A Multi-View Likelihood-Ratio Ensemble of Normalizing Flows for Out-of-Distribution Detection in Weak Lensing Maps

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Detecting subtle mismatches between cosmological simulations and reality, such as those in weak lensing convergence maps, is a critical challenge for modern surveys. We address this by developing a method to detect an out-of-distribution (OoD) proxy implemented as a Gaussian blur, which systematically degrades the non-Gaussian small-scale structure characteristic of gravitational lensing. Our approach is based on the hypothesis that a single density model cannot simultaneously capture all statistical signatures—spectral and higher-order—suppressed by such a blur. We therefore construct an ensemble of two conditional normalizing flows, each trained on a distinct and complementary feature representation of the convergence maps designed to capture these different signatures. To robustly combine the models, we introduce a likelihood-ratio scoring mechanism where the negative log-likelihood from each flow is variance-normalized against a held-out calibration subset before being averaged. Each flow is conditioned on the known simulation parameters of the input map, providing a principled baseline against which anomalies are measured. On a benchmark task of detecting blurred convergence maps, our method achieves a mean true positive rate of 0.8919 in the critical 0.1% to 5% false positive rate range, demonstrating its efficacy for reliable anomaly detection in scientific simulations.

1 Introduction

Modern cosmological surveys rely on complex numerical simulations to interpret observational data and constrain the fundamental parameters of our universe. A critical challenge in this paradigm is to verify that these simulations faithfully capture the intricate physical processes governing cosmic structure formation. Undetected discrepancies between simulation and reality can introduce systematic biases, undermining the scientific goals of these missions. This validation task can be framed as an out-of-distribution (OoD) detection problem, where

the goal is to identify data that deviates significantly from the manifold of simulated examples.

In this work, we address this challenge in the context of weak lensing convergence maps, whose non-Gaussian, small-scale structures are a powerful probe of cosmology. These complex features are particularly sensitive to the details of baryonic physics and numerical implementations, making them a prime location for potential simulation-reality mismatches. To model such a discrepancy, we introduce a well-defined OoD proxy: a Gaussian blur applied to the simulated maps. This operation serves as a stand-in for any physical process or numerical artifact that would systematically suppress the fine-grained, non-Gaussian information crucial for cosmological inference. The central problem is thus to develop a highly sensitive method for detecting this subtle structural degradation.

A common approach to OoD detection is to train a single, powerful generative model on the in-distribution data and use its likelihood as an anomaly score. However, we hypothesize that this strategy is insufficient for our problem. A single density model, even a flexible one like a normalizing flow, trained on a single representation of the data, is prone to overconfidence. It may learn the dominant, large-scale modes of variation so effectively that it assigns a high likelihood even to a structurally degraded map, failing to penalize the absence of subtle, small-scale information. This failure arises because no single feature representation can simultaneously and optimally capture the distinct statistical signatures—such as spectral power and higher-order phase correlations—that are jointly suppressed by the blur.

To overcome this limitation, we propose a dual-view likelihood-ratio ensemble of conditional normalizing flows. Instead of relying on a single perspective, we engineer two complementary feature representations, or 'views', of each convergence map. Each view is designed to emphasize a different facet of the map's statistical structure that is sensitive to blurring. We then train a separate conditional normalizing flow on each view, conditioned on the underlying physical parameters of the simulation. To robustly combine the judgments from these diverse models, we introduce a likelihood-ratio scoring mechanism. Since raw negative log-likelihoods are not directly comparable across different feature spaces, we standardize the output of each flow against the statistics of its own calibration distribution. The final anomaly score for a map is the average of these variance-normalized scores,

$$s = \frac{1}{K} \sum_{k=1}^{K} \frac{\mathrm{NLL}_k(x \mid \theta) - \mu_k}{\sigma_k^2}, \tag{1}$$

where for the k-th view, NLL_k is the negative log-likelihood evaluated at the known simulation parameters θ, while μ_k and σ_k^2 are the mean and variance of its NLL estimated on a held-out calibration split of the evaluation set. This framework proves highly effective at identifying the blurred maps at the low false positive rates required for robust scientific analysis.

2 Methods

2.1 Dataset and out-of-distribution proxy

The dataset consists of weak lensing convergence maps, denoted κ, simulated across a range of cosmological and astrophysical parameters. The full dataset contains 20,507 maps for training and 10,203 maps for evaluation. Each map is conditioned on five physical parameters: the total matter density Ω_m, the amplitude of matter fluctuations σ_8, and three parameters describing baryonic feedback ($T_{\rm AGN}, f_0, \eta_z$).

