A generative modeling approach for interpreting population-level variability in brain structure

Ran Liu[1], Cem Subakan[2], Aishwarya H. Balwani[1], Jennifer Whitesell[3], Julie Harris[3], Sanmi Koyejo[4], Eva L. Dyer[1,5]

[1] School of Electrical & Computer Engineering, Georgia Institute of Technology
[2] Mila, Quebec Artificial Intelligence Institute
[3] Neuroanatomy Division, Allen Institute for Brain Science
[4] Computer Science, University of Illinois at Urbana Champaign
[5] Department of Biomedical Engineering, Georgia Institute of Technology

Abstract. Understanding how neural structure varies across individuals is critical for characterizing the effects of disease, learning, and aging on the brain. However, disentangling the different factors that give rise to individual variability is still an outstanding challenge. In this paper, we introduce a deep generative modeling approach to find different modes of variation across many individuals. Our approach starts with training a variational autoencoder on a collection of auto-fluorescence images from a little over 1,700 mouse brains at 25 micron resolution. We then tap into the learned factors and validate the model's expressiveness, via a novel bi-directional technique that makes structured perturbations to both, the high-dimensional inputs of the network, as well as the low-dimensional latent variables in its bottleneck. Our results demonstrate that through coupling generative modeling frameworks with structured perturbations, it is possible to probe the latent space of the generative model to provide insights into the representations of brain structure formed in deep networks.

Keywords: variational autoencoder · interpretable deep learning · brain architecture and neuroanatomy.

1 Introduction

Understanding how disease, learning, or aging impact the structure of the brain is made difficult by the fact that neural structure varies across individuals [15,6]. Thus, there is a need for better ways to model individual variability that provide accurate detection of structural changes when they occur. Traditional approaches for modeling variability [6,5] require extensive domain knowledge to produce handcrafted features e.g., volumetric covariance descriptors over pre-specified regions of interest (ROIs) [20,14]. However, in high-resolution datasets where micron-scale anatomical features can be resolved, it is unclear i) which features best describe changes of interest across many brains, and ii) how to extract these features directly from images. Thus, unsupervised data-driven solutions for discovering variability across many brains are critical moving forward.

In this work, we introduce a deep learning model and strategy for interpreting population-level variability in high-resolution neuroimaging data (Figure 1). Our model is a regularized variant of the variational autoencoder (VAE) called the β-VAE [8,3], and consists of an *encoder* and a *decoder* which work together to first distill complex images into a low dimensional latent space and next, expand this low-dimensional representation to generate high resolution images. Therefore, to gain insight into what the complete model has learned from the data, we take a *bi-directional* approach to characterizing how latent components are both, impacted by perturbations to specific regions in the input, via the encoder, and consequently impact specific regions of the generated output, via the decoder. Our work provides new strategies for understanding how different brain regions are mapped to latent variables within the network, an important step towards building an interpretable deep learning model that gives insight into how changes in different brain regions may contribute to population-level differences.

We applied this method to a collection of roughly 1,700 mouse brain images at 25 micron resolution from different individuals in the Allen Mouse Connectivity Atlas. By tuning the regularization strength in the β-VAE, we found that it is possible to both generate plausible brain imagery, as well as denoise images in the dataset that are corrupted by a number of artifacts. Our investigation into the latent space of this model also revealed a number of interesting findings; First, we found that information contained within the latent space is often asymmetric, with artifacts and noise being stored in one direction and biologically meaningful variance observed across many individuals in a separate direction within the same latent factor. Second, we found that multiple latent factors appear to generate outputs that vary within specific brain areas and thus have localized impact on generated outputs. Our results demonstrate that the proposed approach can be used to systematically find latent factors that are tuned to specific ROIs, and that generative modeling approaches can be used to reveal informative components of individual variability.

