Detecting Model Misspecification in Cosmology with Scale-Dependent Normalizing Flows

Aizhan Akhmetzhanova [1 2] **Carolina Cuesta-Lazaro** [3 2 4] **Siddharth Mishra-Sharma** [4 2 5 1 †]

Abstract

Upcoming cosmological surveys will generate unprecedented datasets, but validating whether our theoretical models accurately describe the observed Universe remains a fundamental challenge. We present a novel framework combining scale-dependent neural summary statistics with normalizing flows to detect model misspecification in cosmological simulations through Bayesian evidence estimation. By conditioning both compression and evidence networks on smoothing scale, we systematically identify where theoretical models break down in a data-driven manner. We demonstrate our approach using matter and gas density fields from three CAMELS simulation suites with different subgrid physics implementations.

1. Introduction

Observational cosmology now faces a growing number of tensions that challenge the standard ΛCDM model (Abdalla et al., 2022). These anomalies have so far been found by either i) direct comparison of parameter constraints derived from different cosmic epochs, where agreement would validate our understanding of cosmological evolution, or ii) specific parametric extensions for beyond ΛCDM models. In this paper, we present a complementary approach focused on identifying model-independent anomalies. That is, once the baseline model or training dataset has been specified, our approach for detecting anomalous data does not require that the anomalies follow a particular functional form or parameterization.

We take advantage of advances in large datasets of numerical simulations (Pakmor et al., 2023; Maksimova et al., 2021; Villaescusa-Navarro et al., 2021) and high-dimensional inference techniques for solving complex inverse problems. Although most research at this intersection has focused on parameter estimation and optimal information extraction, we propose a shift in perspective to tackle a complementary question: How can we systematically identify significant discrepancies between our theoretical models, as represented by numerical simulations, and the observed Universe?

Finding discrepancies between simulations and observations also addresses the critical need for robust goodness-of-fit metrics in high-dimensional spaces to validate our inference methods, where traditional approaches like chi-square statistics are insufficient due to their limitations with high-dimensional data and Gaussianity assumptions.

Detecting out-of-distribution (OOD) data is crucial to ensuring that trained machine learning systems are applied in a safe and reliable manner, given that small shifts in the data distribution can introduce large biases in the parameters of interest, see (Horowitz & Melchior, 2022; Mudur et al., 2024) for examples from cosmology. To this end, the topic of OOD detection has attracted increasing interest from the machine learning research community (Yang et al., 2024). Our work builds on recent advances in neural density estimation and anomaly detection to develop a scale-dependent framework for identifying model misspecification in cosmological analyses. Moreover, physical models in cosmology often have well-defined domains of validity that depend on spatial scale. By explicitly incorporating scale dependence into our analysis, we can identify at which scales theoretical models begin to break down.

Previously, Dai and Seljak (2024) explored the idea of identifying anomalies and distribution shifts due to scale-dependent systematics at the field-level with Multiscale Flows, which hierarchically decompose cosmological fields into lower-resolution approximations using a wavelet basis

[1]Department of Physics, Harvard University, 17 Oxford Street, Cambridge, MA 02138, USA [2]The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [3]Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA [4]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [5]Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Correspondence to: Aizhan Akhmetzhanova <aakhmetzhanova@g.harvard.edu>.
[†] Currently at Anthropic; worked performed while at MIT/IAIFI.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

and use normalizing flows to model each wavelet component separately and learn the full field-level likelihood function. They applied Multiscale Flows to simulated weak lensing datasets and found that in addition to significantly improving cosmological inference analysis, these flows are also able to detect small-scale OOD shifts such as addition of baryonic effects. Here, we develop an alternative framework for multiscale neural posterior estimation such that we can scale to higher dimensional datasets.

Concurrent to our work, Diao, Dai, and Seljak (2025) has extended (Dai & Seljak, 2024) by approaching the task of model validation as an OOD detection problem. Similarly to Dai and Seljak (2024), they work with simulated weak lensing datasets generated from dark-matter only simulations and corrections from baryonic effects. They examine multiple various statistics on these datasets, and find that continuous time flow models (CFTMs) at the field level consistently achieve the best performance for their OOD detection task.

These studies highlight the relevance of density-based OOD detection methods in cosmology in the context of model validation and anomaly detection. Our work complements the studies above and extends this line of work in new directions. In particular, our contributions are: 1) we develop a general training strategy for obtaining constraints as a function of scale on one single model, 2) we demonstrate that estimating evidence as a function of scale on learned summary statistics which are optimized for cosmological parameter inference can be used for detecting model misspecification through an example using the CAMELS simulations.

2. Methodology

Figure 1 provides a schematic overview of our framework.

Learning Sufficient Neural Statistics: We first extract sufficient summary statistics from high-dimensional cosmological fields using Variational Mutual Information Maximization (VMIM) (Lanzieri et al., 2024; Jeffrey et al., 2021; Ho et al., 2024). A statistic t is sufficient for parameters θ if $I(x, \theta) = I(t, \theta)$, preserving all parameter-relevant mutual information I from the original data x.

VMIM maximizes $I(t, \theta)$ via the loss function $L_{\text{VMIM}} = -\log q(\theta|F_\phi(x); \phi')$, where F_ϕ is a compressor network extracting statistics $t = F_\phi(x)$, and $q(\theta|F_\phi(x); \phi')$ is a normalizing flow-based posterior inference network. We train both networks jointly to learn optimal summary statistics.

Incorporating scale conditioning: Of particular importance in physics is how out-of-distribution metrics vary as a function of scale. This is especially relevant given that we often have accurate descriptions of large-scale phenomena through effective field theories, while missing crucial details at smaller scales with higher frequency components. We incorporate this by conditioning our compressor network on smoothing scale k_s: $t_s = F_\phi(x_s)$, where x_s is a Gaussian-smoothed field. We pass k_s as an additional channel prior to the first convolutional layer, enabling evaluation of statistics at any scale with the same network parameters.

Evidence estimation with normalizing flows: Starting from our learned summary statistics, we use the Bayesian evidence to assess model misspecification. For a model with parameters θ and data x, the evidence is defined as the probability of the data x under the given model marginalized over all model parameters: $p(x) = \int p(x|\theta)p(\theta)d\theta$.

This integral is often intractable in high dimensions, but machine learning approaches can address this challenge (Skilling, 2004; Papamakarios et al., 2021; Jeffrey & Wandelt, 2024; Polanska et al., 2024). In this work, we focus on density estimation using normalizing flows applied to the low-dimensional latent space of our summary statistics t_s. Specifically, after learning $t_s = F_\phi(x_s)$, we train normalizing flows to model the density $p(t_s)$ in this lower-dimensional space, also as a function of smoothing scale. The reduction in dimensionality makes density estimation significantly more tractable and avoids many of the challenges typically associated with density estimation for OOD detection (Nalisnick et al., 2018; Zhang et al., 2021), while our VMIM-optimized summaries retain the key information needed for parameter inference. We refer to the the normalizing flow trained to estimate the density of the compressed representations $p(t_s)$ as the *evidence estimation network*. Once trained, the normalizing flow allows both to estimate the log-evidence $\log p(t_s)$ of a given summary statistic t_s and to sample from the estimated probability distribution of the summary statistics for a given k_s.

