Contrastive AI reveals the structure of individual variation in Autism Spectrum Disorder

Aidas Aglinskas, Joshua K. Hartshorne, & Stefano Anzellotti

Department of Psychology and Neuroscience,
Boston College, Boston, MA 02467

Autism Spectrum Disorder is highly heterogeneous. Identifying systematic individual differences in neuroanatomy could inform diagnosis and personalized interventions. The challenge is that these differences are entangled with variation due to other causes: individual differences unrelated to ASD and measurement artifacts. We used contrastive deep networks to disentangle ASD-specific neuroanatomical variation from individual variation in common with typically developing participants. ASD-specific variation correlated with individual differences in symptoms. The structure of this ASD-specific variation also addresses a long-standing debate about the nature of ASD: individuals do not cluster into subtypes; instead, they are organized along at least two continuous dimensions – one associated with concurrently enlarged temporoparietal junction and somatosensory cortex and compressed superior longitudinal fasciculus, the other associated with changes in the thalamus.

Psychiatric disorders affect hundreds of millions of people worldwide [32]. Heterogeneity is a key obstacle to understanding them: different individuals diagnosed with the same type of disorder often present with unique profiles of behavioral symptoms and with different genetic

variants [66]. We investigated heterogeneity within autism spectrum disorder (ASD): a highly prevalent neurodevelopmental condition [38; 50] characterized by impaired social interactions, restricted patterns of behavior, and communication deficits [70]. Individuals with ASD differ substantially in the severity of social and behavioral symptoms [54; 59]. Moreover, ASD's genetic underpinnings are highly variable, with more than 200 associated copy number variations (CNVs), distributed differently across individuals and affecting multiple distinct pathways [34; 41; 48]. Understanding neuroanatomical heterogeneity within ASD could be pivotal to improving quality of life in affected individuals [26], by leading to more specific diagnoses and targeted behavioral interventions [39].

Recent studies attempted to investigate individual differences in neuroanatomy within the ASD population. However, researchers have struggled to identify systematic neural variation that correlates with symptoms and that replicates across different groups of participants [49; 56; 60; 65]. Different studies found different effects [13; 37; 56] or none at all [5].

We hypothesized that ASD-specific variation is being obscured by the many other factors that lead brains to vary. Brains differ from one to another due to numerous genetic and environmental causes unrelated to ASD [36]. Neuroanatomical data from different individuals, in addition, vary also because of methodological artifacts, such as systematic differences due to the scanners at different scanning sites [30; 43; 45]. ASD-specific variation may be difficult to identify within this mass of irrelevant variation. Researchers have developed several methods for addressing this problem, but they remain unsatisfactory. For instance, case-matching ASD and TD participants works in theory, but it assumes we know exactly what factors we need to match. However, brain anatomy is shaped by a multitude of genetic and environmental factors [36], some of which are unknown. This makes it very difficult (if not impossible) to match them all in practice.

In order to better characterize ASD-specific neuroanatomical variation, we employed an

alternative that disentangles it from variation in common with the general population. Specifically, we modeled structural MRI data using contrastive variational autoencoders (CVAEs, [55]). CVAEs take as inputs samples from two distinct populations, and can be trained to isolate features that capture variation specific to one population from features that are common to both (see Figure S1). We used CVAEs to disentangle features that capture anatomical variation within the ASD group ('ASD-specific') from 'shared' features that capture anatomical variation common to both ASD and TD participants (See Figure 1A).

First, we validated these neuroanatomical features by testing whether the ASD-specific features were differentially related to clinical symptoms while the shared features were differentially related to scanner type and non-clinical individual characteristics. We then replicated the results with a zero-free-parameter generalization to an independent dataset. Next, we explored the structure of the ASD-specific features, applying cluster analysis to determine whether there are distinct subtypes of ASD neuroanatomy. Finally, in order to identify anatomical loci of individual variation, we leveraged the properties of the CVAE to generate counterfactual TD brains closely matched to the brains of the ASD participants ("synthetic twins"). Comparing the anatomy of ASD participants to their synthetic twins allowed us to detect specific brain regions whose alterations vary within the ASD population.

Results

We used the ABIDE dataset – one of the largest available structural MRI datasets including both ASD participants (N = 470) and TD controls (N = 512) [31; 46] – to train a CVAE and a non-contrastive VAE matched in the number of parameters that uses a single latent space but has the same number of latent features as the CVAE. The non-contrastive VAE allows us to test whether associations between neuroanatomy and ASD symptoms can be identified using variational autoencoding alone (effectively, a powerful method for factor analysis) without explicitly

disentangling ASD-specific and shared variation.

Thus, to establish a baseline, we first report the non-contrastive VAE results. We used Representational Similarity Analysis (RSA; [11]) to test whether the VAE's neuroanatomical feature space correlates with individual variation within the ASD participants in key clinical and non-clinical individual characteristics, such as ADOS scores (a numerical measure of ASD symptom severity), Vineland adaptive behavior scores, scanner type, and age. More specifically, we separately calculated the pairwise dissimilarity between participants with respect to the VAE neuroanatomical feature space and with respect to a set of clinical and non-clinical individual characteristics, obtaining a dissimilarity matrix for the VAE feature space and one for each characteristic (Figure 1B). The VAE feature space showed Kendall τ correlations with some of the non-clinical characteristics, such as scanner type (τ = 0.04, t(9) = 16.29, p $<$ .001), age (τ = 0.03, t(9) = 8.27, p $<$ .001) and gender (τ = 0.03, t(9) = 4.71, p = 0.001). While there was some relationship between neuroanatomical feature similarity extracted by VAE and DSM-IV behavioral subtypes (τ = 0.03, t(9) = 4.77, p = 0.001), there was no relationship with autism severity (ADOS total; τ = 0.00, t(9) = -1.08, p = 0.310) or Vineland adaptive behavior scores (τ = 0.00, t(9) = -0.29, p = 0.780). This is consistent with the idea presented above, namely that entangled measures of neuroanatomy (such as VAE features) fail to capture more subtle variations in symptom severity.

We then assessed whether explicitly disentangling ASD-specific and shared neuroanatomical variation with a CVAE would allow us to identify clinically-relevant individual variation. As described above, the CVAE segregates its internal representations into ASD-specific and shared latent features (Figure 1A, Figure S2). The CVAE was trained on the same ABIDE data. Critically, while the CVAE training implicitly makes a binary distinction between ASD and TD subjects, the model is not provided with any of the clinical and non-clinical individual characteristics of interest, such as continuous measures of symptomatology like ADOS. We used RSA

to compare the CVAE's ASD-specific and shared neuroanatomical spaces to each of the clinical and non-clinical individual characteristics. We expected to find that shared features correlate with non-clinical variation common to both ASD and TD participants, while ASD-specific features correlate with clinical ASD variation (Figure 1B).

Scanner type was strongly associated with subject similarity in the shared feature space (τ = 0.11, t(9) = 253.01, p $<$ .001) but not the ASD-specific space (τ = -0.01, t(9) = -14.16, p $<$ .001; shared vs ASD-specific: $\Delta\tau$ = 0.12, t(9) = 124.83, p $<$ .001). Thus, the CVAE was able to factor out a common source of "nuisance" variation in multi-site data [43]. In contrast, measures of ASD clinical symptoms were more associated with the ASD-specific space but generally not the shared space. These include DSM IV behavioral subtypes (ASD-specific: τ = 0.06, t(9) = 30.83, p $<$ .001; shared: τ = 0.02, t(9) = 29.02, p $<$ .001; comparison: $\Delta\tau$ = 0.04, t(9) = 20.04, p $<$ .001), ADOS total score (ASD-specific: τ = 0.01, t(9) = 16.85, p $<$ .001; shared: τ = 0.00, t(9) = -1.50, p = 0.167; comparison: $\Delta\tau$ = 0.01, t(9) = 11.59, p $<$ .001), and Vineland adaptive behavior questionnaire (ASD-specific: τ = 0.05, t(9) = 12.33, p $<$ .001; shared: τ = 0.00, t(9) = 1.17, p = 0.270; comparison: $\Delta\tau$ = 0.05, t(9) = 10.46, p $<$ .001) (see also Figure S4, SOM). Thus, the CVAE was not only able to disentangle neuroanatomical variation specific to ASD from general-population variation, but this neuroanatomical variation was differentially associated with ASD-specific and general-population phenotypic and demographic characteristics.

