Disentangling Hippocampal Shape Variations: A Study of Neurological Disorders Using Mesh Variational Autoencoder with Contrastive Learning

Jakaria RABBI HTTPS://ORCID.ORG/0000-0001-9572-9010
jakaria@ualberta.ca
Department of Computing Science, University of Alberta, Edmonton, Canada.

Johannes Kiechle https://orcid.org/0009-0004-9610-146X
johannes.kiechle@tum.de
Institute for Computational Imaging and AI in Medicine, Technical University of Munich, Munich, Germany.

Christian Beaulieu https://orcid.org/0000-0002-2342-3298
christian.beaulieu@ualberta.ca
Department of Radiology and Diagnostic Imaging & Biomedical Engineering, University of Alberta, Edmonton, Canada.

Nilanjan Ray* https://orcid.org/0000-0002-7588-5400
nray1@ualberta.ca
Department of Computing Science, University of Alberta, Edmonton, Canada.

Dana Cobzas* https://orcid.org/0000-0003-1318-4868
cobzasd@macewan.ca
Department of Computer Science, MacEwan University, Edmonton, Canada.

Abstract

This paper presents a comprehensive study focused on disentangling hippocampal shape variations from diffusion tensor imaging (DTI) datasets within the context of neurological disorders. Leveraging a Mesh Variational Autoencoder (VAE) enhanced with Supervised Contrastive Learning, our approach aims to improve interpretability by disentangling two distinct latent variables corresponding to age and the presence of diseases. In our ablation study, we investigate a range of VAE architectures and contrastive loss functions, showcasing the enhanced disentanglement capabilities of our approach. This evaluation uses synthetic 3D torus mesh data and real 3D hippocampal mesh datasets derived from the DTI hippocampal dataset. Our supervised disentanglement model outperforms several state-of-the-art (SOTA) methods like attribute and guided VAEs in terms of disentanglement scores. Our model distinguishes between age groups and disease status in patients with Multiple Sclerosis (MS) using the hippocampus data. Our Mesh VAE with Supervised Contrastive Learning shows the volume changes of the hippocampus of MS populations at different ages, and the result is consistent with the current neuroimaging literature. This research provides valuable insights into the relationship between neurological disorder and hippocampal shape changes in different age groups of MS populations using a Mesh VAE with Supervised Contrastive loss.

Our code is available at `https://github.com/Jakaria08/Explaining_Shape_Variability`.

Keywords: Disentangled Representation Learning, Mesh Variational Autoencoder, Deep Learning, Contrastive Learning, Neurological Disorders, Medical Imaging, Hippocampal Shape Variations, Diffusion Tensor Image.

---

*. Equal contribution.

1. Introduction

1.1 Motivation and Objective

Advances in shape analysis and disentanglement techniques have contributed significantly to medical imaging, particularly in the 2D and 3D analysis of anatomical structures (Altaf et al., 2019). Latent space refers to a lower-dimensional representation of the complex high-dimensional space inherent in the data. Disentanglement involves the extraction and isolation of independent factors within this latent space, enabling a more interpretable and meaningful representation of anatomical variations from 2D and 3D datasets in the realm of medical imaging (Van der Velden et al., 2022).

Integrating latent space disentanglement techniques in medical image analysis helps reveal hidden factors that contribute to the observed variations in shapes and structures. Within the paradigm of 3D shape analysis, disentangling latent spaces holds profound potential for unraveling the complexities of diseases and age-related variations in brain structures (Kiechle et al., 2023). By isolating and understanding these latent factors, researchers can pave the way for more accurate diagnostic tools, predictive models, and a deeper comprehension of the underlying conditions driving anatomical changes.

Investigating the shape changes of the hippocampus over different age groups, especially in the context of neurological disorders, is complex when longitudinal data is not available, such as multiple Magnetic Resonance Imaging (MRI) scans taken at various ages for the same individual (Kiechle et al., 2023). Nevertheless, as shown through the current study, valuable information about hippocampal morphology and atrophy can still be extracted from existing limited datasets. Our proposed method intends to discern and describe the hippocampal shape variations in individuals across various age groups, differentiating between those with and without neurological conditions such as Multiple Sclerosis (MS) (Valdés Cabrera et al., 2023).

We use 3D mesh representation for our experiments as opposed to images or point clouds (Kiechle et al., 2023). The advantages of 3D mesh representation include its ability to capture complex surface details, providing a high-fidelity representation for studying anatomical shape variability. This representation allows for more intuitive and interpretable shape analysis, directly examining surface geometry for a clearer understanding of morphological changes (Lv et al., 2021). The 3D mesh also facilitates the application of advanced shape analysis techniques, aligning well with methodologies like statistical shape models and deep learning.

We employ Mesh VAE to obtain interpretable latent dimensions and generate valid shapes similar to the training dataset. We use a modified contrastive loss (Frosst et al., 2019) as a latent space disentanglement strategy to isolate two data generative factors: age and disease (MS). Age is represented by continuous values, while disease is labeled by discrete values. This combination of regression and classification tasks is frequently encountered in medical imaging. Our model formulation results in more interpretable latent codes and enables control over the generative process based on the specified factors (both classification and regression). As part of our validation process for the proposed method, we develop a 3D synthetic torus dataset with four factors of variability. During training and testing, we disentangled two of these factors using labels. Our results demonstrate

supervised disentanglement using both classification and regression data, both combined and separately.

1.2 Related Works

The disentanglement of the latent space has been the focus of numerous research works. In this section, we provide an overview of both supervised and unsupervised methods for disentangling latent variables. In section 1.2.1, we discuss the vanilla VAE and related works that enhance the disentanglement performance of the vanilla VAE. Most of these methods are unsupervised and disentangle the latent space without any prior knowledge of which variable disentangles which data-generating factor. We also examine supervised VAEs that disentangle specific latent variables using the data labels, and these methods exhibit strong disentangling performance for specific factors when compared to unsupervised approaches. Furthermore, we present some contrastive learning-based methods that enhance disentanglement, although they are also unsupervised methods. Additionally, we explore some graph autoencoder methods. Finally, in section 1.2.2, we explore disentanglement techniques that employ 3D mesh data with self-supervision and conditional VAEs and are related to our proposed method.

1.2.1 Disentangled Latent Representation VAEs

Variational Autoencoders (VAEs) represent a powerful class of generative models in machine learning that aim to capture the underlying structure of complex data (Kingma et al., 2019). VAEs consist of an encoder network, which maps input data to a probability distribution in a latent space, and a decoder network that reconstructs the input from sampled points in that space. The concept of disentanglement in VAEs addresses the challenge of extracting interpretable and independent features from the latent representation.

Higgins et al. (2016) introduced β-VAE, a framework for obtaining interpretable latent representations from raw image data through unsupervised learning. β-VAE modifies the traditional VAE by incorporating an adjustable hyperparameter, β, which influences the trade-off between latent channel capacity, independence constraints, and reconstruction accuracy. Factor VAE, introduced by Kim and Mnih (2018), addresses the issue of overly compact representations of β-VAE by incorporating a total correlation term in the objective function, promoting more independent and disentangled latent variables. DIP-VAE (Disentangled Inferred Prior VAE), proposed by Kumar et al. (2017), aims at mitigating the learning of trivial latent dimensions by introducing a penalty term that encourages the inferred posterior to have fixed marginals. Another method, β-TC-VAE (β-Total-Correlation-VAE), introduced by Chen et al. (2018), dynamically adapted the β hyperparameter for each latent dimension based on the total correlation, striking a balance between disentanglement and reconstruction accuracy.

Several methods provide insights and theoretical assessments on disentangled representations in VAEs, specifically in β-VAE. Burgess et al. (2018) proposed a training process modification for β-VAE that progressively increases latent code information capacity, facilitating robust learning of disentangled representations without sacrificing reconstruction accuracy. Estermann and Wattenhofer (2023) proposed another training approach for variational auto-encoders named DAVA (Disentangling Adversarial Variational Autoencoder).

DAVA effectively addresses the challenge of hyperparameter selection, reducing dependence on dataset-specific regularization strength. Additionally, Dupont (2018) proposed an unsupervised framework for learning interpretable representations, combining continuous and discrete features using variational autoencoders.

