Title: Structural changes in autism reflect atypical brain network organization and phenotypical heterogeneity: a hybrid deep network approach

Short title: Systematic study of ASD heterogeneity using deep learning

Siddharth Lokray[1], Basilis Zikopoulos[2, 3, 4, 5*], Arash Yazdanbakhsh[1, 2, 3*]

[1]Computational Neuroscience and Vision Laboratory, Department of Psychological and Brain Sciences, Boston University, Boston, MA, 02215 United States
[2]Graduate Program for Neuroscience, Boston University, Boston, MA, United States
[3]Center for Systems Neuroscience, Boston University, Boston, MA, United States
[4]Human Systems Neuroscience Laboratory, Department of Health Sciences, Program in Human Physiology, Boston University, Boston, MA, 02215 United States
[5]Department of Anatomy & Neurobiology, Boston University School of Medicine, Boston, MA, United States

*Corresponding authors

Corresponding authors email addresses

Arash Yazdanbakhsh, (64 Cummington Mall, Department of Psychological & Brain Sciences, Boston University, MA, 02215), Email: yazdan@bu.edu
Basilis Zikopoulos, (635 Commonwealth Avenue, Department of Health Sciences Boston University, Boston MA 02215), Email: zikopoul@bu.edu

Conflict of interest statement
The authors declare no competing interests.

Keywords
Autism Spectrum Disorder, Convolutional Neural Networks, Functional Brain Networks, Deep Neural Networks, Saliency maps, Social Responsiveness Scale, structural Magnetic Resonance Imaging

Acknowledgments
This work was supported by a grant from the Simons Foundation (SFARI Grant 946867) to BZ and AY.

Author contributions
Conceptualization: BZ and AY; Experimental procedures: SL; Analysis: SL, BZ, AY; Funding acquisition: BZ, AY; Writing – original draft: SD, AY; Writing – review & editing: BZ, AY.

Data and code availability
Data supporting the findings of this study are either included in the manuscript and supplementary materials.

Abstract
Autism Spectrum Disorder (ASD) is a developmental disorder characterized by heterogeneity in social and emotional responses, language, and behavior. Assessments such as the Social Responsiveness Scale (SRS) can quantify this variability but understanding underlying mechanisms and identifying distinct and shared atypical organization and function of brain networks remains a challenge. Convolutional Neural Networks (CNNs) have been used to analyze imaging data. However, the relationship between structural brain changes observed in structural MRI (sMRI), the affected brain functional networks inferred from these structural changes, and their connection to ASD phenotypes and scores still requires systematic investigation. In this study, we ensembled 3D CNNs with other artificial intelligence (AI) methods to conduct a comprehensive analysis of macrostructural changes in ASD. We found consistently dominant involvement of (a) the left hemisphere, (b) the frontal and temporal lobe, and (c) the default mode, salience, and language networks in ASD. Our findings highlighted brain network similarities and differences between high and low severity ASD and showed that typically developed brains fall at the low-severity end of the high-to-low severity spectrum of ASD. Our systematic AI approach utilized the phenotypic heterogeneity and spectral nature of ASD to uncover significant structural changes across brain regions and functional networks, correlating the structural, functional, and phenotypical heterogeneity of individuals with ASD. This enabled us to identify known and novel global and local brain region and network changes in ASD in relation to phenotypes and clinical scores that can guide diagnostic subtyping.

Teaser (one sentence summary): We used machine learning approaches to characterize the phenotypic and imaging-derived structural heterogeneity of ASD with the aim to identify the brain regions and networks linked to its cognitive and behavioral phenotypes.

Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder that affects social interactions, emotional responses, language, and speech development, leading to behavioral inflexibility in affected individuals (Amaral, Schumann, and Nordahl 2008; Ecker 2012; Lord et al. 2000). Types, onset, and severity of symptoms can vary significantly; thus, ASD affects individuals differently leading to a display of a variety of phenotypical combinations. This heterogeneity makes it challenging to understand the underlying mechanisms of disruption, pinpoint cortical and subcortical regions that comprise atypical brain networks in ASD, and ultimately accurately diagnose and treat affected individuals. Heterogeneity in ASD can be partially expressed as a function of the cognitive, behavioral, and verbal scores in the Autism Diagnostic Interview (ADI) (Ecker 2012; Lord, Rutter, and Le Couteur 1994). A key question is whether ASD heterogeneity indices, based on ADI, can be reliably and consistently reflected in and correlated with the structural and functional changes in brain networks that can be assessed *in vivo* by imaging techniques, such as Magnetic Resonance Imaging (MRI) (Amaral, Andrews, and Nordahl 2024; Duan et al. 2025) .

Modern AI techniques, including 2- and 3-Dimensional versions of Convolutional Neural Networks (CNNs) and their derivatives, deep attentional, contrastive, and CNN and Long short-term memory (LSTM) hybrids have been used to analyze imaging data (Aghdam, Sharifi, and Pedram 2019; Aglinskas, Hartshorne, and Anzellotti 2022; Ecker et al. 2010; Khosla et al. 2018; Li et al. 2018; Niu et al. 2020; Supekar et al. 2013; Uddin et al. 2013; Zhao et al. 2018a). Most of these studies focused on accurately classifying ASD against typically developed (TD) individuals, using structural imaging datasets, or identified disruptions in brain network activity predominantly using functional imaging (fMRI). For example, Aglinskas et al. (Aglinskas et al. 2022) developed a neural network that identified ASD- and control-specific neuroanatomical variability and correlated ASD anatomical features with symptoms. Using resting state fMRI data, Zhao et al. (Zhao et al. 2018b) developed a neural network that extracted salient brain regions (Zhao et al. 2018b) that were then mapped to outline brain networks that are disrupted in ASD. However, two key relationships remain poorly explored and understood: a) the presence of structural changes and their correlation across brain areas, as a structural pointer to the functionally affected brain networks, and b) correlations between autism diagnostic scores and brain pathology (Uddin et al. 2013).

To address these gaps, we investigated the relationship between structural changes across regions and ASD ADI scores and their association with brain networks to reveal common and divergent structural and functional changes in individuals with high- and low-severity ASD. Our overarching hypothesis is that sMRI, in fact, contains information that reflects functional network and diagnostic score changes in ASD, which can be deciphered, using state-of-the-art AI, to identify salient brain regions and their networks underlying ASD heterogeneity. Our findings suggest that there is indeed partial required information available in macroscopic sMRI that can be used to reliably identify brain regions, connectivity, and networks involved across the heterogeneous spectrum of ASD, and provides a framework for the identification of systematic, correlated neuroanatomical and functional dimensions across ASD.

Results

In this study, we examined sMRI datasets (ABIDE II, Table S1) from individuals with different ages and different ASD severity scores to capture ASD heterogeneity. Our approach involved development and optimization of CNNs that can initially evaluate imaging data, through individual brain sections (2D). This approach then evolved into direct evaluation of entire brain volumes by custom-building a 3D CNN for feature extraction from the 3D MRI volumes. These extracted features were subsequently clustered by k-Nearest Neighbors algorithm (kNN), resulting in a 3D-CNN-kNN ensemble that accurately distinguished low and high severity ASD individuals based on their structural MRIs. We then extracted brain regions salient for ASD heterogeneity utilizing our hybrid 3D-CNN-kNN ensemble and projected the extracted saliency measures on cortical parcellations using the Destrieux atlas (Destrieux et al. 2010). This enabled a more detailed

identification of brain regions with macro-structural changes as highlighted by AI saliency maps, along with the analysis of correlations between these changes, and the severity of ASD. We then identified hierarchical relationships between brain regions, based on our correlation analysis, to associate correlated regions with functional brain networks. We focused, in particular on networks that are affected in ASD, including the default mode network (DMN), salience network (SN), social network, and central executive network (CEN or frontoparietal network), which play key roles in self-awareness, attention, social interactions, and cognitive control**,** which are all impacted in ASD. We also included the Language, Sensorimotor, Dorsal/Ventral Attention, and Limbic/Emotion networks—known to be affected in ASD—in the same analysis.

