Searching for Rare Gems in Astronomy and Cosmology: Methods and Applications

Uros Seljak
UC Berkeley/LBNL

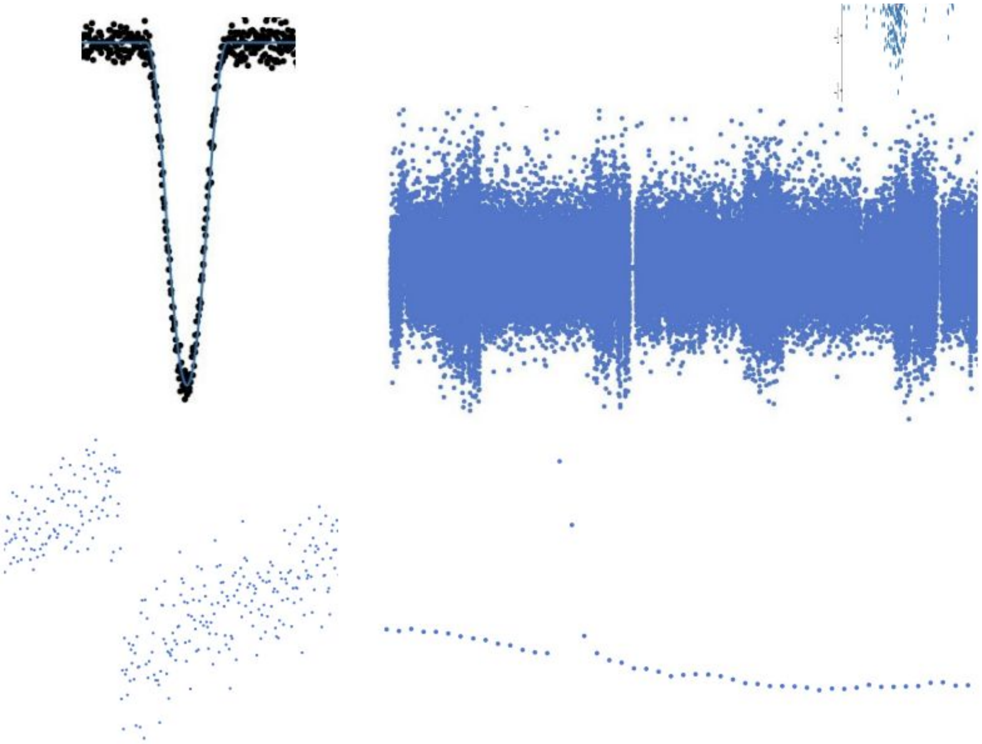
work with Biwei Dai,, Divij Sharma, Jakob Robnik, Vanessa Boehm, George Stein, Zihao Wu

Outline

- Methods:
  - Linear methods: fast and often optimal
  - Noise: correlated, often non-Gaussian
  - Look elsewhere effect: how to account for it
  - Optimal test statistic
  - The role of priors
  - Non-linear methods: dimensionality reduction (e.g. AutoEncoders)
  - Dimensionality preserving (e.g. Normalizing Flows)
  - Anomaly detection (unknown unknowns)
- Applications:
  - Searching for exoplanets and eclipsing binaries
  - Searching for binary black holes
  - Analyzing Large Scale Structure of the Universe

Exoplanet detection in Kepler data: challenges

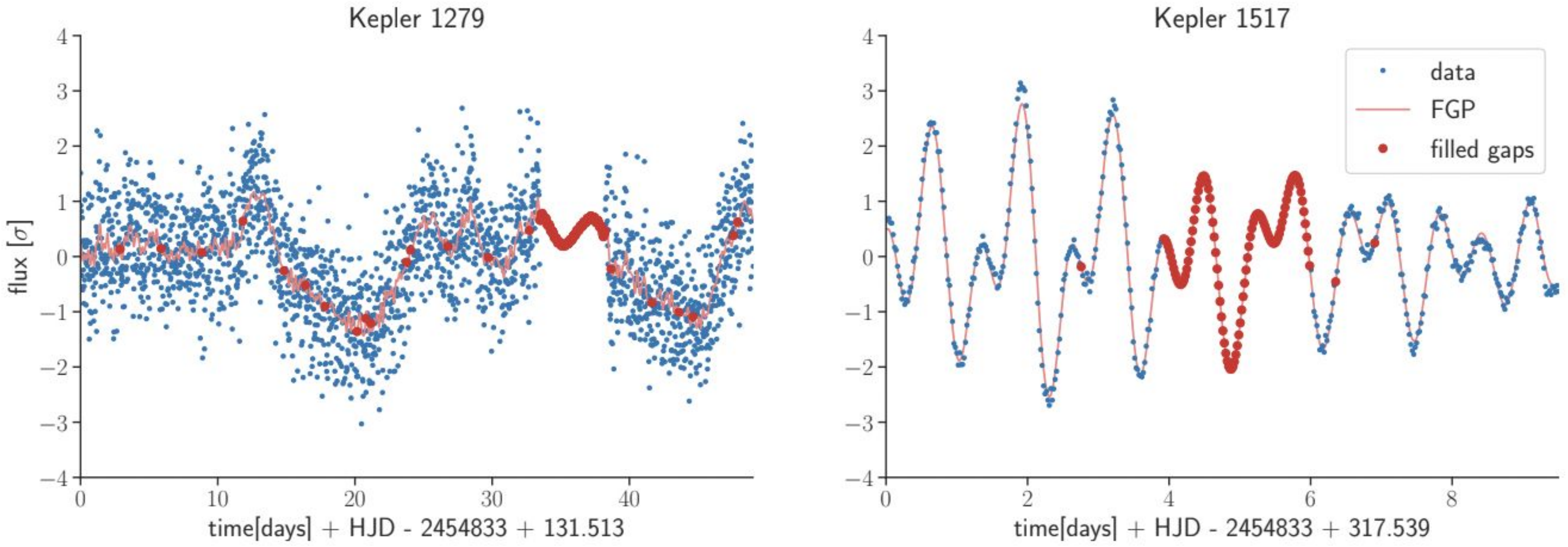
- Non-gaussian outliers
- Stellar variability
- Gaps
- Rolling bands
- Flares, drops
- Eclipsing binaries
- Third light contamination
- Unknown?



Stellar variability

Stars are variable with “red” power spectrum (a lot of power on large scales)

We have to deal with gaps in the data: inpainting



Linear methods

We are searching for a signal that is an unknown amplitude times a known time series profile (known unknown), searched over unknown period and phase using folded analysis for exoplanets

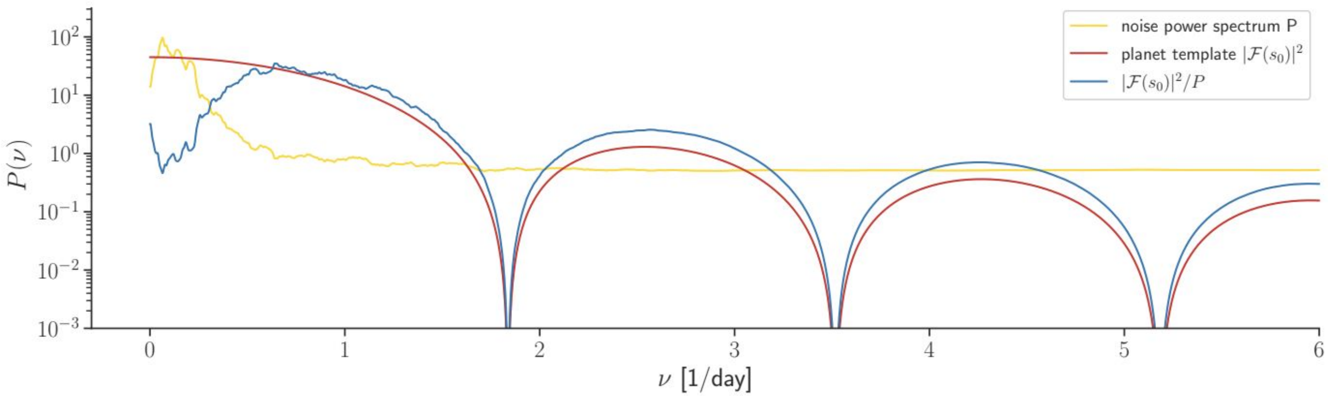
- For Gaussian noise we have an analytic solution: no optimization required, can be very fast

  This is called **matched filter**

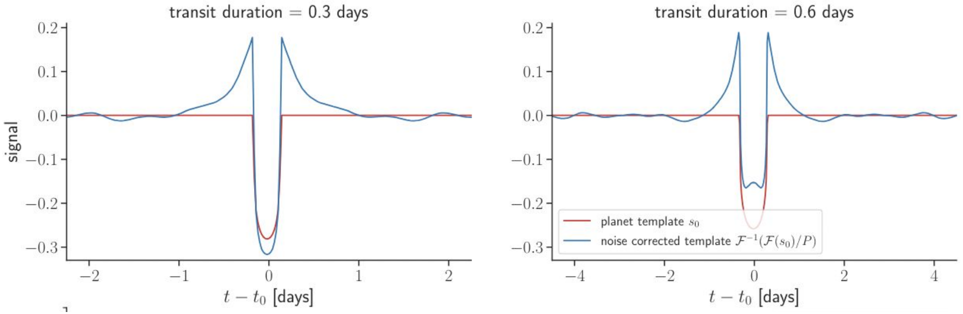
  Often we search over many templates (can be millions for gravity wave searches)

Matched filter for exoplanet detection in Kepler data

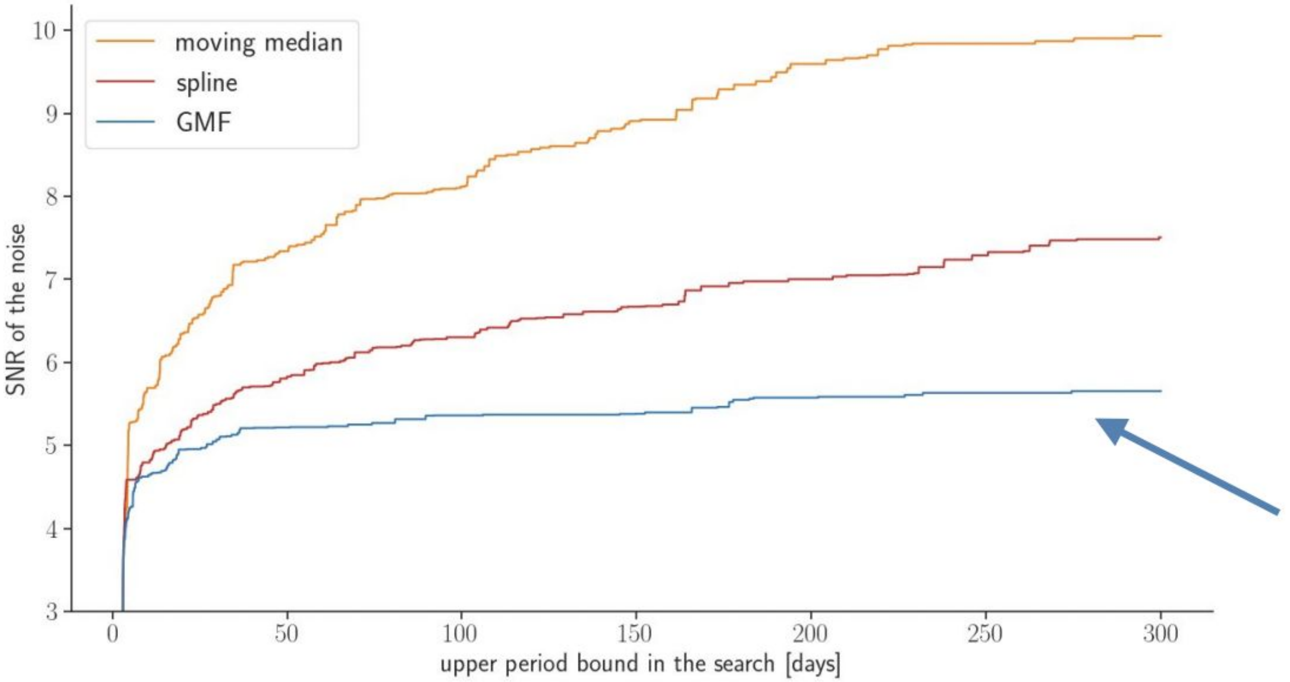
inverse noise weighting: $SNR = \mathcal{F}^{-1}\left\{\frac{\mathcal{F}\{d\}^* \, \mathcal{F}\{s\}}{\mathcal{P}}\right\}$



J. Robnik and U. Seljak. "Matched filtering with non-Gaussian noise for planet transit detections."



Does it matter? Yes, it reduces the number of false positives!