The contributions of this paper include: (i) the creation and specification of a β-VAE that can model high-resolution structural brain images, (ii) a bi-directional approach for revealing relationships between brain regions and latent factors in a deep generative network, and (iii) demonstration that structured perturbations to both image inputs and the latent space can reveal biologically meaningful variability.[1]

2 Methods

2.1 Model details

Low-dimensional models are used throughout machine learning to represent complex data with only a small set of latent variables. In deep learning, a bottleneck, i.e., layer with small width inside the neural network, often enforces a low-dimensional modeling of data. The VAE couples an autoencoder architecture

[1] Code and visualization can be found at: `https://nerdslab.github.io/brainsynth/`

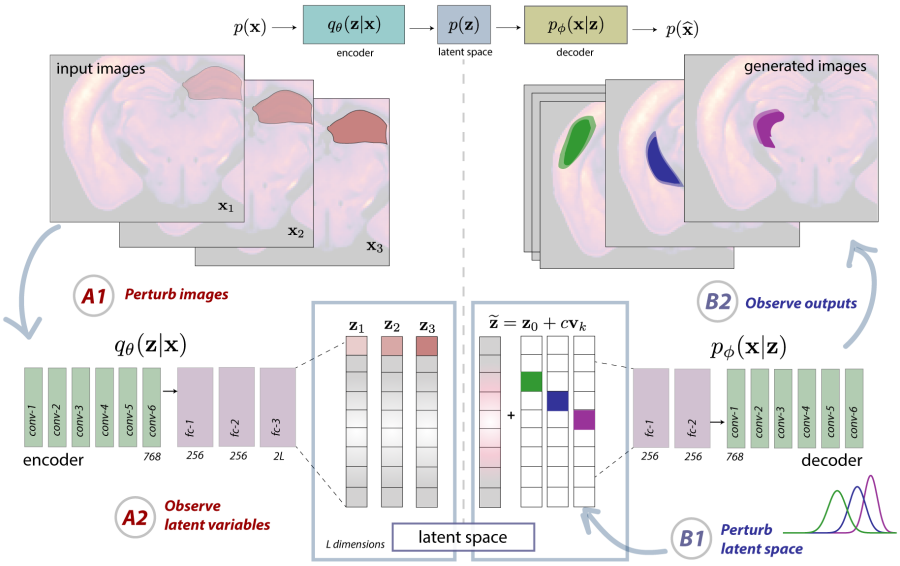

Fig. 1: *Visualization of our bi-directional approach for analyzing variational autoencoders trained to generate brain imagery.* On the left, we show a specific ROI being manipulated in a collection of input images (A1) and how this perturbation might result in a distinct shift in the latent representations (A2) formed from these inputs. On the right, we show the reverse process, where we perturb the latent space (B1) and observe the generated output images (B2).

[9,18] with a variational objective, thus providing a probabilistic view towards the generation of new high-dimensional data samples [10,17]. Much like regular autoencoders, VAEs embed information from the image space $\mathcal{X}$ into a latent space $\mathcal{Z}$ with latent dimension L via an encoder, and transform elements from the latent space into those in the image space via a decoder. The relationship between the encoder, decoder, and latent space can be written as:

$$\text{Encoder} : q(\mathbf{z}|\mathbf{x})p(\mathbf{x}) \mapsto p(\mathbf{z}), \quad \text{Decoder} : q(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \mapsto q(\hat{\mathbf{x}}), \tag{1}$$

where $p(\mathbf{x})$ denotes our dataset's distribution over the high-dimensional image space, $q(\mathbf{z}|\mathbf{x})$ and $q(\mathbf{x}|\mathbf{z})$ are, respectively, the distribution of the estimated encoder and estimated decoder, and $p(\mathbf{z})$ is the assumed prior on latent variables[2].

To train a good encoder (θ) and decoder (ϕ), the VAE aims to maximize the following objective:

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \tag{2}$$

The first term measures the likelihood of the reconstructed samples and the second term measures the KL-divergence between the estimated posterior distribution

[2] For simplicity, the prior is typically assumed to be Gaussian, $\mathbf{z} \sim \mathcal{N}(0, \boldsymbol{I})$.

$q_\phi(\mathbf{z}|\mathbf{x})$ and the assumed prior distribution. When $\beta = 1$, the model simplifies to a vanilla-VAE, whereas when β is a free parameter, the resulting model is referred to as the β-VAE [8]. Increasing the value of β encourages a certain degree of clustering, whereas lowering it encourages dispersion of similar elements in the latent space. Thus, by tuning β correctly, the model can learn to disentangle latent factors [8,3].

