Addressing Misspecified Physical Models: Correcting Underspecified Lens Convergence Models via Data-Driven Updates

Nicolas Payot [1 2 3] **Gabriel Missael Barco** [1 2 3] **Laurence Perreault-Levasseur** [1 2 3 4 5 6] **Yashar Hezaveh** [1 2 3 4 5 6]

Abstract

Reconstructing a galaxy's mass distribution from a single observed lensed image is a challenging nonlinear inverse problem with many degeneracies, typically requiring strong prior assumptions. We propose a deep generative approach that learns a flexible prior over lens mass maps using a Wasserstein Autoencoder (WAE) trained on simulated data. To correct for misspecification of the initial prior, which is trained on a simpler surrogate model, we iteratively refine the prior by generating posterior samples, obtained from new simulated lenses, under the current model and retraining the WAE on these samples. In experiments on synthetic lensed images, this iterative scheme increases the model's data likelihood and yields more accurate recovery of lens parameters such as ellipticity, despite starting from a biased prior lacking this feature. These results demonstrate that data-driven prior adaptation can mitigate model misspecification in nonlinear lensing inversion and potentially improve inference in other complex inverse problems.

1. Introduction

Strong gravitational lensing is a prominent example of an inverse problem in astrophysics. Given a distorted image of a background source, one needs to infer both the foreground mass distribution, *i.e.* the lens that produced the observed deflection, as well as the background source. This problem is very challenging since lensing involves a highly nonlinear mapping for the mass distribution (κ-map) and permits many possible mass profiles and background source configurations that fit the data. To mitigate this degeneracy, lens modelers have long relied on either simple parametric forms or free-form reconstructions. While the parametric models are computationally convenient, they lack the flexibility needed to capture complex or irregular mass distributions. On the other hand, free-form models need to be strongly regularized, a process which is usually done by enforcing manually specified priors (Birrer et al., 2015; Merten, 2016; Galan et al., 2022). These assumptions are usually relatively simple, but hard to encode. Alternatively, flexible, learned priors offer a compelling alternative: by leveraging data or simulations to learn a distribution over realistic mass maps, we can regularize the lens inversion with a richer model of plausible structures (Morningstar et al., 2019; Adam et al., 2023; Wagner-Carena et al., 2023).

In particular, deep generative models like Variational Autoencoders (VAEs) or their extensions like the Flow-VAE or Wasserstein Autoencoders (WAEs) can learn complex distributions of parameters from simulated data and serve as high-dimensional priors (Kingma & Welling, 2014; Rubenstein et al., 2018; Lanusse et al., 2021). VAEs learn a latent representation of the data by maximizing an evidence lower bound, providing an efficient way to encode and sample high-dimensional distributions. WAEs are a more recent variant that minimizes the Wasserstein distance between the model distribution and the data distribution for its optimal transport objective, often yielding improved sample quality and a closer match to the true distribution (Rubenstein et al., 2018). These learned priors can improve lens modeling as they embed learned regularities of mass profiles rather than imposing simplistic forms.

However, a critical difficulty arises when the generative model is trained on an imperfect or mismatched dataset (Huang et al., 2023; Wehenkel et al., 2025). In practice, one does not have access to true κ-maps drawn from the actual distribution of galaxy mass distribution and instead, must rely on imperfect simulations or simplified models. This can lead to model misspecification where the prior encoded by the WAE does not exactly match reality. If not taken into account, such a mismatch could bias posteriors and lead to incorrect inferences (Grünwald, 2012; Grünwald & Van Ommen, 2017; Miller & Dunson, 2019). Recent

[1]Ciela Institute, Montréal, Canada [2]Mila—Quebec Artificial Intelligence Institute, Montréal, Canada [3]Department of Physics, Université de Montréal, Montréal, Canada [4]Center for Computational Astrophysics, Flatiron Institute, New York, USA [5]Perimeter Institute for Theoretical Physics, Waterloo, Canada [6]6 Trottier Space Institute, McGill University, Montréal, Canada. Correspondence to: Nicolas Payot <nicolas.payot@umontreal.ca>.

ML4Astro 2025, Vancouver, CA.

work on posterior-corrected priors has proposed using the observed data itself to iteratively refine a misspecified prior (Bissiri et al., 2016; Barco et al., 2025; Rozet et al., 2024). Their approaches showed that, even when starting from a biased prior, repeatedly retraining the generative model on samples drawn from the posterior obtained with that prior, or by using specialized loss functions, allows the prior to asymptotically converge toward a data-explaining-distribution that attains the same evidence as the ground-truth prior.

Our key contribution is demonstrating that this approach can mitigate prior misspecification even when the lensing forward model is nonlinear and the generative model must learn a distribution over mass maps while being much less flexible than score-based models. In the following, we outline the lensing forward model and our inference method. We then detail the iterative WAE retraining procedure that progressively adapts the κ-map prior towards a distribution explaining the data despite the initial mismatch.

2. Methodology

2.1. Lensing Forward Model

We consider the standard forward model of strong gravitational lensing, in which a background source with intensity distribution s is lensed by a foreground mass distribution characterized by its convergence field κ. When κ, s and η are pixelated, we can state our problem as in Warren & Dye (2003):

$$y = A_\kappa s + \eta \tag{1}$$

where $y \in \mathbb{R}^m$ is an observation, $A_\kappa \in \mathbb{R}^{m \times n}$ is the Jacobian of the forward model describing the distortion to source $s \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ is a noise realization. Although s and κ are usually inferred jointly, we instead fix s to a known 2-D Gaussian centered in the image. This turns the task into recovering only the convergence field κ, which remains a nonlinear problem.

In this work, all lenses are simulated with the package `Caustics` (Stone et al., 2024). We simulate synthetic observations y by drawing convergence maps $\kappa \sim p_*(\kappa)$ from a 6-parameter Elliptical Power-Law (EPL) (Barkana, 1998) mass profile and inject Gaussian noise $\eta \sim \mathcal{N}(0, \sigma_\eta I)$, $\sigma_\eta = 1$.

2.2. WAE Prior

We employ an over-parametrized Wasserstein Autoencoder (WAE) to learn a flexible prior over the κ-maps. The WAE consists of an encoder $E_\phi(\kappa)$ that maps a high-dimensional κ-map to a latent sample, $z \in \mathbb{R}^{16}$, and a decoder $G_\theta(z)$ that generates a reconstructed $\hat{\kappa}$ (Rubenstein et al., 2018). Training is done on a set of misspecified mass distributions, here, drawn from a simpler surrogate model. The initial training set is composed of κ-maps simulated from the Singular Isothermal Sphere (SIS) parametric model, which lacks ellipticity: $\kappa \sim p_{\text{SIS}}(\kappa)$. The parameters for both the true and initial convergence map distributions are given in Annex A.

