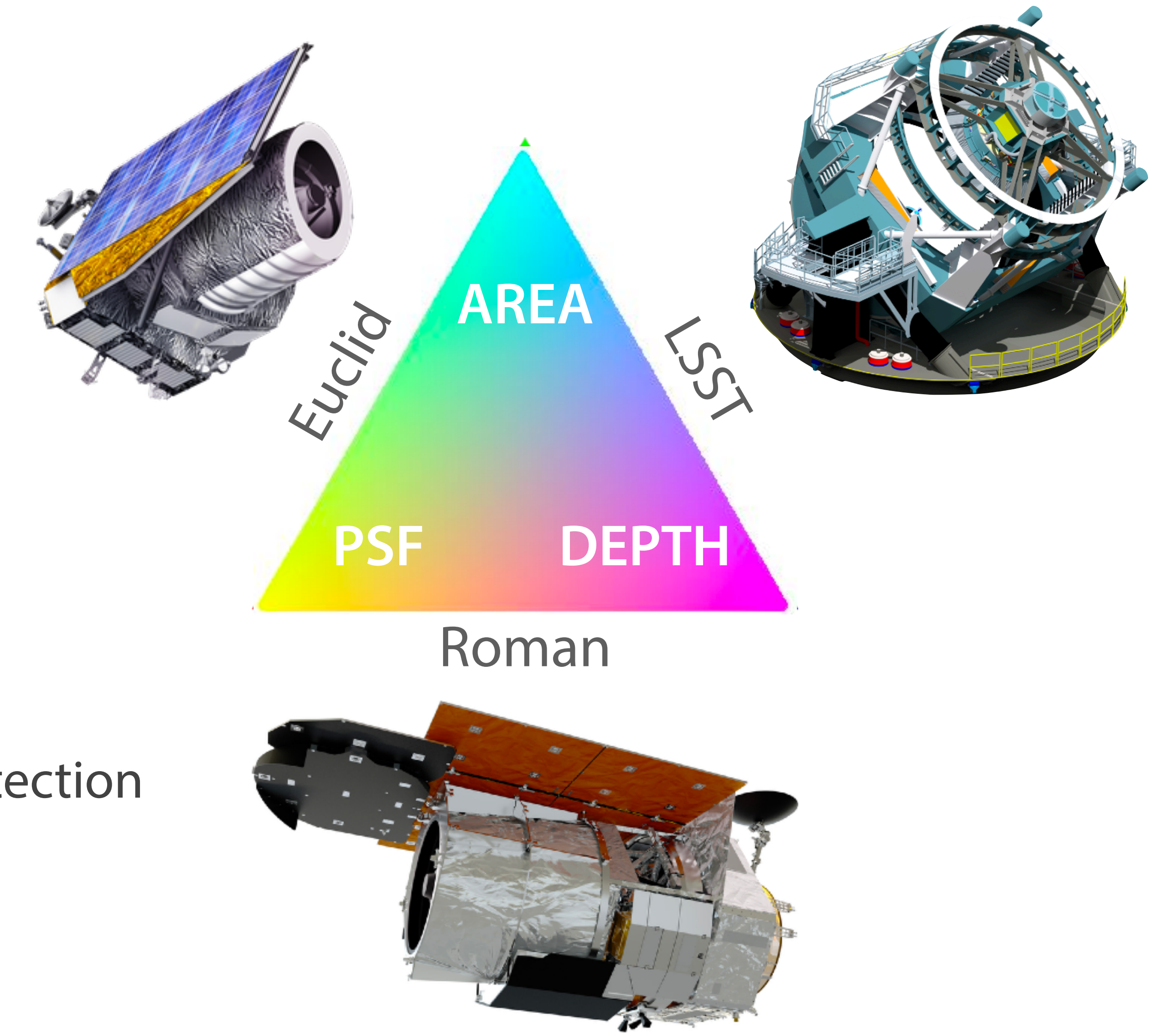




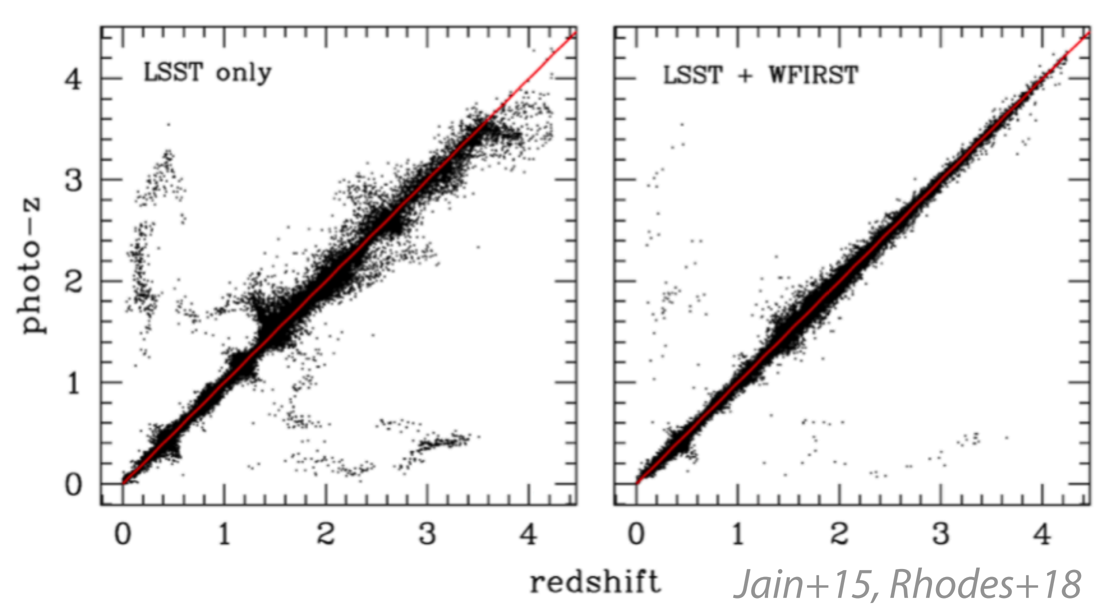
A preview of LSST data from HSC

Context: LSST + Euclid + Roman

- **Space-based weak lensing surveys**
  - ESA's Euclid mission
  - NASA's Nancy Grace Roman telescope
- **Complementary characteristics**
  - Resolution + IR bands
  - Catalog-level combination → eg better photo-z
  - Pixel-level joint processing → cross-calibration, detection and… deblending!

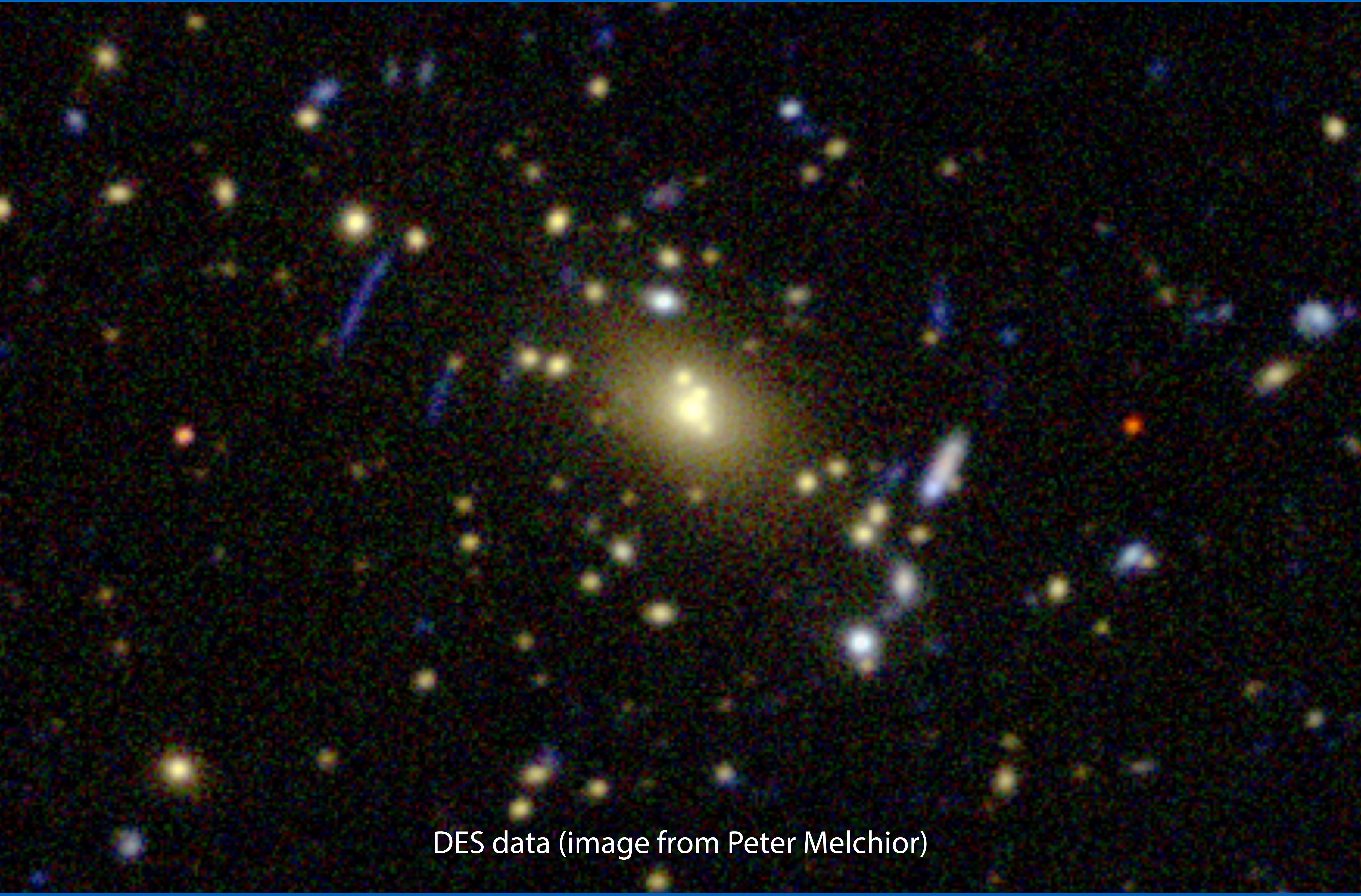


	WFIRST	Euclid	LSST
Start	2024	2021	2022
Duration	2 out of 6 years	5-6 years	10 years
Area	2300 sq. deg.	**15000 sq. deg.**	**18000 sq. deg.**
Footprint	South (within LSST)	Excludes galactic and elliptical planes	South
Passes	~5	1	~500
Bands	4 near-infrared	1 broad optical, 4 NIR	6 optical (ugrizy)
Depth	**27 in NIR**	24.5 in optical and NIR	**25 to 28 in optical**
Seeing	**0.12"**	**0.13"**	0.4"
Spectra	grism	grism	none



Jain+15, Rhodes+18

Space+ground observations



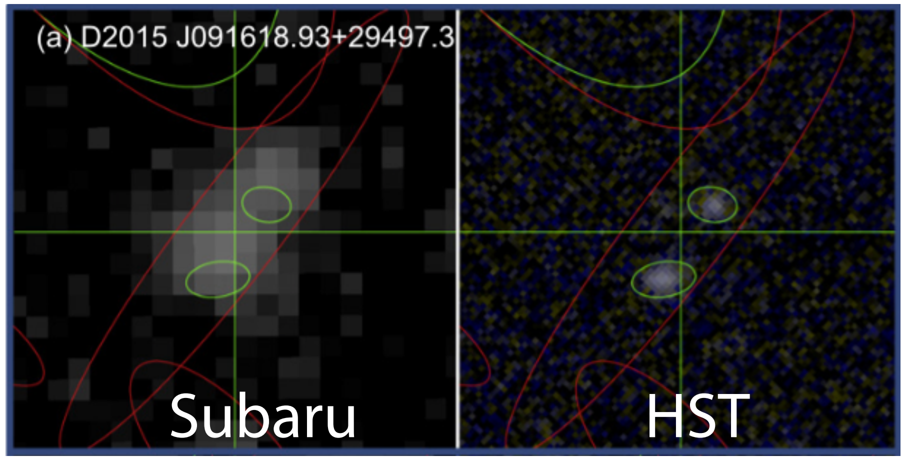
DES data (image from Peter Melchior)

Space+ground observations

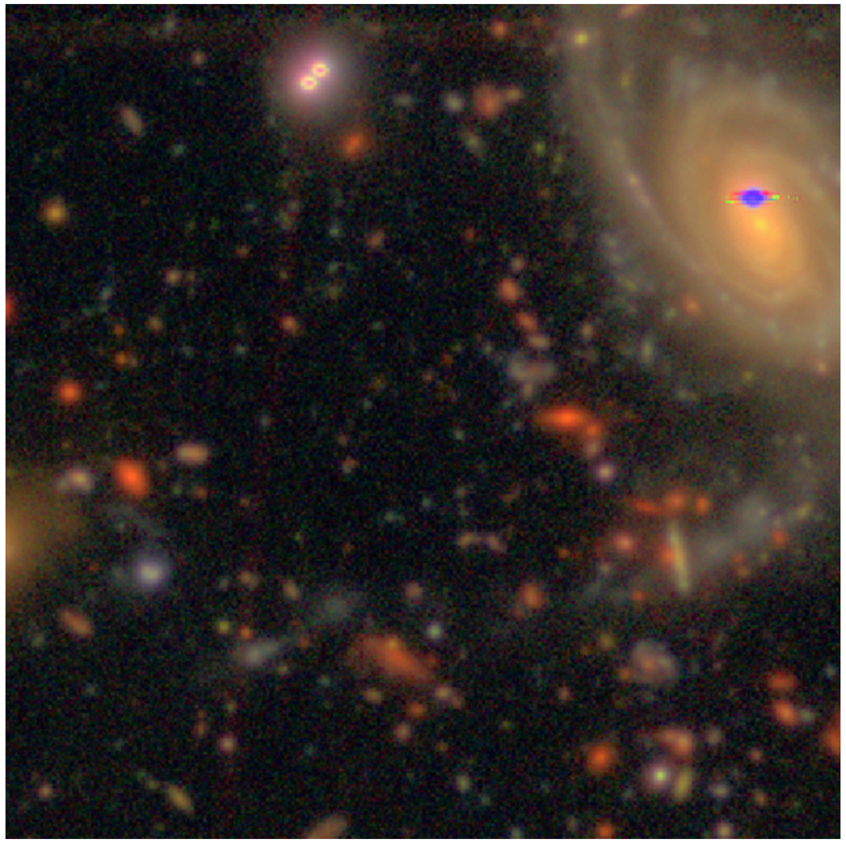
CLASH WFC3/IR data (image from Peter Melchior)

The deblending problem

- **Why is it an issue?**
  - ~50% galaxies are blended at LSST's depth (58% for HSC, see Bosch+17)
  - Discarding them decreases *statistical power* and induces *selection biases*
  - Impacts *shape and color/redshift measurement*, thus all weak lensing science!
- **Why is it difficult?**
  - Modelling morphologies beyond fitting profiles (Sérsic, de Vaucouleurs, exponential, etc)
  - It's impossible… *without making assumptions* (sic Robert Lupton)
  - Strongly tied to detection algorithm (iterative procedure), *ie* unrecognised blends