We use the evidence as an out-of-distribution score by comparing observed and simulated samples. If the observed representations fall outside the distribution of simulated ones, this indicates that parameter constraints may be biased, pointing to missing physics in our simulations. Detecting misspecification in raw data would not necessarily lead to this conclusion, as anomalies in data space might reflect physics irrelevant to our cosmological constraints. By examining how the Bayesian evidence varies across different physical scales, we can identify specific scales where discrepancies between our simulations and observations are most prominent.

However, *any* OOD detection method has a fundamental limitation: observations may appear in-distribution yet still yield biased constraints when different physical effects produce degenerate features in our summary statistics. The most robust solution is combining complementary probes that respond differently to various effects, breaking degeneracies through joint analysis.

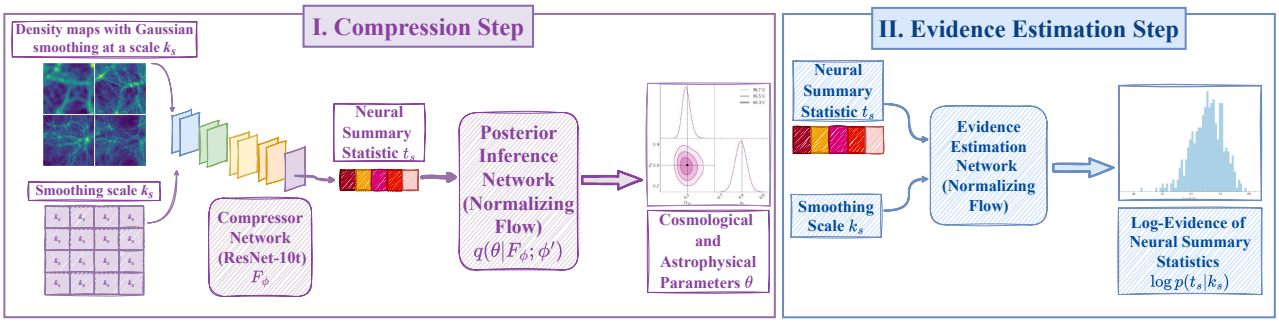

Figure 1. A schematic overview of the framework for OOD detection implemented in this study. Left: We first learn sufficient neural summary statistics of the data by training a compressor network with a Variational Mutual Information Maximization (VMIM) loss. Right: We train a normalizing flow to estimate the evidence of the learned summary statistics as a function of smoothing scale k_s.

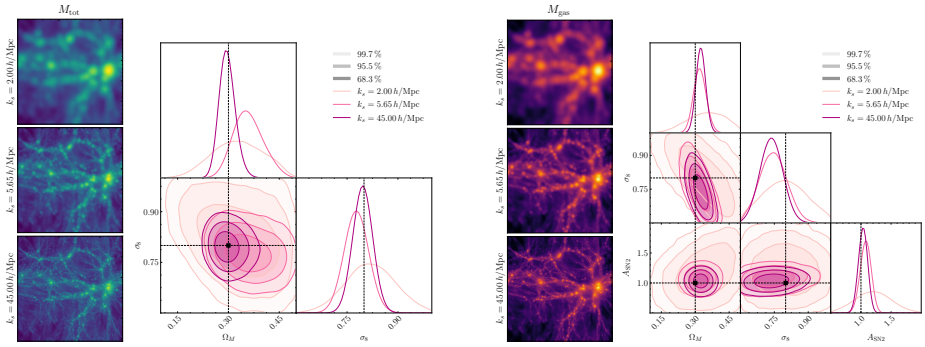

Figure 2. $M_{\rm tot}$ (left) and $M_{\rm gas}$ (right) fields from the Astrid CV set with various levels of Gaussian smoothing and corresponding posteriors over astrophysical and/or cosmological parameters. The true parameter values are indicated with dashed lines. The smoothing scales k_s plotted are 2, 5.65, and 45 $h/{\rm Mpc}$. As we include higher frequency information, the posteriors become tighter and concentrate closer to the true parameter values, validating scale-dependence of our neural summary statistics.

3. Experiments

Dataset and Training: Our dataset consists of total matter density ($M_{\rm tot}$) and gas density ($M_{\rm gas}$) fields at redshift $z = 0$ from the CAMELS project (Villaescusa-Navarro et al., 2021; 2022; Ni et al., 2023) from three suites of hydrodynamical simulations: IllustrisTNG (Weinberger et al., 2017; Pillepich et al., 2018), Astrid (Bird et al., 2022; Ni et al., 2022), and SIMBA (Davé et al., 2019).

As the baseline dataset for training our models, we use the Latin Hypercube (LH) set of the Astrid suite, which varies both cosmological (Ω_M, σ_8) and astrophysical parameters ($A_{\rm SN1}$, $A_{\rm SN2}$, $A_{\rm AGN1}$, and $A_{\rm AGN2}$). The LH set consists of 1000 independent simulations and 15 000 corresponding maps for each astrophysical field of interest. In addition to testing the parameter estimation performance of our models on a holdout test set from the Astrid LH set, we also assess the capabilities of our trained models to detect model misspecification on the Cosmic Variance (CV) sets of the three suites. CV sets are designed to explore the effects of cosmic variance on various cosmological and astrophysical probes: each set consists of 27 simulations (with 405 maps per astrophysical field) with the same values of cosmological and astrophysical parameters, but varied initial conditions.

To mimic a realistic observational scenario, we implement the following experimental design: we use Astrid as the baseline physics model, while treating IllustrisTNG and SIMBA as "mock observable universes" that may be out-of-distribution relative to our baseline. Specifically, we train our density estimation models on the Astrid LH set, then use the CV sets from all three suites to assess model misspecification. This approach mirrors what we would do with actual observations — splitting the observed volume into smaller subvolumes to assess statistical significance, while all subvolumes share the same underlying physical parameters. This setup provides a controlled test of how differences in subgrid physics implementations affect our ability to detect model misspecification and obtain unbiased cosmological constraints.

Detailed description of neural network architectures and training hyperparameters can be found in App. A and B.

Summary statistics validation: In Figure 2, we demonstrate how our summary statistics effectively capture cosmological information across different physical scales. We present parameter constraints as a function of smoothing

scale for $M_{\rm tot}$ and $M_{\rm gas}$ fields from the Astrid test set. For $M_{\rm tot}$, the learned statistics primarily constrain cosmological parameters. $M_{\rm gas}$ fields are able to additionally constrain the supernovae feedback parameter $A_{\rm SN2}$, which reflects the sensitivity of gas distribution to baryonic feedback processes that have less significant impact on the matter distribution.

