# Multi-view variational autoencoders allow for interpretability leveraging digital Avatars: application to the HBN cohort

Corentin Ambroise, Antoine Grigis, Edouard Duchesnay, Vincent Frouin

2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-7358-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISBI53787.2023.10230552

# MULTI-VIEW VARIATIONAL AUTOENCODERS ALLOW FOR INTERPRETABILITY LEVERAGING DIGITAL AVATARS: APPLICATION TO THE HBN COHORT

*Corentin Ambroise, Antoine Grigis, Edouard Duchesnay, Vincent Frouin*

Neurospin, CEA Saclay, University Paris-Saclay, France

**Abstract** – *If neural network-based methods are praised for their prediction performance, they are often criticized for their lack of interpretability. When dealing with multi-omics or multi-modal data, neural network methods must be able learn the independent and joint effect of heterogeneous views while yielding interpretable results intra- and inter-views. In the literature, multi-view generative models exist to learn joint information in a reduced-size latent space. Among these models, multi-view variational autoencoders are very promising. In this work, we demonstrate how they provide a convenient statistical framework to learn the input data joint distribution and offer opportunities for the results interpretation. We design a method that discovers the relationships between one view and others. The generative capabilities of the model enable the exploration of a whole disorder spectrum through the generation of realistic values. While modifying a subject's clinical score, the model retrieves a representation of the subject's brain at this clinical status, so-called digital avatar. By computing associations between cortical regions measures and behavioral scores, we showcase that such digital avatars convey interpretable information in a multi-modal cohort with children experiencing mental health issues.*

## 1. INTRODUCTION

Autism Spectrum conditions (ASD) are recognized to be caused by a combination of genetic and environmental factors, and consequently, its diagnosis, subtyping, or care modalities are difficult to establish. Gaining insights into the disease etiology or generating new research hypotheses are major challenges in neuroscience. Recent approaches propose integrative analyses in population cohorts gathering multimodal data including subjects with and without ASD symptoms. Most current ASD-related multimodal studies usually explore correlates of image-derived features with binary diagnoses like DSVM-IV [1] or scores assessing communication and social skills. Other approaches try to derive subgroups within populations. However, we argue that setting subgroups can never be validated in practice and is often not replicable to other datasets. Multivariate multi-task approaches usually rely on an *a priori* model of variability that involves removing the general variability from the data (e.g., site, sex, age) to focus on the variability of interest. We follow an alternative strategy to study the ASD variability hidden in multi-block data like clinical scores, imaging data, or genotyping either directly (one block) or jointly (several blocks) [2], taking into account both global and subject-specific variabilities.

Recent advances in deep learning have opened up opportunities beyond the models' predictive capacities. These advances help investigate pathology while encompassing the growing amount of medical data. In particular, Artificial Neural Networks (ANN) flexibility make them sensibly adapted to integrate data from different sources and build multi-block multivariate models. Developments of Variational AutoEncoder (VAE) have resulted in frameworks that manage multiple views and model their interactions [3, 4, 5, 6].

We apply a generative VAE-based architecture as an exploration tool on the Healthy Brain Network (HBN) cohort with subjects not recruted solely based on an ASD diagnosis but on the presence of behavioral constructs relevant to the field of ASD [7]. We design a novel interpretability framework built upon the so-called digital avatars to study ASD-related variability. The contributions are two-fold: *i)* we implement a method to exhibit interpretable relationships between views features, and *ii)* we showcase the relevance of these relationships in the HBN cohort.

## 2. RELATED WORKS

**Model Choice.** In machine learning, Canonical Correlation Analysis (CCA), and its Regularized Generalization to multiple blocks (RGCCA) [8], are frameworks for modeling and quantifying the interactions within and between blocks of heterogeneous variables. CCA extensions can use ANNs as projectors to learn non-linear relationships [9] or embed them as a variational framework [10]. More recent methods apply VAEs for multi-modal learning. For instance, conditional VAEs model latent variables and data conditioned on some input label [11]. Other VAE-based methods avoid such supervision by mixing the view information within each view-specific latent space by optionally enforcing a shared sparsity [3] and reconstructing each view from every latent representation. Most recent multi-view VAE-based models learn a joint latent space between views, writing the joint posterior distribution as a product or a mixture of the marginal posteriors

[4, 5]. Finally, the MoPoe-VAE generalizes the two previous methods by combining their assets [6].