The WAE training objective includes a reconstruction term, ensuring $G_\theta(z)$ produces accurate mass maps, and a divergence term that encourages the model's latent-space output to match a prior distribution (Rubenstein et al., 2018). By minimizing the Wasserstein distance between the model output distribution and the empirical data distribution, the WAE learns to generate κ samples that mimic the training set statistics. After training, the decoder $G_\theta(z)$ defines an implicit prior $p_{\text{WAE}}(\kappa) = p_1(\kappa)$. We can draw random latent vectors $z \sim \mathcal{N}(0, I)$ and map them through G_θ to obtain random realistic mass maps. See Annex B for more information on the WAE.

2.3. Posterior Inference

This WAE prior is then used in our inference of κ. Specifically, we treat $p_{\text{WAE}}(\kappa)$ as the prior in a posterior $p(\kappa \mid y, s) \propto p(y \mid \kappa, s) p_{\text{WAE}}(\kappa)$. Directly sampling the posterior is non-trivial because the posterior in the high-dimensional latent space might be complex and multi-modal due to degeneracies. Standard MCMC methods could be applied but might struggle with the higher dimensionality of z and complex likelihood surface (Yao et al., 2022). Instead, we adopt a latent optimization approach. The idea is to find latent vectors z whose decoded mass maps produce simulated lenses that closely match the observation. Specifically, we define a loss function for a given latent z:

$$\mathcal{L}(z) = \frac{1}{2\sigma_\eta^2} \| y_{\text{obs}} - A_{G_\theta(z)} s \|_2^2 + \frac{\sigma_\eta}{2} \| z \|_2^2 \tag{2}$$

where $A_{G_\theta(z)}$ is the forward model for the decoded $\hat{\kappa} = G_\theta(z)$. Minimizing $\mathcal{L}(z)$ seeks a maximum a posteriori (MAP) estimate of z, and corresponding κ, that best explains the observation. We use the gradient-based optimizer Adam to minimize $\mathcal{L}(z)$, leveraging the fact that our forward model and decoder are differentiable. Importantly, to prevent the optimizer from converging to a single mode, we employ a multi-start strategy: we draw $N = 20$ random initial latent vectors $z_i^{(1)}{}_{i=1\ldots N}$ from $p_1(z)$, run N separate Adam optimizers and keep the one with the lowest loss. Then, starting from the MAP estimate, we obtain posterior samples with stochastic gradient Langevin dynamics (SGLD) and a decreasing step size (Welling & Teh, 2011). It should be noted that using a MAP estimate for the latent z in place of a posterior sample would correspond to a hard-EM approach (Laarhoven & Marchiori, 2018). Hard-EM effectively maximizes the complete-data posterior $p(\kappa, z \mid y, s)$ rather than the marginal posterior $p(\kappa \mid y, s)$, hence it does not con-

serve EM's usual guarantees such as the non-decreasing of the marginal likelihood (Gupta & Chen, 2011).

2.4. Iterative Prior Adaptation

The core of our method is an iterative scheme that progressively adapts the WAE prior toward the true distribution of κ-maps despite starting from a misspecified SIS-trained model. The procedure is as follows:

1. **Initial Prior & Inference:** We train the WAE on the surrogate SIS dataset to obtain an initial prior $p_1(\kappa)$. Then, for each observed image y_j, simulated with κ-maps drawn from the true prior, $p_*(\kappa)$, we perform posterior sampling with the current prior $p_1(\kappa)$. This yields a collection of samples $\kappa_{i,j}^{(1)}$ for each observation j. Due to prior misspecification, these samples may not perfectly match the true EPL distribution, but they provide corrective updates toward fitting the data. In this work, we sampled $j = 500$ different observations and obtained the posterior samples for each ($i = 1$), a relatively modest dataset that should suffice here since we are fine-tuning an existing model rather than training a new one from scratch.

2. **Retraining the Prior:** With the new posterior-drawn dataset, the WAE is fine-tuned to learn a new prior $p_2(\kappa)$ that better reflects these posterior samples. Intuitively, this adjusts the prior towards mass configurations that were actually needed to explain the observations, thereby shifting it closer to a distribution better explaining those observations. Importantly, this retraining does not require any knowledge of the true EPL form. It uses only the results of inference on real/simulated data.

3. **Iterate:** Using the updated prior $p_2(\kappa)$, we repeat the posterior inference on fresh observations, though ideally using additional lenses prevents overfitting to specific systems. This E-step / M-step loop, in analogy to EM algorithms, is iterated for a few cycles to obtain $p_n(\kappa)$, $n = 20$.

This iterative adaptation approach is inspired directly by Barco et al. (2025), who applied a similar cycle to source reconstruction with diffusion models, and similarly, by Rozet et al. (2024), who applied it to natural images in an inpainting task. Crucially, while both prior studies used expressive diffusion models, we use a WAE instead, a choice that may limit the method since sampling could wander into the WAE's untrained regions or be constrained by its narrower support.

3. Results

3.1. Likelihood Improvement

When applying the method, the likelihood improves significantly in the first few iterations as the WAE prior adapts to better fit the data. By about the 6$^{\text{th}}$ iteration, however, the gains plateau, suggesting the prior has converged to the most descriptive distribution achievable within the WAE's architectural limits. Beyond iteration 6, further retraining yields marginal improvements at best. This diminishing return is evident in Figure 1, where the log-likelihood curve flattens out.

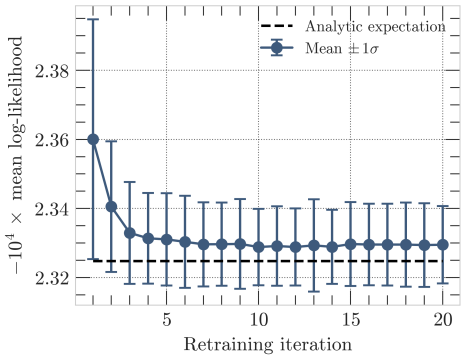

Figure 1. Mean log-likelihood of the training samples for each retraining iteration. Iteration 1 corresponds to the prior encoded by the WAE after its training on samples drawn from the SIS prior.

We assess prior fidelity with the PQMass metric (Lemos et al., 2025), which quantifies the statistical consistency between two sets of samples by estimating the probability that they originate from a common underlying distribution. Per iteration, we draw 1000 posterior samples from the WAE and 1000 from the true EPL prior, partition the image into 100 random Voronoi cells, and compute a χ^2 between the sample counts in matching cells. Averaging this over 1000 such tessellations gives the final statistic, where lower values signal a closer match between learned and target distributions. Figure 2 shows the resulting χ^2_{PQM} curves. The statistic for unprocessed samples slowly increases with retraining, indicating that the WAE-generated samples do not significantly approach the true EPL distribution in fine detail and, in fact, diverge from it. In contrast, when both the WAE and EPL κ-maps are smoothed with a Gaussian kernel ($\sigma = 2$, 7×7), χ^2_{PQM} steadily declines with each iteration, until reaching a plateau. This outcome implies that while the WAE prior is unable to exactly replicate the true distribution, likely due to architectural limitations, sampling limitations, and the contained information in the data, it does converge in terms of large-scale structure. Crucially, smoothed EPL κ-maps yield similar observations to the raw EPL, their residual differences falling below the noise level, indicating that learning the high-frequency information in

the true distribution would be very difficult.

