Supplementary material

Methods

Datasets & Subjects

Two datasets were used in this study: ABIDE I [31; 46] and SFARI VIP [22]. ABIDE I consists of 1112 participants in total, 539 ASD participants (65 female) and 573 TD participants (99 female). ABIDE data were collected across 17 scanning sites. Ages (in years) ranged between 6-56 in the TD group and between 7-64 in the ASD group. The SFARI VIP dataset was originally collected to investigate the genetic bases of ASD. The MRI dataset consisted of 341 subjects total, 235 of whom had confirmed copy-number-variations: 16p11.2 deletion (N = 113), 16p11.2 duplication (N = 91), 1q-deletion (N=14), 1q-duplication (N = 13) and 16p-triplication (N = 4). N=106 Subjects were non-familial controls. More details about the two datasets can be found in [31; 46] (ABIDE I) and in [22] (SFARI VIP).

Quality assurance and exclusion criteria

Structural data were included for N=1099 ABIDE I subjects. Data were excluded if an anatomical scan failed to complete the tissue segmentation procedure (SPM12) or if tissue segmentation was inaccurate. Data from N=8 subjects failed to complete the segmentation, resulting in errors and no output. Data from N=109 subjects resulted in inaccurate segmentation (segmented gray matter and white matter masks were outside the brain). These data were excluded from further analyses, resulting in a final sample size of N = 982 subjects. After the quality assurance (QA) steps described above, ABIDE I consisted of 470 ASD participants (N=55 female), age range 7-64 (M = 17.63, SD = 8.75) and 512 TD participants (82 female) age range 6-56 (M = 17.41, SD = 7.95).

SFARI data was a previously selected subset containing structural and functional data (N = 121). Functional data were used in a separate project (not discussed here). Structural data were subjected to the same quality assurance steps as ABIDE data. Data from all N = 121 subjects successfully passed tissue segmentation procedure and tissue probability masks were manually inspected for segmentation quality. The final sample consisted of N = 121 subjects, N = 26 of whom had 16p11.2 deletion CNV (age range 7-44, M = 15.12, SD = 9.69, 7 diagnosed ASD) and N = 25 had 16p11.2 duplication CNV (age range 5-52, M = 25.28, SD = 14.54, 0 diagnosed ASD).

MRI data pre-processing

The data used as input for the autoencoders remained in native space (they were not normalized to a template). The T1w anatomical images were segmented into separate tissue classes using SPM12. We then used gray matter and white matter probability masks (tissue probability threshold = .8) to extract voxels corresponding to brain tissue. The brain tissue images were then normalized (between 0 and 1) and re-sampled to a target resolution of 64x64x64 voxels before being passed on to further analyses.

Architecture of the autoencoders

The CVAE architecture was based on a modified version of a two stream neural network, originally described in [55]. The inputs were anatomical brain volumes in native space resampled to 64x64x64 resolution (see: MRI data pre-processing). The CVAE consists of an encoder and a decoder. The encoder employed two successive convolutional layers (kernel size: 3, stride: 2, 64 and 128 convolutional filters) for shared features, and two convolutional layers (kernel size: 3, stride: 2, 64 and 128 convolutional filters) for ASD-specific features (see Figure S2, Table S1). These encoders projected the data separately onto parameters of two distinct, 16-dimensional latent distributions (a distribution on shared space, and a distribution on ASD-specific space). This procedure is standard

for variational autoencoders, the CVAE is unique in that it has two separate latent spaces instead of a single one. The decoder layer took as input a 32 dimensional vector (obtained by concatenating the shared and ASD-specific features) and produced as output a reconstructed structural MRI volume. Decoding steps used two deconvolutional layers (using 128 and 64 filters), which reconstructed brain images from latent-space representations. For reconstructing structural volumes of TD participants, the latent 32-dimensional representation was a concatenation of the shared-feature representation (a 16-dimensional vector) and 16 zeros (see [55]).

In addition to the CVAE, we trained a standard variational autoencoder (VAE) for comparison. The standard VAE had the same architecture as one stream of the CVAE - the shared stream - and it lacked ASD-specific features. To equate the dimensionality of the latent space across the two models, the standard VAE had the same amount of hidden nodes in the latent-space as the entire CVAE (32): twice the number of hidden nodes of the CVAE's shared space (Table S2). The CVAE and VAE were trained using the ADAM optimizer [29]. Training was stopped when the reconstruction error (mean-squared-error between input and reconstruction batch) fell below .005 (see also Figure S3 for training loss dynamics).