In our experiments, we used a β-VAE with a deep convolutional structure mimicking the DC-GAN architecture [16] (Figure 1). Our encoder had seven convolutional layers followed by three fully connected layers and used the ReLU activation function throughout. The same structure was mirrored for the decoder. The learning rate and batch size were set to 2e-4 and 64 respectively, resulting in a training time of roughly 4 hours on an Nvidia Titan RTX. After performing a grid search ($\beta = 1 - 20$, $L = 4 - 20$), we selected $L = 8$ and $\beta = 3$ as our model hyper-parameters since they exhibited performance that was relatively stable (i.e., these parameters produced an inflection point in evaluation metrics). The vanilla VAE's performance also exhibits an inflection point at the same latent dimension, which further confirmed that this choice holds for different amounts of regularization. In contrast, PCA continues to decrease its approximation error with higher dimensions; however, high-variance artifacts and other sources of noise are very quickly incorporated into the model when the bottleneck size increases beyond 30 dimensions.

2.2 Bi-directional latent space analysis

As images in our dataset are spatially aligned to an atlas, understanding how different regions of the pixel space are mapped to latent variables within the network can be a critical first step in building an interpretable model that gives insight into how different brain regions may contribute to population-level differences. To do this, we present a bi-directional approach to investigate the interaction between the image space and the β-VAE's latent space (see Figure 1). By understanding how the encoder and decoder work together to represent spatial changes in the data, we can build a more informed look into how brain structure can be modeled effectively within deep networks [11,21].

In one direction, we can map a latent variable's *receptive field* (left, Figure 1), i.e. which pixels in the input space impact each latent factor's activations. If changing the content of a region of the input image does not impact a specific unit, then the manipulated region is not in the unit's receptive field. To model this perturbation, let $\widetilde{\mathbf{x}} = \mathbf{x}_0 + w\mathbf{p}_\ell$ denote the perturbed input image, where $\mathbf{x}_0$ is the original image, $\mathbf{p}_\ell$ is a region specific (spatially localized ROI) perturbation, and w is the perturbation weight. By designing these perturbations to examine the responses of the units to changes in specific brain regions of interest, we can study the regional specificity of different units.

In the other direction, we can map a latent variable's *projective field* (right, Figure 1), or the parts of space that a latent variable affects when a new image is generated. To make this precise, let $\mathbf{v}_k$ be a canonical basis vector with a one in the k^{th} entry and zeros otherwise, c denote the interpolation weight, and $\mathbf{z}_0$

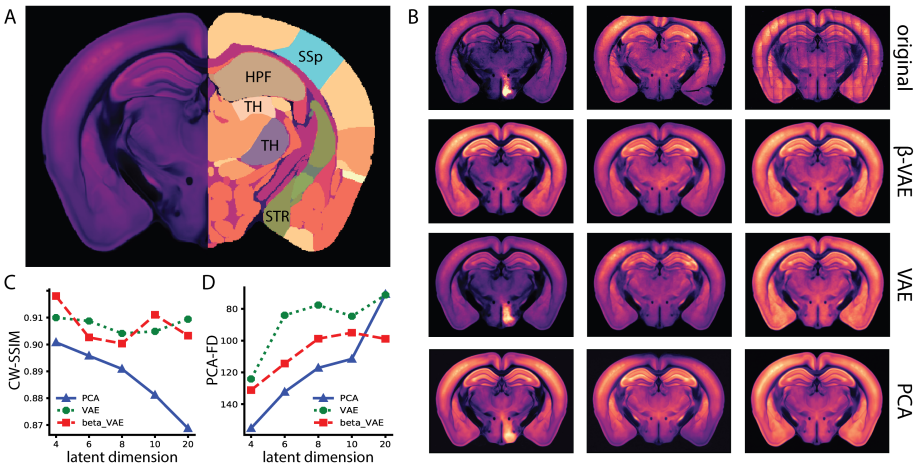

Fig. 2: *Evaluation of image synthesis and denoising performance.* **(A)** The left half of the image shows the average brain template and on the the right, we display simplified annotations for different regions of interest, including somatosensory cortex (SSp), hippocampal formation (HPF), striatum (STR), and parts of the thalamus (TH). **(B)** Examples of corrupted images with physical sectioning and grid-like artifacts along the top row. Below, we display the reconstructions obtained using a β-VAE, VAE, and PCA. The CW-SSIM and PCA-based FD scores for all three models are compared in **(C)** and **(D)**, respectively.

be the distribution mean. To generate an output image, we will first define the latent representation as $\widetilde{\mathbf{z}} = \mathbf{z}_0 + c\mathbf{v}_k$, and then pass this representation through the decoder to generate an image. We can use this synthesis approach to estimate the spatial extent of each factor's projective field by producing outputs across a range of different interpolation weights and then computing the variance of each pixel in the generated images.

