DIFFERENTIABLE VQ-VAE'S FOR ROBUST WHITE MATTER STREAMLINE ENCODINGS

Andrew Lizarraga[1] *Brandon Taraku*[2] *Edouardo Honig*[1] *Ying Nian Wu*[1] *Shantanu H. Joshi*[2,3]

[1] Department of Statistics and Data Science, UCLA, USA
[2] Ahmanson-Lovelace Brain Mapping Center, Department of Neurology, UCLA, USA
[3] Department of Bioengineering, UCLA, USA

ABSTRACT

Given the complex geometry of white matter streamlines, Autoencoders have been proposed as a dimension-reduction tool to simplify the analysis streamlines in a low-dimensional latent spaces. However, despite these recent successes, the majority of encoder architectures only perform dimension reduction on single streamlines as opposed to a full bundle of streamlines. This is a severe limitation of the encoder architecture that completely disregards the global geometric structure of streamlines at the expense of individual fibers. Moreover, the latent space may not be well structured which leads to doubt into their interpretability. In this paper we propose a novel Differentiable Vector Quantized Variational Autoencoder, which are engineered to ingest entire bundles of streamlines as single data-point and provides reliable trustworthy encodings that can then be later used to analyze streamlines in the latent space. Comparisons with several state of the art Autoencoders demonstrate superior performance in both encoding and synthesis.

Index Terms— Streamlines, Diffusion Tractography, Differentiable, Gumbel Distribution, Vector Quantization

1. INTRODUCTION

Autoencoders (AEs), drawing inspiration from traditional factor analysis, have been successfully applied in data compression, segmentation, and representation tasks. However, their application in encoding high-dimensional structures, particularly white matter streamlines, has encountered emerging limitations [1] [2]. While various dimension reduction techniques, such as UMAP and tSNE, have been explored for white matter streamlines [3], recent advancements in encoder architectures, notably Variational Autoencoders (VAEs) [4] [5] and Vector Quantized-VAEs (VQ-VAEs) [6], have shown superior results in dimension reduction tasks. In light of these developments, our research focuses on leveraging these architectures to analyze white matter streamlines, aiming to overcome the shortcomings associated with traditional AEs.

The application of VAEs and VQ-VAEs to streamlines is not without challenges. VAEs, which strive to create meaningful latent encodings, encounter difficulties in optimization since encodings are required to be Gaussian. This process involves minimizing the KL-divergence through the Evidence Lower Bound (ELBO), a task that can result in noisy reconstructions. VQ-VAEs, on the other hand, eliminate the need to optimize the ELBO by using a uniformly distributed codebook of quantized vectors. The distribution for selecting codebook vectors is determined via an arg-minimization problem, making the KL divergence between codebook and selection distribution constant. Despite this advantage, VQ-VAEs introduce the issue of non-differentiable selection of codebook vectors, necessitating a straight-through estimator [6]. This means that the neural network can't backpropogate gradients to adjust the codebook vectors during training.

This has prompted techniques such as using an exponential moving average (VQ-EMA) to adjust and improve utilization of the codebook vectors. However, even with these additional improvements, the reconstruction results can still be noisy. Other proposals to effectively sample the codebook vectors (post-trainnig) to provide high quality image reconstructions, requires swapping the uniform prior on the codebook with a strong prior discovered by another model like PixelNet [6] or a Transformer [7]. But we don't have such luxuries for white matter streamline analysis given that there are so few architectures trained on streamlines and the datasets are typically too small to learn powerful auto-regressive models. To address these issues, we propose a novel Differentiable VQ-VAE (VQ-Diff) which allows for a fully differentiable approach to the quantization step in the traditional VQ-VAE. We demonstrate state of the art results for streamline reconstruction and empirically observe the models robustness to perturbations in the latent space suggesting that geometrically similar streamlines are grouped in similar neighborhoods.

1.1. Contributions

This paper makes the following contributions:

- We propose a novel neural network architecture (VQ-Diff) that improves upon the VQ/VAE models by enabling a differentiable approach.
- Our model avoids the need for optimizing the KL divergence as is done in traditional VAE based models.
- Our model parallels the reconstructive performance with AEs, yet yields a robust and reliable latent encodings of

streamlines.

- Our model demonstrates superior reconstruction performance compared to the state of the art VAEs, VQ-VAEs, and VQ-EMAs.
- We develop and release an open-source PyTorch dataset, derived from the *Tractoinferno* [8] dataset, but offering full latent space encodings based on our model VQ-Diff as well as other competing state of the art models.