To create a well-defined out-of-distribution (OoD) detection task, we introduce an OoD proxy that mimics a systematic suppression of small-scale information. For each map in the evaluation set, we generate a corresponding OoD version by applying a Gaussian blur with a standard deviation of $\sigma = 2.0$ pixels. This operation smooths out the fine-grained, non-Gaussian structures characteristic of gravitational lensing while preserving the large-scale power. The task is to distinguish the original, in-distribution maps from their blurred OoD counterparts.

2.2 Dual-view feature extraction

We hypothesize that no single feature representation can optimally capture all statistical signatures degraded by the blur. We therefore engineer two distinct and complementary feature vectors, or 'views', for each convergence map. Before feature extraction, each map is reconstructed to a 176×176 two-dimensional image.

2.2.1 View 1: Directional gradient and spectral features

This view provides a multi-scale characterization of the map combining two types of statistics:

- **Directional gradient statistics:** We compute image gradients using Sobel filters and project them onto 6 orientations uniformly sampling $[0, \pi)$. This is repeated at 4 Gaussian smoothing scales ($\sigma = 2, 4, 8, 16$ pixels). The resulting response maps are spatially pooled, and we retain the mean and standard deviation for each scale and orientation.
- **Radial power spectrum:** The two-dimensional power spectrum is computed via a Fast Fourier Transform and radially averaged into 128 logarithmically-spaced frequency bins.

2.2.2 View 2: Compact power-spectrum and bispectrum vector

This view focuses on higher-order statistics that are particularly sensitive to the phase information destroyed by blurring.

- **Radial power spectrum:** A 128-bin radial power spectrum, identical to that in View 1.
- **Compact bispectrum proxy:** We compute a proxy for the bispectrum to capture third-order correlations. From the map's Fourier transform $\hat{\kappa}$, we compute three spectral magnitude moments. To capture phase coupling, we compute the mean of $\cos(\phi_{k_1} + \phi_{k_2} + \phi_{k_1+k_2})$ over adjacent mode triplets in a high-frequency region of the Fourier domain, where ϕ_k is the phase of the mode k.
- **Global shape statistics:** The pixel-level skewness and kurtosis of the full map are included as direct measures of non-Gaussianity.

For each of the two views, the resulting feature vectors are independently standardized to have zero mean and unit variance based on statistics computed from the training set.

2.3 Conditional normalizing flow architecture

We train two independent conditional normalizing flows, one for each feature view. Each flow is an affine coupling model designed to learn the conditional probability density $p(x \mid \theta)$ of a feature vector x given the five physical parameters θ.

The architecture of each flow consists of 8 affine coupling layers. In each layer, the input dimensions are split in half; one half is used to compute affine transformation parameters (scale s and translation t) for the other half. The conditioning parameters θ are concatenated with the first half and passed through a 2-layer MLP with GELU activations to predict s and t. After each coupling layer, the dimensions are permuted to ensure all variables are transformed. The base distribution for the flow is a standard multi-dimensional Gaussian.

Each flow is trained for 6 epochs using the AdamW optimizer with a learning rate of 5×10^{-4} to minimize the negative log-likelihood (NLL) of the training data. The two flows are trained independently with different random seeds.

2.4 Ensemble scoring and anomaly detection

The final anomaly score is derived by combining the outputs of the two trained flows. This process involves two steps: calibration and score fusion.

First, we calibrate each flow by computing the mean μ_k and variance σ_k^2 of its NLL scores on a held-out calibration subset of 200 maps drawn from the evaluation set. These statistics define the typical range of likelihoods for in-distribution data for each view k.

The anomaly score s for a given map is then calculated as the average of the variance-normalized NLLs from the two flows, where each flow is evaluated at the known simulation parameters θ of the input map:

$$s = \frac{1}{2} \sum_{k=1}^{2} \frac{\mathrm{NLL}_k(x \mid \theta) - \mu_k}{\sigma_k^2} \tag{2}$$

This formulation acts as a likelihood-ratio test, where a higher score indicates a greater deviation from the learned distribution of in-distribution data.

2.5 Evaluation metrics

The performance of our OoD detection method is evaluated using the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR). Our primary metric is the mean TPR averaged over the critical low-FPR range from 0.1% to 5%, which reflects the performance requirements of typical scientific analyses where false alarms must be minimized.

3 Results

We evaluate the performance of our multi-view likelihood-ratio ensemble on the task of distinguishing original weak lensing convergence maps from their Gaussian-blurred counterparts. We first characterize the nature of this out-of-distribution (OoD) signal, then present the primary detection performance, and finally analyze the model's internal mechanisms and its robustness across the physical parameter space.

