Structured Priors for Disentangling Pathology and Anatomy in Patient Brain MRI

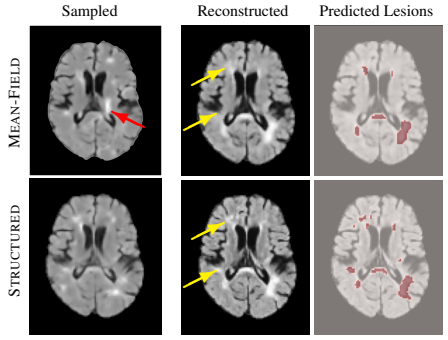
Anjun Hu[1,2] **Jean-Pierre R. Falet**[1,2,3] **Changjian Shui**[1,2] **Brennan S. Nichyporuk**[1,2]
Douglas L. Arnold[3] **Sotirios A. Tsaftaris**[4,5] **Tal Arbel**[1,2]

[1] Center for Intelligent Machines, McGill University, Montréal, Canada
[2] Mila (Québec Artificial Intelligence Institute), Montréal, Canada
[3] Brain Imaging Center, Dept. Neurology and Neurosurgery, McGill University, Montréal, Canada
[4] School of Engineering, University of Edinburgh, UK
[5] The Alan Turing Institute, UK
{anjun,jpfalet,maxshui,arbel}@cim.mcgill.ca,
nichypob@mila.quebec, douglas.arnold@mcgill.ca, s.tsaftaris@ed.ac.uk

Abstract

We propose a structured variational inference model for disentangling observable evidence of disease (e.g. brain lesions or atrophy) from subject-specific anatomy in brain MRIs. With flexible, partially autoregressive priors, our model (1) addresses the subtle and detailed dependencies that typically exist between anatomical and pathological generating factors of an MRI to ensure the validity of generated samples; (2) preserves and disentangles finer pathological details pertaining to a patient's disease state. We additionally demonstrate that, by providing supervision to a subset of latent units, that: (1) a partially supervised latent space achieves a higher degree of disentanglement between evidence of disease and subject-specific anatomy; (2) when the prior is formulated with an autoregressive structure, knowledge from the supervision can propagate to the unsupervised latent units, resulting in more informative latent representations capable of modelling anatomy-pathology interdependencies.

1 Introduction and Background



(a) (b)

Figure 1: (a) A sample drawn from a mean-field model with lesions in clinically invalid locations (red arrow). (b) Mean-field model leads to missed small lesions in the reconstructed image (yellow arrows). Our proposed structured model does not suffer from these issues.

In the context of magnetic resonance imaging (MRI) analysis, various methods have been introduced to disentangle particular structures from the rest of the image under different observations [1–10]. Many of such methods are based on the variational auto-encoder (VAE) [11–13]. However, in brain MRI, observable pathological features (e.g. hyper-intense MS lesions) cannot be easily disentangled from subject-specific anatomical structures (e.g. sulcal pattern, ventricular shape) due to dependencies that exist between the anatomical and pathological generative factors. A straight adoption of VAE for pathology-anatomy disentanglement in patient brain MRIs is, therefore, often unsatisfactory for several reasons. Firstly, mean-field variational inference poses the unlikely assumption that all generating factors are independent [14], thereby

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

failing to capture the dependencies that exist between the anatomical and pathological generative factors. For example, in the case of multiple sclerosis (MS), a chronic neurological disease, T2 hyperintense lesions in the brain are typically located in certain regions and cannot be found in others (e.g. within the ventricles). Good representations in this context must be capable of modelling the spatial distribution of the lesions and their dependencies on surrounding anatomical structures to avoid assigning higher likelihoods to clinically invalid samples, as depicted in Figure 1a. Secondly, VAEs tend to suffer from lower synthesis quality [14, 15]. In cases such as MS, where lesions as small as 3 mm long still represent a significant marker of disease activity [16], failure to capture fine details in the learned representation could lead to significantly poorer performance in downstream tasks (Figure 1b).

In this work, we propose to address these issues by using structured variational inference [17] for fine-grained pathology-anatomy disentanglement in brain MRI. Specifically we model the dependencies that typically exist between pathological and anatomical features via multi-scale VAEs with a hierarchical latent structure (Figure 2). We find that more expressive structured priors indeed lead to higher reconstruction quality and the preservation of important small pathological details. Moreover, with an additional optional supervision objective, the model is shown to achieve high-quality pathology-anatomy disentanglement and to be capable of capturing latent dependency.

2 Methods

Four different parametrisations of a multi-scale VAE with spatial latent variables are evaluated:

(1) VAE: a vanilla multi-scale VAE with mean-field priors at each layer. This model has a "hierarchy" in the sense of having latent representations at various resolution scales. However, it does not have explicit inter-layer dependencies in the prior formulation: all latent variables at each layer of the model are subject to a parameter-free, standard Gaussian prior.