Comparison between anatomical similarity and similarity based on symptoms, demographic properties and scanning site.

Once the the networks were trained, we used an approach inspired by representational similarity analysis (RSA) [10] to probe the information represented in the learned latent spaces (shared, ASD-specific and the latent-space for control VAE). We computed subject similarity according to different demographic variables and behavioral measurements. The modelling procedure differed depending on the measurement instrument (single measurement or measurement battery) and data-scale (categorical- or ratio- valued data). For ratio-valued measurements, such as age and full scale IQ (FIQ), subject distance was the absolute euclidean distance between their measurements (i.e. age difference or FIQ score difference). For categorical variables (such as gender or scanning-site) subject distance was 0 if categorical variables matched (same gender or same scanning-site) and 1 otherwise. For behavioral batteries of tests (such as Vineland, WISC and ADOS) that are composed of multiple instruments (e.g. separate scores for behavioral and social impairments) we applied principal component analysis (PCA) to the test battery and calculated the absolute difference in the first principal component loadings. ADOS scores were included as a battery (ADOS PCA) and also as separate instruments (ADOS total, ADOS social, see Figure 1). To construct RSA models, we used data from all subjects that had a score on that measurement (e.g. ADOS total, Vineland, see table S3).

To establish the robustness of the RSA analyses across different samples extracted from the CVAE's latent distributions, we sampled the latent distribution for each ASD participant for n = 10 times. This effectively generates 10 distinct reconstructions of the dataset, and RSA models were fit to each reconstruction individually. The robustness of the results across samples was then assessed, determining whether the correlations between the representational dissimilarity matrices (RDMs) for the feature spaces and the RDMs for properties such as age, gender, symptoms, and genetics were greater than zero (using one-sample t-tests). Additionally, we used paired sample t-tests to determine whether the RDMs for participants' properties (i.e. age, symptoms) correlated more with the RDMs based on features from the ASD-specific space or with the RDMs based on features from the shared space. To make these tests more stringent, the number of reconstructions was kept intentionally low (large samples can make very small effects significant [27]).

Cluster analysis

To assess the number of clusters in shared, ASD-specific and VAE feature spaces, we followed an elbow plot method. We first concatenated ABIDE and SFARI data to look for clusters across both datasets. For each of the feature spaces, we fitted a Gaussian mixture model with varying numbers of clusters. For each space and for each choice of the number of clusters, we computed Bayesian Information Criterion (BIC) as a measure of model fit

($\log(p)$). For each space, the number of clusters that produced the lowest BIC (measured in log probability) was selected as the optimal number of clusters. To ensure robustness of results - we repeated this procedure for N=100 different samples from the feature distributions, Figure 1 shows mean BIC for each number of clusters, shaded bands indicate the full range (minimum and maximum) of BIC across ten samples of ABIDE data.

Tensor-Based morphometry

We next aimed to identify neuroanatomical features associated with ASD-specific variability represented in the ASD-specific CVAE space. To this end, we associated each point in the ASD-specific latent-space with an ASD-related deformation field, morphing a counterfactual TD brain into the corresponding ASD brain. To calculate this deformation field, we took advantage of the CVAE architecture, and more specifically of the fact that each ASD input datum (brain volume) is projected to both shared and ASD-specific spaces. Using both shared and ASD-specific features together enables us to accurately reconstruct the ASD brain volume used as the input. Importantly, for a given ASD subject, using only the inferred shared features (with ASD-specific features set to zero), we reconstruct a synthetic "TD counterpart" of that brain. This provides a synthetic case-control pair consisting of an ASD brain and a synthetic TD brain matched on sources of shared variation (such as age, gender, scanning-site, and other factors discovered by the CVAE in a data-driven fashion).

Having synthesized case-control pairs for ASD subjects, we next calculated deformation fields transforming the synthetic TD brains into their corresponding ASD brains. These deformation fields reveal subject specific, ASD-related neuroanatomical differences. To accomplish this, we used a nonlinear warping (using inverse-consistent diffeomorphic image registration implemented in ANTs [12; 14; 18]) to transform each synthetic TD brain into its corresponding ASD brain. We then calculated the Jacobian determinant of the transformation matrix as an index of local expansion or contraction [69]. We used SPM12 (two-tailed, two-sample independent t-test with family-wise-error correction for multiple comparisons) to establish statistical significance.