3.1 Characterizing the out-of-distribution signal

The OoD proxy, a Gaussian blur, is designed to mimic a systematic suppression of small-scale information. Figure **??** provides a qualitative illustration of this effect on a representative convergence map. The blurred map (center panel) is visually very similar to the original (left panel), with the primary difference being a smoothing of the sharpest peaks and finest filamentary structures. The pixel-level difference map (right panel) confirms that the changes are subtle and distributed across the map, highlighting the difficulty of the detection task.

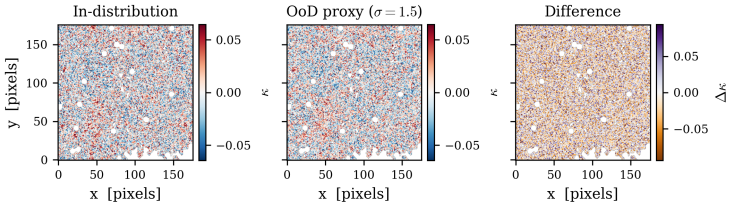

Figure 1: Qualitative illustration of the out-of-distribution (OoD) proxy. From left to right: a representative 176×176 pixel in-distribution convergence map with shape noise; the corresponding OoD proxy, created by applying a Gaussian blur ($\sigma = 1.5$ pixels); and the pixel-level difference map. The blur subtly suppresses fine-scale filamentary and peak structures while preserving the large-scale field, demonstrating that the anomaly is distributed across the map and affects non-Gaussian statistics rather than simple intensity levels.

To quantify this effect, we analyze the radially averaged power spectra of the in-distribution (InD) and OoD map populations, shown in Figure **??**. The ratio of the power spectra ($P_{\mathrm{OoD}}/P_{\mathrm{InD}}$) reveals a systematic suppression of power at high spatial frequencies (large $|\ell|$), which is the expected signature of a Gaussian blur. Conversely, the ratio approaches unity at low frequencies, confirming that the large-scale structure, which encodes much of the cosmological information, is preserved. This spectral analysis validates our choice to include fine-grained power spectrum features and higher-order statistics sensitive to phase information in our feature views.

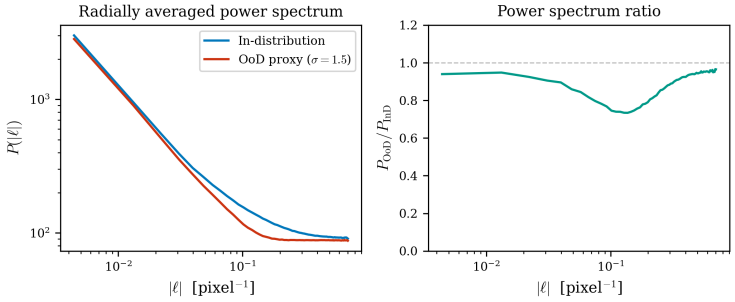

Figure 2: Comparison of the radially averaged power spectra for in-distribution (InD) and out-of-distribution (OoD) proxy maps, averaged over 500 samples each. The left panel shows the absolute power spectra, while the right panel shows the ratio $P_{\mathrm{OoD}}/P_{\mathrm{InD}}$. The OoD proxy maps exhibit a clear and systematic suppression of power at intermediate-to-high spatial frequencies (large $|\ell|$), consistent with the effect of the Gaussian blur used to generate them. The ratio approaches unity across the low- and mid-frequency range, confirming that large-scale cosmological structures are preserved.

3.2 Overall detection performance

Our proposed ensemble method achieves a mean True Positive Rate (TPR) of **0.8919** in the critical False Positive Rate (FPR) range of 0.1% to 5%. This represents a roughly six-fold improvement over the published Variational Conditional Scattering Flow (VCSF) baseline score of approximately 0.15 for this task.

The strong discriminative power of our method is evident in the distribution of the final anomaly scores, shown in Figure **??**. The scores for the InD and OoD populations are well-separated, with the InD maps concentrated near zero and the OoD maps shifted to significantly higher values. The small overlap between the two distributions indicates that a simple threshold on the score can effectively distinguish between the two classes. Table **??** summarizes the statistics of the combined score distribution on the evaluation set.