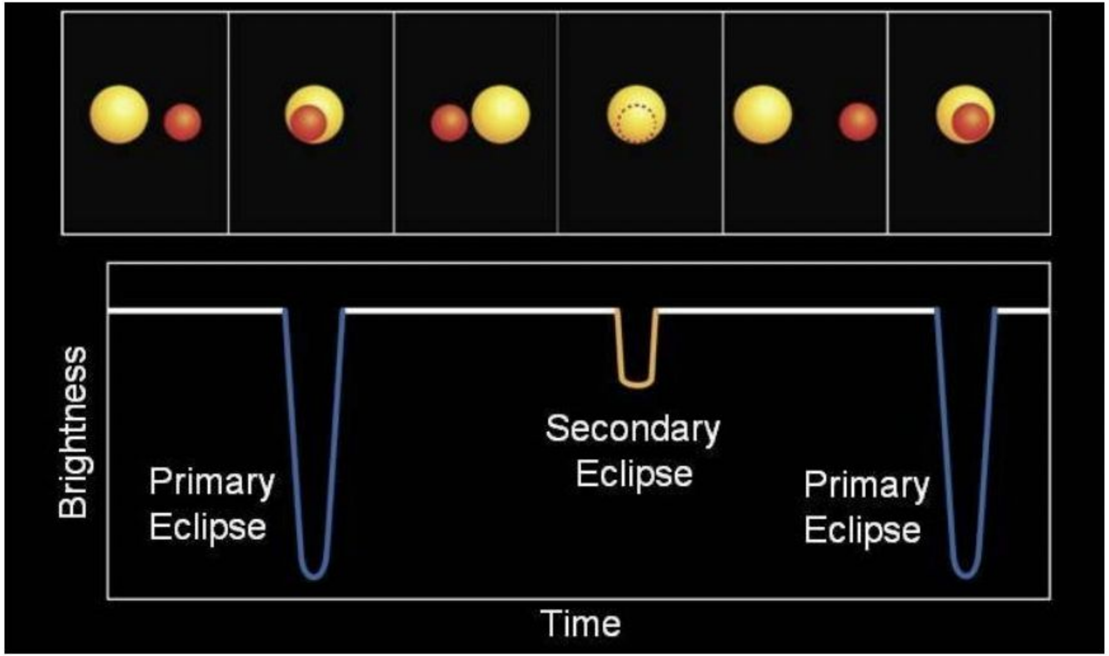


Eclipsing Binaries

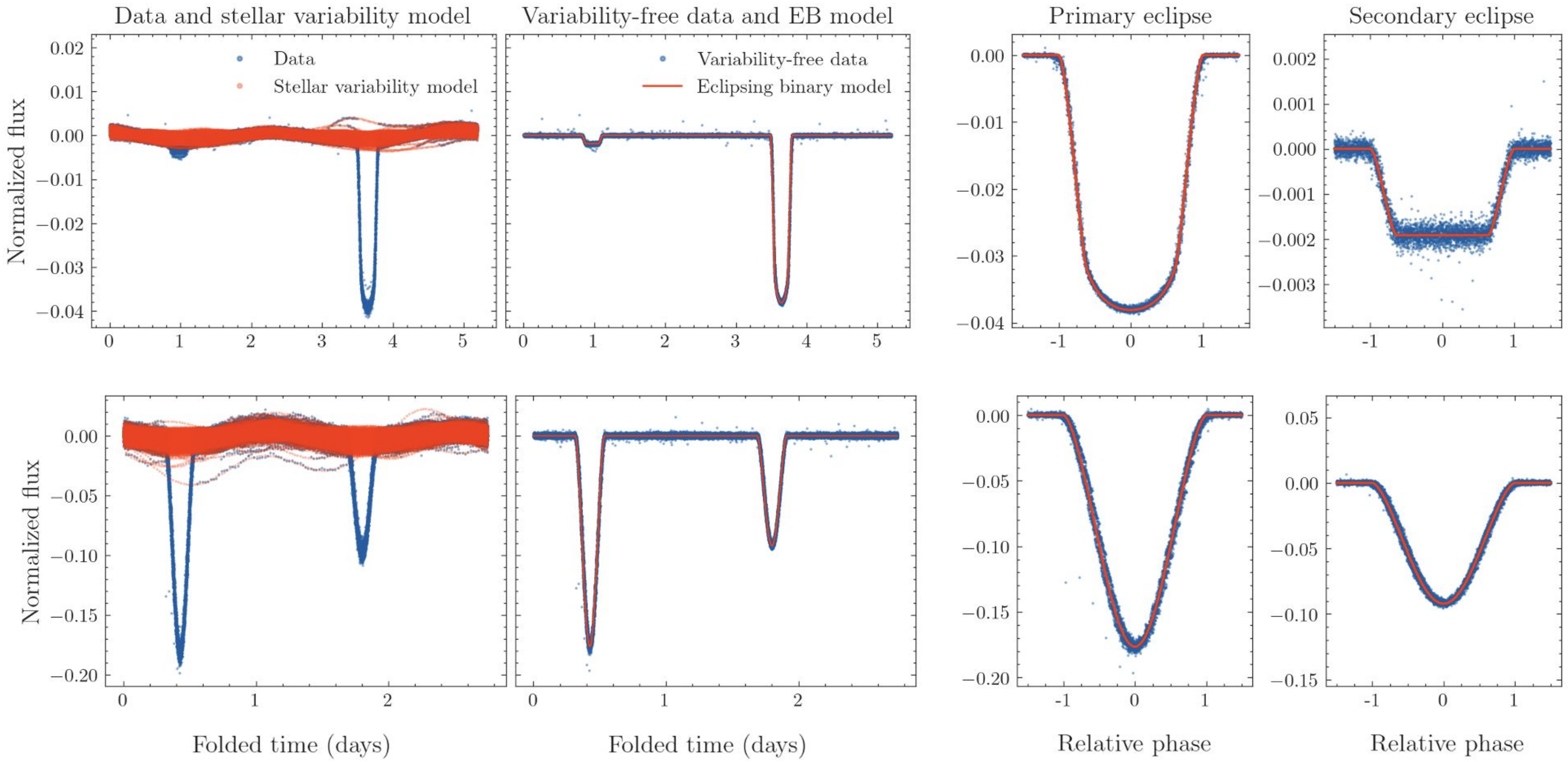
Zihao Wu

- V-shape transits
- Prior odds ← demographics of the small radius ratio eclipsing binaries
- Villanova Kepler Eclipsing Binary Catalog

B. Kirk, et al. "Kepler eclipsing binary stars. VII. The catalog of eclipsing binaries found in the entire Kepler data set." The Astronomical Journal 151.3 (2016): 68.



Wikimedia, NASA



Data and stellar variability model
Variability-free data and EB model
Primary eclipse
Secondary eclipse
Data
Stellar variability model
Variability-free data
Eclipsing binary model
Normalized flux
Normalized flux
Folded time (days)
Folded time (days)
Relative phase
Relative phase

How do we distinguish between exoplanets and eclipsing binaries?

Bayes Factor: ratio of evidences for the two hypotheses

What is **Bayes evidence**: it combines the quality of the fit with the trials factor (Occam's razor, Look Elsewhere effect)

What is **trials factor**? If you try to detect something and you try it many times you need to account for the fact that it can happen by chance

Typically we scan over the prior of the parametrization of the hypothesis: e.g. period, phase, amplitude, transit duration for exoplanets

We developed a new parametrizations for eclipsing binaries

Each time we move by one sigma in each of the parameters we incur a new trials factor

This can be very large (100 million!) for exoplanets where we scan over periods of years, but the error on period and phase is minutes

Bayes factor between null hypothesis and signal

Bayes factor (expensive to compute it) is also useful to quantify the false positive rate (frequency of pure noise events at high SNR), but can be misleading if the noise properties are poorly understood (e.g. non-Gaussian noise)

Even then Bayes Factor can be a powerful test statistic (optimal if the priors are chosen well)

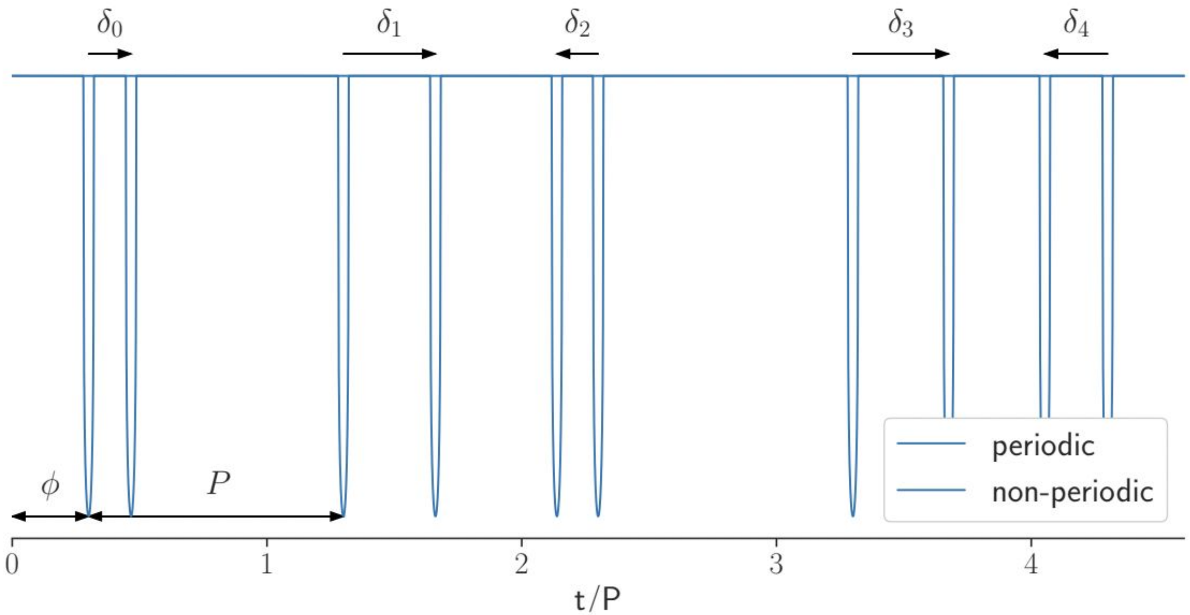
This is important since for SNR test statistic we may have false positive contamination

For example: maybe true signal is lurking at low exoplanet periods, but long periods have larger trials factor and hence produce more false positives at larger SNR: Bayes factor corrects for this

How to quantify false positive rate if you do not have reliable simulations?

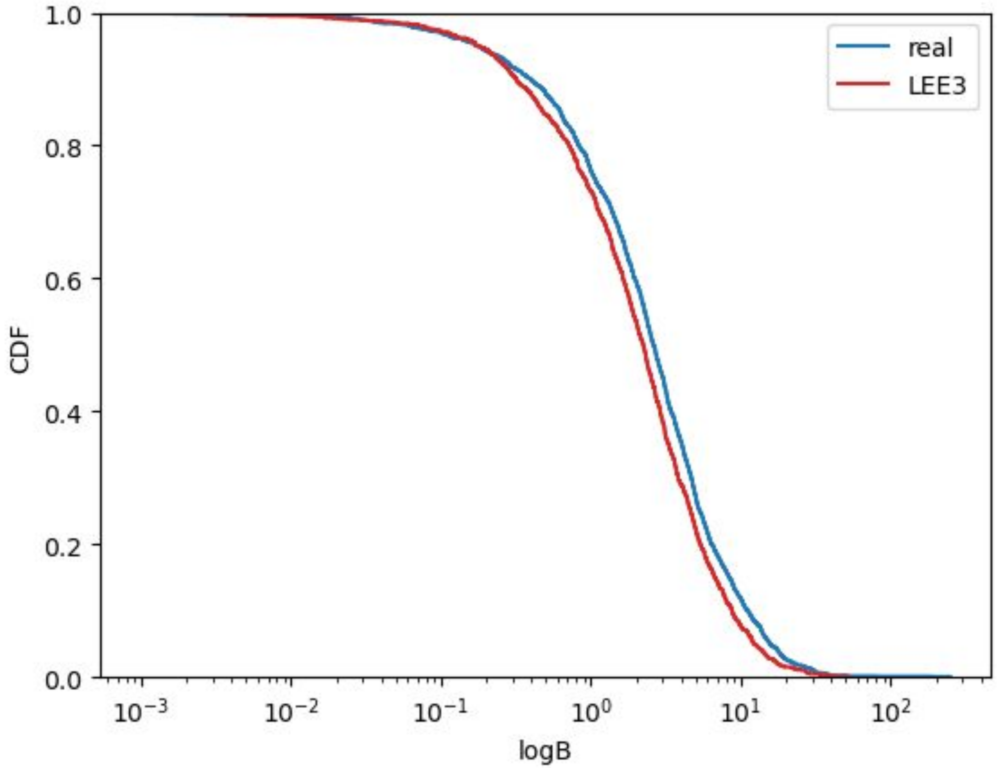
We (Robnik & Seljak, in prep) developed a new method that gives the same false positive rate as the main search, but eliminates the exoplanet signal