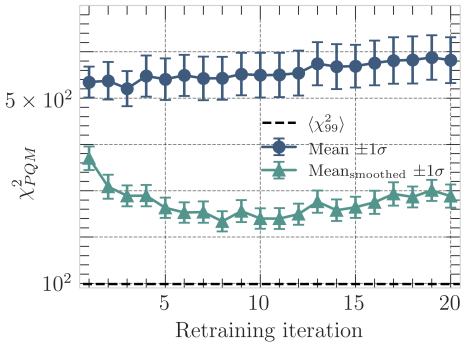

Figure 2. PQMass χ^2 statistics comparing 1000 WAE-generated posterior samples to 1000 samples from the true EPL prior at each retraining iteration. The blue (circles) curve shows χ^2_{PQM} for unprocessed κ-maps. The teal (triangles) curve shows χ^2_{PQM} after applying a Gaussian blur ($\sigma = 2$, 7×7 kernel) to both sets of maps. Applying a Gaussian kernel to both dataset ensures only the features of the κ-maps bigger than the filter are compared. This removes effects of small artifacts originating from the generative model.

3.2. Recovery of Lens Parameters

To evaluate the physical fidelity of the learned models, we fit each of the 500 posterior sample's mass maps with a corresponding parametric EPL model by gradient descent (lens center (x_0, y_0), axis ratio q, orientation ϕ, Einstein radius R_{Ein}, and mass profile slope τ).

We then compared these recovered parameters to the ground-truth values used to generate the simulated data. Figure 3 presents density plots of the error when comparing the recovered to the true parameter values for multiple iterations. We observe that lens positions and shapes are more accurately recovered after a few iterations of our method: the offsets (x_0, y_0) cluster tightly around zero, and the inferred ellipticity and orientation (q and ϕ) for each lens is very close to the true value. This indicates that after sufficient iterations, the generative model has learned to represent elliptical mass distributions, despite starting from a non-elliptical SIS prior. On the other hand, we find a systematic bias in the recovered R_{Ein}: the model tends to under-predict the Einstein radius. In Figure 3, the R_{Ein} error is mostly negative, implying the reconstructed lenses are slightly less massive than the true lenses. This bias is slightly reduced across iterations, but is mostly consistent. Most strikingly, the slope parameter τ is unconstrained. The inferred τ values show little to no correlation with the true τ. Generally, the error for all parameters at iteration 10 is slightly smaller than at iteration 20, which is consistent with the PQMass analysis, where the χ^2_{PQM} started to diverge from the smooth EPL distribution after iteration 10, likely due to overfitting on the limited dataset.

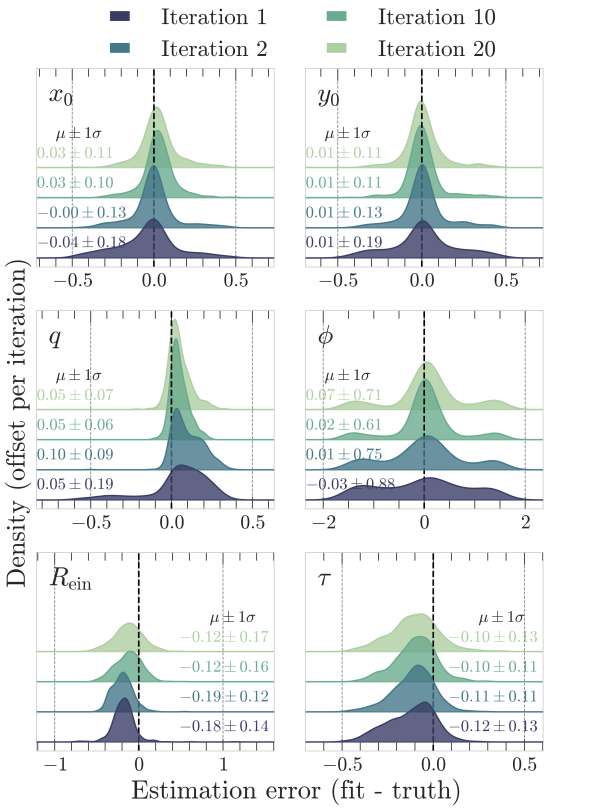

Figure 3. Error of the recovered parameters of 500 posterior samples compared to their true value for six parameters of the EPL mass model. For the lens-center offsets (x_0, y_0) and the Einstein radius (R_{ein}) panels, the axes are in arcseconds ($''$). Across retraining iterations, the error shrinks, indicated by a clustering around 0 (the dotted line). For a more detailed plot, see Annex C.

4. Future Work and Limitations

Future work will employ more sophisticated posterior samples such as latent normalizing flows, flow-conditioned Langevin dynamics, or noise-space diffusion methods, which already reach MCMC-level accuracy with much shorter run times (Holzschuh & Thuerey, 2024; Venkatraman et al., 2025). A more efficient sampler will let us leverage much larger and more diverse lens datasets, which, in turn, will let us train deeper and more flexible generative models.

Extending the iterative scheme to real observations will require adding instrumental effects (PSF convolution, non-Gaussian noise, imperfect calibration, and residual lens light) into the model refinement loop, so that their uncertainties propagate through the posterior. Beyond this, we plan to generalize the framework to jointly reconstruct the unknown source light and lens mass. Complementary generative priors for both components should temper the notorious lens-source degeneracy, although working with this degeneracy may still prove very challenging.

Since the prior shifts away from the ground truth distribution when using the prior-refinement method while still explaining the data better, there exist several prior distributions that explain the data with equal likelihood. Therefore, the adoption of a maximum-entropy selection criterion (e.g. Vetter et al., 2024) could single out the least-informative distribution consistent with the data.

The present work nevertheless inherits important limitations. The reliance on MAP estimates can hide posterior volume and understate uncertainty, while SGLD may mix too slowly. Moreover, because the WAE was initialized on a narrow SIS training set, its support excludes physically plausible maps lying outside of its manifold. Architectural bias can, therefore, limit fidelity after many refinement cycles. Addressing these issues, through provably consistent samplers and richer priors, defines the next milestones toward truly data-driven, unbiased lens modeling.