As expected, incorporating higher frequency information (smaller scales) leads to significantly tighter parameter constraints, confirming that our neural summarizer successfully extracts scale-dependent information. This scale-dependent behavior is crucial for understanding which physical processes dominate at different scales and for detecting model misspecification. In all cases, our method successfully recovers the true parameters within the expected uncertainty, validating the accuracy of our neural posterior estimation approach. A more thorough validation is shown in App. D.

Anomaly detection with neural summaries: We next leverage our neural summary statistics to detect model misspecification between the simulation suites. We evaluate the log-evidence $\log p(t_s)$ of the summary statistics as a function of smoothing scale for the CV sets from the three suites using the evidence estimation network trained exclusively on the Astrid suite, our reference model. We plot the evidence distributions for three different scales in App. E.

To quantify the differences in $\log p(t_s)$ systematically, we compute the mean percentile rank of evidence values for each CV suite relative to the Astrid reference distribution. While the samples within each CV suite are not strictly independent (some of them are projections of the same simulation box), the percentile ranking provides a robust measure of relative model discrepancy. We show these results for $M_{\rm tot}$ and $M_{\rm gas}$ fields in solid lines (right y-axis) in Figure 3 with error bars representing the standard error on the mean percentile.

For $M_{\rm tot}$, SIMBA's percentile rank drops dramatically from over 50% at large scales to about 20% at small scales, indicating increasingly significant model differences. IllustrisTNG shows a much more moderate decrease to around 45%, maintaining this level across intermediate and small scales. This suggests that while both models differ from Astrid, SIMBA implements substantially different small-scale physics that affects the total matter distribution. For $M_{\rm gas}$, both simulations show considerably lower percentile ranks at small scales, dropping below 20%, confirming their gas distributions are highly distinct from Astrid, even at the level of the summary statistics. These results demonstrate our method's effectiveness at quantifying simulation differences across scales and provide valuable insights into which physical scales and components are most affected by modeling choices.

OOD detection and parameter bias: Physical degeneracies between different simulation models can complicate OOD detection, which creates two challenging scenarios: (1) cases where model misspecification remains undetected yet produces biased parameter constraints, and (2) cases where data appear strongly out-of-distribution while still yielding unbiased cosmological constraints.

In Figure 3, we quantitatively analyze the relationship between OOD detection and parameter bias across our simulation suites. We plot both the average percentile rank (our OOD metric) and the Mahalanobis distance (our bias measure) as functions of smoothing scale. We compute the Mahalanobis distance only for cosmological parameters on IllustrisTNG and SIMBA LH test sets, excluding astrophysical parameters, since these represent different physical quantities across subgrid models.

For $M_{\rm gas}$, we observe a notable correlation between the OOD detection strength and parameter bias. As the percentile ranks drops on smaller smoothing scales, the Mahalanobis distance increases appreciably. This correlation suggests that the differences in baryonic physics between simulation suites directly impact the summary statistics most relevant to cosmological parameter inference, making the observables related to the gas fields a potential probe for detecting physically relevant model misspecification.

For $M_{\rm tot}$, however, we find an interesting exception, similar to the case (2) mentioned above. Despite significant OOD detection for SIMBA on smaller scales, the inferred parameter bias remains mild throughout. However, this effect could also be in part due to the limitations of Mahalanobis distance, which, despite being an informative measure of bias, fails to penalize correlated biases in the inferred posteriors.

4. Conclusions

We have presented a general framework for detecting model misspecification in cosmological datasets as a function of scale. We propose a training strategy for constructing neural summary statistics of cosmological fields by training a single neural network model which is conditioned on the smoothing scale applied to the fields. The learned statistics can then be used (1) for obtaining constraints on the parameters of interest, and (2) as an informative low-dimensional representation of the data for Bayesian evidence estimation.

Using fields from CAMELS simulations a proof-of-concept, we demonstrate that this framework allows us to detect the discrepancies in the subgrid physics between different simulations. An application to realistic cosmological observations, such as observations from the DESI survey, would require more complex training data, with large simulation volumes and realistic survey systematics (Lemos et al., 2023b). From the methodological point of view, it would be valuable to consider working with 3-D point clouds, which

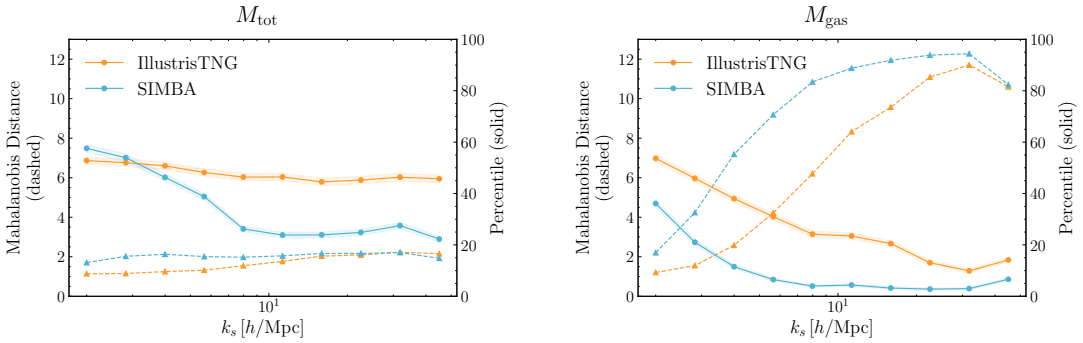

Figure 3. Mahalanobis distance (dashed) and percentiles of the $\log p(t_s)$ values (solid) for the $M_{\rm tot}$ (left) and $M_{\rm gas}$ (right) fields. The parameter bias largely correlates with the strength of OOD detection of the summary statistics in all scenarios, except for SIMBA's $M_{\rm tot}$ fields, for which the inferred parameter bias remains relatively low across the smoothing scales, despite the data appearing more strongly out-of-distribution.

provide a natural representation for galaxy distributions, and to perform the model misspecification analysis either in the compressed latent space of the point clouds or at the field-level by working with, for instance, diffusion-based generative models (Cuesta-Lazaro & Mishra-Sharma, 2024; Nguyen et al., 2024). These additional follow-up studies are necessary to effectively extend this framework to real survey data and to make steps towards improved, more robust modeling of our Universe.

Software and Data

Code used to reproduce the results of this paperis available at `https://github.com/AizhanaAkhmet/model-misspecification.git`. This research made extensive use of the `Jupyter` (Kluyver et al., 2016), `Matplotlib` (Hunter, 2007), `Numpy` (Harris et al., 2020), `PyTorch` (Paszke et al., 2019), `PyTorch-Lightning` (Falcon et al., 2020), `Scipy` (Virtanen et al., 2020), and `WANDB` (Biewald, 2020) packages.