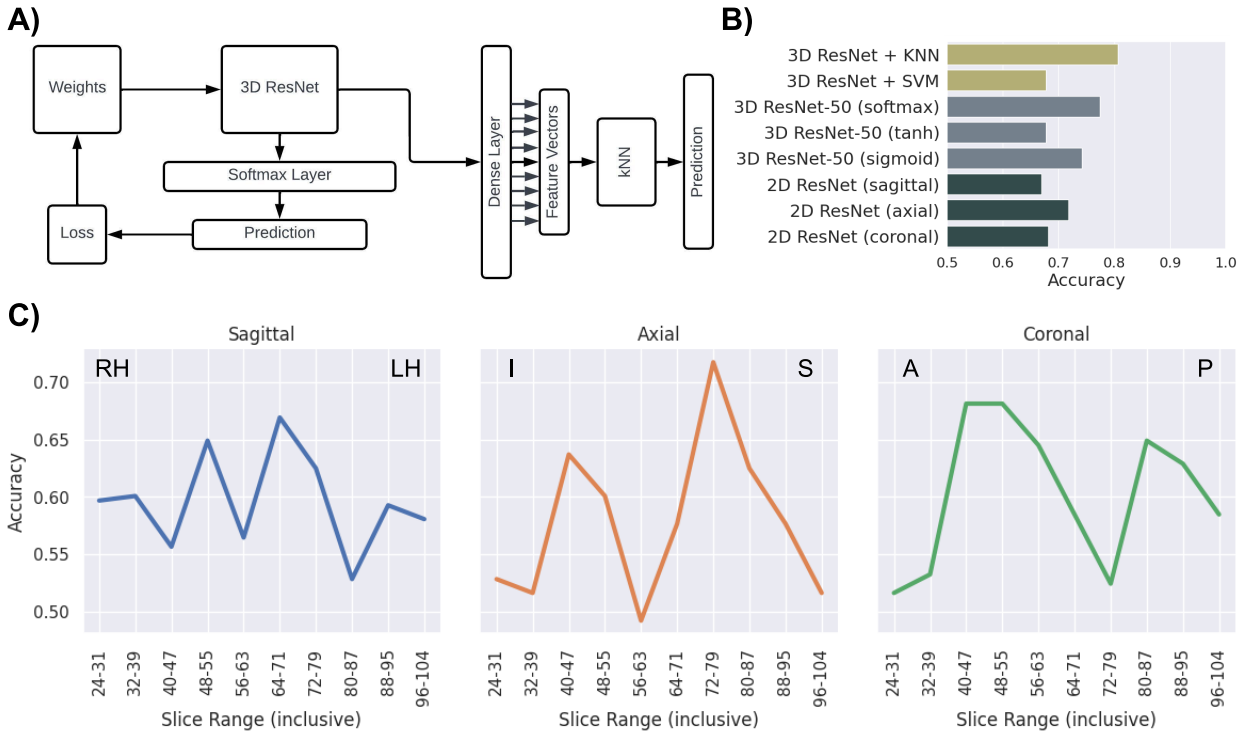

Figure 1: Machine learning network architecture and performance comparison after training. A) 3D ResNet-50 + kNN hybrid architecture. The 3D ResNet-50 was trained initially on the preprocessed structural MRI data, then was passed through a kNN to make the category predictions. B) Performance of 3D ResNet-50 (softmax, tanh, and sigmoid activation), hybrid 3D ResNet-50 (kNN and SVM), and 2D ResNet-50. The performance of the 2D ResNet-50 is representative of the highest performance on the slice ranges. C) Performance of 2D ResNet-50 on the sagittal (slice 24 is the furthest slice on the right hemisphere and slice 103 is furthest slice on the left hemisphere), axial (slice 24 is the most inferior slice and slice 103 is the most superior slice), and coronal (slice 24 is the most anterior slice and slice 103 is the most posterior slice) axes for each slice range. Table S2 provides the accuracy values for each slice range on the sagittal, axial, and coronal axis.

We trained a 2D ResNet-50, our custom developed 3D ResNet-50, and hybrids with kNN and Support Vector Machine (SVM; Fig. 1A), using T1-weighted sMRI scans from individuals with low and high severity ASD and compared their performance (Fig. 1B). The 2D ResNet-50 was

trained separately on the sagittal, axial, and coronal axes, using MRI slices as input, whereas the 3D ResNet-50 and its hybrid variations were comprised of 3D convolution filters and pooling layers, thus, used MRI *volumes* as their inputs rather than MRI slices. After extensive hyperparameter searches, we noted the performance of the 3D network was superior to the 2D networks; hence, extracted saliency maps from the 3D ResNet-50 and overlaid the salient areas in cortical and subcortical regions a) on the FreeSurfer pial surface, and b) in subcortical volumes to parse brain regions with saliency. We ran correlation analysis between the Social Responsiveness Scale (SRS) Total and SRS subscores and salient regions as well as pairwise cross-region saliency correlation analysis to identify the regions and networks correlated with ASD severity at a macrostructural level.

Performance of the 2D ResNet-50 on MRI slice intervals. The performance of the 2D ResNet-50 on each group of slice ranges is shown in Fig. 1B-C. With regard to the performance of the 2D ResNet-50 on the groups of sagittal slices (Fig. S1, Table S2), there were two distinct peaks; one slightly left of medial region of the right hemisphere (slices 64-71) around the thalamus and corpus callosum, and one in the middle of the right hemisphere around the hippocampus, putamen, and amygdala (slices 48-55). On the axial slices (Fig. S2, Table S2), there were two peaks; one around the inferior region (slices 40-47) intersecting the posterior lobe of the cerebellum and a more pronounced peak around the superior region (slices 72-79) intersecting the lateral ventricles and the body of the corpus callosum. The coronal slice groups (Fig. S3, Table S2) showed a solid peak around the anterior region (slices 40-55) crossing the middle frontal gyri, inferior frontal gyri, and the head of the caudate nucleus as well as another peak around the cerebellar region (slices 80-87) crossing the hippocampus, thalamus, and fourth ventricle. The performance of the 2D ResNet-50 on groups of MRI slices provided a basic understanding on what regions of the brain were more useful to differentiate high severity subjects from low severity subjects.

Performance of the 3D ResNet-50 and hybrid variations on MRI volumes. The architecture for the 3D ResNet-50 is shown in Fig. 1A. Three activation functions were tested to push the 3D ResNet-50 to perform to its full potential: softmax, tanh, and sigmoid. Softmax, sigmoid, and tanh performed with an accuracy of 77.41%, 74.19%, and 67.74, respectively. Based on these results, we selected the softmax function as the classification layer activation for the retrieval of saliency maps for each subject. Additionally, the two hybrid neural networks tested were the 3D ResNet-50 + kNN and the 3D ResNet-50 + SVM. The difference between the 3D ResNet-50 and the hybrid 3D ResNet-50 is that the classification layer of the neural network is replaced with a fully connected layer with *n* nodes to output a feature vector into the respective machine learning algorithm (kNN or SVM) of the hybrid network (Fig. 1A). The performance with the kNN is significantly better than with the SVM, as the kNN yielded an accuracy of 80.64% (3.23% better than 3D ResNet-50 only) while the SVM yielded an accuracy of 67.74%. The fully connected layer that outputs the feature vectors was initialized with randomized weights thus, many attempts

were conducted, and the best accuracy was reported. We used the 3D ResNet-50 for the following saliency, cross-region and region-score correlation analyses.