On simulations it gives same FPR as periodic signals



Application to Kepler data

We see a slight excess in the real signal: we can statistically quantify the excess in the regime where individual detections are not possible (important for demographics of habitable zone planets, work in progress)



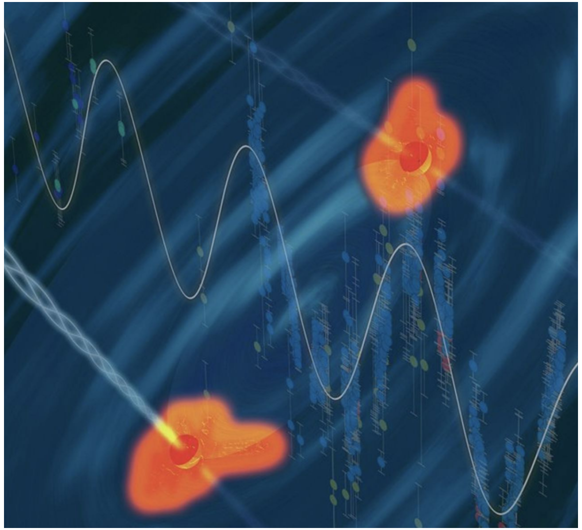
Supermassive Black hole binaries with periodograms in quasar variability data

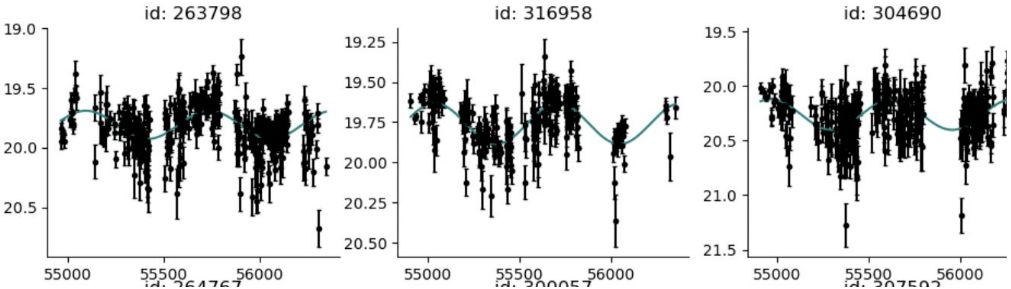
Several groups (e.g. Graham etal, Charisi etal) have claimed a detection of the SMBHB signal (PTF, Catalina)

Problem: false positive rate is quantified using Gaussian correlated noise

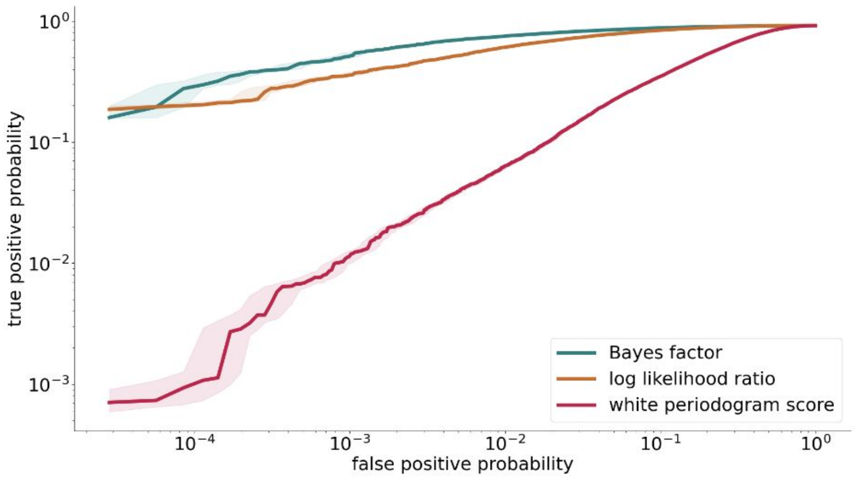
Problem: SNR is not computed using matched filter inverse noise weighting

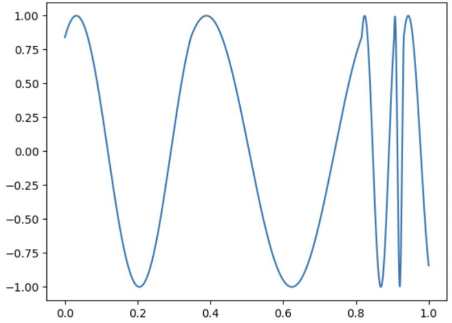
Problem: data sampling very uneven, observed periods are long





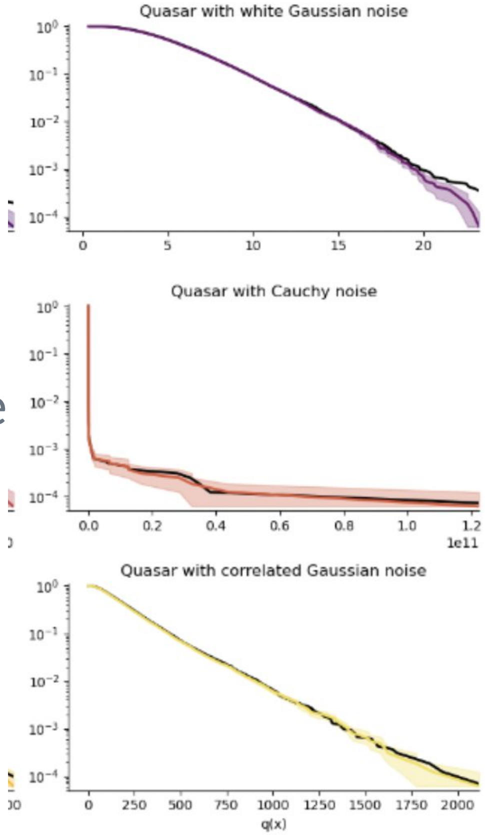
SMBHB Bayes factor has best ROC



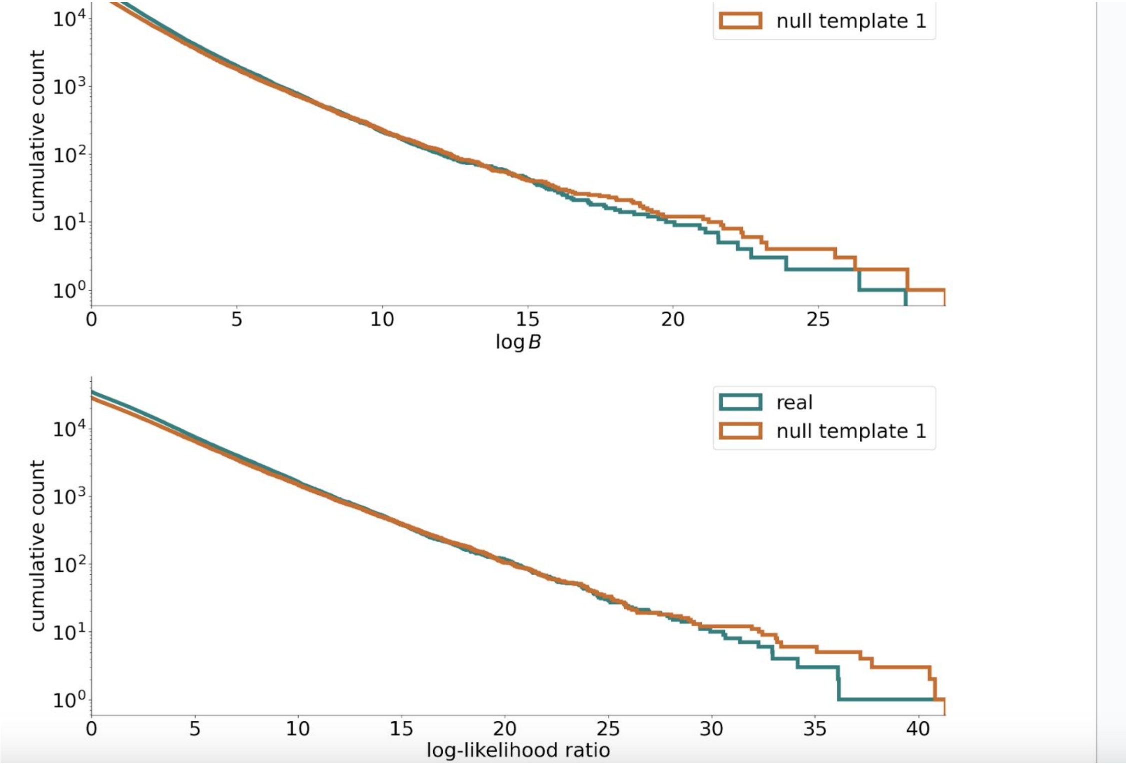


Classic periodogram is almost useless here because of QSO variability

We have modifie the template to eliminate signal



Application to PTF data (preliminary!)



No evidence of SMBHB signal!

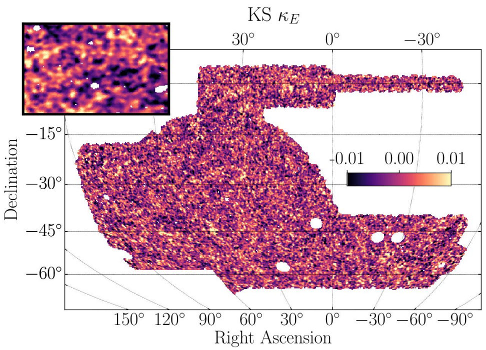
Lessons learned

Searching for rare gems is hard:

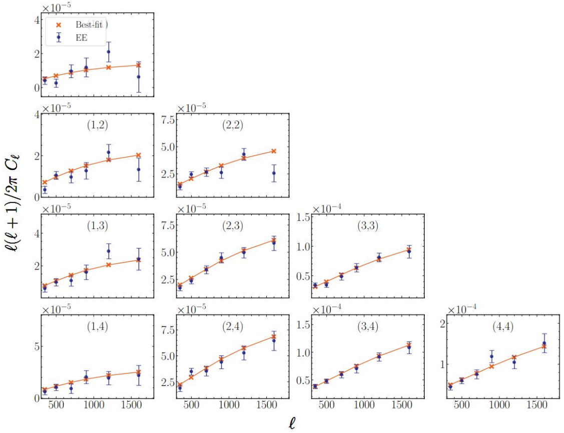
1) Account for Look Elsewhere Effect (trials factor): how many trials have you performed?
2) Estimate priors and ideally to use Bayes Factor as a test statistic even if you use frequentist methods to quantify the false positive rate
3) Use the data directly as a noise simulator to quantify the false positives
4) Try linear methods before doing nonlinear ML methods
5) Bayes Factor search with matched filters is doable even for Rubin SMBHB and Kepler/TESS exoplanets

Cosmological analysis based on summary statistics

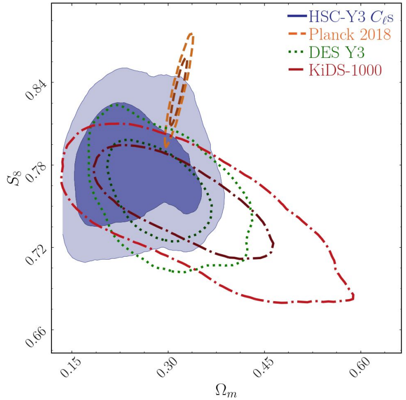
data x



summary statistics S



cosmological parameter y



Compressing data x to summary statistics S

Gaussian likelihood on summary statistics S: p(S|y)

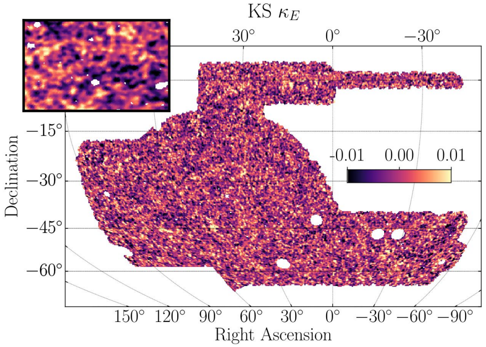
- Cosmological analysis based on two-point summary statistics: p(S|y) ⟶ p(y|S) = p(S|y)p(y)/p(S)
  - For non-gaussian data, usually leads to **information loss**