Some related works utilize supervised disentanglement methods, enhancing disentanglement by utilizing specific latent variables and labels. Ding et al. (2020) discussed methods employing both unsupervised and supervised learning in the context of generative models. They introduced an algorithm, Guided VAE, aimed at achieving controllable generative modeling through latent representation disentanglement learning. Cetin et al. (2023) proposed a supervised approach called Attri-VAE, which employs a VAE to generate interpretable representations of medical images. This method includes an attribute regularization term, associating clinical and medical imaging attributes with different dimensions in the latent space, facilitating a more disentangled interpretation of attributes.

An alternative approach for disentangling the latent space is to use contrastive learning in VAEs that proves beneficial for achieving improved disentanglement and generative performance. In a study by Deng et al. (2020), a methodology is introduced to generate facial images of virtual individuals with controlled and disentangled latent representations for identity, expression, pose, and illumination. The contrastive learning strategy is also applied to train autoencoder priors (Aneja et al., 2021) and masked autoencoders (Huang et al., 2023).

Another type of autoencoder called Graph autoencoder has become a crucial tool for studying graph-structured data. It enables the learning of meaningful representations for tasks such as node classification, link prediction, and clustering. A pioneering contribution to this domain was made by Kipf et al., who introduced the Graph Convolutional Network (GCN) in an autoencoder framework. Their approach, known as the Variational Graph Autoencoder (VGAE) (Kipf and Welling, 2016), combines the power of GCNs to aggregate and propagate node features with a variational autoencoder's capacity for learning interpretable latent representations. Further advancements include the work of Pan et al.'s Adversarially Regularized Graph Autoencoder (ARGA), and Adversarially Regularized Variational Graph autoencoder (ARVGA) (Pan et al., 2018), which incorporates adversarial training to enhance the robustness of learned representations, and Wang et al.'s Marginalized Graph Autoencoder (MGAE) (Wang et al., 2017), which introduced a denoising-based approach specifically tailored for graph clustering tasks.

1.2.2 Disentanglement in 3D Mesh Data using Mesh VAEs

While most disentangling methods are designed for images, researchers also use self-supervised and conditional VAEs to disentangle specific attributes in 3D mesh datasets. One of the research (Foti et al., 2022) introduced a self-supervised method for training a 3D shape VAE aimed at achieving a disentangled latent representation of identity features in 3D generative models for faces and bodies. The approach involves mini-batch feature swapping between various shapes to optimize mini-batch generation and formulate a loss function based on known differences and similarities in latent representations. Sun et al. (2022) proposed a VAE framework to disentangle identity and expression from 3D input faces that have a wide variety of expressions.

The two approaches mentioned for 3D VAE are not applicable to our specific medical domain problem, as we find neither feature swapping nor unsupervised learning appropriate. In our case, we possess labeled data and aim to disentangle multiple latent variables (for classification and regression) with supervision because supervised training increases the disentangling and reconstruction performance according to the previous research we discussed. Our work partially uses the process proposed by Kiechle et al. (2023), who explore a supervised variational Mesh autoencoder to understand and explain the variability in anatomical shapes. However, they only use the excitation-inhibition mechanism for the regression problem and used two additional neural networks for their method, which is different from our proposed method.

The domain of medical imaging often necessitates the disentanglement of various factors, including but not limited to age, diseases, and gender, through both classification and regression techniques. Therefore, we focus on simultaneous classification and regression techniques in the medical imaging domain with better loss functions. Our decision to focus on the two latent factors (classification and regression tasks) in the hippocampal study was primarily guided by their strong biological relevance and interpretability, particularly in the context of neurodegenerative diseases. These factors align with well-understood morphological variations in hippocampal structures, making our model's output more relevant for clinical research. The availability of supervised labels for these specific dimensions further supports our choice, allowing us to validate the model's disentanglement and ensure that the learned representations are meaningful and clinically significant. To the best of our knowledge, no prior work has utilized contrastive loss for both the classification and regression tasks (simultaneously) with a guided mesh VAEs to disentangle multiple latent variables. Our proposed method demonstrates superior disentanglement compared to guided VAE (Ding et al., 2020) and attribute VAE (Cetin et al., 2023) while achieving comparable generative quality and speed.

1.3 Contributions

The main contributions of this paper are summarized as follows:

- We introduce a novel contrastive loss for categorical and continuous labels to improve the disentanglement performance of specific latent variables through supervised learning using 3D mesh data.

- Our unified loss function incorporates both excitation and inhibition mechanisms for classification and regression tasks.

- We apply our novel formulation to analyze anatomical shape variations across various factors, including age and disease (MS), through the generation of 3D shapes.

2. Materials and Methods

Our proposed VAE designed for deep mesh convolution operates with an input comprising 3D mesh vertices denoted as $X = [x_0, x_1, ..., x_{N-1}]^T \in \mathbb{R}^{N \times F}$, where F represents the feature dimension, and N is the total number of vertices per mesh. In the context of 3D mesh data, F is specified as 3, and X contains the coordinates of each vertex.

The network's encoder is based on the SpiralNet++ structure (Gong et al., 2019), where all vertices of an input mesh are interconnected via a spiral trajectory that initiates from a randomly chosen vertex. The execution of spiral convolution operations involves two steps: initially, mesh vertices along the trajectory within a fixed distance are concatenated, a process known as neighborhood aggregation. Following this, the concatenated vertices undergo processing through a multilayer perceptron (MLP) with weight sharing (Kiechle et al., 2023). The decoder module performs a reverse transformation compared to the encoder using the latent space z. Our overall network architecture is shown in Figure 1.

Our method uses SpiralNet++ to exploit the local geometric structure of mesh data, preserving spatial relationships between vertices, which is crucial for accurately modeling the hippocampus and torus data. In contrast, we do not use PointNet-type models that focus on global features and may lose critical local geometric information. SpiralNet++ ensures consistent capture of local features through a fixed template, aligning with our goal of disentangling specific factors within hippocampus data (Gong et al., 2019). The importance of maintaining local mesh structure for accurate shape analysis, as highlighted by Litany et al. (Litany et al., 2018), further justifies the use of SpiralNet++ over PointNet-type models.

In the following section, we present the β-VAE used for our formulation. Then we discuss supervised guided VAE to explore the excitation-inhibition mechanism. However, we implement the mechanism differently. Finally, in section 2.3, we show the formulation of our method.

2.1 β-VAE (Higgins et al., 2016)

Our approach uses the β-VAE as the backbone of our network architecture. The VAE uses the Evidence Lower Bound (ELBO) as its objective function, expressed as:

$$\max_{\theta,\phi}\{L_{ELBO}(\theta,\phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot KL(q_\phi(z|x)\|p(z))\} \tag{1}$$

This equation reflects the balance between maximizing the reconstruction accuracy of the model, quantified by the expected log-likelihood of the observed data (x) given latent variables (z), which is $\log p_\theta(x|z)$, and minimizing the divergence between the posterior distribution of the latent variables under the encoder model $q_\phi(z|x)$ and the prior distribution $p(z)$, scaled by the hyperparameter β. Here, ϕ and θ are the parameters of the encoder and decoder network. This objective function can be decomposed into two primary components: the reconstruction loss and the Kullback-Leibler divergence, forming the VAE loss demonstrated by the following equations:

$$L_{vae} = L_{reconstruction} + \beta \cdot L_{KL} \tag{2}$$

where, L_{vae} is the overall β-VAE loss consisting of reconstruction and KL divergence loss (multiplied with hyperparameter β).

$$L_{reconstruction} = \|\hat{X} - X\|^2 \tag{3}$$

where, X is the input 3D mesh shape and $\hat{X}$ is the reconstructed shape.