5. Acknowledgments

This work is partially supported by Schmidt Futures, a philanthropic initiative founded by Eric and Wendy Schmidt as part of the Virtual Institute for Astrophysics (VIA). The work is in part supported by computational resources provided by Calcul Québec and the Digital Research Alliance of Canada. Y.H. and L.P. acknowledge support from the Canada Research Chairs Program, the National Sciences and Engineering Council of Canada through grants RGPIN-2020-05073 and 05102, and the Fonds de recherche du Québec through grant CFQCU-2024-348060. N.P. acknowledges support from the Natural Sciences and Engineering Research Council of Canada (NSERC) through Canada Graduate Scholarships - Master's (CGS-M) Program.

References

Adam, A., Perreault-Levasseur, L., Hezaveh, Y., and Welling, M. Pixelated Reconstruction of Foreground Density and Background Surface Brightness in Gravitational Lensing Systems Using Recurrent Inference Machines. *The Astrophysical Journal*, 951(1):6, July 2023. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/accf84. URL https://iopscience.iop.org/article/10.3847/1538-4357/accf84.

Barco, G. M., Adam, A., Stone, C., Hezaveh, Y., and Perreault-Levasseur, L. Tackling the Problem of Distributional Shifts: Correcting Misspecified, High-dimensional Data-driven Priors for Inverse Problems. *The Astrophysical Journal*, 980(1):108, February 2025. ISSN 0004-637X, 1538-4357. doi: 10.3847/1538-4357/ad9b92. URL https://iopscience.iop.org/article/10.3847/1538-4357/ad9b92.

Barkana, R. Fast calculation of a family of elliptical gravitational lens models. *The Astrophysical Journal*, 502(2):531, aug 1998. doi: 10.1086/305950. URL https://dx.doi.org/10.1086/305950.

Birrer, S., Amara, A., and Refregier, A. GRAVITATIONAL LENS MODELING WITH BASIS SETS. *The Astrophysical Journal*, 813(2):102, November 2015. ISSN 1538-4357. doi: 10.1088/0004-637X/813/2/102. URL https://iopscience.iop.org/article/10.1088/0004-637X/813/2/102.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. A General Framework for Updating Belief Distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, November 2016. ISSN 1369-7412, 1467-9868. doi: 10.1111/rssb.12158. URL https://academic.oup.com/jrsssb/article/78/5/1103/7040623.

Galan, A., Vernardos, G., Peel, A., Courbin, F., and Starck, J.-L. Using wavelets to capture deviations from smoothness in galaxy-scale strong lenses. *Astronomy & Astrophysics*, 668:A155, December 2022. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/202244464. URL https://www.aanda.org/10.1051/0004-6361/202244464.

Grünwald, P. The Safe Bayesian: Learning the Learning Rate via the Mixability Gap. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Bshouty, N. H., Stoltz, G., Vayatis, N., and Zeugmann, T. (eds.), *Algorithmic Learning Theory*, volume 7568, pp. 169–183. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-34105-2 978-3-642-34106-9. doi: 10.1007/978-3-642-34106-9_16. URL http://link.springer.com/10.1007/978-3-642-34106-9_16. Series Title: Lecture Notes in Computer Science.

Grünwald, P. and Van Ommen, T. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4), December 2017. ISSN 1936-0975. doi: 10.1214/17-BA1085. URL https://projecteuclid.org/journals/bayesian-analysis/volume-12/issue-4/Inconsistency-of-Bayesian-Inference-for-Misspecified-Linear-Models-and-a/10.1214/17-BA1085.full.

Gupta, M. R. and Chen, Y. Theory and use of the em algorithm. *Foundations and Trends® in Signal Processing*, 4(3):223–296, 2011. ISSN 1932-8346. doi: 10.1561/2000000034. URL http://dx.doi.org/10.1561/2000000034.

Holzschuh, B. and Thuerey, N. Flow Matching for Posterior Inference with Simulator Feedback, October 2024. URL http://arxiv.org/abs/2410.22573. arXiv:2410.22573 [cs].

Huang, D., Bharti, A., Souza, A., Acerbi, L., and Kaski, S. Learning Robust Statistics for Simulation-based Inference under Model Misspecification. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=STrXsSIEiq.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014. URL https://openreview.net/forum?id=33X9fd2-9FyZd.

Laarhoven, T. v. and Marchiori, E. Domain Adaptation with Randomized Expectation Maximization, March 2018. URL http://arxiv.org/abs/1803.07634. arXiv:1803.07634 [stat].

Lanusse, F., Mandelbaum, R., Ravanbakhsh, S., Li, C.-L., Freeman, P., and Póczos, B. Deep generative models for galaxy image simulations. *Monthly Notices of the Royal Astronomical Society*, 504(4):5543–5555, May 2021. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stab1214. URL https://academic.oup.com/mnras/article/504/4/5543/6263655.

Lemos, P., Sharief, S., Malkin, N., Salhi, S., Stone, C., Perreault-Levasseur, L., and Hezaveh, Y. PQMass: Probabilistic Assessment of the Quality of Generative Models using Probability Mass Estimation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=n7qGCmluZr.

Merten, J. Mesh-free free-form lensing – I. Methodology and application to mass reconstruction. *Monthly Notices of the Royal Astronomical Society*, 461(3):2328–2345, September 2016. ISSN 0035-8711, 1365-2966. doi: 10.1093/mnras/stw1413. URL https://academic.oup.com/mnras/article-lookup/doi/10.1093/mnras/stw1413.

Miller, J. W. and Dunson, D. B. Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, July 2019. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2018.1469995. URL https://www.tandfonline.com/doi/full/10.1080/01621459.2018.1469995.

Morningstar, W. R., Perreault Levasseur, L., Hezaveh, Y. D., Blandford, R., Marshall, P., Putzky, P., Rueter, T. D., Wechsler, R., and Welling, M. Data-driven Reconstruction of Gravitationally Lensed Galaxies Using Recurrent Inference Machines. *The Astrophysical Journal*, 883:14, September 2019. ISSN 0004-637X. doi: 10.3847/1538-4357/ab35d7. URL https://ui.adsabs.harvard.edu/abs/2019ApJ...883...14M. Publisher: IOP ADS Bibcode: 2019ApJ...883...14M.

Nakagawa, N., Togo, R., Ogawa, T., and Haseyama, M. Gromov-Wasserstein Autoencoders. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sbS10BCtc7.

Rozet, F., Andry, G., Lanusse, F., and Louppe, G. Learning diffusion priors from observations by expectation maximization. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 87647–87682. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9f94298bac4668db4dc77ddb0a244301-Paper-Conference.pdf.

Rubenstein, K. P., Sch{\"o}lkopf, B., and Tolstikhin, I. Wasserstein Auto-Encoders: Latent Dimensionality and Random Encoders. In *Workshop at the 6th International Conference on Learning Representations (ICLR)*, May 2018. URL https://openreview.net/pdf?id=HkL7n1-0b.