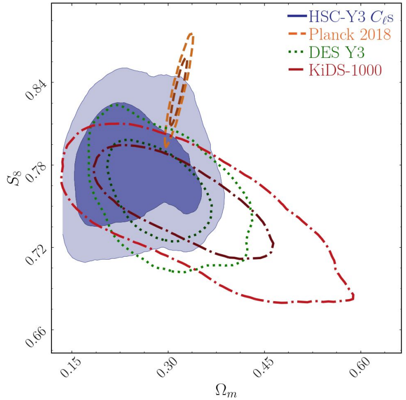
Credit: Jeffrey et al. 2021, Dalal et al. 2023

Field-level cosmological inference

cosmological parameter y

data x

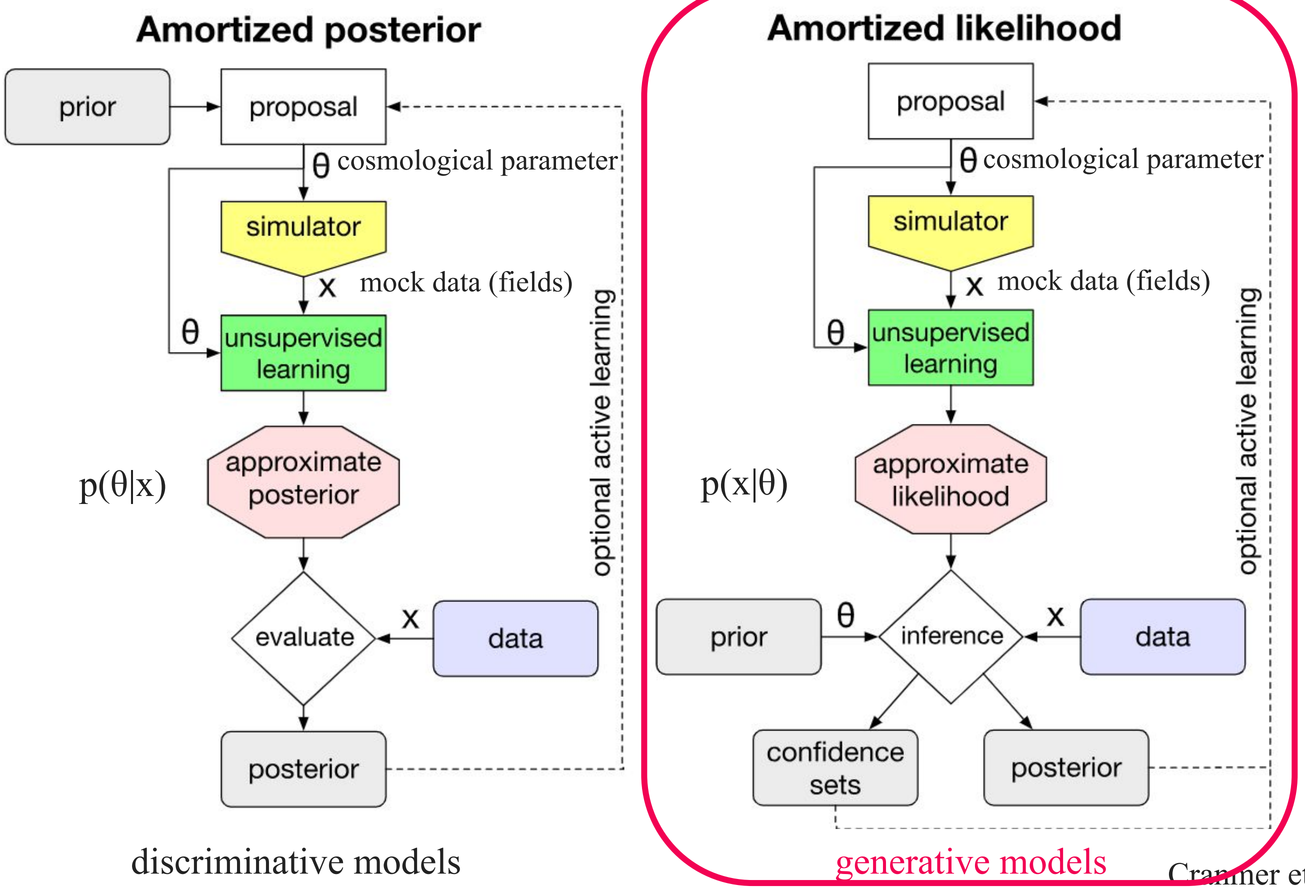




- Field-level inference
  - Pro: **No information loss** due to data compression.
  - Deep learning allows us to directly extract information at the field level (simulation-based inference)

Credit: Jeffrey et al. 2021, Dalal et al. 2023

Simulation Based Inference (SBI)



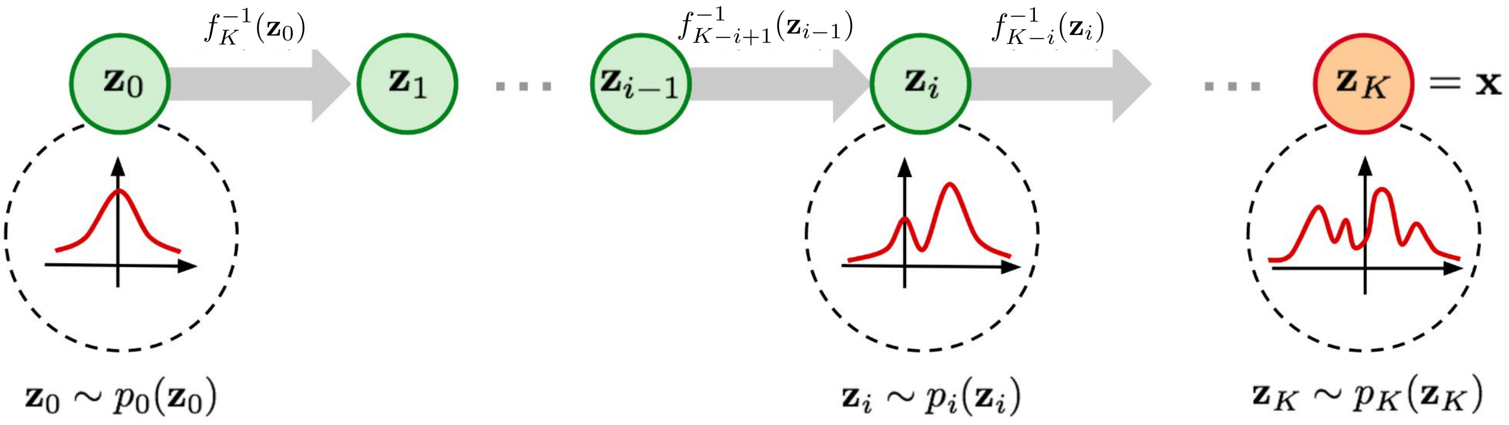
Cranmer et al. 2020

Green box: machine learning models (normalizing flows) that take in $\{x,\theta\}_i$ pairs and estimate $p(x|\theta)$ or $p(\theta|x)$.

Potential issues of SBI:

1. The simulations may not be accurate (distribution shift)
2. The ML model is a black box and lacks interpretability

Normalizing Flows



Credit: https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html

- Bijective mapping f between data x and latent variable z ($z = f(x)$, $z \sim \pi(z)$)
  - **Evaluate density**: $p(x) = \pi(f(x))\, |\det(df/dx)|$
  - **Sample**: $x = f^{-1}(z)$ ($z \sim \pi(z)$)

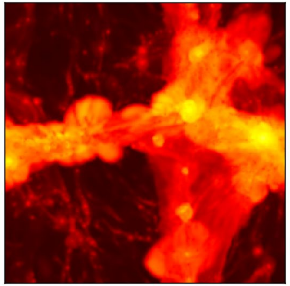
What can Normalizing Flows do for Astronomy?

Normalizing flows provide a powerful framework for high-dimensional density estimation (likelihood) and sampling

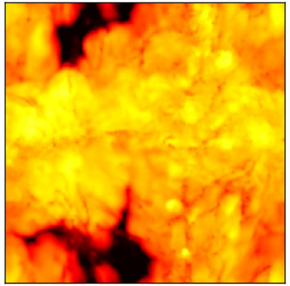
- Extract physical information (simulation-based inference)
- Fast sample generation
- Anomaly detection
  - Detect systematic effect (distribution shift)
  - Search for new physics/asrophysics

Test 1: Goodness-of-fit test / Out-of-distribution detection

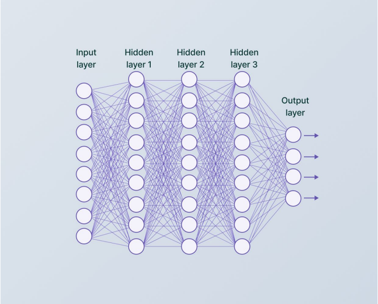
Training simulations

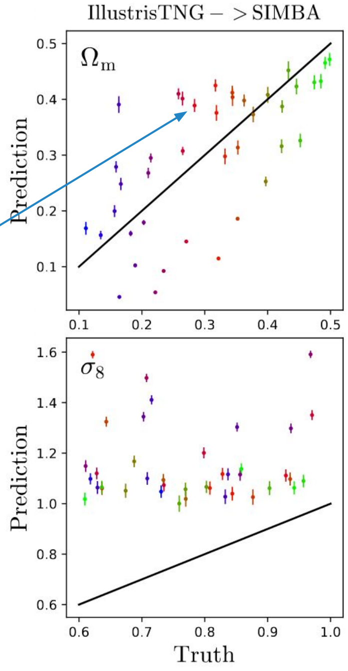


Test data / observation



Discriminative models

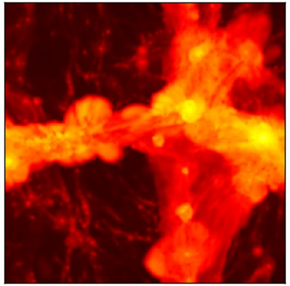




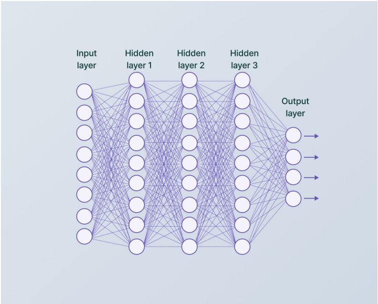
Biased parameter constraints due to distribution shifts, and we don’t know it!

Test 1: Goodness-of-fit test / Out-of-distribution detection

Training simulations

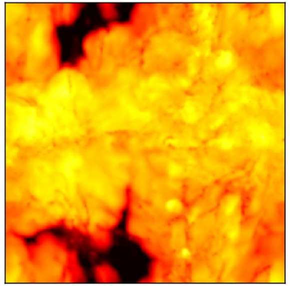


Generative models



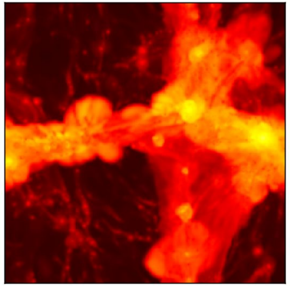
likelihood p(x|y)

Test data / observation

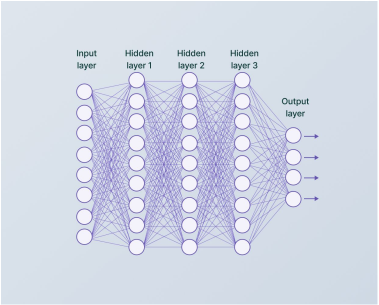


Test 1: Goodness-of-fit test / Out-of-distribution detection

Training simulations

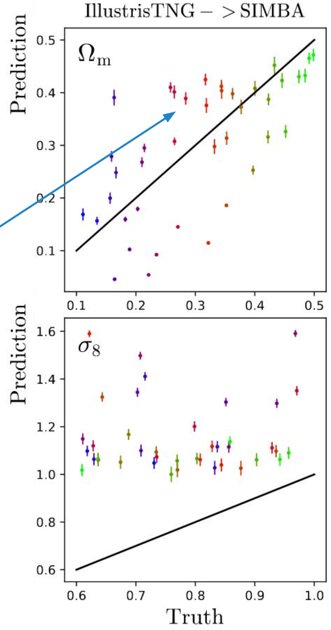


Generative models

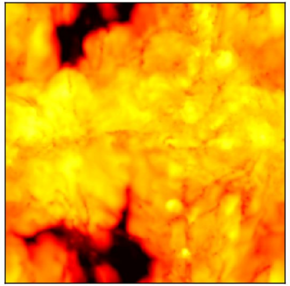


likelihood p(x|y)