**Interpretability.** ANNs are often criticized for their lack of interpretability and explainability. Considerable efforts have been put into filling this gap in recent years, but the challenge remains. For many practical uses, where the focus is on getting interpretable results, ANNs are bearly usable. A recent review on interpretability classifies ANNs into three categories according to the explanation type, nature (active or passive), and scope (local or global) [12]. The proposed method is passive - applicable to a trained model - and brings global interpretation between attribution - associations between input features and output ones - and the hidden semantics intrinsic to the model.

## 3. METHOD

We selected a particular model among the state-of-the-art multi-view VAEs, namely MoPoe-VAE [6]. VAE assumes that some random process involving an unobserved continuous random variable with lower dimension $Z$ enables the data generation [13]. $Z$ is generated from some prior distribution $p_\theta(Z)$ approximated via the posterior $p_\phi(Z|X)$, computed through an encoder. The data $X$ is generated from some conditional distribution $p_\theta(X|Z) = \mathcal{N}(\mu_\theta(Z), \sigma_\theta)$, approximated via the decoder network. $\phi$ and $\theta$ respectively refer to the encoder and decoder parameters. $Z$ is also referred to as the hidden representation of $X$. MoPoe-VAE handles multi-views by modeling different latent spaces posterior distributions, one specific for each input view, and a joint posterior distribution shared between views as the mixture of the products of their marginal posteriors (experts). Its robustness to missing blocks, which is very frequent in multi-view settings, makes it specifically convenient. It also enables the setting of different sizes for each view-specific latent spaces, convenient when views have discrepancies in dimensionality.

### 3.1. Digital avatars to produce interpretability

Considering a trained MoPoe-VAE, we further focus on the hold out test observations (or subjects) $X_i$, $i \in \{1, \ldots, N\}$. We assume that each observation has every view available, i.e. for all $i$, $X_i = (X_i^1, \ldots, X_i^K)$ with K being the number of views (e.g. clinical, imaging or genomic). We note $X_i^k = (X_i^{k,1}, \ldots, X_i^{k,J_k})$ the view $k$ with $J_k$ features. From the trained MoPoe-VAE, the proposed framework will decipher the relationships of one view with others using digital avatars. Let $l \in \{1, \ldots, K\}$ be this view's index. The general idea is to vary one feature $j$ at a time for each feature of view $l$, $j \in \{1, \ldots J_l\}$ and observe how these modifications influence the reconstruction of other views. Without loss of generality and for simplicity, let's assume $J_l = 1$. A first approach, albeit too simplistic, is to linearly or randomly sample values between feature percentiles or even bootstrapping values between different observations. The proposed method offers a simulation scheme to realistically perturb the feature in block $l$ to generate digital avatars through the model, allowing the discovery of interesting relationships with further analyses. The realistic nature of these perturbations derives from the variational aspect of the VAEs. Indeed they quantify population-level uncertainty via a learned shared variance across subjects and is subject-specific via the mean reconstruction. We therefore try to globally assess what the model learned by introducing subject-level perturbations and monitoring their impact on the reconstructed digital avatars. The interpretability method enables a general insight into individual- and cohort-wide effects. We think this approach is a suitable way to model the multi-faceted variability of the subjects. The proposed strategy can be divided into three steps (inference, simulation and association) outlined below, and illustrated in Figure 1.