Stone, C., Adam, A., Coogan, A., Yantovski-Barth, M. J., Filipp, A., Setiawan, L., Core, C., Legin, R., Wilson, C., Barco, G. M., Hezaveh, Y., and Perreault-Levasseur, L. Caustics: A Python Package for Accelerated StrongGravitational Lensing Simulations. *Journal of Open Source Software*, 9(103):7081, November 2024. ISSN 2475-9066. doi: 10.21105/joss.07081. URL https://joss.theoj.org/papers/10.21105/joss.07081.

Tessore, N. and Benton Metcalf, R. The elliptical power law profile lens. *Astronomy & Astrophysics*, 580:A79, August 2015. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201526773. URL http://www.aanda.org/10.1051/0004-6361/201526773.

Venkatraman, S., Hasan, M., Kim, M., Scimeca, L., Sendera, M., Bengio, Y., Berseth, G., and Malkin, N. Outsourced diffusion sampling: Efficient posterior inference in latent spaces of generative models. *CoRR*, abs/2502.06999, February 2025. URL https://doi.org/10.48550/arXiv.2502.06999.

Vetter, J., Moss, G., Schröder, C., Gao, R., and Macke, J. H. Sourcerer: Sample-based maximum entropy source distribution estimation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0cgDDa4OFr.

Wagner-Carena, S., Aalbers, J., Birrer, S., Nadler, E. O., Darragh-Ford, E., Marshall, P. J., and Wechsler, R. H. From Images to Dark Matter: End-to-end Inference of Substructure from Hundreds of Strong Gravitational Lenses. *The Astrophysical Journal*, 942:75, January 2023. ISSN 0004-637X. doi: 10.3847/1538-4357/aca525. URL https://ui.adsabs.harvard.edu/abs/2023ApJ...942...75W. Publisher: IOP ADS Bibcode: 2023ApJ...942...75W.

Warren, S. J. and Dye, S. Semilinear Gravitational Lens Inversion. *The Astrophysical Journal*, 590(2):673–682, June 2003. ISSN 0004-637X, 1538-4357. doi: 10.1086/375132. URL https://iopscience.iop.org/article/10.1086/375132.

Wehenkel, A., Gamella, J. L., Sener, O., Behrmann, J., Sapiro, G., Jacobsen, J.-H., and Cuturi, M. Addressing Misspecification in Simulation-based Inference through Data-driven Calibration, May 2025. URL http://arxiv.org/abs/2405.08719. arXiv:2405.08719 [stat].

Welling, M. and Teh, Y. W. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 681–688, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Yao, Y., Vehtari, A., and Gelman, A. Stacking for non-mixing Bayesian computations: the curse and blessing of multimodal posteriors. *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435. Publisher: JMLR.org.

A. Gravitational Lensing Formulation

The lens mass map produces deflection field $\vec{\alpha}$ via the lensing potential, ψ, itself related to the convergence field by:

$$\kappa = \frac{1}{2}\nabla^2\psi\,. \tag{3}$$

This deflection field maps the source-plane coordinates $\vec{\beta}$ to image-plane coordinates $\vec{\theta}$ through the lens equation:

$$\vec{\beta}(\vec{\theta}) = \vec{\theta} - \vec{\alpha}(\vec{\theta}) = \vec{\theta} - \vec{\nabla}\psi(\vec{\theta}) \tag{4}$$

The lens image $I(\vec{\theta})$ is given by the source light mapped through this distortion: $I(\vec{\theta}) = s(\vec{\beta}(\vec{\theta}))$ when assuming a thin lens. Observational noise is then added to I to obtain y_{obs}. In this work, two different lenses are used. The Singular Isothermal Sphere's (SIS) convergence field is given by:

$$\kappa_{\text{SIS}}(x, y) = \frac{R_{\text{ein}}}{\sqrt{(x - x_0)^2 + (y - y_0)^2}} \tag{5}$$

Similarly, the convergence field for the Elliptical Power-Law (EPL), as in Tessore & Benton Metcalf (2015), is given by:

$$\kappa_{\text{EPL}}(x, y) = \frac{2 - \tau}{2}\left(\frac{R_{\text{ein}}}{q^2(x - x_0)^2 + (y - y_0)^2}\right)^{\tau} \tag{6}$$

with

$$\phi = \arctan(qx, y)\,. \tag{7}$$

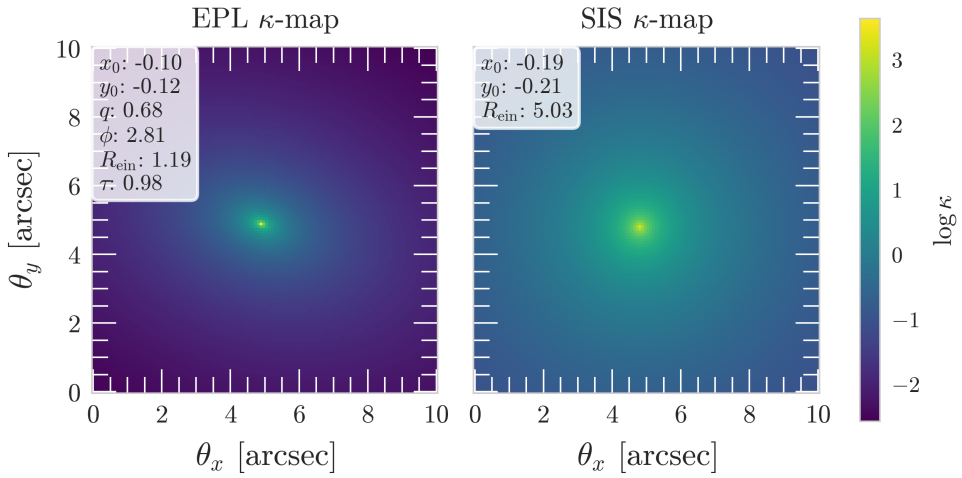

Figure 4. Elliptical Power-Law (EPL) and Singular Isothermal Sphere (SIS) mass profiles from the prior defined in table 1.

Table 1. Priors over the parameters of the Elliptical Power-Law (EPL) mass profile used for the observations and the Singular Isothermal Sphere (SIS) used as the misspecified initial prior.