2. METHODOLOGY

Unlike 2D images, a bundle of white matter streamlines is a collection of curves $B = \{f_i : \mathbb{R} \to \mathbb{R}^3 | i = 1, \ldots, N\}$. As a result, streamline bundles exhibit heterogeneous patterns, making their representation using codebook vectors challenging. Our approach involves composing weighted combinations of codebook vectors $\{e_1, \ldots, e_k\}$, allowing for more flexibility in the VQ-models. Typically, a VQ-model assumes a uniform prior $p(x)$ on the codebook and sets $q(x)$ to be a codebook selection distribution that arises from solving $\text{argmin}_i ||z - e_i||_2^2$. This is done in VQ-models because the KL divergence between $p(x)$ and $q(x)$ will be constant [6] and therefore these models don't need to optimize the ELBO which in turn produces less noisy reconstructions. However for our VQ-Diff model we let $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})$ and $q(x) = \frac{1}{\beta} \exp(-\frac{x}{\beta} - \exp(-\frac{x}{\beta}))$ be a zero-mean Gaussian and Gumbel distribution respectively. Then we take the weighted combination of the codebook vectors given by $s_j = \sum_{i=1}^{k} w_i e_i$ as our latent representation, where $e_i \sim p(x)$ and $w_i \sim q(x)$. Here, we show for the first time that the KL divergence between the Gaussian and Gumbel Distribution is constant: First assume $p(x)$ and $q(x)$ are zero-mean Gaussian and Gumbel distributions, as described earlier. Then the KL divergence is computed as follows:

$$
\begin{aligned}
&D_{KL}(p||q) = E_{p(x)}\left[\log\left(\frac{p(x)}{q(x)}\right)\right] \\
&= E_{p(x)}\left[\log p(x) - \log q(x)\right] \\
&= E_{p(x)}\left[-\frac{1}{2}\log(2\pi\sigma^2) + \log\beta - \frac{x^2}{2\sigma^2} + \frac{x}{\beta} + e^{\frac{-x}{\beta}}\right] \\
&= -\frac{1}{2}\log(2\pi\sigma^2) + \log\beta - \frac{1}{2} + E_{p(x)}\left[e^{\frac{-x}{\beta}}\right] \\
&= \text{const} + \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x}{\beta}\right) \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\
&= \text{const} + \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}\left(x + \frac{\sigma^2}{\beta}\right)^2 - \frac{\sigma^2}{2\beta^2}\right) dx \\
&= \text{const} + \exp\left(\frac{\sigma^2}{2\beta^2}\right) = \text{const}
\end{aligned}
$$

Similar to the VQ-VAE architecture, this is indeed an advantage as we do not need to optimize the ELBO. Additionally, since the Gumbel weighted sum is differentiable we may back-propogate gradients to update the codebook. Moreover, we may choose a flat Gumbel distribution to ensure the network utilizes all the codebook vectors and avoids codebook collapse [6]. In summary, we improve over the VQ-VAE's weaknesses by passing gradients to update the codebook vectors and the flat Gumbel distribution ensures we utilize all of the codebook vectors.

2.1. Architecture

The VQ-Diff architecture is comprised of a ResNet encoder and decoder with the bottleneck being a Gumbel Soft-max assignment of weights. Our architecture is modeled after the VQ-VAE, but as mentioned earlier in Sec. 2, the codebook vectors, e_k, are initialized with a Gaussian distribution, $p(x)$. Additionally instead of solving $\text{argmin}_i ||z - e_i||_2^2$ to assign a codebook vector to the encoded input z, we apply Gumbel Soft-max [9] across all distances $||z - e_i||_2^2$ to assign Gumbel weighted selection of the codebook vectors: $s_j = \sum_{i=1}^{n} w_i e_i$.

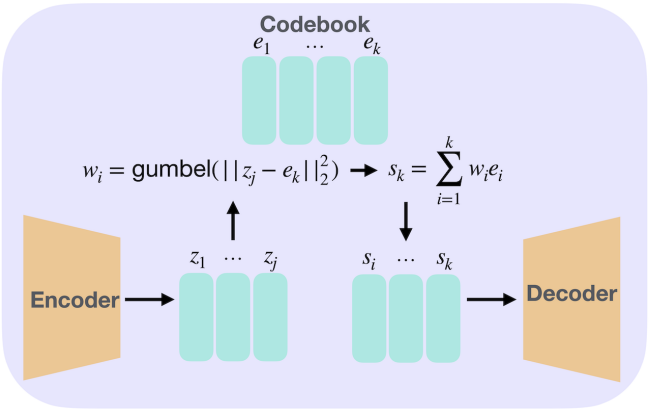

Fig. 2.1 Schematic of the VQ-Diff Architecture.