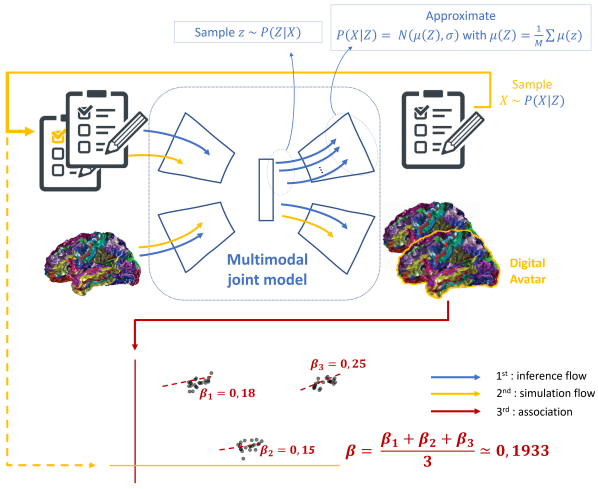


**Fig. 1**. Illustration of the proposed interpretation framework in a clinical cohort setting with two modalities: imaging data and clinical questionnaires. First, the inference flow estimates output distributions via sampling in the latent space. Then, the simulation flow generates realistic perturbed samples of the view we want to study against others (here, the questionnaires) and infer digital avatars through the model. Finally, meaningful inter-view associations are inspected using hierarchical linear regressions.

**Inference flow - estimating output distributions:** We propose to use the output probability distributions $p_\theta(X_i^l|Z_i)$ learned from the data to sample values. This likelihood model will tend to sample likely values in the sense of the model. Therefore, provided our model is properly trained, such values reflects at best the training data. In order to accurately estimate $p_\theta(X_i^l|Z_i)$, we first draws $M = 1000$ latent representations $z_p$, $p \in \{1, \ldots, M\}$, realisations of $Z_i \sim p_\phi(Z|X_i)$ that are sampled for each observation $i$. $p_\theta(X_i^l|Z_i)$ is a Gaussian distribution with two parameters, its mean $\mu_\theta(Z_i)$ and its variance $\sigma_\theta$. Passing the latent rep-

| | eCRF | | Joint | | Image | |
|---|---|---|---|---|---|---|
| **Score** | $\tau$ | $p$ | $\tau$ | $p$ | $\tau$ | $p$ |
| SRS | 0.163 | $<10^{-20}$ | −0.050 | $<10^{-20}$ | −0.051 | $<10^{-20}$ |
| SCARED | 0.189 | $<10^{-20}$ | 0.020 | $10^{-8}$ | −0.014 | $10^{-6}$ |
| ARI | 0.354 | $<10^{-20}$ | −0.014 | $10^{-4}$ | 0.020 | $10^{-10}$ |
| SDQ ha | 0.166 | $<10^{-20}$ | −0.018 | $10^{-7}$ | 0.000 | $10^{0}$ |
| CBCL ab | 0.273 | $<10^{-20}$ | −0.021 | $10^{-9}$ | 0.011 | $10^{-3}$ |
| CBCL ap | 0.252 | $<10^{-20}$ | −0.021 | $10^{-9}$ | −0.008 | $10^{-2}$ |
| CBCL wd | 0.134 | $<10^{-20}$ | −0.009 | $10^{-2}$ | −0.014 | $10^{-5}$ |
| Site | 0.021 | $10^{-6}$ | −0.024 | $10^{-8}$ | 0.128 | $<10^{-20}$ |

**Table 1**. Representational similarity analysis results between the different latent spaces (eCRF, Joint, and Image) and the clinical eCRF scores.

resentations $z_p$ to the decoder, and averaging the $p$ decoded reconstructions offers a good estimation of $p_\theta(X_i^l|Z_i)$ [13].

**Simulation flow - sampling realistic values:** Drawing $T$ samples from $p_\theta(X_i^l|Z_i)$, results in $T$ perturbations $(\hat{X}_{i,t}^l)$ for each observation $i$. We create $T$ new input samples, replacing the original view $l$ value. More formally, we create $(X_{i,1}, \ldots, X_{i,T})$ with

$$X_{i,t} = (X_i^k, \ldots, X_i^{l-1}, \hat{X}_{i,t}^l, X_i^{l+1}, \ldots, X_i^K)$$

We generate the other views from these $T$ perturbed observations with a forward pass in the model, which form the so-called digital avatars.