Supplementary figures

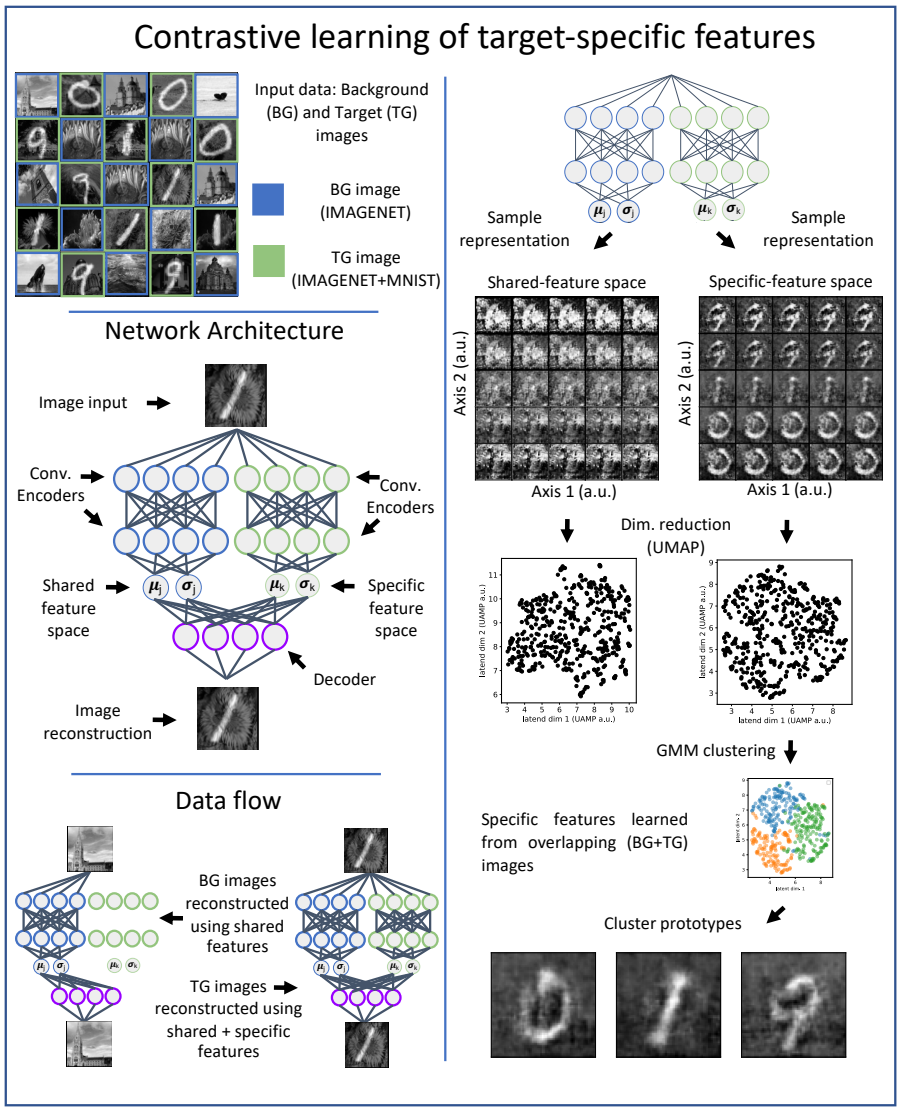

Figure S1: Demonstration of a contrastive variational autoencoder (CVAE) using synthetic data. Synthetic data consisted of background images (selected from ImageNet [15]), and target images (created by superimposing MNIST [3] digits on imagenet images). The Contrastive Variational Autoencoder (CVAE) is a two-stream neural network built to disentangle shared and target-specific features. Background images are reconstructed using shared features, while target images are reconstructed using shared and target-specific features together. The latent spaces of the CVAE (shared and target-specific) learn distinct feature distributions. Clustering shows that the MNIST digits used (0, 1 and 9) occupy distinct parts of the target-specific latent-space.