Acknowledgements

We would like to thank Francisco Villaescusa-Navarro for helpful discussions. AA thanks the LSST-DA Data Science Fellowship Program, which is funded by LSST-DA, the Brinson Foundation, the WoodNext Foundation, and the Research Corporation for Science Advancement Foundation; her participation in the program has benefited this work. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of High Energy Physics of U.S. Department of Energy under Grant No. DE-SC0012567. This work was performed in part at the Aspen Center for Physics, which is supported by National Science Foundation grant PHY-2210452. The computations in this paper were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Impact Statement

This paper presents work whose goal is to advance the fields of Cosmology and Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abdalla, E., Abellán, G. F., Aboubrahim, A., Agnello, A., Akarsu, , Akrami, Y., Alestas, G., Aloni, D., Amendola, L., Anchordoqui, L. A., Anderson, R. I., Arendse, N., Asgari, M., Ballardini, M., Barger, V., Basilakos, S., Batista, R. C., Battistelli, E. S., Battye, R., Benetti, M., Benisty, D., Berlin, A., de Bernardis, P., Berti, E., Bidenko, B., Birrer, S., Blakeslee, J. P., Boddy, K. K., Bom, C. R., Bonilla, A., Borghi, N., Bouchet, F. R., Braglia, M., Buchert, T., Buckley-Geer, E., Calabrese, E., Caldwell, R. R., Camarena, D., Capozziello, S., Casertano, S., Chen, G. C.-F., Chluba, J., Chen, A., Chen, H.-Y., Chudaykin, A., Cicoli, M., Copi, C. J., Courbin, F., Cyr-Racine, F.-Y., Czerny, B., Dainotti, M., D'Amico, G., Davis, A.-C., de Cruz Pérez, J., de Haro, J., Delabrouille, J., Denton, P. B., Dhawan, S., Dienes, K. R., Di Valentino, E., Du, P., Eckert, D., Escamilla-Rivera, C., Ferté, A., Finelli, F., Fosalba, P., Freedman, W. L., Frusciante, N., Gaztañaga, E., Giarè, W., Giusarma, E., Gómez-Valent, A., Handley, W., Harrison, I., Hart, L., Hazra, D. K., Heavens, A., Heinesen, A., Hildebrandt, H., Hill, J. C., Hogg, N. B., Holz, D. E., Hooper, D. C., Hosseininejad, N., Huterer, D., Ishak, M., Ivanov, M. M., Jaffe, A. H., Jang, I. S., Jedamzik, K., Jimenez, R., Joseph, M., Joudaki, S., Kamionkowski, M., Karwal, T., Kazantzidis, L., Keeley, R. E., Klasen, M., Komatsu, E., Koopmans, L. V., Kumar, S., Lamagna, L., Lazkoz, R., Lee, C.-C., Les-

gourgues, J., Levi Said, J., Lewis, T. R., L'Huillier, B., Lucca, M., Maartens, R., Macri, L. M., Marfatia, D., Marra, V., Martins, C. J., Masi, S., Matarrese, S., Mazumdar, A., Melchiorri, A., Mena, O., Mersini-Houghton, L., Mertens, J., Milaković, D., Minami, Y., Miranda, V., Moreno-Pulido, C., Moresco, M., Mota, D. F., Mottola, E., Mozzon, S., Muir, J., Mukherjee, A., Mukherjee, S., Naselsky, P., Nath, P., Nesseris, S., Niedermann, F., Notari, A., Nunes, R. C., Ó Colgáin, E., Owens, K. A., Özülker, E., Pace, F., Paliathanasis, A., Palmese, A., Pan, S., Paoletti, D., Perez Bergliaffa, S. E., Perivolaropoulos, L., Pesce, D. W., Pettorino, V., Philcox, O. H., Pogosian, L., Poulin, V., Poulot, G., Raveri, M., Reid, M. J., Renzi, F., Riess, A. G., Sabla, V. I., Salucci, P., Salzano, V., Saridakis, E. N., Sathyaprakash, B. S., Schmaltz, M., Schöneberg, N., Scolnic, D., Sen, A. A., Sehgal, N., Shafieloo, A., Sheikh-Jabbari, M., Silk, J., Silvestri, A., Skara, F., Sloth, M. S., Soares-Santos, M., Solà Peracaula, J., Songsheng, Y.-Y., Soriano, J. F., Staicova, D., Starkman, G. D., Szapudi, I., Teixeira, E. M., Thomas, B., Treu, T., Trott, E., van de Bruck, C., Vazquez, J. A., Verde, L., Visinelli, L., Wang, D., Wang, J.-M., Wang, S.-J., Watkins, R., Watson, S., Webb, J. K., Weiner, N., Weltman, A., Witte, S. J., Wojtak, R., Yadav, A. K., Yang, W., Zhao, G.-B., and Zumalacárregui, M. Cosmology intertwined: A review of the particle physics, astrophysics, and cosmology associated with the cosmological tensions and anomalies. *Journal of High Energy Astrophysics*, 34:49–211, June 2022. ISSN 2214-4048. doi: 10.1016/j.jheap.2022.04.002. URL http://dx.doi.org/10.1016/j.jheap.2022.04.002.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Bird, S., Ni, Y., Di Matteo, T., Croft, R., Feng, Y., and Chen, N. The astrid simulation: galaxy formation and reionization. *Monthly Notices of the Royal Astronomical Society*, 512(3):3703–3716, March 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac648. URL http://dx.doi.org/10.1093/mnras/stac648.

Cuesta-Lazaro, C. and Mishra-Sharma, S. Point cloud approach to generative modeling for galaxy surveys at the field level. *Physical Review D*, 109(12):123531, 2024.

Dai, B. and Seljak, U. Multiscale flow for robust and optimal cosmological analysis. *Proceedings of the National Academy of Sciences*, 121(9):e2309624121, 2024.

Davé, R., Anglés-Alcázar, D., Narayanan, D., Li, Q., Rafieferantsoa, M. H., and Appleby, S. SIMBA: Cosmological simulations with black hole growth and feedback. , 486(2):2827–2849, June 2019. doi: 10.1093/mnras/stz937.

Diao, K., Dai, B., and Seljak, U. Detecting modeling bias with continuous time flow models on weak lensing maps. *arXiv preprint arXiv:2505.00632*, 2025.

Falcon, W. et al. Pytorchlightning/pytorch-lightning: 0.7.6 release, May 2020. URL https://doi.org/10.5281/zenodo.3828935.

Harris, C. R. et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Ho, M., Bartlett, D. J., Chartier, N., Cuesta-Lazaro, C., Ding, S., Lapel, A., Lemos, P., Lovell, C. C., Makinen, T. L., Modi, C., Pandya, V., Pandey, S., Perez, L. A., Wandelt, B., and Bryan, G. L. Ltu-ili: An all-in-one framework for implicit inference in astrophysics and cosmology. *The Open Journal of Astrophysics*, 7, July 2024. ISSN 2565-6120. doi: 10.33232/001c.120559. URL http://dx.doi.org/10.33232/001c.120559.

Horowitz, B. and Melchior, P. Plausible adversarial attacks on direct parameter inference models in astrophysics, 2022. URL https://arxiv.org/abs/2211.14788.

Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

Jeffrey, N. and Wandelt, B. D. Evidence networks: simple losses for fast, amortized, neural bayesian model comparison. *Machine Learning: Science and Technology*, 5(1):015008, 2024.