In summary, the encoder takes in a bundle B and assigns a collection of latent vectors z_j to each streamline. Then to each latent vector the network assigns a gumbel weighted combination of codebook vectors: $z_j \mapsto s_j = \sum_{i=1}^{n} w_i e_i$, see Fig. 2.1. The goal of the network is to learn a suitable codebook that captures features of streamlines that comprise a bundle. We compare this architecture against an AE, VAE, VQ-VAE and a VQ-EMA all composed of the same ResNet encoder and decoder. The full model implementation can be found at https://github.com/drewrl3v/diff-vq-vae.

2.2. Data

We use the open-access dataset, *Tractoinferno* [8], which consists of 284 datasets acquired from a variety of 3T scanners, to demonstrate the performance of our model. Here, streamline segmentation was performed with multiple techniques, resulting in 30 bundles per subject. In this paper, we only

make use of the streamline coordinates as processed in *Tractoinferno* [8]. Since not all bundles comprise the same number of streamlines, we selected tracts that consistently had over 1000 streamlines which resulted in 12 bundles, namely, the Middle Cerebellar Peduncle (MCP), Right Frontopontine Tract (FPT_R), Right Inferior Longitudinal Fasciculus (ILF_R), Left Inferior Fronto-Occipital Fasciculus (IFOF_L), Left Frontopontine Tract (FPT_L), Left Inferior Longitudinal Fasciculus (ILF_L), Left Parieto-Occipital Pontine Tract (POPT_L), Right Inferior Fronto-Occipital Fasciculus (IFOF_R), Right Parieto-Occipital Pontine Tract (POPT_R), FrontalRostrum of Corpus Callosum (CC_Fr_1), Left Pyramidal Tract (PYT_L), Right Pyramidal Tract (PYT_R).

We then sub-divided each bundle per subject into groups of 64 streamlines and up-sampled the number of points comprising a streamline to be 64 points. Thus a single data-point for our neural network yields a $(64, 3, 64)$ tensor (has a dimension (number of streamlines $\times$ dimension (3) $\times$ number of points). This is done for computational convenience to keep the bundle size consistent and to allow the network to ingest 256 batches of 64 streamlines during training for a total of $16,384$ streamlines per training iteration. Since some tract produce more streamlines than others, we down-sample the number of bundles per tract to ensure there is an equal number of each bundle per tract. This prevents the network from favoring a particular bundle due to its overrepresentation in the training set. We then split the dataset into a 90% train set, 10% validation set. This PyTorch white matter streamline dataset is now open-access, and publicly available under the *Tractoinferno* [8] license at https://github.com/drewrl3v/diff-vq-vae. To the author's knowledge, this is the first such dataset that provides not only our full model and its parameters, but also the encoded streamlines and their latent spaces generated for state of the art models that have been used on streamlines.

2.3. Training

All models were trained for $15,000$ iterations each with a mean-squared error (MSE) loss function penalizing for low reconstructive quality of streamlines. Experimentally we found that a setting a Gumbel temperature of $\beta = 10.0$ and assuming the codebook prior to be Gaussian with variance $\sigma = 2.0$ produced the best results for the VQ-Diff model. All models were ran on an AMD Ryzen Threadripper 3960X 24-Core Processor @ 3.8 GHz machine with a NVIDIA A6000 GPU and are released at: https://github.com/drewrl3v/diff-vq-vae.

3. EXPERIMENTAL RESULTS

3.1. Reconstructive Quality

The Bundle analytic (BUAN) score [10] is a state of the art method for comparing closeness of bundles of streamlines. We used a very low threshold tolerance of 0.05 for the bundle analytic score which makes the metric highly sensitive to minor differences among the bundles. After training we compared the BUAN scores across all the model architectures and bundles. A BUAN score closer to 1.0 signifies perfect reconstruction. Table 3.1 is a record of the average BUAN score across all bundles in the validation set along with the first standard deviation in the BUAN score. As we can see, the VQ-Diff is on par with AE in terms of reconstructive quality, while VAE does not fare so well since the ELBO enforcing a Gaussian latent space is difficult to learn. As suggested in Sec. 2, the VQ architectures, despite performing well in classical image reconstruction tasks perform poorly on streamline data.

