Interpretable and integrative deep learning for discovering brain-behaviour associations with stability analysis

Corentin Ambroise, Antoine Grigis, Josselin Houenou, Vincent Frouin

INTERPRETABLE AND INTEGRATIVE DEEP LEARNING FOR DISCOVERING BRAIN-BEHAVIOUR ASSOCIATIONS WITH STABILITY ANALYSIS

PREPRINT

Corentin Ambroise
NeuroSpin, Institut Joliot, CEA
Université Paris-Saclay
Gif-sur-Yvette 91191, France
corentin.ambroise@cea.fr

Antoine Grigis
NeuroSpin, Institut Joliot, CEA
Université Paris-Saclay
Gif-sur-Yvette 91191, France
antoine.grigis@cea.fr

Josselin Houenou
NeuroSpin, Institut Joliot, CEA
Université Paris-Saclay
Gif-sur-Yvette 91191, France
Pôle de Psychiatrie, AP-HP, Faculté de Médecine de Créteil,
DHU PePsy, Hôpitaux Universitaires Mondor,
Créteil 94000, France
josselin.houenou@aphp.fr

Vincent Frouin
NeuroSpin, Institut Joliot, CEA
Université Paris-Saclay
Gif-sur-Yvette 91191, France
vincent.frouin@cea.fr

ABSTRACT

Relying on traditional classification strategies from a single data source is now recognised as ineffective for understanding, diagnosing, and predicting psychiatric syndromes. Classification targets that rely solely on clinician labels will not capture enough variability. In 2009, the Research Domain Criteria (RDoC) project recommended a more comprehensive approach to studying psychiatric disorders by incorporating diverse data types that cover different levels of life organisation (e.g., imaging, genetics, symptoms). The RDoC principles suggest that a thorough description of a pathology requires consideration of multiple dimensions that may be shared across different psychiatric syndromes and may even contribute to non-pathological variability. Efficient multivariate and multimodal unsupervised learning frameworks hold the promise of providing methodologies for handling and integrating the kind of datasets advocated by the RDoC project. Of particular interest, deep learning offers the ability to learn on multimodal datasets with modality-specific correlation structure. However, it is often disregarded due to its perceived lack of transparency. In this study, we employ a digital avatar procedure as an interpretability module capable of reporting the relationships learned within a multimodal autoencoder. We integrate this procedure into a novel framework that utilises stability selection to identify meaningful and reproducible associations between brain imaging modalities and behaviour. Specifically, we apply this framework to uncover specific brain-behaviour interactions present in the transdiagnostic Healthy Brain Network cohort. The identified brain-behaviour interactions establish connections between cortical measures derived from structural magnetic resonance imaging and electronic clinical record forms assessing psychiatric symptoms. We show that by using incomplete records and automatically isolating variability of interest from that of confounders, this framework is able to find relevant and stable associations.

1 Introduction

Today, psychiatry, in both its diagnostic and therapeutic dimensions, is moving from a paradigm based on the study of syndromes to a new one built on an understanding of their underlying neurobiological mechanisms. Achieving this

goal of precision psychiatry is an ongoing process that requires the combination of scientific advances, technological innovations, and a patient-centered approach [Williams, 2022]. Identifying relationships between behaviours and brain measures is a key aspect of this paradigm. However, behaviours are complex and often result from a combination of genetic, environmental, and psychological factors. In addition, many behaviours observed in mental health conditions can be present in multiple disorders, and the same disorder may manifest with different behaviours in different individuals. This complexity challenges the idea that a single behaviour signature can correspond to a sole disorder. The NIMH Research Domain Criteria (RDoC) [Insel et al., 2010] provides recommendations for properly addressing this complexity. In contrast to traditional approaches that aim to find a diagnosis from a specific score or modality (as outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM) [Association, 2013]), the RDoC promotes dimensional and transdiagnostic approaches. These dimensions include genetic, biological, environmental, and lifestyle factors in the research on personalised psychiatry. In recent years, the transdiagnostic literature in psychopathology has developed along these lines [Fusar-Poli et al., 2019]. Transdiagnostic studies propose to examine a general psychopathological factor, the so-called p-factor [Caspi et al., 2014, Caspi et al., 2023]. This p-factor underlies mechanisms common to several psychiatric syndromes and is usually represented as a single dimension calculated on the basis of different symptoms. The goal of such studies is to search for neural correlates of the p-factor using various biological markers, such as imaging or genetics.

For these novel experimental studies in populations, new methods are being developed to integrate multimodal data, including structural or functional characteristics of the brain, tabular data from report forms, genotyping, or lifestyle conditions. At the same time, cohorts of at-risk individuals and numerous clinical initiatives are collecting datasets consistent with the RDoC framework. In imaging genetics, publicly available cohorts show diverse psychological scores as well as imaging phenotypes such as regional cortical thickness or gyrification. Among the currently available integrative methods, some aim to adjust a classifier, a regressor, a clustering to find biomarkers (selection of combined features) or significant associations (univariate approach). In this paper, our focus will derive from the latter type of approach –integrative association study –, specifically aimed at uncovering associations between measurements coming from multiple modalities. Integrative association studies represent a departure from conventional univariate association analyses that consider a single measurement from one modality and find its associations with other modalities. Instead, we combine multivariate tools with association tools in accordance with the principles of precision psychiatry. These tools allow for modeling intra- and inter-modality structures, enabling the estimation of joint relationships between multiple measures across modalities. Integrative association methods face specific challenges, most notably the problem of missing data. This problem does not only affect data in a sporadic manner but can occasionally strikes all measurements of a modality for a patient. Furthermore, research in integrative association is also very active, especially regarding three axes: first, the integrative capacity of the considered multivariate models –whether they are linear or not –, second their interpretability, and third their generalisability or stability. These three points are important in assessing the contribution of these methods to precision psychiatry and are detailed below.

Integrative capacity. Pioneering works on integrative association models are Canonical Correlation Analysis (CCA) [Hotelling, 1936] or Partial Least Squares (PLS) [Wold et al., 2001]. We retain these two methods as emblematic of integration analysis. Two essential concerns when using these approaches are the limited capacity of linear framework used, and the need to normalise the different data blocks beforehand. Beyond linearly adjustable covariates such as age or sex, the removal of other known confounding factors, such as imaging acquisition site, is more challenging. These factors introduce variability that may entangle the signal of interest in a manner that depends on the recruitment configuration. The need to address normalisation challenges has drawn attention to artificial neural network-based methods. For example, mutli-view AutoEncoders (AEs) have been shown to embed disentanglement capabilities [Lee and Pavlovic, 2021], which in turn could replace normalisation. In line with the idea proposed by [Andrew et al., 2013], we propose to use Deep Learning (DL) for its greater integrative capacity and to leverage this disentanglement to avoid normalisation. Of particular interest, multimodal Variational AE (mVAE) is an extension of traditional VAE, as proposed by [Kingma and Welling, 2014], designed to handle multiple views or modalities. It extends the idea of a VAE to capture the joint distribution of data from different views. The choice of the prior distribution in latent space represents the assumed distribution of the latent variables. Thus, it plays a crucial role in constraining the learning process and influences the structure of the latent space. Classical choices for the prior distribution in an mVAE include standard or multimodal Gaussian distributions [Kingma et al., 2014, Suzuki et al., 2017, Wu and Goodman, 2018, Antelmi et al., 2019]. Recent proposals for prior distributions include Product of Experts (PoE), Mixture of Experts (MoE), or a combination of the two (MoPoE) [Wu and Goodman, 2018, Shi et al., 2019, Sutter et al., 2021]. The rationale for adopting a mVAE lies in its ability to capture shared latent representations that account for intra-view correlation structure and model the inter-view correlation. However, questions remain regarding some limitations of mVAE in modeling only a shared latent space where information from different views is integrated [Daunhawer et al., 2022]. Recent research advocates exploring alternatives involving view-specific latent spaces [Sutter et al., 2021, Lee and Pavlovic, 2021, Daunhawer et al., 2021,

Palumbo et al., 2022]. In this work, we propose to use a MoPoE-VAE as a specific mVAE, which promotes integrative capacity. This model can elaborate modality-specific and shared across modalities latent representations during the training phase. This shared representation conveniently dampens the influence of non-interest factors.

Interpretability. Using an mVAE to perform integrative association analysis comes at the expense of interpretability, a strength of their linear counterparts, such as CCA or PLS. Indeed, these classical methods have been praised for their interpretability. For example, adaptations of CCA and PLS propose to obtain interpretability by imposing latent covariation exploration with various sparsity constraints, including L1 or total variation regularisations [Tenenhaus et al., 2014, Cao et al., 2011]. These implementations not only identify associations between different blocks, but also select the variables within each block that contribute to these correlations. This selection step facilitates the translation of these associations into interpretable knowledge. Bringing interpretation capabilities to mVAE, and more generally in DL, remains an opened and significant challenge. In previous work, we developed an interpretability method to explain the information learned by a mVAE. This method is used to evaluate the effect of controlled variations in the input of one view on the output of other views as generated by the mVAE [Ambroise et al., 2023]. In the present work, we adopt this novel interpretability framework, which is rooted in what we call a Digital Avatar (DA). This framework is used to study brain-behaviour relationships, allowing us to showcase interpretable associations between features across different views. Harnessing the generative capabilities of a trained mVAE, we generate sets of DAs from left-out subjects. Starting from a subject and varying its behavioural score, the model reconstructs the corresponding brain image for each score value. The analysis performed on the obtained sets of DAs yields what we hereafter call a Digital Avatar Analysis (DAA). It allows the exploration of brain-behaviour relationships through linear model fitting for all pairs of behavioural scores and regional image measures. The interpretability of DAs is enhanced by their exclusive reliance on the joint latent representation. However, DAA results are strongly affected by the epistemic variability [Kendall and Gal, 2017] of the mVAE inherent to training stochasticity and weight initialisation. We mitigate the effects of epistemic variability through ensembling, using multiple trained mVAE models with different initialisation parameters on the same training subjects. This strategy, which acts as a regularisation of the DAA, is parameterised by the number of trained models and hereafter referred to as regularised-DAA (r-DAA). In summary, we expect that an ensemble of DAAs will provide a more comprehensive understanding of brain-behaviour associations.

Stability. The stability of results in integrative association studies within a multimodal dataset is a critical consideration to ensure the generalisability of the approach. Traditional approaches, such as sparse CCA, employ cross-validation strategies that often include the definition of a downstream auxiliary predictor for model selection. However, it has been shown that these procedures may not consistently identify stable associations [Labus et al., 2015, Baldassarre et al., 2017, Ing et al., 2019, Mihalik et al., 2020, Chegraoui et al., 2023]. This is particularly true when correlations between views are low or sample sizes are small [Cao et al., 2011, Helmer et al., 2023, Yang et al., 2021, Nakua et al., 2023]. Clearly, the challenge of stability remains an open question, especially in the context of the DAA described above. To address this concern, we propose to use a sound framework. Stability selection is an inspiring technique developed in machine learning that efficiently exploits the regularisation mechanisms embedded in classical optimisation problems such as fitting a classifier [Meinshausen and BÃijhlmann, 2010]. This procedure consistently identifies stable structures upon which the classifier is built, which inherently promotes robustness. In this study, we leverage Meinshausen's concepts regarding the stability of selected features and extend their application to the stability of selected associations. In fact, the DAA procedure, which generates the brain-behaviour associations, is not only affected by the epistemic variability, but is also sensitive to aleatoric variability (characteristics of the train / left-out split of the dataset). To reduce it, the stability selection procedure splits the dataset multiple times to generate different training and left-out splits. This allows the generation of stability paths, which identify stable brain-behaviour associations. The overall goal of this approach is to increase the robustness of our findings. This, in turn, contributes to a more reliable exploration of brain-behaviour relationships in the context of multimodal data analysis.

Overall, we present an empirical framework for extracting a consistent set of brain-behaviour associations from a multimodal dataset containing behavioural reports and brain imaging features. This framework is based on the stable selection of r-DAA output associations. In summary, our methological contributions are threefold. First, we propose a r-DAA that uses a weighted ensembling procedure to consolidate interpretations. Second, we introduce a stability selection procedure to generate a set of robust brain-behaviour associations. Finally, we show that our procedure, which is built on mVAEs, inherits from their ability to handle data with missing modality, and alleviates the need for additional normalisation. The proposed framework is applied to a publicly available cohort of at-risk children with notable behavioural symptoms. The cohort consists of structural MRI data and behavioural scores. Our results reveal interesting transdiagnostic brain-behaviour associations common to several psychiatric syndromes, such as Autism Spectrum Disorder (ASD) or Attention Deficit Hyperactivity Disorder (ADHD).

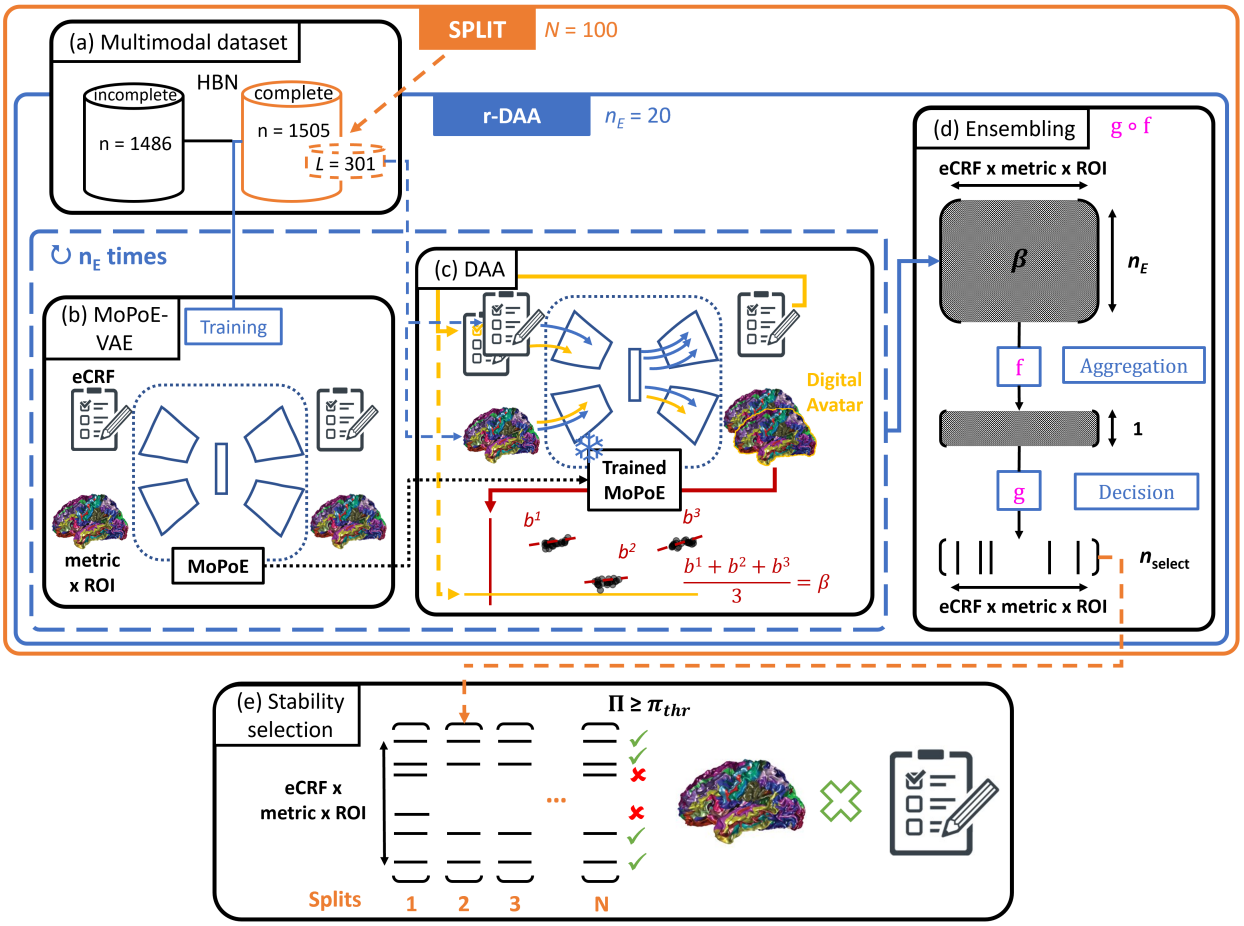


Figure 1: Illustration of the interpretation framework based on DAA. In (a) the HBN cohort with two modalities: imaging data (metric $\times$ ROI) and clinical questionnaires (eCRF). For each split (orange bounding box) of the dataset, complete data are split into training (plain orange removing the dotted part) and left-out (dotted orange part) subjects. Each r-DAA (blue bounding box) procedure uses one split of the dataset. In the inner dotted blue bounding box, we repeat the following n_E times : (b) a MoPoE-VAE is trained using all the subjects with one modality and the training subjects with both. (c) Left-out subjects are used for DAA using this trained model, outputting coefficients β. (d) These n_E DAA outputs are then ensembled with an aggregation and a binary support decision function into n_{select} associations. The r-DAA procedure is repeated for each of the $N = 100$ different splits of the complete dataset. (e) A stability selection procedure is used to retain associations with stability path $\Pi \geq \pi_{\text{thr}}$ high enough. The whole framework outputs stable associations between clinical questionnaires and imaging.