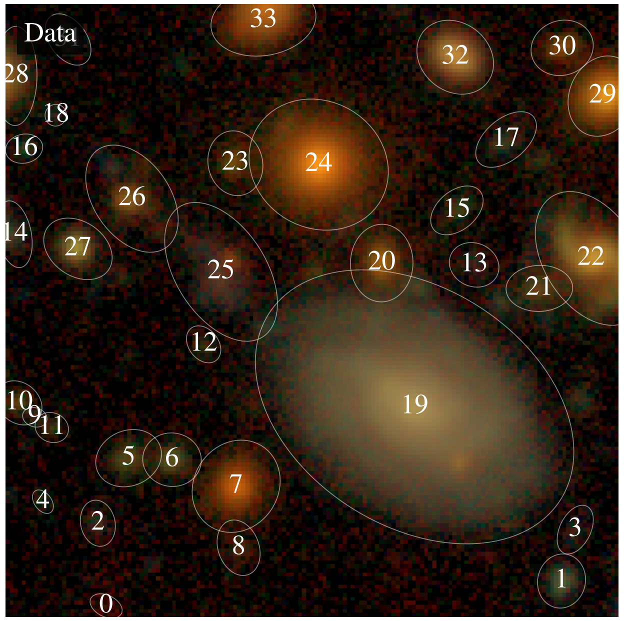


Dawson+15



The deblending problem

- **Existing deblenders**
  - SExtractor (Bertin+96) : segmentation via thresholding
  - SDSS deblender (Lupton, in prep) : symmetry constraint, only one band
  - Inpainting techniques (Zhang+15, Connor+17)
  - MuSCADeT (Joseph+16) : source separation with sparse spatial constraint
  - Multi-Object Fitting (Drlica-Wagner+18) : friends-of-friends + bulge/disk model
  - **SCARLET** (Melchior,Moolekamp+18) > integration in LSST pipeline

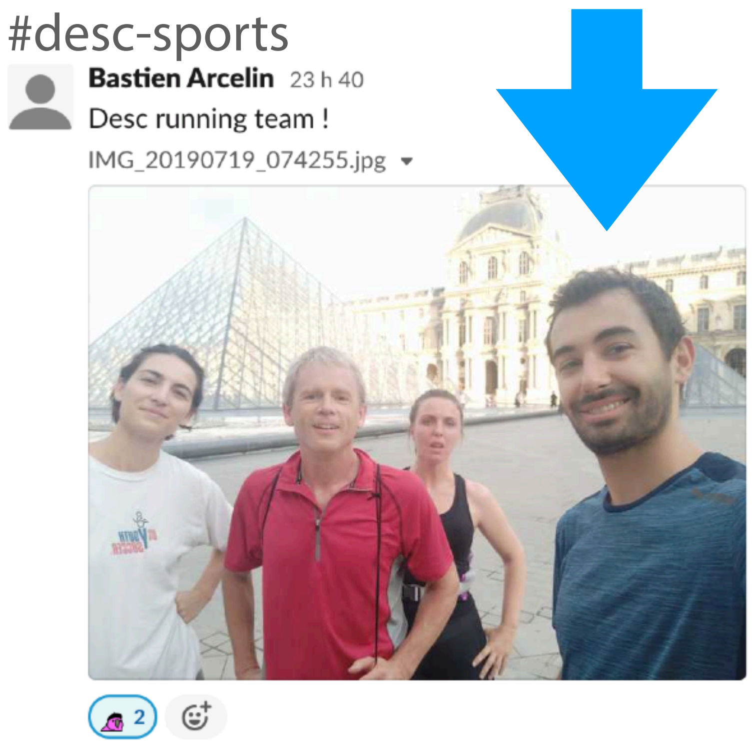


$$= \sum_{k=1}^{K} \underbrace{A_k^\top}_{\text{Profile}} \underbrace{S_k}_{\text{SED}}$$

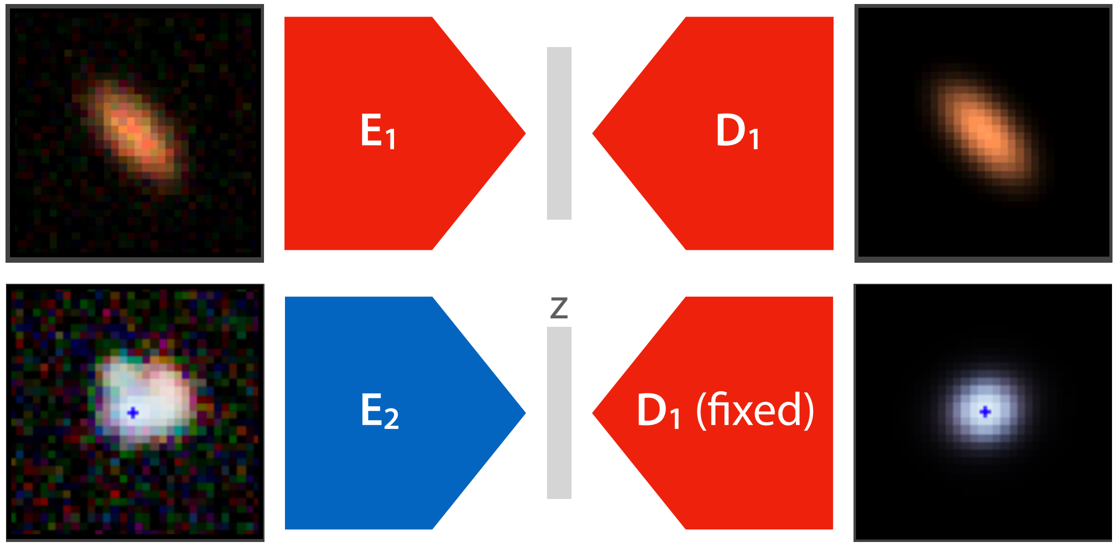
- symmetry and monotonicity constraints on A_k
- bS-DMM constrained minimization
- uses all bands
- λ-dep PSF + correlated noise

This work

- **Goals**
  1. Minimum assumption on galaxy morphology
  2. Fast enough to deal with LSST data (15Tb/night)
  3. Incorporate LSST with Euclid/WFIRST data

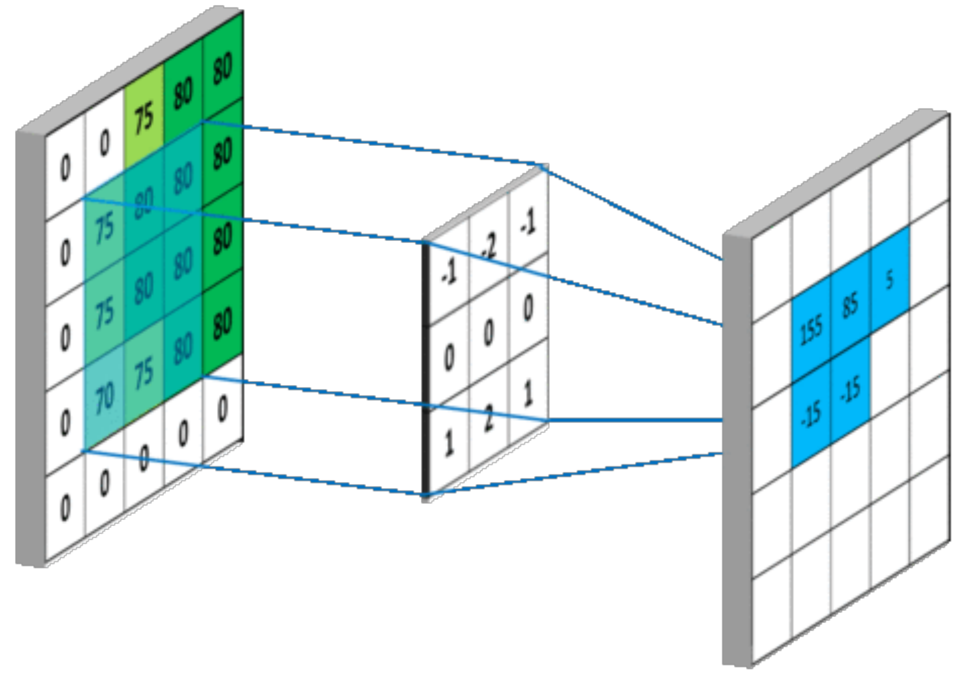


- **This work** with Bastien Arcelin (grad student at APC, Paris)
  - We developed and tested a new method based on **two probabilistic CNNs** sharing weights
    - Network 1 (VAE) learns a *generative model* **D_1**
    - Network 2 (deblender) deblends with **E_2** under constraints
    - Multi-bands/instru used as image channels (like RGB)
    - Super fast once trained
  - **Results**
    - ✓ Training/testing on simulated images with fixed PSF
    - ✓ Accuracy measured by shear/flux recovery
    - ✓ Initial tests on real data

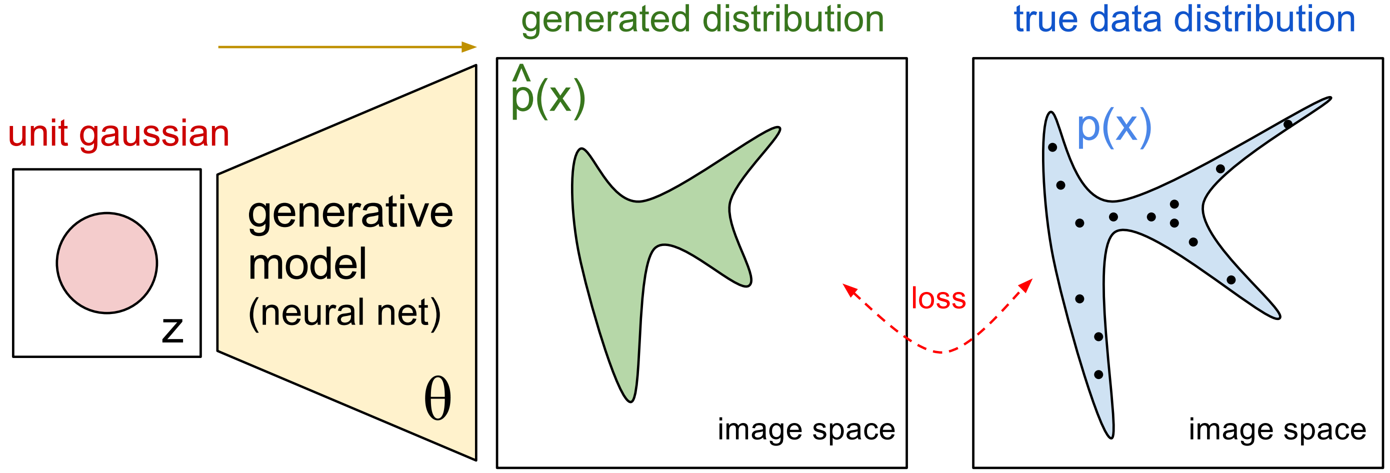


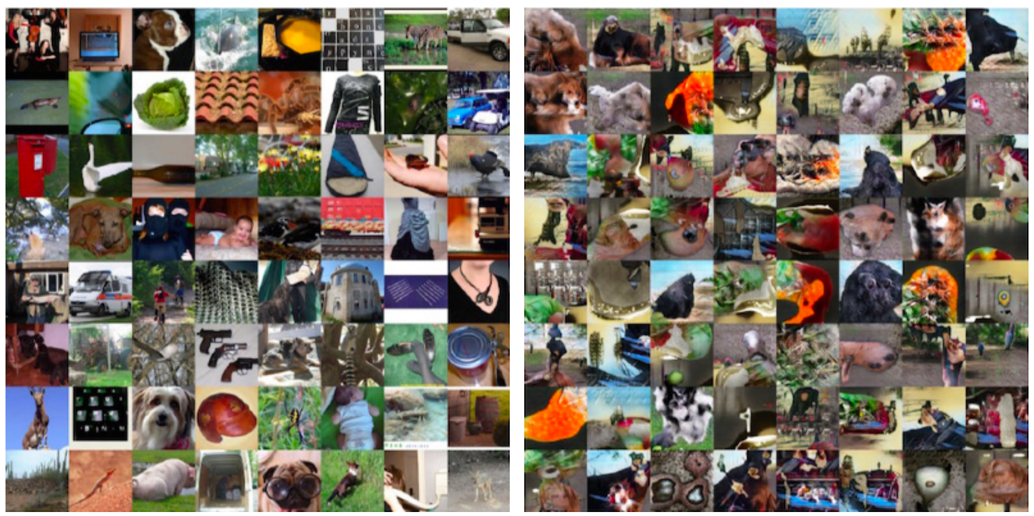
CNNs and generative models

- **Convolutional neural networks**
  - learn image filters to find pattern/recover transformation
  - good at classification, segmentation, tagging… and
- **Generative/bayesian models**
  - Latent variable models, ie linking data ***X*** to ***Z***~*N*(0,𝟙)
    - Generative Adversarial Networks (GAN, Goodfellow+15)
    - Variational Auto Encoders (VAE, Kingma+14)

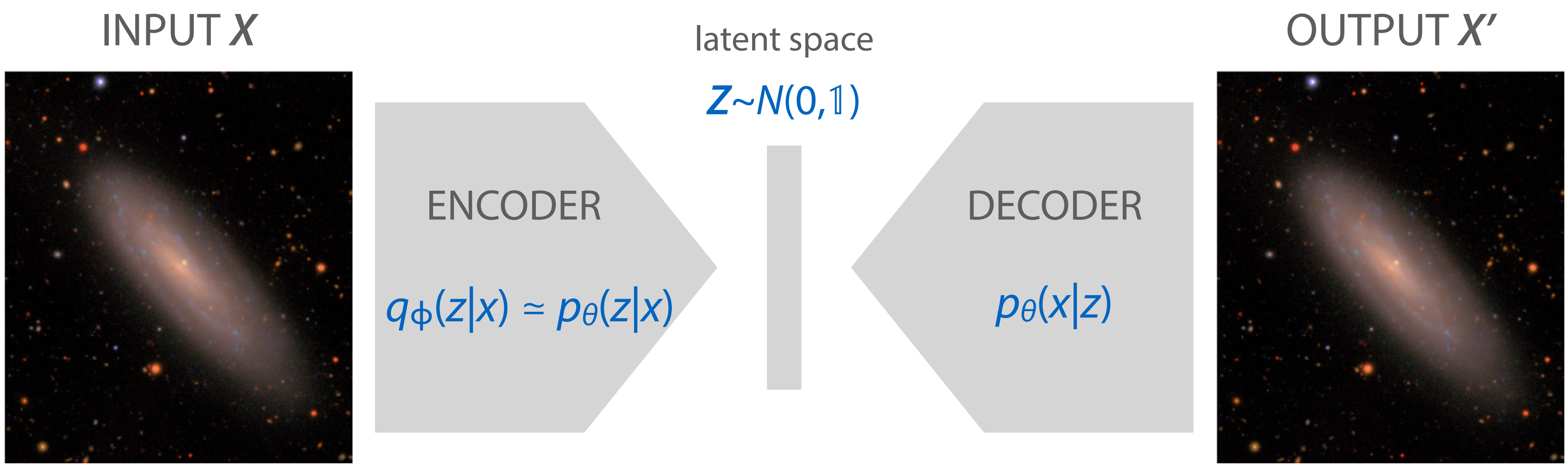


mlnotebook.github.io





Variational autoencoder (VAE)



- **Learn** a *bayesian model* between data ***X*** and latent variables ***Z***~$N(0,\mathbb{1})$

  - DECODER network learns the generative model $p_\theta(x|z)$

  - ENCODER network approximates the posterior with $q_\phi(z|x) \simeq p_\theta(z|x)$

