FAIR Universe

Weak Lensing ML Uncertainty Challenge

Handling uncertainties and distribution shifts for precision cosmology

NeurIPS 2025 Competition Workshop

2025.12.07

Biwei Dai (IAS)

Po-Wen Chang (LBNL)

U.S. DEPARTMENT of ENERGY
BERKELEY LAB
CHA LEARN
université PARIS-SACLAY
UNIVERSITY of WASHINGTON
IAS INSTITUTE FOR ADVANCED STUDY
Berkeley UNIVERSITY OF CALIFORNIA
Stanford University
SLAC NATIONAL ACCELERATOR LABORATORY
UCI University of California, Irvine

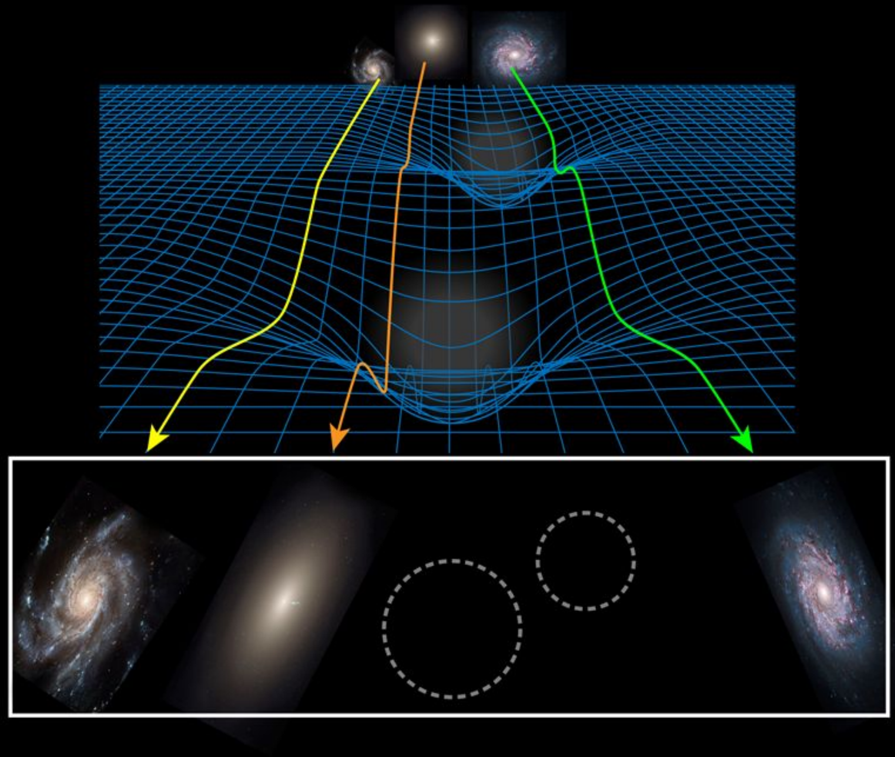
Weak Lensing ML Uncertainty Challenge

FAIR Universe

Overview

Weak gravitational lensing

- The gravity of matter warps the surrounding space-time and causes distortions in the observed shapes of the background galaxies.
- Powerful probe of the matter distribution in our universe from coherent patterns of galaxy shapes.
- Numerous current and upcoming WL surveys: DES, HSC, Euclid, Rubin LSST, Roman, etc.
- Traditional analysis based on two-point correlation functions can only capture limited amount of information from the weak lensing data.
- **AI/ML-based approaches could capture more information hidden in higher-order correlations! WE NEED YOU!**



Weak Lensing ML Uncertainty Challenge

FAIR Universe

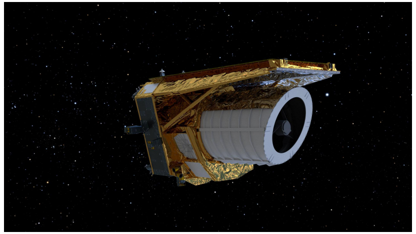
Overview

Weak gravitational lensing

- The gravity of matter warps the surrounding space-time and causes distortions in the observed shapes of the background galaxies.
- Powerful probe of the matter distribution in our universe from coherent patterns of galaxy shapes.
- Numerous current and upcoming WL surveys: DES, HSC, Euclid, Rubin LSST, Roman, etc.
- Traditional analysis based on two-point correlation functions can only capture limited amount of information from the weak lensing data.
- **AI/ML-based approaches could capture more information hidden in higher-order correlations!**
  WE NEED YOU!

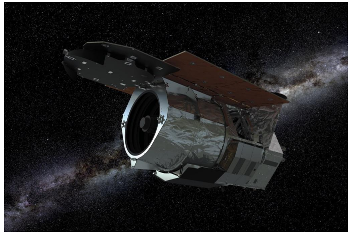
Dark Energy Survey (DES)