Jeffrey, N., Alsing, J., and Lanusse, F. Likelihood-free inference with neural compression of des sv weak lensing map statistics. *Monthly Notices of the Royal Astronomical Society*, 501(1):954–969, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Kluyver, T. et al. Jupyter notebooks - a publishing format for reproducible computational workflows. In *ELPUB*, 2016.

Lanzieri, D., Zeghal, J., Makinen, T. L., Boucaud, A., Starck, J.-L., and Lanusse, F. Optimal neural summarisation for full-field weak lensing cosmological implicit inference. *arXiv preprint arXiv:2407.10877*, 2024.

Lemos, P., Coogan, A., Hezaveh, Y., and Perreault-Levasseur, L. Sampling-based accuracy testing of posterior estimators for general inference. In *International Conference on Machine Learning*, pp. 19256–19273. PMLR, 2023a.

Lemos, P., Parker, L., Hahn, C., Ho, S., Eickenberg, M., Hou, J., Massara, E., Modi, C., Dizgah, A. M., Blancard, B. R.-S., and Spergel, D. Simbig: Field-level simulation-based inference of galaxy clustering, 2023b. URL https://arxiv.org/abs/2310.15256.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL https://arxiv.org/abs/1608.03983.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Maksimova, N. A., Garrison, L. H., Eisenstein, D. J., Hadzhiyska, B., Bose, S., and Satterthwaite, T. P. ¡scp¿abacussummit¡/scp¿: a massive set of high-accuracy, high-resolution n-body simulations. *Monthly Notices of the Royal Astronomical Society*, 508(3):4017–4037, September 2021. ISSN 1365-2966. doi: 10.1093/mnras/stab2484. URL http://dx.doi.org/10.1093/mnras/stab2484.

Mudur, N., Cuesta-Lazaro, C., and Finkbeiner, D. P. Diffusion-hmc: Parameter inference with diffusion-model-driven hamiltonian monte carlo. *The Astrophysical Journal*, 978(1):64, 2024.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Nguyen, T., Villaescusa-Navarro, F., Mishra-Sharma, S., Cuesta-Lazaro, C., Torrey, P., Farahi, A., Garcia, A. M., Rose, J. C., O'Neil, S., Vogelsberger, M., et al. How dreams are made: Emulating satellite galaxy and subhalo populations with diffusion models and point clouds. *arXiv preprint arXiv:2409.02980*, 2024.

Ni, Y., Di Matteo, T., Bird, S., Croft, R., Feng, Y., Chen, N., Tremmel, M., DeGraf, C., and Li, Y. The astrid simulation: the evolution of supermassive black holes. *Monthly Notices of the Royal Astronomical Society*, 513 (1):670–692, February 2022. ISSN 1365-2966. doi: 10.1093/mnras/stac351. URL http://dx.doi.org/10.1093/mnras/stac351.

Ni, Y., Genel, S., Anglés-Alcázar, D., Villaescusa-Navarro, F., Jo, Y., Bird, S., Di Matteo, T., Croft, R., Chen, N., de Santi, N. S. M., Gebhardt, M., Shao, H., Pandey, S., Hernquist, L., and Dave, R. The CAMELS Project: Expanding the Galaxy Formation Model Space with New ASTRID and 28-parameter TNG and SIMBA Suites. , 959(2):136, December 2023. doi: 10.3847/1538-4357/ad022a.

Pakmor, R., Springel, V., Coles, J. P., Guillet, T., Pfrommer, C., Bose, S., Barrera, M., Delgado, A. M., Ferlito, F., Frenk, C., Hadzhiyska, B., Hernández-Aguayo, C., Hernquist, L., Kannan, R., and White, S. D. M. The millenniumtng project: the hydrodynamical full physics simulation and a first look at its galaxy clusters. *Monthly Notices of the Royal Astronomical Society*, 524(2):2539–2555, July 2023. ISSN 1365-2966. doi: 10.1093/mnras/stac3620. URL http://dx.doi.org/10.1093/mnras/stac3620.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance.pdf.

Pillepich, A., Springel, V., Nelson, D., Genel, S., Naiman, J., Pakmor, R., Hernquist, L., Torrey, P., Vogelsberger, M., Weinberger, R., and Marinacci, F. Simulating galaxy formation with the IllustrisTNG model. , 473(3):4077–4106, January 2018. doi: 10.1093/mnras/stx2656.

Polanska, A., Price, M. A., Piras, D., Mancini, A. S., and McEwen, J. D. Learned harmonic mean estimation of the bayesian evidence with normalizing flows. *arXiv preprint arXiv:2405.05969*, 2024.

Skilling, J. Nested Sampling. In Fischer, R., Preuss, R., and Toussaint, U. V. (eds.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735 of *American Institute of Physics Conference Series*, pp. 395–405. AIP, November 2004. doi: 10.1063/1.1835238.

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., Somerville, R. S., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., Narayanan, D., Li, Y., Philcox, O., La Torre, V., Maria Delgado, A., Ho, S., Hassan, S., Burkhart, B., Wadekar, D., Battaglia, N., Contardo, G., and Bryan, G. L. The CAMELS

Project: Cosmology and Astrophysics with Machine-learning Simulations. , 915(1):71, July 2021. doi: 10.3847/1538-4357/abf7ba.

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., Spergel, D. N., S. Somerville, R., Dave, R., Pillepich, A., Hernquist, L., Nelson, D., Torrey, P., Narayanan, D., Li, Y., Philcox, O., La Torre, V., Maria Delgado, A., Ho, S., Hassan, S., Burkhart, B., Wadekar, D., Battaglia, N., Contardo, G., and Bryan, G. L. The camels project: Cosmology and astrophysics with machine-learning simulations. *The Astrophysical Journal*, 915(1):71, July 2021. ISSN 1538-4357. doi: 10.3847/1538-4357/abf7ba. URL http://dx.doi.org/10.3847/1538-4357/abf7ba.

Villaescusa-Navarro, F., Genel, S., Anglés-Alcázar, D., Thiele, L., Dave, R., Narayanan, D., Nicola, A., Li, Y., Villanueva-Domingo, P., Wandelt, B., Spergel, D. N., Somerville, R. S., Zorrilla Matilla, J. M., Mohammad, F. G., Hassan, S., Shao, H., Wadekar, D., Eickenberg, M., Wong, K. W. K., Contardo, G., Jo, Y., Moser, E., Lau, E. T., Machado Poletti Valle, L. F., Perez, L. A., Nagai, D., Battaglia, N., and Vogelsberger, M. The CAMELS Multifield Data Set: Learning the Universe's Fundamental Parameters with Artificial Intelligence. , 259(2):61, April 2022. doi: 10.3847/1538-4365/ac5ab0.

Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020. doi: https://doi.org/10.1038/s41592-019-0686-2.