Results for age, gender, and full-scale IQ (FIQ) were of particular interest, since these are known to differently relate to neuroanatomy in the TD and ASD populations [33; 62]. Thus, it is not surprising that each of these was significantly related to both the ASD-specific space (age: τ = 0.05, t(9) = 48.60, p $<$ .001; gender: τ = 0.02, t(9) = 8.13, p $<$ .001, FIQ: τ = 0.02, t(9) = 20.22, p $<$ .001) and the shared space (age: τ = 0.08, t(9) = 89.29, p $<$ .001, gender: τ = 0.05, t(9) = 35.34, p $<$ .001, FIQ: τ = 0.01, t(9) = 15.57, p $<$ .001), indicating that the CVAE was able to disentangle general effects of age, gender, and FIQ from those that specifically

interact with ASD. As expected, shared features captured greater variation in age and gender than ASD-specific features (age: $\Delta\tau$ = 0.03, t(9) = 24.11, p $<$ .001; gender: $\Delta\tau$ = 0.03, t(9) = 11.90, p $<$ .001). Conversely, variation in full-scale IQ was more related to ASD-specific than shared features ($\Delta\tau$ = 0.01, t(9) = 12.86, p $<$ .001).

In sum, the CVAE was not only able to disentangle individual neuroanatomical variation that is specific to ASD from variation that characterizes the population as a whole, but these patterns of variation were differentially associated with clinical and non-clinical participant characteristics. This contrasts with the control VAE model, whose unitary neuroanatomical space showed weaker correlations with individual characteristics. Indeed, there was no individual characteristic that was most strongly associated with the control VAE space (Figure S4, SOM).

Generalization to an independent dataset.

Generalization to a new dataset that was not used for training is a very stringent test of a model's performance. In the context of diagnosis, generalization across datasets is especially important, because a model trained on one group of participants may need to be used to inform the diagnosis of novel participants that were not included in the training dataset. To test generalization to a new dataset, we applied the ABIDE-trained CVAE to the anatomical scans of participants from the SFARI VIP dataset (N = 121) [22] using a parameter-free fit. SFARI VIP is a particularly interesting test case because it includes information about ASD-relevant CNVs, allowing us to study whether ASD-specific neuroanatomical features correlate with genotype.

We expected that if CVAE features were robust, shared features should have high correlations with properties of scanning site, age and gender, and ASD-features should correlate with ASD-related individual differences such as DSM-IV behavioral subtypes. This is exactly what we found. Compared to ASD-specific features, shared features correlated better with scanner-type ($\Delta\tau$ = 0.09, t(9) = 12.81, p $<$ .001), age ($\Delta\tau$ = 0.06, t(9) = 15.09, p $<$ .001) and gender

($\Delta\tau$ = 0.01, t(9) = 3.17, p = 0.011). In contrast, ASD-specific features, compared to shared features, had higher Kendall tau values for DSM-IV behavioral subtypes ($\Delta\tau$ = 0.01, t(9) = 2.34, p = 0.044) suggesting that CVAE identified population-wide patterns of neuroanatomy, some of which are shared between ASD and non-ASD participants, and some of which are only present in those with ASD. Additionally, the SFARI VIP dataset allowed us to ask whether neuroanatomical differences observed in 16p11.2 deletion and duplication carriers are consistent with patterns of variation in the typically developing population, or whether they match patterns of variation within ASD. Similarity between deletion and duplication CNVs was better reflected in ASD-specific than shared features ($\Delta\tau$ = 0.05, t(9) = 14.54, p $<$ .001).

The nature of variation

Researchers have long debated whether individual differences in ASD are better understood as distinct subtypes or as variation along continuous dimensions [24; 49; 57; 65]. Having identified a space of ASD-specific features makes it possible to test these hypotheses directly. We used Gaussian mixture modelling [2] to identify clusters of subjects in each of the feature spaces, and we used the Bayesian Information Criterion (BIC [1]) to select the optimal number of clusters (Figure 1D).

The distribution of subjects in the VAE feature space, which conflates ASD-specific and shared variation, was consistent with a single cluster (p $<$ .01): one cluster was the optimal solution for 100/100 samples of the latent space (Figure 1D). The CVAE results were more nuanced. The shared features showed evidence of subtypes, revealing more than one cluster (p $<$ .01): the one cluster solution was optimal in 0/100 samples. Instead, the three cluster solution provided the best fit (selected in 88/100 samples), followed by the two cluster solution (12/100 samples) (binomial two-tailed test p $<$ .001). However, the subject distribution in the space of ASD-specific features again suggested continuous variation: the optimal solution assigned all

participants to a single cluster in 100/100 samples ($p < .01$). Thus, the results of cluster analysis show that once disentangled from typical variation, ASD-related neuroanatomical variation is better captured by continuous dimensions, rather than by discrete categories.

Neuroanatomical interpretation

Even though variation in the space of ASD-specific features is better described by continuous dimensions than by distinct categories, different dimensions might nonetheless be associated with different anatomical changes. Whether ASD-specific individual differences in neuroanatomy are focal (affecting one or few regions), or distributed (affecting many regions) - remains an open question. To address it, we tested whether different ASD-specific features are associated with neuroanatomical changes in localized brain regions. We identified loci of anatomical variation between ASD subjects following a three step process. First, for each ASD participant, we generated a simulated brain of a counterfactual "synthetic TD twin" by setting ASD-specific feature values to zero and using the CVAE decoder to reconstruct their brain only using the inferred shared features (see: Methods in Supplementary Materials (SM)). This "synthetic twin" is effectively a case-control for each ASD participant's brain, matched on shared variation features (including scan site artifacts as well as neuroanatomical variation due to age, gender, and other dimensions discovered from the data by the CVAE). In the second step, we estimated a nonrigid transformation morphing the counterfactual TD brain to match the corresponding ASD participant's brain. This produced a vector field describing the differences between the ASD brain and the corresponding TD brain (see: Methods in SM). Finally, we calculated the Jacobian determinant of the vector field. This measure captures the local volumetric compression/expansion needed to morph the simulated TD brain into the corresponding ASD brain. Repeating this procedure for multiple participants with different values of ASD-specific features, we computed interpretable gray and white matter alterations that vary across the ASD

participant population.

To organize the search of interpretable neuroanatomical features, we projected the 16-dimensional space of ASD-specific neuroanatomical features on a two-dimensional subspace (using UMAP, [53]). We then measured systematic variation in the compression/expansion of different brain regions along each of the two principal dimensions by comparing the Jacobian determinant maps of subjects ranking high (N = 50) and low (N = 50) on each dimension (Figure 2). By focusing on the two largest, most informative UMAP dimensions, we simplify interpretation and reduce the number of comparisons. Using more dimensions might reveal additional anatomical loci, but it would diminish power due to the need to correct for multiple comparisons.

Comparing extremes of the x dimension (negative vs. positive), we observed enlarged parietal cortex and compressed superior longitudinal fasciculus (SLF). The affected portion of parietal cortex (734 voxels in size) is centered in secondary somatosensory cortex (SII, MNI x = 62 y = -22 z = 24) and extends posteriorly into the temporoparietal junction (TPJ). The SLF cluster (466 voxels in size) is centered at MNI coordinates x = 32 y = -22 z = 30, and extends anteriorly terminating in the inferior frontal gyrus (BA44/BA45 MNI x = 42 y = 22 z = 30). Anatomical locations of SLF and SII were confirmed using the Juelich Histological Atlas [61]. The TPJ location was confirmed using Neurosynth's [19] meta-analytic activation maps generated using the "theory of mind" search term (meta analysis of 181 studies).