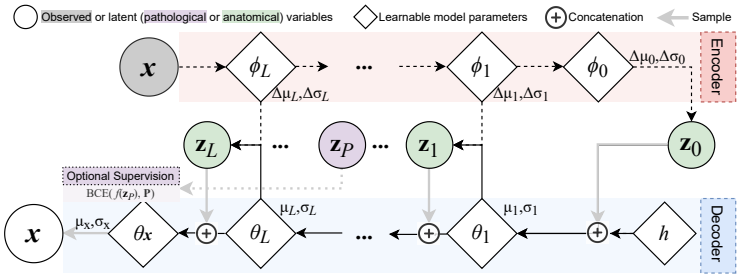


Figure 2: The structured latent space consists of $(L+1)$ disjoint variable groups (layers) that follow a hierarchical structure. Dashed lines are active during inference. Solid lines are active during generation and, if residual parameterisation [18] is used, also during bidirectional inference.

(2) NVAE [18]: a multi-scale VAE with a partially autoregressive prior and hierarchical residual parameterisation for the posterior. More precisely: (i) The prior of the topmost layer $\mathbf{z}_0$ is still a standard Gaussian but the priors of subsequent layers $p_\theta(\mathbf{z}_l|\mathbf{z}_{<l}), l > 0$ are *moving Gaussians* that explicitly depend on the priors of previous layers. (ii) Distributional parameters obtained from the encoder, $(\Delta\mu_l, \Delta\sigma_l)$, are not directly interpreted as posteriors. Instead, they are combined with distributional parameters obtained from preceding layers of the decoder, (μ_l, σ_l), to form "*residual posteriors*" $q_{\phi,\theta}(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x})$ that characterise deviations from the moving priors at each layer. This implies that the model requires *bidirectional inference* since obtaining the residual posterior distribution at layer l involves a forward pass through the decoder up to θ_l.

(3) NVMP: our novel extension of NVAE where we replace the standard-Gaussian prior of the topmost layer of NVAE with a K-component multimodal VamPrior [19] $\frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}_l|\mathbf{u}_k)$ characterized by trainable pseudo-inputs $\mathbf{u}$ and encoder parameters ϕ. The subsequent layers still adhere to the hierarchical residual parameterisation like NVAE. The implication is that only $\mathbf{z}_0$ is multi-modal whereas the lower level "deviations" are assumed to be Gaussians.

(4) NVMP+: a variant of NVMP where we impose an additional KL term between the encoder-driven VamPriors and the decoder priors throughout the latent hierarchy, which means every layer in the hierarchy would enjoy the flexibility of a multi-modal distribution.

Further details on parameterisation and implementation can be found in Appendices A, B.

3 Experiments and Results

We validate our approach on two brain MRI datasets: the publically available Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [20] ($N = 864$), and a proprietary MS dataset from a MS clinical trial ($N = 815$). The central 16 2-D slices of T1-weighted sequences were used for the AD experiments, while the central 24 2-D slices of Fluid Attenuated Inverse Recovery (FLAIR) sequence were used for the MS experiments. Ground-truth T2 lesion labels for the MS experiments were provided. Both datasets were divided into non-overlapping training (60 %), validation (20 %) and testing (20 %) sets. Additional acquisition and pre-processing details are described in Appendix B.

We first train the model under an unsurpervised setting and evaluate the effect of incorporating additional prior structures on synthesis quality. As shown in Table 1, VAEs with more expressive structured priors indeed outperform their mean-field counterpart at the same model capacity in terms of image reconstruction fidelity.

We additionally examine model behaviours in a supervised learning setting depicted in the bottom-left (purple) block in Figure 2 and Figure 5. In this setting, we supplement the MS model with a lesion segmentation objective between a chosen "pathological" latent subset $\mathbf{z}_P$ and ground-truth pathology (lesion) labels $\mathbf{P}$. The rest of the latent space (that remains unsupervised) are regarded as the anatomical latent subsets, denoted as $\mathbf{z}_A$.

Firstly, as one might expect, supervision is shown to enhance latent disentanglement as one may anticipate. Disease-related features in the synthesized images are noticeably more *sensitive* [21] to perturbations in $\mathbf{z}_P$ comapred to $\mathbf{z}_A$, as shown in Figure 6c, Appendix C. Such a disparity in attribute sensitivity is appreciable in unsupervised models, but is made much more pronounced by the selective latent supervision.

Secondly and more importantly, *supervision helps to verify that the model is indeed actively using the latent structures.* In models with autoregressive structures (NVAE, NVMP, NVMP+), knowledge from the supervision is propagated to the unsupervised "anatomical" latent units $\mathbf{z}_A$, as in, those unsupervised latent units attain a higher linear predictability (Lasso regression R^2 scores [22], Table 4) with respect to lesion volume. This is in contrast to the behaviour of the baseline mean-field VAE, where information from the supervision task is constrained within the supervised group $\mathbf{z}_P$. This observation shows that the model is indeed taking advantage of the extra structures brought by the autoregressive priors and the residual parameterisation and hence, indeed capable of modelling the dependencies between anatomical and pathological generating factors.