2 Material and methods

This section introduces the proposed concept of performing stable association discovery in multimodal data using association study and stability selection. The digital avatar-based association study, referred to as DAA, is designed to find associations from a trained mVAE in heterogeneous multimodal datasets with potential missing data. Specifically, a regularised DAA (r-DAA) procedure is employed to mitigate the epistemic uncertainty inherent in training neural networks. Regularisation is achieved by ensembling the results of DAAs to retain meaningful associations. Finally, stability selection is applied to r-DAA by resampling the original dataset. This strategy is employed to handle the aleatoric uncertainty associated with the inherent variability in the data. Associations consistently selected by the r-DAA across these resampling splits are considered stable associations. The proposed approach consists of three nested steps listed from the innermost to the outermost: 1) DAA - training a mVAE and conducting an association study, 2) r-DAA - ensembling results from several DAAs derived on a same left-out set but with different initialisations and training batches, and 3) stability selection - performing several r-DAAs on different left-out sets to select stable associations. These steps are described below and illustrated in Fig. 1, and Supplementary Alg. D.1 and D.2.

2.1 Multimodal dataset

This study uses the Healthy Brain Network (HBN) cohort [Alexander et al., 2017]. This dataset is a multi-center, transdisciplinary clinical study. It includes a variety of assessments, including imaging, and a comprehensive set of psychological and clinical assessments to better understand psychiatric disorders. Inclusion criteria are not diagnosis dependent, but rather encompass an at-risk population with notable behavioural symptoms. Specifically, subjects were selected based on the presence of behavioural constructs related to Autism Spectrum Disorder (ASD) or Attention Deficit Hyperactivity Disorder (ADHD). Consensus diagnoses are not available for the majority of subjects enrolled in the HBN cohort. As such, the dataset allows for the study of the different manifestations of psychiatric syndromes within the data. In particular, it provides an opportunity to explore methods for new biomarker discovery and dimensional analysis.

In a previous study, our group identified seven behavioural assessments that capture the most salient dimensions in ASD patients [Mihailov et al., 2020]. Specifically, a quantitative measure of clinical autistic traits was defined as the parent Social Responsiveness Scale (SRS), hyperactivity levels were determined using the hyperactivity subscale within the Strengths and Difficulties Questionnaire (SDQ-ha), the anxiety was measured using the total score from the Screen for Child Anxiety Related Disorders Parent-Report (SCARED), irritability was defined using the total score of the Affective Reactivity Index Parent-Report (ARI), and finally, levels of depression, aggression, and attention problems were determined using subscales of the same names within the Child Behavior Checklist (CBCL-wd, CBCL-ab, and CBCL-ap, respectively). Filtering out subjects with sporadic missing data in questionnaires results in 2454 individuals.

The MRIs were acquired at four sites. A mobile 1.5T Siemens Avanto on Staten Island, a 3T Siemens Tim Trio at the Rutgers University Brain Imaging Center, and 3T Siemens Prisma at the CitiGroup Cornell Brain Imaging Center and at the CUNY Advanced Science Research Center. Collected T1-weighted images were preprocessed using FreeSurfer [Dale et al., 1999]. All results were manually reviewed in-house. The Euler number is used as a quality metric summarising the topological complexity of the reconstructed cortical surfaces [Rosen et al., 2018]. Specifically, a single Euler number exclusion threshold of -217 is applied to yield 2042 selected subjects. Finally, cortical measures based on three metrics - cortical thickness, curvature, and area - are averaged in the 148 cortical regions of interest (ROIs) defined by Destrieux's parcellation [Destrieux et al., 2010].

The proposed brain-behaviour integrative analysis considers these two blocks of data. First, the electronic Clinical Record Forms (eCRF) view consists of the $p_{eCRF} = 7$ behavioural scores. Second, the ROI view is composed of the $p_{ROI} = 444$ cortical features from the 3 considered metrics across the 148 ROIs. In total, our dataset comprises 2991 subjects. Among them, 1505 have both complete views (referred to as complete dataset in Fig. 1(a) and 1486 have only one of the two views available (referred to as incomplete dataset in Fig. 1(a). Missing views remain a common problem in data integration. The factors contributing to missing data are usually not known in advance. Most traditional data mining and machine learning approaches operate on complete data and fail when data is missing. As fallback strategies, some models rely only on samples with all views available or on an auxiliary inference step that generates missing views. Our goal here is to use a model able to accommodate missing views, in order to use a maximum number of available samples.

2.2 Digital avatar analysis

Multimodal deep learning, which does not impose data normalisation and provides integrative capacitive models, is a promising approach to decipher the relationship between clinical dimensions, neuroimaging and a pathology. It has the potential to support the development of personalised medicine. If the considered model has generative capabilities, the exploration of a disorder spectrum can be achieved through the generation of so-called Digital Avatars (DAs). In the proposed work, we employ a mVAE and specifically use a MoPoE-VAE (see Supplementary A for details). We keep $L = 301$ subjects with complete data for the left-out set and use the remaining subjects, whether having complete or incomplete data, for the training set. The MoPoE-VAE can indeed be trained on data with missing views. Once trained, a MoPoE-VAE cannot directly provide brain-behaviour associations. However, the information learned by the model can be used to discover relevant pairwise associations between views, which we consider in our work to be relevant pairwise associations between each brain imaging feature and each score. Starting from the left-out subjects, we modify each subject's clinical score and generates a set of $T = 200$ DAs. Thus, as being predicted by the model, each DA bears brain imaging features at a given clinical state. From this set of virtual neuroimaging data, a conventional association study is performed to find brain-behaviour relationships. For each score of the eCRF view, we perform p_{ROI} univariate association studies. In a previous publication, we introduced this DA analysis [Ambroise et al., 2023]. Hereafter, we describe in detail how the DAs are obtained and analysed.

For the DAA step, we use the left-out set for which all subjects have complete data. Without loss of generality and for simplicity, we consider the case where the eCRF view has only one score in what follows. Our aim is to create a set of DAs from each of the L left-out subjects. Let $s = (s^1, \ldots, s^L) \in \mathbb{R}^L$ represent the scores from the eCRF view, and $m = [m^1, \ldots, m^L] \in \mathbb{R}^{L \times p_{ROI}}$ represent all cortical measures from the ROI view. Note that this specific case can be generalised to all clinical scores and additional views. The general idea is to modify s into $\hat{s}$ and observe how these modifications affect the reconstructed cortical measures $\hat{m}$. In this way, virtual pairs $(\hat{s}, \hat{m})$ are obtained. A first, although simplistic, approach to perturb s is to sample linearly or randomly between feature percentiles or even to bootstrap values across subjects. Conversely, the proposed approach relies on a simulation scheme to realistically perturb s. The realism of these perturbations stems from the generative aspect of the model. We hypothesise that such an approach, which models the composite variability of subjects measurements, will facilitate the discovery of interesting brain-behaviour relationships. Indeed, the variability in the simulated avatars integrates both subject-specific and population-level variability. The latter source of variability is captured through the learned variance that is shared across the population. Our goal is to comprehensively assess what the model has learned by introducing subject-level perturbations and investigating their effects on the reconstructed digital avatars. The proposed strategy can be divided into three stages (inference, simulation, and association), which are outlined below and illustrated in Fig. 1(c) and Supplementary Fig. S2.

Inference - estimating likelihood distributions
For a given subject $i \in \{1, \ldots, L\}$, we propose to use $p_\theta(s^i|z^i)$, the likelihood distribution of observations conditioned on the latent variable learned from the data, to sample DA score values. In $p_\theta(s^i|z^i)$, θ represents the weights of the eCRF view decoder, and $z^i = (z^i_{\text{eCRF}}, z^i_{\text{Joint}})$ is the latent representation of s^i, consisting of the eCRF specific and joint latent representations, respectively (see Supplementary A). This likelihood model will tend to sample likely values in the sense of the model. Therefore, provided our model is properly trained, such perturbed clinical scores will at best reflect the training data. The estimation of $p_\theta(s^i|z^i)$ is obtained by drawing $D = 1000$ realisations of $z^i \sim q_\phi(z^i|s^i, m^i)$, where ϕ are the weights of the MoPoE-VAE encoders (see Supplementary A). Passing these latent representations to the eCRF decoder provides a good estimate of $p_\theta(s^i|z^i)$ [Kingma and Welling, 2014] (by averaging the D decoded reconstructions). Note that the same strategy can be applied to categorical data by approximating the parameters of a Bernouilli distribution instead of a Gaussian.

Simulation - sampling realistic values
Given a subject $i \in \{1, \ldots, L\}$, T samples are drawn from $p_\theta(s^i|z^i)$, resulting in T perturbations $\hat{s}^i \in \mathbb{R}^T$. Repeating this sampling for all subjects i, the generated perturbed set forms our DA perturbed eCRF observations $\hat{s} = [\hat{s}^1, \ldots, \hat{s}^L] \in \mathbb{R}^{L \times T}$. Importantly, when we generalise to multiple clinical scores, only one score is perturbed at a time. Finally, the perturbed ROI measures are reconstructed using a forward pass by considering $\hat{s}$ and the corresponding ROI features m as input to the model. This results in a set of T perturbed ROI measures representing our DAs $\hat{m} \in \mathbb{R}^{L \times T \times p_{ROI}}$. As a compromise between computational cost and accuracy, we set $T = 200$.

Association - computing inter-view associations
We search for associations between a clinical score and each ROI measure. We refer to the generated DA score values as $\hat{s} \in \mathbb{R}^{L \times T}$, and generated DA ROI measures as $\hat{m} \in \mathbb{R}^{L \times T \times p_{ROI}}$. Associations are obtained using hierarchical regression models [Bryk and Raudenbush, 1992]. Specifically, for each image feature $k \in \{1, \ldots, p_{ROI}\}$, a linear regression of the form $\hat{s}^i = c^i_k + b^i_k \hat{m}^i_k + \epsilon^i_k$ is fitted for each subject i. The resulting slopes are averaged over all subjects as $\beta_k = \frac{1}{L} \sum_{i=1}^{L} b^i_k$. Performing the same analysis for all available eCRF scores results in the association vector $\beta \in \mathbb{R}^p$, where $p = p_{ROI} * p_{eCRF}$. This vector encompasses all potential associations between the p_{eCRF} clinical scores of the eCRF view and the p_{ROI} cortical features of the ROI view.

2.3 Regularised digital avatar analysis

In our seminal paper [Ambroise et al., 2023], we observed that the associations were not stable, even when the model was retrained on the same training set, and evaluated on the same left-out set. This instability is likely due to epistemic variability [Kiureghian and Ditlevsen, 2009, Kendall and Gal, 2017], and has mainly been studied in supervised settings [Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017]. In feature selection using deep learning, studies have shown that ensembling can enhance stability and thus mitigate epistemic variability [Gyawali et al., 2022]. We propose to repeat the previous step n_E times: we train a MoPoE-VAE on the training subjects, then perform the DAA on the obtained model using the left-out set. Note that the training set and left-out set remain the same throughout these n_E procedures, only the random weights initialisations and batches during training vary. This is illustrated in Fig. 1. By subsequently using ensembling, our objective is to identify stable associations from n_E DAAs association matrix $\boldsymbol{\beta} = [\beta^1, \ldots, \beta^{n_E}] \in \mathbb{R}^{n_E \times p}$. Like classical deep ensembling, our approach involves ensembling candidate associations (i.e., model predictions in supervised settings [Lakshminarayanan et al., 2017]) proposed by the n_E DAAs, and regrouped in the $\boldsymbol{\beta}$ matrix. The proposed ensembling procedure requires the definition of an aggregation function

f and a decision function g. The aggregation function summarises the n_E associations coefficients. The decision function generates a binary decision support from the aggregated coefficients, which is designed to retain meaningful associations. Formally, let $f : \mathbb{R}^{n_E \times p} \to \mathbb{R}^p$ be an aggregation function and $g : \mathbb{R}^p \to \{0,1\}^p$ be a decision function. The composition $g \circ f$ forms the proposed ensembling, which is outlined below and illustrated in Fig. 1(d) and Supplementary Alg. D.1(b) and D.2(b).

Aggregation - f
The role of the aggregation function f is to assign an importance to each association from the association matrix $\boldsymbol{\beta}$. Initially, we opt for a classical mean, although alternative functions such as median or maximum could also be relevant. However, the quality of the estimated n_E models is not accounted for by any of these functions. In fact, some models may converge to local minima of the optimised loss function, resulting in less relevant models and representations. Such behaviour can produce outlier associations that affect the stability of selected associations [Gyawali et al., 2022]. Therefore, we employ a weighted average where the weights reflect the quality of the estimated models. The weights are determined by the model's joint latent space ability to capture a significant proportion of the eCRF-related variability. Practically, we use the representational similarity analysis (RSA) [Kriegeskorte et al., 2008], which generates correlations between the joint latent space and the eCRF scores (see Supplementary B and Supplementary Alg. D.2 for details).

Decision - g
The purpose of the decision function g is to obtain binary decision support from the aggregated coefficients on a score-metric basis. This support allows us to choose a subset of the most informative associations for each score-metric pair (s, m), enhancing reproducibility and interpretability. While a simple strategy involves specifying a numeric threshold or defining built-in heuristics to find this threshold, we prefer setting the number of features explicitly rather than using a threshold. This approach provides consistency across score-metric pairs and DAAs. In practice, we choose to select the top $n_{\text{select}} = 12$ associations with the greatest amplitudes.