Comparing extremes of the y dimension (negative vs. positive), we observed deformations corresponding to the thalamus bilaterally - more prominently expressed on the left hemisphere. The deformation cluster spanned 537 voxels and was centered at MNI x = 6.35, y = -12.31 z = .53, $p < .048$ (cluster-level FWE corrected, two-tailed test). We used the Oxford thalamic connectivity probability atlas [4] to determine the anatomical location more precisely. This procedure confirmed that the cluster is centered on a subdivision of the thalamus with predom-

inant connections to prefrontal cortex previously associated with individual differences in ASD symptom severity, such as ADOS scores [42; 68].

Discussion

Behavioral heterogeneity is a defining characteristic of ASD; to understand its biological basis we need to identify systematic neural heterogeneity. Furthermore, individuals with similar behavioral symptoms might have different anatomical alterations that could help clinicians to select distinct, tailored interventions.

Unfortunately, identifying systematic neuroanatomical variation between individuals with ASD has proven challenging [44; 65], due to variation caused by ASD-unrelated factors. Using a contrastive variational autoencoder (CVAE), we revealed a space of neuroanatomical features that captures ASD-specific variation, disentangled from neuroanatomical variation common to both ASD and TD participants. We then found that individual differences in ASD-specific neuroanatomical features are associated with individual differences in diagnosis and ASD-related genotype. Additional analyses identified two distinguishable components of this neural variation, linked to localized anatomical changes. One dimension of neuroanatomical variability is right lateralized and affects the secondary somatosensory cortex (SII), the temporoparietal junction (TPJ), and the superior longitudinal fasciculus (SLF); the other affects the thalamus bilaterally. These findings demonstrate that different variants of ASD can have a concerted impact on multiple anatomical structures, affecting in conjunction cortical areas and white matter tracts.

Analyses further demonstrated that ASD-specific neuroanatomical differences are best characterized by continuous variation along at least two dimensions (Figure 2), rather than by distinct subtypes as suggested by prior work [49; 57; 65]. This evidence indicates that – at least at the level of neuroanatomy – dimensional approaches (such as the Research Domain Crite-

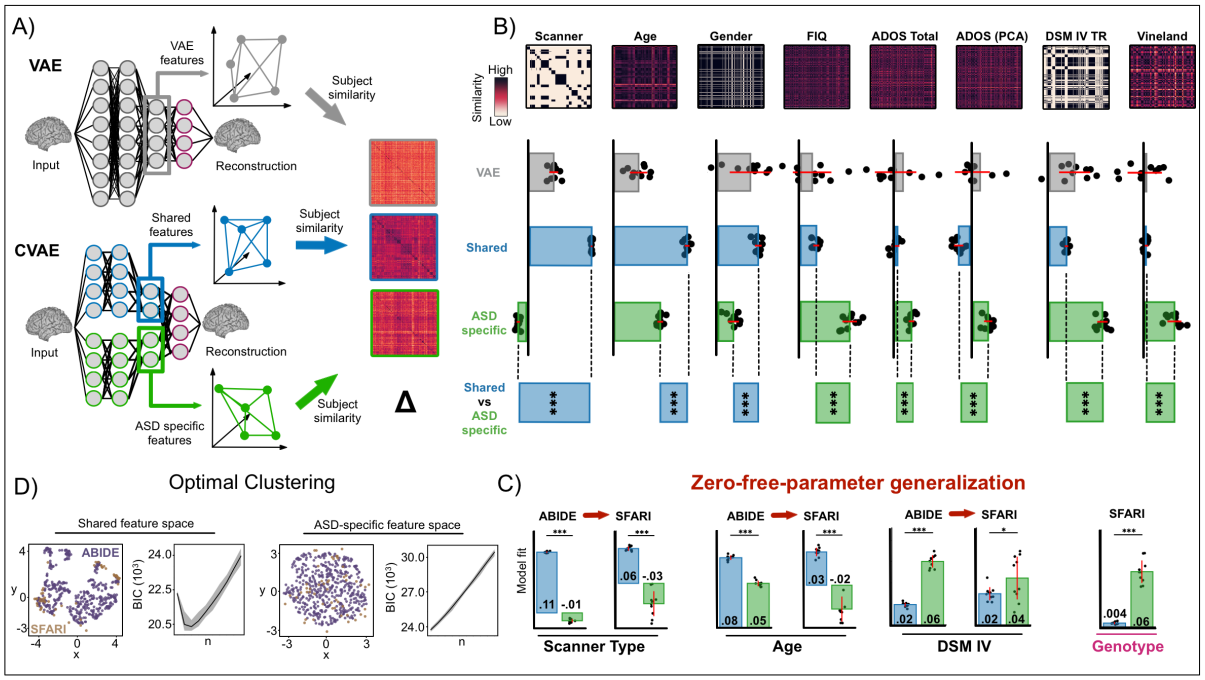

Figure 1: Patterns of individual variation captured in the learned latent spaces. **A)** Representations of the subjects' neuroanatomy in the three latent feature spaces (VAE: gray, CVAE shared: blue, and CVAE ASD-specific: green) are used to construct neuroanatomical similarity matrices. **B)** Neuroanatomical similarity matrices are compared to models of similarity based on different participant properties (e.g. demographic variables, clinical measurements). Significance values indicated by stars, $p < .05$ (*), $p < .0001$ (***). Variables common to TD and ASD participants are best captured by the shared latent neuroanatomical features. Variables associated with ASD-related variation are best captured by the ASD-specific latent neuroanatomical features. Model fit for control model (VAE) is noticeably poorer across all variables. **C)** *Zero-free-parameter generalization*. RSA analyses using SFARI data that was not used to train or tune the CVAE. In both ABIDE and SFARI datasets, the shared space shows higher Kendall τ values with variation in scanner-type and age, and ASD-specific space has higher Kendall τ values with variation associated with ASD severity (DSM IV behavioral subtypes). Participants sharing similar copy-number-variations associated with increased risk of ASD (16p11.2 deletion or duplication) are more similar in ASD-specific, but not shared neuroanatomical features. **D)** *Optimal clustering*. Scatter-plots show inter-subject similarity of ABIDE ASD (Purple dots, N = 492) and SFARI 16p11.2 CNV (Orange dots, deletion N = 26, duplication N = 25) subjects in shared and ASD-specific feature spaces after dimensionality reduction using UMAP. Line plots show Gaussian mixture model fit (Bayesian Information Criterion; BIC) for different numbers of clusters (1-10). Clustered structure can be seen in the shared feature space (optimal clustering n = 3) but not in the ASD-specific feature space (optimal clustering n = 1).