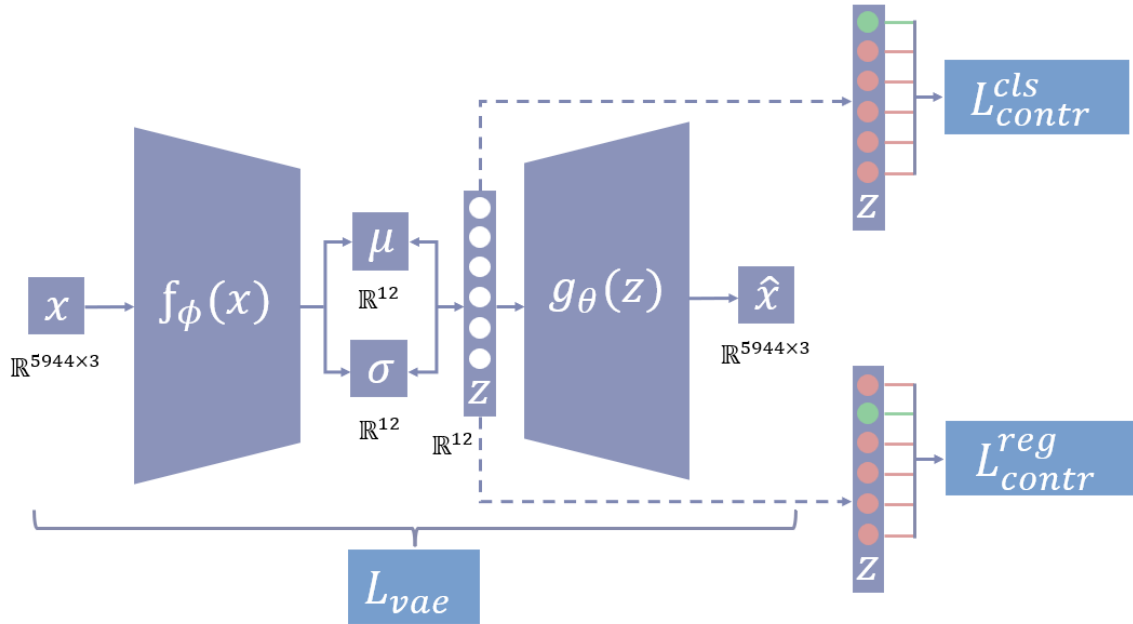$$L_{KL} = KL(q_\phi(z|x)\|p(z)) \tag{4}$$

Figure 1: Overall architecture of our method. We have mesh VAE with an encoder $f_\phi(x)$ and decoder $f_\theta(z)$ where x is the input 3D mesh and z represents the latent space. L_{vae} represents the VAE loss that combines reconstruction and KL divergence loss. Another two losses are classification loss L_{contr}^{cls} and regression loss L_{contr}^{reg}, where a specific latent variable is disentangled for a specific feature (continuous or discrete). We use the first variable for contrastive classification loss (z_1 corresponds to binary labels, and the rest variables are uncorrelated to the labels). The second variable z_2 corresponds to regression loss, and the rest variables are uncorrelated to the continuous labels.

where, L_{KL} measures the KL divergence between the posterior ($q_\phi(z|x)$) and prior ($p(z)$) distributon.

2.2 Supervised Guided VAE (Ding et al., 2020)

The objective function of the supervised guided VAE model is defined as follows:

$$\max_{\theta,\phi}\{L_{ELBO}(\theta,\phi) + L_{Excitation}(\phi,t) - L_{Inhibition}(\phi,t)\} \tag{5}$$

where L_{ELBO} is the evidence lower bound, $L_{Excitation}$ is the excitation loss, and $L_{Inhibition}$ is the inhibition loss calculated using separate feed-forward neural networks. The excitation loss is used to establish a correlation between a specific latent variable, for example, z_1, and a data generative factor like age or scale. The excitation loss is defined as:

$$L_{Excitation}(\phi,t) = \max_\omega\{E_{q_\phi(z_t|x)}[\log p_\omega(y|z_t)]\} \tag{6}$$

where z_t is the latent variable for supervised disentanglement, x is the input data, y is the label, and ω parameterizes the excitation network. The inhibition loss is defined as:

$$L_{Inhibition}(\phi, t) = \max_{\gamma}\{E_{q_\phi(z_{-t}|x)}[\log p_\gamma(y|z_{-t})]\} \tag{7}$$

In this context, z_{-t} represents a composite of latent variables excluding z_t, and γ denotes the parameters of the inhibition network. The methodology involves the training of distinct latent variables (z_t) to establish correlations with specific features within a dataset (y). The inhibition term is designed to avoid unintended associations between other latent variables (z_{-t}) and the labeled output. Our approach follows the excitation-inhibition mechanism like the guided VAE but with different losses without the need for separate neural networks.

2.3 Supervised Contrastive VAE (Ours)

We introduce a contrastive loss based on Frosst et al. (2019) applied to the excitation-inhibition mechanism inspired by Ding et al. (2020) and Kiechle et al. (2023) including a threshold hyperparameter in the regression loss for disentangling latent space of a VAE. The loss function L_{contr} is composed of three parts: L_{vae}, L_{contr}^{cls}, and L_{contr}^{reg} shown in 8. The first part is the loss function of the VAE, while the other two are the contrastive loss functions.

Contrastive loss learns the representations of the data in the latent space, and our first contrastive loss function L_{contr}^{cls} enforces the similarity between the samples of the same class and the dissimilarity between the samples of different classes. In our problem setting, classification inherently deals with discrete labels, where the model's objective is to distinguish between distinct classes. The loss function L_{contr}^{cls} encourages the latent space to separate data points based on these discrete labels. This loss only applies to the latent variable responsible for the classification task. We use it to disentangle the first latent variable (z_1) that correlates with the bump (present or absent) in the torus dataset and disease (healthy or MS) in the hippocampus dataset. Here, the binary labels are represented by y. The loss is demonstrated in Figure 1.

$$L_{contr} = L_{vae} + L_{contr}^{cls} + L_{contr}^{reg} \tag{8}$$

$$L_{contr}^{cls} = -\frac{1}{b}\sum_{i\in 1..b}\log\left(\frac{\sum_{\substack{j\in 1..b \\ j\neq i \\ y_i=y_j}} e^{-\frac{||z_1^i - z_1^j||^2}{T}}}{\lambda_1 \sum_{\substack{k\in 1..b \\ k\neq i}} e^{-\frac{||z_1^i - z_1^k||^2}{T}} + \lambda_2 \sum_{\substack{k\in 1..b \\ k\neq i \\ y_i=y_k}} e^{-\frac{\sum_{d\in 2..d_z}||z_d^i - z_d^k||^2}{(d-1)T}}}\right) \tag{9}$$

In equation 9, the loss function is estimated across the data batch b by sampling a neighboring point z_1^j for each point z_1^i in the latent space. The likelihood of sampling z_1^j depends on the distance between points z_1^i and z_1^j. The loss is represented by the negative logarithm of the probability of sampling a neighboring point z_1^j from the same class (y) as z_1^i. The temperature parameter T regulates the significance assigned to the distances between pairs of points. We implement the inhibition mechanism by introducing a term

$\sum_{\substack{k \in 1..b \\ k \neq i \\ y_i = y_k}} e^{-\frac{\sum_{d \in 2..d_z} ||z_d^i - z_d^k||^2}{(d-1)T}}$, weighted by λ_2 in the denominator of the loss and the formulation ensures that other latent variables from z_2 to z_{d_z} (d_z = number of latent variables) remain uncorrelated with the classification labels. We use all the values from the other dimensions from the latent space within the exponential term and use $d-1$ in the denominator to take the average. The loss formulation acts like excitation unit when $\lambda_2 = 0$ and $\lambda_1 = 1$.