**Association - finding meaningful associations:** Finally, we search for associations of view $l$ with every feature $s$ of other views $k \neq l$. We test for association using hierarchical regression models. More precisely, we fit a linear regression $X_{i,t}^{k,s} = C_i^{k,s} + \beta_i^{k,s} \hat{X}_{i,t}^l + \epsilon_{i,t}^{k,s}$ for every sample $i$ of our dataset. Then we aggregate the slopes $\beta^{k,s} = \frac{1}{N}\sum_{i=1}^{N} \beta_i^{k,s}$ and test for the null hypothesis $H_0$: *$\beta^{k,s}$ is not null*. We kept associations with *p*-values that passed Bonferroni correction ($p_{cor} < 0.05$).

### 3.2. Representational Similarity Analysis.

To investigate the learned latent representations, we derive a Representational Similarity Analysis (RSA) [14] between these representations and some measures on subjects (e.g. clinical scores or other covariates). We compute the subject-pairwise dissimilarity matrices in the latent space (modality-specific and joint) using the euclidean distance. We derive the same subject-pairwise dissimilarity matrices for the target measures. Finally, the Kendall rank correlation coefficient (Kendall $\tau$) enables the comparison of these dissimilarity matrices emphasizing the captured information.

## 4. EXPERIMENTS AND RESULTS

**Experimental setup:** We trained a MoPoe-VAE on the HBN cohort. HBN is a clinical multi-center transdisciplinary study on stratification biomarkers across neurodevelopmental disorders [7]. This study aims to better understand psychiatric disorders using a variety of assessments, including imaging and a comprehensive set of psychological and clinical scores. The inclusion criteria were defined to gather an at-risk population with notable behavioural symptoms, which makes it a good candidate to study ASD with a dimensional approach. We selected seven scores (eCRF view) from different questionnaires assessing different symptoms, namely Social Responsiveness Scale (SRS), hyperactivity (SDQ ha), anxiety (SCARED), depression (ARI), and behavioral disorders with aggressivity (CBCL ab), attention deficit (CBCL ap) or seclusion (CBCL wd). For brain imaging data, we focused on structural MRI (Image view) and, more specifically, on ROI-based averaged measurements composed of the cortical thickness, curvature, and area computed with FreeSurfer [15]. As a cortical parcellation prior, we opted for Destrieux's parcellation [16].

**Training setting:** The input dataset was split into train and test sets while preserving the population statistics regarding the age, sex, and acquisition site distributions using iterative stratification [17], for subjects with both modalities. We considered subjects with at most one missing view (either the eCRFs or Image data), leading to $N = 2991$ samples. Among them, $N = 1505$ subjects had both eCRF and Image data. Then, we trained a MoPoe-VAE on the train set ($N = 2690$) and monitored the training by ensuring it generalized well to the test set ($N = 301$). The encoders have a fully connected architecture with one hidden layer with 256 units and a ReLU activation, and the decoders are linear. The joint la-