Data	Model		NLL↓	PSNR↑	SSIM↑	FID ↓
MS	VAE	(M.1) [11]	2758	25.1	0.72	0.058
	NVAE	(M.2) [18]	2458	25.7	0.75	**0.023**
	NVMP	(M.3) [19]	2374	25.9	0.75	0.031
	NVMP+	(M.4) [19]	**1953**	**26.6**	**0.79**	0.035
AD	VAE	(M.1) [11]	2386	24.6	0.70	0.030
	NVAE	(M.2) [18]	2105	25.2	0.73	0.013
	NVMP	(M.3) [19]	1863	25.5	0.75	0.011
	NVMP+	(M.4) [19]	**842**	**26.8**	**0.80**	**0.007**

Table 1: Reconstruction quality metrics.

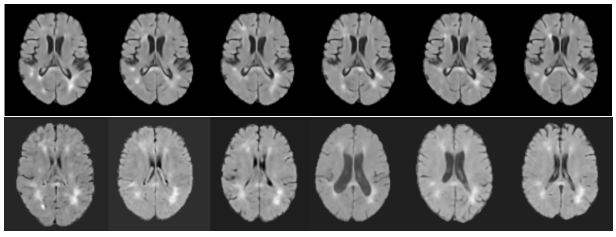


Figure 3: Conditional synthesis

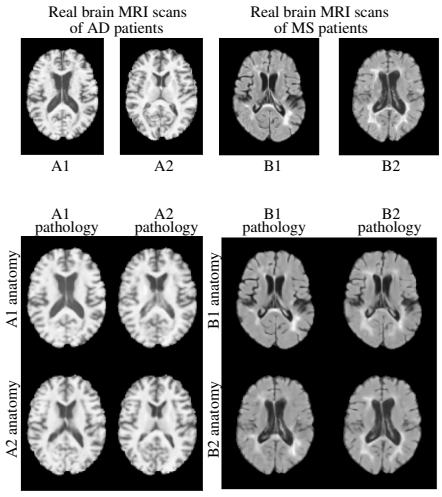


Figure 4: Pathology style-mixing

We qualitatively evaluate pathology-anatomy disentanglement by swapping anatomical and pathological latent features between a pair of subjects in a manner similar to "style-mixing" [23]. As shown in Figure 4 for representative examples, brain atrophy in AD patients (left), and T2 lesions in MS patients (right), are disentangled from the subject's anatomical particularities (such as sulcal pattern), thus enabling the mixing the pathology of one patient with the anatomy of the other. We may

also leverage conditional distributions learned by the model to examine subject-specific pathology distributions. For example, based on learned representations of Subject B1 in Figure 4, we may visualise many possible disease states given this subject's anatomy (top row images in Figure 3 are generated by fixing $\mathbf{z}_A$ to that of Subject B1 and resampling $\mathbf{z}_P$ conditioned on $\mathbf{z}_A^{B1}$) or explore how this subject's lesions would manifest on other subjects' brain anatomies (bottom row images are generated by combining $\mathbf{z}_A$ obtained from other real samples with fixed $\mathbf{z}_P^{B1}$.).

4 Conclusions

We propose hierarchical VAEs with structured priors for learning pathology-anatomy disentangled representations of brain MRIs. Our model can faithfully capture imaging features, including fine-grained details, while accounting for pathology-anatomy dependencies to ensure sample validity. We additionally examine model hevaviours in a supervised learning setting. Supervision is shown to (1) further enhance latent disentanglement; and (2) enable the inspection of information propagation between latent groups for modelling pathology-anatomy interdependencies. Our model allows for robust and controllable brain MRI synthesis rich in high-frequency and pathologically-sound details, which could be meaningful for various downstream tasks.

5 Potential Negative Societal Impact

We propose a deep learning framework for learning pathology-anatomy disentangled representations of brain MRIs, which could facilitate the understanding of relevant diseases. Simultaneously, since the dataset is not large-scale, it could not be representative of the true population and exhibits potential bias towards certain demographics.

Acknowledgments and Disclosure of Funding

The authors are grateful to the International Progressive MS Alliance for supporting this work (grant number: PA-1412-02420), and to the companies who generously provided the clinical trial data that made it possible: Biogen, BioMS, MedDay, Novartis, Roche / Genentech, and Teva. Funding was also provided by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, and a technology transfer grant from Mila - Quebec AI Institute. S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819 / 8 / 25). Supplementary computational resources and technical support were provided by Calcul Québec and the Digital Research Alliance of Canada. Falet, J.-P., was supported by an end MS Personnel Award from the Multiple Sclerosis Society of Canada, by a Canada Graduate Scholarship-Masters Award from the Canadian Institutes of Health Research, and by the Fonds de recherche du Québec - Santé / Ministère de la Santé et des Services sociaux training program for specialty medicine residents with an interest in pursuing a research career, Phase 1. This work was made possible by the end-to-end deep learning experimental pipeline developed in collaboration with our colleagues Justin Szeto, Eric Zimmerman, and Kirill Vasilevski. Additionally, the authors would like to thank Louis Collins and Mahsa Dadar for preprocessing the MRI data.