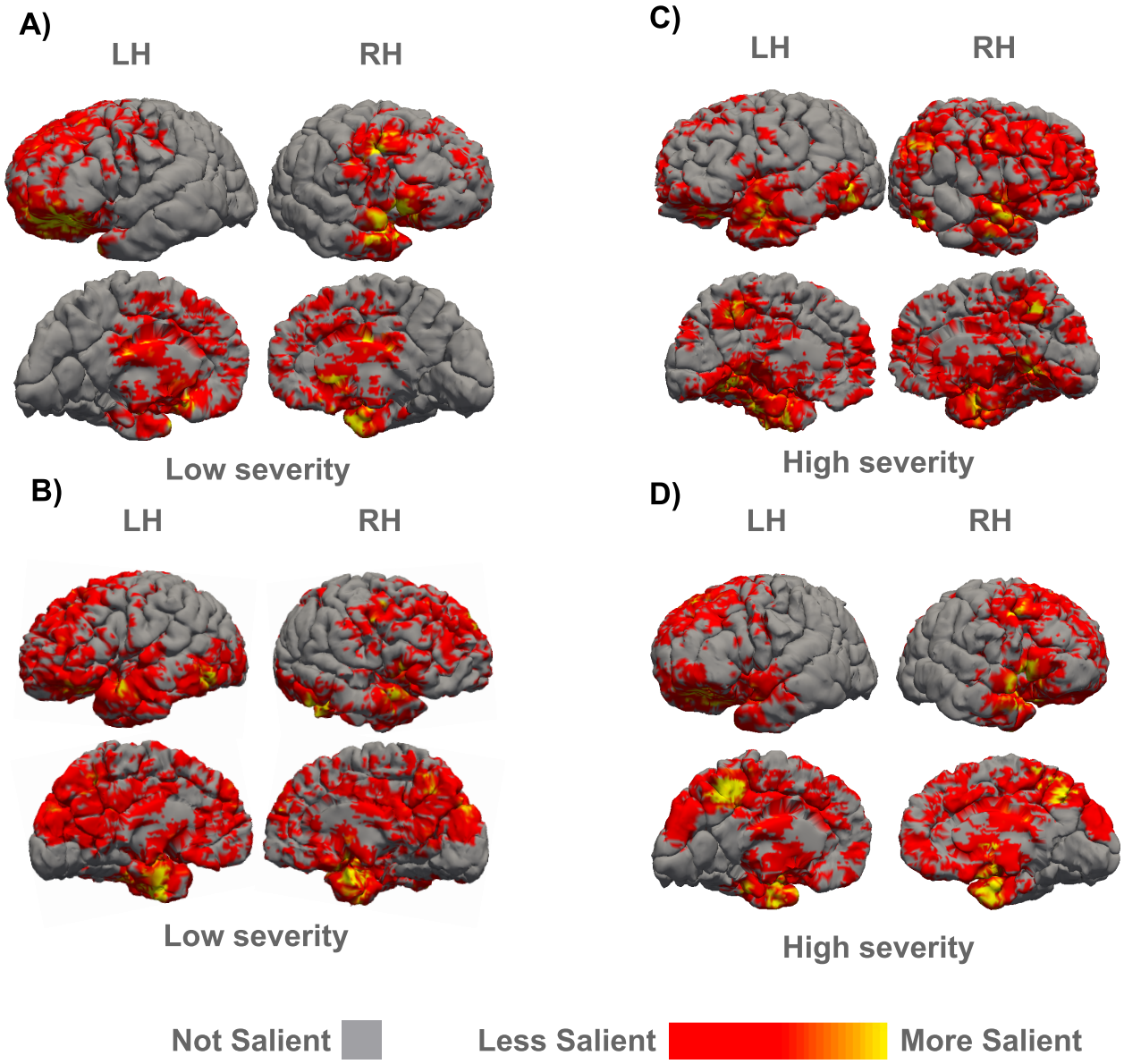

Figure 2: Saliency maps for network classification. Each case saliency map (averaged on sagittal plane) generated from the 3D ResNet-50 was overlaid on the case pial surfaces (gray), which was generated through Freesurfer's recon-all pipeline. Lateral (top row) and medial (bottom row) views of subject's left hemisphere (LH) and right hemisphere (RH) were visualized by FreeView, Freesurfer's visualization software. Yellow marked areas highlight more salient regions whereas red marked areas display moderately salient regions. A) Low severity case with SRS-Total Score of 44. B) Low severity case with SRS-Total Score of 42. C) High severity case with SRS-Total Score of 83. D) High severity case with SRS-Total Score of 87.

Saliency in cortical and subcortical regions. Sixteen saliency maps were obtained by TensorFlow gradient tape within the 3D ResNet-50 and were overlaid on their respective pial

surface (Fig. 2). Salient regions were observed in the dorsolateral prefrontal cortex (BA areas 8, 9, and 46) as well as the orbitofrontal cortex (BA 11, and 13). There are also significant salient regions in the temporal lobe (BA 28, 35, and 36) including the amygdala. Both Broca's Area (BA 44, and 45) and Wernicke's Area (BA 22) show regions with moderate saliency. The dorsal cingulate (BA 24, and 32), subgenual cingulate (BA 25), and corpus callosum had distinguishable, moderately salient regions, that were well delineated. Detailed saliency maps overlaid over the MRI scans are shown in Figures S1-3.

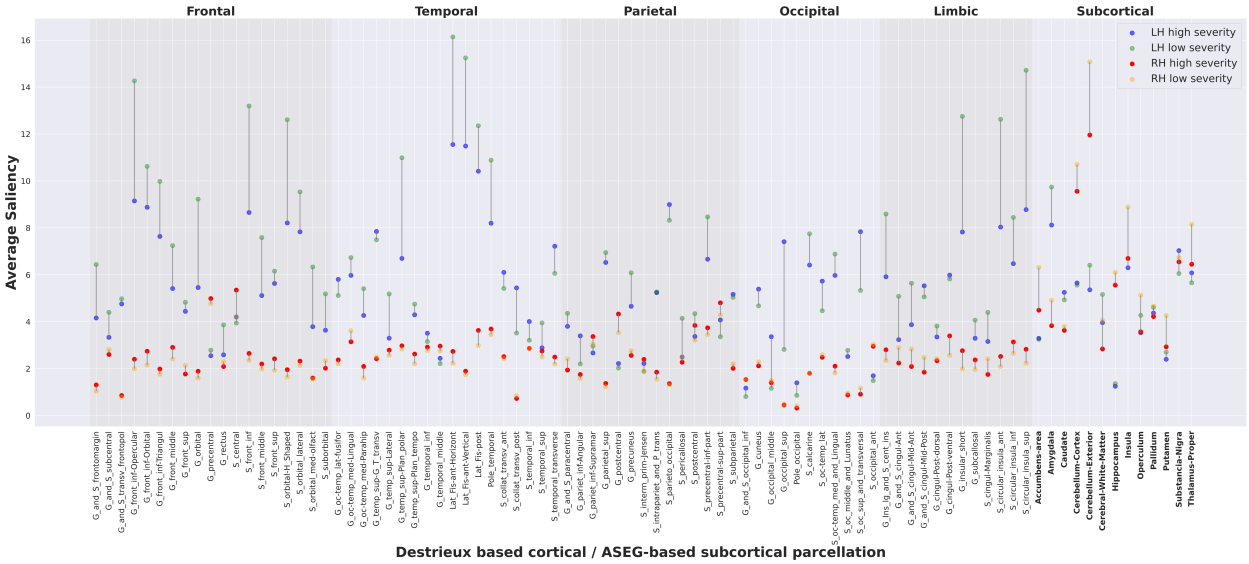

Figure 3: Differences in average saliency between high severity and low severity subjects in the left and right hemispheres for each cortical region labeled based on parcellation by Destrieux atlas and by FreeSurfer's automatic subcortical segmentation of a brain volume (aseg atlas, bolded regions). Average saliency in high severity subjects in the left hemisphere (LH, blue) are connected to the average saliency in low severity subjects in the left hemisphere (green) and average saliency in high severity in the right hemisphere (RH, red) is connected to the low average saliency in lower severity subjects in the right hemisphere (yellow). In summary, cold colors represent the LH and the hot colors represent the RH and high/low saturation colors represent high/low severity. Note an overall larger average and broader range of saliency in the left hemisphere. The high and low severity values of the brain stem and fornix are 4.72 and 4.79, and 3.34 and 3.55 respectively. These regions are not included in this figure as they do not have a hemisphere attribute.