	EPL	SIS
x_0 ($''$)	$U[-0.25, 0.25]$	$U[-0.5, 0.5]$
y_0 ($''$)	$U[-0.25, 0.25]$	$U[-0.5, 0.5]$
q	$U[0.4, 1.0]$	-
ϕ	$U[0, \pi]$	-
R_{ein} ($''$)	$U[0.5, 2.5]$	$U[0.5, 10]$
τ	$U[0.75, 1.25]$	-

B. WAE Training

We trained a Wasserstein Autoencoder (WAE) on 2×10^5 simulated 128×128 pixels κ-maps drawn from a spherical Singular Isothermal Sphere (SIS) prior. This WAE learned a latent representation of SIS mass distributions, which initially cannot capture elliptical features. The network consists of mirrored convolutional encoder–decoder blocks:

- **Encoder** E_ϕ: four stride-2 convolutions (3×3 kernels, ReLu), halving spatial resolution at each step while doubling the feature depth starting at 128. Linear layer outputs a latent vector $z \in \mathbb{R}^{16}$.
- **Decoder** G_θ: starting from the 16-D latent sample, a linear projection reshapes to a $8\times8\times1024$ tensor, followed by bicubic upsampling interleaved with convolutions that halve the channel count at each scale. Bicubic interpolation avoids the checkerboard artifacts we observed with transposed convolutions or bilinear layers, at the price of decoding time.

We observe that a 16-dimensional latent space is sufficient to encode the desired mass distribution information (EPL κ-maps). Bicubic layers raise runtime but preserve the smoothness expected for this surface-density field.

Instead of an adversarial discriminator, we enforce the latent distribution to match a standard Gaussian via Maximum Mean Discrepancy (MMD) (Rubenstein et al., 2018; Nakagawa et al., 2023). Therefore, the loss is:

$$\mathcal{L}_{\text{WAE}} = \mathbb{E}_{\kappa\sim p_{\text{data}}}\left[\ell\big(\kappa, G_\theta \circ E_\phi(\kappa)\big)\right] + \lambda\,\text{MMD}(q_Z, p_Z), \tag{8}$$

where ℓ is the reconstruction loss, the L-2 loss in our case. q_Z is the aggregated posterior produced by E_ϕ, and $p_Z = \mathcal{N}(0, I)$, a Gaussian in 16 dimensions. For any kernel k,

$$\text{MMD}(P, Q) = \mathbb{E}_{x,\,x'\sim P}[k(x, x')] + \mathbb{E}_{y,\,y'\sim Q}[k(y, y')] - 2\mathbb{E}_{x\sim P,\,y\sim Q}[k(x, y)]. \tag{9}$$

We use the inverse-multiquadratic kernel:

$$k_{\text{IMQ}}(z, z') = \frac{c}{c + \|z - z'\|_2^2}, \qquad c = 2d\,(d = 16), \tag{10}$$

whose slow tail decay yields a more discriminative divergence than an RBF kernel. During the first training phase, we multiply the reconstruction term by 10, effectively reducing λ so the autoencoder first learns to copy the maps accurately before being pushed to enforce its latent samples to follow a Gaussian distribution.

We note that the posterior sampling step was computationally expensive: due to the slow bicubic upsampling operations in our architecture, obtaining MAP estimates for 500 lenses for all iterations required roughly 3 GPU-months of compute (A100 40Gb).

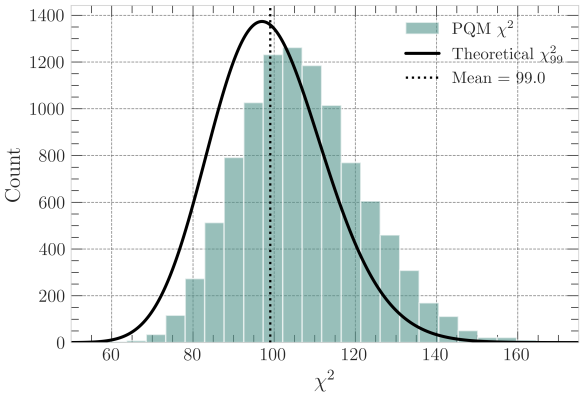

Figure 5. PQMass comparison between 1000 samples drawn from the prior learned by the WAE, *i.e.* $\kappa \sim p_1(\kappa)$ and 1000 samples drawn from the initial SIS prior, $\kappa \sim p_{\text{SIS}}(\kappa)$. 10000 re-tessellations were used to obtain a mean of 106.98 ± 15.34.

After training, we wanted to evaluate how closely the WAE's learned prior distribution $p_1(\kappa)$ resembles the original SIS prior $p_{\text{SIS}}(\kappa)$ that it was intended to model. Figure 5 summarizes one such comparison using the PQMass statistic. We generated 1000 random κ-maps from the trained WAE (*i.e.* $z \sim p_Z$ latent prior, $\kappa = G_\theta(z)$ giving $\kappa \sim p_1(\kappa)$) and compared them to 1000 κ-maps drawn from the true SIS prior ($\kappa \sim p_{\text{SIS}}(\kappa)$). We performed 10,000 re-tessellations. The result was a mean PQMass value of 106.98 ± 15.34 for the WAE samples vs SIS samples comparison, with an expected mean of 99 and a standard deviation of 14.07. This indicates a close, but not perfect, match of the samples from the two sets.

C. Posterior Sample Fitting

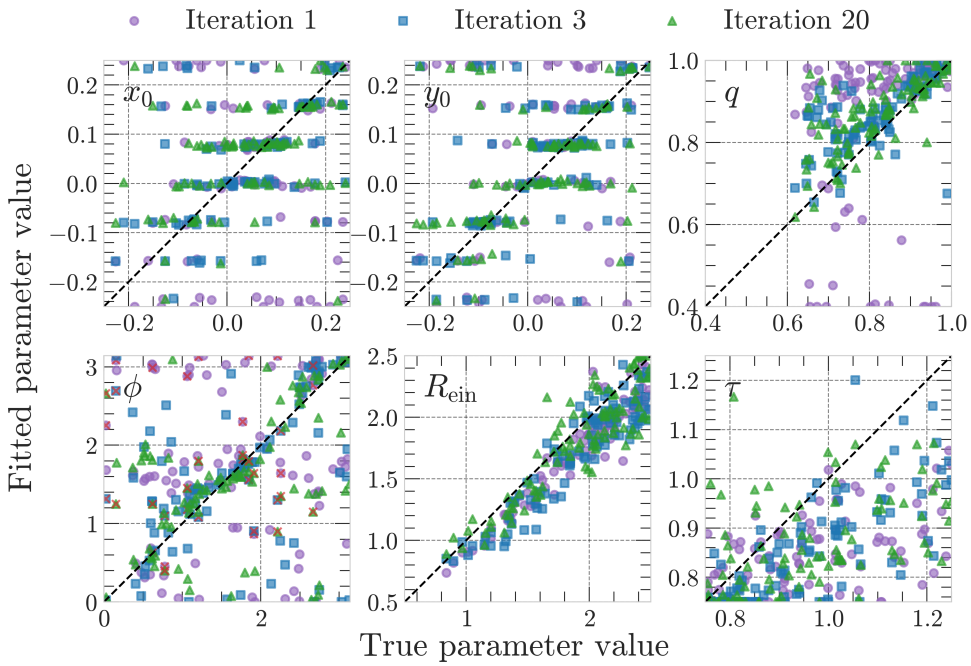

Figure 6. True parameter values against the fitted parameter values for six parameters of the EPL mass model: lens-center offsets (x_0, y_0), axis ratio q, position angle ϕ, Einstein radius R_{ein}, and radial slope τ. The black dashed line marks perfect recovery. The progressive clustering of points along this line illustrates how iterative retraining steadily reduces the bias and scatter of the recovered parameters. For the lens-center offsets and the Einstein radius panels, the axes are in arcseconds ($''$). Only 100 posterior samples' parameter recovery are shown for better clarity. Red crosses in the ϕ panel indicate lenses for which the true axis-ratio is bigger than 0.95 ($q > 0.95$).