References

[1] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q O'Neil, and Sotirios A Tsaftaris. A tutorial on learning disentangled representations in the imaging domain. *arXiv preprint arXiv:2108.12043*, 2021.

[2] Jana Fragemann, Lynton Ardizzone, Jan Egger, and Jens Kleesiek. Review of disentanglement approaches for medical applications–towards solving the gordian knot of generative models in healthcare. *arXiv preprint arXiv:2203.11132*, 2022.

[3] Dan Hu, Han Zhang, Zhengwang Wu, Fan Wang, Li Wang, J. Keith Smith, Weili Lin, Gang Li, and Dinggang Shen. Disentangled-multimodal adversarial autoencoder: Application to

infant age prediction with incomplete multimodal neuroimages. *IEEE Transactions on Medical Imaging*, 39(12):4137–4149, 2020. doi: 10.1109/TMI.2020.3013825.

[4] Qiushi Yang, Xiaoqing Guo, Zhen Chen, Peter Y. M. Woo, and Yixuan Yuan. D2-net: Dual disentanglement network for brain tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022. doi: 10.1109/TMI.2022.3175478.

[5] Lei Zhou, Joseph Bae, Huidong Liu, Gagandeep Singh, Jeremy Green, Dimitris Samaras, and Prateek Prasanna. Chest radiograph disentanglement for covid-19 outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 345–355. Springer, 2021.

[6] Chen Qin, Bibo Shi, Rui Liao, Tommaso Mansi, Daniel Rueckert, and Ali Kamen. Unsupervised deformable registration for multi-modal images via disentangled representations. In *International Conference on Information Processing in Medical Imaging*, pages 249–261. Springer, 2019.

[7] Qingsong Xie, Yuexiang Li, Nanjun He, Munan Ning, Kai Ma, Guoxing Wang, Yong Lian, and Yefeng Zheng. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. *IEEE Transactions on Medical Imaging*, pages 1–1, 2022. doi: 10.1109/TMI.2022.3192303.

[8] Lianrui Zuo, Blake E Dewey, Aaron Carass, Yihao Liu, Yufan He, Peter A Calabresi, and Jerry L Prince. Information-based disentangled representation learning for unsupervised mr harmonization. In *International Conference on Information Processing in Medical Imaging*, pages 346–359. Springer, 2021.

[9] Xiaofeng Liu, Fangxu Xing, Georges El Fakhri, and Jonghye Woo. A unified conditional disentanglement framework for multimodal brain mr image translation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 10–14. IEEE, 2021.

[10] Xiao Liu, Spyridon Thermos, Alison O'Neil, and Sotirios A Tsaftaris. Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 307–317. Springer, 2021.

[11] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

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

[13] 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. 2016.

[14] C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(08):2008–2026, aug 2019. ISSN 1939-3539. doi: 10.1109/TPAMI.2018.2889774.

[15] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P. Xing. On unifying deep generative models. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rylSzl-R-`.

[16] Liqun Wang, H-M Lai, A J Thompson, and D H Miller. Survey of the distribution of lesion size in multiple sclerosis: implication for the measurement of total lesion load. *Journal of Neurology, Neurosurgery & Psychiatry*, 63(4):452–455, 1997. ISSN 0022-3050. doi: 10.1136/jnnp.63.4.452. URL `https://jnnp.bmj.com/content/63/4/452`.

[17] Matthew Hoffman and David Blei. Stochastic Structured Variational Inference. In Guy Lebanon and S. V. N. Vishwanathan, editors, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 361–369, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/hoffman15.html.

[18] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e3b21256183cf7c2c7a66be163579d37-Paper.pdf.

[19] Jakub M. Tomczak and Max Welling. VAE with a vampprior. *CoRR*, abs/1705.07120, 2017. URL http://arxiv.org/abs/1705.07120.

[20] Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005.

[21] Michal Rolínek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). *CoRR*, abs/1812.06775, 2018. URL http://arxiv.org/abs/1812.06775.

[22] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018. URL https://arxiv.org/abs/1812.04948.

[24] Suhang You, Kerem C. Tezcan, Xiaoran Chen, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 540–556. PMLR, 08–10 Jul 2019. URL https://proceedings.mlr.press/v102/you19a.html.

[25] Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. *CoRR*, abs/1507.02672, 2015. URL http://arxiv.org/abs/1507.02672.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of NAACL-HLT*, pages 240–250, 2019.