On the other hand, Regression deals with continuous targets in our setting. We employ a threshold parameter (Th) in equation 10 to simulate a classification problem using continuous targets. This is a hyperparameter that controls the granularity of the regression task. It divides the values into separate bins and is completely different from the classification loss function. This loss only applies to the latent variable responsible for the regression task. We formulate the loss function to address our regression problem of disentangling z_2 (depicted in figure 1) based on continuous labels such as ages or scales. The loss function categorizes data objects into the same class if their labels fall within a specified range ($|y_i - y_j| \leq Th$) determined by the threshold.

$$L_{contr}^{reg} = -\frac{1}{b} \sum_{i \in 1..b} \log \left(\frac{\sum_{\substack{j \in 1..b \\ j \neq i \\ |y_i - y_j| \leq Th}} e^{-\frac{||z_2^i - z_2^j||^2}{T}}}{\lambda_1 \sum_{\substack{k \in 1..b \\ k \neq i}} e^{-\frac{||z_2^i - z_2^k||^2}{T}} + \lambda_2 \sum_{\substack{j \in 1..b \\ j \neq i \\ |y_i - y_j| \leq Th}} e^{-\frac{\sum_{d \in 1,3..d_z} ||z_d^i - z_d^j||^2}{(d-1)T}}} \right) \quad (10)$$

Overall, we employ the modified soft nearest neighbor losses (SNNL) (Frosst et al., 2019) as part of the inhibition-excitation mechanism, aiming to disentangle specific latent variables associated with distinct data generative factors. Our approach is scalable, allowing extension to more than two latent variables to disentangle additional data-generative factors of interest. While our model can disentangle other latent variables except the targetted ones, it does not enforce supervision for those variables. The SNNL focuses on enhancing latent representations' quality by promoting similarity among embeddings and assigning probabilities to all samples.

There exists an alternate contrastive loss, InfoNCE (Oord et al., 2018), which is formulated as a binary classification task, distinguishing positive from negative pairs. It compels the model to learn representations where positive pairs are more similar to each other than to negative pairs, effectively maximizing mutual information between positive pairs. However, we opt for a modified SNNL due to its less explicit differentiation between positive and negative pairs, emphasizing the creation of a smoother, probabilistic representation of similarity.

The denominator of our modified SNNL (equation 9 and 10) involves a sum over the exponential terms of all latent representations (z_1 and z_2) of the samples in the dataset, encompassing both positive and negative samples, and it encourages the model to assign higher probabilities to positive pairs without enforcing a strict binary distinction, as InfoNCE does. Therefore, our model and loss function can address both classification (equation 9) and regression problems (equation 10) using SNNL. The inclusion of an extra term in the denominator, weighted by λ_2, enhances the probability of attaining a more disentangled

representation. This is achieved by considering all latent representations in the variables not intended for disentanglement for a specific data generative factor.

3. Experiments and Analysis

3.1 Datasets

In this section, we provide an overview of the datasets utilized in this study, namely the hippocampus and synthetic data. All models compared in the results section are assessed using the data from both datasets.

3.1.1 Synthetic Torus Dataset

We have a hippocampus dataset that only includes a single scan per subject, it lacks the necessary ground truth to establish the relationship between shape and age. Furthermore, the data only offers scans for healthy and MS populations separately. Longitudinal data, on the other hand, can provide insight into the hippocampal shape of individual subjects over time, taking into account their MS status. Consequently, for evaluation purposes, synthetic data representing a torus with a bump (varying in size and presence/absence) is utilized following the method introduced in an article by Kiechle et al. (2023). We introduce four types of variability (scale of the torus, different noises, presence, and height of the bump) but only two variabilities (bump presence and torus scale) are disentangled in the latent space. We generate 5000 torus data by varying generative factors for our experiments. In figure 2, the color difference illustrates the dissimilarity between original and generated torus shapes, highlighting the variations in torus shapes by adjusting the values of the two latent variables controlling torus bump size and total scale.

3.1.2 Hippocampus Dataset

We utilize a neuroimaging dataset that incorporates diffusion tensor imaging (DTI) scans. The high-resolution data displays a voxel size of 1 mm isotropic, is acquired at 3 Tesla, and consists of volumes measuring $220 \times 216 \times 20$ mm^3 (Solar et al., 2021). This dataset encompasses scans from 204 healthy subjects spanning an age range between 32 to 71 years, with 112 females. Additionally, we have scans from subjects with MS (43 subjects aged between 32 to 71, with 35 females and the rest being males) (Valdés Cabrera et al., 2023).

The segmentation of the hippocampus in each scan for healthy subjects is conducted automatically (Efird et al., 2021) and manual segmentation is used for the MS subjects, followed by a series of preprocessing steps. Initially, the volumetric representations (i.e., voxel-based) underwent conversion into 3D mesh representations using a marching cubes algorithm (Lorensen and Cline, 1998). Subsequently, Laplacian surface smoothing and rigid alignment via an iterative closed point algorithm were applied to eliminate rotational artifacts. To ensure uniform topology across instances, Deformetrica (Durrleman et al., 2014) was employed to establish point correspondence across the meshes due to the assumption of meshes having the same topology (Gong et al., 2019). The result of this process is a collection of diffeomorphic deformation maps that illustrate the relationship between a computed mean atlas and the individual subject meshes. Each mesh is characterized by 5944 vertices
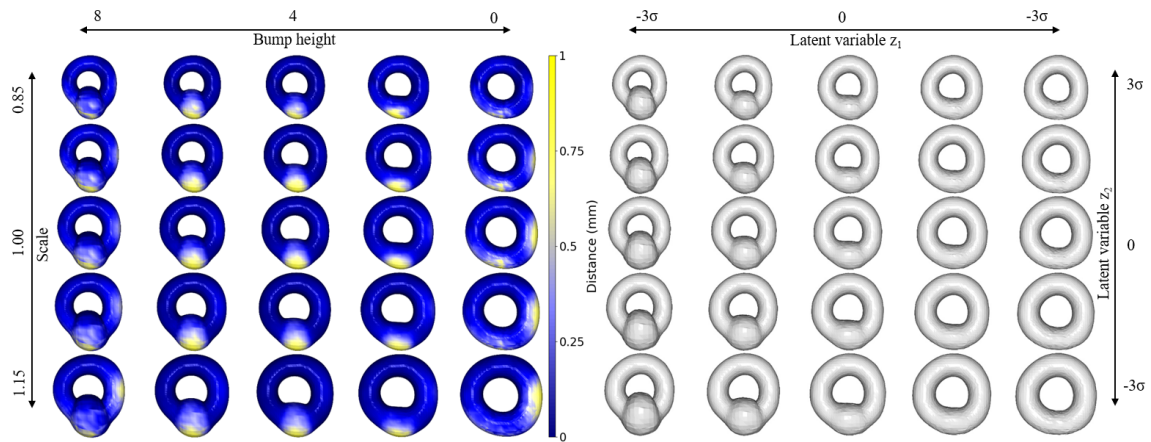
Figure 2: On the left side, we show the combination of reconstructions and original torus meshes from the synthetic dataset using our proposed model. The dark blue indicates a very small deviation between the reconstruction and the original mesh. We show two variabilities in the matrix of images: bump height and scale. On the right side, we show the decoder's output by varying the disentangled latent variable z_1 and z_2 in the x and y axis while holding the other latent variables constant at a mean value which is zero.
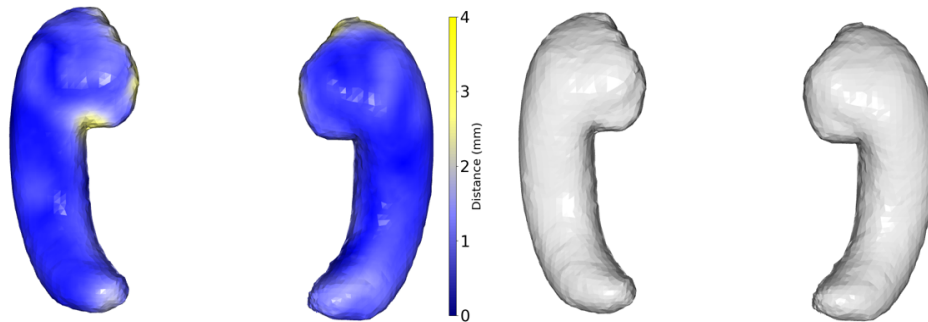

Figure 3: On the left side of the figure, we show the combination of reconstructions and original hippocampus (left and right hippocampus) meshes from the dataset using our proposed model. The dark blue indicates a very small deviation between the reconstruction and the original mesh. On the right side, we show the original hippocampus data.

and 11880 faces. In figure 3, the color difference illustrates the dissimilarity between original and generated hippocampus shapes, and on the right side, original hippocampus data is shown.