Figure 6 presents the evolution of the fitted parameters depending on the retraining iteration. The method yields a very noticeable improvement in the κ-map shape parameters: the axis-ratio, q, and the orientation, ϕ. When the axis-ratio of a EPL is close to 1, its orientation is loosely unconstrained, thus, we don't expect to recover the orientation ϕ. Such points are shown in Figure 6 with red crosses.

The horizontal banding apparent in the lens-center panels reflects the pixelized nature of the WAE output: because each reconstructed κ-map is defined on a discrete grid, the fitted EPL center is effectively locked to the brightest pixel. Since the κ-maps field of view is 10 arcseconds wide, the pixel resolution is 0.078125 arcseconds.

D. Posterior Samples

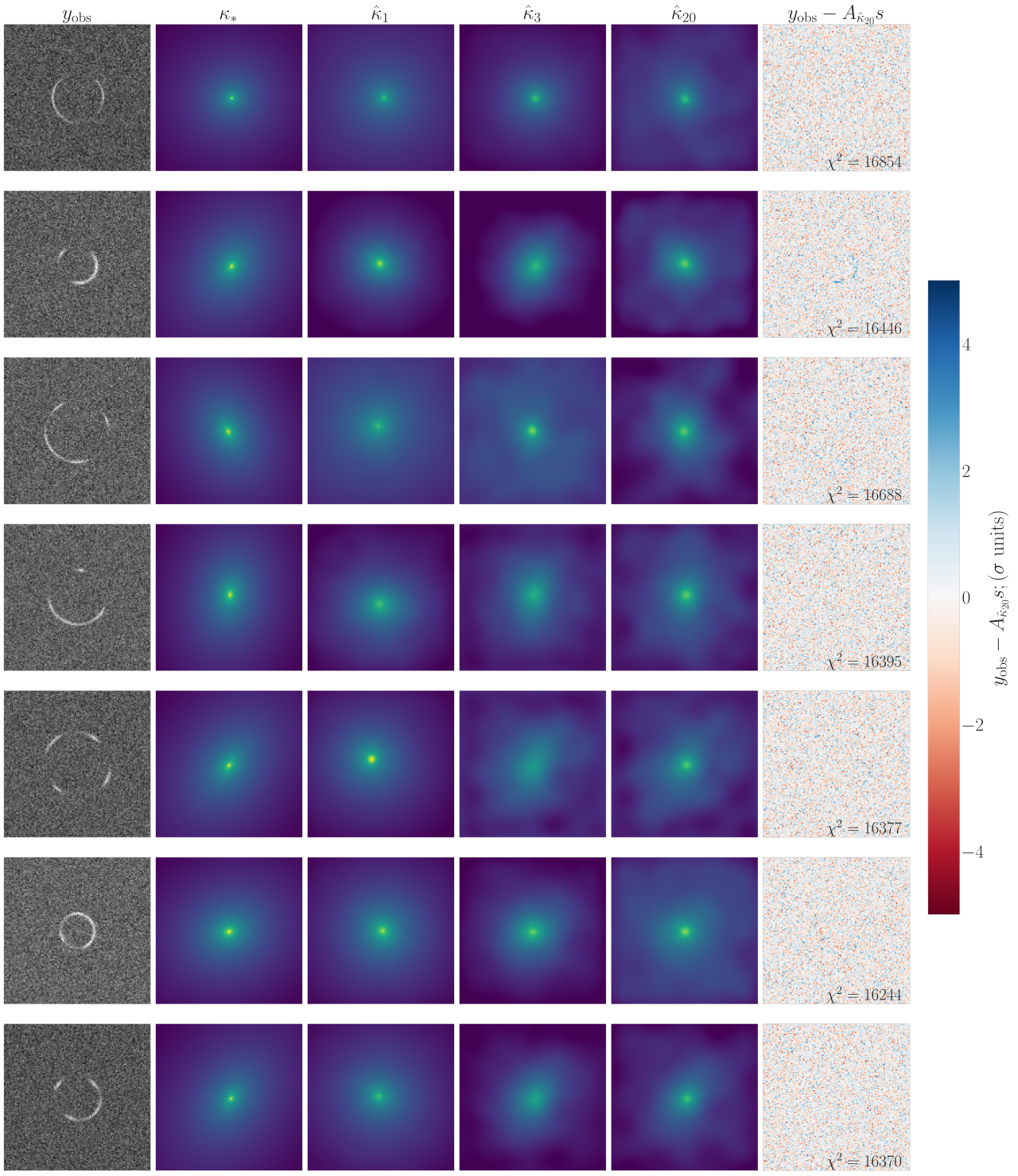

Figure 7. Posterior samples for different retraining steps. Columns from left to right contain the simulated observations, EPL κ-maps from the EPL prior, posterior samples from the WAE after retraining steps 1, 3, and 20, and the residuals for the posterior sample at retraining step 20. The κ-maps are plotted in log-space and normalized per sample (row). Ideal $\chi^2 = 16384$.