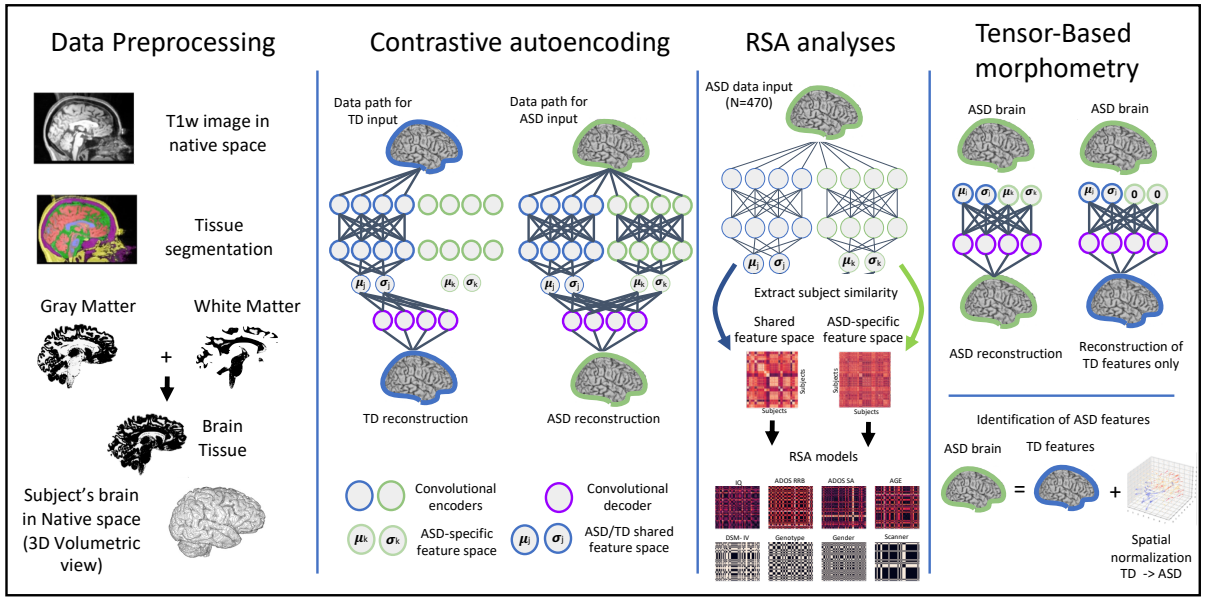

Figure S2: Data analysis pipeline. *Data Preprocessing*: Minimal preprocessing was employed before entering data into the CVAE. Anatomical T1w scans of TD (N = 470) and ASD (N = 512) subjects were segmented into tissue classes, and gray matter and white matter masks were used to reconstruct brain tissue (i.e. brain extraction). *Contrastive Autoencoding*: We trained a Contrastive Variational Autoencoder to disentangle ASD-specific variation from variation shared with TD participants. *RSA analyses* : To investigate what kind of information is represented in latent-spaces, we compared subject similarity in shared and ASD-specific spaces to subject similarity models derived from behavioral, demographic and genetic data. *Tensor-based morphometry*: To identify neuroanatomical correlates of ASD, we performed longitudinal Tensor-based morphometry using synthesized brains. For each ASD subject, we reconstructed their anatomical brain volume using either shared and ASD-specific features (full reconstruction) or only shared features. We then analysed the Jacobian determinant of the vector fields required to morph TD brains into ASD brains.