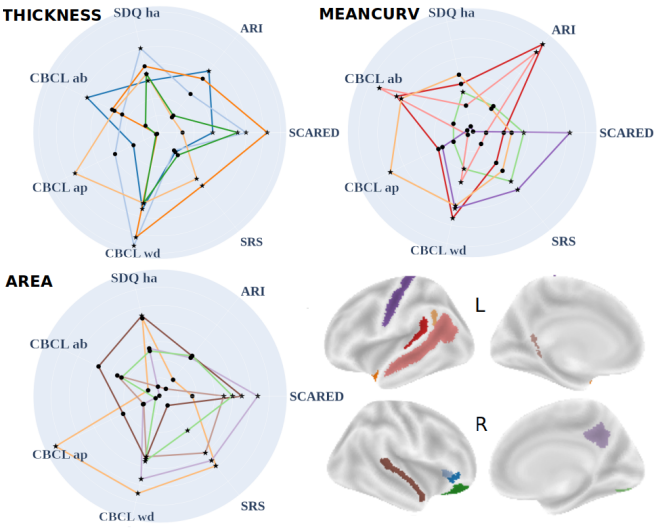


**Fig. 2**. Most significant associations for thickness (top left), curvature (top right) and area (bottom left). The radars show the magnitude of the coefficient corresponding to the association and the brain plots show the location of the associated regions. Stars indicate meaningful associations.

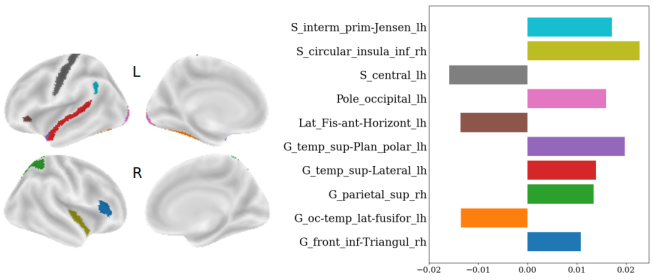


**Fig. 3**. Significant associations between thickness and SRS. The left panel shows the regions locations and corresponding coefficients are shown on the right.

tent space has a dimension of 20, and specific latent spaces for eCRF and Image views have dimensions of 3 and 20, respectively. We use an Adam optimizer [18] with a learning rate of $2 \times 10^{-3}$. The model has a learnable parameter for the variance of each reconstructed feature, which does not depend on the input sample. All the data are z-scored before being passed to the network. For all the associations' analyses that follow the training, we only use test subjects (for which we have both modalities, $N = 301$). The code to reproduce the experiments and results is made available here.

**Latent representations evaluation:** The RSA results on the modality-specific and joint latent representations are given in Table 1. These results show that the eCRF latent variables contain most information about each questionnaire score, the joint space captures some of it, and the Image space a little bit less. This supports the fact that imaging contains some symptom related variability.

To assess the site effect captured by our model, we used RSA between site and latent representations from the different latent spaces. The RSA indicated that the Image latent space strongly encodes the site, while the joint latent space is less impacted, as well as the eCRF latent space. This suggest that most site effect is captured by the Image latent space and the joint latent space is partially spared, which is good as it is the only latent space explored in our associations analysis. Further, to ensure our association discovery method is not biased towards the site, we conducted ANOVA analyses to test for differences in distributions of association coefficients between sites and did not find any significant effect.

**ROI-score meaningful associations:** We focus on the 301 left-out subjects with all views to study the influence of each clinical score on the cortical features' reconstructions. We draw $T = 200$ samples for each subject and each eCRF score to generate as many digital avatars. Then, we computed associations as described in Section 3.1. To study statistical significance, we bootstrapped 20 times the whole procedure by using 150 randomly sampled subjects of the test set and kept only stable associations reproduced in more than $70\%$ of these experiments. In Figure 2, we represented the ROIs significantly associated with multiple eCRF scores across metrics (five associations per metric). We also showed the coefficient corresponding to these associations' strengths in the radar plots. This figure highlights some commonly reported regions in the ASD literature [19]. In particular, we could observe that the left and right temporal areas (pink and brown) are associated with different anxiety and seclusion related scores, in curvature and area respectively. The central sulcus of the left hemisphere (deep purple) is associated with anxiety related symptoms in curvature. Some other frontal regions (blue) are associated in thickness with seclusion, anxiety and hyperactivity, as well as with other symptom scores.