[28] Arash Vahdat, William Macready, Zhengbing Bian, Amir Khoshaman, and Evgeny Andriyash. Dvae++: Discrete variational autoencoders with overlapping transformations. In *International conference on machine learning*, pages 5035–5044. PMLR, 2018.

[29] Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, 2016.

A Model and Objective Formulation Details

In this work, we examine four ways to construct the ELBO objective for multi-level VAEs with spatial latent variables [24], shown in Table 3. Each model consists of an inference model and a generative model. A top-down decoder likelihood $p_\theta(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$ generates observations $\mathbf{x}$ with samples drawn from the prior $p(\mathbf{z})$ defined over latent variables $\mathbf{z}$. The inference model, on the other hand, is a bottom-up encoder that approximates the posterior distribution $q_{\phi,\theta}(\mathbf{z}|\mathbf{x})$ to allow for tractable optimization through the maximization of the Evidence Lower BOund (ELBO).

The model accepts image space observations $\mathbf{x} \in \mathbb{R}^{w_\mathbf{x} \times h_\mathbf{x}}$ drawn from a dataset $\mathbf{x} \sim \mathcal{D}$ as inputs. We use $L+1$ disjoint groups of spatial latent variables [24] at various resolutions scales, denoted as $\mathbf{z} = \{\mathbf{z}_l \in \mathbb{R}^{w_l \times h_l}; l \in \{0, 1, ..., L\}\}$. The inference model and the generative model are joint distributions as follows:

$$
\begin{aligned}
q_{\phi,\theta}(\mathbf{z}|\mathbf{x}) &:= q_{\phi_L,\theta_L}(\mathbf{z}_L|\mathbf{x}) \prod_{l=0}^{L-1} q_{\phi_l,\theta_l}(\mathbf{z}_l|\mathbf{z}_{l+1}) \\
p_\theta(\mathbf{x}, \mathbf{z}) &:= p_{\theta_x}(\mathbf{x}|\mathbf{z}_L)p(\mathbf{z}_0) \prod_{l=1}^{L} p_{\theta_l}(\mathbf{z}_l|\mathbf{z}_{l-1})
\end{aligned}
\tag{1}
$$

Layer	Model	Inference	Generation
$\mathbf{z}_{l=0}$	VAE (M.1) [11] NVAE (M.2) [18] NVMP (M.3) [19] NVMP+ (M.4) [19]	$q_\phi(\mathbf{z}_0) := \mathcal{N}(\Delta\mu_0(\mathbf{x}), \Delta\sigma_0(\mathbf{x}))$	$p(\mathbf{z}_0) := \mathcal{N}(\mathbf{0}, \mathbf{I})$ $p_{\phi,\mathbf{u}}(\mathbf{z}_0) := \frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}_0|\mathbf{u}_k)$
$\mathbf{z}_{l>0}$	VAE (M.1) [11]	$q_{\phi_l}(\mathbf{z}_l) := \mathcal{N}(\Delta\mu_{\phi_{>l}}(\mathbf{x}), \Delta\sigma_{\phi_{>l}}(\mathbf{x}))$	$p_{\theta_l}(\mathbf{z}_l) := \mathcal{N}(\mathbf{0}, \mathbf{I})$
	NVAE (M.2) [18] NVMP (M.3) [19] NVMP+ (M.4) [19]	$q_{\phi,\theta}(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x}) := \mathcal{N}\Big(\mu_\theta(\mathbf{z}_{<l}) + \Delta\mu_\phi(\mathbf{x}),$ $\sigma_\theta(\mathbf{z}_{<l}) \cdot \Delta\sigma_\phi(\mathbf{x})\Big)$	$p_\theta(\mathbf{z}_l|\mathbf{z}_{<l}) :=$ $\mathcal{N}(\mu_{\theta_{<l}}(\mathbf{z}_{<l}), \sigma_{\theta_{<l}}(\mathbf{z}_{<l}))$