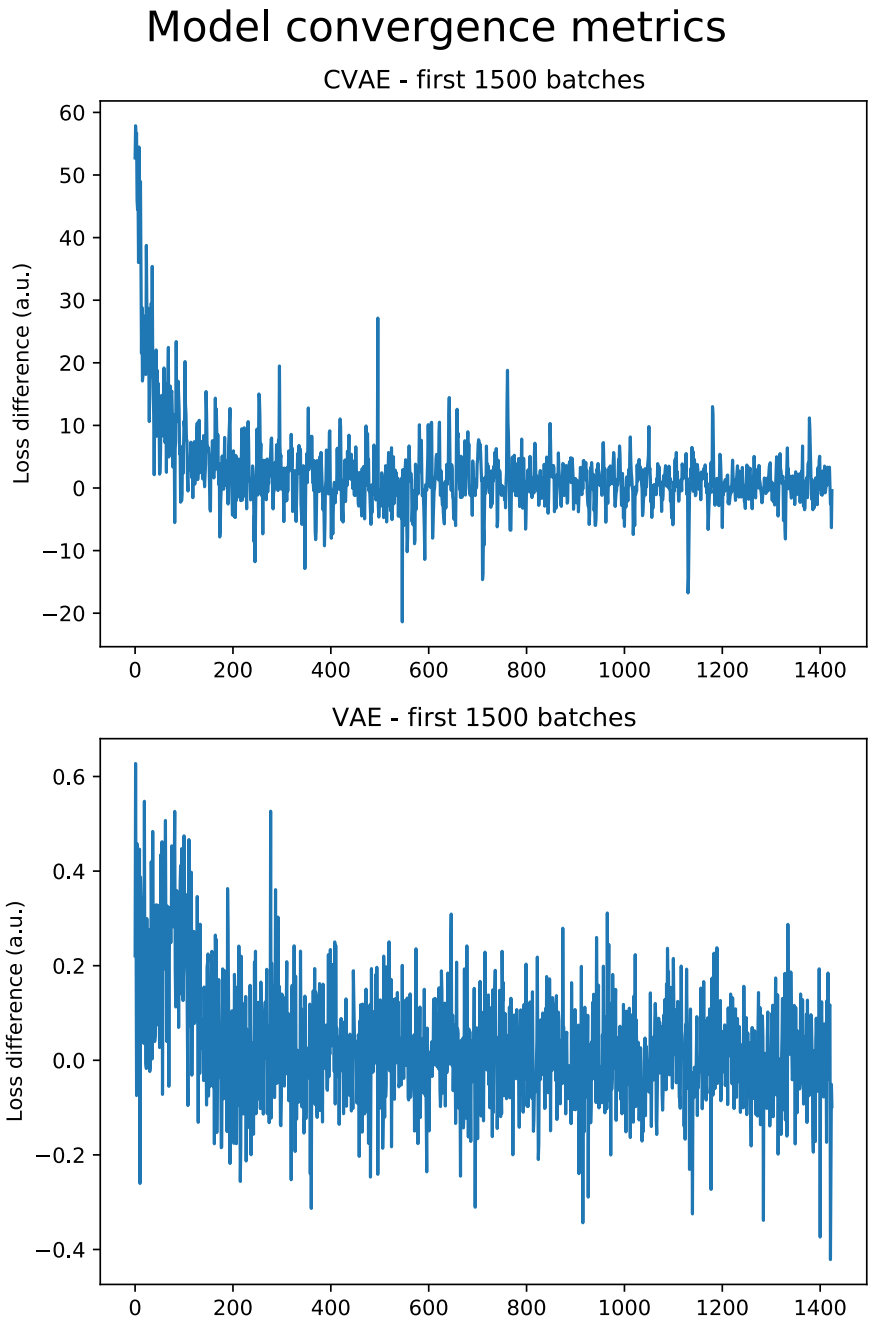

Figure S3: Training loss for the contrastive variational autoencoder (CVAE) and control variational autoencoder (VAE). The figure shows training loss difference for the first 1500 batches. Training was stopped when mean-squared-error between input batch and reconstruction fell below .005.