3 Results

3.1 Dataset and pre-processing

To build a generative model of brain structure, we utilized registered images from 1,723 individuals within the Allen Institute for Brain Science's (AIBS) Mouse Connectivity Atlas [13].[6] The connectivity atlas consists of 3D image volumes acquired using serial 2-photon tomography (STP) collected from whole mouse brains (0.35 μm x 0.35 μm x 100 μm resolution, 1TB per experiment). Rather than using the fluorescence signal obtained from the viral tracing experiments (green channel), we obtained the auto-fluorescence signal acquired from each of

[6] The MCA is accessible through the Allen Institute's Python-based SDK [1] (`http://connectivity.brain-map.org/`)

the injected brains (red channel), which captures brain structure and information about overall cell density and axonal projection patterns. Our models were trained on 2D coronal sections extracted from near the middle of the brain (slice 286 out of 528) in each of the individuals in our dataset. This particular coronal slice was selected because it reveals key brain areas, including the hippocampus (HPF), regions of thalamus (TH), and parts of striatum (STR) (Figure 2A). The images were then downsampled from 0.35 microns to 25 microns, and centre-cropped to produce an image of size 320×448. In order to mitigate the effects of leakage of fluoresence signal, we pre-processed the data by adjusting each image's overall brightness to the dataset's average brightness and then set high intensity pixels 3.8 times over the average to this maximum value.

3.2 Evaluations and comparisons

To evaluate the image generation capability of our β-VAE model, we compared its performance with a vanilla-VAE and PCA. We first sought to examine each model by seeing how it performed when supplied with images containing three different types of artifacts: (i) corrupted bright areas due to leakage from the fluorescence signal's green channel, (ii) physical sectioning artifacts (missing data), and (iii) grid artifacts from scanning (Figure 2B, Supp. Materials S1). In these and other examples, we found that the β-VAE did the best job of removing artifacts from data while still preserving relevant biological variance. The ability of the β-VAE to reject artifacts is particularly pronounced in the case of classes (i, ii), where both PCA and VAE fail to reject the signal leaking into the channel of interest and cannot recover missing data. We observe that the β-VAE tends to learn a more accurate distribution over the dataset, while the vanilla-VAE overfits to the noise, and PCA does not deviate much from the mean in terms of its structural details.

To quantify the quality of images generated by the different models, we computed two metrics used to evaluate generative model outputs viz. the complex wavelet structural similarity (CW-SSIM) [19] (Supp. Material Sec. 1.1), and the PCA-based Frechet distance (PCA-FD) [7,12] (Supp. Material Sec. 1.2). When studying these metrics for different bottleneck sizes, we found that both for the β-VAE and the vanilla-VAE, latent dimensions in the range $L = 8 - 10$ produced stable performance (where scores plateau) before decreasing in accuracy. Analysis of the CW-SSIM scores along with visual inspection of the generated images, revealed that PCA is unable to capture high-dimensional textural details for low dimensions and quickly begins to represents artifacts and noise when the size of the latent space is increased. The PCA-FD scores, on the other hand, suggest that both VAE models capture more variability across the data and better match the overall global distribution of population-level variance. However, the β-VAE appears to successfully capture variability without reconstructing artifacts due to the explicit regularization that we utilized in training. These results provide initial evidence that regularization, in this case with a β-VAE model, is helpful for striking a balance between denoising, representing fine scale structures, and capturing the data's global distributional properties.

3.3 Interpreting the latent factors

After confirming that our model can generate high quality images and denoise data, we next explored its interpretability with the bi-directional analysis method described in Section 2.2 (Figure 1). We first examined the *projective field* of each latent factor. In this case, our goal was to produce three heatmaps to reveal which parts of the image space are impacted by changing a specific latent factor with either a (i) a small negative, (ii) small positive, or (iii) a large interpolation weight. Sorting the interpolations in this way allows us to generate three images that can be stacked into different channels of a color image to visualize the impact of all three types of perturbations on the image domain jointly (Figure 3A). Upon further inspection of the images that resulted from this analysis (Supp. Material S4), we observed that localized noise artifacts (type i) were synthesized at the extrema of the interpolation space. Interestingly, we observed asymmetries in the representations: Type (i) artifacts, while not usually recovered by the decoder, were more likely recapitulated when moving far into the space of negative interpolation weights (Supp. Material S4). In contrast, small interpolation weights appeared to highlight biologically meaningful variance that aligns with key ROIs including the barrel fields of somatosensory cortex, hippocampus, and retrosplenial areas in cortex. These results provide initial evidence that VAE models can be used to decompose biological variability in complex data, even in the presence of different types of noise and artifacts.

