Skepthical review: *Robust Detection of Simulation Mismatch in Weak Lensing Maps with Conditional Scattering-Flows*

Summary

This paper introduces the Variational Conditional Scattering-Flow (VCSF) framework to detect simulation mismatch in weak-lensing convergence maps by casting it as an out-of-distribution (OoD) detection problem (Sec. 1–2). Each map is encoded via a Wavelet Scattering Transform (WST) into **417** non-Gaussian coefficients, compressed with PCA to **3** dimensions ($>$ **97%** variance explained), and then "whitened" using an intra-cosmology covariance intended to reduce sensitivity to baryonic nuisance parameters (Sec. 2.2, Sec. 3.1). A conditional Neural Spline Flow (NSF) models the density $p(z|\theta)$ of the resulting **3**D features conditioned on five physical parameters ($\Omega_m, S_8, T_{\mathrm{AGN}}, f_0, \Delta z$), and an anomaly score is defined by profiling over parameters via gradient-based minimization of NLL, initialized by an MLP regressor (Sec. 2.3). Experiments on the NeurIPS **2025** FAIR Universe challenge data evaluate OoD detection using Gaussian-blurred maps ($\sigma = \mathbf{1.5}$ pixels) as a proxy mismatch, reporting ROC AUC **0.925** and pAUC **0.1488** over $\mathbf{FPR} \in [\mathbf{0.001}, \mathbf{0.05}]$ with visually clear InD/OoD score separation (Sec. 3.3), plus suggestive robustness to extreme AGN settings (Sec. 3.4). The pipeline is conceptually coherent and promising, but key components are under-specified (especially whitening and θ-optimization), the OoD evaluation is narrow (single blur proxy within one simulation suite, with potential leakage concerns), and baseline comparisons/metrics are not yet presented in a fully quantitative and reproducible way—making some broader claims in Sec. 1 and Sec. 4 premature without additional validation.

Strengths

- Timely and well-motivated framing: simulation mismatch detection for weak-lensing maps as OoD detection is relevant to upcoming precision cosmology surveys (Sec. 1).
- A clear, modular pipeline combining established tools (WST $\rightarrow$ PCA $\rightarrow$ whitening $\rightarrow$ conditional normalizing flow) in a way that is easy to reason about and potentially extensible (Sec. 2.2–2.3).
- Use of WST features is a principled choice to capture non-Gaussian, multiscale morphology beyond power spectra, with desirable stability properties under certain perturbations (Sec. 2.2).
- The observed compressibility of **417** WST coefficients to **3** PCs explaining $>$ **97%** of the variance is interesting and potentially scientifically informative (Sec. 3.1).
- Likelihood-based scoring that profiles over physical parameters is aligned with the goal of flagging structural anomalies while not penalizing valid parameter changes (Eq. (1), Sec. 2.3).

- Reported low-FPR detection performance on the chosen proxy OoD is strong and clearly visualized (ROC AUC **0.925**; pAUC **0.1488** for $\mathbf{FPR} \in [\mathbf{0.001}, \mathbf{0.05}]$; Fig. 3, Sec. 3.3).
- Figures generally support interpretability: training dynamics (Fig. 1), parameter-prediction behavior motivating the MLP initializer (Fig. 2), and InD/OoD score separation plus robustness plots (Fig. 3).

Major issues

1. **OoD evaluation is too narrow to substantiate claims about general "simulation mismatch" (or "new physics") detection: experiments use a single OoD proxy (Gaussian blur) and all maps come from a single simulation suite, with InD testing performed on a held-out subset of the same dataset (Sec. 2.1, Sec. 3, Sec. 4).** Moreover, OoD samples appear to be derived from the same underlying clean maps as the InD split (blurred versions), which can make the task closer to detecting a known corruption of familiar content rather than detecting genuinely unseen simulator behavior.

   Recommendation: In Sec. 3, broaden OoD tests beyond isotropic blur to cover qualitatively different structural shifts (e.g., anisotropic/PSF-like convolution, scale-dependent filtering that is not equivalent to smoothing, sharpening/artifacts, altered noise models, or baryonic-like transformations that modify profiles non-uniformly). If possible, evaluate on an independent simulation suite/code or the official challenge test protocol. If additional data are not feasible, explicitly narrow claims in Sec. 1 and Sec. 4 to "blur-like small-scale suppression within this simulation framework," add a limitations paragraph in Sec. 4, and clearly state whether OoD maps share the same underlying realizations as InD validation maps (and whether this could inflate apparent separability).