We used the Destrieux atlas (Destrieux et al. 2010) to create masks of cortical and subcortical regions for 150 subjects (Destrieux et al. 2010) (75 cortical regions per hemisphere) and Freesurfer's automatic subcortical segmentation (ASEG, 13 subcortical regions per hemisphere as well as brain stem and fornix that were not separated across hemispheres). We calculated the average saliency value per region across all subjects (Fig. 3). Overall, the left hemisphere generally had a) higher saliency value, with an average saliency of 5.85, while the average saliency of the right hemisphere was 2.31 (p = 5.19e-17 , df=148), and b) broader saliency range that was estimated by the average saliency range of each region in the left and right hemispheres across high and low severity subjects (p = 7.53e-12, df=148). In the left hemisphere, the opercular, orbital, and triangular regions of the inferior frontal gyrus, the horizontal, vertical, and posterior ramus of the lateral sulcus, the anterior and superior segments of the circular sulcus of the insula, and the

short insular gyri, temporal pole, and orbital sulci were among the regions with high/moderate saliency. High saliency regions in the right hemisphere included the cerebellum, basal ganglia, thalamus, insula, postcentral and precentral gyrus, the central sulcus, and the temporal pole.

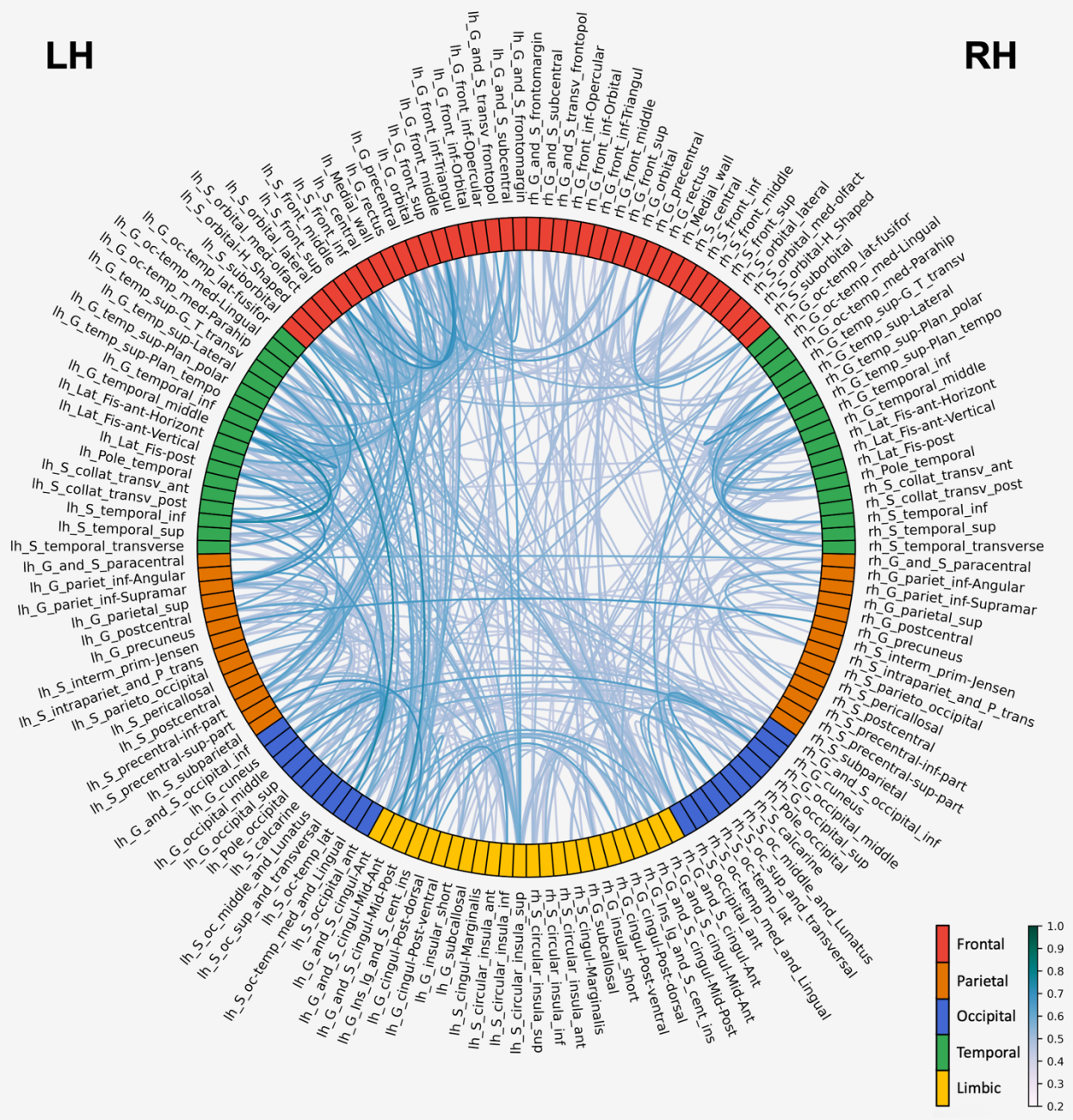

Figure 4: Circular connectivity plot highlighting significant pairwise saliency correlations (critical α value = 0.01) between cortical brain regions across high and low severity MRI cases. Darker lines represent higher Kendall τ whereas lighter lines represent lower Kendall τ values. The color of each brain region corresponds to the lobe the region is located in. The strongest correlations were mostly between regions within the same lobe however, cingulate regions showed several strong cross-lobe correlations.