2.4 Stability selection

Stability selection, derived from penalised machine learning [Meinshausen and BÃijhlmann, 2010], is designed to reinforce feature selection through the definition of stability paths. Stability paths represent the probability of selecting each feature when training the same algorithm with some regularisation parameter on different random splits of a dataset. We extend this methodology to our brain-behaviour association study. In our context, the regularisation parameter is the number of models n_E considered in the ensembling step. We propose to repeat the r-DAA (see Section 2.3) N times, using different splits of the original dataset, and setting the regularisation parameter $n_E \in \{1, \dots, 20\}$. This approach allows us to model aleatoric uncertainty inherent to population variability [Kendall and Gal, 2017]. The implementation comprises three aspects outlined below and illustrated in Fig. 1(e) and Supplementary Alg. D.1(c) and D.2(c). First, we define a valid splitting strategy. Then, we repeat the r-DAA N times by varying the regularisation parameters n_E, enabling us to estimate stability paths. Finally, we define a criterion to assess the stability of the results obtained across the N splits.

Splitting strategy
Our dataset consists of two parts: 1505 subjects with complete data and 1486 subjects with incomplete data (see Section 2.1). It's important to note that subjects with missing views can only be used to train the model. For each split, out of the 2991 subjects available, 2690 are used in the training set. The latter includes the 1486 subjects with incomplete data and 1204 randomly selected subjects with complete data. The remaining $L = 301$ subjects with complete data form the left-out set (see Fig. 1). To maintain population statistics, including age, sex, and acquisition site distributions, we employ shuffled iterative stratification [Sechidis et al., 2011]. Although using the entire incomplete dataset in each training iteration may appear as a limitation, it is effectively mitigated by employing different shuffled batches at each training epoch. We opt for using $N = 100$ splits.

Stability paths
Each r-DAA with regularisation parameter n_E produces a decision support $S^j(n_E) \in \{0,1\}^p$ for a given split $j \in \{1, \dots, N\}$. The stability paths $\Pi^{n_E} \in [0,1]^p$ are the probability for each association to be selected across the N splits. It is obtained by averaging all the calculated binary decision supports as:

$$\Pi^{n_E} = \frac{1}{N} \sum_{j=1}^{N} S^j(n_E) \tag{1}$$

Stability criterion
The stability criterion analyses the stability paths and defines which associations are considered stable. If the prob-

Score	eCRF $\bar{\tau}$ (↑)	Joint $\bar{\tau}$ (↑)	ROI $\bar{\tau}$ (↑)
SRS	**0.302**	**0.018**	−0.003
SCARED	**0.101**	**0.032**	−0.004
ARI	**0.256**	**0.05**	**0.017**
SDQ ha	**0.326**	**0.058**	0.005
CBCL ab	**0.406**	**0.03**	0.008
CBCL ap	**0.443**	**0.026**	−0.002
CBCL wd	**0.152**	**0.06**	0.008
Site	0.004	0.011	**0.156**
Age	−0.005	**0.05**	**0.10**
Sex	0.014	0.003	**0.067**

Table 1: Representational similarity analysis results between the different latent spaces (eCRF, Joint, and ROI) and the clinical eCRF scores, and some covariates of interest : the imaging acquisition site (Site), the patient age (Age) and sex (Sex). $\bar{\tau}$ is the corresponding average Kendall τ across models and splits. We computed the p-value associated with this statistic. Values in bold indicate their significativity, i.e. median corrected p-value < 0.01, see Supplementary B for details.

ability of an association happens to be greater than a user-defined threshold $\pi_{\text{thr}} \in [0, 1]$ for a specific regularisation parameter n_E, then the corresponding association is selected. Finally, we define the set of stable features as:

$$S^{\text{stable}} = \{k : \max_{n_E \in \{1,\dots,20\}} \Pi_k^{n_E} \geq \pi_{\text{thr}}\} \tag{2}$$

More simply, this means that associations with a high probability of selection for any regularisation parameter are retained as stable associations. Conversely, those with a low probability of selection are dropped.

3 Results

3.1 Models inspection

To evaluate the information encoded in the various latent spaces of the trained models, we employ the RSA [Kriegeskorte et al., 2008] tool introduced in Supplementary B. The RSA results in Table 1 depict the relationships between different latent spaces (eCRF, Joint, and ROI) and each eCRF score, alongside other relevant covariates like age, sex, and image acquisition site. In short, we computed each of these Kendall τ correlations for every $N = 100$ splits and corresponding $n_E = 20$ models. The reported correlations $\bar{\tau}$ are averaged across these $N * n_E$ values. See Supplementary B for further details.
Notably, each eCRF score strongly correlates with the eCRF-specific latent space, as highlighted in pink in Table 1. This is noteworthy since the models successfully learned in a single one-dimensional space, capturing a significant amount of variability associated with all eCRF view scores. This outcome is expected, considering the correlations between the eCRF scores (see Supplementary Fig. S1). Additionally, it underscores that this latent space is not informing about age, sex, or image acquisition site. In contrast, the latent space specific to the ROI view does not significantly correlate with the eCRF scores with the only exception of ARI, yet with a small correlation. Note that it significantly correlates with age, sex, and image acquisition site, as shown in the cells highlighted in orange. Finally, we highlight in salmon in Table 1 what the joint latent space has learned. These representations moderately and significantly correlate with each score but not with acquisition site or sex. Moreover, it appears to correlate with age. It seems that the brain-behaviour relationships, modeled in the MoPoE-VAE joint latent space, relies on information related to age.

3.2 Stability selection from r-DAA

By increasing the number of models and applying a stability selection procedure, we aim to increase the stability of the sets of brain-behaviour associations supported by our multimodal dataset. In the following, we consider the SRS score associations with the thickness metric. Figure 2 illustrates the associations trajectories taken by the considered 148 ROIs. This illustration is inspired by the figure of merit proposed in the seminal paper

[Meinshausen and BÃijhlmann, 2010]. In Figure 2(a), we display trajectories of aggregated β values (i.e., effect sizes), using a mean function for f, against the number of model n_E. Some trajectories become more prominent than others as the number of models increases. However, they are hardly distinguishable from other trajectories, which remain densely packed with low values. This general aspect is displayed here for one of the $N = 100$ stability selection splits. This suggests that we can isolate some, but not all, stable associations by increasing the number of models aggregated in the r-DAA.

Figure 2(b) presents the stability paths Π for each ROI with respect to the number of models n_E. This probability is computed as described in Eq. 1 and summarises the $N = 100$ stability selection splits. Here, the aggregation function f is the mean function. The results highlights that the most stable ROIs, characterised by higher probabilities Π, are easily identifiable. The stable paths reach a plateau with the number of models (i.e., the considered regularisation hyperparameter) between 5 and 10 models, depending on the considered ROI.

Figure 2(c) is similar to Figure 2(b), but the procedure uses the RSA-weighted average as the aggregation function f instead of a simple average (see Supplementary Alg. D.2). The results suggest that stability is only slightly affected by the choice of the aggregation function f. We keep the RSA-weighted average because it may be more robust to outlier models. We apply a threshold of $\pi_{\text{thr}} = 0.4$ (as illustrated by the dashed red line) to retain stable associations, displayed in colors. These colors correspond to the same ROIs throughout the different plots. Comparing Figure 2(b) or Figure 2(c) with Figure 2(a) reveals that a high coefficient amplitude is not always a good indicator of stability. In fact, the ROIs represented by pink or light purple paths in Figure 2(a) are indistinguishable from other black dotted lines (i.e., unselected ROIs). However, these ROIs convey some of the most stable associations ($\Pi \simeq 0.7$ when $n_E = 20$). Note that the variability of the SRS score is not the best captured by our models on our dataset (see Table 1). To test the influence of the selected score and metric on the stability paths, we generate similar figures by examining the association coefficients between the SDQ-ha score and the area metric, as well as the stability paths. Very similar observations can be made, as shown in Supplementary E.

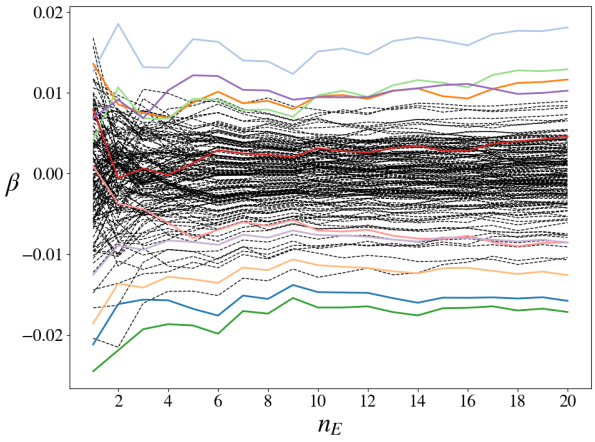
3.3 Transdiagnostic-factor spatial support

At the core of our approach is a MoPoE-VAE network designed to construct a latent space consolidating joint information between the entire ROI view and the entire eCRF view, along with ROI- and eCRF-specific latent spaces. The interpretation of joint latent representations amounts to extracting brain-behaviour associations that effectively represent this joint information. This strategy supports the research on the general psychopathology dimension introduced by Caspi [Caspi et al., 2014, Caspi et al., 2023]. In line with these ideas, our goal is to identify the brain regions and metrics that underpin these similarities. These brain-behaviour associations retained by our approach for each score-metric pair (s, m) are listed in see Supplementary F.

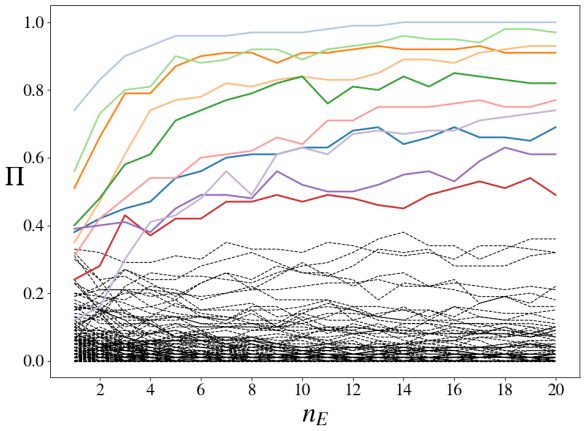
In this section, we focus on transdiagnostic associations whose ROIs exhibit co-association with exactly four specific scores. These regions could be considered as components of a transdiagnostic spatial support for mental disorders. We focus on four specific scores of interest: SRS, which assesses social interaction and is correlated with autism disorder; SCARED, which assesses fear and anxiety issues and is correlated with anxiety disorders; SDQ-ha, which is related to hyperactivity disorders; and CBCL-wd, an indicator of depression, a symptom common to most of these pathologies. For each cortical measure, Figure 3 highlights the ROIs found to be associated with these four scores. Among these selected transdiagnostic regions, many belong to the pericalosal and cingulate regions. These regions, whether considered with the area or curvature metrics, are consistently associated with each examined score. This means that the associations between selected ROIs and scores show similar covariations. In the identified pericalosal and cingulate regions, a decrease in both area and curvature metrics is associated with an increase in the SRS, SCARED, or CBCL-wd scores. However, the opposite is observed for the SDQ-ha score (Figure 3(b)-(c)). Looking at the thickness metric and grouping the curvature and area metrics, we find two disjoint sets of regions among the selected transdiagnostic regions. First, the left and right occipital poles for the thickness metric. Second, the cingulate regions for the curvature/area metrics. Only the right cingulate mid / posterior region (R.Ci.Mid.Post.) and left cingulate posterior dorsal gyrus (L.G.Ci.Post.Dors) seem to be related to the four scores for both metrics. Finally, looking at the SDQ-ha score and comparing to the grouped SRS, SCARED and CBCL-wd scores, we find that the associations identified have systematically opposite signs. Overall, the transdiagnostic regions are mostly bilateral and present a spatially smooth pattern of association with the eCRF scores.

3.4 Transdiagnostic-factor spatial support in ASD and ADHD

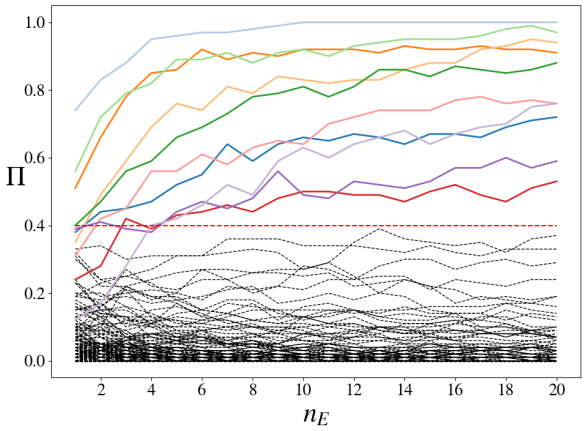
Recall that our method, with the presented setting, is designed to find transdiagnostic associations, in line with the general psychopathology dimension. Among the structural biomarkers involved in ASD (or ADHD), we can study the ones that can attributed to transdiagnostic factors. In this section, we focus on transdiagnostic associations whose