  - Trained by maximizing marginal distribution of ***X***, log p(***X***), lower bound (ELBO, Kingma+14)

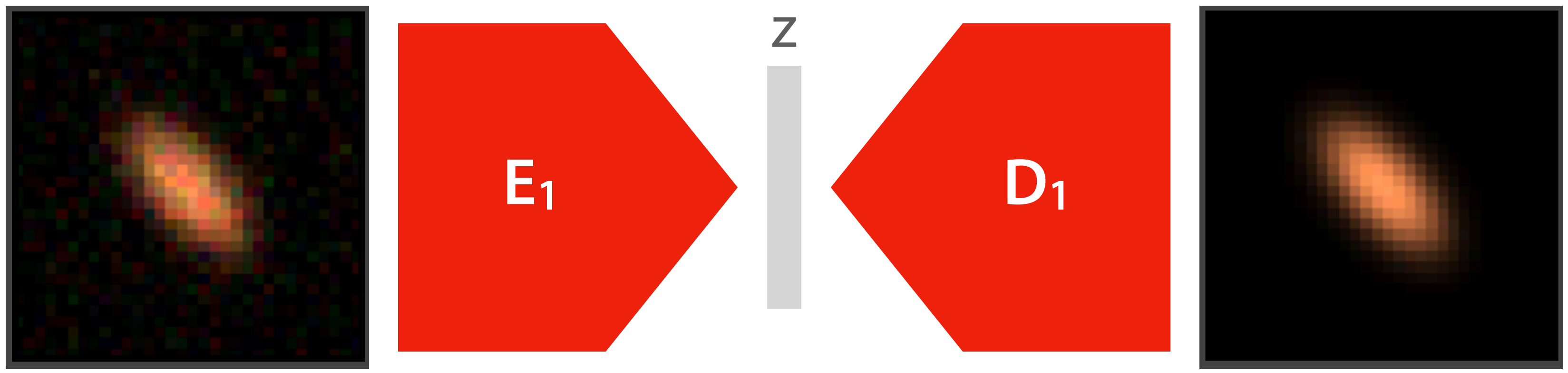
$$\log p(x) \geq -D_{\mathrm{KL}}\left(q_\phi(z|x)||\mathcal{N}(\mathbf{0},\mathbf{1})\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]$$

~regularization reconstruction ***X*** *vs* ***X′***

- **Architecture** ENCODER = CNN + dense layers, DECODER in mirror

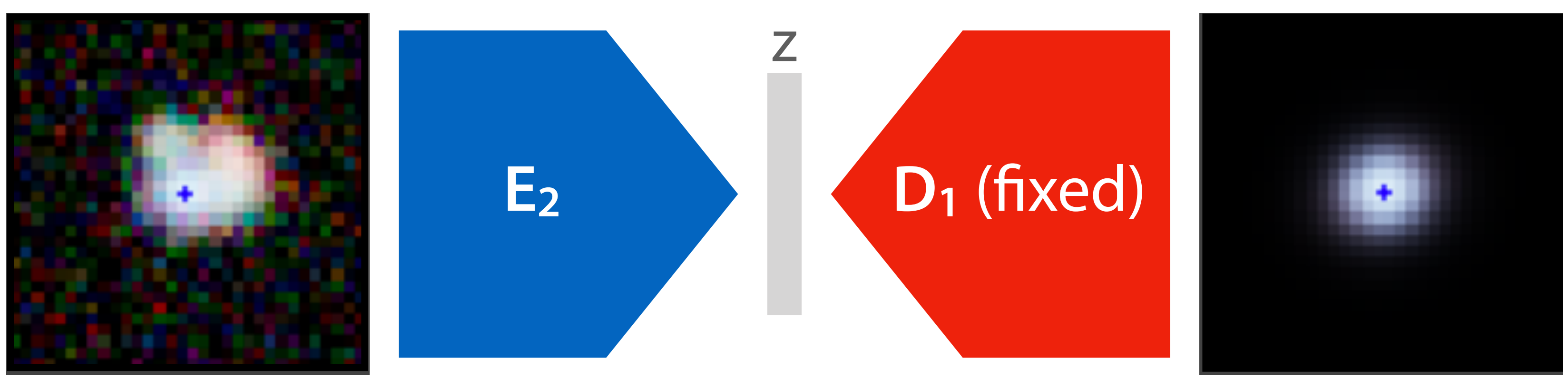
2-step method

1. **Generative model of isolated galaxy images** with a variational auto-encoder



- VAE encodes noisy images of isolated and ~centred galaxies in unsupervised latent space

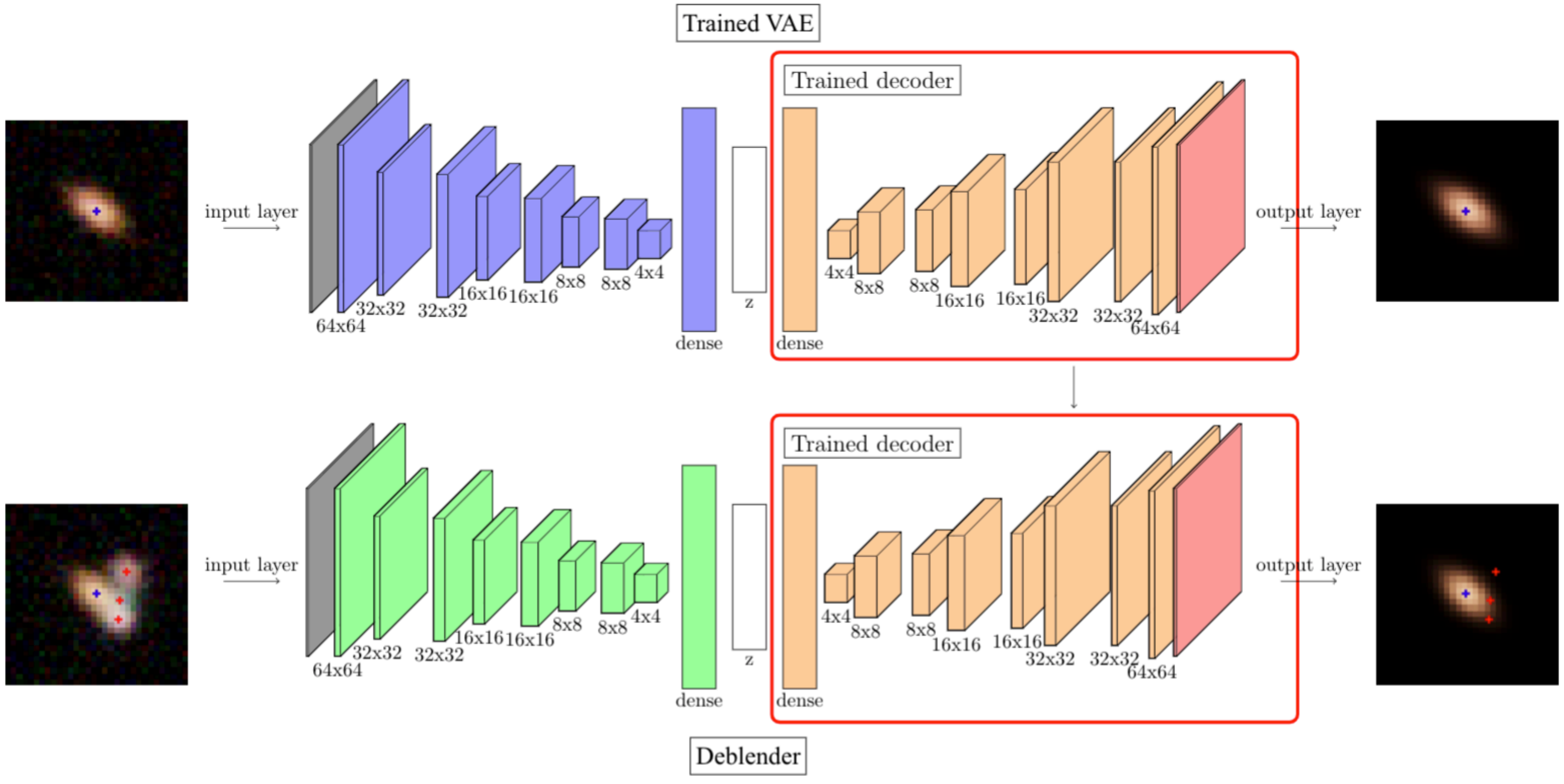
2. **Deblender for central galaxy** with fixed generative model



- Encoder learns to approximate $P(z_{\text{center}}|X_{\text{blended}}) \rightarrow$ output $X' = \text{DECODER}(z_{\text{center}})$
- Deblending is *constrained* by prior in latent space
- Validated by reproduction of shapes and fluxes

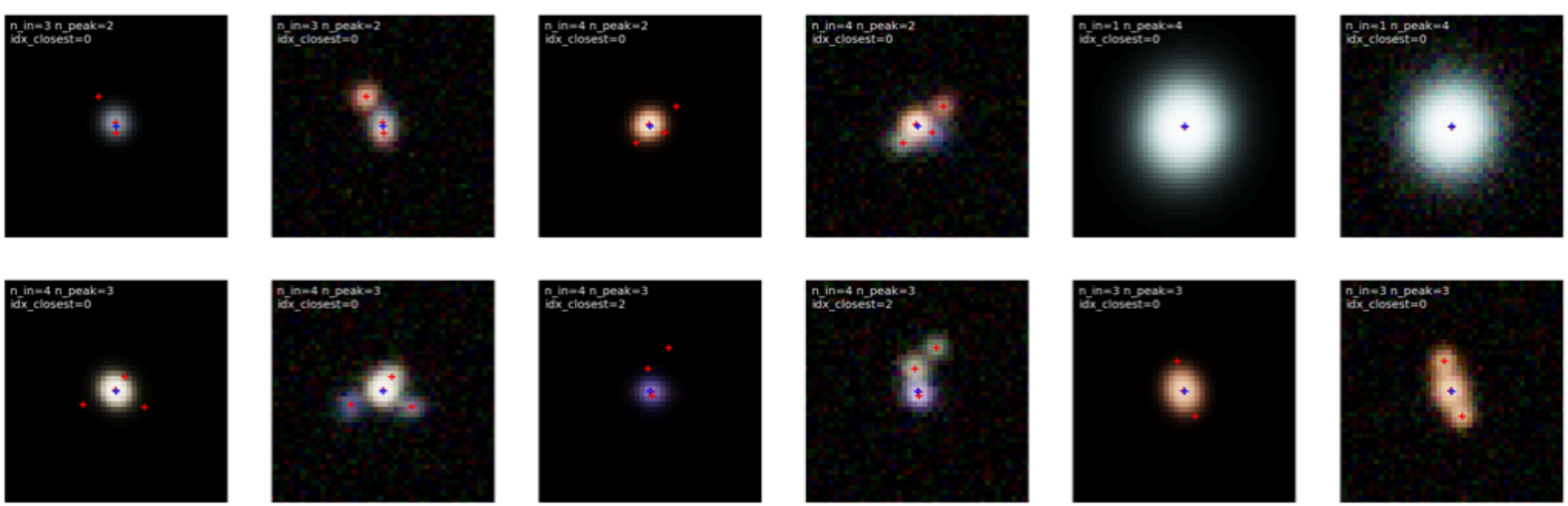
VAE/deblender architecture

- **Data** INPUT = noisy isolated/blended images (6 or 10 channels)
  OUTPUT = isolated noiseless images (6 or 10 channels)
- **Architecture** CNN β-VAE with 32 latent variables
  - $\beta=10^{-2}$ to improve reconstruction
  - PReLU activations (~lossless), >5M parameters
  - posterior $q_\phi(z|x)=N(\mu(x),\sigma(x))$
  - likelihood $p_\theta(x|z)$="continuous Bernouilli" with tuned normalisation

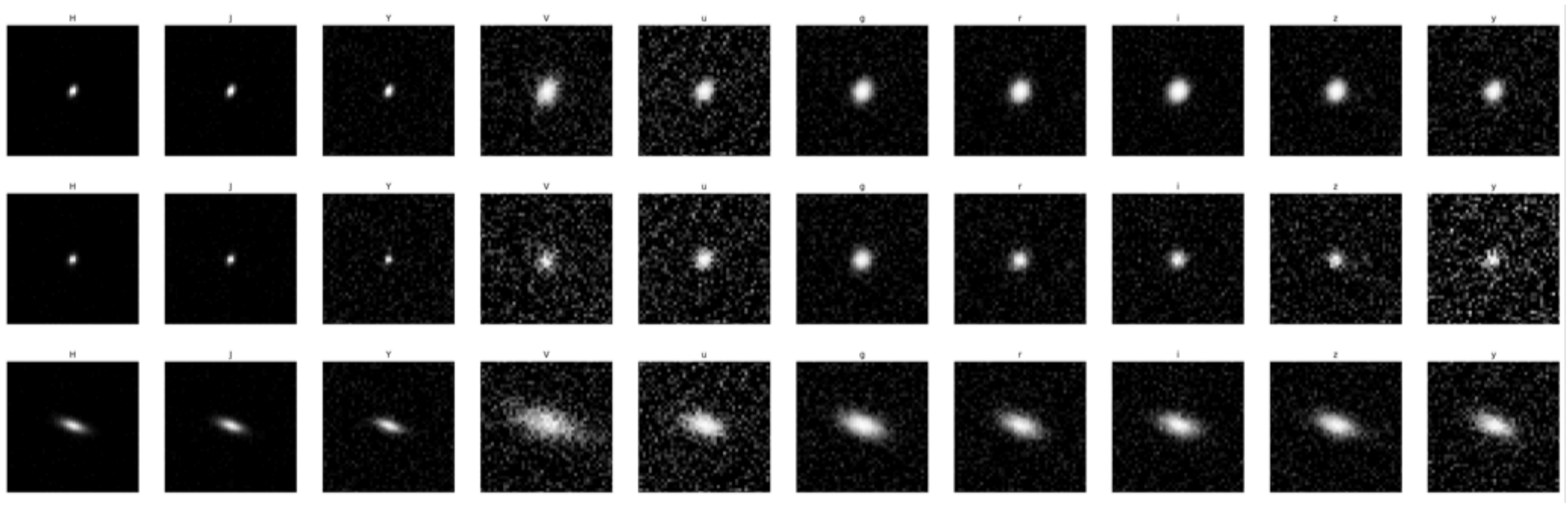


Training samples

- **Catalog** — Parametric chromatic `GalSim` images from COSMOS $i<25.2$
  - 100k/10k for training/testing
- **Bands/exposures** — **6 LSST bands *ugrizy***, 824 15s exposures (~10 years) in total (uneven)
  - **Euclid VIS + 3 NIR**, 4 450s exposures
- **PSF** — Fixed PSF, Kolmogorov 0.65” for LSST, Moffat 0.18” (0.22”) for Euclid
- **Noise** — Poisson noise with fiducial sky background values
- **Decentering** — 1- perfectly centred
  2- uniformly decentered by half an LSST pixel
  3- centered on brightest peak in *r* (simplistic `photutils` peak finder)