Hyper Suprime-Cam (HSC) Subaru Strategic Survey



Euclid telescope

Rubin Observatory LSST



Roman space telescope

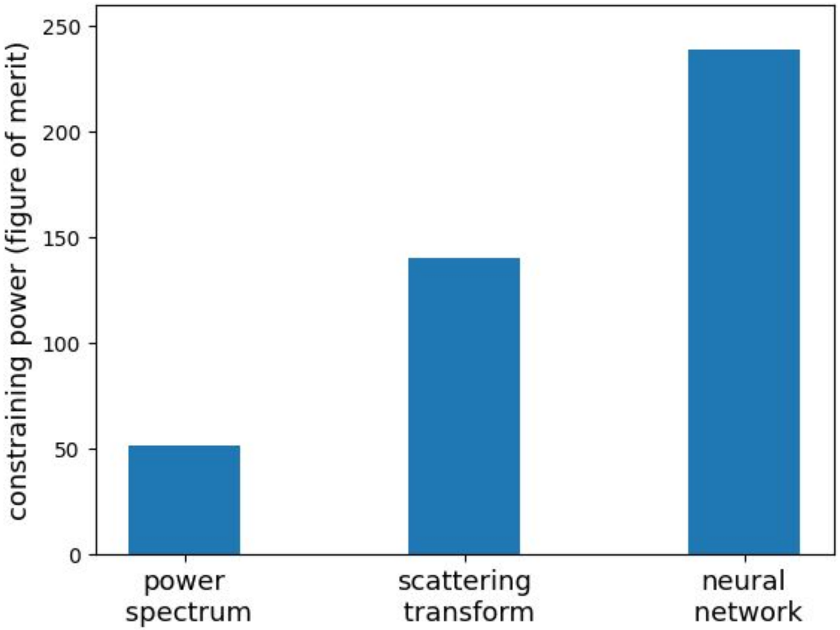
Weak Lensing ML Uncertainty Challenge

FAIR Universe

Overview

Weak gravitational lensing

- The gravity of matter warps the surrounding space-time and causes distortions in the observed shapes of the background galaxies.
- Powerful probe of the matter distribution in our universe from coherent patterns of galaxy shapes.
- Numerous current and upcoming WL surveys: DES, HSC, Euclid, Rubin LSST, Roman, etc.
- Traditional analysis based on two-point correlation functions can only capture limited amount of information from the weak lensing data.
- **AI/ML-based approaches could capture more information hidden in higher-order correlations!** **WE NEED YOU!**



The Challenges of Simulation Based Inference in Cosmology

The need of benchmark dataset

- Many different summary statistics and ML models are proposed
- Most people test their methods on their own dataset with different setups, making it hard to compare different methods and understand their pros and cons

Method Name	Type	Reference	Improvement over Two-Point Statistics
Bispectrum (3-point correlation)	Summary Statistic	Multiple foundational papers	10–20% tighter constraints; breaks parameter degeneracies
Trispectrum (4-point correlation)	Summary Statistic	Multiple foundational papers	Further lifts degeneracies; improves error estimation
Peak Counts	Summary Statistic	Multiple, incl. MCALens	Up to 157% improvement with advanced mass mapping
Wavelet Peak Counts / Starlet Transform	Summary Statistic	Multiple wavelet analysis papers	Tighter constraints; nearly diagonal covariance
Minkowski Functionals	Summary Statistic	Morphological statistics papers	70% tighter constraints when combined with Betti numbers
Betti Numbers / Persistent Homology	Summary Statistic	Topology-based analysis papers	70% tighter constraints when combined with second moments
Probability Distribution Function (PDF)	Summary Statistic	Density field PDF papers	Extracts non-Gaussian info inaccessible to 2PCF
Void Statistics	Summary Statistic	Void analysis papers	Complementary to peaks; adds unique information
Scattering Transform	Summary Statistic	Recent mathematical framework	Up to 2x higher constraining power than peak counts/CNNs
3PCF Multipoles	Summary Statistic	3PCF multipole analysis	20% improvement; quadrupole most constraining
Cumulant Correlators, Skew/Kurt-Spectra	Summary Statistic	Higher-order moment analysis	Improves parameter constraints; captures non-Gaussianity
Convolutional Neural Networks (CNNs)	ML Model	\[1802.01212], \[1902.03663], \[1906.03156]	2–9x stronger constraints; 4–7x lower parameter scatter
Information Maximising Neural Networks (IMNN)	ML Model	\[2407.10877]	Up to 100% of full-field FoM; outperforms MSE-based (81%)
Multiscale Flow (Normalizing Flow)	ML Model	\[2403.03490]	2.7–7.8x stronger than power spectrum; \~2x higher than peaks/CNNs
Simulation-Based Inference (SBI)	ML Model	\[2409.17975], \[2409.01301]	Enables high-dimensional stats; combines HOS for improved constraints
Neural Posterior Estimation (NPE)	ML Model	SBI framework papers	Outperforms traditional stats; direct posterior estimation
Neural Likelihood Estimation (NLE)	ML Model	\[2409.17975]	Best among implicit methods for full-field inference
Diffusion Models	ML Model	\[2312.00000]	Outperforms GANs in denoising (qualitative)
Generative Adversarial Networks (GANs)	ML Model	GAN application papers	Lower quality than diffusion models for cosmological stats
Hybrid Summary Statistics (Neural + Physics-based)	Hybrid	\[2407.18909]	At least as much as power spectrum, up to 2x in some regimes
Field-Level Inference + SBI (Shear-to-Cosmology)	Hybrid	\[2511.22851]	\~2x higher FoM than convergence-based; 36.4% over shear 2pt
Nearest Neighbour Stats + Hybrid NN	Hybrid	\[2511.13393]	CDFs nearly 2× better than 2ptCF; 24× more efficient than point cloud methods
Combined HOS + Neural Compression	Hybrid	\[2409.01301]	30% improvement in Ωm error, 21% in σ8 over power spectrum
PCA Denoising + ML Compression	Hybrid	\[2511.22851]	36.4% improvement in FoM over standard shear 2PCF
Physically-Informed NN Architectures	Hybrid	\[2407.18909]	Same/better performance with fewer parameters/simulations
Lognormal & GPTG Models	Hybrid	GPTG modeling papers	2–5x better than lognormal; matches higher-order stats

(Table generated by ChatGPT)

Weak Lensing ML Uncertainty Challenge

FAIR Universe

The Challenges of Simulation Based Inference in Cosmology

Small training size

- Cosmological simulations are expensive! Each simulation evolves hundreds of billions of particles from the early universe to the present day
- In most cases we are in the low training data regime.
- ML approaches are powerful but can be data-hungry
- We need special treatment to reduce the sample complexity:
    - Domain knowledge (e.g., symmetry, summary statistics)
    - ML techniques (e.g., weight sharing, ensembles)
    - Pre-training
    - …