(a) Coefficients



(b) Stability paths using uniform ensembling

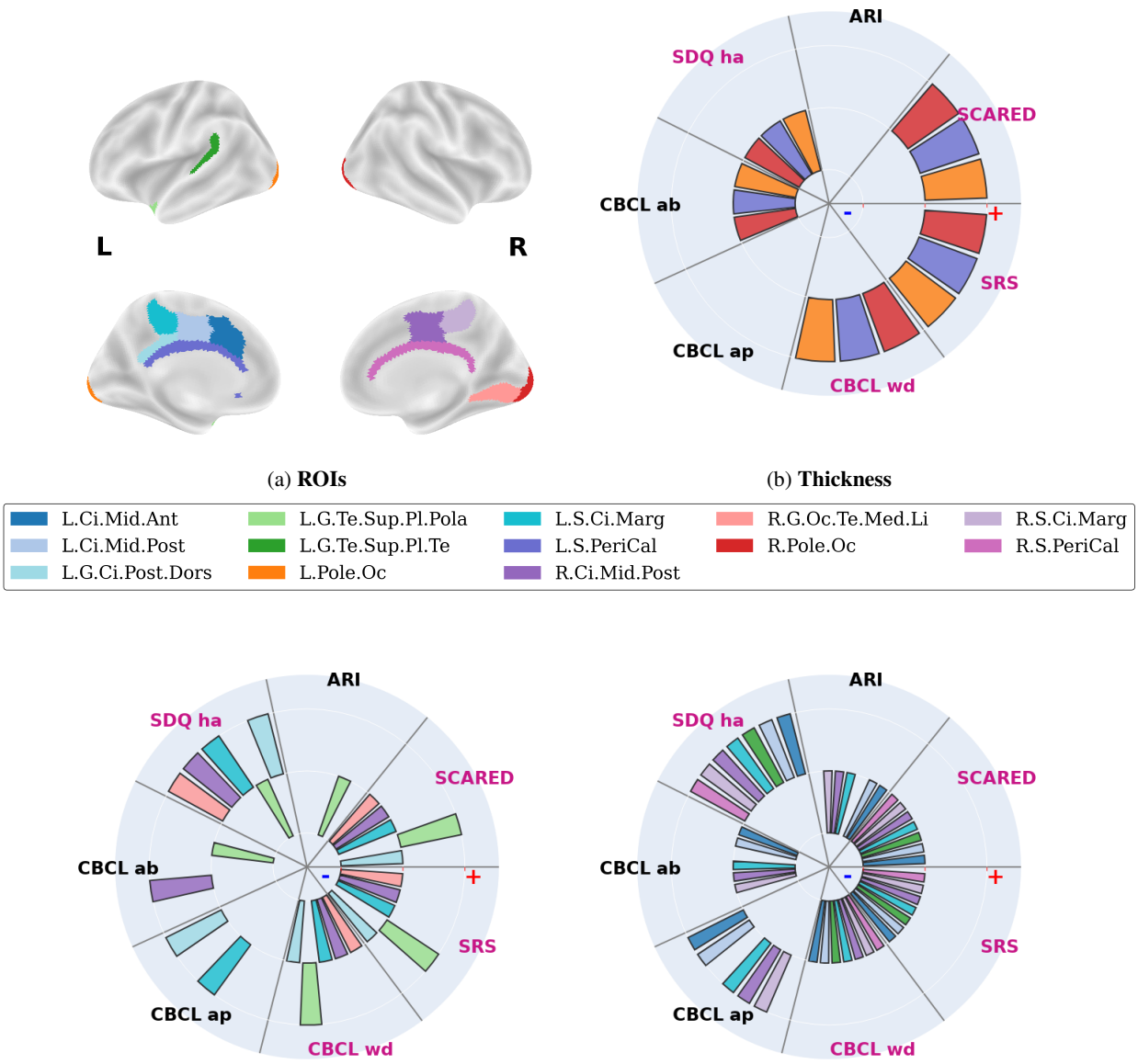


(c) Stability paths using weighted ensembling

Figure 2: Investigation of the ROIs associated with SRS in thickness. Each line corresponds to a ROI, dotted black ones are not selected as below the threshold $\pi_{\text{thr}} = 0.4$ when using as aggregation function a mean weighted by RSA correlations. The ROIs thus selected are colored and the colors are consistent across the plots. The red horizontal dotted line highlights the threshold $\pi_{\text{thr}} = 0.4$. (a) Mean coefficients aggregated across models for a given split over the $N = 100$ splits, plotted against the number of models. (b) ROIs stability paths Π when using an uniformly weighted mean to aggregate the coefficients output from DAA on each model in the r-DAA, against the number of models used. (c) Same as the last one, except that the mean was weighted using RSA correlations. The latter strategy is used to select the ROIs associated in thickness with the SRS score, with $\Pi > \pi_{\text{thr}} = 0.4$.

ROIs are specifically associated with one symptom/score. We perform this analysis for two scores. First, the SRS score, which is considered as a proxy for ASD [Constantino et al., 2003]. Although the SRS score alone may not be sufficient, clinicians often use it to determine a patient's status. This score has also a high correlation with diagnosis (unpublished results based on our databases) in datasets adhering to diagnosis-balanced inclusion criteria such as ABIDE I (SRS-1, 0.82), ABIDE II (SRS-1, 0.86 and SRS-2, 0.72), or EU-AIMS (SRS-2, 0.85). With the same general caveats, we use the SDQ-ha score as a proxy for ADHD-related symptoms and diagnosis [Goodman, 1997]. The retained associations can be found in Supplementary F and are summarised below.

Markers related to SRS (ASD). In Figure 4, we display ROIs associated in thickness and area with the score SRS. Interestingly, the selected ROIs display a rather symmetrical pattern. With the thickness metric, the SRS appears to positively covary with the bilateral occipital poles, bilateral pericallosal region, and a right prefrontal region, while negatively covarying with the bilateral temporal sulci and right postcentral region. With area metrics, the SRS appears to positively covary with the left prefrontal cortex, left subcallosal gyrus and bilateral circular inferior insular sulci, while negatively covarying with the bilateral cingulate and left parieto-temporal cortex.

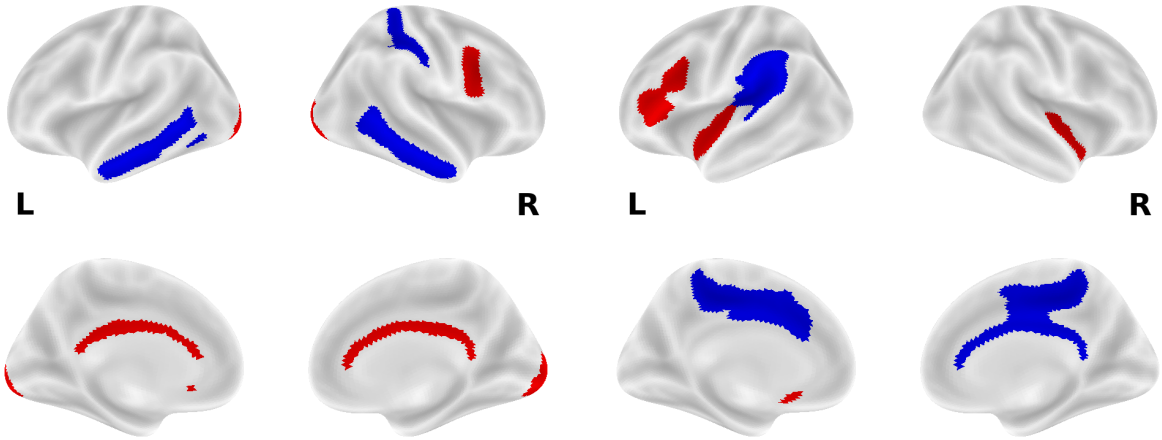


(a) **ROIs** (b) **Thickness**

(c) **Curvature** (d) **Area**

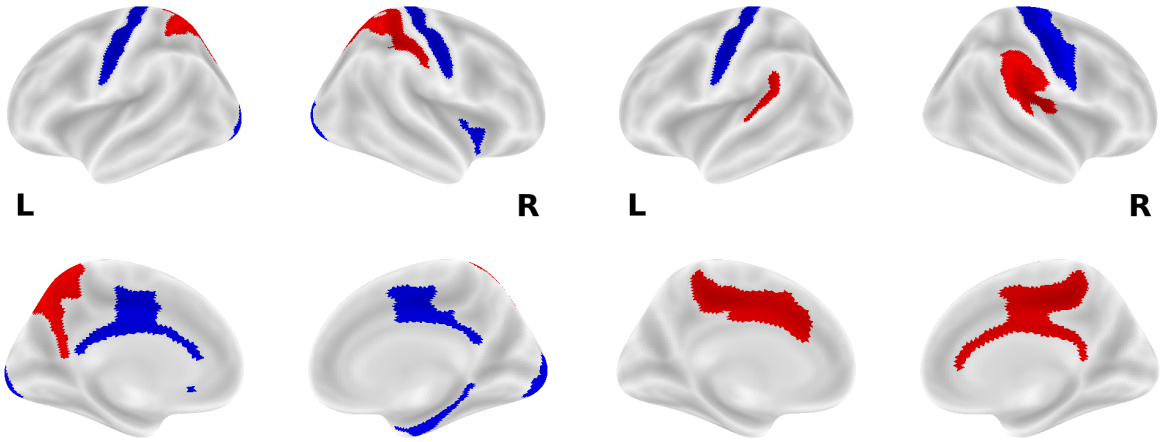
Figure 3: Display of ROIs associated with selected eCRF scores for a transdiagnostic assessment, e.g. SRS, SDQ-ha, SCARED and CBCL-wd, for each cortical measure. (a) ROIs color-coded, corresponding to color in the following polar plots. Lateral and medial view are displayed for each hemisphere (left hemisphere on the left hand side and right hemisphere on the right). The polar plots indicate the sign of each association (negative when the bar is close to the center, postive when it is close to the edge of the circle). Only retained associations are displayed. Each polar plot correspond to a metric : (b) displays assocations in thickness, (c) associations in curvature and (d) in area. L: left, R: right, Ci: cingulate, Oc: occipital, Te: temporal, S: sulcus, G: gyrus, Mid: middle, Post: posterior, Dors: dorsal, med: medial, Inf: inferior, Marg: marginal, Ant: anterior, PeriCal: pericallosal, PostCe: postcentral, Sup: superior, Li: lingual, Pl: plan, Pola: polar.

Marker related to SDQ (ADHD). In Figure 5, we display ROIs found associated in thickness and area with the score SDQ-ha. Once again the figure displays symmetrical patterns. With the thickness metric, the SDQ-ha appears to



(a) **Thickness** (b) **Area**

Figure 4: Display of ROIs associated with the SRS score.



(a) **Thickness** (b) **Area**

Figure 5: Display of ROIs associated with the SDQ-ha hyperactivity score.

positively covary notably with postcentral and superior parietal regions, including the left precuneus, while negatively covarying with central sulci, cingulate and occipital cortices. With the area metric, the SDQ-ha appears to positively covary notably with bilateral parieto-temporal cortex and bilateral cingulate cortex, while negatively covarying with bilateral central sulci and right precentral gyrus. Of note, the central sulci appear affected by the SDQ-ha score in the same direction for both area and thickness metrics.

4 Discussion

In this paper, we present an interpretability method dedicated to deep learning-based integration models using mVAEs. We demonstrate its application in the study of the transdiagnostic dimension within the HBN cohort, integrating neuroimaging data and psychological assessments in an at-risk population with notable behavioural symptoms. Our method is endowed with a stability selection procedure to retain associations and we investigate its effects. A sufficient number of models is the only need to achieve stability. Finally, with only a prior on the expected number of associations, our method enables the identification of stable associations between measures of the brain cortical surface and symptom scores.

Deep mVAE model are expressive multi-modal integration tools

We take advantage of the versatile definition of the latent space in the MoPoE-VAE to choose a representation setting that naturally disentangles specific and shared sources of variation between view-specific and joint latent spaces. Other work already identified the disentanglement achieved with this type of architecture [Lee and Pavlovic, 2021, Qiu et al., 2022], but we apply it here to neuroimaging and score questionnaires integration. We also leverage this source separation using our interpretation module based on digital avatars. By mitigating or eliminating confounding effects such as the MRI acquisition site in shared representation, we expect interpretations to be unaffected, without performing any standardisation or harmonisation of the data prior to learning. Moreover, the joint latent spaces that significantly correlate with age highlight how our models handle age information, although it is never explicitly provided. This contrasts with conventional approaches using age residualisation.

Our framework based on the MoPoE-VAE can conveniently handle incomplete data, a common issue in multi-view integration. The requirement of complete data often significantly reduces the number of subjects that can be used and impairs the statistical power of most studies. This resilience made it possible for us to use an openly available cohort with minimal missing data control and include nearly all available subjects. This has provided us with a substantial sample size collected from multiple acquisition sites. This characteristic should improve reproducibility and replicability when employing such an approach in multi-modal neuroimaging studies, aligning with state-of-the-art guidelines [Klapwijk et al., 2021].

We also show that MoPoE-VAE networks, when equipped with a stable interpretation module, can provide associations between variables of different views. This stable interpretation module can be configured with only one a priori parameter which is the expected number of associations ($n_{\text{select}} = 12$ here).

Deep mVAE models for transdiagnostic studies

Transdiagnostic researches [Fusar-Poli et al., 2019] focus a common risk factor of pathology in psychiatry, often referred to as p-factor [Caspi et al., 2014], computed using various clinical assessments and summarised in a unique dimension. These approaches address the fact that many psychiatric illnesses share common symptoms and comorbidities, and that the heterogeneity of the population cohorts makes it difficult to obtain reproducible results based on a single diagnosis. Studies hypothesise a common etiology, that would explain some underlying mechanisms common to multiple psychiatric disorders. They usually express this global pathological variability as p factor, and then inspect its association with brain imaging to find potential biomarkers for these common mechanisms.

Deep learning MoPoE-VAE models are trained from data in an at-risk population cohort in which subjects are assessed with questionnaires, expressing symptoms of several psychiatric syndromes (SRS, SDQ-ha, etc). This MoPoE-VAE training is organised to integrate all available data to create distinct representations of shared and specific information. This shared representation contains multivariate variability linking multiple symptom scores to imaging. Our stable interpretation pipeline inspects this joint latent space to produce transdiagnostic associations between behaviour and cortical measures. By displaying regions linked with multiple symptom scores, we capture regions often reported as associated with transdiagnostic factors in the literature.

In particular, results from our transdiagnostic perspective in section 3.3 highlight fairly symetrical regions, mostly associated in area and curvature in cingulate regions and in thickness in the occipital poles. Studies in functional MRI already identified cingulo-opercular network (CON) [Sheffield et al., 2017] and default mode network (DMN), and particularly the anterior cingulate areas [Gong et al., 2016, MacNamara et al., 2017] as related to transdiagnostic factors. Other fMRI studies identified anterior / middle cingulate and occipital areas, among other, associated with general p-factor [Elliott et al., 2018], as well as in studies of transdiagnostic population with other factors [Feldker et al., 2017, Tong et al., 2022], and in particular with ASD and ADHD [Bush et al., 1999, D'Cruz et al., 2016, Lukito et al., 2020].

Alongside fMRI studies, many studies report structural alterations associations with transdiagnostic factors, such as changes in cortical thickness or grey matter volume of the cingulate and occipital cortices, as part of more global patterns [Clementz et al., 2016, Yin et al., 2022, Parkes et al., 2021]. Such findings were highlighted as well in specific ASD [Oblak et al., 2010, Chien et al., 2021, Mihailov et al., 2020, Ecker et al., 2022] or ADHD [Amico et al., 2011, He et al., 2015, Bayard et al., 2020] studies, as well as research studying these diseases jointly [Rommelse et al., 2017, Lukito et al., 2020]. Moreover, occipital lobe grey matter reduction was specifically identified as linked with increases in p-factor [Romer et al., 2018, Romer et al., 2021]. Finally, cingulate related white matter tracks fractional anisotropy extracted from diffusion MRI were found associated with transdiagnostic factors as well [Stefanik et al., 2018].

Cingulate regions are well-known to be implicated in cognition, emotion processing [Bush et al., 2000, Vogt et al., 1992, Denson et al., 2009], while occipital regions are responsible for primary visual processing. Since these functions are often altered when expressing psychiatric symptoms, their implication in a common mechanism

to multiple syndromes is very likely. Our results support the hypothesis that cingulate and occipital regions and their related functional or structural networks are important in general psychopathology and warrant further investigation.

5 Perspectives

The method described in this paper focuses on an integrative multi-view approach based on deep learning and featuring interpretation capacity. We discussed the main benefits of its use and we list below some perspectives related to some limitations of the proposed approach, while highlighting more general remarks regarding the integration of the ever growing volume and variety of available data in neuroscience.

Effect size. In our transdiagnostic study, the stable associations found are characterised by rather small effect sizes. This underscores the difficulty of exploring the common processes that hypothetically contribute to the etiology of multiple psychopathologies. While individual effect sizes may be modest, the cumulative evidence supports the existence of transdiagnostic factors. One could expect an increase of the observed effect sizes by considering functional MRI (fMRI) data instead of structural MRI. Association studies using fMRI offer numerous advantages, including direct measurement of brain activity and localisation of function. Current research with fMRI shows strong associations compared to structural MRI [Oblong et al., 2023] and would be worth investigating using our tool.