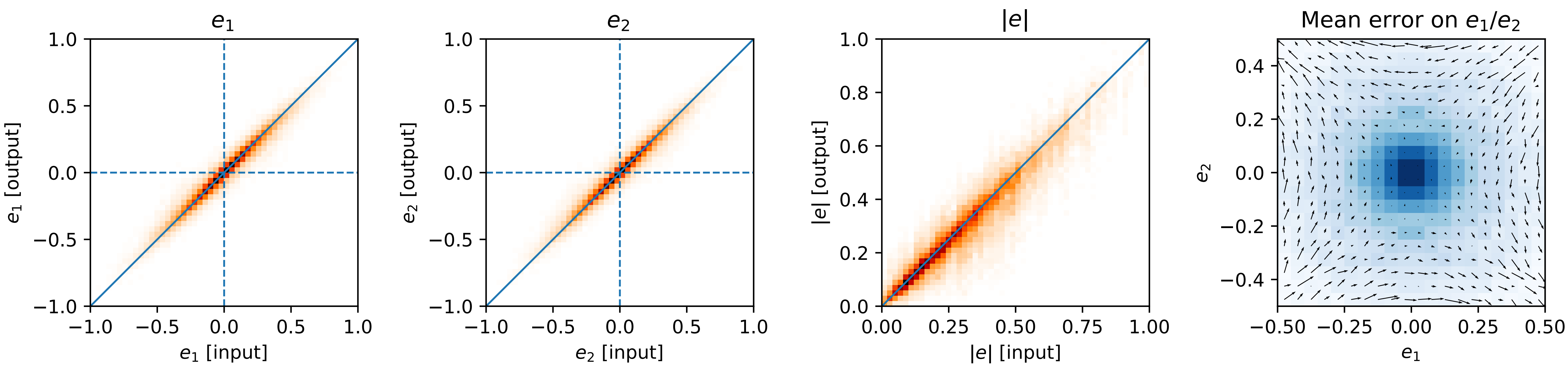


Training samples - isolated galaxies

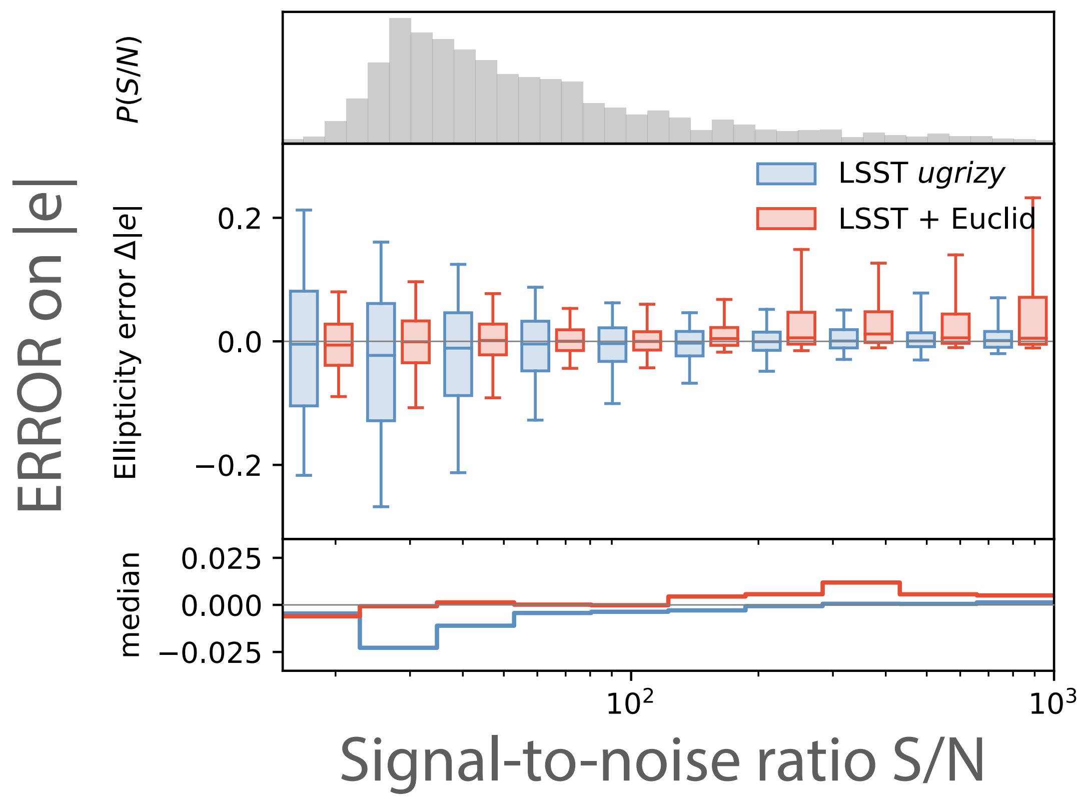


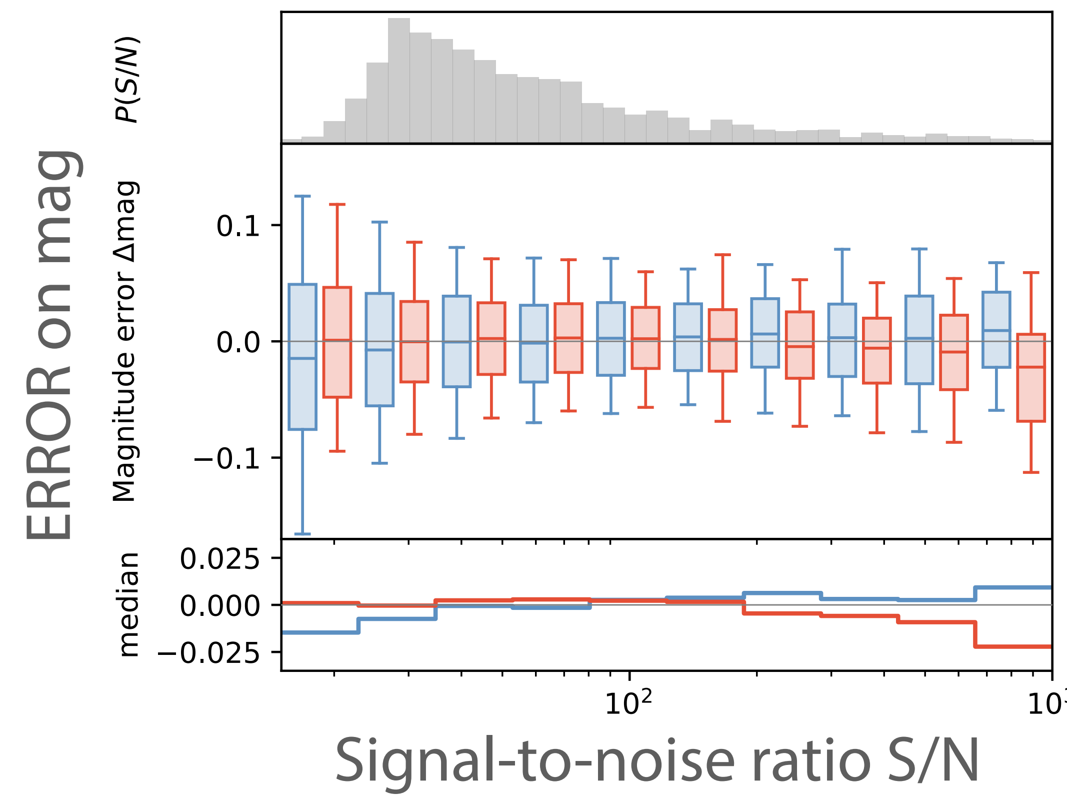
Results : validating generative model

- Comparison input/output ellipticities



- Analysis of reconstruction errors as function of S/N (or mag, distance, etc)

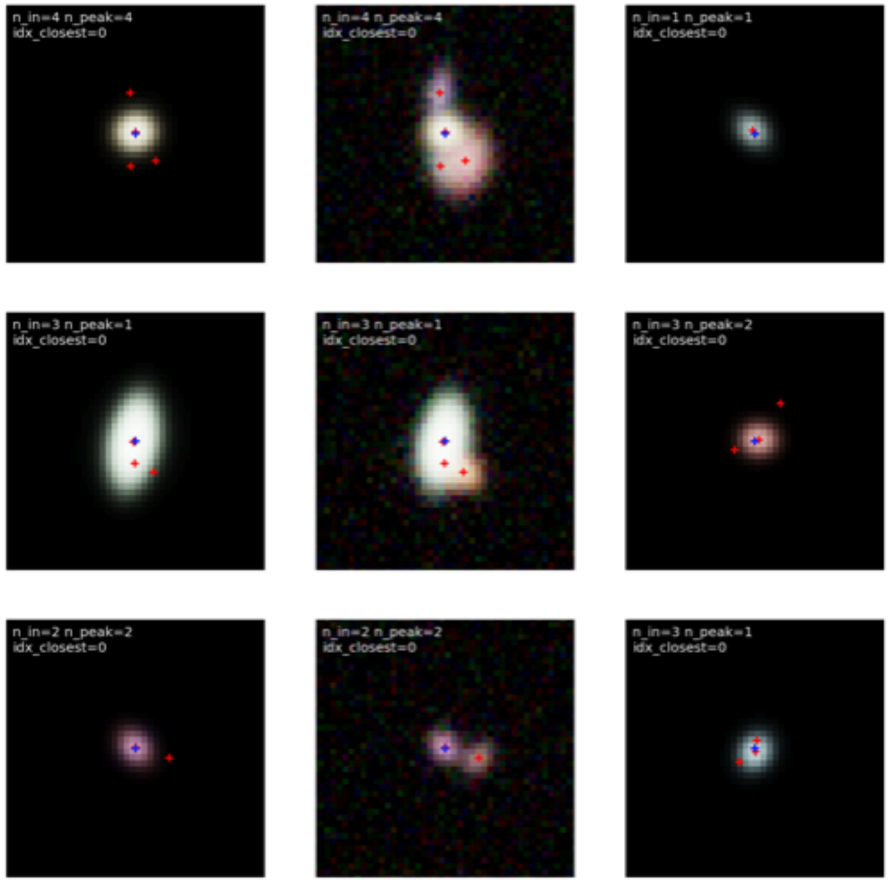


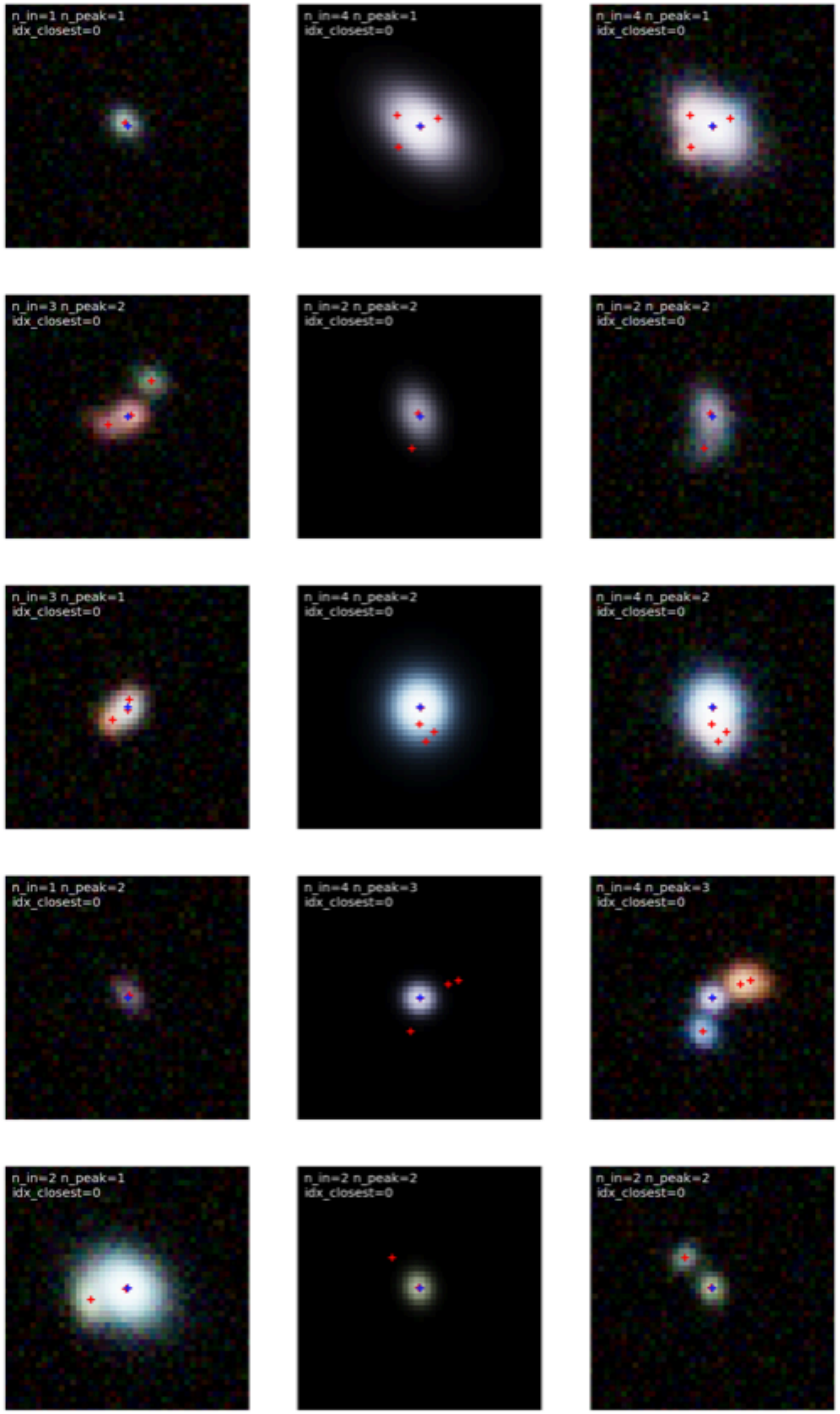


Training samples - blended galaxies

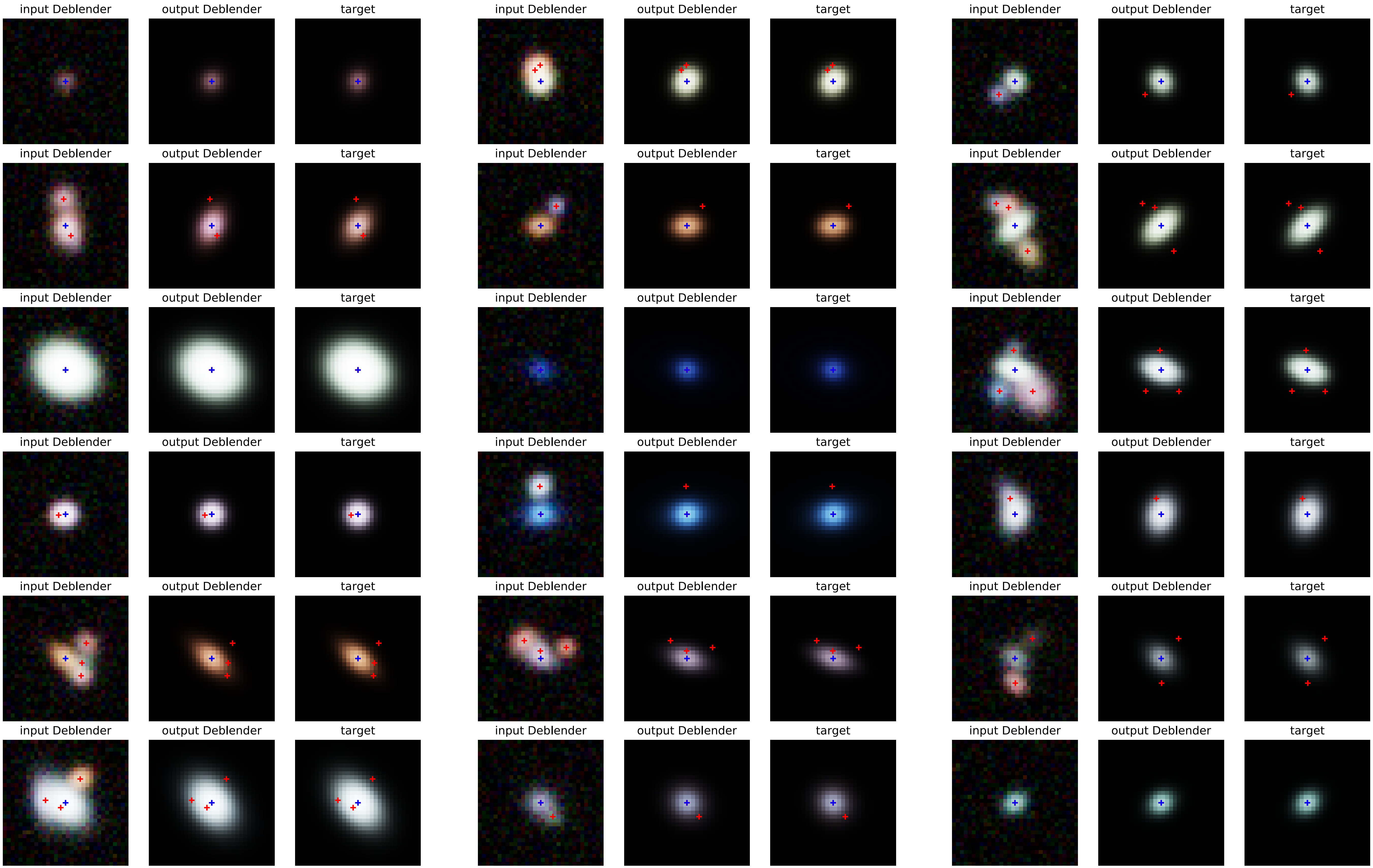
- **Generation of artificial blends**
  - 25% of [1,2,3,4] galaxies
  - Brightest galaxy centred (3 decentering methods)
  - Exclusion area of PSF $\theta_{\mathrm{fwhm}}/2=0.3''$ between centers
  - Total blendedness metrics B_{tot} (from Scarlet)