2. **The whitening step (central to nuisance-robustness claims) is under-specified and only partially validated: the "intra-cosmology covariance" and whitening transform are not defined mathematically, it is unclear whether whitening is global or cosmology-dependent, what samples enter the covariance estimate, how ill-conditioning is handled, and how much dependence is actually removed for all nuisance parameters ($T_{\mathrm{AGN}}, f_0, \Delta z$), not just one (Sec. 2.2, Sec. 3.1, Sec. 3.4).** There is also a conceptual ambiguity: whitening aims to suppress nuisance dependence, yet the flow is still conditioned on the nuisance parameters (Sec. 2.2–2.3).

   Recommendation: Expand Sec. 2.2 with explicit equations and notation: define Σ precisely (e.g., how you average/condition over cosmology vs. nuisance), define $\boldsymbol{\mu}$ if centering is applied, and state $z_{\mathrm{white}} = \Sigma^{-1/2}(z_{\mathrm{PCA}} - \boldsymbol{\mu})$. Specify the estimator (sample covariance vs. shrinkage/diagonal loading), numerical method for $\Sigma^{-1/2}$ (SVD/eigendecomposition), and whether the transform is global or varies with (Ω_m, S_8). In Sec.

3.1/3.4 add quantitative diagnostics before/after whitening: correlations or mutual information between features/scores and each parameter ($\Omega_m, S_8, T_{\rm AGN}, f_0, \Delta z$). Include an ablation (WST→PCA vs. WST→PCA→whiten) reporting both OoD metrics and parameter-sensitivity to justify the step and clarify how it complements (rather than conflicts with) conditioning in Sec. 2.3.

3. **The anomaly score relies on profiling over θ via gradient-based optimization, but the optimization is not fully specified or validated: only "10 Adam steps" from an MLP initialization are described, without learning rate, β parameters, constraints/ranges, parameterization/normalization, boundary handling, restarts, or evidence that 10 steps approximates $\min_\theta$ NLL (Sec. 2.3, Sec. 3.3).** If θ is unconstrained or under-optimized, scores can be biased (either inflated NLL due to poor minimization or deflated NLL by drifting to unphysical regions). Fig. 2 also suggests the MLP initializer may collapse for Δz and be noisy for nuisance parameters, which could materially affect profiling quality.

   Recommendation: In Sec. 2.3 explicitly define the feasible set Θ for Eq. (1) (e.g., box constraints matching the training priors) and describe the constrained optimization method (projection, squashing transforms, or penalty). Provide optimizer hyperparameters, step schedule, and whether the reported score is the minimum along the trajectory or the final iterate. In Sec. 3.3 add a convergence/sensitivity study: (i) number of steps (0/5/10/20/50), (ii) initialization (MLP vs. random vs. oracle/true θ if available), (iii) effect on AUC/pAUC and on score distributions. For Fig. 2, add quantitative metrics per parameter (RMSE/MAE/bias) and demonstrate that downstream profiling remains robust even when the initializer is poor (especially for Δz).

4. **Baseline comparisons are not yet quantitative and reproducible, weakening claims of superiority: baselines are described qualitatively and some numbers appear not clearly tied to the exact same dataset/split/metric, with no table detailing implementations, hyperparameter tuning, or uncertainty estimates (Abstract, Sec. 2.4, Sec. 3.3, Sec. 4).**

   Recommendation: Add a dedicated baseline table (Sec. 2.4 or Sec. 3.3) with fully specified methods and results computed under the same protocol: e.g., power spectrum + Gaussian likelihood; WST (**417**) + Gaussian; WST→PCA(**3**) + Gaussian; unconditional flow; conditional flow without whitening; conditional flow without θ-optimization (plug-in MLP θ); and an oracle θ score if available. Report ROC AUC and pAUC (with bootstrap CIs) for each method on the same InD/OoD split. Clearly distinguish your own runs from numbers taken from external sources, and temper claims accordingly.

5. **Key implementation details needed for reproducibility are missing or ambiguous across preprocessing and models: WST configuration (wavelet family, scales/orientations, padding, normalization, spatial averaging), PCA fitting protocol (training-only vs. full data; standardization), whiten-**

ing estimation protocol, and NSF/MLP architectures/training hyperparameters (Sec. 2.1–2.3, Sec. 3.1–3.3). This also connects to potential evaluation leakage if PCA/whitening are fit using validation data.

