We thank the reviewers for their insightful feedback! We are encouraged that they appreciated the novelty of $\mathcal{M}$-flows (**R1**, **R3**, **R4**), their ability to learn the data manifold and a tractable density on it from samples (**R4**), and the exact invertibility on the manifold (**R3**, **R4**). We are glad that the reviewers liked our discussion of the subtleties of maximum-likelihood training (**R1**, **R2**) and appreciated the benefits of the new training scheme that separates manifold and density updates (**R1**, **R2**, **R3**, **R4**). They also found our experiments diverse (**R1**, **R2**) and convincing (**R2**, **R4**), noticed our improvements to the PIE baseline (**R2**) and commended the writing and pedagogical examples (**R2**). We answer some questions below, but will incorporate all feedback in the final version.

Why use normalizing flows for dimensionality reduction (R3)? Our primary goal was to construct a tractable probabilistic model for data on an n-dimensional manifold embedded in the data space, but in addition the coordinates of the learned manifold make a great candidate for dimensionality reduction: the flow approach ensures that for data exactly on the manifold, the compression to these variables is lossless, and the decoder is by construction the exact inverse of the encoder (unlike in autoencoders). Our goal is *not* to reduce the dimensionality further below n.

What are the convergence properties of the proposed training method (R4)? We wholeheartedly agree with **R4** that this needs more discussion. The convergence of the model to the correct manifold shape (defined by f) and to the true density on the manifold (defined by f and h) can be analyzed separately. 1) The ability of $\mathcal{M}$-flows to converge to the correct *manifold* is essentially the same as the considerations for autoencoders, with an additional architectural requirement of invertibility. For data on a manifold that can be described by a single chart with the latent space dimensionality (which we assume in this first work), this does not pose a restriction (related to the fact that all submanifolds that satisfy modest regularity conditions can be expressed as level sets of bijections). 2) If f has converged and learned the manifold, then learning the *density* on the manifold is a n-dimensional density estimation task. By implementing h as a flow that is a universal density approximator, we ensure that the $\mathcal{M}$-flow model can express any density on the manifold (up to some regularity assumptions). We do not study the convergence properties in detail, but argue that the loss will learn the correct distribution in the infinite capacity, asymptotic training limit. In that spirit, the argument is much like the initial claims for the ability of GANs to learn the distribution on a data manifold.

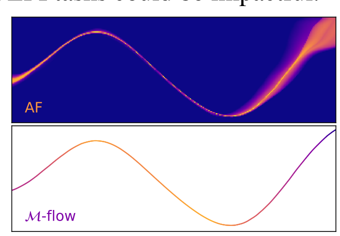
Is the sequential or alternating training scheme better (R4)? Unclear. We compare both approaches in the polynomial surface experiment, but did not find a clear advantage.

Comparison to autoencoders (AEs), VAEs, GANs (R1, R2, R3). We agree that a comparison to (V)AEs and GANs on generative and dimensionality reduction tasks would be very interesting. We also really like **R1**'s suggestion of comparing to an $\mathcal{M}$-flow-like model with a non-invertible decoder. In the paper we focused on AF and PIE because (like $\mathcal{M}$-flows) they allow for exact likelihood evaluations, which is crucial for inference tasks.

It would be nice to have a different metric to compare the models (R1). We agree, but do not know any single metric of all relevant aspects. Since likelihood values on different manifolds cannot be compared, we chose to study different metrics of data generation (manifold distance, FID scores, physics closure tests), manifold quality (reconstruction error when projecting to the manifold), inference (MMD between true and estimated posterior, log posterior evaluated at the true parameter point), and OOD detection (ROC AUC between in-distribution and OOD test samples).

In the particle physics task, it seems unfair to compare on a new metric of inference of underlying parameters (R1). Instead of "unfair" we would characterize it as a different metric that is often more relevant in a scientific context. *Likelihood-free inference* (LFI) is its own thriving research area with applications from neuroscience to epidemiology. Many state-of-the-art methods do not involve learning the likelihood, so the quality of likelihood estimation is not admissable to compare them. The good performance of $\mathcal{M}$-flows on LFI tasks could be impactful.

Did the AF baseline learn the manifold in the polynomial surface task (R1)? No. On the right we show the AF and $\mathcal{M}$-flow log likelihood along the two-dimensional slice $x_0 = 0$ through the 3-D data space. The AF density is sharply peaked around the true data manifold and most of the probability mass is very close to it. Still, it has non-zero support off the manifold, especially in regions of low density. In contrast, the $\mathcal{M}$-flow exactly learns a two-dimensional manifold. We thank **R1** for the suggestion of showing this explicitly and will include more results in the final version.



No PIE results for CelebA (R2). These were not completed in time for the submission. We have the answers now: on CelebA, PIE achieves FID scores of 75.7 ± 5.1, substantially worse than our $\mathcal{M}$-flow results and the AF baseline.

$\mathcal{M}$-flows did not outperform AF on CelebA (**R1**, **R2**). Yes. As **R2** pointed out, here we do not know the manifold dimension. Due to limited resources, we have not scanned over this hyperparameter or optimized the architecture. The good performance of $\mathcal{M}$-flows in datasets with known manifold dimension makes us optimistic that such a tuning will improve the results, but we are not in a position to make this claim at this time. We share **R2**'s hope that our results will spark more research along these lines, and were excited to see some steps at the recent ICML INNF+ workshop.