$$B_{\mathrm{tot}} = 1 - \frac{\langle I_{\mathrm{centered}}, I_{\mathrm{centered}} \rangle}{\langle I_{\mathrm{centered}}, I_{\mathrm{total}} \rangle}$$



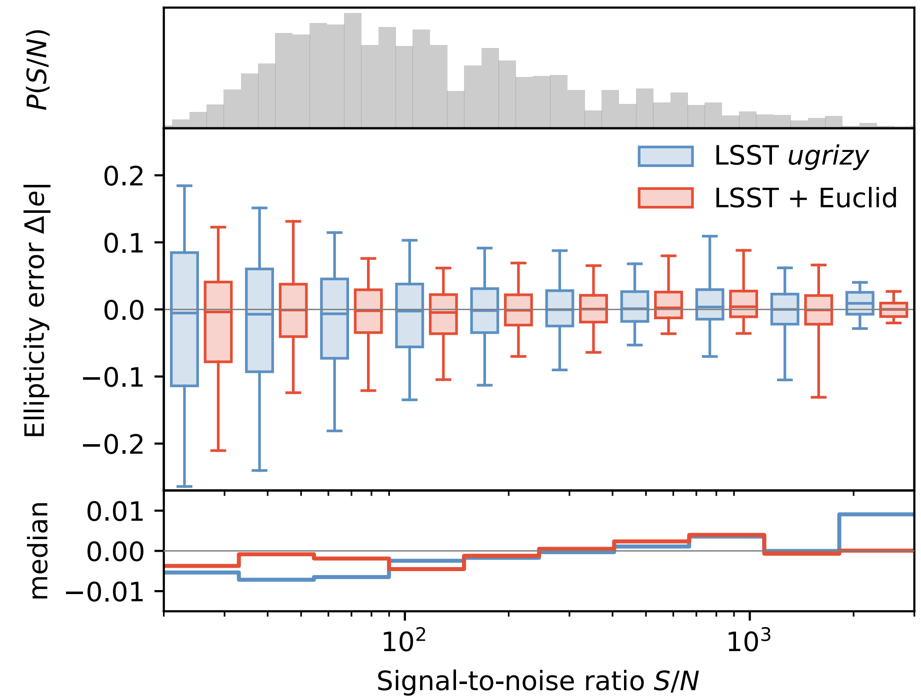


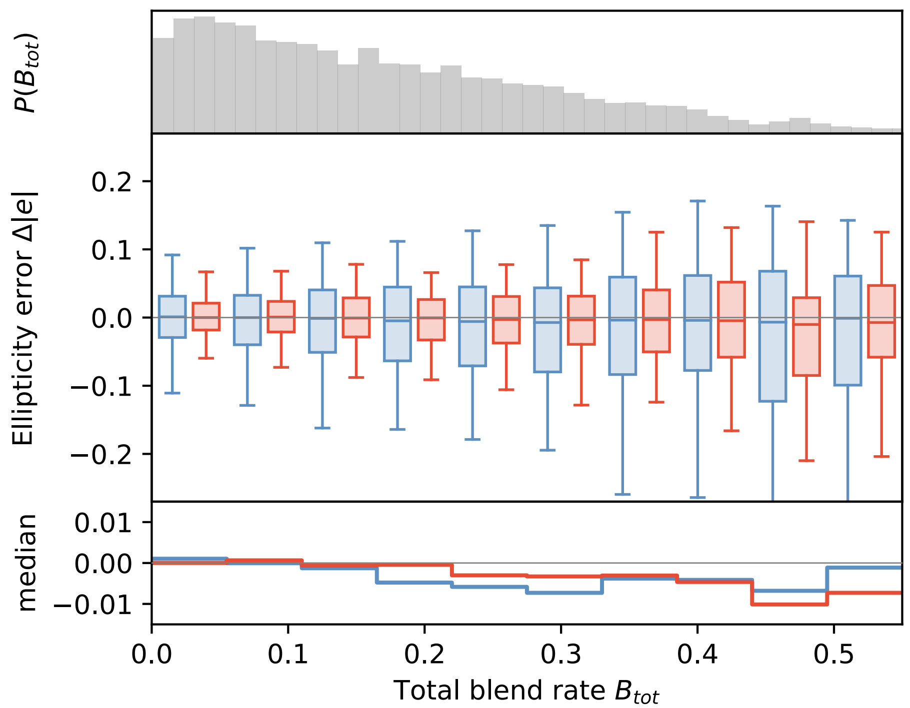
Results - deblender

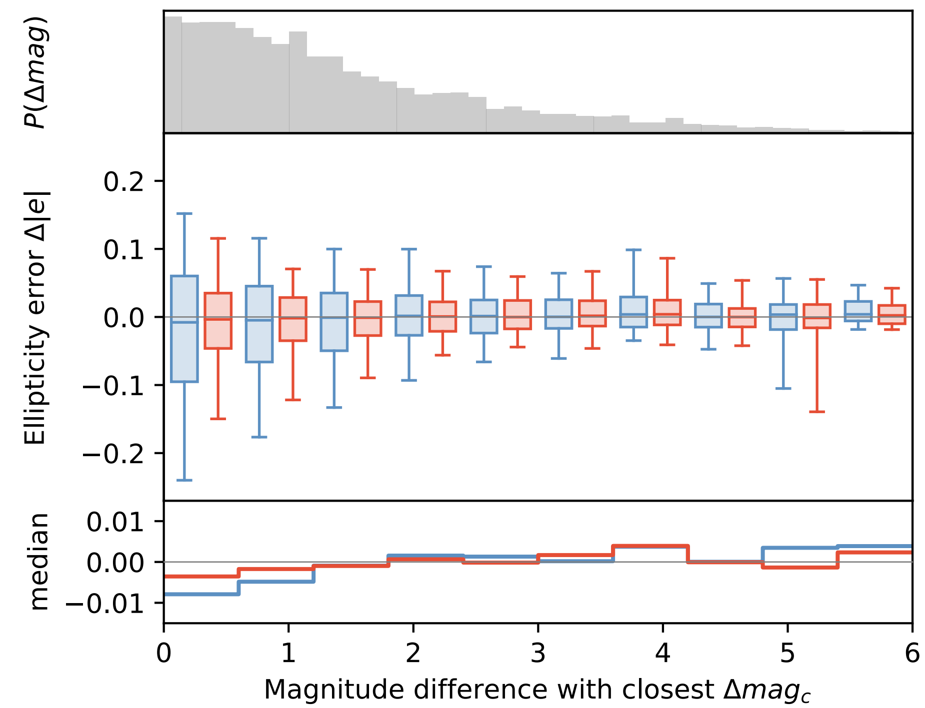


Deblender performances

- **Analysis of ellipticity errors**
  - Median errors within ±0.01, stable across $10<S/N<3000$, $0<B_{\mathrm{tot}}<1$
  - 30% smaller error distribution with LSST+Euclid
  - Ellipticity biases of 5.6%(1.6%) for LSST(+Euclid)
  - Shear multiplicative bias of 4-6% on sample

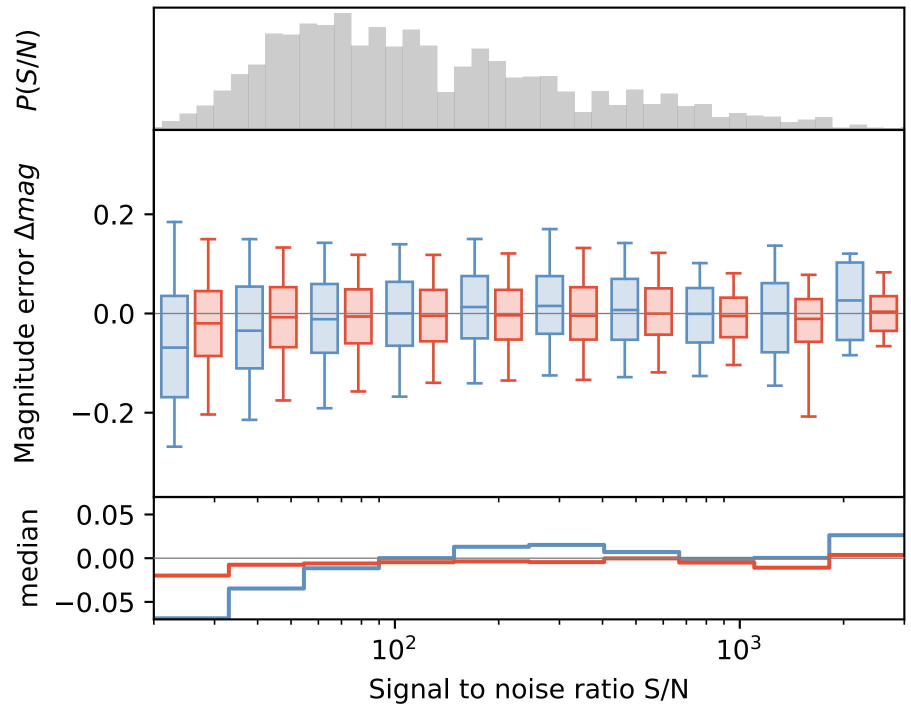


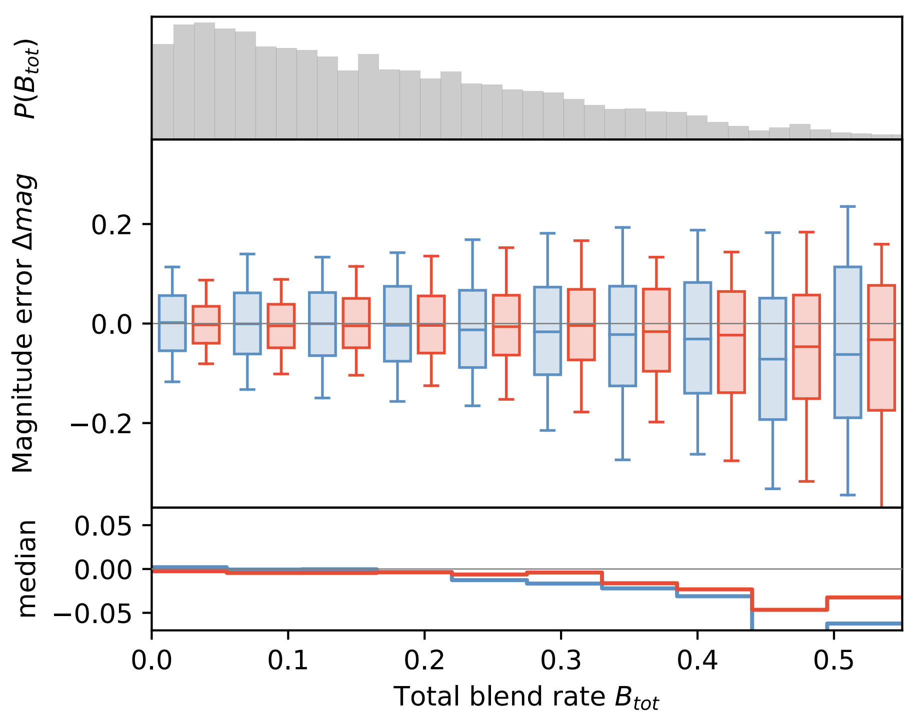


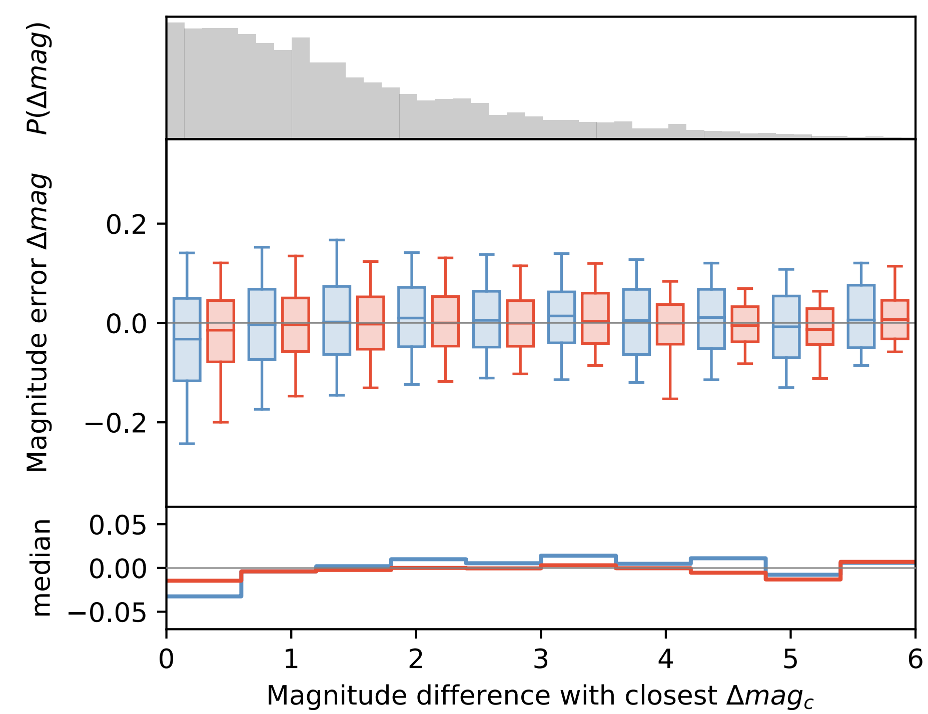


Deblender performances

- **Analysis of magnitude errors**
  - Median errors within ±0.05, stable across $10<S/N<3000$, $0<B_{\mathrm{tot}}<1$
  - 20% smaller error distribution with LSST+Euclid