3.2 Implementation Details

Our mesh VAE architecture is similar to the SpiralNet++ (Gong et al., 2019). The encoder module comprises four spiral convolution layers with output channel sizes of [8, 8, 8, 8] and a latent channel size of 12. We use latent channel size as a parameter and find that a size of 12 gives the best results in terms of disentanglement. Sizes smaller than 12 reduce reconstruction accuracy, while sizes larger than 12 compromise disentanglement accuracy. We test sizes of 4, 8, 16, 32, and 12 balanced the trade-off most effectively, providing optimal performance for both disentanglement and reconstruction. The decoder module mirrors the transformations of the encoder. We set $\beta = 0.0015$, according to the parameter tuning results, and employ dilated spiral convolution with subsampling that enhances overall performance. A dilation factor of 2 and a spiral sequence length of 45 are used and those are selected by the parameter tuning process. In the numerical experiments, we adopt an 80/10/10 split for training/validation/testing. The ADAM optimizer is utilized with a batch size of 16, an initial learning rate of 3.6×10^{-4}, and a training horizon of 300 epochs. We use a scheduler to decay the primary learning rate by a factor of 0.77 every epoch. The proposed contrastive loss functions use temperature $T = 181$ and threshold $Th = .035$, and all the hyperparameter values are tuned using a hyperparameter optimization framework, Optuna (Akiba et al., 2019). We run all the experiments of models on Nvidia Titan RTX GPUs.

3.3 Evaluation Metrics

In this section, we present the evaluation metrics used to assess the performance of our proposed model. We focus on disentanglement, regression, and classification aspects, employing a variety of metrics suitable for each task.

3.3.1 Separated Attribute Predictability (SAP) (Kumar et al., 2017)

SAP score measures how well the model disentangles different attributes or factors of variation. It quantifies the ability to predict individual attributes from the learned representations. The computation of SAP score involves creating a score matrix, denoted as S, of dimensions $R^d \times k$ with d latent variables and k data generative factors. Each entry (i, j) in this matrix signifies the linear regression score for predicting the j-th factor using solely the i-th latent code. The R^2 value of the regression, denoted as S_{ij}, represents the predictability. Subsequently, for each column in S (corresponding to a factor), the SAP score is determined as follows:

$$SAP = \frac{1}{M} \sum_{i}^{M} (S_i^* - S_i^+) \tag{11}$$

In this equation, S_i^* is the highest score, S_i^+ is the second highest, and M denotes the number of considered factors.

3.3.2 Pearson correlation coefficient (PCC) (Cohen et al., 2009)

We use PCC to calculate the correlation between the values of a specific latent variable and the feature labels (continuous) that the variable is disentangling. The Pearson correlation

coefficient, denoted as r_{xy}, quantifies the linear relationship between two continuous variables. It measures how well the data points align along a straight line. The formula for Pearson correlation is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{12}$$

where N is the total number of data points. x_i and y_i represent the values of the two variables for the i-th data point. $\bar{x}$ and $\bar{y}$ denote the means of the x and y values, respectively.

3.3.3 POINT BISERIAL CORRELATION (PBC)(BROWN, 2001)

PBC is used for the correlation between the values of a specific latent variable and the feature labels (binary) that the variable is disentangling. Point biserial correlation measures the association between a binary attribute and a continuous variable and is defined by the following equation:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s}\sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{13}$$

where $\bar{X}_1$ and $\bar{X}_0$ are the means of the continuous variable for positive and negative classes, respectively, s is the pooled standard deviation, n_1 and n_0 are the sample sizes for positive and negative classes and n is the total sample size.

3.3.4 ACCURACY (ACC.)

Accuracy measures the proportion of correctly classified instances. We use K-nearest neighbor (Imandoust et al., 2013) for accuracy calculation for our classification task. The values of the specific latent variables are used to predict the labels.

3.3.5 MEAN SQUARED ERROR (MSE)

MSE quantifies the average squared difference between predicted and actual values. The outcomes of K-nearest neighbor (Imandoust et al., 2013) are used for MSE calculation for our regression task. The values of the specific latent variables are used to predict discrete labels.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{14}$$

where N is the total data points and y_i and $\hat{y}_i$ are the original and predicted labels.

3.3.6 RECONSTRUCTION ERROR (REC. ERR.)

Rec. Err. measures the dissimilarity (euclidean distance in 3D) between original and reconstructed 3D mesh shapes.

$$\text{Reconstruction Error} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i|_2^2 \tag{15}$$

where N is the total data points and x_i and $\hat{x}_i$ are the original and reconstructed mesh shapes.

3.3.7 1-Nearest Neighbor Accuracy (1-NNA)(Yang et al., 2019)

1-NNA evaluates the quality of learned representations by comparing nearest neighbors in the learned space using Chamfer Distance (CD) and Earth Mover's Distance (EMD). We calculate 1-NNA accuracy using the 3D coordinates from the original data and the 3D coordinates generated from the decoder of our model. We generate the latent variables from our learned distribution of z values from the training set.

Let S_g be the set of generated point clouds, and S_r be the set of reference point clouds with $|S_r| = |S_g|$, S_{-X} as the union of S_r and S_g excluding the element X, and let N_X represent the nearest neighbor of X within S_{-X}. The 1-NN accuracy, denoted as 1-NNA, for the 1-NN classifier is expressed as follows:

$$\text{1-NNA}(S_g, S_r) = \frac{\sum_{X \in S_g} I[N_X \in S_g] + \sum_{Y \in S_r} I[N_Y \in S_r]}{|S_g| + |S_r|}, \tag{16}$$

where $I[\cdot]$ represents the indicator function. In this context, each sample is classified by the 1-NNA classifier as either belonging to S_r or S_g based on the label of its nearest sample. If S_g and S_r are drawn from the same distribution, the accuracy of this classifier should approach 50% with an adequate number of samples. The proximity of the accuracy to 50% reflects the similarity between S_g and S_r, indicating the model's effectiveness in capturing the target distribution.

CD quantifies the dissimilarity between two point sets by measuring the average distance from each point in one set to its nearest neighbor in the other set. It is defined as follows:

$$\text{CD}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} |x - y|_2^2 + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} |y - x|_2^2 \tag{17}$$

where, X and Y are the two point sets. $|X|$ and $|Y|$ represent the cardinalities of sets X and Y, respectively. $|x - y|_2^2$ denotes the Euclidean distance between points x and y.

EMD, also known as Wasserstein distance, measures the minimum cost required to transform one point distribution into another. It considers the global distribution of points and accounts for both spatial arrangement and quantity. EMD is defined as:

$$\text{EMD}(X, Y) = \min_{\gamma} \sum_{(x,y) \in \gamma} c(x, y) \tag{18}$$

where, γ represents a transport plan that maps points from set X to set Y. $c(x, y)$ is the cost of transporting point x to point y.

3.4 Results

In this section, we present a comprehensive evaluation of our proposed model, Supervised Contrastive VAE (SC VAE), using the evaluation metrics discussed in the previous section. The comparison section compares our method with two baselines and two SOTA methods, using synthetic Torus and Hippocampus (containing both healthy and MS subjects) datasets. The comparison is based on disentanglement, correlation, prediction, reconstruction, and data-generative performance. The baseline models are β-VAE (Higgins et al.,

2016) and β-TCVAE (Chen et al., 2018) while we use Supervised Guided VAE (SG VAE) (Ding et al., 2020) and Attribute VAE (Cetin et al., 2023) as SOTA methods.

We provide an ablation study of our method, demonstrating the significance of the inhibition term. Furthermore, we present individual SAP scores for the discrete and continuous labels, when our model is trained to disentangle them separately. Then the training and test time comparisons are shown for all the five models. Lastly, we demonstrate the implementation of our model to analyze 3D hippocampus shape changes due to MS and aging.

3.4.1 Comparison

We compare the models in terms of disentanglement score (SAP), the correlation between the latent variables and labels (Corr.), accuracy (Acc.), and MSE score in predicting the labels from the values of the latent variables using the K-nearest neighbor classifier. The reported SAP scores in table 1 are calculated by averaging the SAP scores for z_1 and z_2 variables and the models are trained simultaneously for classification and regression tasks for all the scores.