Table 2: Model Parameterisations

Model	ELBO
VAE (M.1)	$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \sum_{l=0}^{L} \mathrm{KLD}[q_\phi(\mathbf{z}_l|\mathbf{x})||\mathcal{N}(\mathbf{0}, \mathbf{I})]\Big]$
NVAE (M.2)	$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Big[\mathbb{E}_{\mathbf{z}\sim q_{\phi,\theta}(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KLD}[q_\phi(\mathbf{z}_0|\mathbf{x})||\mathcal{N}(\mathbf{0}, \mathbf{I})]$ $- \sum_{l=1}^{L} \mathbb{E}_{q_{\phi,\theta}(\mathbf{z}_{<l}|\mathbf{x})} \mathrm{KLD}[q_{\phi,\theta}(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x})||p_\theta(\mathbf{z}_l|\mathbf{z}_{<l})]\Big]$
NVMP (M.3)	$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Bigg[\mathbb{E}_{\mathbf{z}\sim q_{\phi,\theta}(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KLD}\big[q_\phi(\mathbf{z}_0|\mathbf{x})||\frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}_0|\mathbf{u}_k)\big]$ $- \sum_{l=1}^{L} \mathbb{E}_{q_{\phi,\theta}(\mathbf{z}_{<l}|\mathbf{x})} \mathrm{KLD}[q_{\phi,\theta}(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x})||p_\theta(\mathbf{z}_l|\mathbf{z}_{<l})]\Bigg]$
NVMP+ (M.4)	$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\Bigg[\mathbb{E}_{\mathbf{z}\sim q_{\phi,\theta}(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KLD}\big[q_\phi(\mathbf{z}_0|\mathbf{x})||\frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}_0|\mathbf{u}_k)\big]$ $- \sum_{l=1}^{L} \mathbb{E}_{q_{\phi,\theta}(\mathbf{z}_{<l}|\mathbf{x})} \Big[\mathrm{KLD}[q_{\phi,\theta}(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x})||p_\theta(\mathbf{z}_l|\mathbf{z}_{<l})] + \mathrm{KLD}[p_\theta(\mathbf{z}_l|\mathbf{z}_{<l})||\frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}_l|\mathbf{u}_k)]\Big]\Bigg]$

Table 3: ELBOs for all four model parameterisation options examined in this work.

A.1 Vanilla Multi-Scale VAE

The most straightforward baseline (M.1) is a "vanilla multi-scale VAE". With this parameterisation, the priors for each latent group is a parameter-free standard Gaussian (2b). Although there is a "hierarchy" in the sense of various resolution scales of the latent spaces, this model is not "hierarchical" in its distributional parametrisation. There is no explicit inter-group dependency in the prior formulation nor explicit information sharing between the encoder and the decoder. We have $\forall l \in \{0, 1, ..., L\}$:

$$
q_{\phi_l}(\mathbf{z}_l) := \mathcal{N}(\Delta\mu_{\phi_{>l}}(\mathbf{x}), \Delta\sigma_{\phi_{>l}}(\mathbf{x})) \tag{2a}
$$

$$
p_{\theta_l}(\mathbf{z}_l) := \mathcal{N}(\mathbf{0}, \mathbf{I}) \;\; \forall l \in \{0, 1, ..., L\} \tag{2b}
$$

A.2 NVAE

NVAE (M.2) is a hierarchical model with residual normal parameterisation proposed by [18] and [25]. The features that set this model apart from its vanilla counterpart are the explicit information sharing between the encoder and the decoder networks, as well as its partially auto-regressive nature.

Firstly, unlike conventional VAEs, decoder parameters θ in NVAE not only characterize the generative distribution p_θ (3b) but are also a part of the inference model and hence play an important role in characterizing the posterior distribution $q_{\phi,\theta}$ (3a). For latent groups other than the topmost one $l > 0$, the inference model is *bidirectional*. It estimates the *relative variational* posteriors (3a) that characterize the deviation from priors obtained from preceeding layers of the decoder. With this design, KL optimization is expected to be simpler than when posteriors predict the absolute mean and variances at each layer.

$$q_{\phi,\theta}(\mathbf{z}_l|\mathbf{z}_{<l}, \mathbf{x}) := \mathcal{N}\Big(\mu_{\theta_{\leq l}}(\mathbf{z}_{<l}) + \Delta\mu_{\phi_{\geq l}}(\mathbf{x}), \sigma_{\theta_{\leq l}}(\mathbf{z}_{<l}) \cdot \Delta\sigma_{\phi_{\geq l}}(\mathbf{x})\Big) \tag{3a}$$

$$p_\theta(\mathbf{z}_l|\mathbf{z}_{<l}) := \mathcal{N}(\mu_{\theta_{\leq l}}(\mathbf{z}_{<l}), \sigma_{\theta_{\leq l}}(\mathbf{z}_{<l})) \tag{3b}$$

Secondly, NVAE is considered to be partially auto-regressive and hence a more expressive prior than the standard mean-field parametrization. While the prior for each group $p(\mathbf{z}_l)$ is dependent on those of the preceding layers $p(\mathbf{z}_{<l})$, each element within the same latent group $\mathbf{z}_l := \{z_l^i, ...\}$ still adhere to the independence assumption as $\forall l \in \{0, 1, ..., L\}, z_l^i \perp\!\!\!\perp z_l^j; i \neq j$. We can hence calculate the relative KLD loss for each element (4) with a simple analytic expression:

$$\mathrm{KLD}[q_{\phi,\theta}(z_l^i|\mathbf{x})||p_\theta(z_l^i)] = \frac{1}{2}\left(\frac{(\Delta\mu_l^i)^2}{(\sigma_l^i)^2} + (\Delta\sigma_l^i)^2 - \log(\sigma_l^i)^2 - 1\right) \tag{4}$$

A.3 NVMP

We propose two extensions to NVAE by incorporating a VamPrior [19] into the hierarchical VAE setup for extra flexibility in the hierarchical latent structure. We refer to them as NVMP (M.3) and NVMP+ (M.4). VamPrior and LVAE share the same philosophy that coupling the prior with the posterior would ease the training by fostering better "collaboration" between the prior and variational posterior despite the seemingly opposite approaches taken by the two works (VamPrior incorporates encoder parameters in the trainable prior whereas LVAE decoder parameters are involved in the formulation of variational posteriors). This similarity motivates us to combine the two approaches to achieve a greater extent of "communication" between the priors and the posteriors.