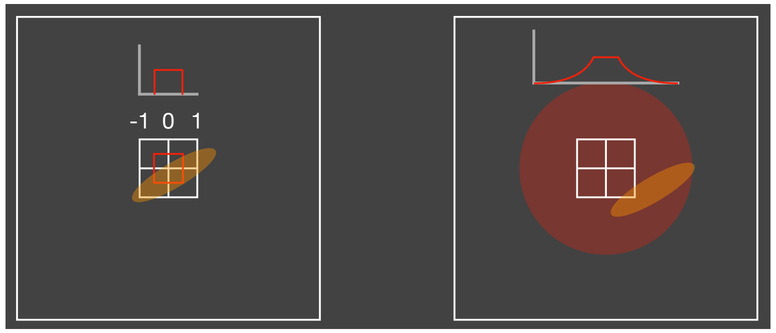




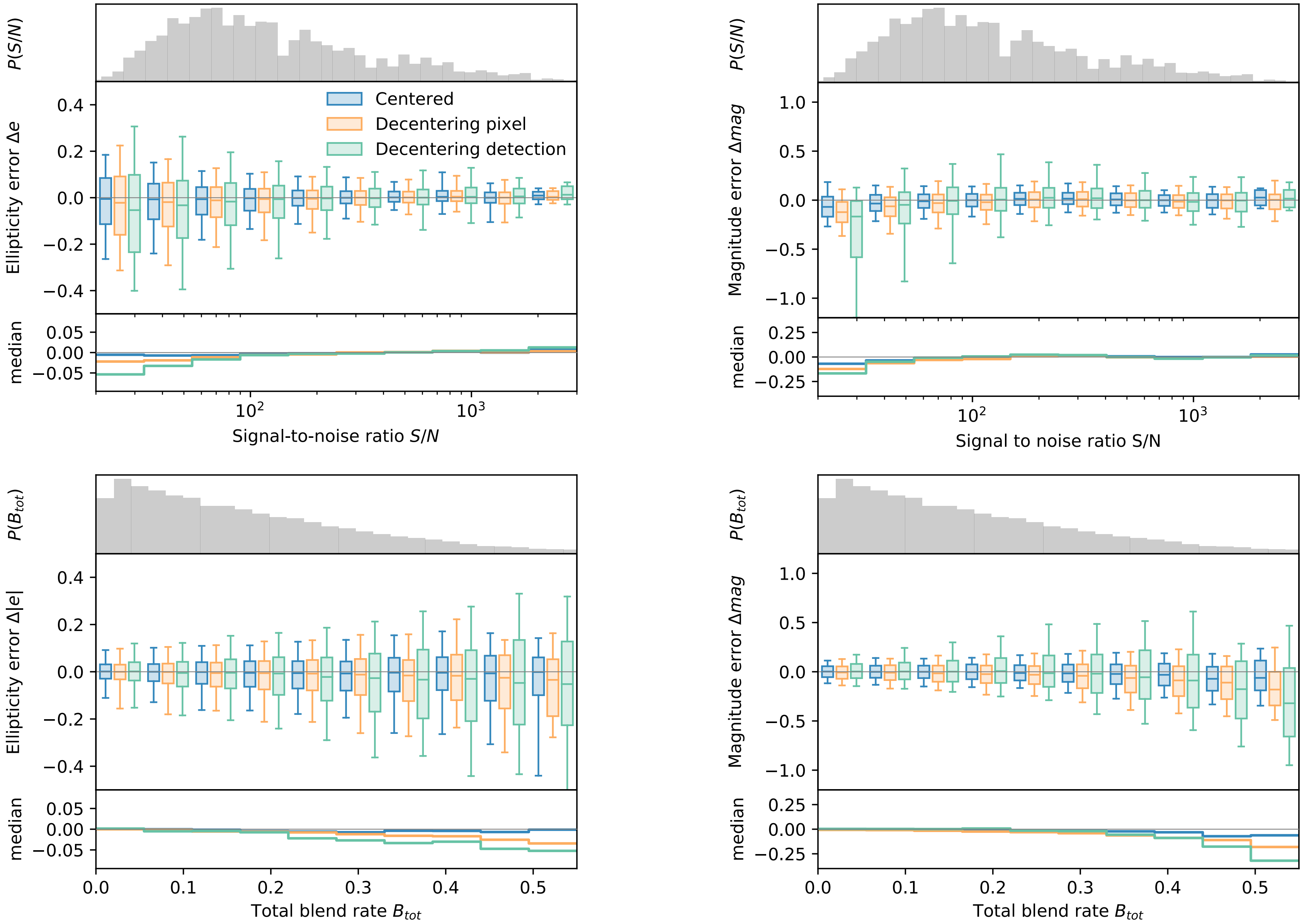


Impact of decentering

1. Perfectly centered on post stamp
2. (pixel) Uniform decentering within a pixel around center
3. (detection) Center detected with simple peak finder (in *r* only)



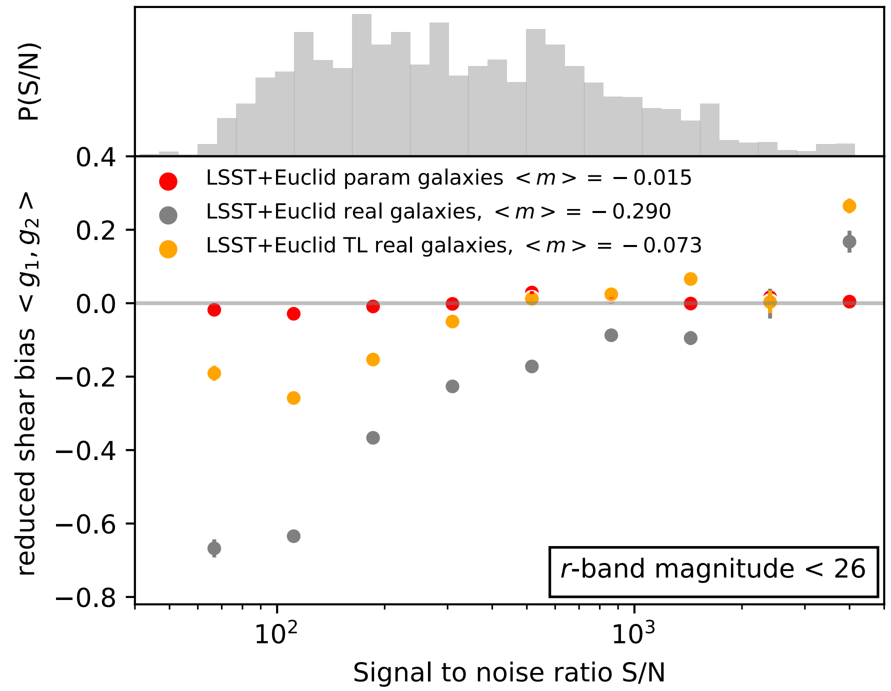
pixel detection

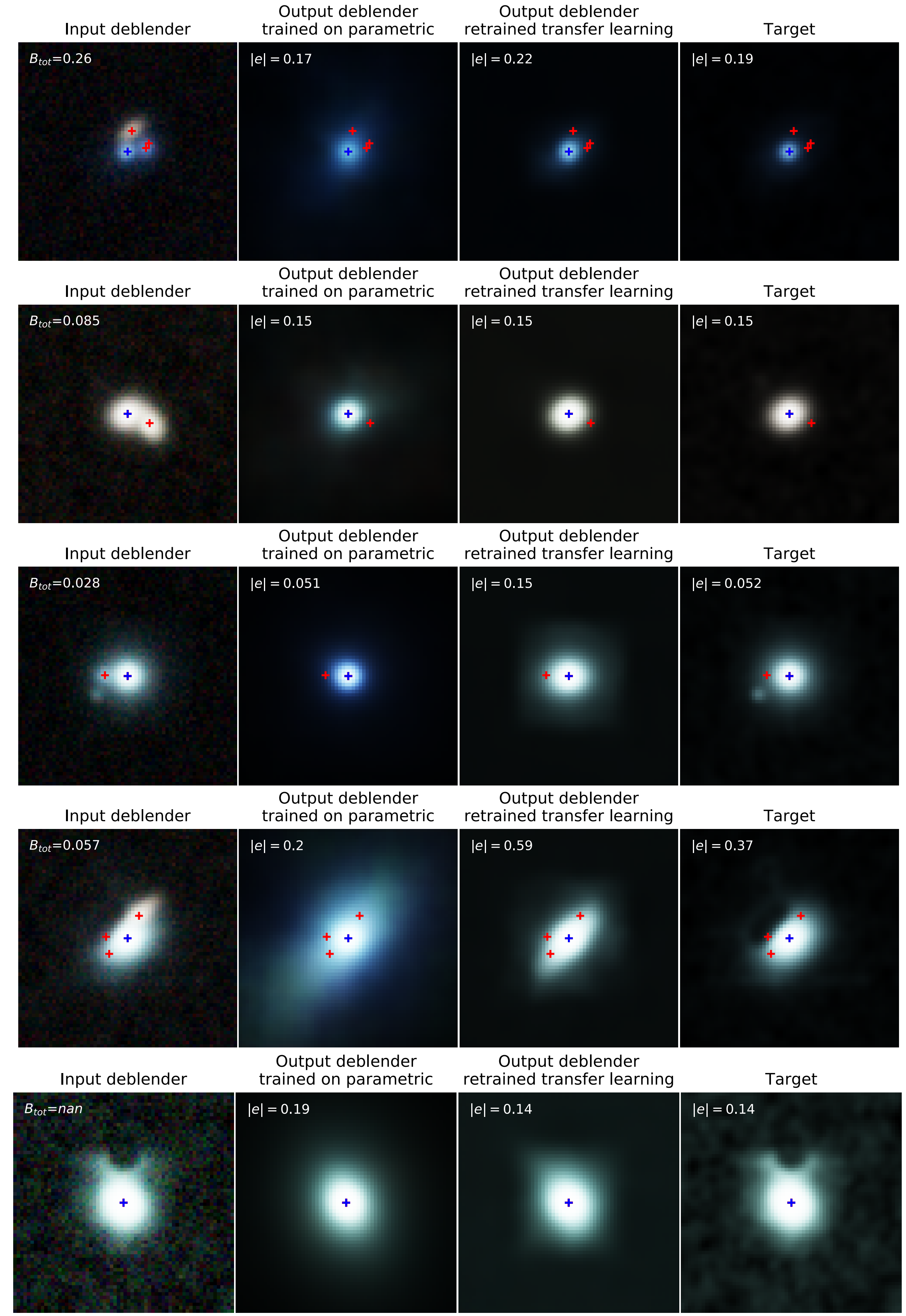


- Median errors still low
- Spread of error increase
- Shear biases degrade to 8%
- Biases (<1σ) only at very low S/N or B_{tot}>0.45

What about real data?

- **Challenges** to build a training sample
  - Clean sample of isolated galaxies?
  - Selection bias
- **Transfer learning** test
  - COSMOS i<25.2 *real* images *r*<26 with added noise
  - Clear blend and postprocessing + correlated noise
  - Shear bias divided by 2 with TL