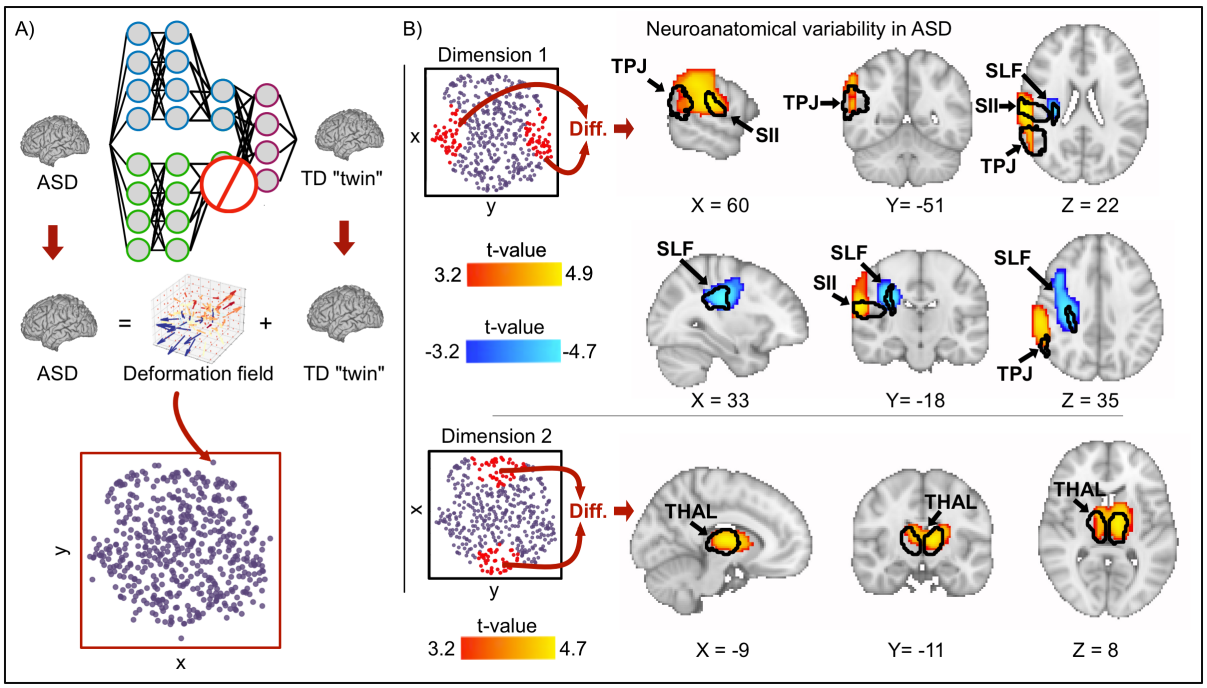

Figure 2: Anatomical loci of individual variation within the ASD group. **A)** For each ASD subject we calculated a synthetic "TD-twin" brain that is matched on ASD-unrelated neuroanatomical features, by setting to zero the features capturing ASD-specific variation. We then computed a nonrigid transformation required to morph the synthetic "TD-brain" into the corresponding ASD brain, obtaining a deformation field. **B)** Comparing the deformation fields along different dimensions in the space of ASD specific features (after UMAP dimensionality reduction) revealed individual variation in specific anatomical loci. Differences along the x-axis revealed an expansion of the parietal lobe, accompanied by a compression of the superior longitudinal fasciculus (SLF). The parietal effects were centered on the secondary somatosensory cortex (SII) and extended posteriorly into temporoparietal junction (TPJ). Along the y-axis we observed expansion of the thalamus (THAL), and specifically in subregions of thalamus connected with prefrontal cortex. Anatomical locations were confirmed using the Juelich histological atlas (SLF, SII), Oxford Thalamic Connectivity atlas (Thalamus) and Neurosynth meta-analytic activation map for theory-of-mind tasks (TPJ).

ria: RDoC, [71]) can provide a better account of individual variation than discrete diagnostic categories.

Individual differences in ADOS and Vineland scores were uniquely associated with ASD-specific neuroanatomical features, and differences introduced by the types of fMRI scanners used were uniquely associated with shared features. Of particular interest were properties that are shared by TD and ASD participants, but are linked to distinct development patterns in the ASD population: age, gender and FIQ [62]. These properties showed positive Kendall τ values for both shared and ASD-specific anatomical features, suggesting that they have an impact on both. Previous studies have documented how the brain matures differently in ASD, and that boys and girls with ASD have different ASD-related neuroanatomical differences [62]. Higher FIQ scores are associated with thicker cortex in TD participants, but with thinner cortex in ASD [62]. The significant correlations between the ASD-specific space and these properties indicate that its features are able to capture ASD-specific patterns of age, gender and FIQ related variability. More broadly, the ability of the CVAE to capture interactions between ASD and development makes it a promising approach to investigate other developmental disorders as well.

Our analyses suggested a key role for two dimensions of neuroanatomical variation. These two dimensions organize a variety of prior findings within a unified framework. One dimension of variability was characterized by a concurrent expansion of the parietal cortex, covering somatosensory and temporoparietal cortices, and compression of the SLF. Greater somatosensory cortex gray matter volume is associated with ASD-related tactile sensitivity [7; 35; 47]. Impaired anisotropy in the SLF, as revealed by diffusion tensor imaging, correlates with language delay in ASD [21] and successful behavioral intervention improving ASD-associated language outcomes proportionately improves the integrity of the SLF tract [58]. Our results showing that somatosensory cortex and SLF abnormalities covary within ASD help elucidate established be-

havioral observations that hypo-reactivity to sensory stimuli (including touch) is associated with impaired language development in children with ASD [9; 25].

The posterior part of the parietal cluster showed substantial overlap with a portion of right TPJ (rTPJ) within the Theory of Mind network: a key region for social cognition. Social impairments are a defining symptom of ASD [70]. Evidence of structural abnormalities is pivotal in establishing this region as a potential locus of social cognition related symptoms in ASD, but reports of altered cortical structure in rTPJ have been scarce [16; 23]. In addition, functional MRI studies have found largely similar amounts of activity in the rTPJ of ASD and TD participants [28]. However, recent work using multivariate analyses identified atypical development of rTPJ representations in ASD participants [64]. Our findings provide converging evidence of rTPJ alterations in ASD, and suggest that the observed differences in fMRI response patterns are likely due to abnormal gray matter development in this region (and not just to upstream alterations affecting the inputs received by rTPJ).

Another axis of variability was localized entirely in the thalamus, consistent with prior observation of substantial thalamic variability in ASD [6; 17; 20] such as above-average and below-average thalamic volumes in 16p11.2 duplication and deletion CNV [34]. While the relationship between thalamic variability and specific ASD symptom domains remains unclear [8; 51; 68], genetic studies have observed thalamic impairments to be common across multiple ASD associated mutations, particularly copy number deletions [52; 63; 67].

Considering ASD as a case study, the links between the present results and the previous literature provide an illustration of how computational psychiatry [40], and more specifically large-scale studies of individual differences that disentangle disorder-specific neural variation, can contribute to our understanding of complex disorders, enabling a synthesis of findings across multiple specialized studies focusing on particular facets.

To reach a more comprehensive understanding of ASD, several future steps are needed.

First, structural variation does not exhaust the ways in which brains can differ: results may differ for functional connectivity, diffusion tensor imaging, or evoked responses. Critically, the methods presented here can be straightforwardly applied to other data modalities. Second, the scope of our investigation of genotype and its relation to neuroanatomical variation was limited by the data available. Datasets with richer genetic information could reveal groups of patients with distinct genetic alterations affecting common pathways. Finally, finer grained measures of behavior will be needed to gain a deeper understanding of the interplay between genetics, anatomy and symptoms.

Our findings were enabled by a key methodological innovation: the use of CVAEs to parcel individual variation into shared variation and variation specific to ASD. As compared to classical case-control matching, this approach has the advantage that it can discover anatomical features that vary also among TD participants in a data-driven fashion. Importantly, this makes it possible (in theory) to control for confounds that are difficult to measure explicitly, such as ASD-unrelated genetic variation and effects of the environment [36]. The results in the present study show that the CVAE is effective in practice, and the comparison with a VAE matched in the number of parameters demonstrates that the contrastive architecture is essential. The features discovered by the CVAE generalize well to an independent dataset – this is especially important for the study of psychiatric disorders, because a model trained with a given dataset may need to be applied to data from new participants to inform diagnosis. Importantly, this novel approach for identifying disorder-specific variability in neuroimaging data can be applied to a variety of neurodevelopmental disorders, paving the way to the search for neurally-informed personalized interventions.

Acknowledgements

This work was supported by a grant from the Simons Foundation (SFARI 614379), awarded to

Joshua Hartshorne and Stefano Anzellotti.

References

[1] Ishiguro, Makio / others u.a.(1981): *A Bayesian approach to the analysis of the data of crustal movements*.

[2] Yang, Ming Hsuan / Ahuja, Narendra(1998): *Gaussian mixture model for human skin color and its applications in image and video databases*In: Storage and retrieval for image and video databases VII458–466.