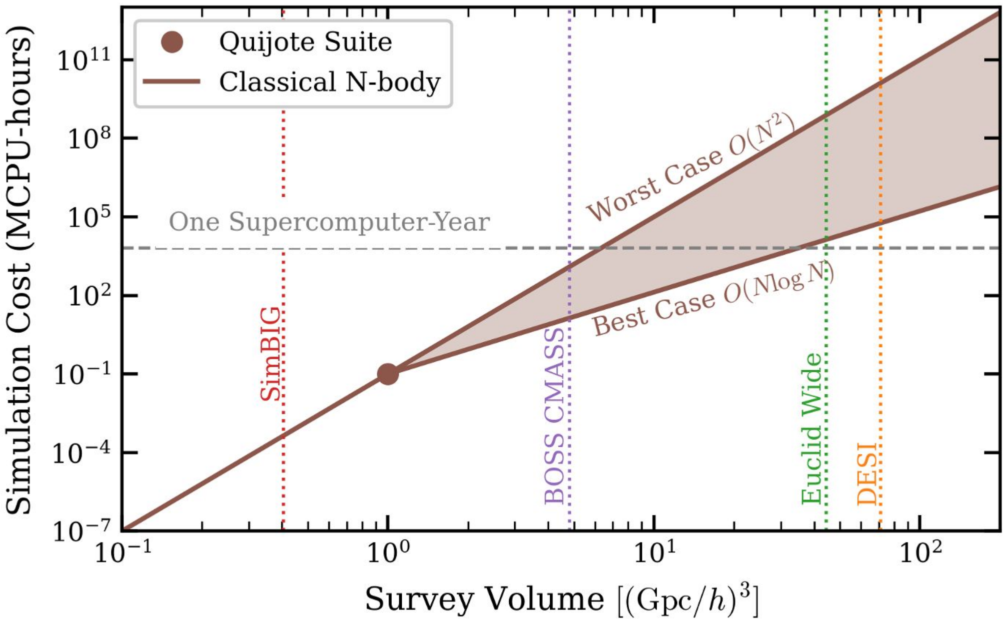


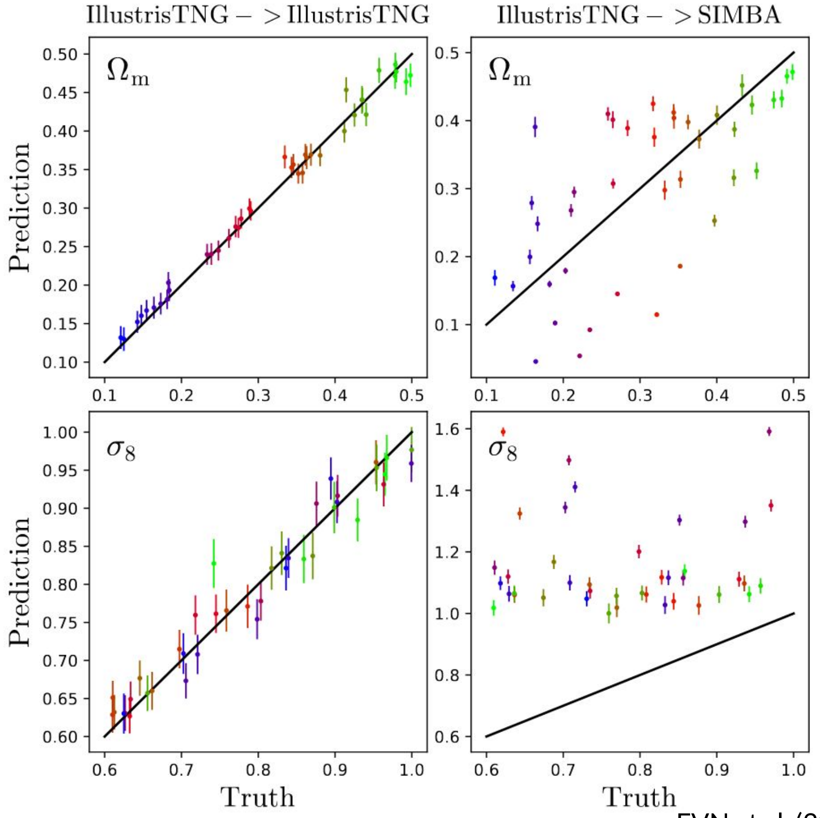
Image credit: Matthew Ho

Weak Lensing ML Uncertainty Challenge

The Challenges of Simulation Based Inference in Cosmology

Distribution shift

- SBI assumes that the simulations it trained on overlap with reality
- There are many systematic effects that we don't have good models (known unknowns)
- Unknown unknowns
- Such distribution shift could lead to significant bias in data analysis
- This is tackled in Phase 2 (anomaly detection).



FVN et al. (2021a)

Weak Lensing ML Uncertainty Challenge

FAIR Universe

The Goals of this Data Challenge

- To encourage groups with expertise in AI and cosmology to develop, test, and validate their model under realistic SBI setups
- To provide a benchmark that helps the community evaluate the performance of different approaches
- To understand the information content of weak lensing maps (Phase 1)
- To improvement the robustness under distribution shifts (Phase 2)
- To facilitate the deployment of DL approaches into survey analysis pipelines

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Competition Tasks

The competition tasks are structured into **two phases**:

- **Phase 1: Cosmological Parameter Estimation**

  Participants will develop models that:

  - Accurately infer cosmological parameters $(\hat{\Omega}_m, \hat{S}_8)$ from the weak lensing image data.
  - Quantify uncertainties via the 68% confidence intervals of the parameters of interest $(\hat{\sigma}_{\Omega_m}, \hat{\sigma}_{S_8})$.