Recommendation: In Sec. 2.2–2.3 (or an Appendix), provide a complete specification: WST library and parameters; whether coefficients are averaged and how; PCA preprocessing (centering/scaling), and explicitly state PCA/whitening are fit on the training split only and then frozen. For NSF: number of coupling layers, spline bins, base distribution, conditioning network structure, hidden sizes/activations, optimizer settings, batch size, epochs, early stopping, weight decay. For MLP: architecture, loss, training settings, and validation metrics per parameter. Add a short ablation in Sec. 3.1/3.3 showing OoD performance vs. number of PCs (e.g., **2/3/5/10**) to justify choosing **3** beyond variance explained.

6. **Metric/reporting ambiguities affect interpretability: (i) pAUC definition/normalization is unclear and the stated random-guess baseline appears inconsistent with "mean TPR over FPR $\in [0.001, 0.05]$" (Sec. 2.4, Sec. 3.3); (ii) NLL values are strongly negative (Fig. 1, Sec. 3.2–3.3) but the exact NLL convention (per-sample/per-dimension, log base, reduction) is not stated, and claims such as "highly negative NLL indicates sharply peaked density" are not well-grounded without calibration; (iii) results appear to be shown for a single training run without seed variability.**

   Recommendation: In Sec. 2.4 write the exact pAUC formula used (including any division by interval width and any normalization) and correct/justify the random-guess baseline under that same definition. For NLL, specify the precise quantity plotted (per-sample vs. averaged; per-dimension recommended), log base, and why negative values are expected for continuous densities; consider reporting bits/dim or NLL/dim and adding a simple likelihood sanity check (e.g., shuffled features or a Gaussian baseline). Report mean±std (or CI) over multiple random seeds for key metrics (AUC/pAUC) and, where feasible, show variability in training curves (Fig. 1) or summarize it in text.

Minor issues

1. Some text uses "marginalized over θ" even though Eq. (1) defines a profiled objective (min over θ / max likelihood), which is conceptually different from marginal likelihood integration (Sec. 2.3, Sec. 3.3, Abstract, Sec. 4).

   Recommendation: Replace "marginalized" with "profiled" throughout, or explicitly define and (if intended) evaluate a true marginal score $-\log \int p(z_x|\theta)p(\theta)\,d\theta$. Keep terminology consistent across Methods, Results, and Conclusions.

2. The paper title/method name uses "Variational," but the presented θ step is closer to profiling/optimization than standard variational inference (no ELBO is defined), which may confuse readers (Sec. 2.3, Sec. 4).

Recommendation: Briefly justify the term "Variational" (e.g., variational optimization over θ) or rename/rephrase to avoid implying ELBO-based inference, especially if targeting an ML audience.

3. Dataset/protocol details are currently too brief for readers to assess realism and to reproduce: total map count, resolution and angular scale, parameter ranges/sampling, independence of realizations, noise model, and exact split construction beyond "20% validation" (Sec. 2.1, Sec. 3).

   Recommendation: Expand Sec. 2.1 with a concise but complete dataset/protocol summary, including map resolution and angular scale, parameter ranges for ($\Omega_m, S_8, T_{\rm AGN}, f_0, \Delta z$), sampling scheme, noise model, and whether the split is stratified by parameters. Explicitly state how OoD blurred maps are generated relative to the split (and whether they reuse the same realizations as InD validation).

4. Computational cost and scalability are not discussed, but are important because inference includes WST computation and per-map θ-optimization (Sec. 2–3).

   Recommendation: Add a short runtime paragraph (Sec. 3.3 or Sec. 4): WST extraction time per map, NSF training time/hardware, and per-map inference time including MLP+optimization. Comment on parallelization and how cost scales with map resolution and dataset size.

5. Overfitting/calibration discussion for the conditional density model is minimal; it is unclear how well-calibrated likelihoods are on held-out InD data beyond the training/validation curves (Sec. 3.2).

   Recommendation: Add one or two lightweight diagnostics (Sec. 3.2): train vs. validation NLL with seed variability, a simple coverage/PIT-style check if feasible, or comparison to a Gaussian baseline in the same feature space to contextualize likelihood quality.

6. Broader impact/usage guidance is limited: false positives/negatives in OoD flags could affect downstream cosmological inference if used operationally (Sec. 4).

