We would first like to thank the reviewers for their time and consideration.

Both **R2 and R3** wish us to elaborate on claims of **universality**. Firstly, we thank the reviewers for drawing our attention to the imprecise statement on L156-159, which should read 'a *differentiable* monotonic spline with sufficiently many bins can approximate any ~~continuous~~ *differentiable* monotonic function on the specified interval'. By definition of the derivative, a differentiable monotonic function is locally linear everywhere, and can thus be approximated by a piecewise linear function arbitrarily well given sufficiently many bins. Since rational-quadratic functions are differentiable, they are also locally linear, and in the limit of number of bins, each segment becomes linear. Satisfying differentiability everywhere means that continuous densities which are transformed by the spline do not become discontinuous, so restricting ourselves to this class of monotonic functions is reasonable, and we also note that all monotonic functions are differentiable almost everywhere regardless. Finally, as **R3** notes, for a fixed and finite number of bins, this universality of course does not hold, but our limit argument is similar to that used both for universality of neural networks in general, as well as that used by Neural Autoregressive Flows for their monotonic neural networks.

R2 correctly notes that families of functions with **infinite Taylor-series expansion** may not necessarily be very flexible. We will change this statement to highlight the precise flexibility that is added (ability to set derivatives and heights at each knot location), which if achieved by increasing the polynomial degree would make it more difficult to accurately invert the function.

R2 asks about **fixing the boundary derivatives** of the splines to match the linear tails with slope 1. If the derivatives at the boundaries of the splines are not set to match the values in the tails, the transformation is still continuous, but its derivative can have jump discontinuities at the boundary points. Since the derivatives of the transformations are the quantities which are accumulated by the change of variables to specify the exact log-likelihood, discontinuous derivatives mean that the log-likelihood training objective becomes discontinuous, which in our experience manifested itself in numerical issues and failed optimization. We observed this with both our proposed rational-quadratic transformations, as well as when modifying the quadratic splines of Muller et al. to include linear tails.

Regarding the **VAE experiments**, **R3** asks why we have chosen tasks that do not seem to require flexible density models of the latent space. Our intention when choosing the tasks was indeed to showcase the flexibility of the proposed flow; to that end we picked challenging image datasets with a sufficient number of datapoints to be able to learn a complex model of the latent space. Some of the typical benchmarks are only a fraction of the size, hence it is unlikely that additional flexibility would be helpful. This is supported by results in FFJORD and Sylvester Normalizing Flows, where flexible models do not always outperform simpler models on these datasets. We agree with **R3** that it is important to find more challenging applications of VAEs that would truly test the latent-space modelling; finding such novel benchmarks is outside of the scope of this work, but is an exciting research direction that would follow up on our findings.

R1 asks about the **coupling layer which transforms all variables**. This addition provided marginal rather than significant improvement, perhaps due to the fact that two successive coupling transforms which do not transform all variables can perform the same function. Nevertheless, we used this technique in all coupling-layer models except for the image experiments, where we found the marginal improvement not worth the extra parameter cost.

All reviewers mention the **importance of Appendix A**, and suggest some of its detail should be included in the main text. We agree with the reviewers that this detail is essential for clear specification of our model, and we will move some of this material to the main text.

Both **R2 and R3** ask for a discussion of **computational complexity and potential drawbacks** of the method. The operations added by our proposed transforms do not depend on the input dimensionality, meaning the transforms scale well to high-dimensional problems, as demonstrated empirically. The only non-constant operation is the binning of the inputs according to the knot locations, which can be efficiently performed in $\mathcal{O}(\log_2 B)$ time for B spline segments with binary search (since the knot locations are sorted). Moreover, due to the increased flexibility of the spline transforms, sequences of transforms required to build flexible flows are shorter (which we also observe empirically), reducing the computational cost. We agree it is important to discuss these considerations in the paper, and intend to add a similar discussion to the final version. One potential drawback of the proposed method is a more involved implementation; we alleviate this by providing an extensive appendix with technical details, and a reference implementation in PyTorch. Recently another group has independently reimplemented the transform in another framework using these resources.

New results: Finally, since submission we have extended our generative image-modelling experiments to include baseline results for all rational-quadratic models, where we replace the spline transforms with affine transforms, keeping all other choices fixed. Rational-quadratic splines improve upon the affine baseline in 3 out of 4 tasks, while matching the baseline score on the 5-bit CIFAR-10 dataset. This provides evidence that the parameter-efficiency results that we have observed in comparison with the full Glow model are not resulting from any changes unrelated to the elementwise transforms, and that rational-quadratic splines do indeed increase the flexibility of the coupling transforms.