Scoring metrics:

KL divergence between the true Gaussian-like posterior distribution and the Gaussian with the predicted mean and standard deviation:

$$\text{score}_{\text{inference}} = -\frac{1}{N_{\text{test}}} \sum_i^{N_{\text{test}}} \left\{ \frac{\left(\hat{\Omega}_{m,i} - \Omega_{m,i}^{\text{truth}}\right)^2}{\hat{\sigma}^2_{\Omega_m,i}} + \frac{\left(\hat{S}_{8,i} - S_{8,i}^{\text{truth}}\right)^2}{\hat{\sigma}^2_{S_8,i}} + \log\left(\hat{\sigma}^2_{\Omega_m,i}\right) + \log\left(\hat{\sigma}^2_{S_8,i}\right) + \lambda \left[\left(\hat{\Omega}_{m,i} - \Omega_{m,i}^{\text{truth}}\right)^2 + \left(\hat{S}_{8,i} - S_{8,i}^{\text{truth}}\right)^2\right] \right\}$$

$\lambda \equiv 10^3$: penalty factor for bad point estimates

- **Phase 2: Out-of-Distribution Detection**

  Participants will develop models that:

  - Identify test data samples inconsistent with the training distribution (OoD detection).
  - Provide probability estimates indicating data conformity to training distributions.

Binary cross-entropy:

$$\text{score}_{\text{OoD}} = \frac{1}{N_{\text{test}}} \sum_i^{N_{\text{test}}} \left[y_i \log\left(\hat{p}_{\text{InD},i} + \epsilon\right) + (1 - y_i) \log\left(1 - \hat{p}_{\text{InD},i} + \epsilon\right) \right]$$

where $\hat{p}_{\text{InD},i} \in [0, 1]$, $y_i = 1$ if the dataset is InD, $y_i = 0$ if the dataset is OoD, and ϵ is a small positive constant to avoid a score of $-\infty$.

Weak Lensing ML Uncertainty Challenge

FAIR Universe

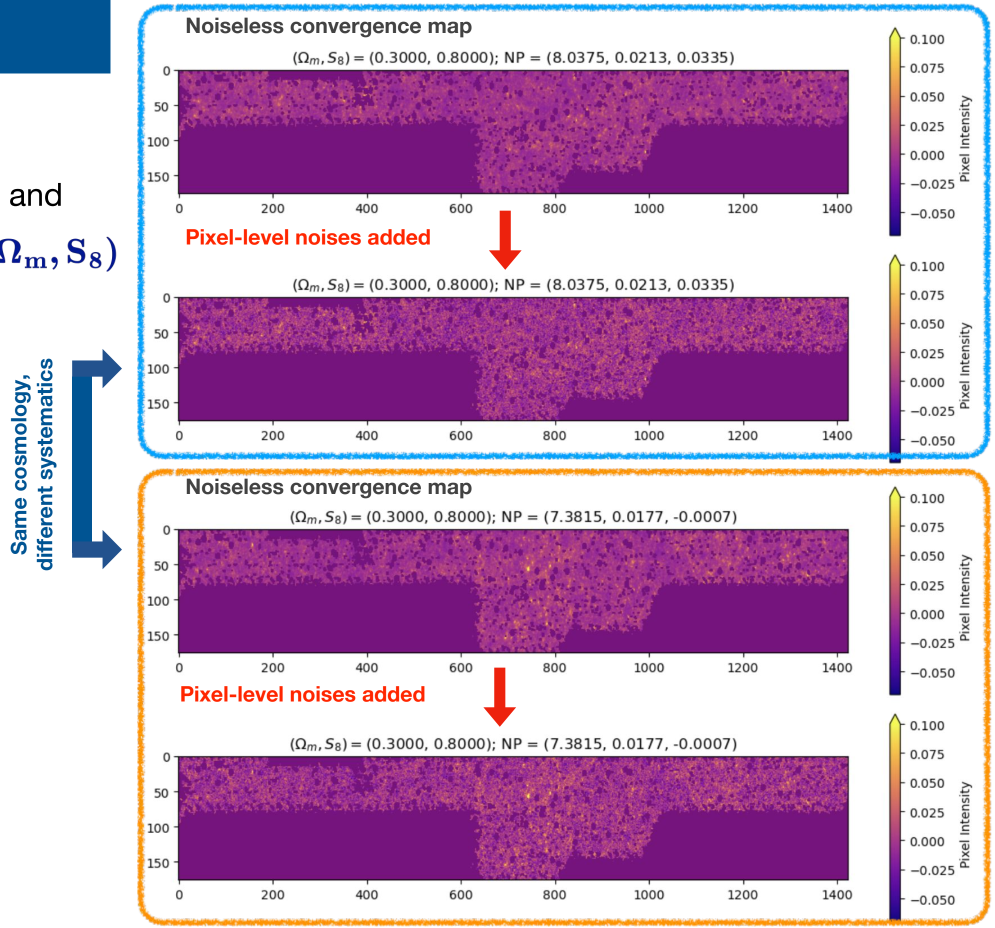
Dataset

- Mock galaxy catalogs predicted with N-body simulations and ray-tracing algorithms at 101 cosmological parameters (Ω_m, S_8)
- Pixelized 2D weak lensing images: **convergence maps**
- The model must take into account the systematic uncertainties from **3 realistic systematic effects**