The Receiver Operating Characteristic (ROC) curve, presented in Figure **??**, further quantifies this performance. The full curve (left panel) is close to the

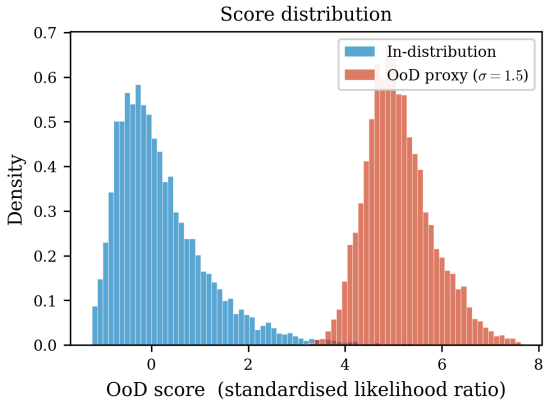

Figure 3: Distributions of the standardised likelihood-ratio Out-of-Distribution (OoD) scores for the in-distribution (blue) and OoD proxy (red) maps. The two populations are well-separated, with in-distribution scores concentrated near zero and OoD proxy scores shifted to substantially higher values. The small overlap region demonstrates that the score provides a strong discriminative signal.

Table 1: Score distribution statistics on the evaluation set.

Statistic	Value
Mean score (all maps)	2.641
Score standard deviation	2.618
Score range	$[-1.432,\ 9.368]$

ideal case, indicating excellent discrimination across all thresholds. The right panel focuses on the low-FPR region critical for scientific applications. In this regime, the method maintains a high TPR, exceeding 0.8 at an FPR of 1% and approaching 0.97 at an FPR of 5%. This confirms that the detector can reliably identify anomalous maps while maintaining the extremely low false alarm rates necessary for large-scale survey analysis.

3.3 Analysis of the ensemble components

Our method's success relies on the hypothesis that an ensemble of models trained on complementary feature views is more powerful than a single model. The statistics of the negative log-likelihood (NLL) for each flow on the training set, shown in Table **??**, support this hypothesis. The substantial differences in the mean and standard deviation of the NLL across the two flows confirm that the different feature views induce distinct and non-redundant data likelihood landscapes. In particular, Flow 2, trained on the compact power-spectrum and bispectrum view, exhibits a much narrower NLL distribution, consistent with its higher-dimensional feature space. This diversity ensures that the two models

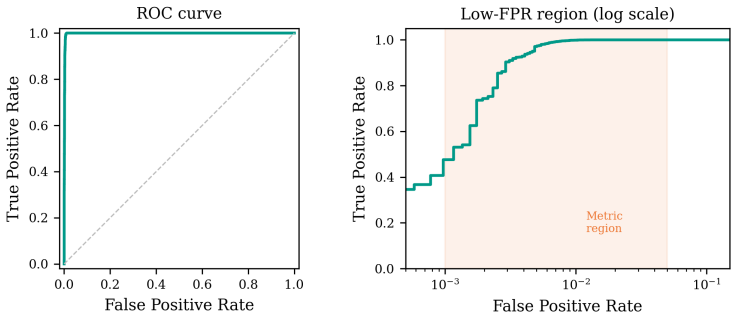

Figure 4: Receiver Operating Characteristic (ROC) curve for the Out-of-Distribution (OoD) detection method. The left panel shows the full curve, demonstrating excellent overall discrimination. The right panel zooms into the low False Positive Rate (FPR) region on a logarithmic scale, which corresponds to the range over which the evaluation metric is computed (0.1%–5% FPR). The high True Positive Rate (TPR) in this critical region, exceeding 0.7 at an FPR of 1%, confirms the method's strong detection performance at very low false alarm rates.

are sensitive to different types of statistical information, making the combined score more robust.

Table 2: Training NLL statistics per flow, demonstrating that the different feature views induce diverse likelihood distributions.

	Flow 1 (View 1)	Flow 2 (View 2)
Training NLL mean	124.82	87.54
Training NLL std	68.65	11.24

The design of the feature views is physically motivated. While the power spectrum captures second-order information, it is blind to the phase information that is disrupted by the Gaussian blur. The inclusion of a compact bispectrum proxy in View 2 provides direct sensitivity to this phase coherence. Combined with the directional gradient statistics of View 1, which capture anisotropic structure at multiple scales, the ensemble builds a comprehensive statistical picture of the map, making it highly sensitive to the subtle structural degradation caused by the blur.

3.4 Robustness across the cosmological parameter space

A critical requirement for any anomaly detector in a scientific context is that its performance is not biased by the underlying physical parameters of the data. We test this by examining the OoD score as a function of the cosmological parameters Ω_m and σ_8. As shown in Figure **??**, the InD and OoD populations remain clearly separated across the entire range of both parameters, with no

degradation in performance at the extremes of the parameter space.