   Recommendation: Add a brief limitations/broader-impact paragraph in Sec. 4 describing how OoD flags should be validated and not over-interpreted as "new physics," and discussing conservative thresholding and cross-checks in an analysis pipeline.

Very minor issues

1. Figure/caption presentation can be made more accessible and actionable: reliance on "left/right" panel references, small fonts, overplotting in Fig. 2, and missing direct annotations of best epoch/early stopping in Fig. 1 (Sec. 3.2–3.4; Fig. 1–3).

   Recommendation: Label subpanels (a,b,c,...) and refer to them explicitly in text; enlarge fonts and increase DPI; in Fig. 1 annotate the best validation epoch and early-stopping point; in Fig. 2 use transparency/hexbin and add per-panel metrics and/or

residual insets.

2. Minor notation and formatting inconsistencies (e.g., T_{AGN} vs. T_{AGN}, Δz notation, inconsistent hyphenation of "out-of-distribution", occasional heading-level/LaTeX artifacts, and stray HTML entities like ">") (Sec. 1–4).

   Recommendation: Standardize parameter notation and key terms throughout, fix heading levels, and remove HTML/line-break artifacts via a careful proofreading pass.

3. The mapping defining the final feature vector is not stated as a single explicit expression, which makes the pipeline slightly harder to follow (Sec. 2.2–2.3).

   Recommendation: Add one explicit definition such as $z_x := \mathbf{whiten}(\mathbf{PCA}(\mathbf{WST}(x)))$ (with centering specified) and use that notation consistently in Sec. 2.3 and Eq. (1).

Key statements and references

- • **On the benchmark weak lensing dataset with Gaussian-blurred maps used as a structural out-of-distribution proxy, the Variational Conditional Scattering-Flow (VCSF) method achieves a full ROC AUC of 0.9250 and a partial AUC of 0.1488 in the low false-positive-rate regime FPR $\in [0.001, 0.05]$, substantially outperforming standard baselines such as power-spectrum-based Gaussian statistics (which yield pAUC ≈ 0.05) and CNN-based parameter-estimation approaches (which yield pAUC ≈ 0.06–0.08).**
- *Reference(s):* (none)
- • **Using Wavelet Scattering Transform features with $J = 4$ spatial scales and $L = 8$ angular orientations (417 coefficients per map), Principal Component Analysis reveals that the top 3 principal components capture 97.35% of the total variance in the scattering coefficients (with individual explained variance ratios of 60.67%, 34.06%, and 2.62%), indicating that the complex morphological signatures of cosmology and baryonic feedback lie on a highly constrained, low-dimensional manifold.**
- *Reference(s):* (none)
- • **Training a Conditional Neural Spline Flow on the 3-dimensional whitened scattering features conditioned on the five physical parameters $\theta = \Omega_m, S_8, T_{\mathrm{AGN}}, f_0, \Delta z$ yields a best validation Negative Log-Likelihood of -3.2034 at epoch 9, with early stopping at epoch 19, demonstrating that the model learns a sharply peaked conditional density over the in-distribution feature space.**
- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper is largely methodological/descriptive with one explicit equation defining the anomaly score as an optimization over conditioning parameters. Most other mathematical operations (PCA compression, whitening via inverse square-root covariance, conditional density modeling with a flow, and pAUC definition) are described conceptually without full formal definitions or derivations, limiting the depth of symbolic verification.

Checked items

1. ✔ **Anomaly score as profile NLL** (Eq. (1), Sec. 2.3, p.4)
    - **Claim:** Defines $\mathrm{Score}(x) = \min_\theta[-\log p(z_x|\theta)]$ as the anomaly score for map x.
    - **Checks:** symbol/definition consistency, objective-function consistency
    - **Verdict:** PASS; confidence: high; impact: critical
    - **Assumptions/inputs:** $p(z|\theta)$ is a valid conditional probability density over z (continuous density)., z_x denotes the whitened feature vector extracted from x., The minimization is over a specified parameter domain Θ (not stated).
    - **Notes:** Mathematically coherent as a profile (maximum-likelihood) scoring rule: $\min_\theta(-\log p) = -\max_\theta \log p$. The only missing piece is the domain Θ of θ, which affects existence/meaning of the minimum (treated separately).
2. ✔ **Gradient ascent vs descent equivalence** (Abstract (p.1) vs Sec. 2.3 (p.4) vs Conclusions (p.8))
    - **Claim:** Parameters are optimized via gradient ascent to maximize likelihood, and equivalently via gradient descent to minimize NLL.
    - **Checks:** algebraic equivalence, sign consistency
    - **Verdict:** PASS; confidence: high; impact: minor
    - **Assumptions/inputs:** $\mathrm{NLL}(\theta) := -\log p(z_x|\theta)$ is differentiable in θ via the conditional flow.
    - **Notes:** Maximizing log-likelihood is equivalent to minimizing NLL, so the ascent/descent descriptions are consistent.
3. ✖ **Use of "marginalized over θ" vs min over θ** (Sec. 3.3 (p.6-7): "marginalized over θ"; Eq. (1) (p.4))
    - **Claim:** Text claims the score is an NLL marginalized over θ, while the equation defines a minimum over θ.
    - **Checks:** definition consistency