MCMC

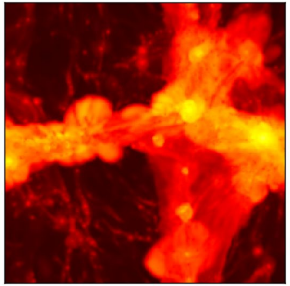


Test data / observation

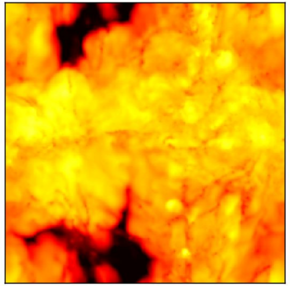


Test 1: Goodness-of-fit test / Out-of-distribution detection

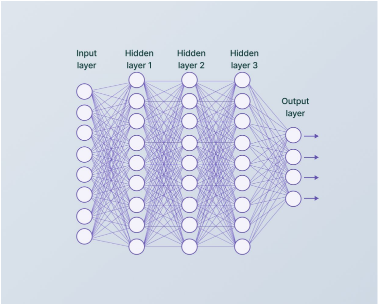
Training simulations



Test data / observation



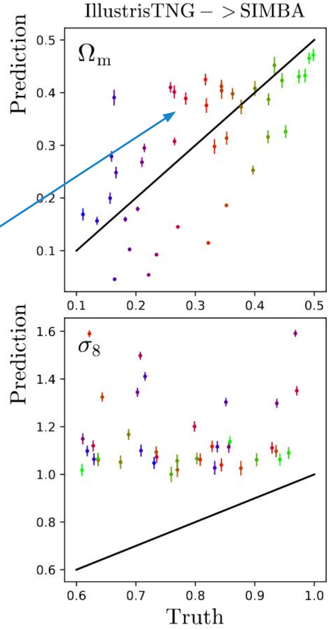
Generative models



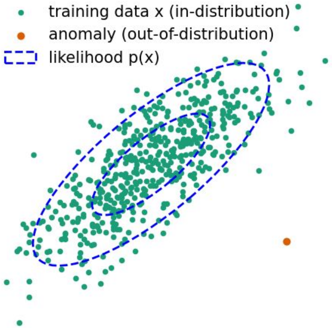
likelihood p(x|y)

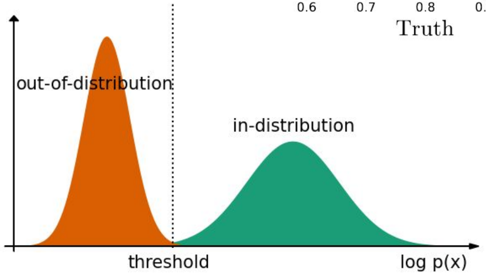
Generative NF models enable goodness-of-fit test to improve the robustness of analysis.

MCMC



The test data / observation doesn't look like training data, so we shouldn't trust our analysis!





Multiscale consistency test with Multiscale Flow

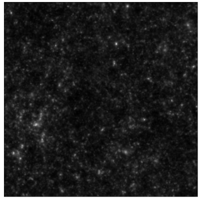
- Motivation: Multiscale analysis for robust constraints
  - Different scales are governed by different physics / systematics: the numerical / astrophysical effects normally happens on small scales, and PSF may influence very large scales
  - Separate and compare the information (likelihood) of different scales, and identify the part of the data that is contaminated by systematics

- Wavelet decomposition: recursively apply low-pass filters (scaling functions) and high-pass filters (wavelet functions) to the data. In each iteration, the data x_n with resolution 2^n is decomposed into a low-resolution approximation x_{n-1}, and detail coefficients of the remaining signal $x_{n-1,extra}$

Dai, B. and Seljak, U., 2023. Multiscale Flow for Robust and Optimal Cosmological Analysis. *arXiv preprint arXiv:2306.04689*.

Multiscale flow

- Consider a cosmological field with 256^2 resolution:

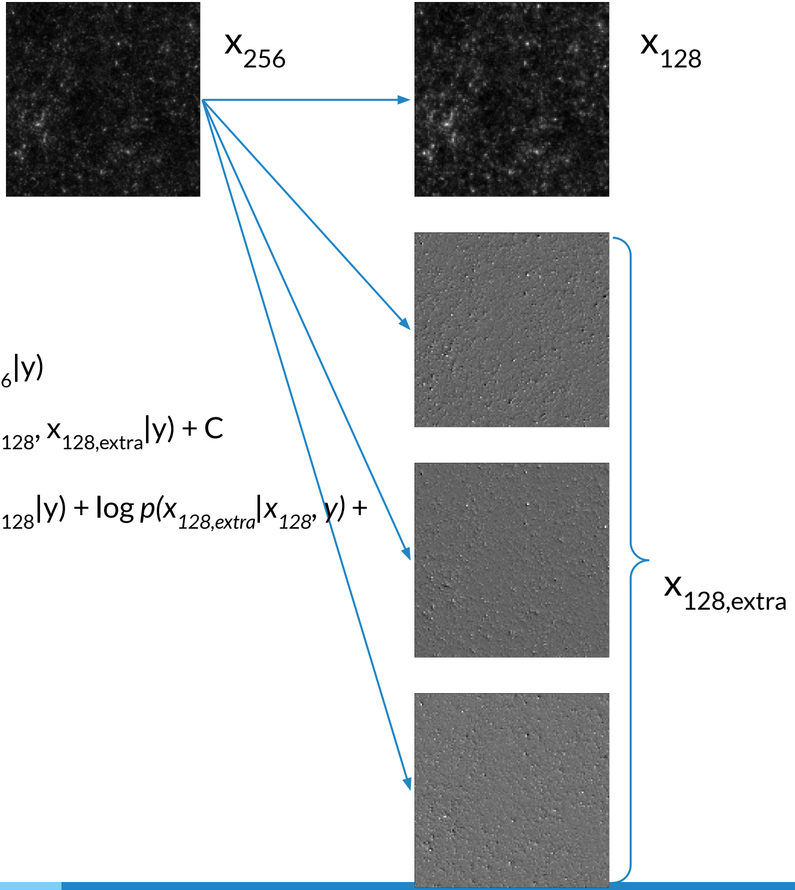


x_{256}

$\log p(x_{256}|y)$

Multiscale flow

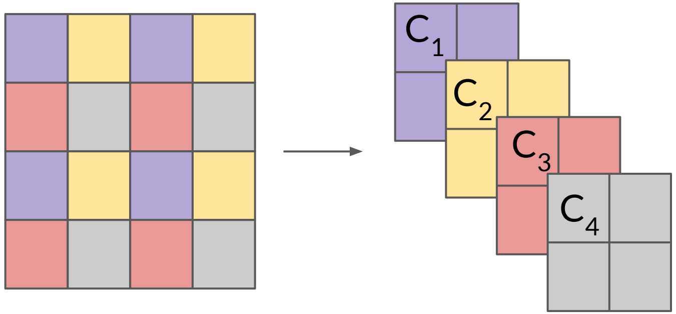
- Consider a cosmological field with 256^2 resolution:



log $p(x_{256}|y)$

$= \log p(x_{128}, x_{128,extra}|y) + C$

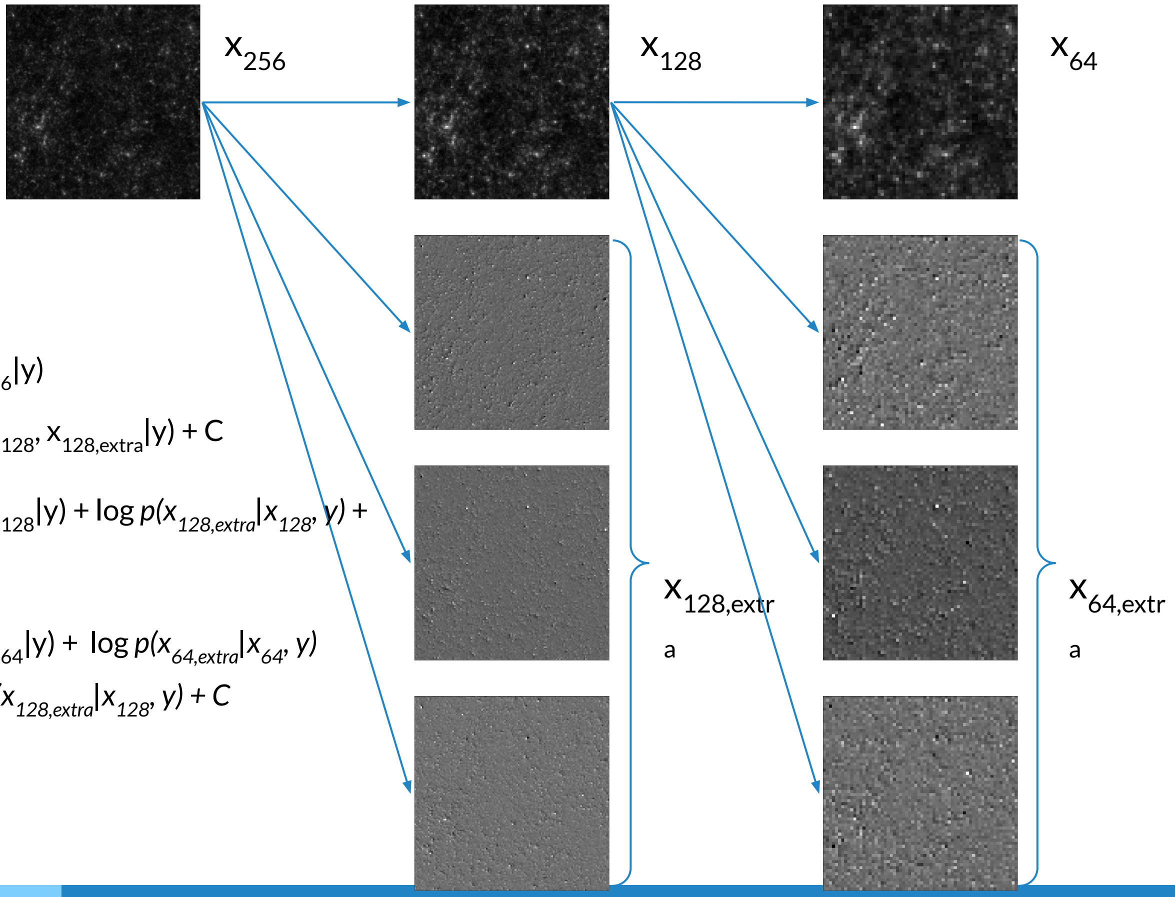
$= \log p(x_{128}|y) + \log p(x_{128,extra}|x_{128}, y) + C$



$$\begin{bmatrix} x_{128} \\ x^1_{128,\text{extra}} \\ x^2_{128,\text{extra}} \\ x^3_{128,\text{extra}} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{bmatrix}$$

Multiscale flow

- Consider a cosmological field with 256^2 resolution:



$\log p(x_{256}|y)$

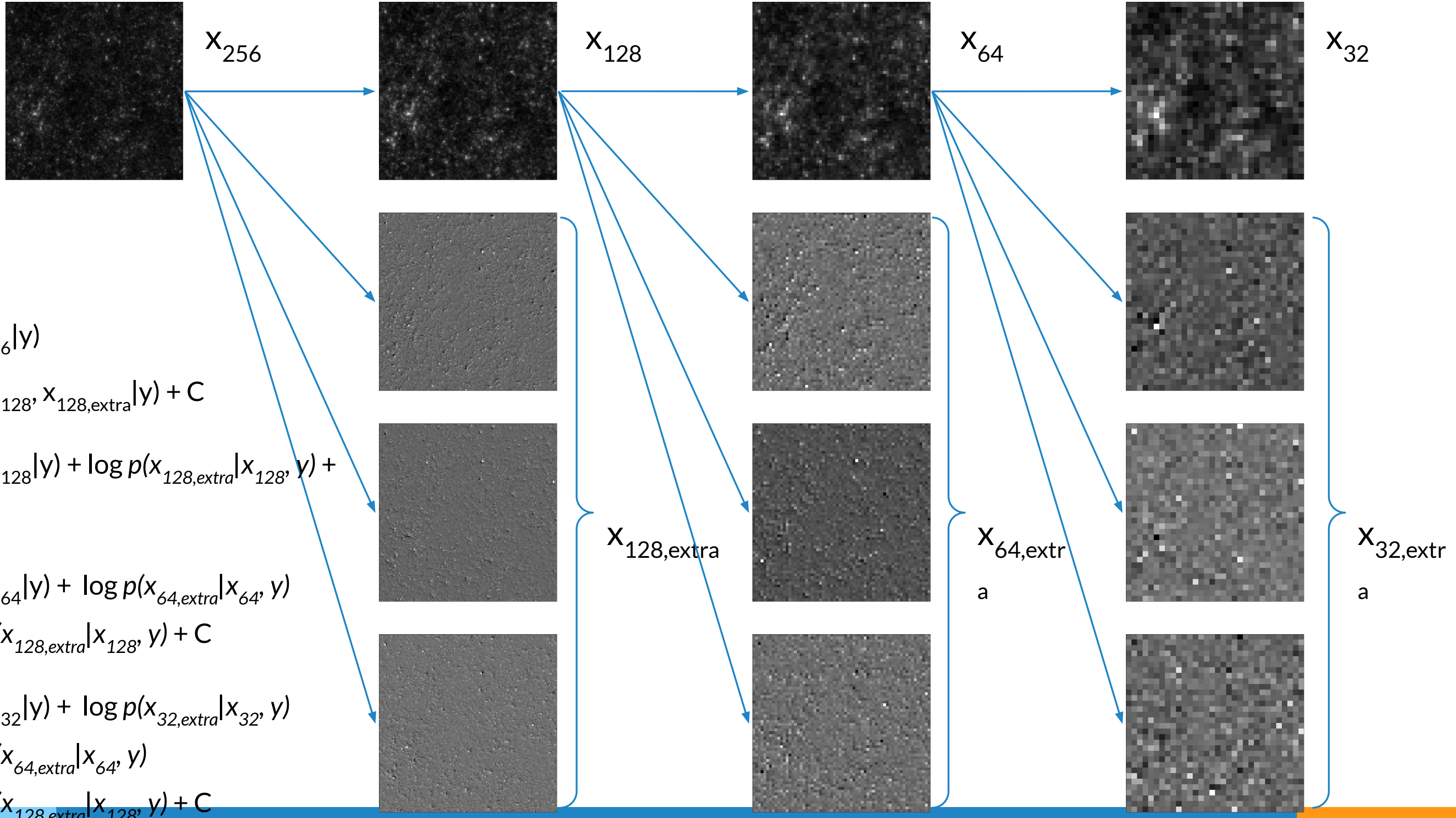
$= \log p(x_{128}, x_{128,extra}|y) + C$