The results, presented in Table 1, include SAP score (average of classification and regression SAP scores), correlation, accuracy, and MSE for all five models across both torus and hippocampus datasets. Our model demonstrates superior performance in SAP scores for both datasets while achieving comparable or better results in terms of correlation, accuracy, and MSE compared to the other models.

In Table 2, we present the reconstruction error and 1-NNA scores using CD and EMD for all five models across both datasets. Lower values indicate better performance for both reconstruction and 1-NNA scores. Our model demonstrates superior 1-NNA scores for the torus dataset using EMD. For both datasets, our model's scores, except for 1-NNA on the torus dataset, are either better or comparable to those of supervised models. However, baseline models exhibit better performance in terms of reconstruction error and the quality of data generation (1-NNA). These results align with expectations, as increased disentanglement in the latent space poses a challenge for models to simultaneously reduce reconstruction error and maintain high-quality data generation capabilities.

To summarize, our proposed method performs better than all four methods in terms of disentanglement, while also showing comparable or better results in terms of prediction and correlation. Additionally, our method performs similarly in terms of reconstruction error and data generation quality when compared to supervised disentangled methods.

3.4.2 Ablation Study

Our ablation study shows the significance of the inhibition term using different metrics. In Table 3, we show the ablation study, providing the scores of SAP, correlation, accuracy, MSE, and reconstruction errors for both torus and hippocampus datasets. Introducing the additional denominator term (inhibition when $\lambda_1 = 1$ and $\lambda_2 = 1$ in equations 9 and 10) results in an increase in SAP score compared to using the SNN loss without any modification (w/o inhibition when $\lambda_1 = 1$ and $\lambda_2 = 0$ in equations 9 and 10) for both datasets. Meanwhile, the scores of other metrics remain comparable. Here, lower MSE and reconstruction error scores signify improved performance.

Table 1: Comparison among models using SAP scores, correlation, accuracy, and MSE utilizing the hippocampus and synthetic torus dataset. A higher score is better (↑) for all the metrics except MSE (a lower score is better for MSE (↓)).

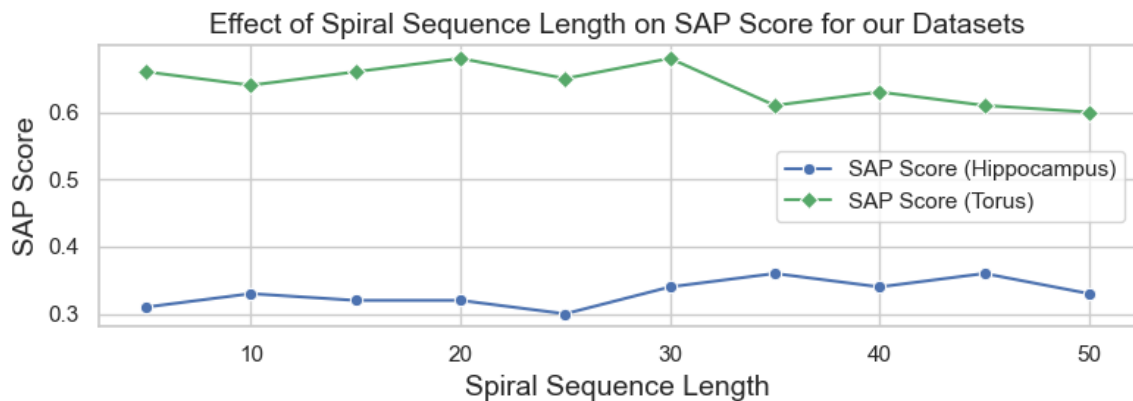| Model	Dataset	SAP (↑)	Corr. (↑)	Acc. (↑)	MSE (↓)
β-VAE	Torus	0.43	0.48	64.47	0.074
	Hippocampus	0.09	0.38	53.98	0.091
β-TCVAE	Torus	0.45	0.48	71.95	0.071
	Hippocampus	0.11	0.39	53.23	0.093
SG VAE	Torus	0.64	**0.78**	**100**	**0.013**
	Hippocampus	0.31	0.69	98.09	0.029
Attribute VAE	Torus	0.66	0.74	**100**	0.017
	Hippocampus	0.32	0.66	98.13	0.028
SC VAE (Ours)	Torus	**0.69**	0.75	**100**	0.016
	Hippocampus	**0.36**	**0.70**	**98.31**	**0.025**

Figure 4: Effect of Spiral Sequence Length on SAP Score for our Torus and Hippocampus Datasets.

Additionally, we conducted an ablation study to verify the effect of neighborhood information on disentanglement in mesh-based convolutional neural networks. The plot in

Table 2: Comparison of reconstruction and generative quality among models using both of our datasets. Reconstruction Error (Rec. Err.) and 1-NNA scores using CD and EMD are calculated for every model. Lower scores are better ($\downarrow$) for each model.

Model	Dataset	Rec. Err. ($\downarrow$)	1-NNA(%, $\downarrow$) CD	EMD
β-VAE	Torus	**0.25**	**51.37**	56.25
	Hippocampus	**0.85**	56.58	55.96
β-TCVAE	Torus	0.28	52.35	54.71
	Hippocampus	0.86	**56.35**	**55.43**
SG VAE	Torus	0.33	52.78	56.33
	Hippocampus	1.08	61.79	60.05
Attribute VAE	Torus	0.36	59.38	57.81
	Hippocampus	1.09	59.38	58.37
SC VAE (Ours)	Torus	0.38	51.56	**53.12**
	Hippocampus	1.07	58.69	59.76

figure 4 demonstrates that varying the spiral sequence length affects the SAP score for both the hippocampus and torus datasets. The non-linear relationship between spiral length and SAP scores shows that optimal performance occurs at specific lengths, implying that neighborhood connectivity impacts disentanglement. Therefore, the length of the spiral, as neighborhood information, has a significant effect on disentanglement. On a scale of 1, the SAP score varies in the range of 0.07 and 0.06 for the torus and hippocampus data, respectively. The scores would not change much or remain similar if a reasonable spiral length does not affect disentanglement. This justifies our use of SpiralNet++, which leverages such connectivity effectively.

3.4.3 Training and Test Time

We present the training (seconds per epoch) and test times (seconds per test set) for both the synthetic torus (80% of 5000 instances during training, and 10% each for test and validation) and hippocampus datasets (80% of 553 instances during training, and 10% each for test and validation). Guided VAE exhibits the longest training time due to additional parameters introduced by neural networks for excitation and inhibition mechanisms. Our method requires slightly more time than Attribute VAE during training, while the two

Table 3: Ablation Study of SC VAE by including and excluding the inhibition term in the denominator of our loss function. Higher scores are better (↑) for SAP, Corr. and Acc. and lower scores are better (↓) for MSE and Rec. Err.

Model	Dataset	SAP (↑)	Corr. (↑)	Acc. (↑)	MSE (↓)	Rec. Err. (↓)
SC VAE (w/o inhibition)	Torus	0.66	**0.77**	**100**	**0.016**	**0.37**
	Hippocampus	0.31	**0.70**	98.11	0.027	**1.07**
SC VAE (w/ inhibition)	Torus	**0.68**	0.75	**100**	**0.016**	0.38
	Hippocampus	**0.36**	**0.70**	**98.31**	**0.025**	**1.07**

baseline methods outperform supervised methods in terms of training time. During test time, our method demonstrates performance comparable to other methods.

Table 4: Training and Test Time (in seconds) for Torus and Hippocampus Dataset. A lower score is better for all the models.