[3] LeCun, Yann / Bottou, Léon / Bengio, Yoshua / Haffner, Patrick(1998): *Gradient-based learning applied to document recognition*, 11: 2278–2324.

[4] Behrens, T E J u.a.(2003): *Non-invasive mapping of connections between human thalamus and cortex using diffusion imaging*, 7: 750–757.

[5] Hrdlicka, Michal u.a.(2005): *Subtypes of autism by cluster analysis based on structural MRI data*, 3: 138–144.

[6] Hardan, Antonio Y / Girgis, Ragy R / Adams, Jason / Gilbert, Andrew R / Keshavan, Matcheri S / Minshew, Nancy J(2006): *Abnormal brain size effect on the thalamus in autism*, 2-3: 145–151.

[7] Rojas, Donald C / Peterson, Eric / Winterrowd, Erin / Reite, Martin L / Rogers, Sally J / Tregellas, Jason R(2006): *Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms*56.

[8] Hardan, Antonio Y / Minshew, Nancy J / Melhem, Nadine M / Srihari, Sumana / Jo, Booil / Bansal, Rahul / Keshavan, Matcheri S / Stanley, Jeffrey A(2008): *An MRI and proton spectroscopy study of the thalamus in children with autism*, 2: 97–105.

[9] Takarae, Yukari / Luna, Beatriz / Minshew, Nancy J / Sweeney, John A(2008): *Patterns of visual sensory and sensorimotor abnormalities in autism vary in relation to history of early language delay*, 6: 980–989.

[10] Kriegeskorte, Nikolaus / Mur, Marieke / Bandettini, Peter(2008): *Representational similarity analysis - connecting the branches of systems neuroscience*4.

[11] Kriegeskorte, Nikolaus / Mur, Marieke / Bandettini, Peter A(2008): *Representational similarity analysis-connecting the branches of systems neuroscience*4.

[12] Avants, B B / Epstein, C L / Grossman, M / Gee, J C(2008): *Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain*, 1: 26–41.

[13] Stanfield, Andrew C / McIntosh, Andrew M / Spencer, Michael D / Philip, Ruth / Gaur, Sonia / Lawrie, Stephen M(2008): *Towards a neuroanatomy of autism: a systematic review and meta-analysis of structural magnetic resonance imaging studies*, 4: 289–299.

[14] Avants, Brian B / Tustison, Nick / Song, Gang(2009): *Advanced normalization tools (ANTS)*, 365: 1–35.

[15] Deng, Jia / Dong, Wei / Socher, Richard / Li, Li Jia / Li, K. / Fei Fei, Li(2009): *ImageNet: A large-scale hierarchical image database*248-255.

[16] Zoccante, Leonardo u.a.(2010): *Increased left parietal volumes relate to delayed language development in autism: a structural mri study*, 4: 217–221.

[17] Tamura, Ryu / Kitamura, Hideaki / Endo, Taro / Hasegawa, Naoya / Someya, Toshiyuki(2010): *Reduced thalamic volume observed across different subgroups of autism spectrum disorders*, 3: 186–188.

[18] Hua, Xue u.a.(2011): *Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry*, 1: 5–14.

[19] Yarkoni, Tal / Poldrack, Russell A / Nichols, Thomas E / Van Essen, David C / Wager, Tor D(2011): *Large-scale automated synthesis of human functional neuroimaging data*, 8: 665–670.

[20] Ecker, Christine u.a.(2012): *Brain anatomy and its relationship to behavior in adults with autism spectrum disorder: a multicenter magnetic resonance imaging study*, 2: 195–209.

[21] Bakhtiari, Reyhaneh u.a.(2012): *Differences in white matter reflect atypical developmental trajectory in autism: A Tract-based Spatial Statistics study*, 1: 48–56.

[22] Consortium, Simons VIP / others u.a.(2012): *Simons Variation in Individuals Project (Simons VIP): a genetics-first approach to studying autism spectrum and related neurodevelopmental disorders*, 6: 1063–1067.

[23] Knaus, Tracey A / Tager Flusberg, Helen / Foundas, Anne L(2012): *Sylvian fissure and parietal anatomy in children with autism spectrum disorder*, 4: 327–339.

[24] Grzadzinski, Rebecca / Huerta, Marisela / Lord, Catherine(2013): *DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes*, 1: 12.

[25] Baranek, Grace T / Watson, Linda R / Boyd, Brian A / Poe, Michele D / David, Fabian J / McGuire, Lorin(2013): *Hyporesponsiveness to social and nonsocial sensory stimuli in children with autism, children with developmental delays, and typically developing children*, 2: 307–320.

[26] Georgiades, Stelios / Szatmari, Peter / Boyle, Michael(2013): *Importance of studying heterogeneity in autism*, 2: 123–125.

[27] Lin, Mingfeng / Lucas, Henry C / Shmueli, Galit(2013): *Research Commentary—Too Big to Fail: Large Samples and the p-Value Problem*, 4: 906–917.

[28] Dufour, Nicholas / Redcay, Elizabeth / Young, Liane / Mavros, Penelope L / Moran, Joseph M / Triantafyllou, Christina / Gabrieli, John DE / Saxe, Rebecca(2013): *Similar brain activation during false belief tasks in a large sample of adults with and without autism*, 9: e75468.

[29] Kingma, Diederik P / Ba, Jimmy(2014): *Adam: A Method for Stochastic Optimization*.

[30] Di Martino, A u.a.(2014): *The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism*, 6: 659–667.

[31] Di Martino, Adriana u.a.(2014): *The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism*, 6: 659–667.

[32] Steel, Zachary / Marnane, Claire / Iranpour, Changiz / Chey, Tien / Jackson, John W / Patel, Vikram / Silove, Derrick(2014): *The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013*, 2: 476–493.

[33] Zielinski, Brandon A u.a.(2014): *Longitudinal changes in cortical thickness in autism and typical development*, Pt 6: 1799–1812.

[34] Qureshi, Abid Y u.a.(2014): *Opposing brain differences in 16p11.2 deletion and duplication carriers*, 34: 11199–11211.

[35] Duerden, Emma G / Card, Dallas / Roberts, S Wendy / Mak Fan, Kathleen M / Chakravarty, M Mallar / Lerch, Jason P / Taylor, Margot J(2014): *Self-injurious behaviours are associated with alterations in the somatosensory system in children with autism spectrum disorder*, 4: 1251–1261.

[36] Gu, Jenny / Kanai, Ryota(2014): *What contributes to individual differences in brain structure?*262.

[37] Sussman, D u.a.(2015): *The autism puzzle: Diffuse but not pervasive neuroanatomical abnormalities in children with ASD*170–179.

[38] Boat, Thomas F / Wu, Joel T(2015): *Mental disorders and disabilities among low-income children*.

[39] Higdon, Roger u.a.(2015): *The promise of multi-omics and clinical data integration to identify and target personalized healthcare approaches in autism spectrum disorders*, 4: 197–208.

[40] Huys, Quentin JM / Maia, Tiago V / Frank, Michael J(2016): *Computational psychiatry as a bridge from neuroscience to clinical applications*, 3: 404–413.

[41] An, Joon Yong / Claudianos, Charles(2016): *Genetic heterogeneity in autism: From single gene to a pathway perspective*442–453.

[42] Schuetze, Manuela / Park, Min Tae M / Cho, Ivy Yk / MacMaster, Frank P / Chakravarty, M Mallar / Bray, Signe L(2016): *Morphological Alterations in the Thalamus, Striatum, and Pallidum in Autism Spectrum Disorder*, 11: 2627–2637.

[43] Auzias, G / Takerkart, S / Deruelle, C(2016): *On the Influence of Confounding Factors in Multisite Brain Morphometry Studies of Developmental Pathologies: Application to Autism Spectrum Disorder*, 3: 810–817.

[44] Riddle, Kaitlin / Cascio, Carissa J / Woodward, Neil D(2017): *Brain structure in autism: a voxel-based morphometry analysis of the Autism Brain Imaging Database Exchange (ABIDE)*, 2: 541–551.