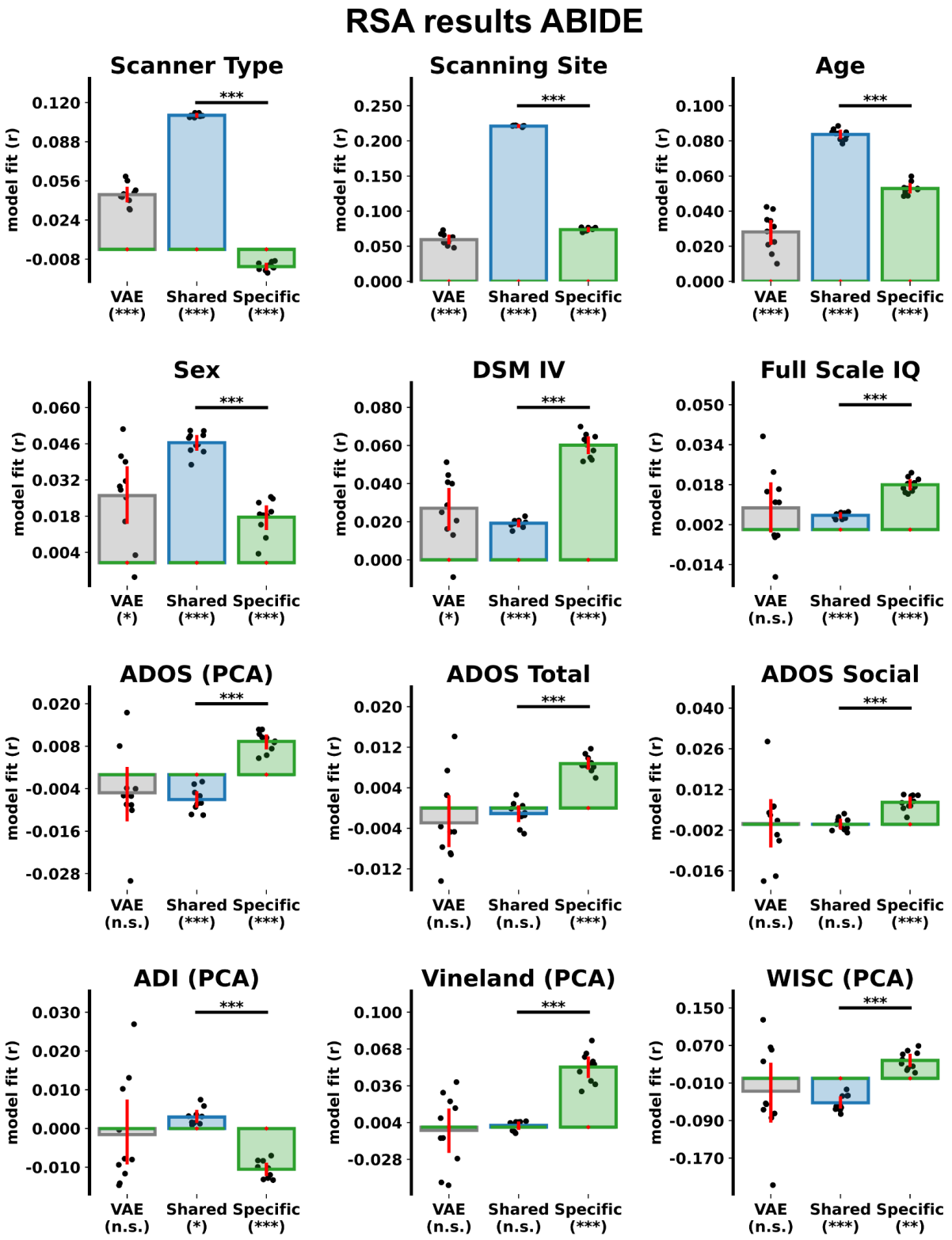

Figure S4: Representational Similarity Analysis (RSA) results (ABIDE dataset). Bar plots show representational similarity between neuroanatomical features (VAE, CVAE shared and CVAE ASD-specific) and patient properties. Significance indicated by stars (* $p < .05$, *** $p < .0001$,)