We next asked whether we could understand properties about each unit's *receptive field*. To do so, we selected a set of high-quality images without obvious artifacts, applied masks to remove all content from different ROIs, and then modulated their intensity with perturbation weights w. We fed these perturbed images into the encoder (Supp. Material S2), computed the latent representations, and fit a Gaussian to the resulting latent codes across all image examples ($n = 832$) (Figure 3C, Supp. Materials S3). The results of this perturbation analysis revealed multiple units that are strongly modulated by changes in some brain regions but not others, and that exhibit localized receptive fields. We found that perturbations to the hippocampus (HPF) impacted almost all of the latent variables, and striatum also has wide reaching impacts. This seems to align with the fact that variability in these areas is more complex and thus it is necessary to encode this variance over multiple factors.

The impact of perturbing a specific ROI on a latent factor could be further quantified by computing the KL-divergence between the activation distributions for two extreme perturbations (strong negative or positive scaling of missing data in ROI). We computed this *impact score* for all 6 brain ROIs and all 8 latent factors in the trained network (displayed as a 6x8 matrix in Figure 3B, further visualized in Figure 3C). This matrix quantifies the impact that missing information from a ROI has on activations in each latent variable in the model. One interesting result from our analysis is that, in some cases, the receptive field and projective field may not be spatially aligned (see Factor 8, HPF). Our results reveal that receptive and projective fields can be asymmetric, and thus it

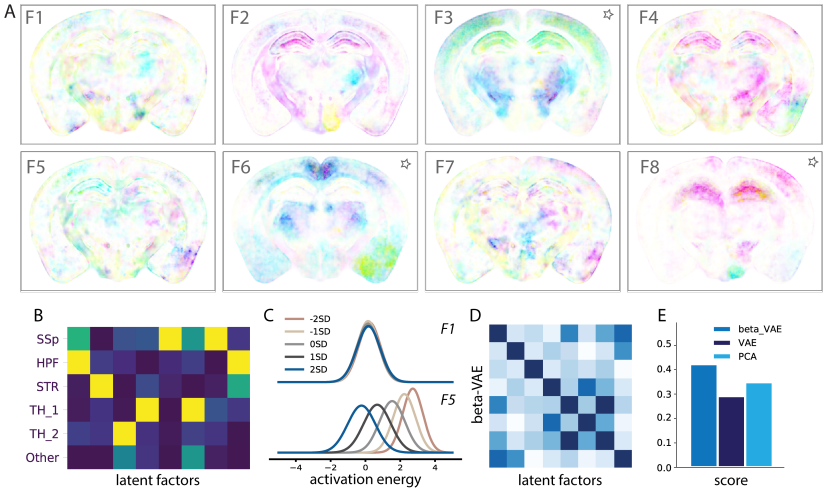

Fig. 3: *Model interpretation.* **(A)** A visualization of how changing a latent factor impacts the generated output images. Here, cyan and magenta represent the pixel level variance in images generated from interpolation weights in the quartile above and below average, respectively, and yellow represents pixels that vary with high interpolation weights. For all factors, the interpolation weight is varied from $[-7, 7]$ with a step size of 0.005. **(B)** For each ROI, we compute the KL-divergence between each factor's response to extreme ROI specific perturbations (blue is low impact, yellow is high impact). **(C)** We show how perturbing the image brightness in HPF region impacts the activation distribution for two factors (F1, F5). **(D)** The covariance of the impact matrix in (B) measures the similarity between how different latent variables impact specific ROIs. **(E)** The disentanglement score for PCA, the VAE, and β-VAE provide a measure of how uncorrelated factors are in terms of their impact on specific brain regions.

is critical to map input-output relationships from the image space to the latent space and back again.