To further explore the results, we plotted in Figure 3 all the regions significantly associated with the SRS score regarding the thickness feature. The SRS is a well-established behavioural marker, highly correlated to the ASD diagnostic. ASD subjects have a higher SRS score. Two trends appear: thickness in the right insula and the left temporal superior, right parietal, and frontal regions increases with SRS, while thickness in the central sulcus decreases with SRS (i.e., highlighting a thicker, respectively thinner cortical ribbon in ASD at these locations). Those regions are commonly reported in the ASD literature, but it should be noted that they are listed here as implicated in an association with a score regardless of typical or neurodiverse development.

## 5. CONCLUSION AND FUTURE WORKS

Overall, we present and showcase how to derive from a multi-view VAE some significant relationships between views. The multi-view VAE was trained on the HBN cohort using all the subjects, including those with missing records, and was able to learn information at the population and subject level. With our interpretation framework, we exhibit in HBN view-specific features associations that convey interpretable information and potential biomarkers. This framework reworks the classical association studies between score and structural imaging features: we highlight some relationships in an integrative approach while taking into account subject- and population-specific effects. Learning from the whole cohort, composed of at-risk individuals and typically developing controls, our method exhibits associations holding for the entire symptoms range. Unlike a linear regression analysis between regional thickness and a score, we did not specify any covariates. This approach achieves association search by questioning the subjects specificities thanks to their associated digital avatars that reflect cohort-level and contextual subject variabilities.

Our work presents limitations. We varied one feature at a time to allow simplified interpretation. We could argue that such a setting is irrealistic because the underlying biological process may be related to complex score interactions (in HBN, scores are correlated to each others $0.75 > r > 0.17$). Future work will account for these relationships when varying features.

## 6. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the Child Mind Institute.

## 7. ACKNOWLEDGEMENT

The authors acknowledge the Healthy Brain Network initiative at the Child Mind Institute in New York City, NY, USA. The authors declare that they have no conflict of interest.

## 8. REFERENCES

[1] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-IV*, APA, Washington, DC, 4th ed. edition, 1994.

[2] T. Insel et al., "Research domain criteria (RDoC): toward a new classification framework for research on mental disorders," *Am J Psychiatry*, vol. 167, no. 7, pp. 748–751, July 2010.

[3] L. Antelmi et al., "Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds. 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 302–311, PMLR.

[4] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. 2018, vol. 31, Curran Associates, Inc.

[5] Y. Shi et al., "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. 2019, vol. 32, Curran Associates, Inc.

[6] T. Sutter et al., "Generalized multimodal ELBO," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.

[7] L. M. Alexander et al., "An open resource for transdiagnostic research in pediatric mental health and learning disorders," *Scientific Data*, vol. 4, no. 1, pp. 170181, Dec 2017.

[8] A. Tenenhaus and M. Tenenhaus, "Regularized Generalized Canonical Correlation Analysis," *Psychometrika*, vol. 76, no. 2, pp. 257–284, Mar. 2011.

[9] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, Eds., Atlanta, Georgia, USA, 17–19 Jun 2013, vol. 28 of *Proceedings of Machine Learning Research*, pp. 1247–1255, PMLR.

[10] W. Wang et al., "Deep variational canonical correlation analysis," 2017.

[11] D. P. Kingma et al., "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 3581–3589.

[12] Y. Zhang et al., "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, oct 2021.

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.

[14] N. Kriegeskorte et al., "Representational similarity analysis - connecting the branches of systems neuroscience," *Frontiers in Systems Neuroscience*, vol. 2, 2008.

[15] A. M. Dale et al., "Cortical surface-based analysis: I. segmentation and surface reconstruction," *NeuroImage*, vol. 9, no. 2, pp. 179–194, 1999.

[16] C. Destrieux et al., "Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature," *NeuroImage*, vol. 53, no. 1, pp. 1–15, 2010.

[17] K. Sechidis et al., "On the stratification of multi-label data," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 145–158.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[19] C. Ecker et al., "Interindividual Differences in Cortical Thickness and Their Genomic Underpinnings in Autism Spectrum Disorder," *AJP*, p. appi.ajp.2021.2, Sept. 2021.