$= \log p(x_{128}|y) + \log p(x_{128,extra}|x_{128}, y) + C$

$= \log p(x_{64}|y) + \log p(x_{64,extra}|x_{64}, y) + \log p(x_{128,extra}|x_{128}, y) + C$

Multiscale flow

- Consider a cosmological field with 256^2 resolution:



$\log p(x_{256}|y)$

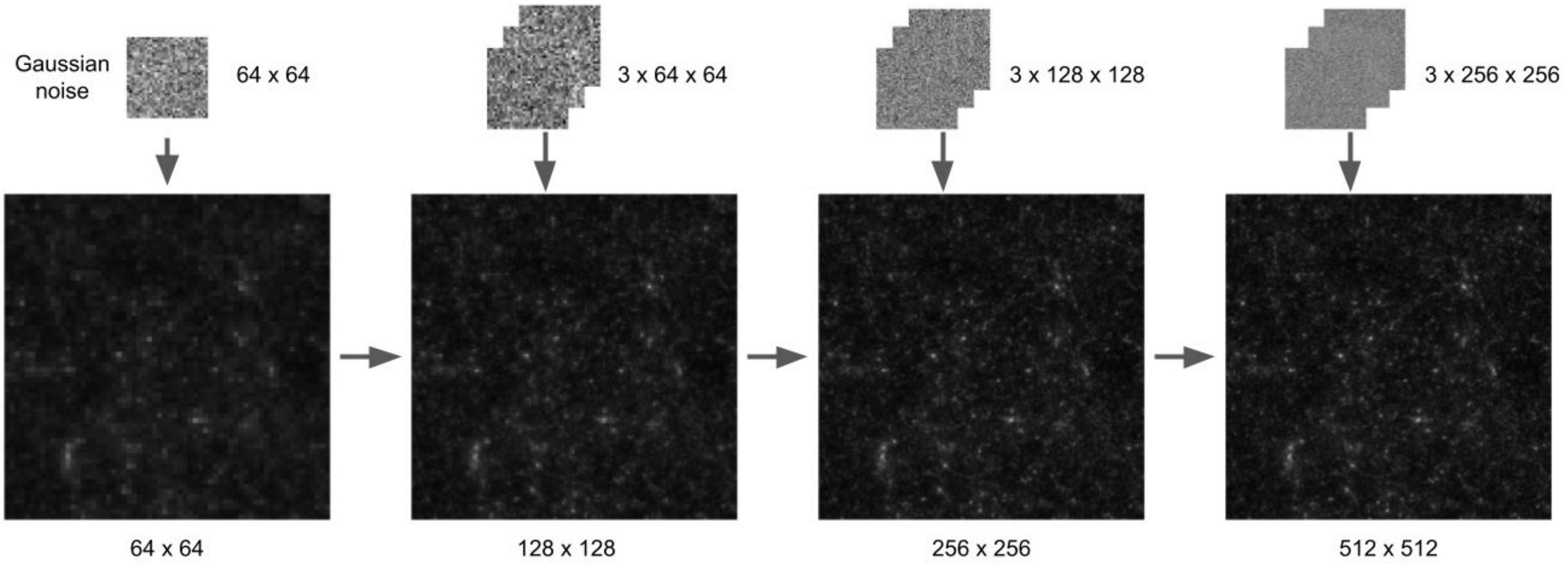
$= \log p(x_{128}, x_{128,extra}|y) + C$

$= \log p(x_{128}|y) + \log p(x_{128,extra}|x_{128}, y) + C$

$= \log p(x_{64}|y) + \log p(x_{64,extra}|x_{64}, y) + \log p(x_{128,extra}|x_{128}, y) + C$

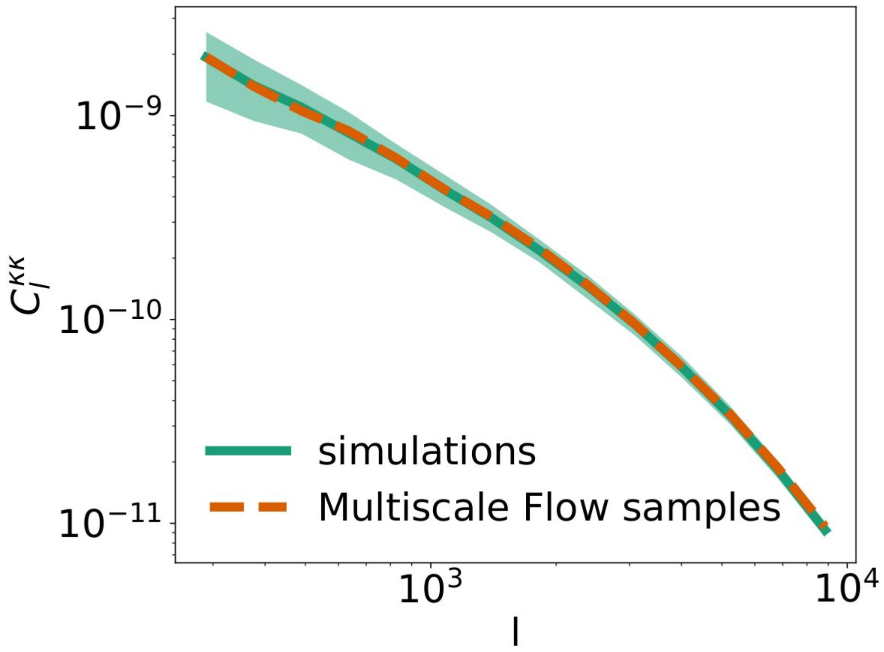
$= \log p(x_{32}|y) + \log p(x_{32,extra}|x_{32}, y) + \log p(x_{64,extra}|x_{64}, y) + \log p(x_{128,extra}|x_{128}, y) + C$

Sample generation & super-resolution

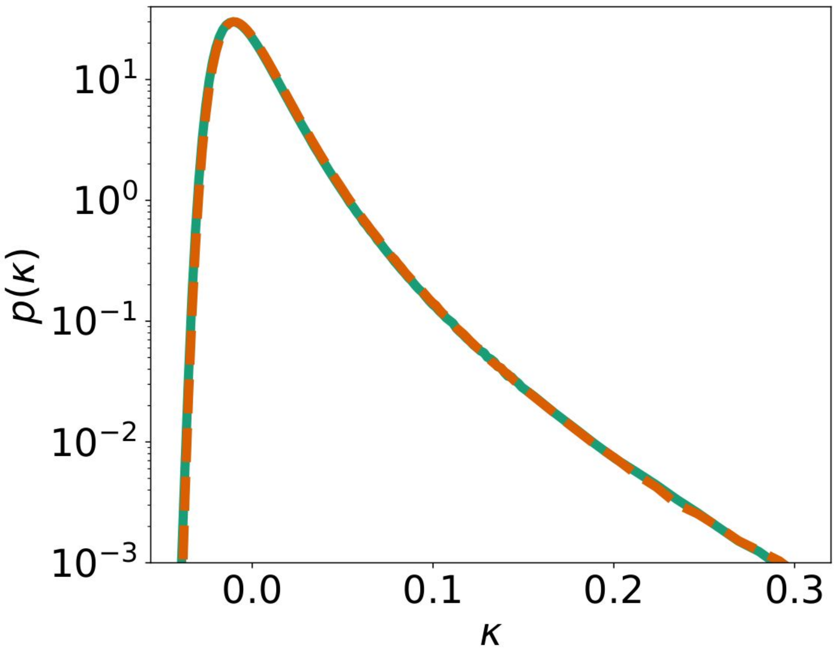


Sample generation & super-resolution

- power spectrum

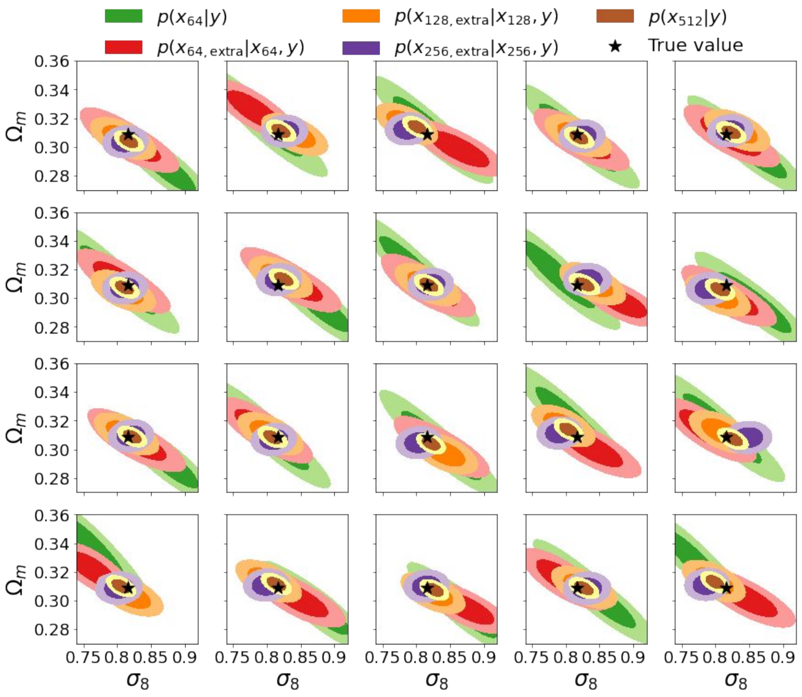


- kappa probability distribution



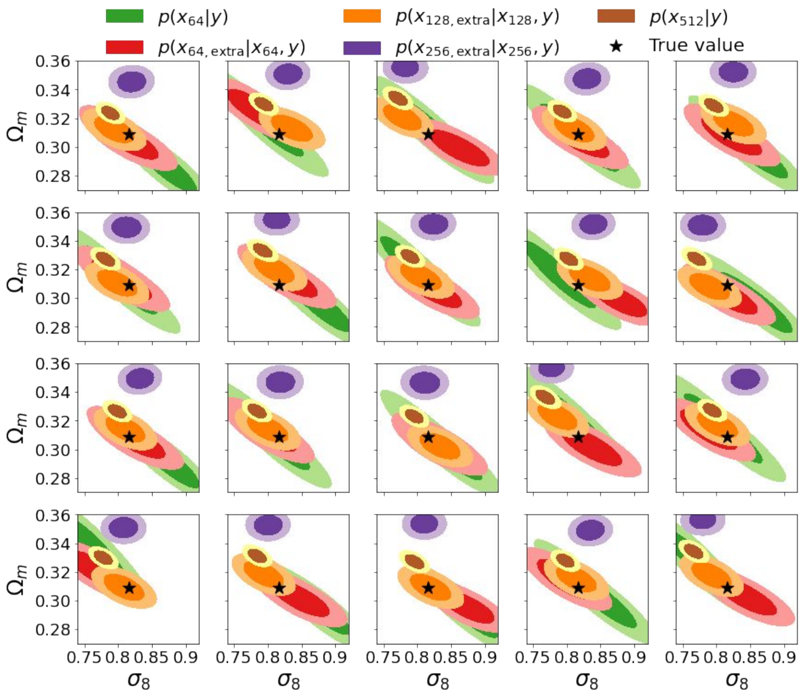
Distribution shift detection — noise miscalibration

- Consistent posteriors from different scales



noise miscalibration

- Inconsistent small scale posterior

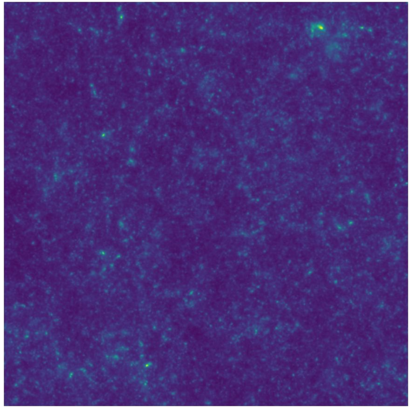


Interpretability

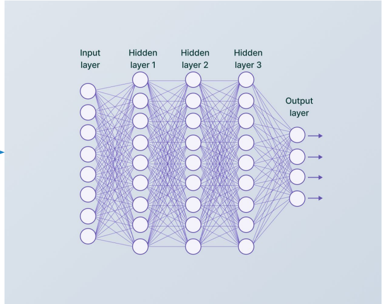
“Where is the extra information coming from?”