- **Verdict:** FAIL; confidence: high; impact: moderate
- **Assumptions/inputs:** "Marginalized" would mean integrating out θ with a measure/prior, not optimizing over θ.
- **Notes:** Marginalization corresponds to $-\log \int p(z_x|\theta)p(\theta)\,d\theta$ (or similar), whereas Eq. (1) is a profile likelihood score. The paper is internally consistent in using optimization operationally, but the term "marginalize" is mathematically inaccurate as written.

4. ⚠ **Feasible set for θ in the score definition** (Eq. (1), Sec. 2.3, p.4)

   - **Claim:** The minimization is "over the entire parameter space" without specifying constraints/bounds.
   - **Checks:** well-posedness of optimization statement, missing assumptions
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** Physical parameters typically have bounded support in the dataset/problem setup., Without constraints, the argmin may leave the physically meaningful region.
   - **Notes:** The mathematical meaning of $\mathbf{min}_\theta$ depends on the domain Θ. If θ is unconstrained, the optimum could occur outside the physical/data-supported range, changing the intended interpretation. The paper does not state Θ or constraint handling.

5. ✔ **Pipeline mapping for z_x (feature extraction + PCA + whitening)** (Sec. 2.2-2.3, p.3-4)

   - **Claim:** Compute WST features (**417**-d), compress to top-**3** PCA components, then apply a whitening transform to obtain **3**-d features z_x.
   - **Checks:** dimensional/shape consistency, symbol/definition consistency
   - **Verdict:** PASS; confidence: high; impact: minor
   - **Assumptions/inputs:** PCA produces a **3**D vector in $\mathbb{R}^3$., Whitening uses a $\mathbf{3} \times \mathbf{3}$ linear transform.
   - **Notes:** As described, all linear operations are dimensionally compatible: $\mathbf{417} \to \mathbf{3}$ then apply a $\mathbf{3} \times \mathbf{3}$ matrix to remain in $\mathbb{R}^3$.

6. ⚠ **Whitening via inverse square-root covariance** (Sec. 2.2, p.3)

   - **Claim:** Compute an intra-cosmology covariance isolating nuisance variance; apply its inverse square root to PCA features to decorrelate from nuisance parameters.
   - **Checks:** missing definitions, linear-algebra consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: moderate
   - **Assumptions/inputs:** A covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$ is defined and positive definite (or regularized)., Whitening is linear: $z_{\text{white}} = \Sigma^{-1/2} z$ (possibly centered).

- **Notes:** Using $\Sigma^{-1/2}$ is standard for whitening, but the paper does not define the covariance (what is conditioned on, how "intra-cosmology" is computed/aggregated, whether centering is applied, and whether Σ is regularized/invertible). Without this, the stated invariance/decoupling claim cannot be verified from the text.

7. ✔ **Conditional density notation $p(\text{features}|\theta)$ after whitening** (Sec. 2.3, p.4)
   - **Claim:** Train a conditional flow to learn $p(\text{features}|\theta)$ where θ includes cosmological and baryonic parameters, with features being the 3D whitened representation.
   - **Checks:** notation consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** The modeled random variable is the whitened 3D feature vector z., Conditioning vector θ is 5D.
   - **Notes:** Consistent at the notation level: z is 3D and θ is 5D throughout. Conceptually, whitening to remove dependence on some components of θ may reduce the need to condition on them, but this is not an internal mathematical contradiction.