Data harmonisation. Several recent works have shown that machine learning models are strongly biased by the MRI acquisition site and do not generalise well to new MRI images from sites that have never been seen before [Glocker et al., 2019, Wachinger et al., 2021]. This problem is due to differences in scanner manufacturers, specifications, settings, and hardware. While traditional residualisation technique applied to remove the site effect marginally improves the performance of machine learning models, it does not bring any improvement for deep learning models [Dufumier et al., 2024]. In this work, we show that the influence of non-interest factors, in particular the site effect, can be effectively eliminated by disentangling modality-specific and joint latent representations. Nevertheless data harmonisation remains an ongoing research area [Dinsdale et al., 2021, Bashyam et al., 2022]. For example, the OpenBHB challenge [Dufumier et al., 2022] on brain age prediction with site effect removal could bring new harmonisation techniques developed by the community.

Views with different internal correlation structure. As mentioned above, a wide variety of assessments are available in large population cohorts. Most often, each type of data is exploited separately for regression, classification and segmentation. Deep learning has been extensively used on imaging data and has demonstrated considerable benefits over traditional methods for some tasks, such as the segmentation of medical anomalies and anatomical structures [Menze et al., 2015, Henschel et al., 2020]. But in the medical field, many assessments, such as imaging or genomic data, can be more naturally represented as graphs due to their underlying biological properties (e.g. functional / structural connectomes or protein interactions). DL specific operators [Bronstein et al., 2017, Ghosal et al., 2022] have been developed to take into account such correlation structure when learning from such data. In the current study, we integrate cortical measures on ROIs as tabular data. It would be highly relevant to return to brain surface data and model their actual correlation structure using dedicated convolution operators [Zhao et al., 2019]. The same applies to other available modalities such as genomic data.

Dimensional discrepancy of the views. The demonstrated integrative capacity of mVAEs comes from modelling modality-specific and joint latent spaces. Handling views with very different number of variables is an ongoing research question. For example, genotyping data may have millions of variables. Training a mVAE in such setting is challenging. Intuitively, learned modality specific latent spaces will have different sizes, somehow proportional to their input size, but this is not so clear for the joint representations. Pioneering work defines a sparse multichannel VAE [Antelmi et al., 2019]. It leverages a variational dropout regularisation [Molchanov et al., 2017] that identifies an optimal number of joint latent dimensions.

Disentanglement. In the development of digital avatars, transparency of the trained NN models is a key feature. Importantly, the chosen mVAE model must learn disentangled representations that separate modality specific and joint variability. Indeed, confounding factor like site effect are usually specific to one modality, and such disentanglement would ensure the joint latent is devoid of such unwanted effects. Recent works propose contrastive VAE which uses deep encoders to capture higher-level semantics [Aglinskas et al., 2022], as compared to its former linear conterpart contrastive PCA [Abid et al., 2018]. From two encoders, they typically structure the learned latent space into two parts containing the background (i.e., common to the studied population) and salient (i.e., specific to a pathology) variabilities. Followup works show that classical VAE losses alone can not effectively separate joint and salient variability, and that further constraints and regularisations are needed to satisfy the assumptions of the generative process and to promote disentanglement in the latent space [Abid and Zou, 2019, Choudhuri et al., 2020, Weinberger et al., 2022, Zou et al., 2022]. In the case of mVAEs, the same observations hold. Modeling modality-specific and joint latent spaces somehow separates joint from specific

variability [Lee and Pavlovic, 2021], as observed in this study. However, to ensure disentanglement, additional regularisations during training are needed [Daunhawer et al., 2021].

Replication. The design of a replication study will contribute to the generalisability of our findings. In particular, it would be interesting to apply our approach to cohorts such as the Dunedin Longitudinal Study [Caspi et al., 2014, Romer et al., 2021], the ABCD Study [Casey et al., 2018, Karcher and Barch, 2021] or the Duke Neurogenic Study [Romer et al., 2018, Elliott et al., 2018]. These studies were not explicitly focused on transdiagnostic research, but their design, including comprehensive assessments with a multidisciplinary context, has provided valuable insights into the understanding of psychiatric disorders from a transdiagnostic perspective.

Data and Code availability

The code to reproduce the experiments and results is made available here. The data is accessible at `https://fcon_1000.projects.nitrc.org/` .

Author Contributions

C.A. designed the experiments, C.A. conducted the experiments, V.F. et A.G. provided critical feedback, V.F. supervised the project, A.G. pre-processed the data, C.A, V.F. and A.G. redacted the manuscript, all authors reviewed the manuscript.

References

[Abid et al., 2018] Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. (2018). Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134.

[Abid and Zou, 2019] Abid, A. and Zou, J. Y. (2019). Contrastive variational autoencoder enhances salient features. *CoRR*, abs/1902.04601.

[Aglinskas et al., 2022] Aglinskas, A., Hartshorne, J. K., and Anzellotti, S. (2022). Contrastive machine learning reveals the structure of neuroanatomical variation within autism. *Science*, 376(6597):1070–1074.

[Alexander et al., 2017] Alexander, L. M. et al. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4(1):170181.

[Ambroise et al., 2023] Ambroise, C., Grigis, A., Duchesnay, E., and Frouin, V. (2023). Multi-view variational autoencoders allow for interpretability leveraging digital avatars: Application to the hbn cohort. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

[Ambroise et al., 2021] Ambroise, C., Mihailov, A., Frouin, V., and Grigis, A. (2021). Multi-modal latent variable model could help individuals stratification: application to hbn cohort. In *OHBM*.

[Amico et al., 2011] Amico, F., Stauber, J., Koutsouleris, N., and Frodl, T. (2011). Anterior cingulate cortex gray matter abnormalities in adults with attention deficit hyperactivity disorder: A voxel-based morphometry study. *Psychiatry Research: Neuroimaging*, 191(1):31–35.

[Andrew et al., 2013] Andrew, G., Arora, R., Bilmes, J., and Livescu, K. (2013). Deep canonical correlation analysis. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1247–1255, Atlanta, Georgia, USA. PMLR.

[Anonymous, 2023] Anonymous (2023). Benchmarking multimodal variational autoencoders: Cdsprites+ dataset and toolkit. In *ICLR*.

[Antelmi et al., 2019] Antelmi, L. et al. (2019). Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 302–311. PMLR.

[Association, 2013] Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5âĎć*. American Psychiatric Publishing, a division of American Psychiatric Association, Washington, DC ;, 5th edition. edition.

[Baldassarre et al., 2017] Baldassarre, L., Pontil, M., and MourÃčo-Miranda, J. (2017). Sparsity is better with stability: Combining accuracy and stability for model selection in brain decoding. *Frontiers in Neuroscience*, 11.

[Bashyam et al., 2022] Bashyam, V. M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Singh, A., Habes, M., Fan, Y., Masters, C. L., Maruff, P., Zhuo, C., VÃűlzke, H., Johnson, S. C., Fripp, J., Koutsouleris, N., Satterthwaite, T. D., Wolf, D. H., Gur, R. E., Gur, R. C., Morris, J. C., Albert, M. S., Grabe, H. J., Resnick, S. M., Bryan, N. R., Wittfeld, K., BÃijlow, R., Wolk, D. A., Shou, H., Nasrallah, I. M., Davatzikos, C., and The iSTAGING and PHENOM consortia (2022). Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *Journal of Magnetic Resonance Imaging*, 55(3):908–916.

[Bayard et al., 2020] Bayard, F., Nymberg Thunell, C., Abé, C., Almeida, R., Banaschewski, T., Barker, G., Bokde, A. L. W., Bromberg, U., Büchel, C., Quinlan, E. B., Desrivières, S., Flor, H., Frouin, V., Garavan, H., Gowland, P., Heinz, A., Ittermann, B., Martinot, J.-L., Martinot, M.-L. P., Nees, F., Orfanos, D. P., Paus, T., Poustka, L., Conrod, P., Stringaris, A., Struve, M., Penttilä, J., Kappel, V., Grimmer, Y., Fadai, T., van Noort, B., Smolka, M. N., Vetter, N. C., Walter, H., Whelan, R., Schumann, G., Petrovic, P., and the IMAGEN Consortium (2020). Distinct brain structure and behavior related to adhd and conduct disorder traits. *Molecular Psychiatry*, 25(11):3020–3033.

[Bronstein et al., 2017] Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42.

[Bryk and Raudenbush, 1992] Bryk, A. and Raudenbush, S. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications.

[Bush et al., 1999] Bush, G., Frazier, J. A., Rauch, S. L., Seidman, L. J., Whalen, P. J., Jenike, M. A., Rosen, B. R., and Biederman, J. (1999). Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fmri and the counting stroop. *Biological Psychiatry*, 45(12):1542–1552.

[Bush et al., 2000] Bush, G., Luu, P., and Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences*, 4(6):215–222.

[Cao et al., 2011] Cao, K. L., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinform.*, 12:253.

[Casey et al., 2018] Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., Chaarani, B., Mejia, M. H., Hagler Jr., D. J., Cornejo, M. D., Sicat, C. S., Harms, M. P., Dosenbach, N. U. F., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J. R., Kuperman, J. M., Fair, D. A., and Dale, A. M. (2018). The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54.

[Caspi et al., 2014] Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., and Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, 2(2):119–137. PMID: 25360393.

[Caspi et al., 2023] Caspi, A., Houts, R. M., Fisher, H. L., Danese, A., and Moffitt, T. E. (2023). The general factor of psychopathology (p): Choosing among competing models and interpreting p. *Clinical Psychological Science*, 12(1):53–82. Publisher: SAGE Publications Inc.

[Chegraoui et al., 2023] Chegraoui, H., Guillemot, V., Rebei, A., Gloaguen, A., Grill, J., Philippe, C., and Frouin, V. (2023). Integrating multiomics and prior knowledge: a study of the Graphnet penalty impact. *Bioinformatics*, 39(8):btad454.

[Chien et al., 2021] Chien, Y.-L., Chen, Y.-C., and Gau, S. S.-F. (2021). Altered cingulate structures and the associations with social awareness deficits and cntnap2 gene in autism spectrum disorder. *NeuroImage: Clinical*, 31:102729.

[Choudhuri et al., 2020] Choudhuri, A., Makkuva, A. V., Rana, R., Oh, S., Chowdhary, G., and Schwing, A. (2020). Towards principled objectives for contrastive disentanglement.

[Clementz et al., 2016] Clementz, B. A., Sweeney, J. A., Hamm, J. P., Ivleva, E. I., Ethridge, L. E., Pearlson, G. D., Keshavan, M. S., and Tamminga, C. A. (2016). Identification of distinct psychosis biotypes using brain-based biomarkers. *American Journal of Psychiatry*, 173(4):373–384. PMID: 26651391.

[Constantino et al., 2003] Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., Metzger, L. M., Shoushtari, C. S., Splinter, R., and Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *J. Autism Dev. Disord.*, 33(4):427–433.

[Dale et al., 1999] Dale, A. M. et al. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194.

[Daunhawer et al., 2022] Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., and Vogt, J. E. (2022). On the limitations of multimodal vaes. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

[Daunhawer et al., 2021] Daunhawer, I., Sutter, T. M., MarcinkeviÄ■s, R., and Vogt, J. E. (2021). Self-supervised Disentanglement of Modality-Specific and Shared Factors Improves Multimodal Generative Models. In Akata, Z., Geiger, A., and Sattler, T., editors, *Pattern Recognition*, volume 12544, pages 459–473. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

[D'Cruz et al., 2016] D'Cruz, A.-M., Mosconi, M. W., Ragozzino, M. E., Cook, E. H., and Sweeney, J. A. (2016). Alterations in the functional neural circuitry supporting flexible choice behavior in autism spectrum disorders. *Translational Psychiatry*, 6(10):e916–e916.

[Denson et al., 2009] Denson, T. F., Pedersen, W. C., Ronquillo, J., and Nandy, A. S. (2009). The Angry Brain: Neural Correlates of Anger, Angry Rumination, and Aggressive Personality. *Journal of Cognitive Neuroscience*, 21(4):734–744.

[Destrieux et al., 2010] Destrieux, C. et al. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53(1):1–15.

[Dinsdale et al., 2021] Dinsdale, N. K., Jenkinson, M., and Namburete, A. I. (2021). Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage*, 228:117689.

[Dufumier et al., 2024] Dufumier, B., Gori, P., Petiton, S., Louiset, R., Mangin, J.-F., Grigis, A., and Duchesnay, E. (2024). Exploring the potential of representation and transfer learning for anatomical neuroimaging: application to psychiatry. working paper or preprint.

[Dufumier et al., 2022] Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., and Duchesnay, E. (2022). Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. *NeuroImage*, 263:119637.

[Ecker et al., 2022] Ecker, C., Pretzsch, C. M., Bletsch, A., Mann, C., Schaefer, T., Ambrosino, S., Tillmann, J., Yousaf, A., Chiocchetti, A., Lombardo, M. V., Warrier, V., Bast, N., Moessnang, C., Baumeister, S., DellâĂŹAcqua, F., Floris, D. L., Zabihi, M., Marquand, A., Cliquet, F., Leblond, C., Moreau, C., Puts, N., Banaschewski, T., Jones, E. J., Mason, L., Bölte, S., Meyer-Lindenberg, A., Persico, A. M., Durston, S., Baron-Cohen, S., Spooren, W., Loth, E., Freitag, C. M., Charman, T., Dumas, G., Bourgeron, T., Beckmann, C. F., Buitelaar, J. K., and Murphy, D. G. (2022). Interindividual differences in cortical thickness and their genomic underpinnings in autism spectrum disorder. *American Journal of Psychiatry*, 179(3):242–254. PMID: 34503340.

[Elliott et al., 2018] Elliott, M. L., Romer, A., Knodt, A. R., and Hariri, A. R. (2018). A connectome-wide functional signature of transdiagnostic risk for mental illness. *Biological Psychiatry*, 84(6):452–459. Translating Biology to Treatment in Schizophrenia.

[Feldker et al., 2017] Feldker, K., Heitmann, C. Y., Neumeister, P., Tupak, S. V., Schrammen, E., Moeck, R., Zwitserlood, P., Bruchmann, M., and Straube, T. (2017). Transdiagnostic brain responses to disorder-related threat across four psychiatric disorders. *Psychological Medicine*, 47(4):730âĂŞ743.

[Fusar-Poli et al., 2019] Fusar-Poli, P., Solmi, M., Brondino, N., Davies, C., Chae, C., Politi, P., Borgwardt, S., Lawrie, S. M., Parnas, J., and McGuire, P. (2019). Transdiagnostic psychiatry: a systematic review. *World Psychiatry*, 18(2):192–207.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

[Ghosal et al., 2022] Ghosal, S., Chen, Q., Pergola, G., Goldman, A. L., Ulrich, W., Weinberger, D. R., and Venkataraman, A. (2022). A biologically interpretable graph convolutional network to link genetic risk pathways and imaging phenotypes of disease. In *International Conference on Learning Representations*.

[Glocker et al., 2019] Glocker, B., Robinson, R., Castro, D. C., Dou, Q., and Konukoglu, E. (2019). Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects.

[Gong et al., 2016] Gong, Q., Hu, X., Pettersson-Yeo, W., Xin Xu, S., Crossley Karmelic, N., Min Wu, H., and Mechelli, A. (2016). Network-level dysconnectivity in drug-naïve first-episode psychosis: Dissociating transdiagnostic and diagnosis-specific alterations. *Neuropsychopharmacology*.