[45] Di Martino, Adriana u.a.(2017): *Enhancing studies of the connectome in autism using the autism brain imaging data exchange II*170010.

[46] Di Martino, Adriana u.a.(2017): *Enhancing studies of the connectome in autism using the autism brain imaging data exchange II*, 1: 1–15.

[47] Wang, Jia u.a.(2017): *Increased Gray Matter Volume and Resting-State Functional Connectivity in Somatosensory Cortex and their Relationship with Autistic Symptoms in Young Boys with Autism Spectrum Disorder*588.

[48] Tordjman, S / Cohen, D / Coulon, N / Anderson, GM / Botbol, M / Canitano, R / Roubertoux, PL(2017): *Reframing autism as a behavioral syndrome and not a specific mental disorder: Implications of genetic and phenotypic heterogeneity*210.

[49] Hong, Seok Jun / Valk, Sofie L / Di Martino, Adriana / Milham, Michael P / Bernhardt, Boris C(2018): *Multidimensional neuroanatomical subtyping of autism spectrum disorder*, 10: 3578–3588.

[50] Christensen, Deborah L u.a.(2018): *Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2012*, 13: 1.

[51] Hegarty, John P u.a.(2018): *A proton MR spectroscopy study of the thalamus in twins with autism spectrum disorder*153–160.

[52] Martin Brevet, Sandra u.a.(2018): *Quantifying the Effects of 16p11.2 Copy Number Variants on Brain Structure: A Multisite Genetic-First Study*, 4: 253-264.

[53] McInnes, Leland / Healy, John / Melville, James(2018): *Umap: Uniform manifold approximation and projection for dimension reduction*.

[54] Lombardo, Michael V / Lai, Meng Chuan / Baron Cohen, Simon(2019): *Big data approaches to decomposing heterogeneity across the autism spectrum*, 10: 1435–1450.

[55] Abid, Abubakar / Zou, James(2019): *Contrastive Variational Autoencoder Enhances Salient Features*.

[56] Chen, Heng / Uddin, Lucina Q / Guo, Xiaonan / Wang, Jia / Wang, Runshi / Wang, Xiaomin / Duan, Xujun / Chen, Huafu(2019): *Parsing brain structural heterogeneity in males with autism spectrum disorder reveals distinct clinical subtypes*, 2: 628–637.

[57] Chen, Heng / Uddin, Lucina Q / Guo, Xiaonan / Wang, Jia / Wang, Runshi / Wang, Xiaomin / Duan, Xujun / Chen, Huafu(2019): *Parsing brain structural heterogeneity in males with autism spectrum disorder reveals distinct clinical subtypes*, 2: 628–637.

[58] Saaybi, Stephanie u.a.(2019): *Pre- and Post-therapy Assessment of Clinical Outcomes and White Matter Integrity in Autism Spectrum Disorder: Pilot Study*877.

[59] Zheng, Shuting / Hume, Kara A / Able, Harriet / Bishop, Somer L / Boyd, Brian A(2020): *Exploring Developmental and Behavioral Heterogeneity among Preschoolers with ASD: A Cluster Analysis on Principal Components*, 5: 796–809.

[60] Zabihi, Mariam u.a.(2020): *Fractionating autism based on neuroanatomical normative modeling*, 1: 384.

[61] Amunts, Katrin / Mohlberg, Hartmut / Bludau, Sebastian / Zilles, Karl(2020): *Julich-Brain: A 3D probabilistic atlas of the human brain's cytoarchitecture*, 6506: 988–992.

[62] Bedford, Saashi A u.a.(2020): *Large-scale analyses of the relationship between sex, age and intelligence quotient heterogeneity and cortical morphometry in autism spectrum disorder*, 3: 614–628.

[63] Moreau, Clara A u.a.(2020): *Mutations associated with neuropsychiatric conditions delineate functional brain connectivity dimensions contributing to autism and schizophrenia*, 1: 5272.

[64] Richardson, Hilary / Gweon, Hyowon / Dodell Feder, David / Malloy, Caitlin / Pelton, Hannah / Keil, Boris / Kanwisher, Nancy / Saxe, Rebecca(2020): *Response patterns in the developing social brain are organized by social and emotion features and disrupted in children diagnosed with autism spectrum disorder*12–29.

[65] Hong, Seok Jun / Vogelstein, Joshua T / Gozzi, Alessandro / Bernhardt, Boris C / Yeo, B T Thomas / Milham, Michael P / Di Martino, Adriana(2020): *Towards neurosubtypes in autism*.

[66] Moreno De Luca, Daniel / Martin, Christa Lese(2021): *All for one and one for all: heterogeneity of genetic etiologies in neurodevelopmental psychiatric disorders*71–78.

[67] Modenato, Claudia u.a.(2021): *Effects of eight neuropsychiatric copy number variants on human brain structure*, 1: 399.

[68] Ayub, Rafi / Sun, Kevin L / Flores, Ryan E / Lam, Vicky T / Jo, Booil / Saggar, Manish / Fung, Lawrence K(2021): *Thalamocortical connectivity is associated with autism symptoms in high-functioning adults with autism and typically developing adults*, 1: 93.

[69] (2004): *Chapter 36 - Morphometry*. In: Frackowiak, Richard S J / Friston, Karl J / Frith, Christopher D / Dolan, Raymond J / Price, Cathy J / Zeki, Semir / Ashburner, John T / Penny, William D (Hg.), Human Brain Function (Second Edition). BurlingtonAcademic Press: 707–722.

[70] Association, American Psychiatric / others u.a. (2013): *Diagnostic and statistical manual of mental disorders (DSM-5®)*. , American Psychiatric Pub.

[71] Insel, Thomas / Cuthbert, Bruce / Garvey, Marjorie / Heinssen, Robert / Pine, Daniel S / Quinn, Kevin / Sanislow, Charles / Wang, Philip (2010): *Research domain criteria (RDoC): toward a new classification framework for research on mental disorders*
.