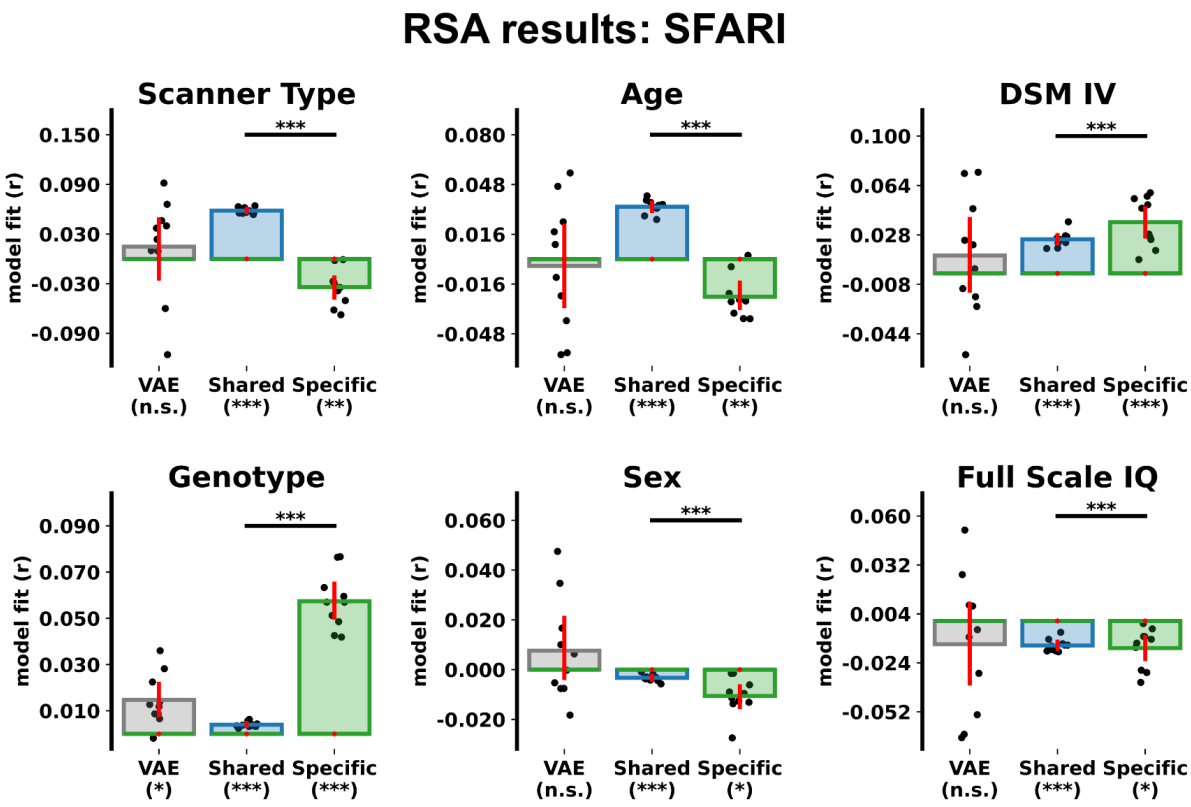

Figure S5: Representational Similarity Analysis (RSA) results (independent SFARI data). Bar plots show representational similarity between neuroanatomical features (VAE, CVAE shared and CVAE ASD-specific) and patient properties. (* $p < .05$, ** $p < .001$, *** $p < .0001$,).

Table S1: Architecture of the CVAE model.

Layer	Output Shape	# Param	Connected to
	Shared encoder		
inputs (InputLayer)	[(None, 64, 64, 64,1)]	0	
conv3d (Conv3D)	(None, 32, 32, 32, 64)	1792	inputs[0][0]
conv3d_1 (Conv3D)	(None, 16, 16, 16, 128)	221312	conv3d[0][0]
flatten (Flatten)	(None, 524288)	0	conv3d_1[0][0]
dense (Dense)	(None, 128)	67108992	flatten[0][0]
z_mean (Dense)	(None, 16)	2064	dense[0][0]
z_log_var (Dense)	(None, 16	2064	dense[0][0]
z (Lambda)	(None, 16)	0	z_mean[0][0] z_log_var[0][0]
	ASD-specific encoder		
inputs (InputLayer)	[(None, 64, 64, 64,1)]	0	
conv3d_2 (Conv3D)	(None, 32, 32, 32, 64)	1792	inputs[0][0]
conv3d_3 (Conv3D)	(None, 16, 16, 16, 128)	221312	conv3d_2[0][0]
flatten_1 (Flatten)	(None, 524288)	0	conv3d_3[0][0]
dense_1 (Dense)	(None, 128)	67108992	flatten_1[0][0]
s_mean (Dense)	(None, 16)	2064	dense_1[0][0]
s_log_var (Dense)	(None, 16	2064	dense_1[0][0]
s (Lambda)	(None, 16)	0	s_mean[0][0] s_log_var[0][0]
	Decoder		
z_sampling (InputLayer)	[(None, 32)]	0	
dense_2 (Dense)	(None, 128)	4224	z_sampling (InputLayer)
dense_3 (Dense)	(None, 524288)	67633152	dense_2 (Dense)
reshape (Reshape)	(None, 16, 16, 16, 128)	0	dense_3 (Dense)
conv3d_transpose (Conv3DT)	(None, 32, 32, 32, 32)	110624	reshape (Reshape)
conv3d_transpose_1 (Conv3DT)	(None, 64, 64, 64, 16)	13840	conv3d_transpose (Conv3DT)
decoder_output (Conv3DT)	(None, 64, 64, 64, 1)	433	conv3d_transpose_1 (Conv3DT)

Shared encoder parameters	67,336,224
ASD-specific encoder parameters	67,336,224
Decoder parameters	67,762,273
CVAE total parameters	202,434,721

Table S2: Architecture of the VAE model.