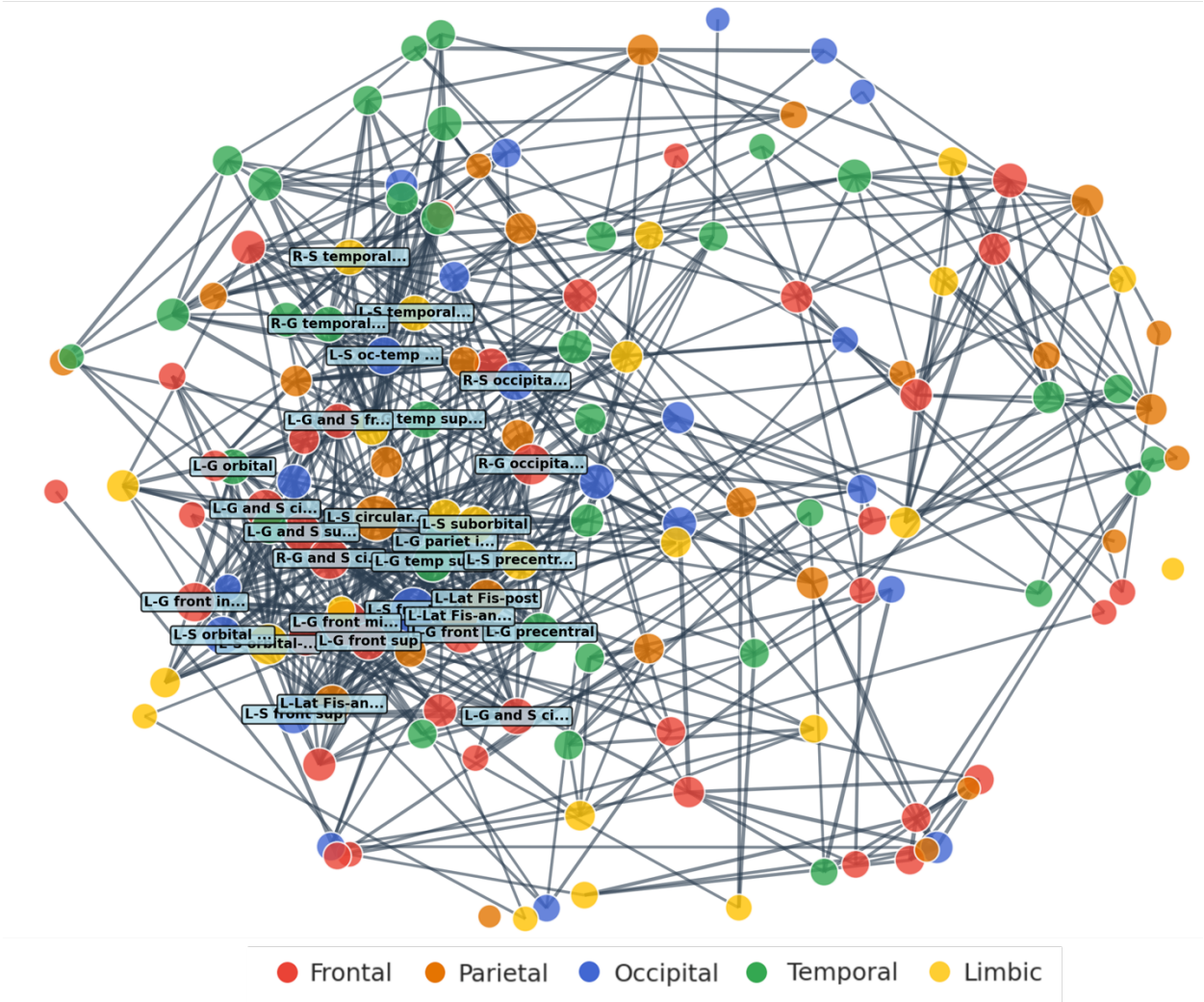

Figure 5: Spring Cluster layout graph for corticocortical saliency correlations. Connectivity brain regional analysis dependent on region-region Tau correlations depicted in Figure 4. Each node represents a region in the Destruiex atlas, highlighted with their corresponding lobe color. Size of each node is dependent on how many significant correlations ($\alpha < 0.01$ and where $\tau > 0.4$) the respective region takes part in. The distance between correlated regions is weighted by $1-\tau$, where highly correlated regions are closer together. The algorithm arranges the nodes to minimize the weights between the regions. This leads to clustering of highly correlated regions (top 30 regions are named) towards the center and less correlated regions to the periphery.

Correlation between corticocortical saliency and corticosubcortical saliency. Kendall τ rank-based correlation between cortical regions in the Destrieux atlas were conducted to observe the relationship in paired saliencies within the brain. The top 500 significant correlations sorted by the τ value are shown in Fig. 4. Each connecting thread indicates the presence of significant ($p < 0.01$) correlation with $\tau > 0.4$ as a midlevel threshold. To maximize and optimize inference from these complex connectivity maps, based on regional saliency correlations, we took the following steps. First, we used a stepwise adaptive threshold approach (from 0 to 0.6 with steps of 0.1) for

examination of cross-regional correlations τ. We found that there were only few cross-regional correlations beyond threshold 0.6. The reason for the extensive, stepwise examination of τ thresholds, is because there were salient regions that were significantly correlated with *many* other regions but with low τ, and also regions correlated with a *few* other regions but with a higher value of τ. To distinguish between the two groups, we refer to the first as the regions *correlated with many* and the second, as the *highly correlated* regions groups, respectively. Both groups reflect extreme and in between scenarios that offer insights about the affected brain networks in ASD. We generated Spring Cluster layout graphs (Fruchterman and Reingold 1991; Hamaguchi, Marumo, and Takeda 2025; NetworkX Developers 2025), which arranged cortical regions based on the number of significant cross-regional Kendall Tau correlations that each region maintained with other regions (Fig. 5). The algorithm created a network where each cortical region was a node, and significant correlations above the threshold became binary connections between regions. Regions were positioned such that those with a greater number of significant correlations with other regions were pulled toward the network center due to having more attraction forces acting upon them. This created a natural hierarchy, where regions serving as potential network hubs appeared in the center while regions with a smaller number of significant correlations appeared at the periphery. This spatial organization allowed identification of cortical regions that may function as structural pathology hubs in ASD, based on their tendency to show significant correlations with many other brain regions rather than the magnitude of individual correlation values. Fig. 5 shows the Spring Cluster layout with τ threshold of 0.4, approximately midway within 0.1-0.6. At this threshold, the following regions were placed more toward the center, because they had a larger number of significant correlations with other regions: left inferior frontal sulcus, left superior segment of the circular sulcus of the insula, left middle frontal gyrus, left supramarginal gyrus, right anterior part of the cingulate gyrus and sulcus (ACC), right middle occipital gyrus, left orbital sulci, left opercular part of the inferior frontal gyrus, left posterior ramus of the lateral sulcus, left temporal plane of the superior temporal gyrus. For the full range of thresholds (0.1-0.6) to cover both groups of "correlated with many" and "highly correlated regions", see supplementary table S3.