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

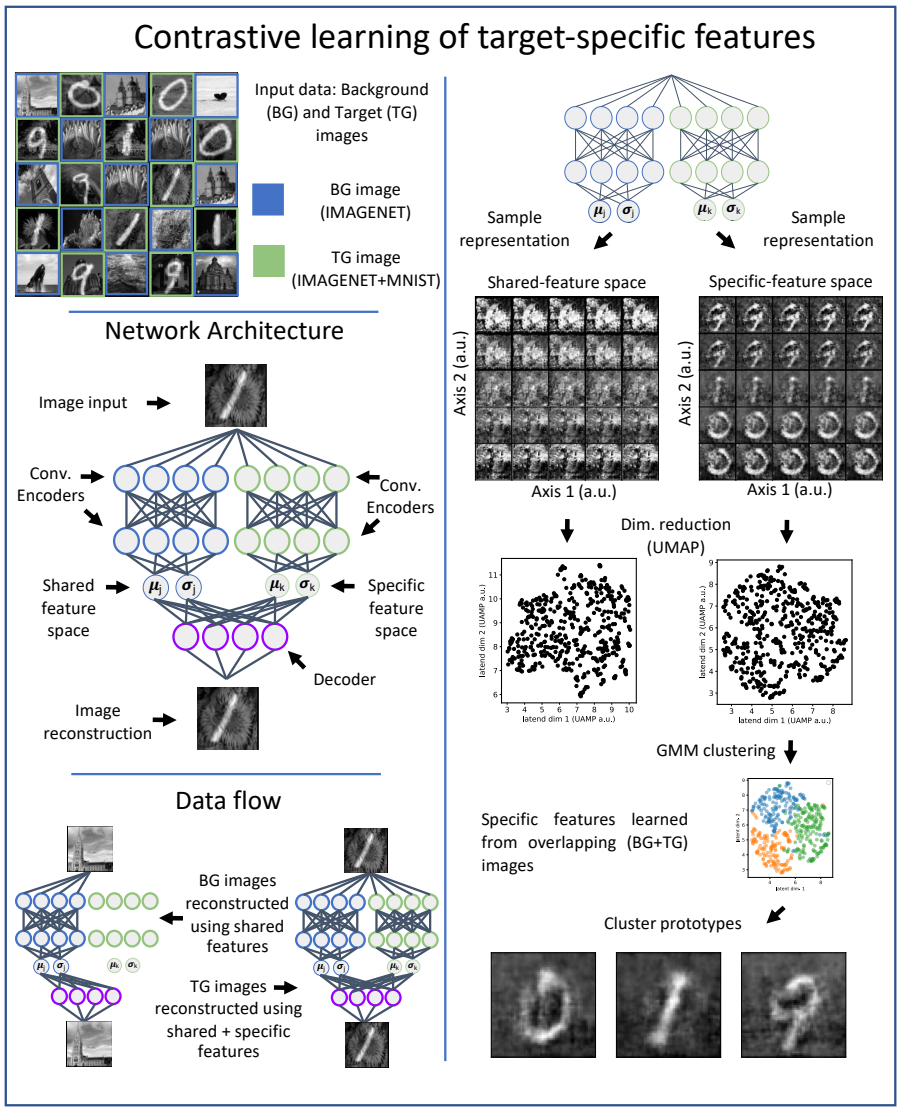

Figure S1: Demonstration of a contrastive variational autoencoder (CVAE) using synthetic data. Synthetic data consisted of background images (selected from ImageNet [15]), and target images (created by superimposing MNIST [3] digits on imagenet images). The Contrastive Variational Autoencoder (CVAE) 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.

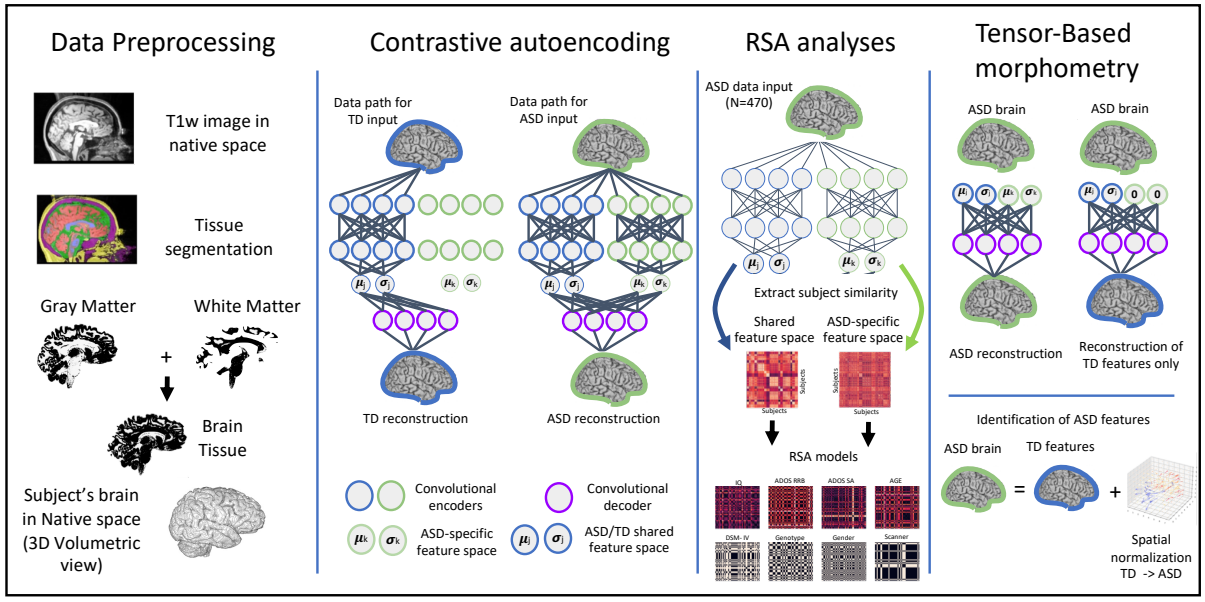

Figure S2: Data analysis pipeline. *Data Preprocessing*: Minimal preprocessing was employed before entering data into the CVAE. Anatomical T1w scans of TD (N = 470) and ASD (N = 512) subjects were segmented into tissue classes, and gray matter and white matter masks were used to reconstruct brain tissue (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.

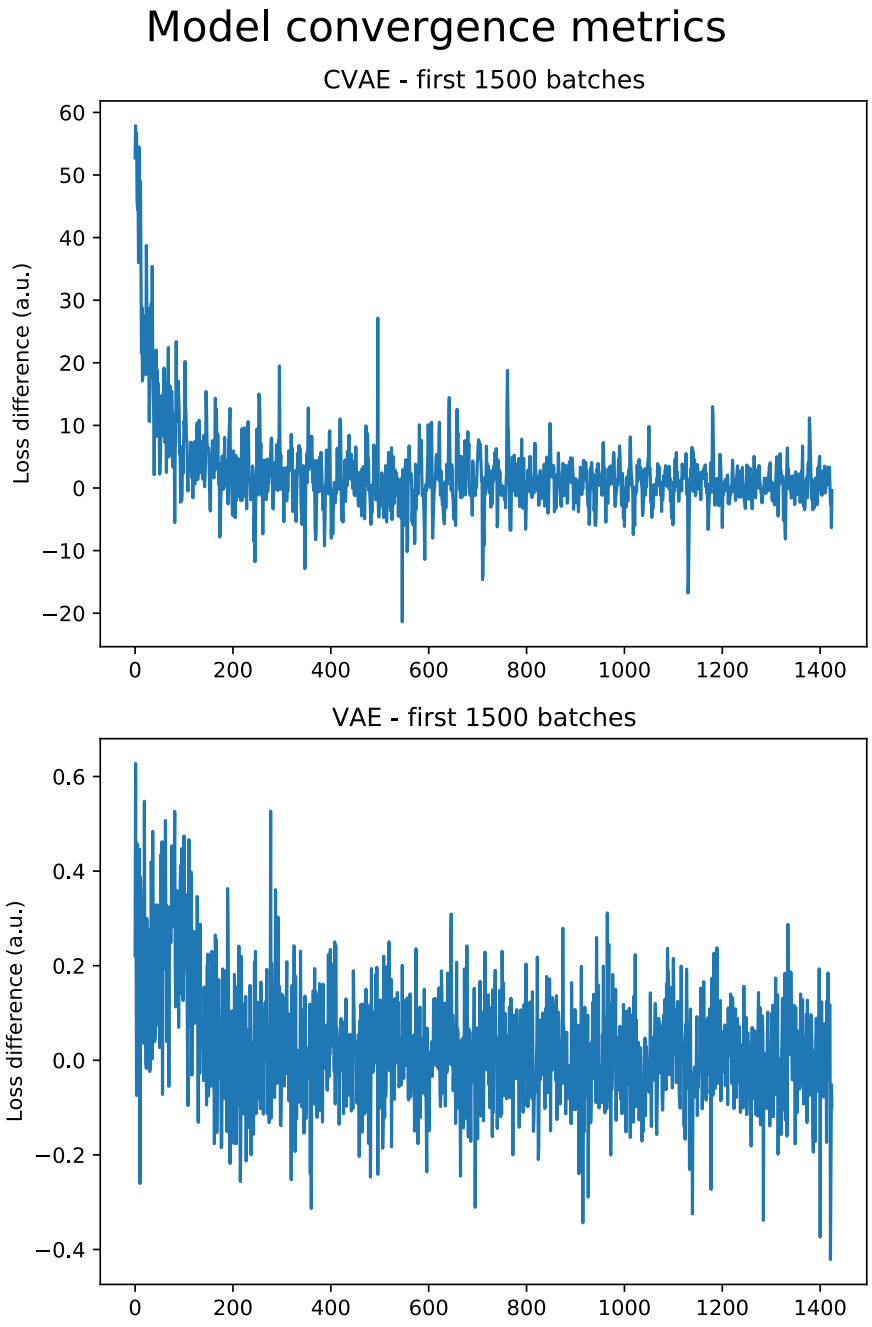

Figure S3: Training loss for the contrastive variational autoencoder (CVAE) and control variational autoencoder (VAE). The figure shows training loss difference for the first 1500 batches. Training was stopped when mean-squared-error between input batch and reconstruction fell below .005.