In NVAE or Equation (M.3), we replace the standard Gaussian prior for the topmost latent group with a VamPrior, $\frac{1}{K}\sum_{k=1}^{K} q_\phi(\mathbf{z}_l|\mathbf{u}_k)$, that is, a K-component multi-modal distribution characterized by trainable pseudo-inputs $\mathbf{u}$ as well as the encoder parameters ϕ. We retain the residual Gaussian parameterisation for subsequent layers ($\mathbf{z}_{>0}$).

In NVAE+ or Equation (M.4), one more KL term between the encoder-driven VamPriors and the decoder priors is imposed for the entire hierarchy. In this case, the decoder "priors" are regarded as "intermediate posteriors" and encouraged to imitate the encoder-driven multi-modal distribution throughout the hierarchy. We postulate that this configuration adds an extra layer of information sharing between the encoder and the decoder networks which can potentially lead to further improvement in representation quality.

B Implementation and Training Details

All MRI sequences were acquired at a resolution of 1 mm $\times$ 1 mm $\times$ 1 mm, Each 2-D slice was downsampled to a resolution of 2 mm $\times$ 2 mm. These were standardized to have zero-mean and unit variance.

We compare the four parameterisations in Table 3 with a 5-layer model ($L = 4$) the exact same capacity. For each dataset, the latent space capacity is set to $\{\mathbf{z}_L \in \mathbb{R}^{w_\mathbf{x} \times h_\mathbf{x} \times 2}, \mathbf{z}_{L-1} \in \mathbb{R}^{(w_\mathbf{x}/2)\times(h_\mathbf{x}/2)\times 2}, ..., \mathbf{z}_0 \in \mathbb{R}^{(w_\mathbf{x}/2^L)\times(h_\mathbf{x}/2^L)\times 2}\}$. We use the Adam optimizer [26] with a learning rate of 5e-5 and s weight decay of 1e-8. Two loss re-weighting mechanisms are used in our training