Models		β-VAE	β-TC-VAE	SG VAE	Attribute VAE	SC VAE (Ours)
Training (Sec./Epoch)	Torus	**6.93**	7.89	23.14	9.8	10.50
	Hippocampus	**0.28**	0.29	0.91	0.33	0.49
Test (Sec./Set)	Torus	**0.24**	**0.24**	0.26	0.25	0.25
	Hippocampus	**0.02**	**0.02**	0.03	0.03	0.03

3.4.4 SAP SCORES FOR CLASSIFICATION AND REGRESSION TASKS

We present the mean SAP score within the comparison section. Figure 5 presents the classification and regression scores for both torus and hippocampus datasets when trained independently (All the other results in different sections were obtained by training classification and regression tasks simultaneously). Our approach demonstrates superior performance in classification and regression tasks for both datasets compared to other methods, except for

Attribute VAE, which does a slightly better job in the regression task specifically for the Torus dataset.
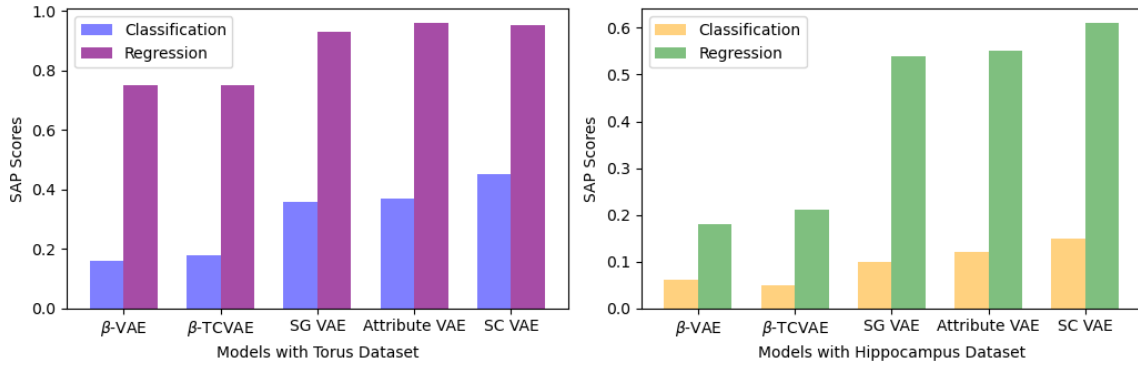

Figure 5: SAP scores (Classification and regression are separated) for different models using synthetic torus (left) and hippocampus (right) datasets.
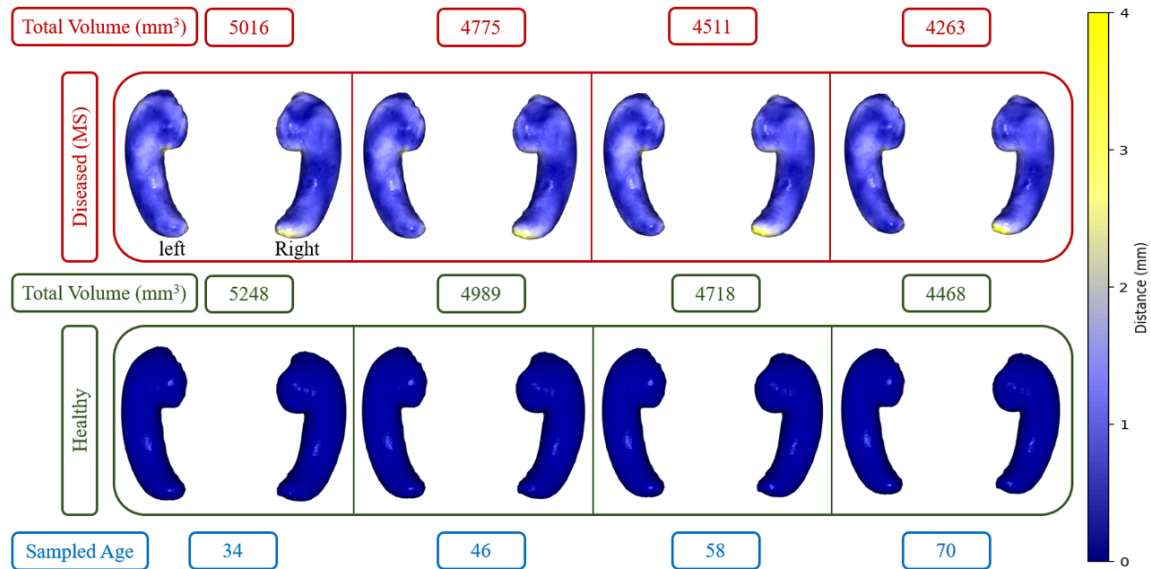
3.4.5 MS vs Hippocampal Volume Across Age Groups

Figure 6: Volume changes (between healthy and MS) are depicted in the first row by the intensity of the blue color and yellow represents the highest change in millimeters. The second row shows the healthy hippocampus. Ages are calculated by mapping the latent values and age range of the subjects of MS.

We employ our trained model to produce hippocampus shapes and assess the data-generating capabilities of our model within the domain of medical data. Our trained model demonstrates the ability to capture volume changes according to different data generative factors like age and diseases. By mapping the latent variable z_2 values within the range of -3σ to +3σ onto the age range of MS subjects, we select ages with intervals. Subsequently, we obtain the shapes of the healthy hippocampus by fixing the z_1 value at -3σ and the MS hippocampus by maintaining the z_1 value at 3σ.
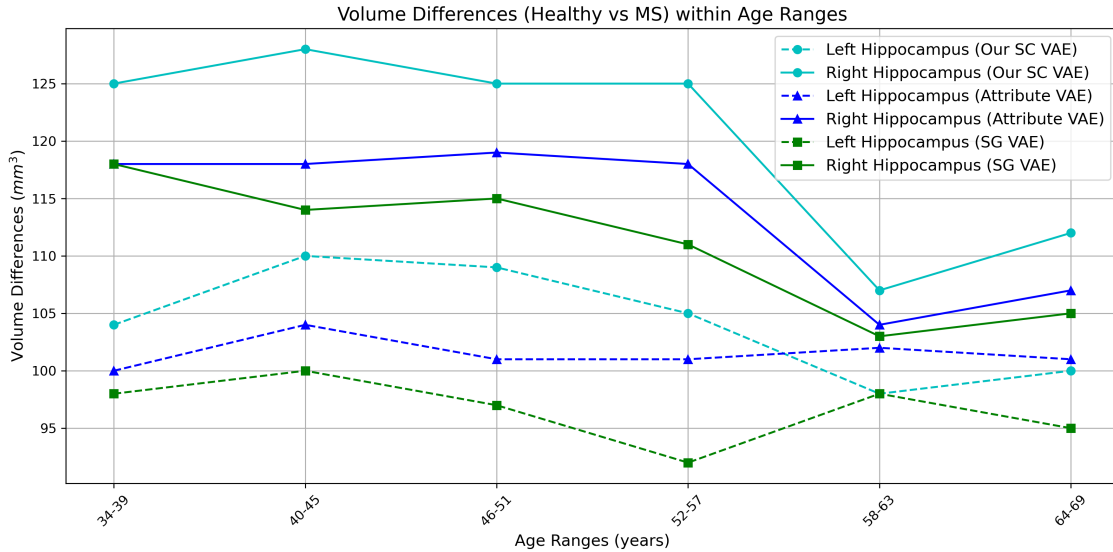

Figure 7: Volume differences between healthy and MS hippocampus are shown using six age ranges. Volume differences for the left and right hippocampus are shown separately for all three supervised disentanglement methods (SG VAE, Attribute VAE, and our SC VAE). The plot depicts a higher right hippocampus volume difference across ages derived by all three methods. The volumes are calculated by generating 10 shapes (using the three models) from each range, and then taking the average. Our method shows the highest volume difference ($\sim$4.5%) among the methods and is closer (compared to other methods) to the average trend (9%) found in a study by Valdés Cabrera et al. (2023) that calculated hippocampus volumes from DTI.

In Figure 6, we present the results for MS vs healthy controls. The lower row displays healthy hippocampus shapes for four sample ages, while the upper row depicts MS hippocampus shapes at those ages. Volume changes (between healthy and MS) are depicted in the first row by the intensity of the blue color and yellow represents the highest change in millimeters. The figure illustrates a noticeable decrease in hippocampus size with advancing age, particularly affecting the right hippocampus in MS. The tail of the right hippocampus has more atrophy than other regions.