  2 baryonic effect uncertainties

  1 photometric redshift uncertainty

  along with **pixel-level noises**



Weak Lensing ML Uncertainty Challenge

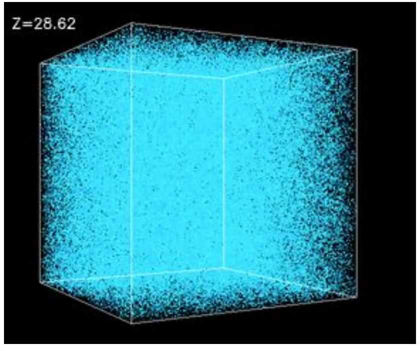
FAIR Universe

Dataset Generation Pipeline

random seed
cosmology

IC

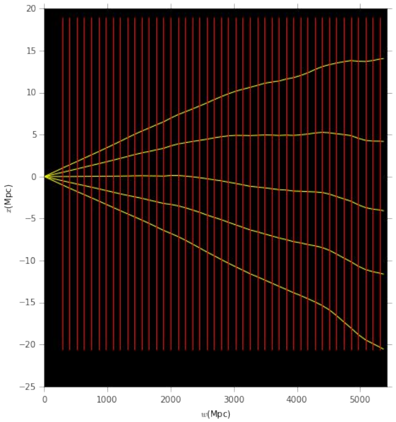
cosmology



N-body simulations: simulate the matter distribution in our universe

Systematic effect
(baryonic feedback, photo-z)

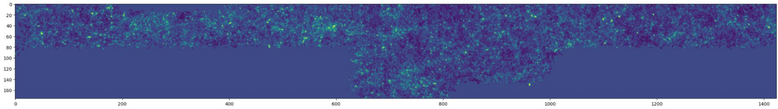
Randomly translate, rotate, and flip the matter planes



Ray-tracing: simulate the weak lensing signal

$\left\{ \begin{array}{c} \mathrm{Galaxy}^{i}\text{\_pos} \\ \mathrm{Galaxy}^{i}\text{\_shape} \end{array} \right\}_{i}$

Galaxy catalog



Pixelize and reconstruct convergence map

Image credit: https://cosmicweb.uchicago.edu/filaments.html
https://lenstools.readthedocs.io/en/latest/raytracing.html

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Dataset

The participants will be provide with:

- **Public training set:**
  - Image data; shape = (101, 256, 1424, 176)
  - Label shape = (101, 256, 5)

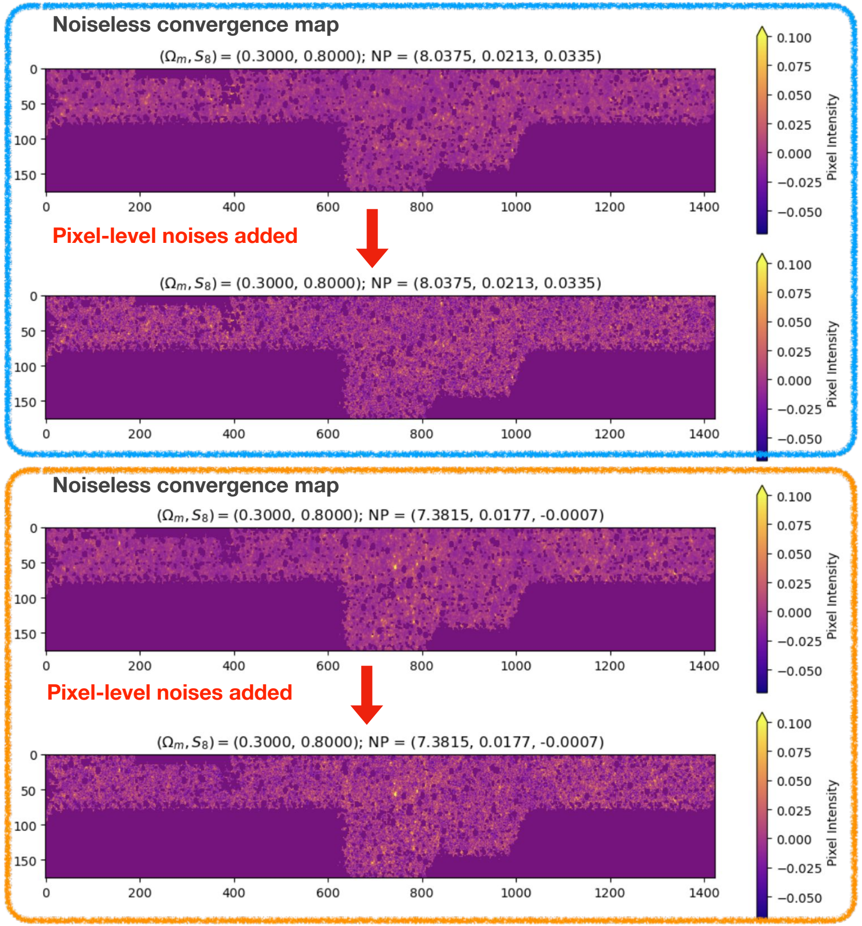
101 = Realizations of cosmological models; each characterized with 2 parameters of interest $(\mathbf{\Omega_m, S_8})$

256 = Realizations of 3 nuisance parameters for systematics (1) and (2)

(1424, 176) = Image dimension

5 = 2 parameters of interest $(\mathbf{\Omega_m, S_8})$ + 3 nuisance parameters for systematics (1) and (2)

- **The provided training set is noiseless.** Participants can generate **pixel-level noise** to augment their training data using a simple `add_noise` function we provide



Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 1 Dataset

The participants will be provide with:

- **Test set:**
  - Image data; shape = (N_{test}, 1424, 176)

    N_{test} = Number of test images

    (1424, 176) = Image dimension
  - The test images are generated with random cosmological parameters, random nuisance parameters, and random pixel-level noises.

Phase 1 Evaluation

The true parameters $(\Omega_m^{truth}, S_8^{truth})$ of the public test set are unknown to the participants.

Participants submit their predictions of

- **Cosmological parameters** $(\hat{\Omega}_m, \hat{S}_8)$
- **Their uncertainties** $(\hat{\sigma}_{\Omega_m}, \hat{\sigma}_{S_8})$

to **Codabench**, our competition platform.

The model performance was then evaluated with the hidden ground truth based on our scoring metrics.