To quantify the separability or *disentanglement* of a model's latent space relative to known brain structures, we examined whether regional perturbations impact different factors in unique ways. We thus computed the covariance between each latent factor's impact scores to reveal their similarity and defined the disentanglement score s as a measure of how far this matrix is from diagonal, where $s = Tr(\mathbf{A})/(\sum_{ij} \mathbf{A}_{ij} - Tr(\mathbf{A}))$ and $\mathbf{A}$ denotes the covariance matrix of interest. A comparison between the β-VAE, VAE, and PCA in terms of their scores revealed that the β-VAE achieved the best disentanglement among three models (Figure 3E). This provides evidence that the β-VAE model can capture variance across a few key brain areas while also providing good separation across different latent factors. In contrast, the vanilla-VAE appears to have factors with much lower disentanglement. PCA on the other hand, provides better disentanglement due to its orthogonality constraints but still doesn't separate brain areas as well as the β-VAE model.

4 Discussion

This work presents a novel data-driven approach for learning population-level differences across high-resolution microscopy images collected from many individuals. Our key contribution is a new method for interpreting factors that drive variance in a deep generative model for brain image synthesis.

In our current study, we used a β-VAE model because of its simplicity and flexibility; however, there are other interpretable VAE variants that have been proposed to facilitate disentanglement [3,21,4] that we could apply our approach to. As our interpretability approach is quite general, one could also potentially use it to visualize and interpret latent representations and/or biomarkers found in other instances of representation learning in neuroscience [2] and medical imaging [15]. Moving forward, interpretability approaches that can probe and model collective responses across many units will be important for revealing complex interactions between features, as well as inspiring new approaches for modeling variability in large high-dimensional datasets.

Acknowledgements: This work was supported by NSF award number IIS-1755871 (ELD, AHB), an Alfred P. Sloan Fellowship (ELD, RL), and NIH Award No. 1R24MH114799-01 (ELD).

References

1. Allen Institute for Brain Science. Allen Mouse Brain Connectivity Atlas. Available from: `connectivity.brain-map.org` (2011)
2. Balwani, A.H., Dyer, E.L.: A deep feature learning approach for mapping the brain's microarchitecture and organization. bioRxiv (2020)
3. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in beta-vae. arXiv preprint arXiv:1804.03599 (2018)
4. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. pp. 2610–2620 (2018)
5. DuPre, E., Spreng, R.N.: Structural covariance networks across the life span, from 6 to 94 years of age. Network Neuroscience **1**(3), 302–323 (2017)
6. Hafkemeijer, A., Möller, C., Dopper, E.G., Jiskoot, L.C., van den Berg-Huysmans, A.A., van Swieten, J.C., van der Flier, W.M., Vrenken, H., Pijnenburg, Y.A., Barkhof, F., Scheltens, P.: Differences in structural covariance brain networks between behavioral variant frontotemporal dementia and alzheimer's disease. Human Brain Mapping **37**(3), 978–988 (2016)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
8. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: Beta-vae: learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations **2**(5), 6 (2017)
9. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)

10. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
11. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: International Conference on Machine Learning. pp. 4114–4124 (2019)
12. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are gans created equal? a large-scale study. In: Advances in Neural Information Processing Systems. pp. 700–709 (2018)
13. Oh, S.W., Harris, J.A., Ng, L., Winslow, B., Cain, N., Mihalas, S., Wang, Q., Lau, C., Kuan, L., Henry, A.M., Mortrud, M.T.: A mesoscale connectome of the mouse brain. Nature **508**(7495), 207 (2014)
14. Pagani, M., Bifone, A., Gozzi, A.: Structural covariance networks in the mouse brain. NeuroImage **129**, 55–63 (2016)
15. Prescott, J.W.: Quantitative imaging biomarkers: the application of advanced image processing and analysis to clinical and preclinical decision making. Journal of Digital Imaging **26**(1), 97–108 (2013)
16. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
17. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: International Conference on Machine Learning. pp. 1278–1286 (2014)
18. Rifai, S., Vincent, P., Muller, X., Glorot, X., Bengio, Y.: Contractive auto-encoders: explicit invariance during feature extraction. International Conference on Machine Learning pp. 833–840 (2011)
19. Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K.: Complex wavelet structural similarity: a new image similarity index. IEEE Transactions on Image Processing **18**(11), 2385–2401 (2009)
20. Vandenberghe, M.E., Hérard, A.S., Souedet, N., Sadouni, E., Santin, M.D., Briet, D., Carré, D., Schulz, J., Hantraye, P., Chabrier, P.E., et al.: High-throughput 3d whole-brain quantitative histopathology in rodents. Scientific Reports **6**, 20958 (2016)
21. Zhao, S., Song, J., Ermon, S.: Infovae: balancing learning and inference in variational autoencoders. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5885–5892 (2019)