Weinberger, R., Springel, V., Hernquist, L., Pillepich, A., Marinacci, F., Pakmor, R., Nelson, D., Genel, S., Vogelsberger, M., Naiman, J., and Torrey, P. Simulating galaxy formation with black hole driven thermal and kinetic feedback. , 465(3):3291–3308, March 2017. doi: 10.1093/mnras/stw2944.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey, 2024. URL https://arxiv.org/abs/2110.11334.

Zhang, L. H., Goldstein, M., and Ranganath, R. Understanding failures in out-of-distribution detection with deep generative models, 2021. URL https://arxiv.org/abs/2107.06908.

A. Model architectures

A.1. Compressor network

The compressor network takes as an input 256×256 2D fields and outputs summary statistics vector t. For the compressor network, we use a ResNet-10-T model as implemented in (Wightman, 2019)[1]. This model is a lightweight variation of the ResNet architecture (He et al., 2016), with fewer trainable parameters (4.94 million parameters), which we found to perform better on our dataset than larger models, such as ResNet-18 (11.2 million parameters) or ResNet-34 (21.3 million parameters). The model consists of four residual blocks, with one convolutional layer in each block, followed by average pooling and downsampling. Throughout the network, we use circular convolutions to account for the periodic boundary conditions of the input fields. The final output of the network is a summary statistics vector t, the dimensionality of which we fix to 40.

A.2. Posterior inference network

The posterior inference network is a Masked Autoregressive Flow (MAF), which is conditioned on the summary statistic vector t from the compressor network, and is trained to predict a vector of all 6 cosmological and astrophysical parameters. The MAF consist of 8 transforms. Each transform a masked MLP (multi-layer perceptron) composed of 3 layers with 256 hidden features in each layer and ReLU activation functions between the layers. To help facilitate the training, each parameter is normalized with respect to its range such that its value lies between [-1, 1].

A.3. Evidence estimation network

The architecture of the evidence estimation network is identical to that of the posterior inference network used in the compression step: the network is a MAF composed of 8 transforms, where each transform is a masked MLP with 3 layers of 256 hidden features and ReLU activation functions between the layers.

B. Implementation and training

We follow the same training procedure for both total matter density and gas density fields. We find that the similar sets of hyperparameters works well for the two astrophysical fields, although it is possible that other fields which track more distinct physical quantities, such as gas temperature or gas pressure, would require significantly different set of hyperparameters for training.

We split the dataset for each field into training, validation, and test sets, following a random 90/5/5 split, which reserves 13500 maps (900 distinct cosmologies) for training and 500 maps (50 cosmologies) for validation and testing. For consistency, we use the same dataset split for training both the compressor network, which is trained together with the parameter inference network, and the evidence estimation network. To aid the training, we work with the fields in the log space and standardize them prior to the training. We also apply data augmentations in the form of random rotations and mirror flips to help the models learn the relevant symmetries.

The compressor network is trained jointly with a posterior inference network to learn the optimal summary statistics. We train the two networks using the AdamW optimizer (Loshchilov & Hutter, 2019; Kingma & Ba, 2017), with peak learning rate 1×10^{-4} for $M_{\rm tot}$ (7×10^{-5} for $M_{\rm gas}$) and cosine annealing learning rate schedule (Loshchilov & Hutter, 2017) for 300 epochs with a batch size of 100. For the downstream task of evidence estimation we select the checkpoint with the lowest validation loss.

The evidence estimation network is trained with similar settings: we use the AdamW optimizer with peak learning rate of 7×10^{-3} for $M_{\rm tot}$ (5×10^{-3} for $M_{\rm gas}$) and cosine annealing scheduler to train the network for 100 epochs with a batch size of 100. Similarly to the first part of the training, we save the checkpoint with the lowest loss on the validation set.

C. Parameter constraints from $M_{\rm tot}$ and $M_{\rm gas}$ without Gaussian smoothing

In Figures 4 and 5 we plot mean values and $1-\sigma$ uncertainties for cosmological (Ω_M, σ_8) and astrophysical parameters (A_{SN1}, A_{SN2}, A_{AGN1}, A_{AGN2}) inferred from the total matter density and gas density maps from the test set of Astrid LH

[1] https://github.com/huggingface/pytorch-image-models

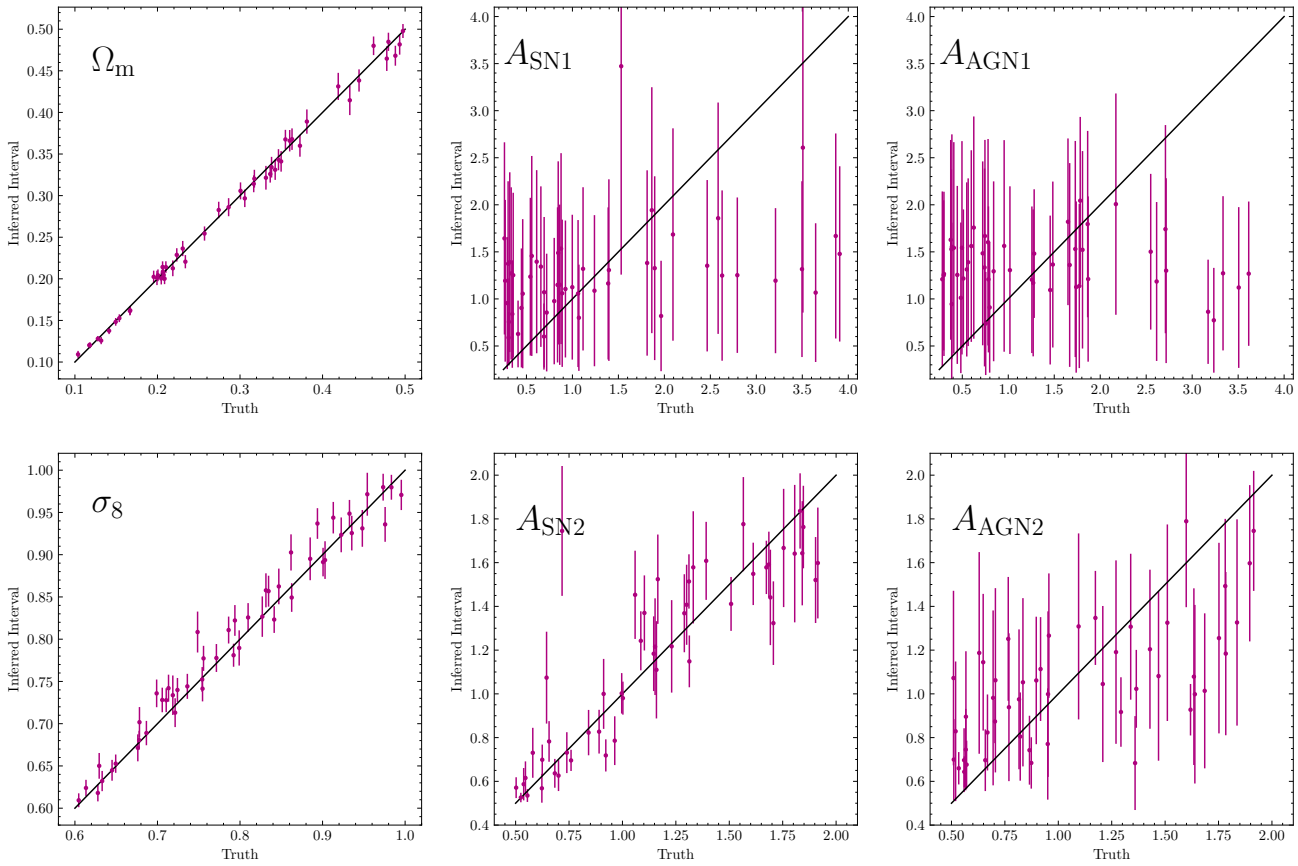

Figure 4. Predicted parameters against ground-truth parameters for total matter density fields (without Gaussian smoothing) of the LH test set of the Astrid suite. The mean values and $1-\sigma$ uncertainties are computed over 10 000 posterior samples for each input field. The parameter inference network is able to constrain cosmological parameters to a high degree of both precision and accuracy.