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Limitations of the Current Data Challenge

- To make the competition more accessible, we simplified the dataset to reduce the training size below 10 GB (e.g., single redshift bin, one subfield, convergence maps instead of galaxy catalog, ignore some systematic effects such as IA).
- The loss function is somewhat ad-hoc.
- The public test set on Codabench contains different realizations of the same 101 cosmologies as the training set, which may have increased the chance of overfitting on the 101 cosmologies when using the public leaderboard score as guidance for model optimization, although it was not our intention.
- The limited number of cosmological models in the second test set.
- Comments and suggestions are welcome to improve the dataset as a permanent benchmark!

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 1 Final Winners

Leaders in the public leaderboard are further evaluated on a **holdout dataset** that contains two sets of cosmologies:

- **(i)** New realizations of the cosmologies that were seen in the the test and training dataset
- **(ii)** New realizations of the cosmologies that were ***not seen*** in the test and training dataset

Holdout dataset = **(i)** **Rewarding the original cosmologies** + **(ii)** **Rewarding the unseen cosmologies**

- We present the final results in three separate leaderboards to reward both cases

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 1 Final Winners

1. Final leaderboard evaluated solely on (i):

RANK	PARTICPANT	FINAL SCORE	MEAN MSE (STANDERDIZED)	MEAN COVERAGE
1st	cmbagent	11.7029	0.1033	0.7000
2nd	eiffl	11.6535	0.1038	0.7087
3rd	Shubhojit	11.5987	0.1032	0.6583

We will award the prizes to **cmbagent**, **eiffl**, and **Shubhojit** for extraordinary performance on the original cosmologies.

🏆 **Cmbagent –** Erwan Allys, Boris Bolliet, Tom Borret, Celia Lecat, Andy Nilipour, Sebastien Pierre, Licong Xu

🏆 **Transatlantic Dream Team (eiffl) –** Noe Dia, Sacha Guerrini, Wassim Kablan, François Lanusse, Julia Linhart, Laurence Perreault-Levasseur, Benjamin Remy, Sammy Sharieff, Andreas Tersenov, Justine Zeghal

🏆 **shubhojit –** Shubhojit Naskar

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 1 Final Winners

2. Final leaderboard evaluated solely on (ii):

RANK	PARTICPANT	FINAL SCORE	MEAN MSE (STANDERDIZED)	MEAN COVERAGE
1st	Shubhojit	11.3606	0.0968	0.6619
2nd Tie	THUML	11.0511	0.1051	0.6733
2nd Tie	jagoncalves	11.0367	0.1073	0.6683
2nd Tie	andry834	11.0014	0.1076	0.7228
2nd Tie	jhu_suicee	10.9892	0.1067	0.6451
2nd Tie	eiffl	10.9883	0.1074	0.6818

We recognize **Shubhojit** for the achievement in the best model generalization, with a score clearly separated from the other participants. The other five participants on the leaderboard cannot be separated in a significant way due to the limited samples of (ii).

🏆 **shubhojit** – Shubhojit Naskar

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 1 Final Winners

3. Final leaderboard from the average of the score obtained on (i) and (ii):

RANK	PARTCIPANT	FINAL SCORE	MEAN MSE (STANDERDIZED)	MEAN COVERAGE
1st	Shubhojit	11.4796	0.1000	0.6601
2nd	eiffl	11.3209	0.1056	0.6953
3rd	THUML	11.2848	0.1060	0.6789

We will award the prizes to **Shubhojit**, **eiffl**, and **THUML** for demonstrating excellent performance on both new and old cosmologies.

- 🏆 **shubhojit** – Shubhojit Naskar
- 🏆 **Transatlantic Dream Team (eiffl)** – Noe Dia, Sacha Guerrini, Wassim Kablan, François Lanusse, Julia Linhart, Laurence Perreault-Levasseur, Benjamin Remy, Sammy Sharieff, Andreas Tersenov, Justine Zeghal
- 🏆 **THUML** – Mingsheng Long, Yuezhou Ma, Haonan Shangguan, Yuanxu Sun, Huikun Weng, Haixu Wu, Hang Zhou

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 1 Jury Prizes & Special Mentions

🏅 **Transatlantic Dream Team (eiffl)** – Noe Dia, Sacha Guerrini, Wassim Kablan, François Lanusse, Julia Linhart, Laurence Perreault-Levasseur, Benjamin Remy, Sammy Sharieff, Andreas Tersenov, Justine Zeghal

For their illuminating analysis of diverse approaches on tackling the limitations of this challenge

🏅 **Cmbagent** – Erwan Allys, Boris Bolliet, Tom Borret, Celia Lecat, Andy Nilipour, Sebastien Pierre, Licong Xu