Our findings align with previous research by Roosendaal et al. (2010) and Hulst et al. (2015), which reported a larger reduction in the volume of the right hippocampus compared

to the left hippocampus due to MS. Additionally, the overall hippocampus volume is lower in the MS population according to our results and the reduction of volume is also reported in the findings of Valdés Cabrera et al. (2023). However, the volume reduction ($\sim$4.5%) by our method is lower than Valdés Cabrera et al. (2023) (9%) who used DTI for calculating volumes, and it is expected because our classification (presence or absence of MS) SAP score is lower than the regression (age) SAP score. Therefore, we need more data for the MS population for more accurate results. We also show volume differences in Figure 7 for the three supervised disentangled methods (SG VAE, Attribute VAE, and our SC VAE) and our method shows the highest volume difference, which is closer (compared to other methods) to the average trend found in (Valdés Cabrera et al., 2023). Additionally, we conducted a one-sample t-test (Chorin et al., 2020) to examine whether there was a significant difference in hippocampal volume between healthy and MS populations, and we found the differences are significant ($p < .001$). In Figure 7, the plot illustrates the greater decrease in the volume of the right hippocampus compared to the left hippocampus in patients with MS. We generate the plot by generating 10 shapes (using the three models) from each range, calculating the volume and then take the average volume.

4. Discussion and Conclusion

4.1 Discussion

Our study contributes to the field of medical imaging and shape analysis in several ways. The proposed method enhances disentanglement performance for categorical and continuous labels in the context of 3D mesh data. The analysis of anatomical shape variations across various factors, including age and disease (MS), through the generation of 3D shapes, provides valuable insights into the relationship between neurological disorders and hippocampal shape changes.

Despite the promising results, there are limitations to our approach. The generalization of our model to diverse populations and datasets needs further exploration. We also need to improve the disentanglement performance of disease status for better prediction and reconstruction. Additionally, the mesh convolution technique we used, as outlined by Gong et al. (2019), necessitates the registration of meshes to a template mesh. Consequently, it is crucial to investigate methods that do not rely on assumptions about mesh topology for the analysis of complex shapes. Also, Our classification and regression losses are mostly similar, with only minor differences. Therefore, a unified formulation could be utilized to compactly represent the loss function. Future work would involve the incorporation of longitudinal data and exploration of the generalizability of the proposed method.

4.2 Conclusion

In this paper, we propose a novel approach for disentangling 3D mesh shape (hippocampal or synthetic) variations from DTI or synthetic datasets and applied in the context of neurological disorders. Our method, which uses a Mesh VAE enhanced with Supervised Contrastive Learning, exhibits superior disentanglement capabilities, particularly in identifying age and disease status in patients with MS. Additionally, our method demonstrates

comparable or better performance in all other metrics. The validity of our method is also demonstrated by a synthetic torus dataset.

We aim to extract meaningful representations of anatomical structures, providing insights into the complexities of diseases and age-related variations in the hippocampus by integrating novel and efficient latent space disentanglement techniques. Our method demonstrates the extraction of valuable insights into hippocampal morphology and atrophy linked to age and MS, even in the face of challenges posed by the absence of longitudinal data in limited datasets.

Acknowledgments

Ethical Standards

No ethics approval was required for the synthetic data analysis. The neuroimaging study was approved by the Human Research Ethics Board at the University of Alberta

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

The GitHub repository contains the script for generating synthetic data. Although the hippocampus data is confidential and cannot be shared, the preprocessing scripts are provided and can be utilized for any publicly available MRI data that includes hippocampus segmentation.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem Khalid Janjua. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572, 2019.

Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in neural information processing systems*, 34:480–493, 2021.

James Dean Brown. Point-biserial correlation coefficients. *Statistics*, 5(3):12–6, 2001.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in *beta*-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Irem Cetin, Maialen Stephens, Oscar Camara, and Miguel A González Ballester. Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders. *Computerized Medical Imaging and Graphics*, 104:102158, 2023.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Ehud Chorin, Matthew Dai, Eric Shulman, Lalit Wadhwani, Roi Bar-Cohen, Chirag Barb-haiya, Anthony Aizer, Douglas Holmes, Scott Bernstein, Michael Spinelli, et al. The qt interval in patients with covid-19 treated with hydroxychloroquine and azithromycin. *Nature medicine*, 26(6):808–809, 2020.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009.

Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and con-trollable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5154–5163, 2020.

Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7920–7929, 2020.

Emilien Dupont. Learning disentangled joint continuous and discrete representations. *Advances in neural information processing systems*, 31, 2018.

Stanley Durrleman, Marcel Prastawa, Nicolas Charon, Julie R Korenberg, Sarang Joshi, Guido Gerig, and Alain Trouvé. Morphometry of anatomical shape complexes with dense deformations and sparse parameters. *NeuroImage*, 101:35–49, 2014.

Cory Efird, Samuel Neumann, Kevin G Solar, Christian Beaulieu, and Dana Cobzas. Hippocampus segmentation on high resolution diffusion mri. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1369–1372. IEEE, 2021.

Benjamin Estermann and Roger Wattenhofer. Dava: Disentangling adversarial variational autoencoder. *arXiv preprint arXiv:2303.01384*, 2023.

Simone Foti, Bongjin Koo, Danail Stoyanov, and Matthew J Clarkson. 3d shape variational autoencoder latent disentanglement via mini-batch feature swapping for bodies and faces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18730–18739, 2022.

Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss. In *International conference on machine learning*, pages 2012–2020. PMLR, 2019.

Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Hanneke E Hulst, Menno M Schoonheim, Quinten Van Geest, Bernard MJ Uitdehaag, Frederik Barkhof, and Jeroen JG Geurts. Memory impairment in multiple sclerosis: relevance of hippocampal activation and hippocampal connectivity. *Multiple Sclerosis Journal*, 21(13):1705–1712, 2015.

Sadegh Bafandeh Imandoust, Mohammad Bolandraftar, et al. Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International journal of engineering research and applications*, 3(5):605–610, 2013.

Johannes Kiechle, Dylan Miller, Jordan Slessor, Matthew Pietrosanu, Linglong Kong, Christian Beaulieu, and Dana Cobzas. Explaining anatomical shape variability: Supervised disentangling with a variational graph autoencoder. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Thomas N Kipf and Max Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Or Litany, Alex Bronstein, Michael Bronstein, and Ameesh Makadia. Deformable shape completion with graph convolutional autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1886–1895, 2018.

William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

Chenlei Lv, Weisi Lin, and Baoquan Zhao. Voxel structure-based mesh reconstruction from a 3d point cloud. *IEEE Transactions on Multimedia*, 24:1815–1829, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Shirui Pan, Ruiqi Hu, Guodong Long, Jing Jiang, Lina Yao, and Chengqi Zhang. Adversarially regularized graph autoencoder for graph embedding. *arXiv preprint arXiv:1802.04407*, 2018.

Stefan D Roosendaal, Hanneke E Hulst, Hugo Vrenken, Heleen EM Feenstra, Jonas A Castelijns, Petra JW Pouwels, Frederik Barkhof, and Jeroen JG Geurts. Structural and functional hippocampal changes in multiple sclerosis patients with intact memory function. *Radiology*, 255(2):595–604, 2010.

Kevin Grant Solar, Sarah Treit, and Christian Beaulieu. High resolution diffusion tensor imaging of the hippocampus across the healthy lifespan. *Hippocampus*, 31(12):1271–1284, 2021.

Hao Sun, Nick Pears, and Yajie Gu. Information bottlenecked variational autoencoder for disentangled 3d facial expression modelling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 157–166, 2022.

Diana Valdés Cabrera, Gregg Blevins, Penelope Smyth, Derek Emery, Kevin Grant Solar, and Christian Beaulieu. High-resolution diffusion tensor imaging and t2 mapping detect regional changes within the hippocampus in multiple sclerosis. *NMR in Biomedicine*, 36 (9):e4952, 2023.

Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022.

Chun Wang, Shirui Pan, Guodong Long, Xingquan Zhu, and Jing Jiang. Mgae: Marginalized graph autoencoder for graph clustering. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 889–898, 2017.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4541–4550, 2019.