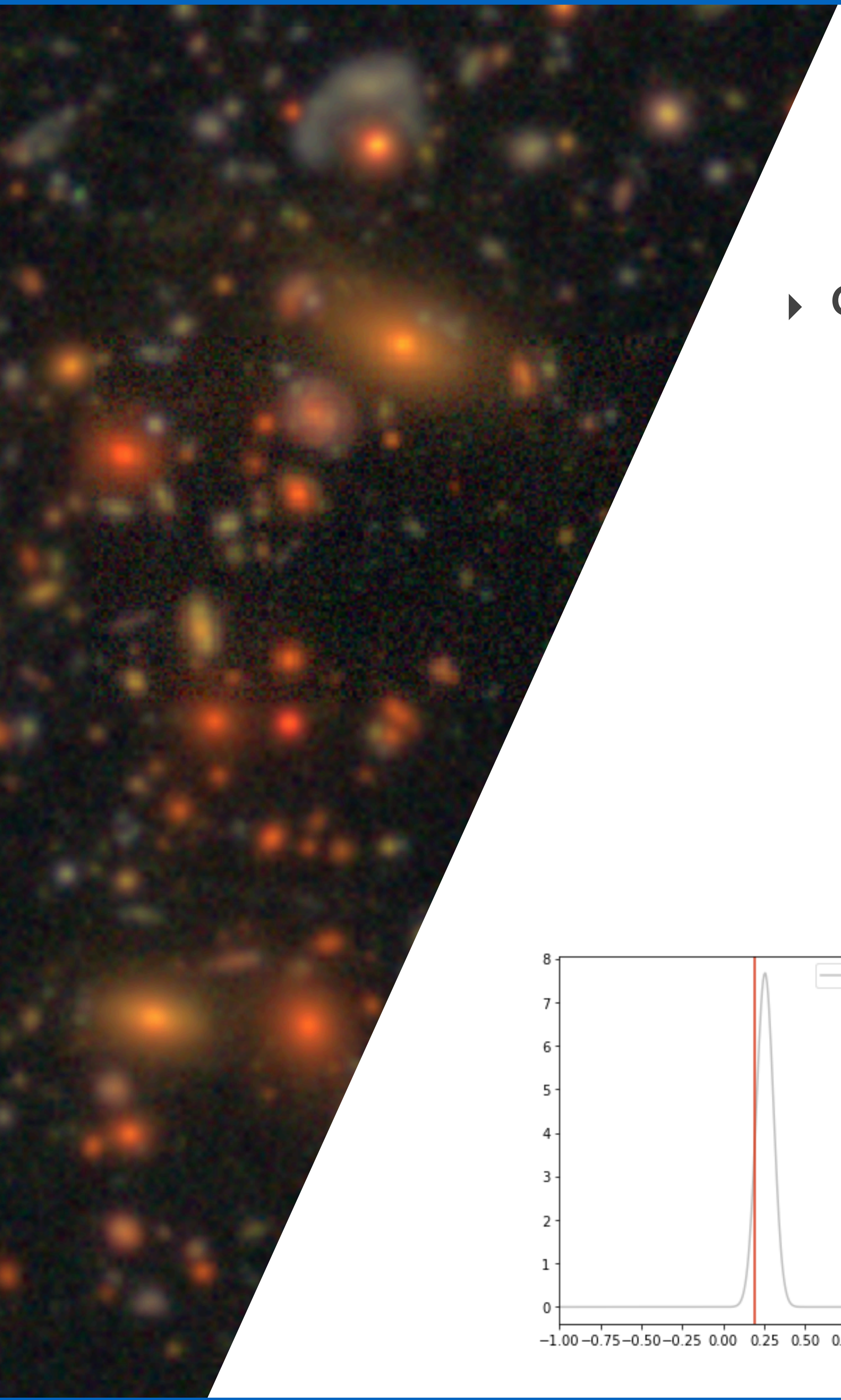




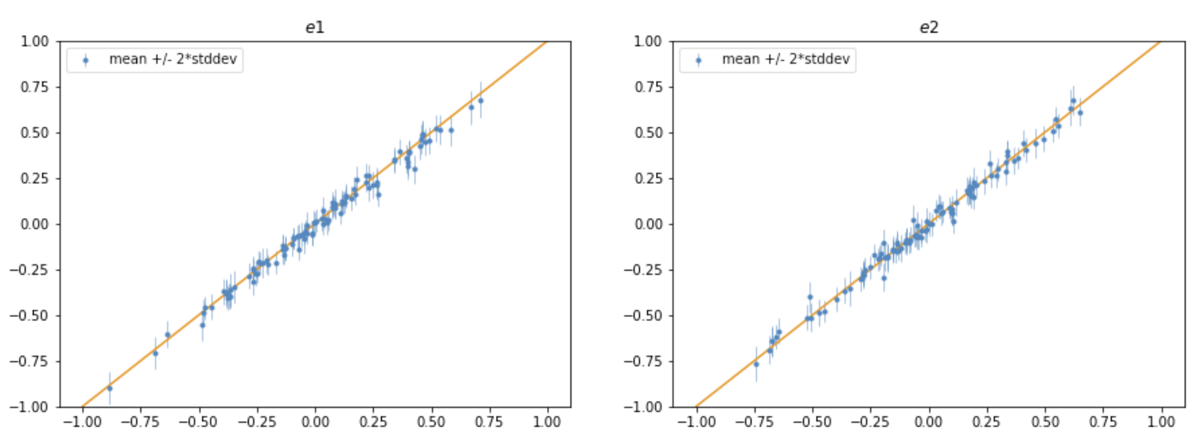
Summary

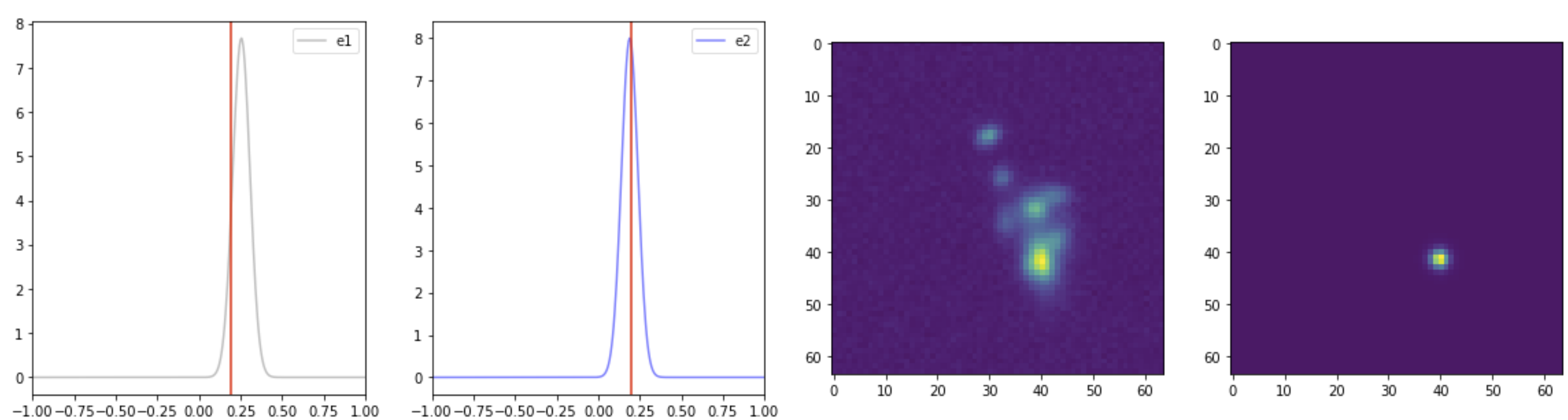
- **Deep-learning model with VAE/deblender architecture**
    - Data-driven model with CNNS minimal assumptions on morpho
    - Detection/deblending: need no info about neighbours but center → iterative
- **Extensive testing of deblender performances on simulated images**
    - Median errors on |e|<0.01 to 0.05, on mag<0.05 to 0.20
    - Ellipticity bias ~5%, shear bias 4-6% *before calibration*
    - Performances tied to detection/centering algorithm
- **Multi-band/multi-instrument** approach for LSST+Euclid
    - Significant improvement (20-30%) with Euclid VIS+NIR
- **Training with real images**
    - Encouraging results from transfer learning!

What's next?



- **Real data**
  - Simulated objects injection with HSC data
- **Go fully bayesian!**
  - Bayesian neural networks to directly obtain posteriors on shape/flux parameters, ie provide $\boldsymbol{P(e_{\mathrm{central}},z|I_{\mathrm{blended}})}$

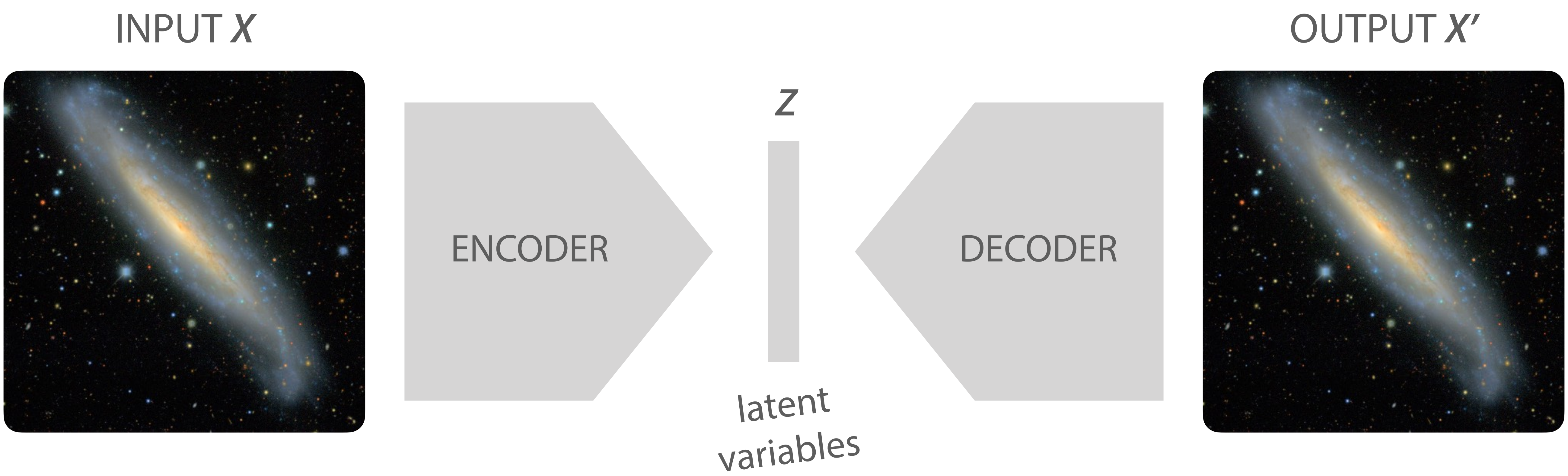




THANKS FOR LISTENING! :^)

EXTRA SLIDES

Variational autoencoder (VAE)



Variational autoencoder (VAE)

INPUT ***X***

ENCODER

$q_\phi(z|x) \simeq p_\theta(z|x)$

[posterior]

$Z \sim N(0,\mathbb{1})$

[prior]

latent variables

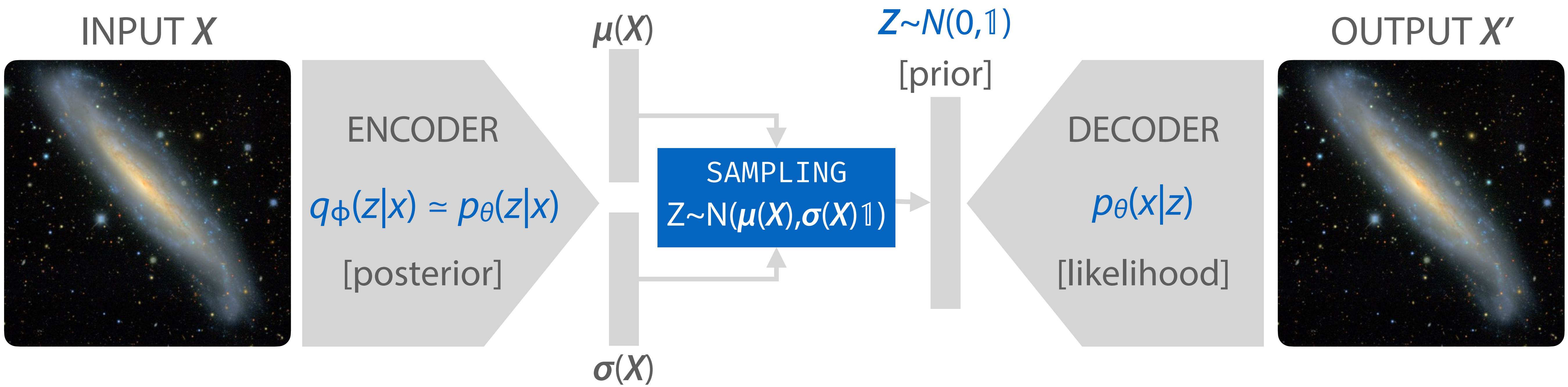
DECODER

$p_\theta(x|z)$

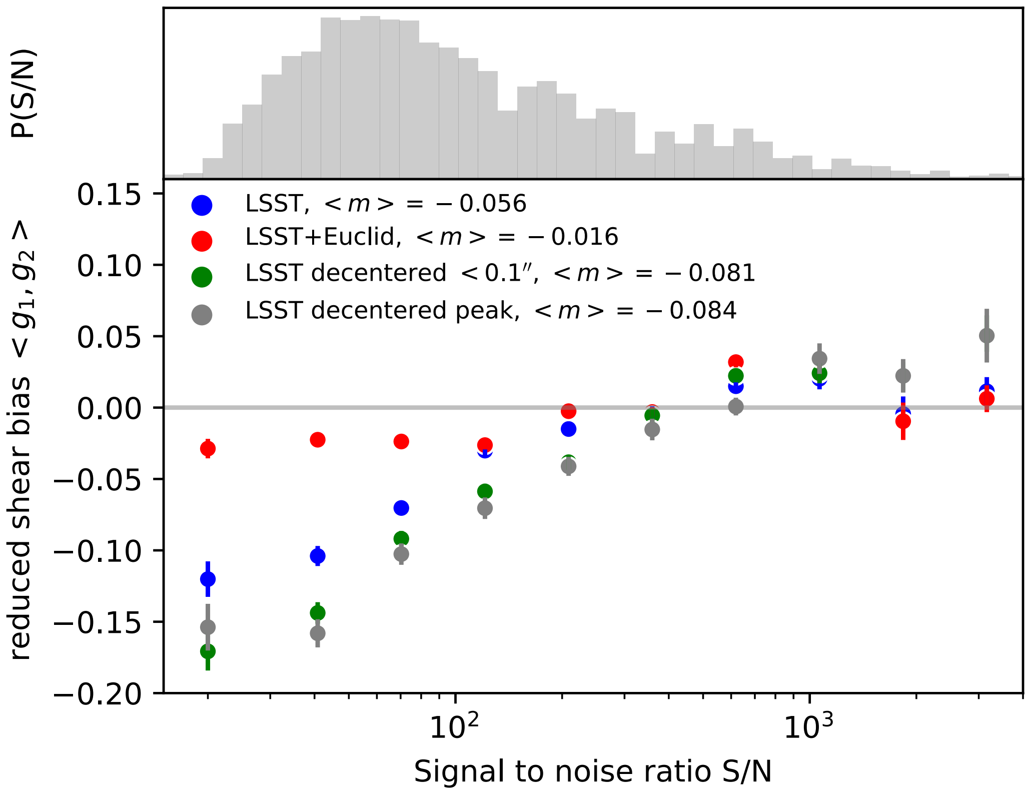
[likelihood]

OUTPUT ***X'***

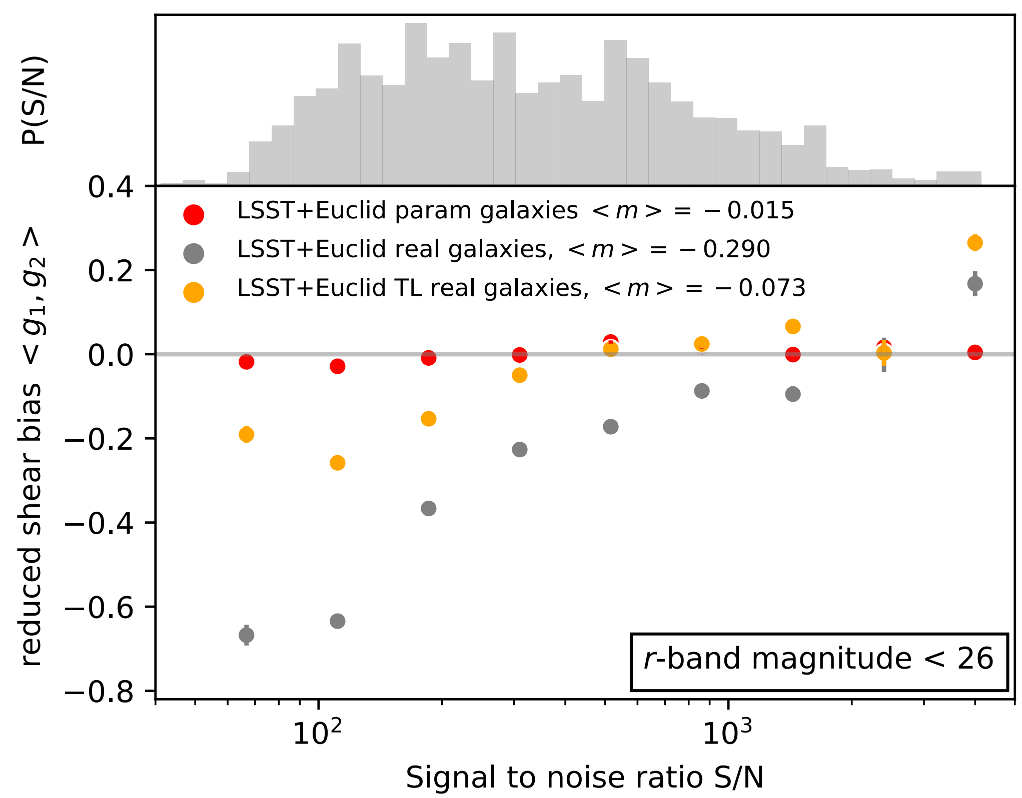
Variational autoencoder (VAE)