[Goodman, 1997] Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38(5):581–586.

[Gyawali et al., 2022] Gyawali, P. K., Liu, X., Zou, J., and He, Z. (2022). Ensembling improves stability and power of feature selection for deep learning models. In Knowles, D. A., Mostafavi, S., and Lee, S.-I., editors, *Proceedings of the 17th Machine Learning in Computational Biology meeting*, volume 200 of *Proceedings of Machine Learning Research*, pages 33–45. PMLR.

[He et al., 2015] He, N., Li, F., Li, Y., Guo, L., Chen, L., Huang, X., Lui, S., and Gong, Q. (2015). Neuroanatomical deficits correlate with executive dysfunction in boys with attention deficit hyperactivity disorder. *Neuroscience Letters*, 600:45–49.

[Helmer et al., 2023] Helmer, M., Warrington, S., Mohammadi-Nejad, A.-R., Ji, J. L., Howell, A., Rosand, B., Anticevic, A., Sotiropoulos, S. N., and Murray, J. D. (2023). On stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *bioRxiv*.

[Henschel et al., 2020] Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., and Reuter, M. (2020). Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012.

[HOMMEL, 1988] HOMMEL, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386.

[Hotelling, 1936] Hotelling, H. (1936). RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, 28(3-4):321–377.

[Ing et al., 2019] Ing, A., Sämann, P. G., Chu, C., Tay, N., Biondo, F., Robert, G., Jia, T., Wolfers, T., Desrivières, S., Banaschewski, T., Bokde, A. L. W., Bromberg, U., Büchel, C., Conrod, P., Fadai, T., Flor, H., Frouin, V., Garavan, H., Spechler, P. A., Gowland, P., Grimmer, Y., Heinz, A., Ittermann, B., Kappel, V., Martinot, J.-L., Meyer-Lindenberg, A., Millenet, S., Nees, F., van Noort, B., Orfanos, D. P., Martinot, M.-L. P., Penttilä, J., Poustka, L., Quinlan, E. B., Smolka, M. N., Stringaris, A., Struve, M., Veer, I. M., Walter, H., Whelan, R., Andreassen, O. A., Agartz, I., Lemaitre, H., Barker, E. D., Ashburner, J., Binder, E., Buitelaar, J., Marquand, A., Robbins, T. W., Schumann, G., and IMAGEN Consortium (2019). Identification of neurobehavioural symptom groups based on shared brain mechanisms. *Nature Human Behaviour*, 3(12):1306–1318.

[Insel et al., 2010] Insel, T. et al. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. *Am J Psychiatry*, 167(7):748–751.

[Karcher and Barch, 2021] Karcher, N. R. and Barch, D. M. (2021). The abcd study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology*, 46(1):131–142.

[Kendall and Gal, 2017] Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

[Kiureghian and Ditlevsen, 2009] Kiureghian, A. D. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.

[Klapwijk et al., 2021] Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M., and Mills, K. L. (2021). Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Developmental Cognitive Neuroscience*, 47:100902.

[Kramer, 1991] Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *Aiche Journal*, 37:233–243.

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

[Labus et al., 2015] Labus, J. S., Van Horn, J. D., Gupta, A., Alaverdyan, M., Torgerson, C., Ashe-McNalley, C., Irimia, A., Hong, J.-Y., Naliboff, B., Tillisch, K., and Mayer, E. A. (2015). Multivariate morphological brain signatures predict patients with chronic abdominal pain from healthy control subjects. *PAIN*, 156(8).

[Lakshminarayanan et al., 2017] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[Lee and Pavlovic, 2021] Lee, M. and Pavlovic, V. (2021). Private-Shared Disentangled Multimodal VAE for Learning of Latent Representations. In *CVPR*, pages 1692–1700.

[Lukito et al., 2020] Lukito, S., Norman, L., Carlisi, C., Radua, J., Hart, H., Simonoff, E., and Rubia, K. (2020). Comparative meta-analyses of brain structural and functional abnormalities during cognitive control in attention-deficit/hyperactivity disorder and autism spectrum disorder. *Psychological Medicine*, 50(6):894âĂŞ919.

[MacNamara et al., 2017] MacNamara, A., Klumpp, H., Kennedy, A. E., Langenecker, S. A., and Phan, K. L. (2017). Transdiagnostic neural correlates of affective face processing in anxiety and depression. *Depression and Anxiety*, 34(7):621–631.

[Meinshausen and BÃijhlmann, 2010] Meinshausen, N. and BÃijhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

[Menze et al., 2015] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.-A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ã., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M. S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.-C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., and Van Leemput, K. (2015). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

[Mihailov et al., 2020] Mihailov, A., Philippe, C., Gloaguen, A., Grigis, A., Laidi, C., Piguet, C., Houenou, J., and Frouin, V. (2020). Cortical signatures in behaviorally clustered autistic traits subgroups: a population-based study. *Transl. Psychiatry*, 10(1):207.

[Mihalik et al., 2020] Mihalik, A., Ferreira, F. S., Moutoussis, M., Ziegler, G., Adams, R. A., Rosa, M. J., Prabhu, G., de Oliveira, L., Pereira, M., Bullmore, E. T., Fonagy, P., Goodyer, I. M., Jones, P. B., Hauser, T., Neufeld, S., Romero-Garcia, R., St Clair, M., VÃl'rtes, P. E., Whitaker, K., Inkster, B., Ooi, C., Toseeb, U., Widmer, B., Bhatti, J., Villis, L., Alrumaithi, A., Birt, S., Bowler, A., Cleridou, K., Dadabhoy, H., Davies, E., Firkins, A., Granville, S., Harding, E., Hopkins, A., Isaacs, D., King, J., Kokorikou, D., Maurice, C., McIntosh, C., Memarzia, J., Mills, H., OâĂŹDonnell, C., Pantaleone, S., Scott, J., Fearon, P., Suckling, J., van Harmelen, A.-L., Kievit, R., Shawe-Taylor, J., Dolan, R., and MourÃčo-Miranda, J. (2020). Multiple holdouts with stability: Improving the generalizability of machine learning analyses of brainâĂŞbehavior relationships. *Biological Psychiatry*, 87(4):368–376. Innovations in Clinical Neuroscience: Tools, Techniques, and Transformative Frameworks.

[Molchanov et al., 2017] Molchanov, D., Ashukha, A., and Vetrov, D. P. (2017). Variational dropout sparsifies deep neural networks. In *ICML*, pages 2498–2507.

[Nakua et al., 2023] Nakua, H., Yu, J.-C., Abdi, H., Hawco, C., Voineskos, A., Hill, S., Lai, M.-C., Wheeler, A. L., McIntosh, A. R., and Ameis, S. H. (2023). Comparing the stability and reproducibility of brain-behaviour relationships found using canonical correlation analysis and partial least squares within the ABCD sample. *bioRxivorg*.

[Oblak et al., 2010] Oblak, A. L., Gibbs, T. T., and Blatt, G. J. (2010). Decreased gabab receptors in the cingulate cortex and fusiform gyrus in autism. *Journal of Neurochemistry*, 114(5):1414–1423.

[Oblong et al., 2023] Oblong, L. M., Llera, A., Mei, T., Haak, K., Isakoglou, C., Floris, D. L., Durston, S., Moessnang, C., Banaschewski, T., Baron-Cohen, S., Loth, E., DellâĂŹAcqua, F., Charman, T., Murphy, D. G. M., Ecker, C., Buitelaar, J. K., Beckmann, C. F., Ahmad, J., Ambrosino, S., Auyeung, B., Banaschewski, T., Baron-Cohen, S., Baumeister, S., Beckmann, C. F., BÃűlte, S., Bourgeron, T., Bours, C., Brammer, M., Brandeis, D., Brogna, C., de Bruijn, Y., Buitelaar, J. K., Chakrabarti, B., Charman, T., Cornelissen, I., Crawley, D., DellâĂŹAcqua, F., Dumas, G., Durston, S., Ecker, C., Faulkner, J., Frouin, V., GarcÃl's, P., Goyard, D., Ham, L., Hayward, H., Hipp, J., Holt, R. J., Johnson, M. H., Jones, E. J. H., Kundu, P., Lai, M.-C., DâĂŹardhuy, X. L., Lombardo, M. V., Loth, E., Lythgoe, D. J., Mandl, R., Marquand, A., Mason, L., Mennes, M., Meyer-Lindenberg, A., Moessnang, C., Mueller, N., Murphy, D. G. M., Oakley, B., OâĂŹDwyer, L., Oldehinkel, M., Oranje, B., Pandina, G., Persico, A. M., Price, J., Rausch, A., Ruggeri, B., Ruigrok, A. N. V., Sabet, J., Sacco, R., CÃąceres, A. S. J., Simonoff, E., Spooren, W., Tillmann, J., Toro, R., Tost, H., Waldman, J., Williams, S. C. R., Wooldridge, C., Ilioska, I., Mei, T., Zwiers, M. P., Forde, N. J., and The EU-AIMS LEAP Group (2023). Linking functional and structural brain organisation with behaviour in autism: a multimodal EU-AIMS longitudinal european autism project (LEAP) study. *Molecular Autism*, 14(1):32.

[Palumbo et al., 2022] Palumbo, E., Daunhawer, I., and Vogt, J. E. (2022). MMVAE+: Enhancing the Generative Quality of Multimodal VAEs without Compromises. In *ICLR*.

[Parkes et al., 2021] Parkes, L., Moore, T. M., Calkins, M. E., Cook, P. A., Cieslak, M., Roalf, D. R., Wolf, D. H., Gur, R. C., Gur, R. E., Satterthwaite, T. D., and Bassett, D. S. (2021). Transdiagnostic dimensions of psychopathology explain individuals' unique deviations from normative neurodevelopment in brain structure. *Translational Psychiatry*, 11(1):232.

[Qiu et al., 2022] Qiu, L., Lin, L., and Chinchilli, V. M. (2022). Variational Interpretable Learning from Multi-view Data. arXiv:2202.13503 [cs, stat].

[Romer et al., 2018] Romer, A. L., Knodt, A. R., Houts, R., Brigidi, B. D., Moffitt, T. E., Caspi, A., and Hariri, A. R. (2018). Structural alterations within cerebellar circuitry are associated with general liability for common mental disorders. *Molecular Psychiatry*, 23(4):1084–1090.

[Romer et al., 2021] Romer, A. L., Knodt, A. R., Sison, M. L., Ireland, D., Houts, R., Ramrakha, S., Poulton, R., Keenan, R., Melzer, T. R., Moffitt, T. E., Caspi, A., and Hariri, A. R. (2021). Replicability of structural brain alterations associated with general psychopathology: evidence from a population-representative birth cohort. *Molecular Psychiatry*, 26(8):3839–3846.

[Rommelse et al., 2017] Rommelse, N., Buitelaar, J. K., and Hartman, C. A. (2017). Structural brain imaging correlates of asd and adhd across the lifespan: a hypothesis-generating review on developmental asd–adhd subtypes. *Journal of Neural Transmission*, 124(2):259–271.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241.

[Rosen et al., 2018] Rosen, A. F. G., Roalf, D. R., Ruparel, K., Blake, J., Seelaus, K., Villa, L. P., Ciric, R., Cook, P. A., Davatzikos, C., Elliott, M. A., Garcia de La Garza, A., Gennatas, E. D., Quarmley, M., Schmitt, J. E., Shinohara, R. T., Tisdall, M. D., Craddock, R. C., Gur, R. E., Gur, R. C., and Satterthwaite, T. D. (2018). Quantitative assessment of structural image quality. *Neuroimage*, 169:407–418.

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

[Sheffield et al., 2017] Sheffield, J. M., Kandala, S., Tamminga, C. A., Pearlson, G. D., Keshavan, M. S., Sweeney, J. A., Clementz, B. A., Lerman-Sinkoff, D. B., Hill, S. K., and Barch, D. M. (2017). Transdiagnostic Associations Between Functional Brain Network Integrity and Cognition. *JAMA Psychiatry*, 74(6):605–613.

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

[Stefanik et al., 2018] Stefanik, L., Erdman, L., Ameis, S. H., Foussias, G., Mulsant, B. H., Behdinan, T., Goldenberg, A., O'Donnell, L. J., and Voineskos, A. N. (2018). Brain-behavior participant similarity networks among youth and emerging adults with schizophrenia spectrum, autism spectrum, or bipolar disorder and matched controls. *Neuropsychopharmacology*, 43(5):1180–1188.

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

[Suzuki et al., 2017] Suzuki, M., Nakayama, K., and Matsuo, Y. (2017). Joint multimodal learning with deep generative models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net.

[Tenenhaus et al., 2014] Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., and Frouin, V. (2014). Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583.

[Tong et al., 2022] Tong, X., Xie, H., Carlisle, N., Fonzo, G. A., Oathes, D. J., Jiang, J., and Zhang, Y. (2022). Transdiagnostic connectome signatures from resting-state fmri predict individual-level intellectual capacity. *Translational Psychiatry*, 12(1):367.

[Vincent et al., 2010] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.

[Vogt et al., 1992] Vogt, B. A., Finch, D. M., and Olson, C. R. (1992). Functional Heterogeneity in Cingulate Cortex: The Anterior Executive and Posterior Evaluative Regions. *Cerebral Cortex*, 2(6):435–443.

[Wachinger et al., 2021] Wachinger, C., Rieckmann, A., and PÃűlsterl, S. (2021). Detect and correct bias in multi-site neuroimaging datasets. *Medical Image Analysis*, 67:101879.

[Weinberger et al., 2022] Weinberger, E., Beebe-Wang, N., and Lee, S.-I. (2022). Moment matching deep contrastive latent variable models. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2354–2371. PMLR.

[Williams, 2022] Williams, L. M. (2022). Special report: Precision psychiatryâĂŤare we getting closer?

[Wold et al., 2001] Wold, S., SjÃűstrÃűm, M., and Eriksson, L. (2001). Pls-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130. PLS Methods.

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

[Yang et al., 2021] Yang, Q., Zhang, X., Song, Y., Liu, F., Qin, W., Yu, C., and Liang, M. (2021). Stability test of canonical correlation analysis for studying brain-behavior relationships: The effects of subject-to-variable ratios and correlation strengths. *Human Brain Mapping*, 42(8):2374–2392.

[Yin et al., 2022] Yin, S., Hong, S.-J., Di Martino, A., Milham, M. P., Park, B.-Y., Benkarim, O., Bethlehem, R. A. I., Bernhardt, B. C., and Paquola, C. (2022). Shared and distinct patterns of atypical cortical morphometry in children with autism and anxiety. *Cerebral Cortex*, 32(20):4565–4575.

[Zhao et al., 2019] Zhao, F., Xia, S., Wu, Z., Duan, D., Wang, L., Lin, W., Gilmore, J. H., Shen, D., and Li, G. (2019). Spherical u-net on cortical surfaces: Methods and applications.

[Zou et al., 2022] Zou, K., Faisan, S., Heitz, F., and Valette, S. (2022). Joint disentanglement of labels and their features with vae. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1341–1345.

A Multimodal Variational AutoEncoders

This appendix presents details on the multiview variational autoencoder (mVAE) models, used to perform the Digital Avatar Analysis, once they are trained.