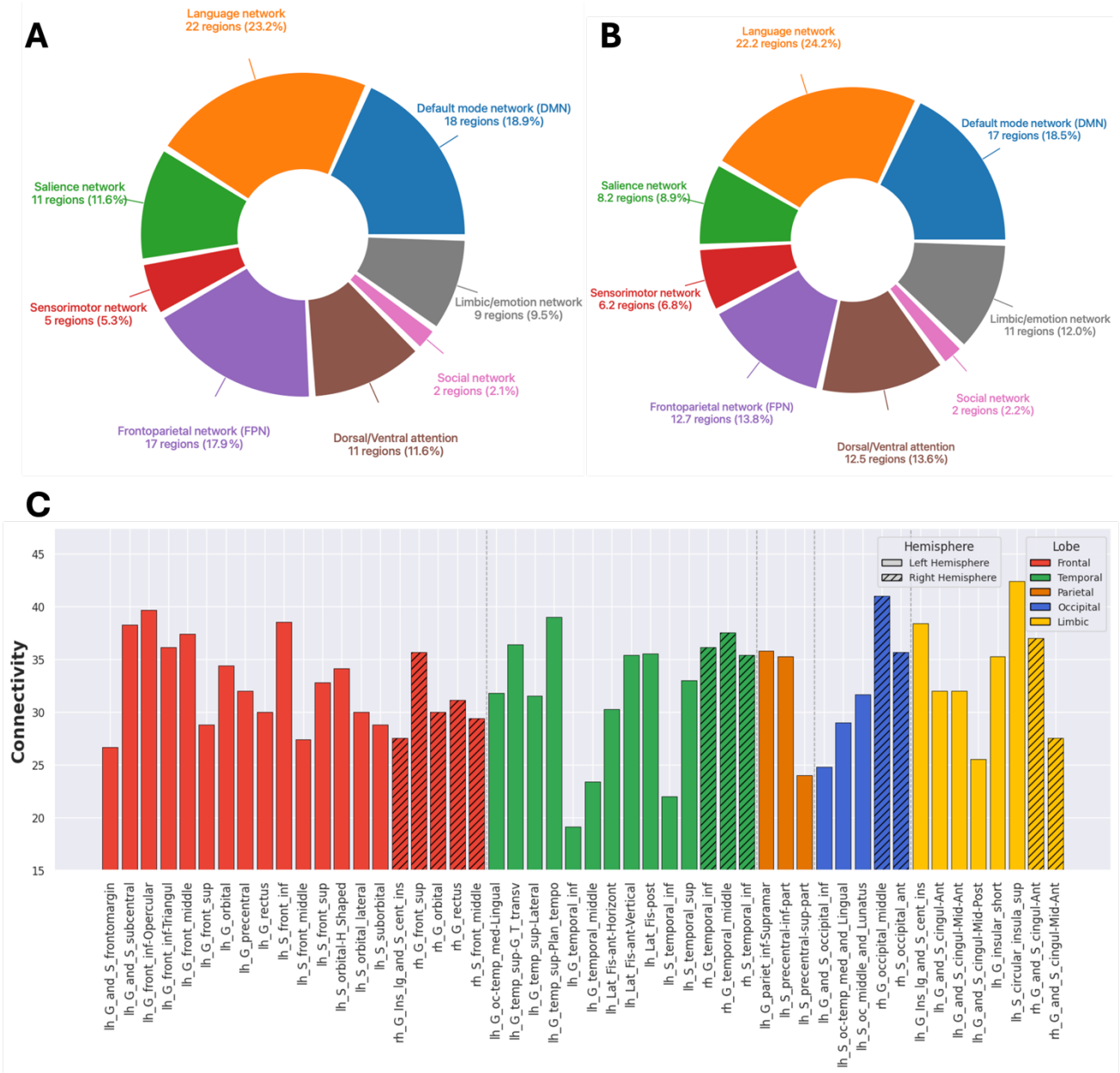

Figure 6: Degrees of involvement of functional brain networks and lobes in ASD. **A.** The degree of involvement for each brain network was derived from the pairwise brain region saliency correlation analysis shown in Fig. 4. Specifically, we considered significant correlations ($p < 0.01$) with τ exceeding a set threshold (e.g., 0.4) and selected the top n brain regions (e.g., 50) with the highest number of such correlations. From this list, we counted how many regions belonged to each functional network (raw counts shown beside the pie chart) and calculated the proportion of each network's count relative to the total, expressed as a percentage. **B.** The analysis in **A** was repeated and averaged across thresholds ranging from 0.1 to 0.6 in increments of 0.1. These measures indicate the global engagement of regions within functional networks with the rest of the brain in ASD. **C.** The number of connections/correlations per brain region was summed across the top 50 areas from each threshold-based table in S3, then divided by six (the number of τ thresholds between 0.1 and 0.6) to calculate the average number of correlations/connections per region across all thresholds. Areas were subsequently re-ranked based on these averaged values to identify the top 50 regions involved in ASD according to averaged structural correlations. The largest number of regions involved in ASD are found in the frontal and temporal lobes in the left hemisphere.

To better understand the organization and likely functional effects of salient regions that showed a high number of correlations with other regions or high correlation values in ASD, we counted the number of salient regions belonging to 4 main brain networks, consistently affected in ASD, i.e., Default Mode (DMN), Salience (SN), Frontoparietal (FPN, or central executive network), and Social networks, as well as 4 additional networks, i.e., Language, Sensorimotor, Dorsal/Ventral attention, and Limbic/Emotion networks, which have been associated with ASD dysfunction. For example, using $\tau = 0.4$, Fig 6A pie chart shows the relative involvement of the 8 networks, highlighting the relative representation of regions that are part of each of these 8 networks, which are well-established in ASD pathology. The raw counts, such as 22 for the Language network (LN) is the count of brain regions belonging to LN in S3 table with threshold $\tau = 0.4$, and the ratio 22/95 = 23.2% indicates the ratio of LN involvement in which 95 is the total brain region count across all 8 networks in the pie chart.

For the overall examination of brain network involvement across the full range of τ thresholds (0.1-0.6), we averaged the brain region counts and their ratios per network across the 6 pie charts, similar to the one in Fig 6A, for the 6 thresholds, 0.1-0.6 (Fig. 6B). Overall, the Language and Default Mode Networks consistently showed the highest average representation and association with ASD, among functional networks, followed by the Frontoparietal, Dorsal/Ventral attention, Limbic/emotion, and Salience, networks. The Sensorimotor and Social networks were minimally represented, reflecting their more distributed or overlapping nature in the anatomical brain regions.

Additionally, we summed the number of connections/correlations per brain region across the top 50 areas from each threshold-based table in S3, and divided by 6 (number of τ thresholds between 0.1-0.6) to reflect the average number of correlations/connections per brain region across tables covering all thresholds. We then, re-ranked the areas based on these averaged correlation/connection values to identify the top 50 regions involved in ASD based on averaged structural correlations (Fig. 6C). This analysis showed that pairwise saliency correlation across cases was dominant in the left hemisphere, both in terms of number of inferred structural connections per region (higher values in the average connection columns) and in terms of the number of regions as central hubs (more table rows for left than right). About 40% of the

correlations were between regions in the same lobe (per hemisphere), with the left frontal lobe exhibiting the highest number of interregional correlations (12.6%).

Fig. 7 visualizes the Tau correlations between the cortical and subcortical regions. The strongest and most numerous correlations of the subcortical regions were with the cortical regions in the left hemisphere [true for all lobes except for limbic, which had balanced left-right correlation with subcortical regions]. 77.27% of the subcortical to cortical correlations were connected to regions in the left hemisphere while 9.09% of the correlations were connected to regions in the right hemisphere, with the rest being within the subcortical regions.