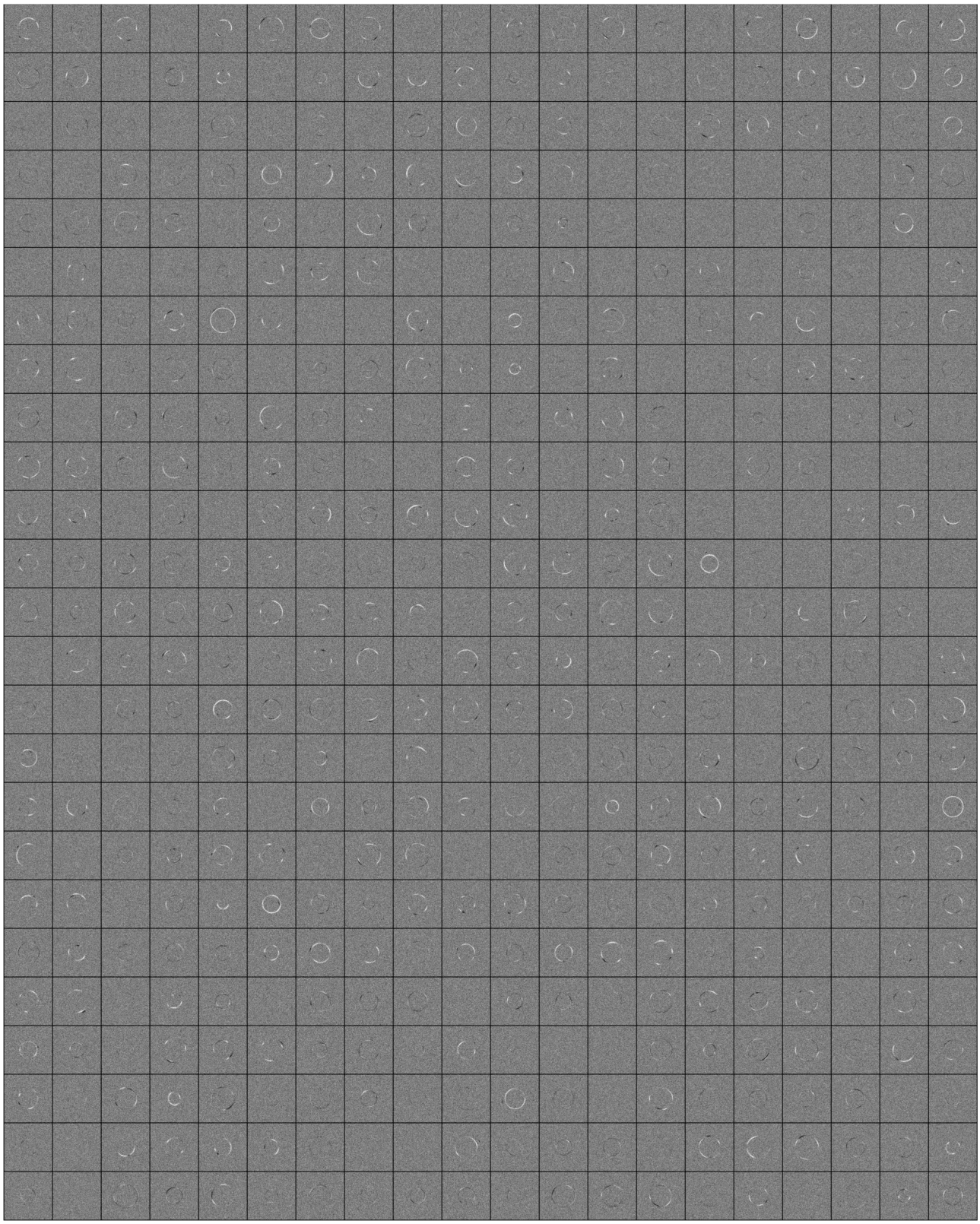

Figure 8. Residual of 500 posterior samples obtained during step 1. Residuals are normalized between $-5\sigma_\eta$ and $5\sigma_\eta$.

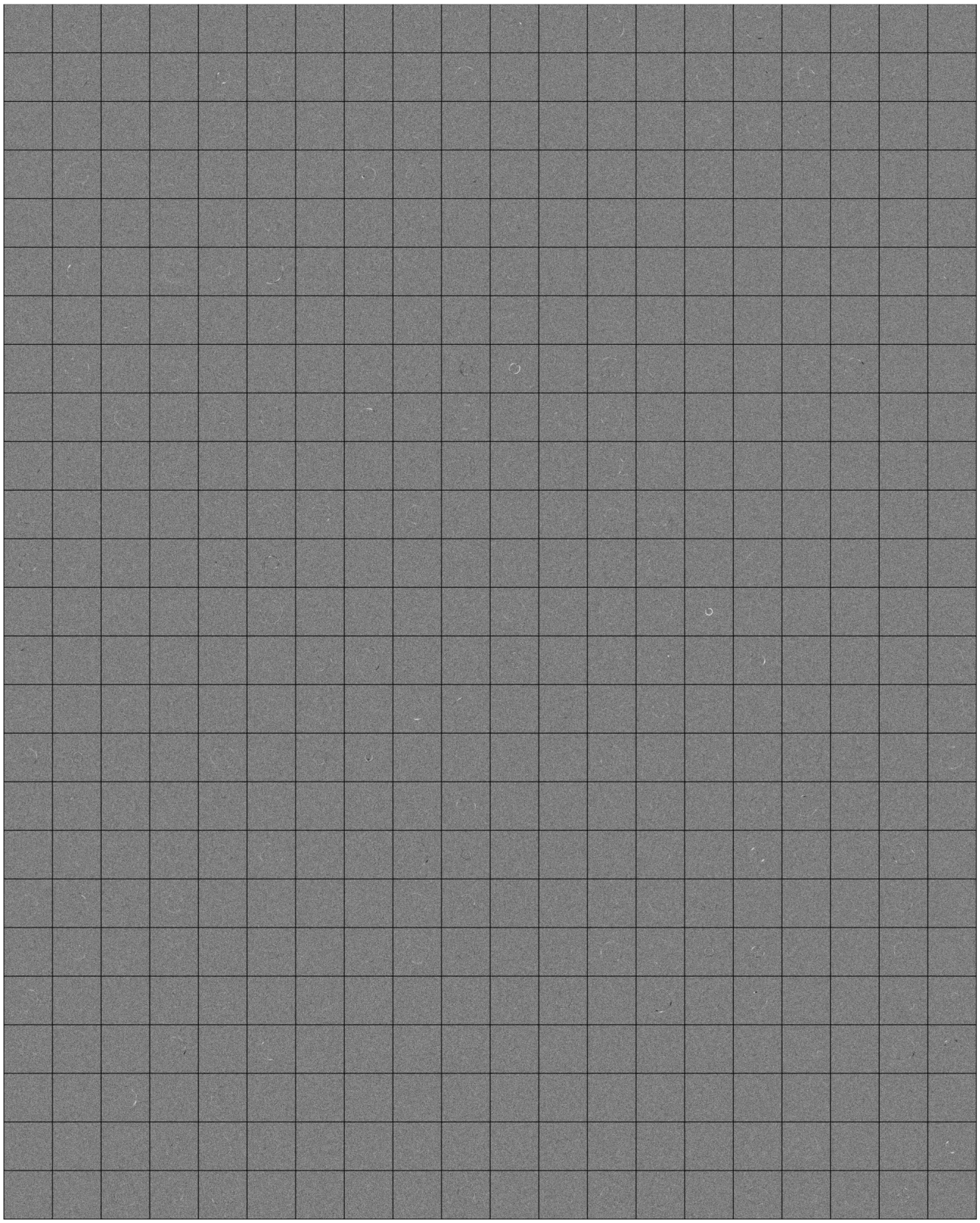

Figure 9. Residual of 500 posterior samples obtained during step 20. Residuals are normalized between $-5\sigma_\eta$ and $5\sigma_\eta$.