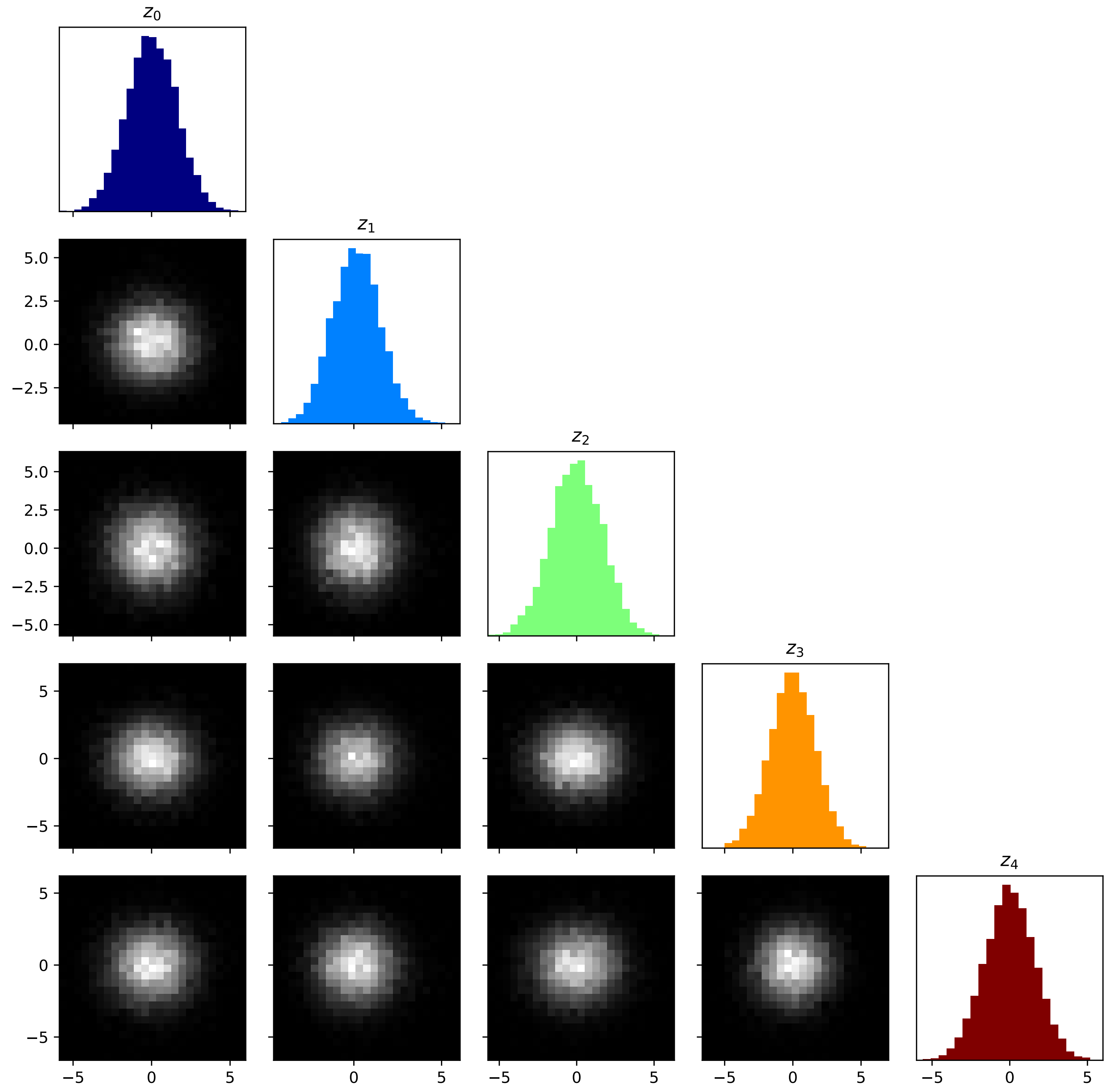
Shear bias



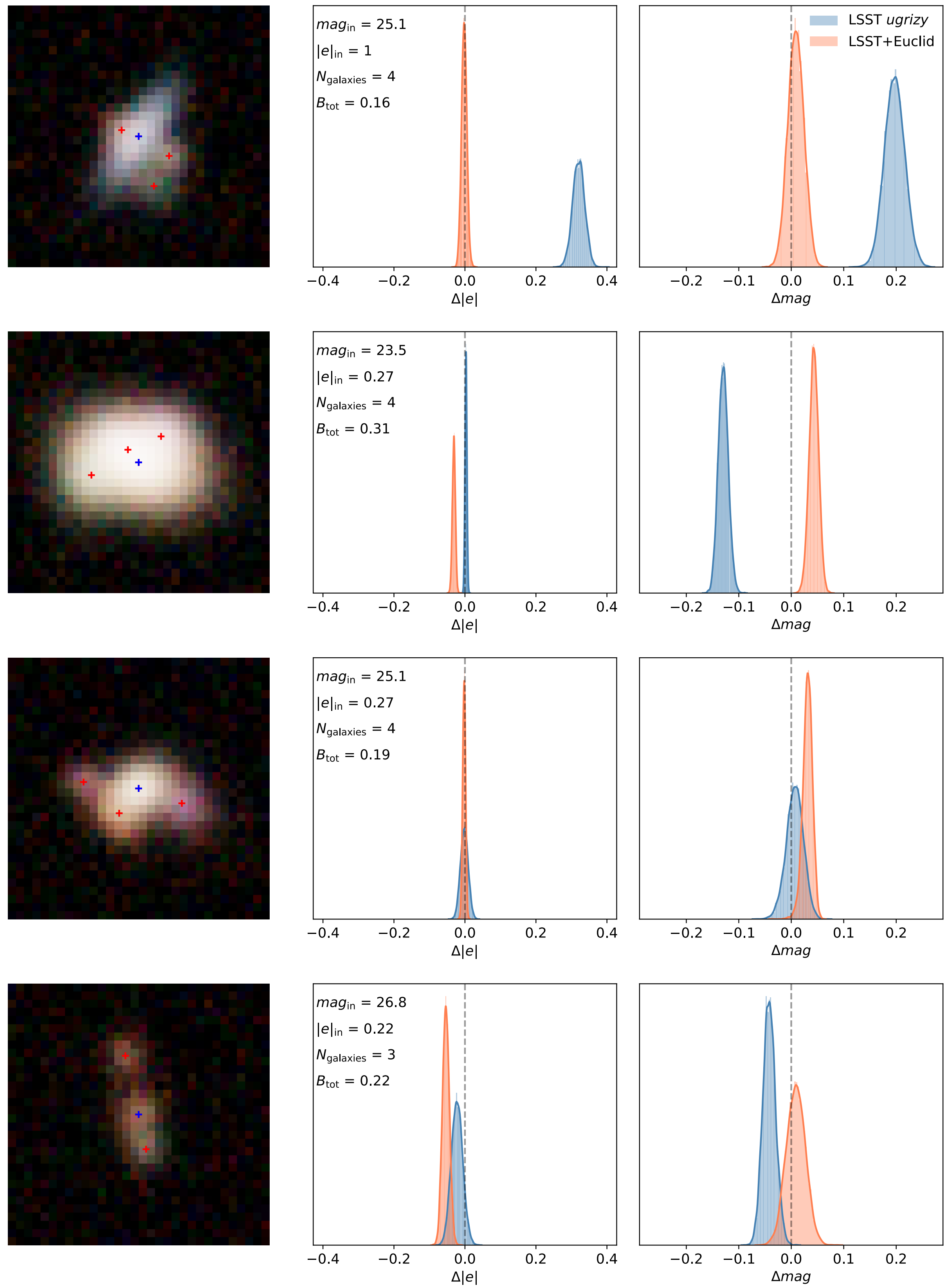
P(S/N)
reduced shear bias $<g_1, g_2>$
LSST, $<m> = -0.056$
LSST+Euclid, $<m> = -0.016$
LSST decentered $<0.1''$, $<m> = -0.081$
LSST decentered peak, $<m> = -0.084$
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
10^2
10^3
Signal to noise ratio S/N

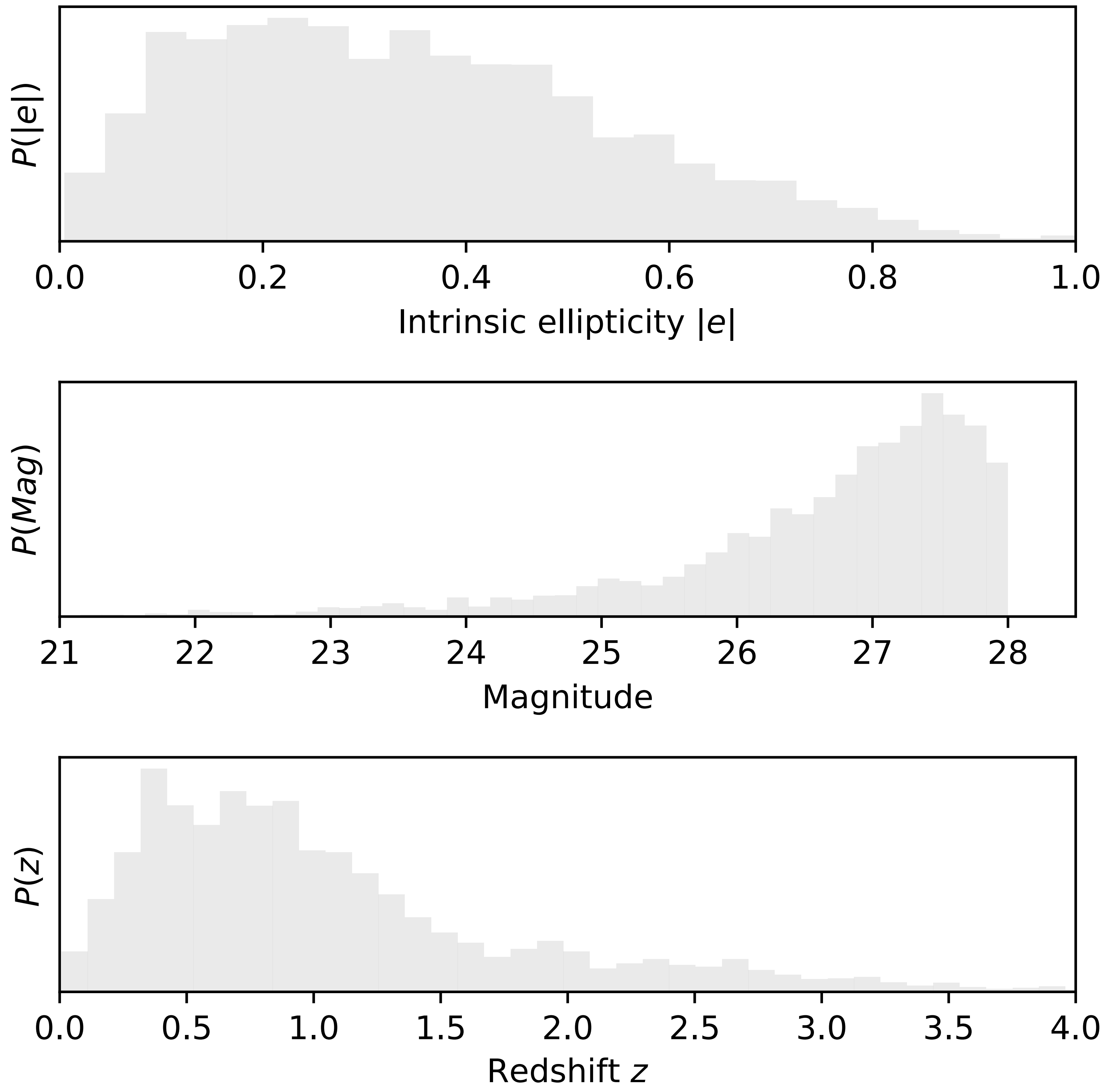


P(S/N)
reduced shear bias $<g_1, g_2>$
LSST+Euclid param galaxies $<m> = -0.015$
LSST+Euclid real galaxies, $<m> = -0.290$
LSST+Euclid TL real galaxies, $<m> = -0.073$
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
r-band magnitude < 26
10^2
10^3
Signal to noise ratio S/N

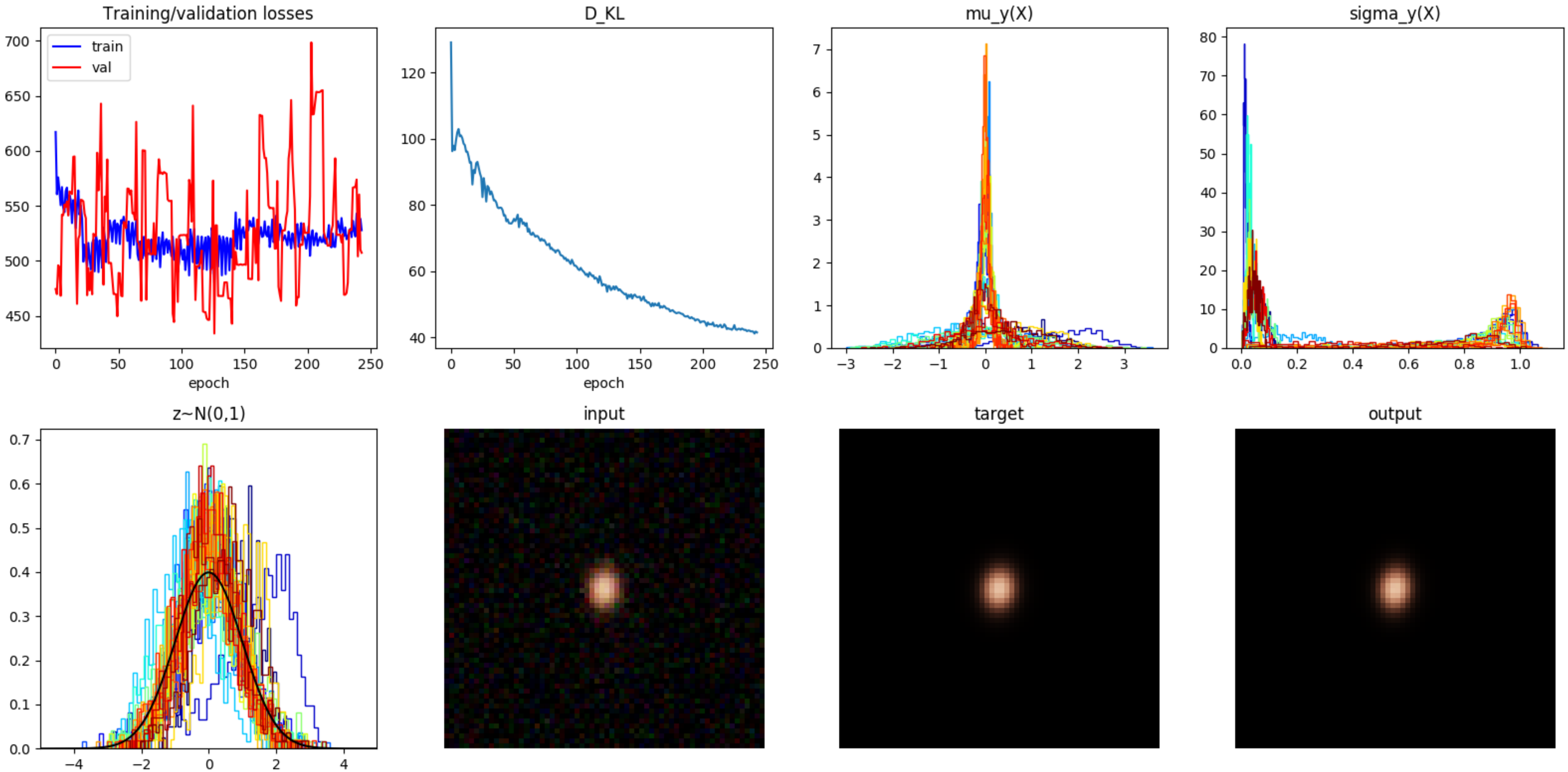

z_0
z_1
z_2
z_3
z_4
5.0
2.5
0.0
−2.5
−5.0
5
0
−5

Probabilistic output





P(|e|)
0.0
0.2
0.4
0.6
0.8
1.0
Intrinsic ellipticity |e|
P(Mag)
21
22
23
24
25
26
27
28
Magnitude
P(z)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Redshift z


Training/validation losses
train
val
epoch
D_KL
epoch
mu_y(X)
sigma_y(X)
z~N(0,1)
input
target
output