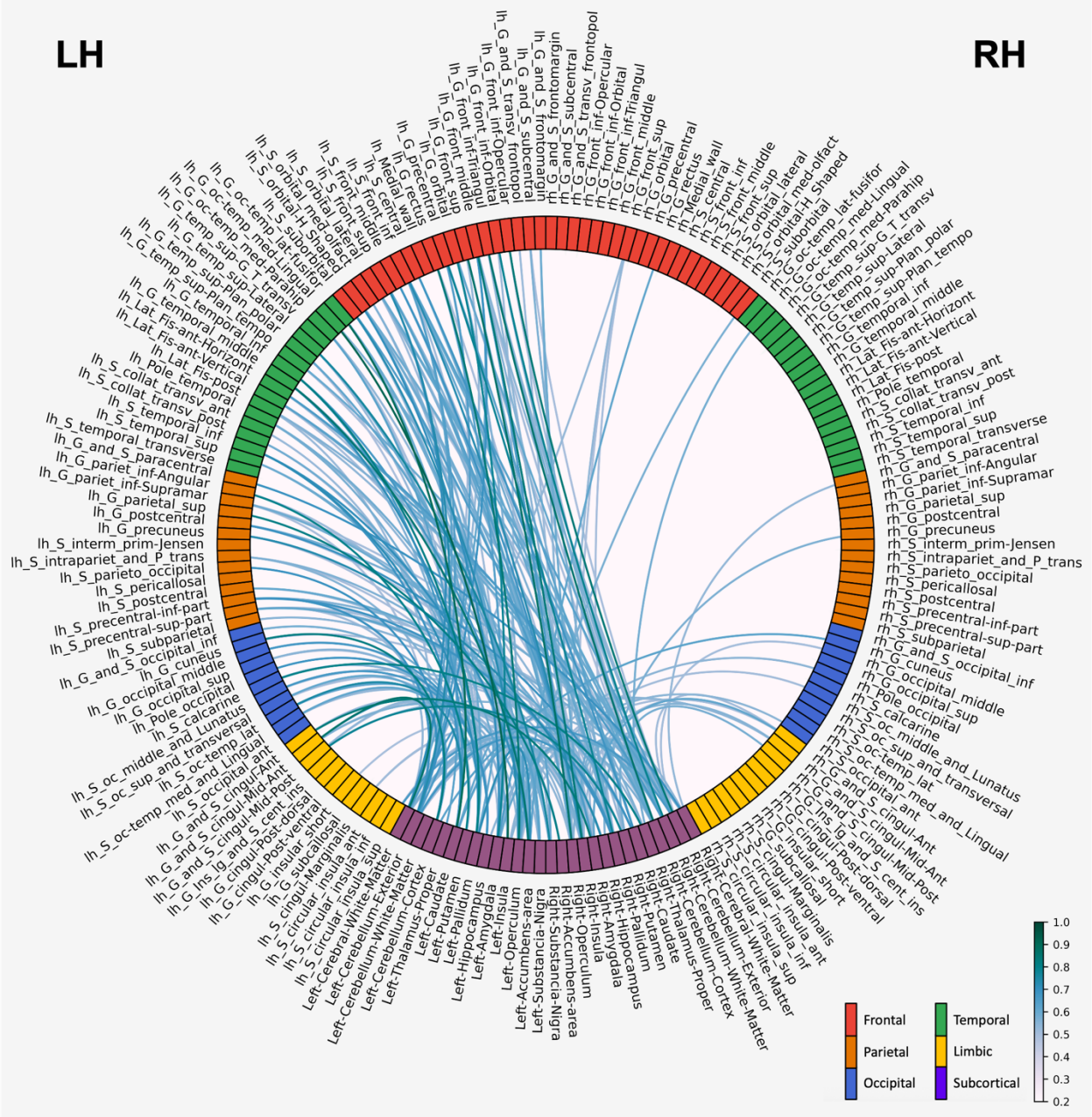

Figure 7: Significant pairwise saliency correlations (critical α value = 0.01) between cortical brain regions and subcortical brain regions across high and low severity MRI cases. This figure is organized like Fig. 4, with darker lines representing higher Kendall τ and lighter lines representing lower Kendall τ values. Color of each brain region corresponds to lobe the region is located in. Notably, the corticosubcortical correlations have higher τ values than the corticocortical correlations, shown in Fig. 4.

Correlations regarding regional saliency averages. We investigated correlations between average and standard deviation of saliency within either LH or RH regions from cases either with high or low severity, resulting in an 8×8 grid accounting for LH vs RH (2) × Average (Avg) vs

Standard deviation (STD) (2) × High vs Low severity (2) scatter plots and distribution histograms (Fig. 8). Overall, the patterns of distribution and scatter indicate more proportional area saliency average and standard deviation in LH in high severity cases, less proportionality in RH in high severity, and no proportionality of area saliency average in one hemisphere and standard deviation of equivalent area in opposite hemisphere in high and low severity cases. This comparison showed that LH structural modifications in high severity cases generate STD proportional to Avg saliency across subjects, indicative of proportional modifications across regions, such as inflated or compressed saliency. The degree of proportionality was smaller in low severity LH and high and low severity RH, and nonexistent across LH and RH. Together, these findings revealed systematic heterogeneity, especially pronounced in LH in high severity ASD and indicate that the heterogeneous structural impact of ASD on each left-brain region was proportional to the mean structural impact on that left brain region across affected individuals. In this regard, the structural saliency of all regions within each lobe in the LH for high and low severity were not normally distributed (see diagonal panels with distribution plots in Fig. 8), especially compared to the right hemisphere. This strongly suggests that structural aberrations in LH regions were not an artifact of random distribution across MRI scans and there were macro-structural effects reflected in MRI saliency. This is consistent with the dominant involvement of left hemisphere in ASD.

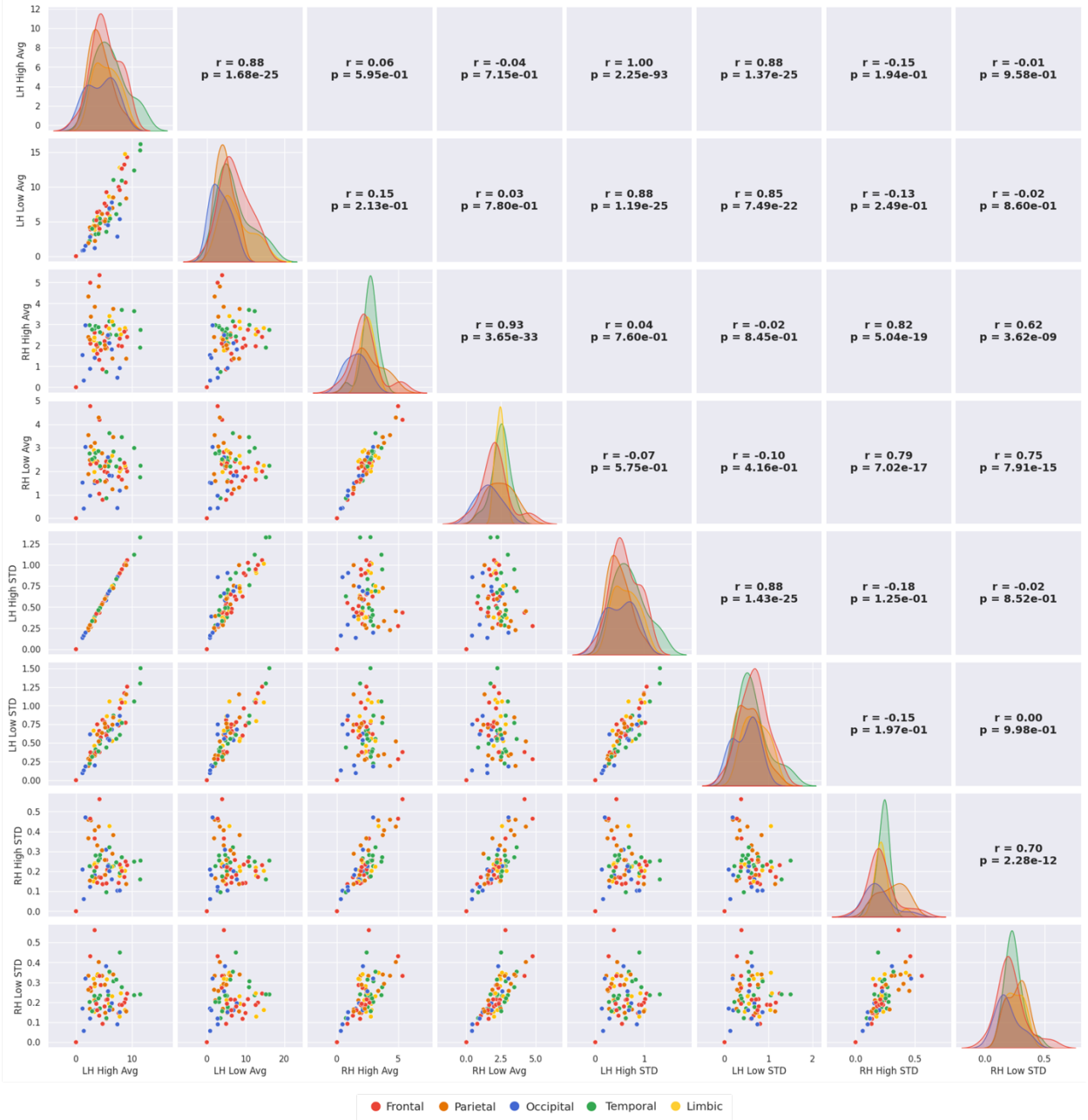

Figure 8. Scatter plots and distributions of the average regional saliency and the associated standard deviation for each cortical brain region. Plots are organized on an 8×8 grid, based on specifications in Fig 3. Each point in the scatter plot belongs to a brain region from Destrieux atlas similar to Fig. 3, total of 75 regions. Scatter plot axes are restricted to the left (LH) or right (RH) hemisphere, high (High) or low (Low) severity score cases, and the metrics include average (Avg) and standard deviation (STD) across the targeted cases (high or low severity). For example, an axis label of LH High STD refers to regions in the left hemisphere of all high severity cases averaged for each brain region. Therefore, each scatter plot has 75 points corresponding to one of the 75 brain regions listed in Fig. 3. Brain regions belonging to each lobe are color coded in all scatter and distribution (diagonal of the grid) plots. Key findings include a) maximum trend of *linear* scatter within LH high score, less in LH low score, RH high and low scores, and isotropic or no trend scatter across LH and RH (either with high or low score), b) distribution plots on the grid diagonal are more normal like for RH and low severity and max deviated from normal for LH high severity. For interpretations, see the text.

