We thank the reviewers for their time, helpful feedback, and advice. Overall, the reviewers praised the originality and clarity of the work. We thank them for their kind words, and hope to address any remaining concerns below.

(R1) Line 48 may be misleading. We agree and propose the following replacement: "We show that replacing VAE latent space components, which traditionally assume a Euclidean metric over the latent space, by their hyperbolic generalisation helps to represent and discover hierarchies." In particular, the prior and posterior probability densities are defined w.r.t. the volume induced by the metric tensor, the decoder (as opposed to concurrent work) treats latent variables as points in the Poincaré ball (by computing geodesic distances to hyperplanes) and the encoder projects points via the exponential map.

(R2) asked for evidence that the "hyperbolic" decoder was helpful. We would like to point out that we conducted an ablation study on this point, whose results are summarised in Figure 5. Admittedly, the figure and the related explanations are a bit condensed. We will improve that for the next version.

In more detail, we compared three decoders: (i) a standard "vanilla" multilayer perceptron (implicitly relying on the flat Euclidean geometry), (ii) a MLP precomposed by the logarithm map defined at the centre of the ball (can be seen as a linearisation of the manifold) and (iii) a decoder with a "hyperbolic" layer – described in Section 2.3 – which generalises a linear layer by computing geodesic distances to hyperplanes. Figure 5 shows improvements in terms of marginal log-likelihood estimates relative to the MLP baseline (i) for different latent dimensions and computed on the MNIST dataset. This ablation study shows that linearising the Poincaré ball through the logarithm map (i.e. decoder (ii)) before feeding latent variables through an MLP improves the performance compared to a vanilla MLP decoder. Yet, the composition of the logarithm map and a linear layer (i.e. decoder (i)) is not as good as a "hyperbolic" layer (i.e. decoder (iii)) which directly rely on the geometry of the Poincaré ball. Though, the differences in performances shrinks when the latent dimension increases.

For a detailed explanation of the analogy between trees and the hyperbolic space, we recommend reading section II of Krioukov et al. (2010). The analogy is not limited to the two-dimensional case. Although, as shown in De Sa et al. (2018), a two dimensional hyperbolic space is sufficient to embedded trees with arbitrarily low reconstruction error, if one has access to an arbitrarily high number of bits of precision.

(R4) Empirical comparison to related methods. We agree that comparing our method against concurrent work is indeed important. Unfortunately, the code or necessary experimental details have not been released, preventing us from doing so.

(R1, R4) Clarifications and minor typos. Thank you for pointing out some minor typos and places where clarifications could be useful. We will correct the typos in the text and move Figure 5 to the next page.

To clarify for (R1), $\mathcal{L}_{IWAE}$ refers to the IWAE unbiased estimate of the marginal likelihood introduced in Burda et al. (2015). We used 5000 samples in our experiments. We will include this definition in the next draft, thank you for pointing out our omission.

(R4) asked for additional high-level guidance in Appendix B. Thank you for the suggestion, we will reorder the subsections and write better connections between them so as to ease the reading.

(R4) Possible generalisations. Thank you for your suggestions. Indeed, we believe spherical distributions can be extended in a similar fashion. One could consider a wrapped Student-t as $Z \sim \exp_{\mu_\#} S_t(0, \nu)$, or a Riemannian Student-t with density (w.r.t. to the measure induced by the metric tensor) proportional to $\left(1 + d_M(z, \mu)^2/\nu\right)^{(-\nu+1)/2}$. As you point out, one could put a wrapped Gaussian process prior on the Poincaré ball to break the independence assumption between latent variables in VAEs. Concerning the limiting behaviour of the hyperbolic normal distributions, it appears that they are different, as the dimension space goes to infinity. Though in the context of dimensionality reduction, we believe that hyperbolic spaces are mostly useful in the low-dimensional setting.

References

Burda, Y., Grosse, R. B., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *CoRR*, abs/1509.00519.

De Sa, C., Gu, A., Re, C., and Sala, F. (2018). Representation Tradeoffs for Hyperbolic Embeddings. *arXiv.org*.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguna, M. (2010). Hyperbolic Geometry of Complex Networks. *arXiv.org*, (3):253.