suite without Gaussian smoothing.

We estimate the parameters from the fields following the same approach and using the same model architectures as in App. A, but without conditioning the summary statistics on the smoothing scale k_s. We use the same training/validation/test split and training strategy as in Section B, with the peak learning rate of $7e-5$ for both fields.

We find that matter density fields can primarily constrain the cosmological parameters, but have considerably less predictive power for the astrophysical parameters, while the gas density field can be used to also place tight constraints on the supernovae feedback parameter A_{SN2}. Therefore, when validating our compressor and posterior inference networks in Appendix D, we estimate coverage for cosmological parameters only for $M_{\rm tot}$ fields and for cosmological parameters and A_{SN2} parameter for $M_{\rm gas}$ fields.

D. Posterior calibration test

We examine whether the posterior inference networks for $M_{\rm tot}$ and $M_{\rm gas}$ fields are well-calibrated using the TARP method (Lemos et al., 2023a).

Given a distance metric $d : U \times U \to R$ and a reference point θ_r, a TARP credible region for a parameter-data pair $(\theta^\star, x^\star)$ and a posterior estimator $q(\theta|x)$ is defined as follows:

$$\{\theta \in U \,|\, d(\theta, \theta_r) \leq d(\theta^\star, \theta_r)\}, \tag{1}$$

which in turn defines a TARP confidence level $1-\alpha_{\rm TARP}$ as the integral of the estimated posterior $q(\theta|x)$ over that credible

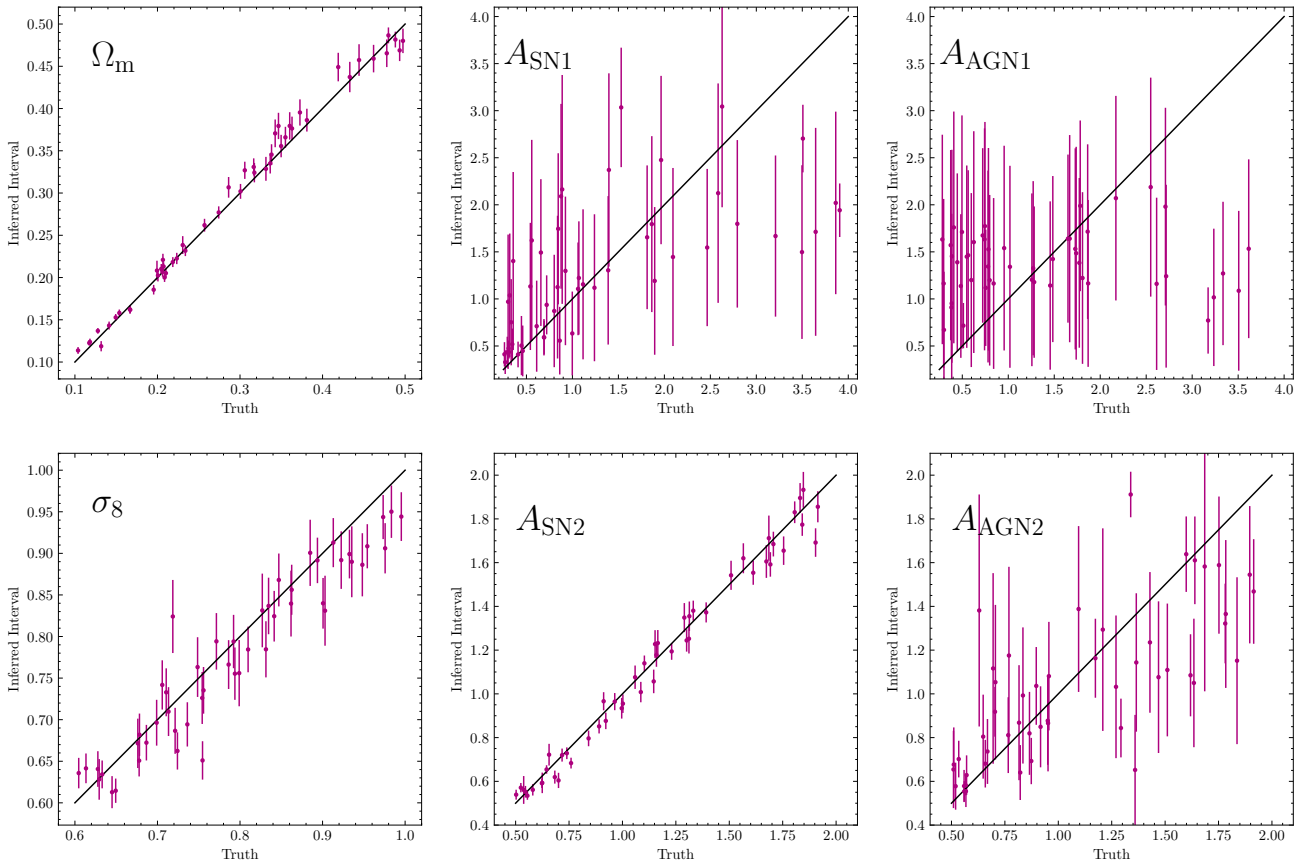

Figure 5. Predicted parameters against ground-truth parameters for gas density fields (without Gaussian smoothing) of the LH set of the Astrid suite. The mean values and $1-\sigma$ uncertainties are computed over 10 000 samples for each input field. In addition to the cosmological parameters, gas density fields from Astrid can also place tight constraints on the supernovae feedback parameter A_{SN2}.

region. We use the Euclidian distance as our metric. The expected coverage plotted on Figure 6 is computed as an average coverage across multiple parameter-data pairs. For well-calibrated posteriors, the expected coverage should match the corresponding the credible level for all credible levels $(1-\alpha_{\rm TARP}) \in [0,1]$. The details of algorithm for computing the TARP coverage are presented in full in (Lemos et al., 2023a). We use `tarp` package with Euclidean distance metric to evaluate the expected coverage for our posterior inference networks.