Layer	Output Shape	# Param	Connected to
	Encoder		
encoder_input (InputLayer)	(None, 64, 64, 64,1)]	0	
conv3d (Conv3D)	(None, 32, 32, 32, 96)	2688	encoder_input[0][0]
conv3d_1 (Conv3D)	(None, 16, 16, 16, 192)	497856	conv3d[0][0]
flatten (Flatten)	(None, 786432)	0	conv3d_1[0][0]
dense (Dense)	(None, 128)	100663424	flatten[0][0]
z_mean (Dense)	(None, 32)	4128	dense[0][0]
z_log_var (Dense)	(None, 32)	4128	dense[0][0]
z (Lambda)	(None, 32)	0	z_mean[0][0] , z_log_var[0][0]
	Decoder		
z_sampling (InputLayer)	[(None, 32)]	0	
dense_1 (Dense)	(None, 128)	4224	z_sampling (InputLayer)
dense_2 (Dense)	(None, 786432)	101449728	dense_1 (Dense)
reshape (Reshape)	(None, 16, 16, 16, 192)	0	dense_2 (Dense)
conv3d_transpose (Conv3DT)	(None, 32, 32, 32, 192	995520	reshape (Reshape)
conv3d_transpose_1 (Conv3DT)	(None, 64, 64, 64, 96)	497760	conv3d_transpose (Conv3DT)
decoder_output (Conv3DT)	(None, 64, 64, 64, 1)	2593	conv3d_transpose_1 (Conv3DT)

Encoder parameters	101,172,224
Decoder parameters	102,949,825
VAE total parameters	204,122,049

Table S3: Supplementary RSA analyses.

Property	N subjects	Specific > VAE	Shared > VAE
		ABIDE	
ADI (PCA)	284	$\Delta\tau$ = -0.01, t(9) = -2.00, p = 0.077	$\Delta\tau$ = 0.00, t(9) = 1.11, p = 0.296
ADOS (PCA)	284	$\Delta\tau$ = 0.01, t(9) = 3.52, p = 0.007	$\Delta\tau$ = 0.00, t(9) = -0.48, p = 0.643
ADOS Social	345	$\Delta\tau$ = 0.01, t(9) = 1.77, p = 0.110	$\Delta\tau$ = 0.00, t(9) = -0.06, p = 0.957
ADOS Total	369	$\Delta\tau$ = 0.01, t(9) = 4.40, p = 0.002	$\Delta\tau$ = 0.00, t(9) = 0.64, p = 0.536
Age	470	$\Delta\tau$ = 0.02, t(9) = 6.16, p < .001	$\Delta\tau$ = 0.06, t(9) = 15.23, p < .001
DSM IV	456	$\Delta\tau$ = 0.03, t(9) = 4.96, p < .001	$\Delta\tau$ = -0.01, t(9) = -1.38, p = 0.202
Full-Scale IQ	428	$\Delta\tau$ = 0.01, t(9) = 1.87, p = 0.094	$\Delta\tau$ = 0.00, t(9) = -0.61, p = 0.554
Scanner Type	470	$\Delta\tau$ = -0.06, t(9) = -19.73, p < .001	$\Delta\tau$ = 0.06, t(9) = 22.89, p < .001
Scanning Site	470	$\Delta\tau$ = 0.01, t(9) = 5.62, p < .001	$\Delta\tau$ = 0.16, t(9) = 59.85, p < .001
Sex	470	$\Delta\tau$ = -0.01, t(9) = -1.37, p = 0.203	$\Delta\tau$ = 0.02, t(9) = 3.07, p = 0.013
Vineland (PCA)	69	$\Delta\tau$ = 0.06, t(9) = 5.05, p < .001	$\Delta\tau$ = 0.00, t(9) = 0.42, p = 0.681
WISC (PCA)	22	$\Delta\tau$ = 0.07, t(9) = 2.15, p = 0.060	$\Delta\tau$ = -0.02, t(9) = -0.76, p = 0.468
		SFARI	
Full-Scale IQ	50	$\Delta\tau$ = 0.00, t(9) = -0.20, p = 0.848	$\Delta\tau$ = 0.00, t(9) = -0.06, p = 0.957
Genotype	51	$\Delta\tau$ = 0.04, t(9) = 6.49, p < .001	$\Delta\tau$ = -0.01, t(9) = -2.97, p = 0.016
Scanner Type	49	$\Delta\tau$ = -0.05, t(9) = -2.35, p = 0.043	$\Delta\tau$ = 0.04, t(9) = 2.25, p = 0.051
Sex	51	$\Delta\tau$ = -0.02, t(9) = -2.19, p = 0.056	$\Delta\tau$ = -0.01, t(9) = -1.68, p = 0.127