Multimodal (or multiview) VAEs are extensions of the well-established VAE [Kingma and Welling, 2014] that allow joint integration and reconstruction of two or more views. These two VAE models are themselves derived from AEs which attempt to find a compressed representation that contains as much information as possible about the input data. They are based on the assumption that the input data can be generated from a lower dimensional space [Kramer, 1991]. AEs have been extensively studied [Vincent et al., 2010] and used in various applications such as medical image segmentation [Ronneberger et al., 2015]. In the variational version of anAE, the main idea is to learn a probability distribution over the low dimensional latent space. Finally, multimodal VAEs allow the construction of view-specific and shared latent spaces via the different modeling schemes. This is of great interest in medicine, for example to separate common sources of variability from disease-specific one [Wu and Goodman, 2018, Shi et al., 2019, Sutter et al., 2021].

The selected model
Let $x = \{x_1, \ldots, x_K\}$ be an observation of K views, where each x_k can have different dimensions J_k. The general form of the latent of the variable is $\{z, z_1, \ldots, z_k, \ldots, z_K\}$ with z denoting the d-dimensional latent variable shared by all x_k, and z_k the view-specific d_k-dimensional latent variables. We assume the following generative process for the observation set:

$$\begin{aligned} z &\sim p(z) \\ z_k &\sim p(z_k) \\ x_k &\sim p_\theta(x_k|z, z_k) \end{aligned}$$

where $p(z), p(z_k)$ are prior distributions for the latent variables and $p_\theta(x_k|z, z_k)$ is a likelihood distribution of the observations conditioned on the latent variable. The generative process $p_\theta(x)$ is approximated by its Evidence Lower Bound (ELBO) . The bottleneck of this approximation problem comes to computing the posterior $p_\theta(z, z_k|x)$. A first usual assumption about $p_\theta(z|x)$ considers that it contains all the shared information between the views, and we can write $p_\theta(z, z_k|x) = p_\theta(z|x) \times p_\theta(z_k|x_k)$. Other assumptions are needed since these posteriors are not analytically tractable. The posterior distributions are classically approximated by parametrized distributions $q_\phi(z|x)$ (the so-called joint variational posterior) and $q_\phi(z_k|x_k)$ respectively.

State-of-the-art models solve this problem with different ways of modeling the joint variational posterior distribution $q_\phi(z|x)$. These approaches compose $q_\phi(z|x)$ using the Gaussian experts $q_\phi(z|x_k)$ unimodal posteriors, which may be combined in various ways. In Product of Experts (PoE) approach [Wu and Goodman, 2018], $q_\phi(z|x)$ is defined as a product of individual Gaussian experts, assuming the conditional independence of the experts, hence remains Gaussian. The product operator used in PoE may dampen the specific variability of a given block described by gaussian unimodal posterior, and other work proposes to model [Shi et al., 2019] $q_\phi(z|x)$ using a Mixture of Experts (MoE), assuming each unimodal experts contributes equally to the joint posterior. MoPoE approaches defines $q_\phi(z|x)$ as a Mixture of Product of Experts. Each product of experts is computed over the different subsets of views [Sutter et al., 2021]. We use the latter strategy, referred to as MoPoE, as it is a generalization of the two previous approaches. Each strategy has different ways of handling missing views, that may differ between train and test time. MoPoE simply uses the subsets of available views for each observation. As regards to the unimodal posteriors $q_\phi(z_k|x_k)$, we classicaly use Gaussian distributions.

The MoPoE optimizes a generalized multimodal ELBO objective for learning view-specific and a joint distribution of multiple views x with potential missing data, defined as:

$$\begin{aligned} L_{MoPoE}(\theta, \phi; x) = &\sum_{i=1}^{K} E_{q_\phi(z,z_k|x)} \left[log(p_\theta(x_k|z, z_k)\right] \\ &- \sum_{i=1}^{K} D_{KL}\left(q_\phi(z_k|x_k)||p_\theta(z_k)\right) - D_{KL}\left(\underbrace{\frac{1}{2^K}\sum_{x_p \in \mathcal{P}(x)} \tilde{q}_\phi(z|x_p)}_{=q_\phi(z|x)} ||p_\theta(z)\right) \end{aligned}$$

There are 2^K different subsets contained in the powerset $\mathcal{P}(x)$, and $\tilde{q}_\phi(z|x_p) = \prod_{x_k \in x_p} q_\phi(z|x_k)$ is the product of Gaussian expert posteriors $q_\phi(z|x_k)$, defined for the subset of views in x_p.

These formulae show our specific implementation of the MoPoE which handles multi-views by modelling view-specific latent space posterior distributions, and a joint posterior distribution shared between views as a mixture of the products of their marginal posteriors (experts). This implementation differs from the one benchmarked in [Daunhawer et al., 2022, Anonymous, 2023] which consider only one joint latent (MoPoE with its initial formulation [Sutter et al., 2021]). These benchmarks also favor the generative abilities when we essentially consider its representation space in our intrepretation framework.

Model specification and training
In our work, the encoders with parameters ϕ are each defined as multilayer perceptrons (MLPs), each with one hidden layer of 256 units and a ReLU activation. The decoders with parameters θ are linear (a single fully connected layer). Note that every decoder has a learnable variance parameter for each reconstructed feature that does not depend on the input observation. This allows the decoders to learn a population-level variability. To handle inputs of different sizes, the dimensions of the view-specific latent spaces were set individually. Based on a previous study [Ambroise et al., 2021], we chose $d_{eCRF} = 1$ and $d_{ROI} = 20$. Defining the dimension of the shared latent space jointly with the dimensions $(J_1, \ldots, J_K)$ of the input views seems to be a reasonable criterion, we use the following rule:

$$d < min_{k \in \{1,\ldots,K\}}\{J_k - d_k\}$$

To spread the variability of the observations across the view-specific and shared latent representations, we choose $d = 3$. All data are z-scored before being fed into the network. The model is trained with an Adam optimization, a initial learning rate of 2×10^{-3} and a batch size of 256.

B Representational Similarity Analysis

Assessing model quality in an unsupervised setting remains an open issue. To investigate the learned latent representations, we derive a Representational Similarity Analysis (RSA) [Kriegeskorte et al., 2008] 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 τ) enables the comparison of these dissimilarity matrices emphasizing the captured information. RSA is used twice in the present work. First we use it as a component of our framework to weight the contribution of a model when aggregating its extracted associations in the regularized digital avatar analysis (r-DAA) (see Section 2.3 and Supplementary alg. D.2). Second we consider it to evaluate the information contained in the different latent representations (see Section 3.1).

Model quality scoring used in r-DAA
We hypothesize that the latent representations of a good model to understand eCRF symptom scores should contain some amount of symptoms-related variability. We evaluate each trained MoPoE-VAE using RSA output Kendall τ between the model's joint latent space and each eCRF score (τ_s), defined as $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{|S|}) \in [-1, 1]^{|S|}$. For each of the n_E models, we use it's average Kendall $\hat{\tau} = \frac{1}{|S|} \sum_{s=1}^{|S|} \tau_s$ across eCRF scores as it quality score, used to weight the aggregation function in the ensembling of the n_E corresponding DAAs (see section 2.3). This is detailed in the Supplementary alg D.2 as well.

Assessing the latent representation spaces with RSA
We report in Table 1 aggregated results from RSAs applied to each split and corresponding models. We report the average $N \times n_E$ Kendall τ between each different measure and latent space. In the table we report the significance of these correlations. We compute it by considering the p-values of the Kendall τ statistics (corresponding to the rejection of the null hypothesis). We correct them for multiple testing using the Hommel [HOMMEL, 1988] correction for multiple dependent tests and multiple independent test using Bonferroni. Then we consider the median of the corrected p-values. We chose the value of 1% for the significance threshold.

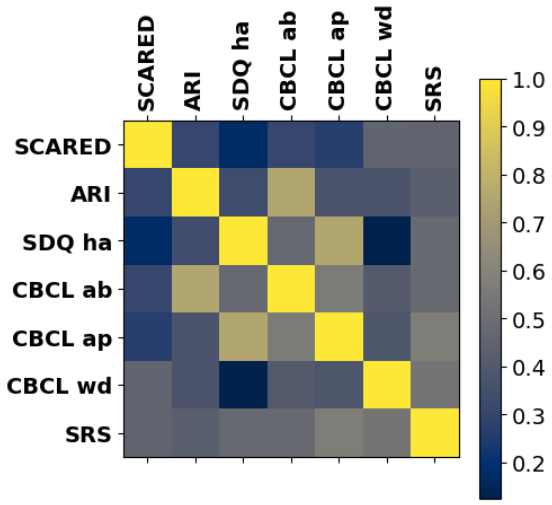


Figure S1: Correlations between the 7 eCRf scores.

C Digital Avatar Analysis

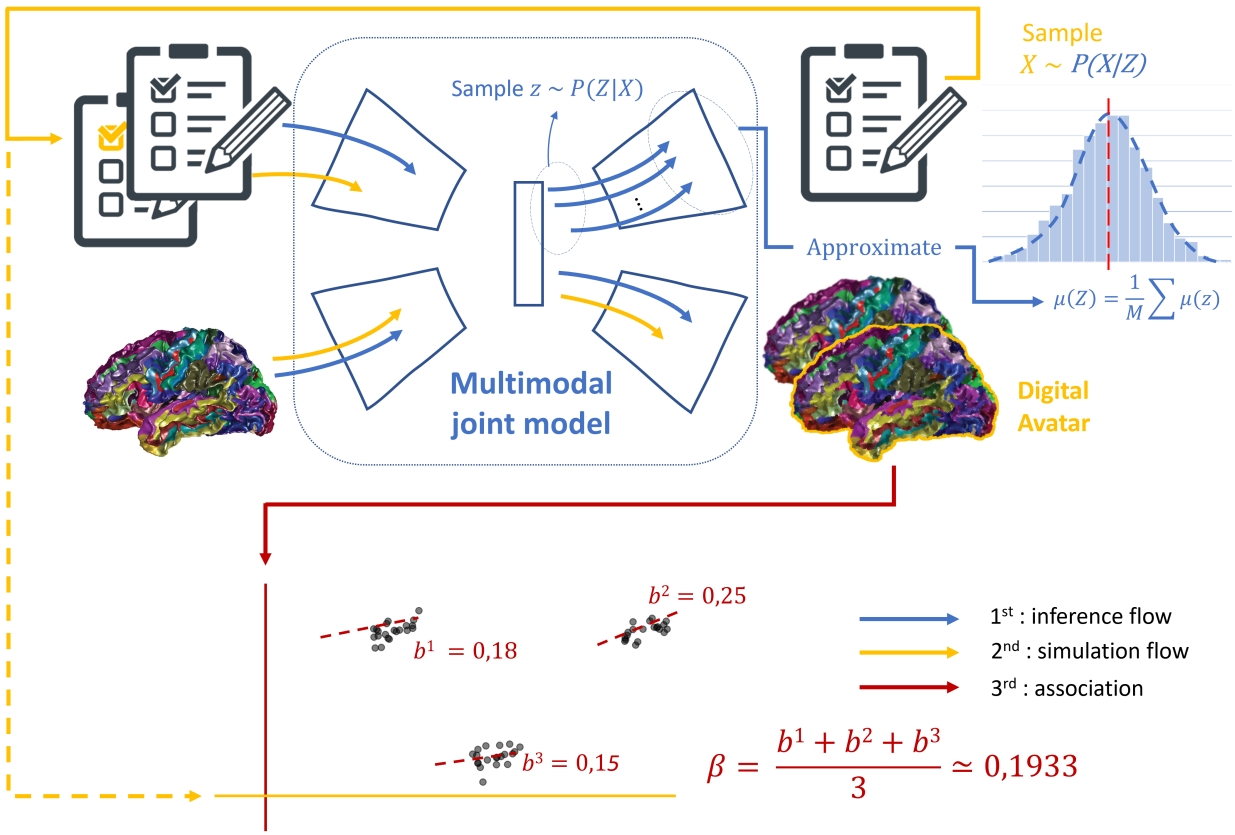


Figure S2: Illustration of the interpretation framework based on DAA in a clinical cohort setting with two modalities: imaging data and clinical questionnaires. First, the inference flow estimates output distributions via sampling in the latent space. Then, the simulation flow generates realistic perturbed 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.

D Algorithm pseudocodes

We display below the pseudocode of the two procedures developed for our association discovery pipeline, relying on r-DAA and stability selection.

D.1 Finding associations with a non-weighted r-DAA

Algorithm 1 r-DAA using a simple average function for aggregation

Input R: a set of ROIs
M: a set of metrics
S: a set of eCRF scores
X: a set with N random train (2690) / left-out (301) splits

$p \leftarrow |R| * |M| * |S|$ ▷ num of features
$n_E \in \{1, \dots, 20\}$ ▷ num of models
$N \leftarrow 100$ ▷ num of splits
$\pi_{\text{thr}} \leftarrow 0.4$ ▷ stability path thr
for $j \in \{1, \dots, N\}$ **do**
  $x^j_{\text{train}}, x^j_{\text{left-out}} \leftarrow X[j]$ ▷ a random split
  for $n_E \in \{1, \dots, 20\}$ **do**
    (a) DAA: Fit a $MoPoE^j_{n_E}$ model on x^j_{train} using init random weights θ^{n_E}
    (a) DAA: Compute the DAA with $x^j_{\text{left-out}}$ association weights $\beta^j_{n_E} \in \mathbb{R}^p$
  end for
end for
$f \leftarrow$ mean function ▷ aggregation func
for $n_E \in \{1, \dots, 20\}$ **do**
  for $j \in \{1, \dots, N\}$ **do**
    (b) ensembling: Compute aggregation scores $f((\beta^j_1, \dots, \beta^j_{n_E})) \in \mathbb{R}^p$
    (b) ensembling: $A^j_{n_E} \leftarrow f((\beta^j_1, \dots, \beta^j_{n_E}))$
    (b) ensembling: Compute decision support $g(A^j_{n_E}) \in \{0, 1\}^p$
    (b) ensembling: $S^j(n_E) \leftarrow g(A^j_{n_E})$
  end for
  (c) Stability selection: Compute the selection probability of each feature $\Pi^{n_E} \in \mathbb{R}^p$
  (c) Stability selection: $\Pi^{n_E} \leftarrow \frac{1}{N} \sum_{j=1}^{N} S^j(n_E)$
end for
$S^{\text{stable}} \leftarrow \{k : \max_{n_E \in \{1,\dots,20\}} \Pi^{n_E}_k > \pi_{\text{thr}}\}$ ▷ stable associations

Output S^{stable}

D.2 Finding associations with a weighted r-DAA

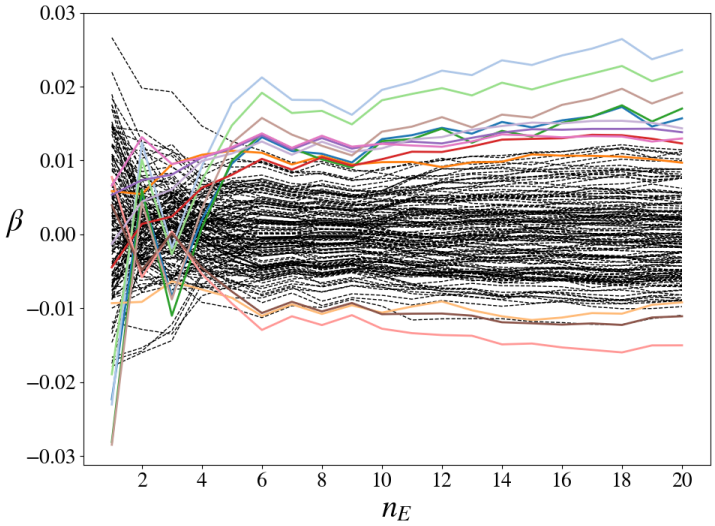
Algorithm 2 r-DAA using a RSA weighted average function for aggregation