8. ⚠ **pAUC definition as mean TPR over an FPR interval** (Sec. 2.4, p.4)
   - **Claim:** Defines pAUC as mean TPR for $\text{FPR} \in [0.001, 0.05]$.
   - **Checks:** definition clarity
   - **Verdict:** UNCERTAIN; confidence: low; impact: minor
   - **Assumptions/inputs:** ROC curve TPR(FPR) exists as a function along the threshold sweep.
   - **Notes:** Mathematically consistent if interpreted as a normalized integral over the interval, but the exact formula (normalized vs. unnormalized area, continuous vs. discrete approximation) is not provided, so the definition cannot be checked precisely from the PDF text alone.

Limitations

- The PDF text contains almost no explicit derivations beyond Eq. (1); PCA/whitening and flow likelihood expressions are described verbally without formulas, preventing step-by-step algebra verification.
- Definitions required to verify key linear-algebra claims (exact covariance used for whitening; centering; regularization; parameter-domain constraints) are omitted, leading to several UNCERTAIN verdicts.
- No unit system or dimensional analysis for physical parameters/features is provided; only shape/notation consistency could be checked.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

Executed **15** candidate numeric checks: **6** PASS, **1** FAIL, **8** UNCERTAIN. Verified several internal arithmetic relationships (dataset split, PCA variance, dimensionality reduction ratio, rounding consistency, and NLL contrast heuristic). One substantive inconsistency was found in the stated random-guess pAUC baseline for the specified low-FPR interval.

Checked items

1. ✔ **C1** (Page 3, Section 2.1 Dataset and OoD proxy)
    - **Claim:** Held out **20%** of the training data (**5,376** samples).
    - **Checks:** percent_to_count_consistency
    - **Verdict:** PASS
    - **Notes:** Back-computed total training samples = $5376/0.2 = 26880$ (integer) and verified $5376/26880 = 0.2$.
2. ⚠ **C2** (Page 3, Section 2.2 Feature extraction and decorrelation)
    - **Claim:** Scattering network with $J = 4$ and $L = 8$ yields a **417**-dimensional feature vector.
    - **Checks:** repeated_value_consistency
    - **Verdict:** UNCERTAIN
    - **Notes:** Cross-location repeated-consistency could not be verified from the available execution context.
3. ✔ **C3** (Page 3, Section 2.2 and Page 5, Section 3.1)
    - **Claim:** Top **3** principal components capture **97.35%** of total variance; individual explained variance ratios are **60.67%**, **34.06%**, **2.62%**.
    - **Checks:** sum_of_percentages
    - **Verdict:** PASS
    - **Notes:** $60.67 + 34.06 + 2.62 = 97.35$ matches the reported top-**3** total within rounding tolerance.
4. ✔ **C4** (Page 3, Section 2.2 and Page 5, Section 3.1)
    - **Claim:** PCA reduces **417** WST coefficients to **3** principal components (**3**-dimensional PCA features).
    - **Checks:** dimensionality_reduction_ratio
    - **Verdict:** PASS
    - **Notes:** Computed $compression_ratio = 3/417 = 0.0071942446$ and $percent_reduction = 99.28057554\%$.
5. ⚠ **C5** (Page 4, Section 2.3 Conditional density estimation and anomaly scoring)

- **Claim:** Conditional NSF uses **5** transform layers and hidden dimensions $[128, 128]$.
- **Checks:** repeated_hyperparameter_consistency
- **Verdict:** UNCERTAIN
- **Notes:** Cross-location repeated-consistency could not be verified from the available execution context.