For their novel approach leveraging an AI agententic workflow for science

🏅 **andry834** – Andry Rafaralahy
azhang81 – Anday Zhang

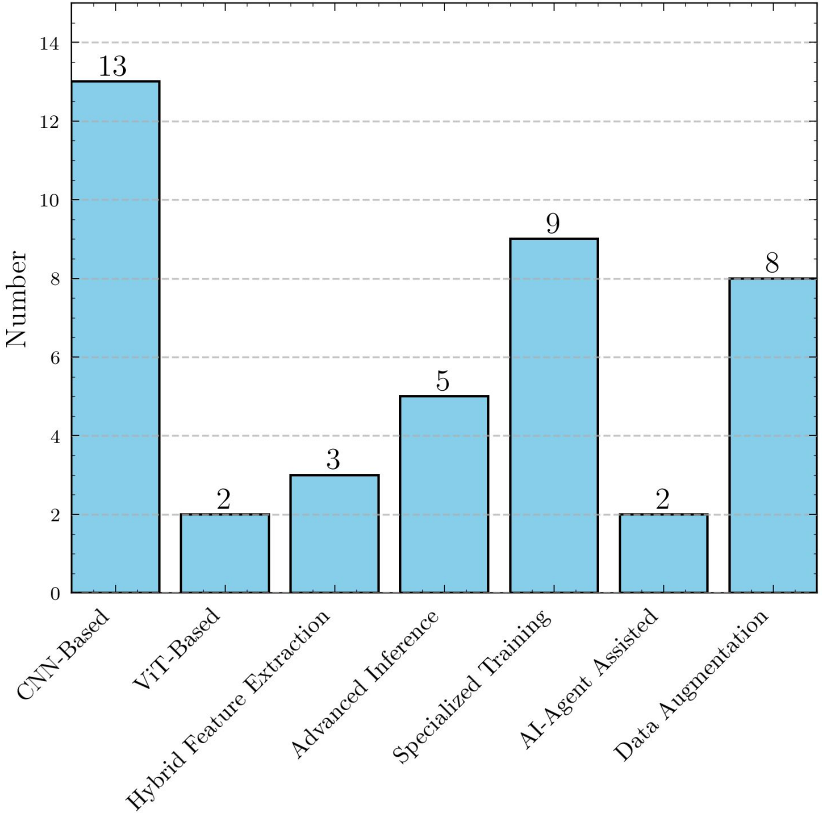
For their innovative methods and model architectures for this challenge

🎉 **Congratulations to all the winning teams!**

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Final Submitted Phase 1 Solutions



Architecture: CNN-based, ViT-based

Hybrid Feature Extraction: Combined deep learning with fixed mathematical or physics-based extractors (e.g., Scattering Transforms, Handcrafted Cosmology Features)

Advanced Inference: Used methods beyond direct regression, such as Simulation-Based Inference, Normalizing Flows, or MCMC sampling to estimate posteriors

Specialized Training: Unique optimization strategies like Reinforcement Learning, Denoising U-Nets, Robust Outlier Filtering, Custom Loss functions, or Post-hoc Uncertainty Calibration

AI-Agent Assisted: Explicitly utilized Large Language Models (LLMs) or automated agents for code generation and architecture search

Data Augmentation: Geometric, Domain-specific synthetic

Note: The best score achieved by higher-order statistics on the public leaderboard seems to be 9.1654 (43th place)

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Competition Tasks

The competition tasks are structured into **two phases**:

- **Phase 1: Cosmological Parameter Estimation**

  Participants will develop models that:

  - Accurately infer cosmological parameters $(\hat{\Omega}_m, \hat{S}_8)$ from the weak lensing image data.
  - Quantify uncertainties via the 68% confidence intervals of the parameters of interest $(\hat{\sigma}_{\Omega_m}, \hat{\sigma}_{S_8})$.

- **Phase 2: Out-of-Distribution Detection**

  Participants will develop models that:

  - Identify test data samples inconsistent with the training distribution (OoD detection).
  - Provide probability estimates indicating data conformity to training distributions.

Scoring metrics:

KL divergence between the true Gaussian-like posterior distribution and the Gaussian with the predicted mean and standard deviation:

$$\text{score}_{\text{inference}} = -\frac{1}{N_{\text{test}}}\sum_{i}^{N_{\text{test}}}\left\{\frac{\left(\hat{\Omega}_{m,i}-\Omega_{m,i}^{\text{truth}}\right)^2}{\hat{\sigma}_{\Omega_m,i}^2}+\frac{\left(\hat{S}_{8,i}-S_{8,i}^{\text{truth}}\right)^2}{\hat{\sigma}_{S_8,i}^2}+\log\left(\hat{\sigma}_{\Omega_m,i}^2\right)+\log\left(\hat{\sigma}_{S_8,i}^2\right)+\lambda\left[\left(\hat{\Omega}_{m,i}-\Omega_{m,i}^{\text{truth}}\right)^2+\left(\hat{S}_{8,i}-S_{8,i}^{\text{truth}}\right)^2\right]\right\}$$

$\lambda = 10^3$: penalty factor for bad point estimates

Binary cross-entropy:

$$\text{score}_{\text{OoD}} = \frac{1}{N_{\text{test}}}\sum_{i}^{N_{\text{test}}}\left[y_i\log\left(\hat{p}_{\text{InD},i}+\epsilon\right)+(1-y_i)\log\left(1-\hat{p}_{\text{InD},i}+\epsilon\right)\right]$$

where $\hat{p}_{\text{InD},i} \in [0,1]$, $y_i = 1$ if the dataset is InD, $y_i = 0$ if the dataset is OoD, and ϵ is a small positive constant to avoid a score of $-\infty$.

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 2 Dataset

The participants will be provide with:

- **Public test set:**
  - Image data; shape = (6000, 1424, 176)

    6000 = Number of test images

    (1424, 176) = Image dimension

  - **A fraction of test data will be generated with different physical models (OoD)**, leading to some distribution shifts with respect to the test data in Phase 1

* Final dataset may be subject to change

Phase 2 Evaluation

Participants will submit their predictions of **in-distribution (InD) probability** of each test instance to our Codabench.

The model performance was then evaluated with the hidden ground truth labels **(y=1 for InD; y=0 for OoD)** based on our scoring metrics.

Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 2 Dataset

Can you tell which instances below are OoD?