Input R: a set of ROIs
M: a set of metrics
S: a set of eCRF scores
X: a set with N random train (2690)/ left-out (301) splits

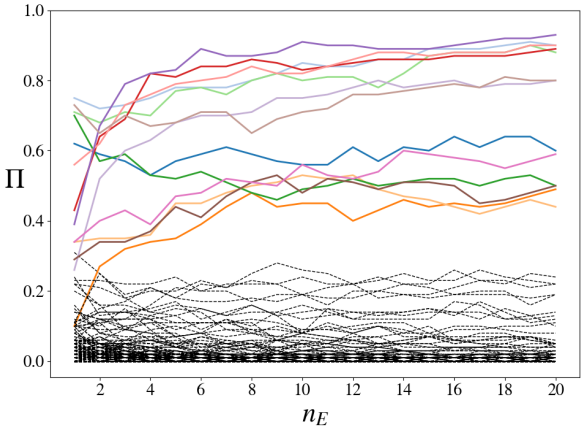
$p \leftarrow |R| * |M| * |S|$ ▷ num of features
$n_E \in \{1, \ldots, 20\}$ ▷ num of models
$N \leftarrow 100$ ▷ num of splits
$\pi_{\text{thr}} \leftarrow 0.4$ ▷ stability path thr
for $j \in [1, N]$ **do**
  $x^j_{\text{train}}, x^j_{\text{left-out}} \leftarrow X[j]$ ▷ a random split
  for $n_E \in \{1, \ldots, 20\}$ **do**
(a) DAA
    Fit a $MoPoE^j_{n_E}$ model on x^j_{train} using init random weights θ^{n_E}
    Compute the DAA with $x^j_{\text{left-out}}$ association weights $\beta^j_{n_E} \in \mathbb{R}^p$
    Compute the left-out subjects pairwise joint latent representation dissimilarity matrix $C^j_{n_E} \in \mathbb{R}^{301 \times 301}$
      $C^j_{n_E} \leftarrow [distance(z_i, z_k) \text{ for } (i,k) \in \{1, \ldots, 301\}^2] \in \mathbb{R}^{301 \times 301}$
    for $s \in S$ **do**
      Compute the left-out subjects pairwise eCRF score dissimilarity matrix $S^j_{n_E} \in \mathbb{R}^{301 \times 301}$
        $S^j_{n_E} \leftarrow [distance(s_i, s_k) \text{ for } (i,k) \in \{1, \ldots, 301\}^2] \in \mathbb{R}^{301 \times 301}$
      Compute the Kendall $\tau^j_{n_E}(s) \in \mathbb{R}$ between upper triangular parts of $C^j_{n_E}$ and $S^j_{n_E}$
    end for
    Compute average Kendall $\tau^j_{n_E} \in \mathbb{R}$ across the eCRF scores
      $\tau^j_{n_E} \leftarrow \frac{1}{|S|} \sum_{s \in S} \tau^j_{n_E}(s)$
  end for
end for
$f \leftarrow$ weighted mean function ▷ aggregation func
for $n_E \in \{1, \ldots, 20\}$ **do**
  for $j \in \{1, \ldots, N\}$ **do**
(b) ensembling
    Compute aggregation scores $f((\beta^j_1, \ldots, \beta^j_{n_E}), (\tau^j_1, \ldots, \tau^j_{n_E})) \in \mathbb{R}^p$
      $A^j_{n_E} \leftarrow f((\beta^j_1, \ldots, \beta^j_{n_E}), (\tau^j_1, \ldots, \tau^j_{n_E})) = \frac{1}{n_E} \sum_{i=1}^{n_E} \tau^j_i \beta^j_i$
    Compute decision support $g(A^j_{n_E}) \in \{0, 1\}^p$
      $S^j(n_E) \leftarrow g(A^j_{n_E})$
  end for
(c) Stability selection
  Compute the selection probability of each feature $\Pi^{n_E} \in \mathbb{R}^p$
    $\Pi^{n_E} \leftarrow \frac{1}{N} \sum_{j=1}^{N} S^j(n_E)$
end for
$S^{\text{stable}} \leftarrow \{k : \max_{n_E \in \{1, \ldots, 20\}} \Pi^{n_E}_k > \pi_{\text{thr}}\}$ ▷ stable associations

Output S^{stable}

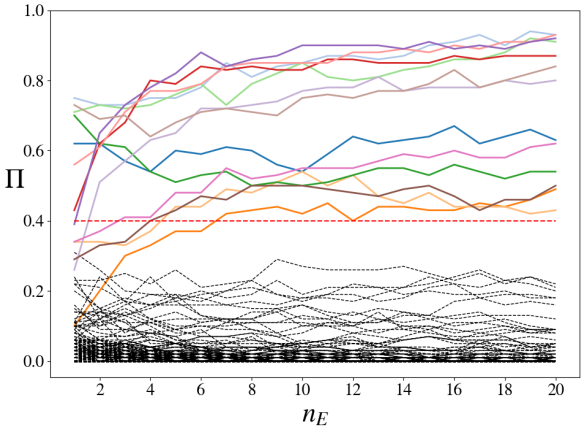
E Additional stability plots



(a) Coefficients



(b) Stability paths using uniform ensembling



(c) Stability paths using weighted ensembling

Figure S3: Investigation of the ROIs associated with SDQ in area. Each line corresponds to a ROI, dotted black ones are not selected as below the threshold $\pi_{thr} = 0.4$ when using as ensembling function a mean weighted by model ratings. The ROIs selected are colored and the color is consistent across the plots. The red horizontal dotted line highlights the threshold $\pi_{thr} = 0.4$. (a) Aggregated coefficients with a simple mean for a given split over the $N = 100$ splits, plotted against the number of models aggregated with the mean. (b) Feature wise stability Π when using an uniformly weighted mean to aggregate the coefficients output from r-DAA for each model, against the number of models used. (c) Same as the last one, except that the mean was weighted using model ratings. The latter strategy is used to select the ROIs associated in area with the SDQ score, with $\Pi > \pi_{thr} = 0.4$.

F All associations

Score	SRS	SCARED	ARI	SDQ ha
Metric				
Thickness	L.G.Te.Mid L.Pole.Oc L.S.PeriCal L.S.Te.Inf R.G.Te.Mid R.Pole.Oc R.S.PeriCal R.S.PostCe R.S.PreCe.Inf. R.S.Te.Inf	L.Ci.Mid.Post L.G.Ci.Post.Dors L.G.Pa.Sup L.G.Te.Mid L.Pole.Oc L.S.PeriCal L.S.Te.Inf R.G.Oc.Te.Med.PH R.G.Pa.Sup R.G.Te.Mid R.Pole.Oc R.S.Inter.Prim.Jens R.S.PeriCal R.S.PreCe.Inf. R.S.Te.Inf	L.FrMarg L.Ci.Mid.Post L.G.Oc.Te.Med.PH L.G.PostCe L.G.Te.Inf R.G.Oc.Te.Med.PH R.G.PostCe R.G.Te.Inf R.Lat.Fis.Ant.Vert R.Pole.Te R.S.Fr.Inf R.S.Fr.Mid	L.Ci.Mid.Post L.G.Pa.Sup L.G.Precu L.Pole.Oc L.S.Ce L.S.PeriCal R.Ci.Mid.Post R.G.Ci.Post.Dors R.G.Ins.St R.G.Oc.Te.Med.PH R.G.Pa.Sup R.Pole.Oc R.S.Ce R.S.PostCe
Meancurv	L.Ci.Mid.Post L.G.Ci.Post.Dors L.G.Ci.Post.Ventr L.G.Te.Sup.Pl.Pola L.S.Ci.Marg L.S.PeriCal R.Oc.Inf R.Ci.Mid.Post R.G.Ci.Post.Ventr R.G.Oc.Te.Med.Li R.G.Te.Sup.Pl.Pola R.Pole.Oc R.S.Circ.Ins.Sup R.S.PeriCal	L.Ci.Mid.Post L.G.Ci.Post.Dors L.G.Ci.Post.Ventr L.G.Te.Sup.Pl.Pola L.S.Ci.Marg L.S.Pa.Oc L.S.PeriCal R.Oc.Inf R.Ci.Mid.Post R.G.Ci.Post.Ventr R.G.Oc.Te.Med.Li R.Pole.Oc R.S.Circ.Ins.Sup R.S.PeriCal	L.G.Ins.St L.G.SubCal L.G.Te.Sup.Pl.Pola L.S.Oc.Te.Med.Ling L.S.Pa.Oc L.S.PreCe.Sup. R.Transv.FrPole R.G.PostCe R.G.Te.Sup.Pl.Pola R.S.Ce R.S.PeriCal	L.G.Ci.Post.Dors L.G.Ins.Lg.S.Cent.Ins L.G.Te.Sup.Pl.Pola L.S.Ci.Marg L.S.Pa.Oc L.S.SubPa R.Oc.Inf R.Ci.Mid.Post R.G.Oc.Te.Med.Li R.G.Pa.Inf.Ang R.G.Pa.Inf.Supr
Area	L.Ci.Mid.Ant L.Ci.Mid.Post L.G.Fr.Inf.Tri L.G.Pa.Inf.Supr L.G.SubCal L.G.Te.Sup.Pl.Te L.Lat.Fis.Post L.S.Ci.Marg L.S.Circ.Ins.Inf L.S.Fr.Inf R.Ci.Mid.Post R.S.Ci.Marg R.S.Circ.Ins.Inf R.S.PeriCal	L.ParaCe L.Ci.Mid.Ant L.Ci.Mid.Post L.G.Te.Sup.Pl.Te L.Lat.Fis.Post L.S.Ci.Marg L.S.Orb.Med.Olfact R.Ci.Mid.Post R.G.Rect R.G.Te.Sup.Pl.Te R.S.Ci.Marg R.S.Orb.Med.Olfact R.S.PeriCal	L.Ci.Mid.Ant L.Ci.Mid.Post L.G.PostCe L.G.SubCal L.S.Ce L.S.Ci.Marg L.S.Te.Sup R.Ci.Mid.Post R.G.PreCe R.S.Ce R.S.Ci.Marg	L.Ci.Mid.Ant L.Ci.Mid.Post L.G.Te.Sup.Pl.Te L.S.Ce L.S.Ci.Marg R.Ci.Mid.Post R.G.Pa.Inf.Supr R.G.PreCe R.G.Te.Sup.Pl.Te R.Lat.Fis.Post R.S.Ce R.S.Ci.Marg R.S.PeriCal

Score	CBCL ab	CBCL ap	CBCL wd
Metric			
Thickness	L.G.Oc.Te.Med.PH L.G.PostCe L.G.Te.Inf L.Pole.Oc	L.FrMarg L.Ci.Mid.Post L.G.Oc.Te.Med.PH L.G.Te.Inf	L.Ci.Mid.Post L.G.Pa.Sup L.Pole.Oc L.S.PeriCal

	L.S.Fr.Inf L.S.PeriCal L.S.Te.Inf R.G.Oc.Te.Med.PH R.G.PostCe R.G.Te.Inf R.G.Te.Mid R.Lat.Fis.Ant.Vert R.Pole.Oc	L.S.Fr.Mid R.Ci.Mid.Post R.G.Ci.Post.Dors R.G.Ins.St R.G.Oc.Te.Med.PH R.G.Te.Inf R.Pole.Te R.S.Ce R.S.Fr.Mid	L.S.PreCe.Sup. L.S.Te.Inf R.Ci.Mid.Post R.Pole.Oc R.S.Ce R.S.PostCe R.S.PreCe.Inf.
Meancurv	L.Ci.Mid.Post L.G.SubCal L.G.Te.Sup.Pl.Pola L.S.Pa.Oc L.S.PeriCal L.S.PreCe.Sup. R.ParaCe R.Ci.Mid.Post R.G.Te.Sup.Pl.Pola R.S.Circ.Ins.Sup R.S.PeriCal	L.G.Ci.Post.Dors L.G.Ins.Lg.S.Cent.Ins L.G.SubCal L.S.Ci.Marg L.S.Pa.Oc L.S.SubPa R.G.Oc.Te.Med.PH R.G.Pa.Inf.Ang R.G.Pa.Inf.Supr R.G.PostCe R.S.Ce R.S.Fr.Sup	L.ParaCe L.G.Ci.Post.Dors L.G.Ins.Lg.S.Cent.Ins L.G.Te.Sup.Pl.Pola L.S.Ci.Marg L.S.Pa.Oc R.Ci.Mid.Post R.G.Oc.Te.Med.Li R.G.Pa.Inf.Ang R.Pole.Oc R.S.Ci.Marg R.S.PeriCal
Area	L.Ci.Mid.Ant L.Ci.Mid.Post L.G.Fr.Inf.Tri L.G.PostCe L.G.SubCal L.S.Ci.Marg L.S.Fr.Inf L.S.Te.Sup R.Ci.Mid.Post R.S.Ci.Marg R.S.Circ.Ins.Inf	L.Ci.Mid.Ant L.Ci.Mid.Post L.S.Ce L.S.Ci.Marg L.S.Te.Sup R.Ci.Mid.Post R.G.Fr.Mid R.G.Oc.Te.Med.PH R.G.PreCe R.G.Te.Sup.Pl.Te R.S.Ce R.S.Ci.Marg	L.ParaCe L.Ci.Mid.Ant L.Ci.Mid.Post L.G.Fr.Inf.Tri L.G.Te.Sup.Pl.Te L.Lat.Fis.Post L.S.Ci.Marg L.S.Circ.Ins.Inf L.S.Fr.Inf R.Ci.Mid.Post R.G.Pa.Inf.Supr R.G.PreCe R.G.Te.Sup.Pl.Te R.Lat.Fis.Post R.S.Ci.Marg R.S.PeriCal

Table 2: Retained associations for each score and metric. Blue indicates a negative association and red denote a positive one. L: left, R: right, S: sulcus, G: gyrus, Lat: lateral, Ci: cingul, Pa: parietal, Ce: central, Oc: occipital, Te: temporal, Fr: front, Orb: orbital, Ins: insula, Post: posterior, Mid: middle, Ant: anterior, Med: medial, Sup: superior, Inf: inferior, Ventr: ventral, Dors: dorsal, PostCe: postcentral, PreCe: precentral, ParaCe: paracentral, PeriCal: pericallosal, Marg: marginal, Pl: plan, Tri: triangular, Circ: circular, Supr: supramarginal, Ang: angular, PH: parahippocampal, SubPa: subparietal, Transv: transverse, FrPole: frontopolar, Pola: polar, SubCal: subcallosal, St: short, Fis: fissure, Olfact: olfactory, Rect: rectus, Precu: precuneus, FrMag: fronto-marginal, FrMarg: frontomargin, Inter: intermedius, Li: lingual, Jens: Jensen, Vert: vertical, Prim: primus.