6. ✔ **C6** (Page 4, Section 2.3; Page 3, Section 2.1; Page 5, Section 3.2)
   - **Claim:** Physical parameter vector θ is five-dimensional: $\Omega_m, S_8, T_{\mathrm{AGN}}, f_0, \Delta z$.
   - **Checks:** set_cardinality_consistency
   - **Verdict:** PASS
   - **Notes:** Counted **5** listed elements and matched the declared **5**-dimensional θ .
7. ⚠ **C7** (Page 4, Section 2.3)
   - **Claim:** Perform **10** steps of gradient descent using Adam optimizer.
   - **Checks:** repeated_constant_consistency
   - **Verdict:** UNCERTAIN
   - **Notes:** Cross-location repeated-consistency could not be verified from the available execution context.
8. ⚠ **C8** (Page 5, Section 3.2 and Page 6, Figure 1 caption)
   - **Claim:** Best validation NLL of $\mathbf{-3.2034}$ at epoch **9**; early stopping at epoch **19**.
   - **Checks:** repeated_value_consistency
   - **Verdict:** UNCERTAIN
   - **Notes:** Could not confirm exact matching across the two cited locations from the available execution context.
9. ⚠ **C9** (Page 4, Section 2.4; Page 7, Figure 3 caption; Page 8, Conclusions)
   - **Claim:** pAUC computed over $\mathbf{FPR} \in [\mathbf{0.001}, \mathbf{0.05}]$.
   - **Checks:** repeated_interval_consistency
   - **Verdict:** UNCERTAIN
   - **Notes:** Could not verify repeated-interval consistency across all cited locations from the available execution context.
10. ✔ **C10** (Page 7, Section 3.3 and Figure 3 caption)
    - **Claim:** Full ROC AUC is **0.925** (also written as **0.9250**).
    - **Checks:** rounding_consistency
    - **Verdict:** PASS
    - **Notes:** **0.925** and **0.9250** are numerically identical within tolerance.

11. ⚠ **C11** (Page 1, Abstract; Page 7, Figure 3 caption; Page 7, Section 3.3; Page 8, Conclusions)
    - **Claim:** Partial AUC (pAUC) is $\mathbf{0.1488}$ in the low-FPR regime.
    - **Checks:** repeated_value_consistency
    - **Verdict:** UNCERTAIN
    - **Notes:** Could not verify identical repetition across all cited locations from the available execution context.
12. ✖ **C12** (Page 7, Section 3.3)
    - **Claim:** Random-guess threshold for pAUC is ($\sim \mathbf{0.05}$) given FPR regime $[\mathbf{0.001}, \mathbf{0.05}]$.
    - **Checks:** baseline_value_recomputation
    - **Verdict:** FAIL
    - **Notes:** Under the stated definition (mean TPR over the FPR interval) and random guess $\mathbf{TPR} = \mathbf{FPR}$, expected baseline is $\mathit{mean}(\mathbf{FPR}) = (\mathbf{0.001} + \mathbf{0.05})/\mathbf{2} = \mathbf{0.0255}$, not $\sim \mathbf{0.05}$; suggests a different definition/normalization.
13. ✔ **C13** (Page 8, Section 3.4 Robustness to baryonic nuisance parameters)
    - **Claim:** Mean NLL scores: InD $= -\mathbf{3.6826}$; high-$T_{\mathrm{AGN}} = -\mathbf{3.7734}$; OoD $= -\mathbf{0.6947}$; claimed 'nearly identical' for InD vs high-T_{AGN} and 'in stark contrast' to OoD.
    - **Checks:** difference_magnitude_check
    - **Verdict:** PASS
    - **Notes:** Computed |InD-high$T_{\mathrm{AGN}}| = \mathbf{0.0908}$ and |InD-OoD$| = \mathbf{2.9879}$; ratio $= \mathbf{32.91}$ ($> \mathbf{10}$ heuristic threshold), supporting the qualitative contrast.
14. ⚠ **C14** (Page 6-7, Figure 3 description (text), Page 7 Section 3.3)
    - **Claim:** Extreme AGN feedback subset defined as $T_{\mathrm{AGN}} > \mathbf{8.3}$.
    - **Checks:** threshold_repeated_value_consistency
    - **Verdict:** UNCERTAIN
    - **Notes:** Could not verify repeated threshold consistency across all definitions from the available execution context.
15. ⚠ **C15** (Page 3, Section 2.1 Dataset and OoD proxy)
    - **Claim:** Gaussian blur applied with standard deviation $\sigma = \mathbf{1.5}$ pixels.
    - **Checks:** repeated_parameter_consistency
    - **Verdict:** UNCERTAIN
    - **Notes:** Could not verify repeated-mention consistency of σ across the document from the available execution context.

Limitations

- Only the provided parsed PDF text was used; no external datasets or internet sources were consulted.
- No values were extracted from plots/images (e.g., ROC curves or training curves) beyond the numbers explicitly stated in the text/captions.
- Several performance claims (AUC/pAUC, convergence quality, regression quality) are not directly verifiable without underlying validation data or tabulated ROC/score outputs.
- Multiple intended cross-location repeated-consistency checks (e.g., repeated hyperparameters/thresholds/metrics across sections and captions) were not verifiable from the available execution context.