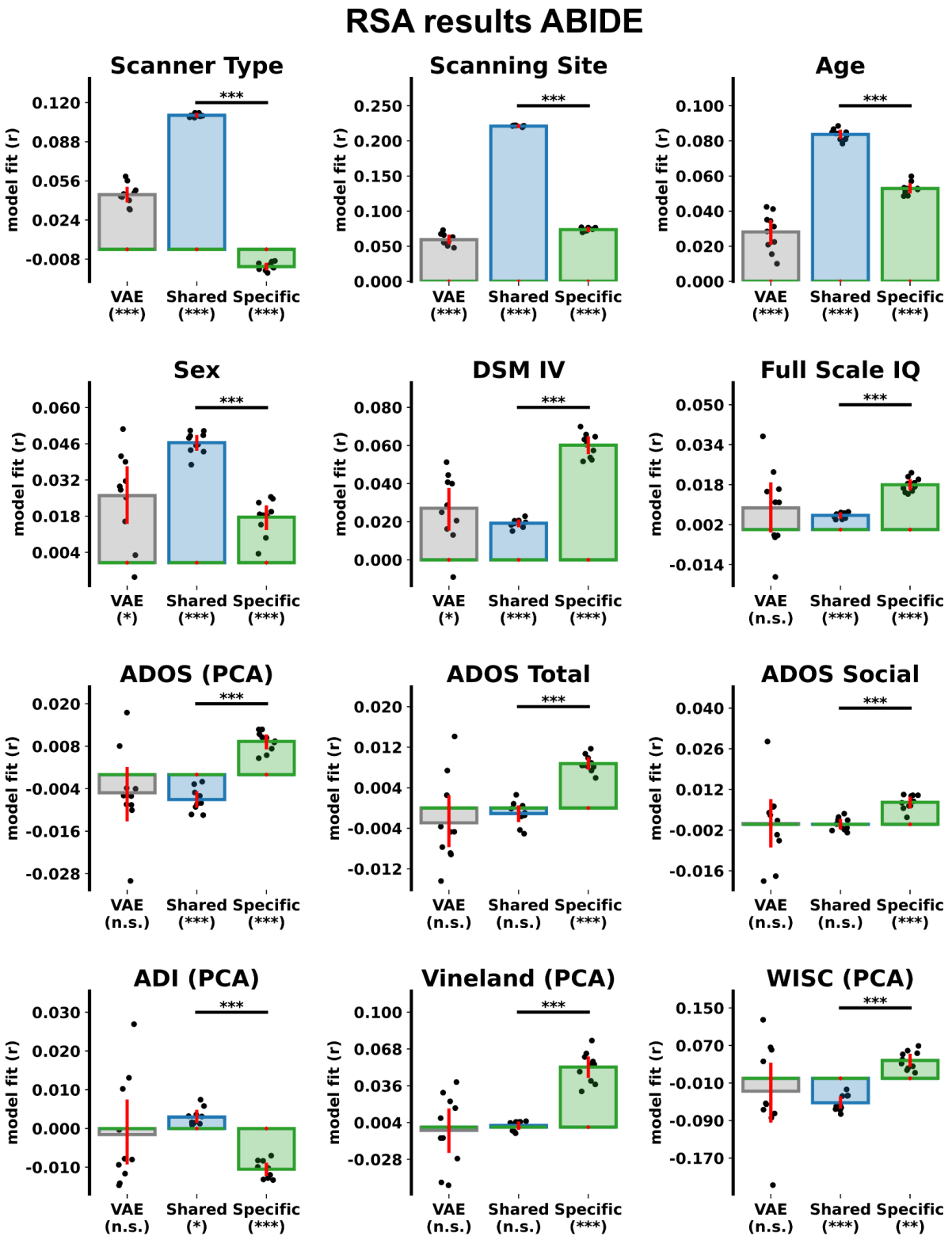

Figure S4: Representational Similarity Analysis (RSA) results (ABIDE dataset). Bar plots show representational similarity between neuroanatomical features (VAE, CVAE shared and CVAE ASD-specific) and patient properties. Significance indicated by stars (* $p < .05$, *** $p < .0001$,)

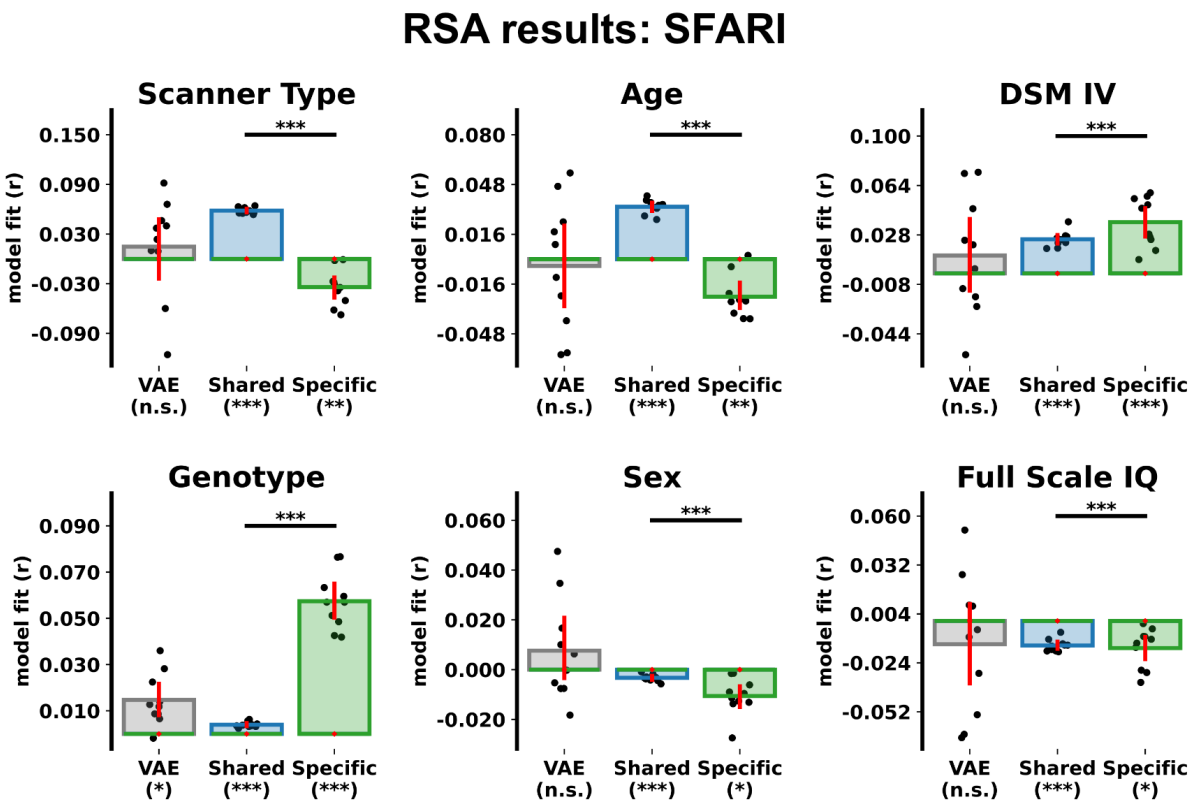

Figure S5: Representational Similarity Analysis (RSA) results (independent SFARI data). Bar plots show representational similarity between neuroanatomical features (VAE, CVAE shared and CVAE ASD-specific) and patient properties. (* $p < .05$, ** $p < .001$, *** $p < .0001$,).

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