This is because image intensities have a neighborhood structure and may be assumed to be piecewise continuous with more relaxed geometric constraints, while bundles of streamlines are composed of several individual fibers and have intrinsically complicated geometry[1].

Table 1. BUAN Scores Across Architectures

Bundle Name	VQ-Diff (Ours)	AE	VAE	VQ-VAE	VQ-EMA
PYT_R	0.9988 ± 0.0038	0.9999 ± 0.0003	0.4989 ± 0.1923	0.7154 ± 0.1227	0.6849 ± 0.126696
PYT_L	0.9988 ± 0.0039	0.9999 ± 0.0007	0.4764 ± 0.2045	0.7045 ± 0.1179	0.6778 ± 0.119968
POPT_R	0.9990 ± 0.0034	0.9999 ± 0.0008	0.4649 ± 0.2166	0.6867 ± 0.1494	0.6562 ± 0.150048
POPT_L	0.9989 ± 0.0037	0.9999 ± 0.0009	0.4815 ± 0.2145	0.6869 ± 0.1397	0.6508 ± 0.144388
ILF_R	0.9910 ± 0.0114	0.9999 ± 0.0008	0.2726 ± 0.2019	0.3409 ± 0.1631	0.2836 ± 0.159151
ILF_L	0.9910 ± 0.0112	0.9999 ± 0.0007	0.2399 ± 0.2018	0.3228 ± 0.1561	0.2408 ± 0.149510
IFOF_R	0.9888 ± 0.0133	0.9988 ± 0.0012	0.2808 ± 0.2366	0.3437 ± 0.1970	0.2793 ± 0.188292
IFOF_L	0.9855 ± 0.0164	0.9975 ± 0.0016	0.2340 ± 0.2130	0.3143 ± 0.1773	0.2296 ± 0.162993
FPT_R	0.9976 ± 0.0058	0.9999 ± 0.0008	0.2832 ± 0.2102	0.5063 ± 0.1795	0.4357 ± 0.188509
FPT_L	0.9971 ± 0.0074	0.9999 ± 0.0006	0.2689 ± 0.2261	0.4983 ± 0.2025	0.4335 ± 0.209599
CC_Fr_1	0.9974 ± 0.0110	0.9999 ± 0.0004	0.3750 ± 0.2033	0.3380 ± 0.1742	0.2458 ± 0.153786
MCP	0.9986 ± 0.0043	0.9999 ± 0.0008	0.3593 ± 0.2528	0.4989 ± 0.2298	0.3740 ± 0.232267

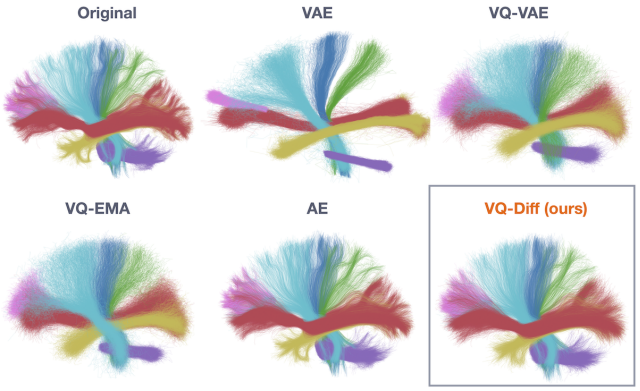

Fig. 3.1 Full subject reconstructions across architectures.

3.2. Visualizing the Latent Space

We visualize the latent spaces for each model by approximating the topology of the space by performing a tSNE [11] projection of the latent vectors, z_j. We see in Fig. 3.2 that the VQ-Diff is able to cluster respective bundle latent vectors and keep them roughly separated from other clusters. The VAE attempts to encode all latent vectors as Gaussian, which makes it difficult for the model to separate out categories, so we see a greater mixture of the latent vectors. The VQ-EMA is able to cluster the latent vectors but the clusters are more sparse. The AE on the other hand manages to cluster some of the latent vectors, but also mixes many of them in the center of the AE plot