“You need to show why the other cosmological models are ruled out”

Input WL map



Generative models



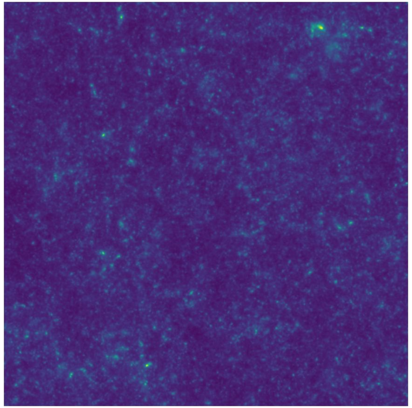
MCMC

$\sigma_8 = 0.76 \pm 0.02$

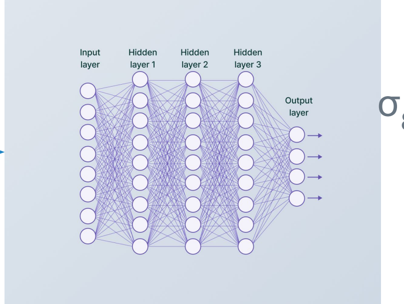
“Where is the extra information coming from?”

“You need to show why the other cosmological models are ruled out”

Input WL map



Generative models

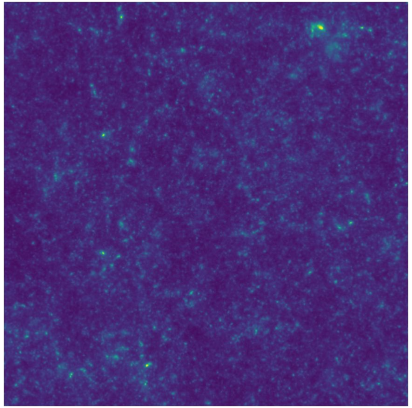


$\sigma_8 = 0.816$

MCMC

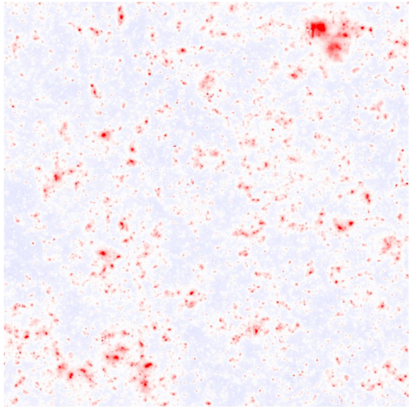
$\sigma_8 = 0.76 \pm 0.02$

Generated sample



The same realization (latent code) as the input map, but assuming a different cosmology

Difference



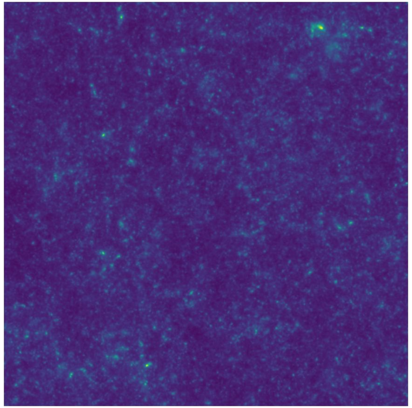
Generated sample - input map

“Where is the extra information coming from?”

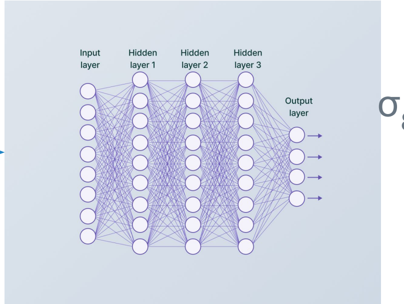
“You need to show why the other cosmological models are ruled out”

Generative models can visualize where the information is coming from, and how the constraints are made.

Input WL map



Generative models

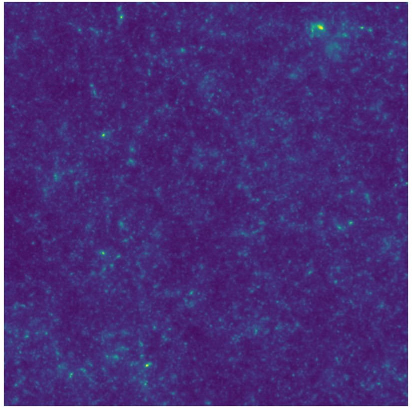


$\sigma_8 = 0.816$

MCMC

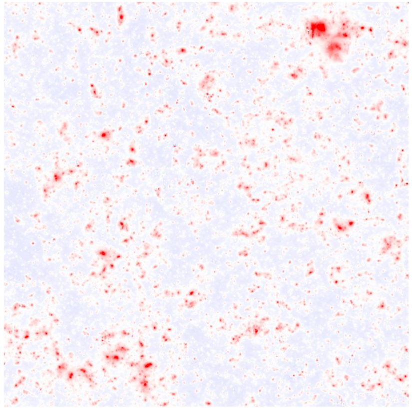
$\sigma_8 = 0.76 \pm 0.02$

Generated sample



The same realization (latent code) as the input map, but assuming a different cosmology

Difference



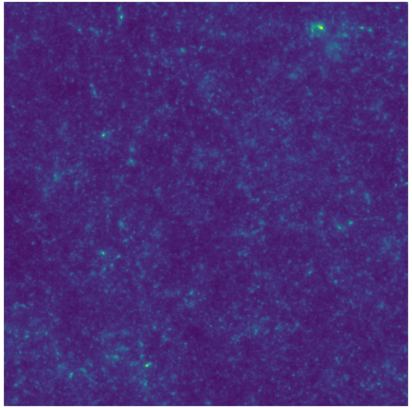
Generated sample - input map

“Where is the extra information coming from?”

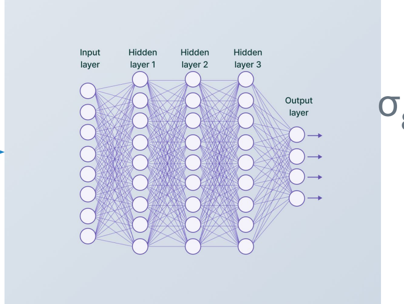
“You need to show why the other cosmological models are ruled out”

My model tells me that the halos from high σ_8 cosmology are too massive!

Input WL map



Generative models

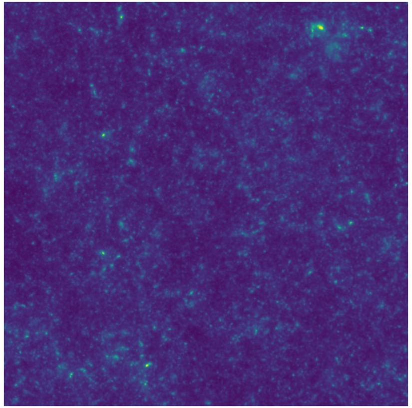


$\sigma_8 = 0.816$

MCMC

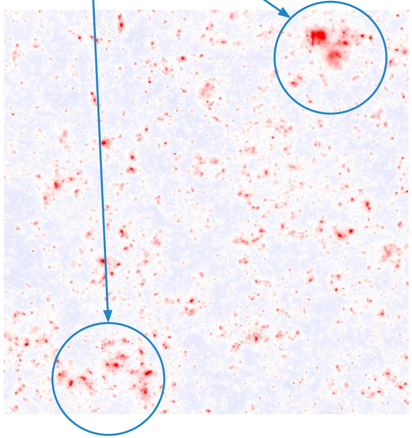
$\sigma_8 = 0.76 \pm 0.02$

Generated sample



The same realization (latent code) as the input map, but assuming a different cosmology

Difference

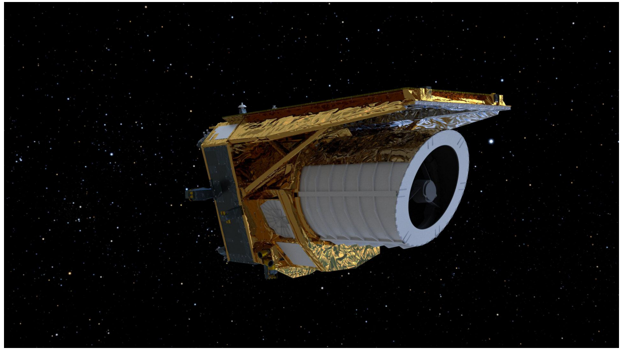


Generated sample - input map

Numerous weak lensing surveys are underway

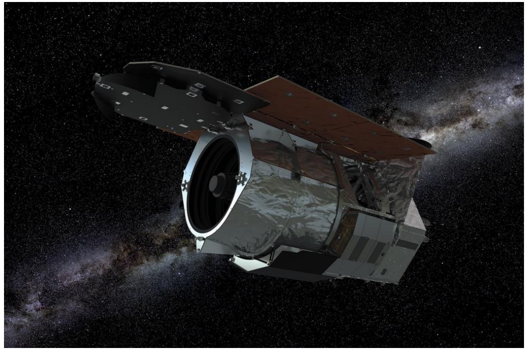
Dark Energy Survey (DES)

Hyper Suprime-Cam (HSC) Subaru Strategic Survey



Euclid telescope

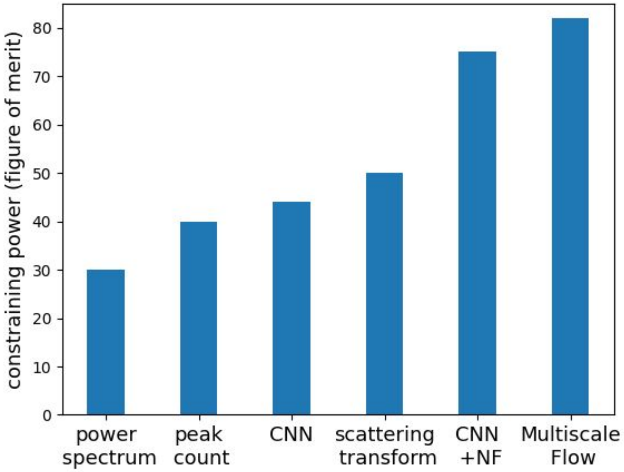
Rubin Observatory LSST



Roman space telescope

Performance on mock weak lensing maps

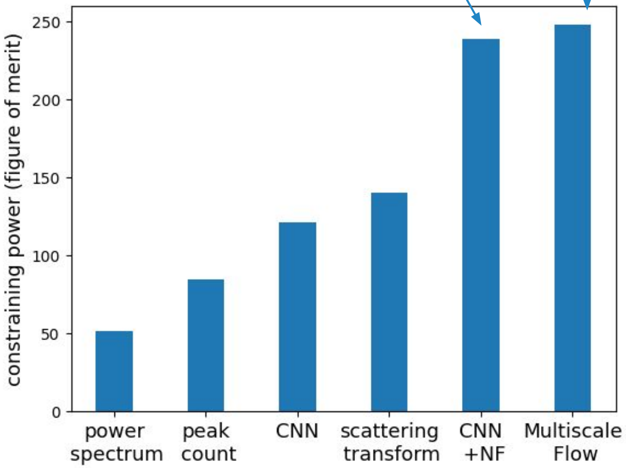
- Current surveys (n_g=10 arcmin^{-2})