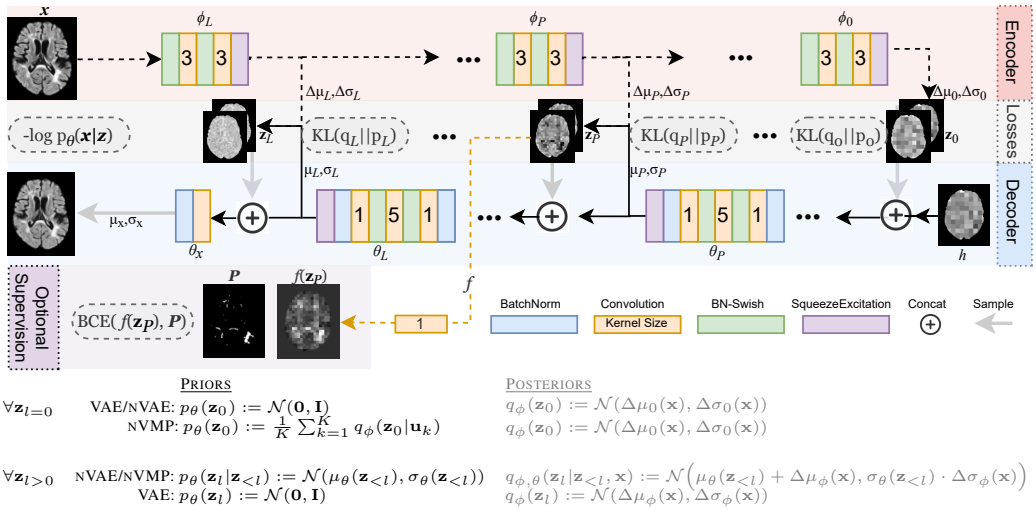


Figure 5: Network Architecture

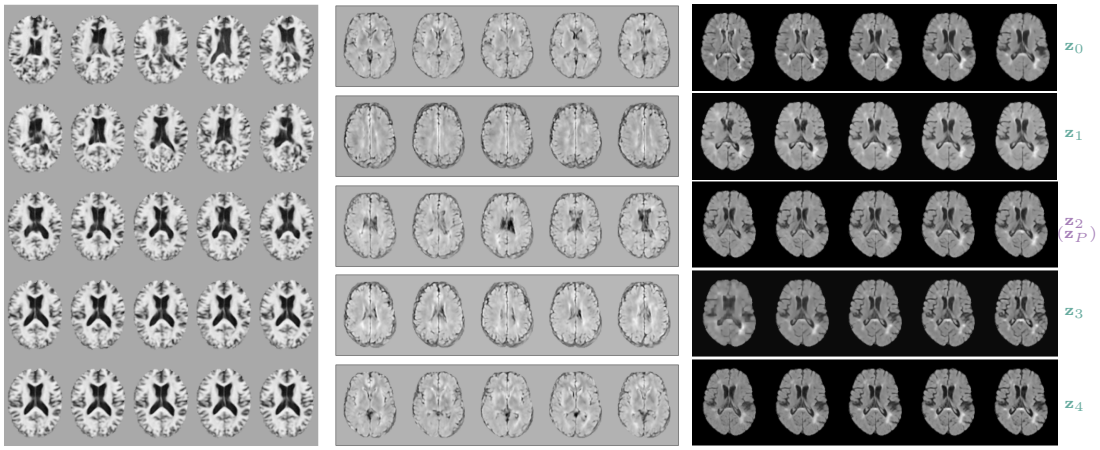
procedure: (1) We use a linear annealing schedule [27] for KLD losses with a cycle length of 10000 iterations. The initial KLD learning rate is set to 2e-7. (2) To avoid posterior collapse, we use a KL Balancing trick suggested by [18]. We re-scale each KL term of the hierarchy with a coefficient proportional to the size of each latent layer as well as the KLD value of that layer. This mechanism encourages more balanced information attribution to each latent layer [28, 29].

C Additional Results

As discussed in 3, we evaluate layer-wise latent pathology informativeness of MS models by examining each layer's linear predictability of a salient pathological attribute, T2 lesion volume. To quantify linear predictability, we train Lasso regressors ($\alpha = 10$) with latent representations obtained from each individual latent layer of each model and compute each Lasso regressors's R^2 scores with respect to ground truth T2 lesion volume. Rows 1-4 show the R^2 scores of the unsupervised models, which are generally poor. Rows 5-8 show the same metrics for supervised models where supervision is provided to $\mathbf{z}_2$ ($\mathbf{z}_P$) as an additional lesion segmentation objective. Models with autoregressive structures (NVAE, NVMP, NVMP+) benefit more from the supervision - knowledge from the supervision is propagated to the unsupervised "anatomical" latent units, resulting in higher R^2 scores even in the unsupervised latent subsets. This shows that the model is indeed actively using the latent structures.

Table 4: (MS) Layer-wise latent informativeness with respect to T2 lesion volume

Model	$\mathbf{z}_0$	$\mathbf{z}_1$	$\mathbf{z}_P$	$\mathbf{z}_3$	$\mathbf{z}_4$
VAE	0.20	0.08	0.28	0.01	0.00
NVAE	0.11	0.12	0.00	0.00	0.00
NVMP	0.06	0.23	0.01	0.01	0.00
NVMP+	0.03	0.00	0.20	0.00	0.00
$\mathbf{z}_P$-supervised VAE	0.00	0.00	0.62	0.08	0.01
$\mathbf{z}_P$-supervised NVAE	0.31	0.20	0.65	0.02	0.00
$\mathbf{z}_P$-supervised NVMP	0.54	0.23	0.63	0.02	0.01
$\mathbf{z}_P$-supervised NVMP+	0.21	0.37	0.56	0.09	0.00



(a) (b) (c)

Figure 6: (a) (AD) Variations captured by each layer of the model. Images at the top row are fully resampled at each level of the hierarchy. On each subsequent row n, we show the residual variation of layer n by fixing latent codes at the top $(n-1)$ layers.
(b) (MS) Clusters discovered by VamPrior.
(c) (MS) Layer-wise pathological attribute sensitivity visualised by individually scaling each layer in the latent hierarchy, from $\mathbf{z}_0$ (top row) to $\mathbf{z}_4$ (bottom row). In this particular example, The appearance of the hyper-intense MS lesions in the synthesised images is relatively insensitive to multiplicative perturbation in all but one latent layer, $\mathbf{z}_2$. The layer with the highest pathological attribute sensitivity, $\mathbf{z}_2$, is hence considered to be a disentangled "pathological" latent subset $\mathbf{z}_P$.
We note that even in the unsupervised setting, disease-related features in the synthesized images are noticeably more sensitive to changes in a small subset of latent variables than the rest, which allows us to identify such a subset as $\mathbf{z}_P$ and the rest as $\mathbf{z}_A$ (anatomical latent subsets) in a post-hoc manner. Such disparity in pathological attribute sensitivity is much more pronounced in the "selective supervision" setting (bottom-left purple block in Figure 2 and Figure 5), where the additional supervision is given to a chosen layer $\mathbf{z}_P$.