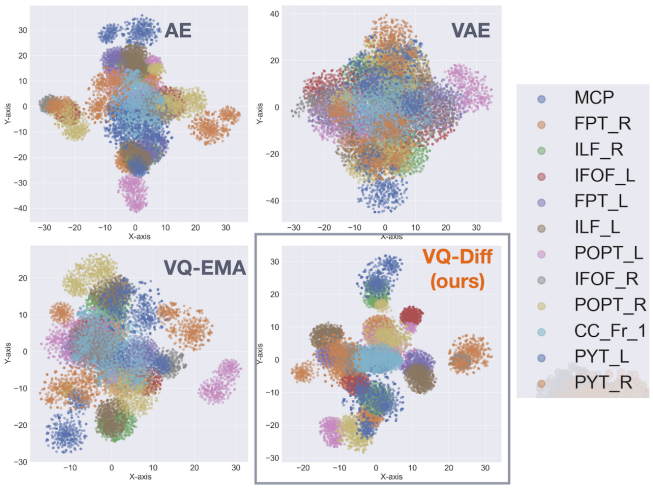

Fig. 3.2 Visualization of the latent space.

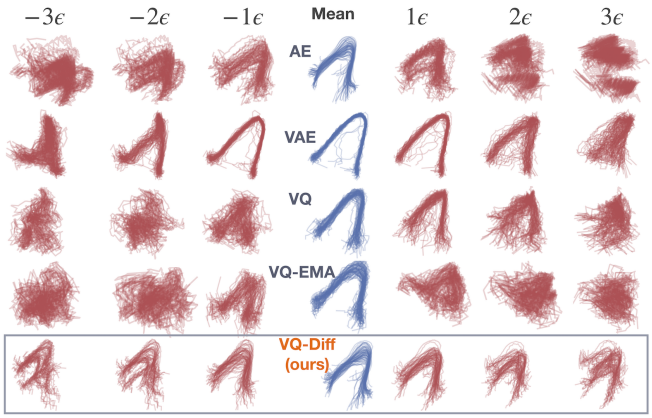

Fig. 3.3 Latent perturbations around the mean for the MCP bundle.

(Fig. 3.2). It is noted that tSNE doesn't necessarily represent distances in the projected representation in Fig. 3.2. To better understand the geometry, we instead isolate a latent vector and perturb it with noise. If the topology of the latent space is well regularized, then similar bundles should be mapped to a similar neighborhood. This means that reconstructions coming from a perturbed latent vector shouldn't drastically differ from the reconstruction coming from the original latent vector. The VQ-Diff plot in Fig. 3.2, which is tightly clustered for bundles of the same type, but is able to achieve a separation across different bundle types, suggests that it is robust to such perturbations. We perform an experiment to test this tolerance to perturbations in Sec. 3.3.

3.3. Perturbation Analysis and Synthesis

The reconstructive results of the VQ-Diff and AE are very promising. Given the strong reconstructive results for AE, Zhong et al. [1] and Legarreta et al. [12, 13] have suggested that the latent space can be used to perform statistical analysis of streamlines. To explore the feasibility of these ideas, we perform perturbation analysis of the underlying encoded latent vectors z_j. We choose the MCP bundle for demonstration purposes as it displays wide geometric variation in the population.

Across all models we map the same MCP bundle for the same subject to their corresponding latent vector representations z_j, then we perturb the vector by a small quantity: $z_j \mapsto z_j + \varepsilon$. We then pass $z_j + \varepsilon$ through each bottle neck layer for each architecture and reconstruct the bundle for each model architecture. As we see in Fig. 3.3, given a selected latent vector for the MCP streamline bundle, the AE performs poorly when the latent vectors are perturbed by small noise, $\varepsilon = 0.5$, while the VAE model performs better than AE as expected. The VQ-VAE and VQ-EMA models behave better for extremely small perturbations at the mean, but quickly degrade in quality with increasing ε. The VQ-Diff model demonstrates superior tolerant to such perturbations across all models. This suggests that the geometry of the latent space for the VQ-Diff groups similar streamlines in the same neighborhood and in turn is a very robust latent representation that can be used for more reliable distance analysis.

4. DISCUSSION AND CONCLUSION

In this work we provide the following: A new open source PyTorch dataset derived from the *Tractoinferno* dataset, a novel Neural Network Architecture that has the state of the art reconstructive performance of an AE, while also ensuring more robust latent representations. We demonstrate that the common assumption that the latent space of streamlines preserves local features does not hold for AEs. To the author's knowledge this is the first study to analyze more recent AE architectures for streamline analysis. We also observed that, while the reconstructive performance of the VAE is not on par with the AE or VQ-Diff, the Gaussian regularization on its latent space ensures that similar streamlines are within a neighborhood of the selected latent vector for MCP.