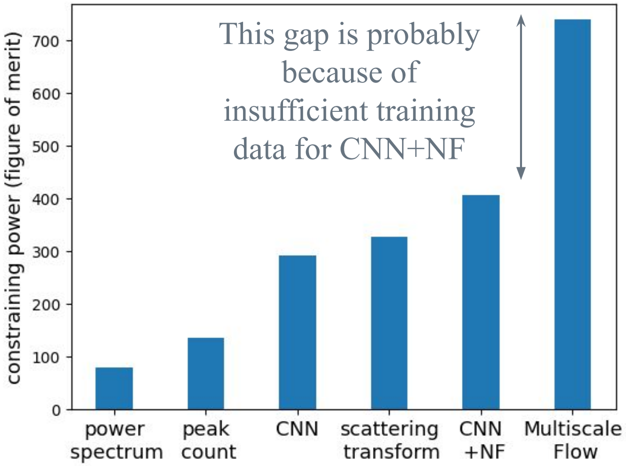
- Upcoming surveys (n_g=30 arcmin^{-2})

Discriminative model (learns posterior)

Generative model (learns likelihood)



- Optimistic scenario for a future-generation space-based survey (n_g=100 arcmin^{-2})



Ribli et al. 2019

Cheng et al. 2021
Allys et al. 2021

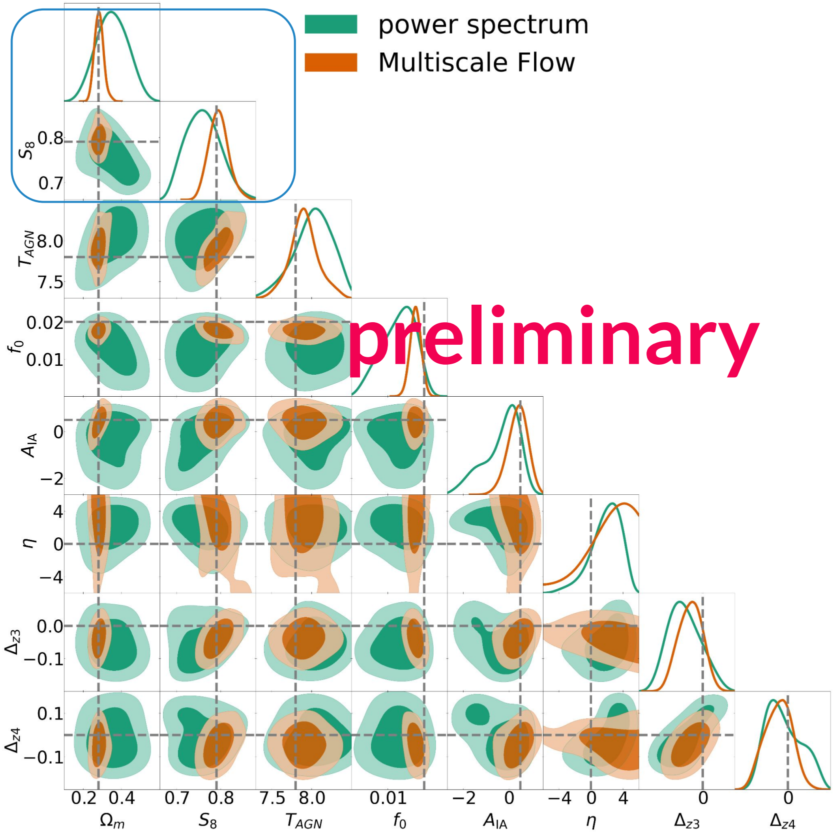
Sharma, **Dai** & Seljak, in prep.

Dai & Seljak 2024

For current and upcoming surveys, generative and discriminative models lead to similar performance, potentially suggesting both may have extracted the full information content from the data

HSC weak lensing analysis with Multiscale Flow

Cosmological constraints

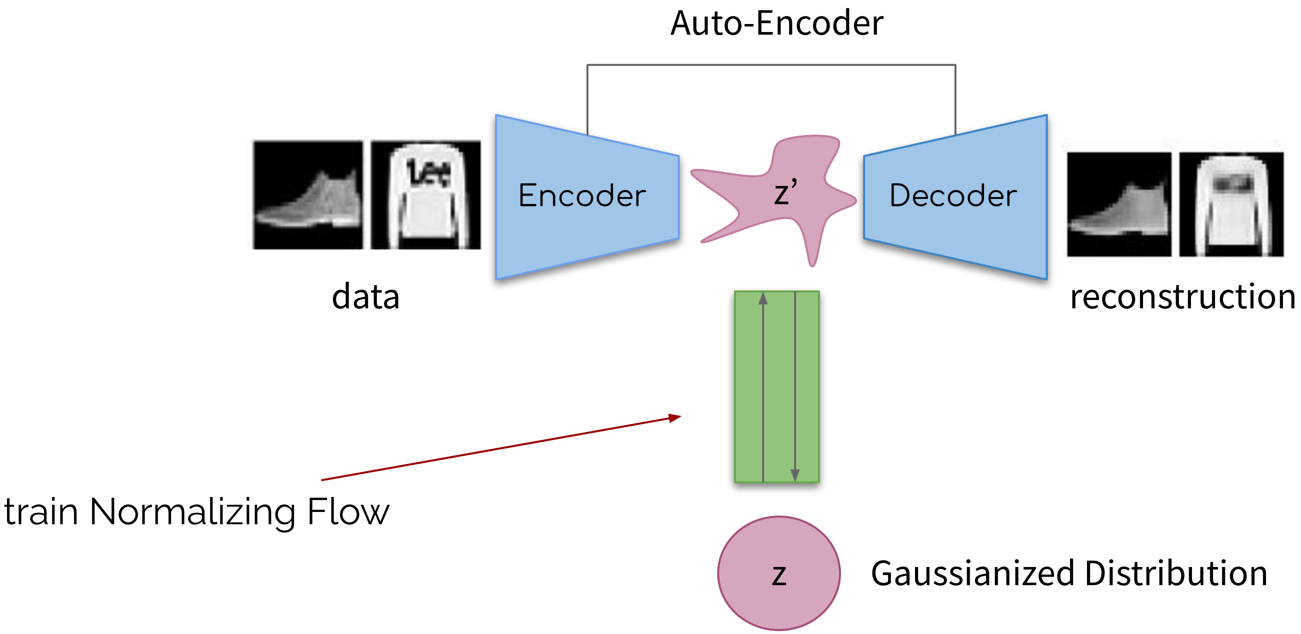


- Tests on mock data: significant improvement compared to traditional power spectrum analysis, after considering various systematic uncertainties

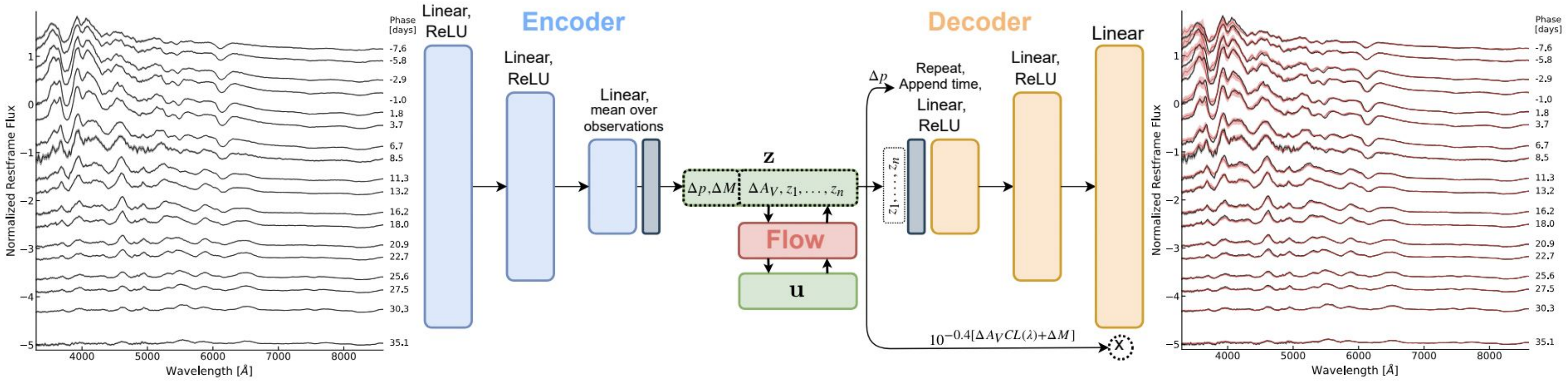
- From left to right:
  - the mean present-day matter density
  - a measure of the homogeneity of the Universe
  - 2 effective baryonic parameter
  - 2 intrinsic alignment parameter
  - 2 parameter of redshift estimation uncertainty

Probabilistic Auto-Encoder (PAE)

Boehm and Seljak 2020 (arxiv: 2006.05479)



PAE for SN1A spectroscopy

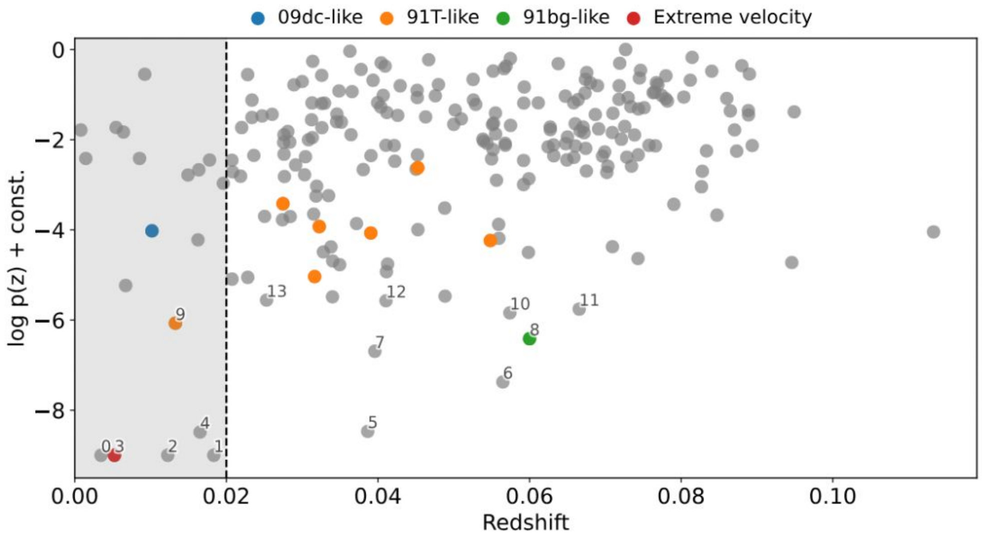


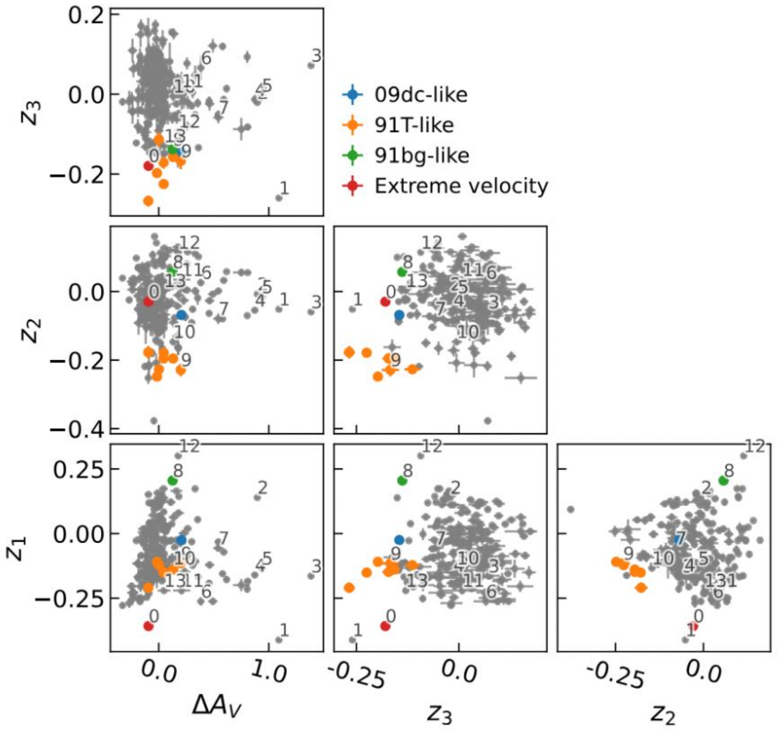
PAE gives a generative model for SN1A
Inpainting of incomplete data
Posterior analysis for distance modulus
Anomaly detection

Better than SALT2 in residuals, 4-5% distance error

Stein, Seljak etal 2022

PAE density and latent space position for anomaly detection in SN1A spectra





Lessons learned

1) In cosmology we seek hidden information in non-Gaussian correlations of the data: **hidden gems are in correlations**
2) **Discriminative learning versus generative learning**: generative harder to train, but gives sample generation (simulations), likelihoods and outlier detection
3) For generative models (e.g. MultiScale Flow) one can use likelihood and scale dependent signal to identify anomalies
4) We are starting to see first applications of ML to cosmology data in weak lensing (CNN, scattering transforms, MSF), with significant gains relative to baseline summary statistic (power spectrum)