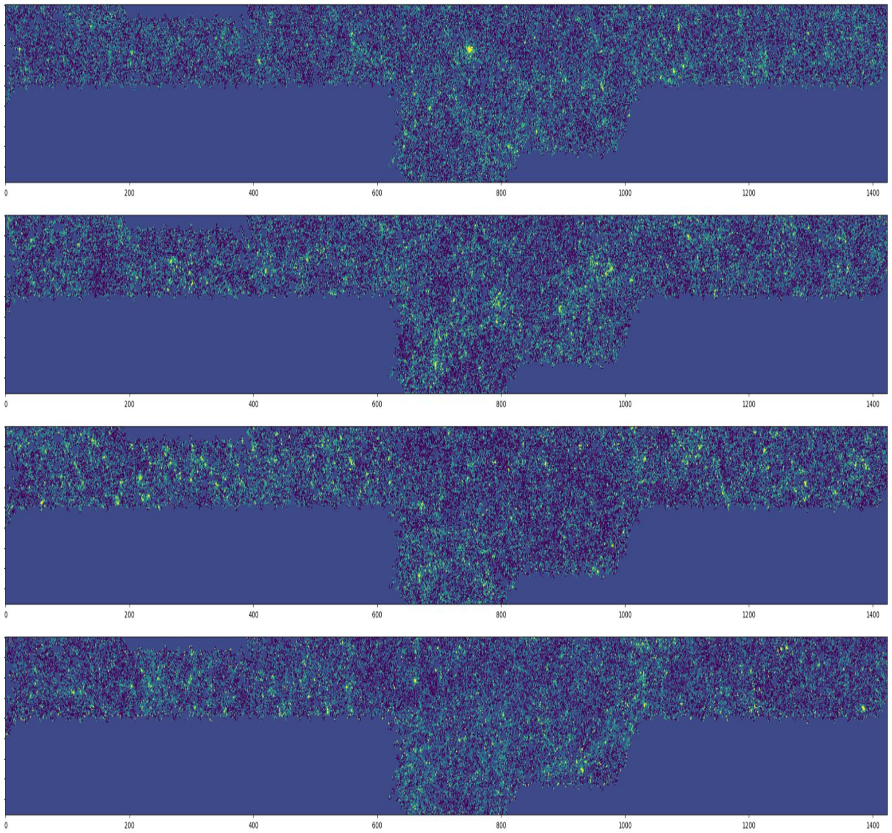
The participants will be provide with:

- **Public test set:**
  - Image data; shape = (6000, 1424, 176)

    6000 = Number of test images

    (1424, 176) = Image dimension

  - **A fraction of test data will be generated with different physical models (OoD)**, leading to some distribution shifts with respect to the test data in Phase 1



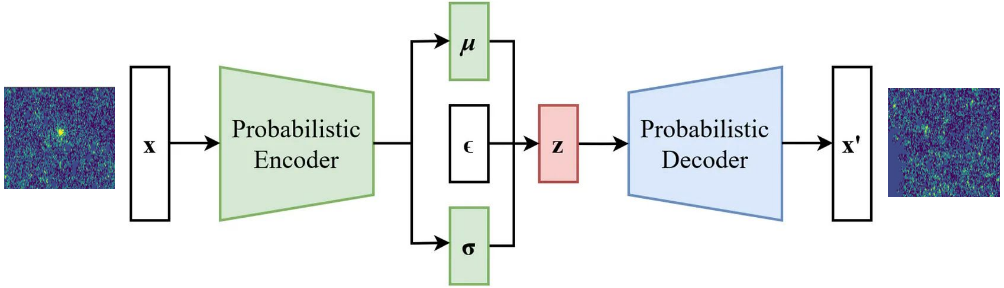
* Final dataset may be subject to change

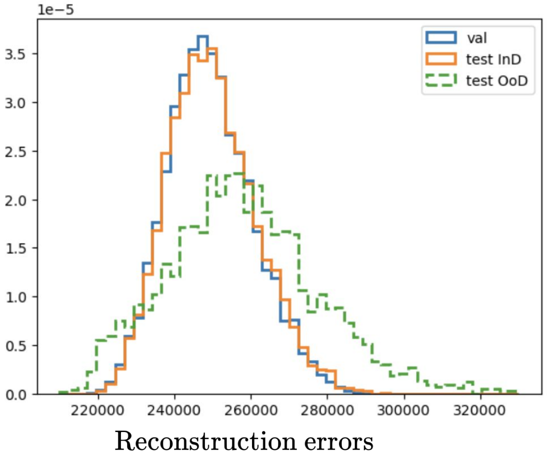
Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 2 Example Baselines

Autoencoder
Reconstruction error



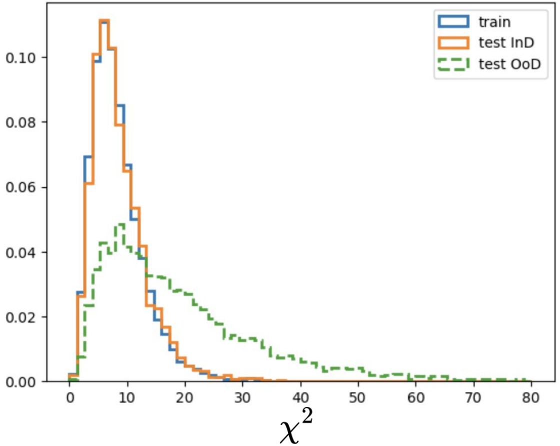


Then use Sellke–Bayarri–Berger method to calibrate p-value to obtain a lower bound of the Bayes Factor

Phase-1 baseline
Chi-square distribution

$$\chi^2(\Theta) = [d_{\text{obs}} - \mu(\Theta)]^T \,\text{Cov}^{-1}(\Theta)\, [d_{\text{obs}} - \mu(\Theta)]$$

Summary statistic:
matter power spectrum, CNN outputs…



Weak Lensing ML Uncertainty Challenge

FAIR Universe

Phase 2 Status and Timeline

Pre-register for the Phase 2 competition today!

- Please register with your **affiliation/company email address.**
- **Not yet open for submission. But you will receive a notification when the Phase 2 officially starts!**
- More information will be available on Codabench soon.
- Tackle impactful cosmology problem and win our monetary prizes!

Phase 2 competition website on Codabench

Envisioned competition schedule (UTC)		
Competition Phase	**Date**	**Description**
Phase 2	Mid December 2025 – Mid March 2026	Open submissions
	Mid March 2026 – End March 2026	Evaluating top submissions on hidden dataset
	End March 2026	Announcement of winners