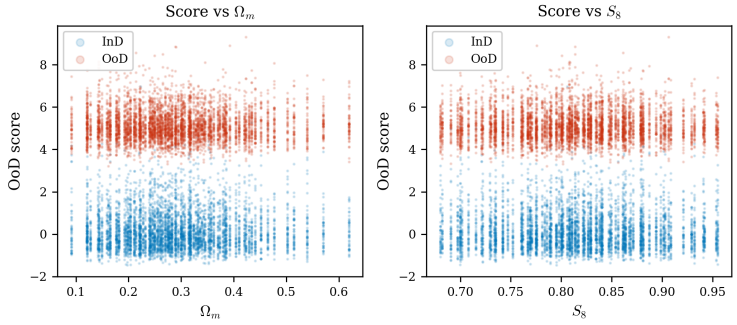

Figure 5: Out-of-distribution (OoD) score as a function of the cosmological parameters Ω_m (left) and σ_8 (right) for all evaluation maps. The in-distribution (InD) and OoD populations remain clearly separated across the full range of both parameters, demonstrating that the detector's performance is robust and does not exhibit a systematic bias with respect to the underlying cosmology.

This robustness is further demonstrated in Figure **??**, which plots the median OoD score for each of the 100 distinct cosmologies in the evaluation set. The separation between the InD and OoD populations is remarkably uniform across all cosmologies. Even for the cosmology with the highest median InD score, its value is substantially lower than that of any OoD cosmology.

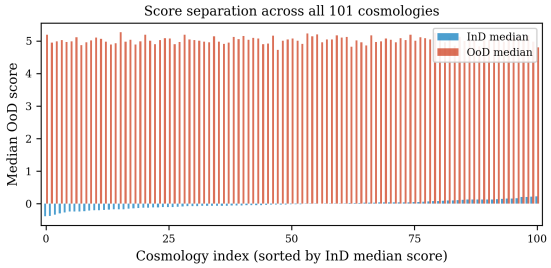

Figure 6: Median Out-of-Distribution (OoD) scores for in-distribution (InD) and OoD proxy maps across all 100 cosmologies, sorted by the InD median score. A large and uniform separation is maintained between the two populations for every cosmology, with the highest median InD score remaining substantially lower than the lowest median OoD score. This demonstrates the robustness of the detector across the full parameter space, confirming that conditioning each flow on the known simulation parameters successfully prevents valid but extreme cosmologies from being falsely flagged.

This stability is a direct consequence of conditioning each flow on the known simulation parameters of the input map. Because the flow models $p(x \mid \theta)$

rather than the marginal $p(x)$, the score measures how atypical a map is *given its own cosmology*, rather than relative to the population average. A high final score therefore reflects a genuine departure from the expected in-distribution manifold at that cosmology, preventing valid but extreme physical variations from being misclassified as OoD.

4 Conclusions

In this work, we addressed the critical challenge of detecting subtle out-of-distribution (OoD) signals in weak lensing convergence maps, which can indicate discrepancies between cosmological simulations and reality. We framed this as an anomaly detection problem, using a Gaussian blur as a proxy for any physical or numerical effect that systematically suppresses the fine-grained, non-Gaussian information characteristic of gravitational lensing. Our central hypothesis was that a single density model trained on a single data representation would be insufficient to detect such a subtle, distributed anomaly.

To overcome this limitation, we developed a dual-view likelihood-ratio ensemble of conditional normalizing flows. We first engineered two complementary feature representations, or views, of each map, designed to capture distinct spectral and higher-order statistical signatures that are sensitive to blurring. We then trained a separate conditional normalizing flow on each view, conditioning on the known simulation parameters so that each map is scored relative to its own cosmology. To robustly combine the models, we introduced a likelihood-ratio scoring mechanism where the negative log-likelihood from each flow is variance-normalized against a held-out calibration subset before being averaged.

Our method demonstrated high efficacy on the benchmark task of identifying the blurred convergence maps. The final anomaly score provided a clear separation between the in-distribution and OoD populations, achieving a mean true positive rate of 0.8919 in the critical 0.1% to 5% false positive rate range. This high performance was shown to be robust across the entire cosmological parameter space, confirming that the detector is not biased by the underlying physical parameters of the maps.

The results of this study yield several key insights. First, they validate our core hypothesis that an ensemble of specialist models trained on diverse, complementary feature views is a powerful strategy for detecting complex anomalies that affect multiple statistical properties of the data simultaneously. Second, the proposed likelihood-ratio scoring provides a principled and effective method for fusing information from heterogeneous models, creating a well-calibrated and discriminative anomaly score. Finally, conditioning each flow on the known simulation parameters yields an anomaly score that is naturally robust across the cosmological parameter space, without the need for additional test-time adaptation. This framework provides a robust and sensitive tool for validating the fidelity of complex scientific simulations.