Overall, the VQ-Diff stands out as a highly robust architecture, having the potential to be trained across diverse MR image modalities. This flexibility underlines its potentially substantial impact in the field of medical imaging. In contrast, while the VQ-VAE and VQ-EMA exhibit limitations in effectively capturing the variability of streamlines in their codebooks, leading to lower reconstructive quality, they do offer a slightly more robust approach in terms of latent representations compared to the AE. This distinction highlights the unique strengths and weaknesses of these architectures, underscoring the VQ-Diff's strengths as a particularly valuable tool in medical imaging applications.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted using human subject data collected retrospectively and made available as an open-source project, *Tractoinferno* [8] https://openneuro.org/datasets/ds003900/versions/1.1.1/download. Thus ethical approval was not required under the licence attached to the open-source dataset.

6. ACKNOWLEDGMENTS

This research was supported by the NIH NIAAA (National Institute on Alcohol Abuse and Alcoholism) awards R01-AA025653 and R01-AA026834 (SHJ) and was partially supported by the NSF DMS-2015577 award (YNW).

7. REFERENCES

[1] Shenjun Zhong, Zhaolin Chen, and Gary Egan, "Auto-encoded Latent Representations of White Matter Streamlines for Quantitative Distance Analysis," *Neuroinformatics*, vol. 20, no. 4, pp. 1105–1120, Oct 2022.

[2] Andrew Lizarraga, Katherine L. Narr, Kirsten A. Donalds, and Shantanu H. Joshi, "StreamNet: A WAE for White Matter Streamline Analysis," in *Proceedings of the First International Workshop on Geometric Deep Learning in Medical Image Analysis*, Erik Bekkers, Jelmer M. Wolterink, and Angelica Aviles-Rivero, Eds. 18 Nov 2022, vol. 194 of *Proceedings of Machine Learning Research*, pp. 172–182, PMLR.

[3] Bramsh Qamar Chandio, Tamoghna Chattopadhyay, Conor Owens-Walton, Julio E. Villalon Reina, Leila Nabulsi, Sophia I. Thomopoulos, Eleftherios Garyfallidis, and Paul M. Thompson, "FiberNeat: Unsupervised White Matter Tract Filtering," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 5055–5061.

[4] Diederik P. Kingma and Max Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2014.

[5] Yixue Feng, Bramsh Q Chandio, Sophia I Thomopoulos, Tamoghna Chattopadhyay, and Paul M Thompson, "Variational autoencoders for generating synthetic tractography-based bundle templates in a low-data setting," *bioRxiv*, pp. 2023–02, 2023.

[6] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu, "Neural Discrete Representation Learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[7] Patrick Esser, Robin Rombach, and Björn Ommer, "Taming Transformers for High-Resolution Image Synthesis," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12868–12878, 2020.

[8] Philippe Poulin, Guillaume Theaud, Francois Rheault, Etienne St-Onge, Arnaud Bore, Emmanuelle Renauld a nd Louis de Beaumont, Samuel Guay, Pierre-Marc Jodoin, and Maxime Descoteaux, "TractoInferno: A large-scale, open-source, multi-site database for machine learning dMRI tractography," *bioRxiv*, 2021.

[9] Eric Jang, Shixiang Gu, and Ben Poole, "Categorical Reparameterization with Gumbel-Softmax," in *International Conference on Learning Representations*, 2017.

[10] Bramsh Qamar Chandio, Shannon Leigh Risacher, Franco Pestilli, Daniel Bullock, Fang-Cheng Yeh, Serge Koudoro, Ariel Rokem, Jaroslaw Harezlak, and Eleftherios Garyfallidis, "Bundle analytics, a computational framework for investigating the shapes and profiles of brain pathways across populations," *Scientific Reports*, vol. 10, no. 1, pp. 17149, Oct 2020.

[11] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[12] Jon Haitz Legarreta, Laurent Petit, François Rheault, Guillaume Theaud, Carl Lemaire, Maxime Descoteaux, and Pierre-Mar oin, "Filtering in tractography using autoencoders (FINTA)," *Medical Image Analysis*, vol. 72, pp. 102126, 2021.

[13] Jon Haitz Legarreta, Laurent Petit, Pierre-Marc Jodoin, and Maxime Descoteaux, "Generative Sampling in Bundle Tractography using Autoencoders (GESTA)," *Medical Image Analysis*, vol. 85, pp. 102761, 2023.