We compute the expected coverage for a wide range of smoothing scales which are logarithmically spaced from $k_{s,\rm min} = 2\,h/{\rm Mpc}$ to $k_{s,\rm max} = 45\,h/{\rm Mpc}$ and find the posteriors to be fairly well-calibrated for all of the scales within this range. As noted in the previous section, we compute the coverage only for the parameters for which the density fields have constraining power for.

E. Log-evidence distribution $\log p(x_s)$

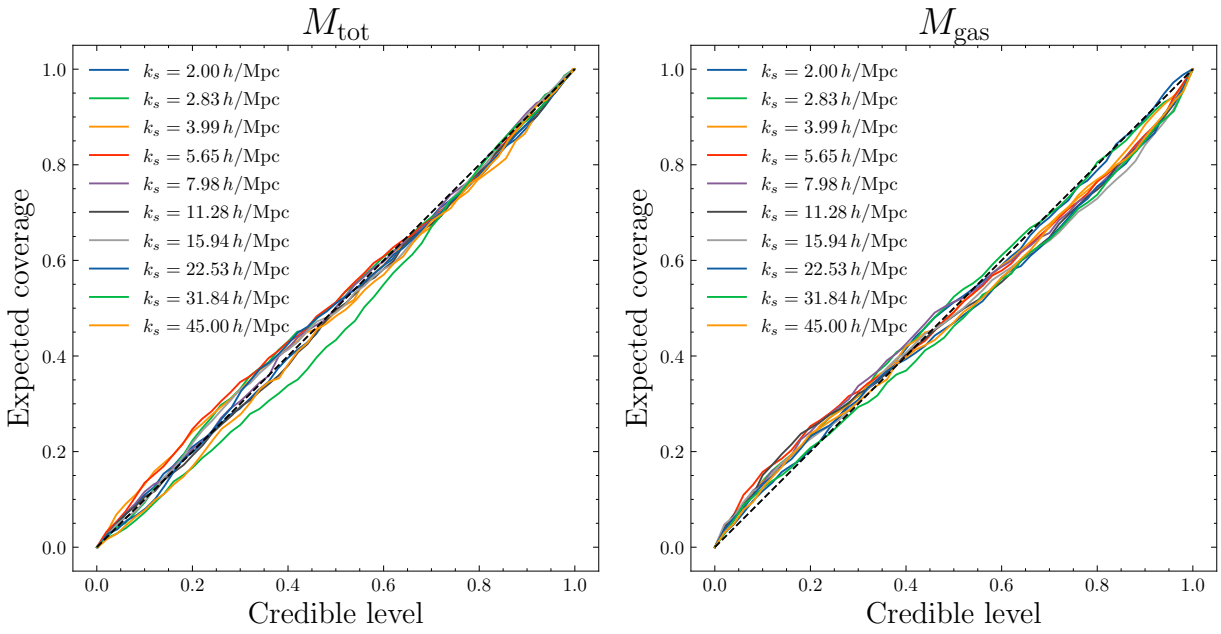

Figure 6. Expected coverage versus confidence levels computed for a range of smoothing scales k_s. Left: for total matter density fields, the coverage is computed for cosmological parameters only. Right: for gas density fields the coverage is computed for cosmological and supernovae feedback parameter A_{SN2} which the gas density fields are also sensitive to. The coverage is computed on the fields from the LH test set of the Astrid suite with the TARP method (Lemos et al., 2023a).

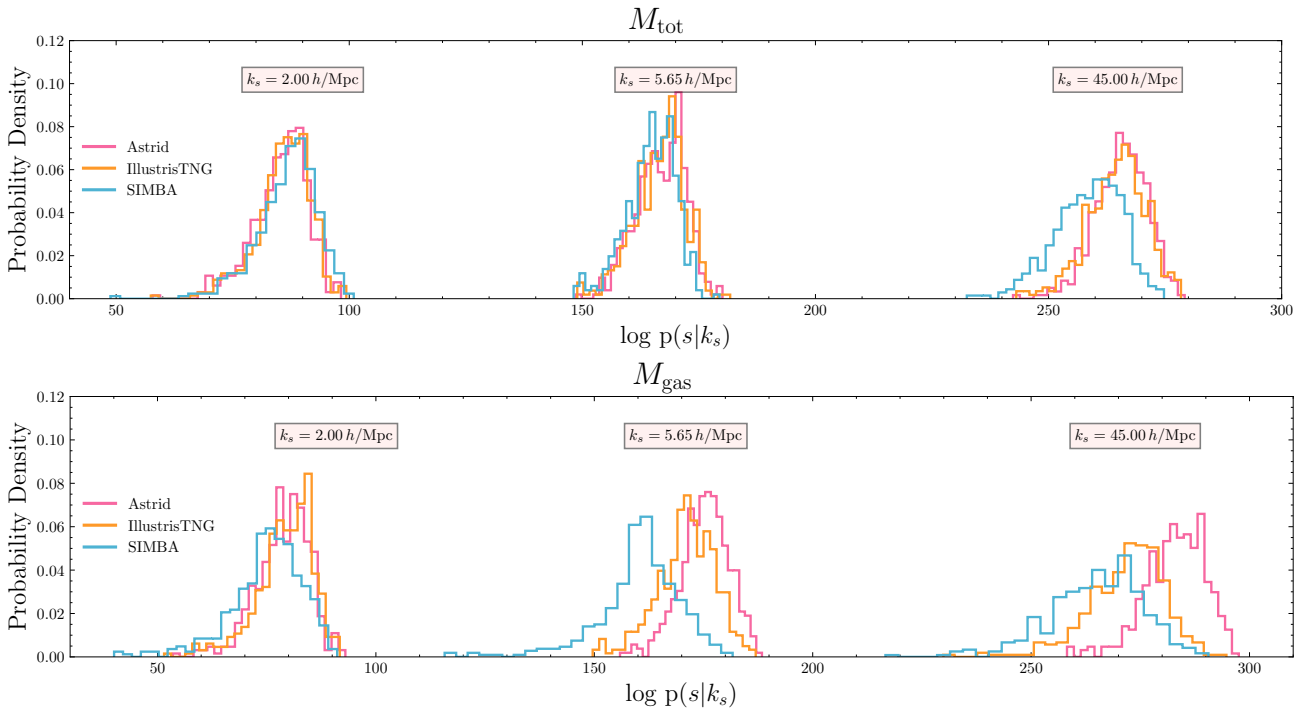

Figure 7. Comparison of the distribution of $\log p(t_s)$ values (the log-evidence of the learned neural summary statistics) for three different smoothing scales (2, 5.65, and 45 $h/{\rm Mpc}$) for the CV sets of Astrid, IllustrisTNG, and SIMBA suites. The top plot corresponds to the total matter density field, the bottom plot corresponds the gas density field. At large smoothing scales ($k_s = 2h/{\rm Mpc}$), the distributions for the three different suites largely overlap. As we include more information from smaller scales ($k_s \downarrow$), the differences between the distributions become more pronounced, in particular for the gas density fields.