Correlating regional saliency and ASD severity across cases. We analyzed correlations using Kendall rank correlation coefficient τ between the SRS subscores (cognition, awareness, communication, mannerisms, and motivation) and the saliency of each brain region across low to high severity ASD cases. For these analyses, we considered 150 regions (2×75 for both left and right hemispheres) labeled using the Destrieux atlas parcellation and found that cognition, awareness, communication, mannerisms, and motivation subscores were significantly ($p < 0.05$) correlated with a distinct number of salient regions in each lobe of the left and right hemisphere (8/7, 12/0, 8/4, 22/7, 7/2, respectively; Table 1).

Sub-score	Left Hemisphere	Right Hemisphere
Cognition	Suborbital sulcus, Superior frontal gyrus, Frontomarginal sulcus and gyrus, Transverse frontopolar gyri and sulci, Transverse temporal sulcus, Angular gyrus, Superior parietal lobule, Paracentral lobule and sulcus	Superior frontal sulcus, Orbital part of inferior frontal gyrus, Suborbital sulcus, Inferior frontal sulcus, Horizontal ramus of lateral sulcus, Angular gyrus, Short insular gyri
Awareness	Lateral orbital sulcus, Medial orbital sulcus, Orbital gyri, Fronto-marginal gyrus and sulcus, Suborbital sulcus, Transverse frontopolar gyri and sulci, Superior frontal gyrus, Paracentral lobule and sulcus, Horizontal ramus of lateral sulcus, Superior parietal gyrus, aMCC, Short insular gyri	None
Commun-ication	Middle frontal sulcus, Suborbital sulcus, Fronto-marginal sulci and gyri, Transverse frontopolar gyri and sulci, Superior frontal gyrus, Vertical ramus of lateral sulcus, Superior parietal lobule, Short insular gyri	Orbital part of inferior frontal gyrus, Suborbital sulcus, aMCC, Short insular gyri
Mannerism	Opercular part of inferior frontal gyrus, Lateral orbital sulcus, Middle frontal sulcus, Middle frontal gyrus, Suborbital sulcus, Fronto-marginal gyrus and sulcus, Transverse frontopolar gyri and sulci, Superior frontal gyrus, Horizontal ramus of lateral sulcus, Lingual gyrus, Vertical ramus of lateral fissure, Angular gyrus, Postcentral sulcus, Superior parietal lobule, Paracentral lobule and sulcus, Middle occipital sulcus, Superior occipital sulcus, Cuneus, Calcarine sulcus, aMCC, Superior circular sulcus of insula, Short insular gyri	Opercular part of inferior frontal gyrus, Suborbital sulcus, Paracentral lobule and sulcus, Subparietal sulcus, Lingual part of the medial occipito-temporal gyrus, aMCC, Marginal cingulate sulcus
Motivation	Suborbital sulcus, Fronto-marginal gyrus and sulcus, Superior frontal gyrus, Transverse frontopolar gyri and sulci, Superior parietal lobule, Paracentral lobule and sulcus, Short insular gyri	Orbital part of inferior frontal gyrus, Short insular gyri

Table 1. Brain regions associated with each ASD sub-score, grouped by hemisphere. Font colors reflect anatomical lobe classification (Frontal lobe, red; Temporal lobe, green; Parietal lobe, brown; Occipital lobe, blue; Limbic regions, yellow).

In summary, comparison across lobes revealed saliency correlation of more regions of the frontal lobe with SRS subscores, both in left and right hemispheres, as well as the dominance of the left hemisphere, in terms of the number of regions that had significant correlations with SRS subscores.

Typically developed vs. low severity brains. We used the hybrid 3D-CNN we developed and trained to classify sMRI scans only from individuals with high- and low-severity ASD to now classify scans from control, neurotypical subjects (therefore with no SRS score). Scans from this control group were predicted with 86.47% accuracy as low severity subjects by the 3D ResNet50 trained on low and high severity ASD subjects. Saliency values of all regions mapped according to the Destrieux atlas for the low severity vs. control subjects showed a linear relationship, suggesting that the two groups share similar defining characteristics and are easily distinguishable

from the high severity group (Supp. Fig. S4). This high resemblance between the control group and low severity allowed us to effectively treat the structural aberrations in ASD along with typical structural variations in neurotypical brain as a spectrum.

Discussion

Advances in genetic, molecular, histopathological, and imaging studies, as well as clinical diagnosis of neurodevelopmental, neurodegenerative, and mental disorders, provide a wealth of data that can facilitate fine quantitative characterization of their heterogenous nature. Combined with major developments in the field of explainable AI, through techniques that can highlight the salient input features contributing to classification, make it possible to systematically interrogate and identify the underlying structural and functional correlates of such heterogeneities

Within this framework, in this study we developed explainable AI approaches and applied these to study *in vivo* structural MRI sequences from individuals with ASD from the ABIDE II dataset with the goal to correlate key structural changes in the autistic brain with the functional, cognitive, and behavioral heterogeneity of ASD. We parsed heterogeneity through a) correlations between the phenotype spectrum, in terms of diagnostic subscores, with the AI derived spectrum of structural and anatomical saliency, and b) pairwise correlations between the AI derived spectrum of structural saliency as a pointer to connection and brain network changes in ASD. Using this approach, we developed and utilized state-of-the-art AI networks, algorithms, and their ensembles to find an optimal combination that achieved the highest accuracy in low vs. high severity ASD classification. Then, based on the best AI ensemble, we extracted salient brain regions that contributed to the classification and measured a) the correlation of their saliency level with the cognitive and behavioral heterogeneity, and b) their pairwise saliency correlation as pointers to brain functional networks and their heterogenous modifications in ASD. After extensive testing of 2D and 3D CNNs and AI algorithms, we found that the 3D ResNet-50, as a feature extractor for the structural MRI, and the kNN algorithm, as a classifier of the features, formed the optimal ensemble that yielded the highest accuracy for classifying ASD severity from structural MRI. Then based on this optimal ensembled AI, we extracted the regions of interest, mapped through cortical and subcortical segmentation, by generating MRI saliency maps using a CNN gradient method.

Our findings about brain structural changes in ASD can be summarized as a) differential involvement of left vs right hemispheres with more involvement of left hemispheric regions and networks, b) more extent (connection number/correlation count) in the ASD involvement of the left brain, c) more left brain structural changes correlated with ASD subscores and phenotype, and therefore d) more reflection of ASD heterogeneity in left brain correlated with phenotypical and structural changes. Among brain lobes and networks, the frontal and temporal lobes and the language, default mode, central executive, and dorsal/ventral attention, limbic emotion, and salience networks showed most structural changes in ASD. Our systematic Al approach leveraged the phenotypical heterogeneity and spectral nature of ASD to uncover structural brain region

heterogeneity in ASD, as well as correlations between areas and across structural and phenotypical/diagnostic heterogeneity. This revealed global, regional, and network changes in ASD associated with symptom severity and diagnostic scores. Finally, our structural analysis showed that the neurotypical brain fell on the low-end tail of the low-high severity ASD spectrum.

The constellation and configuration of high vs. low saliency and correlations we found, revealed a large number of networks and areas that are affected in ASD, including but not limited to multiple brain regions that are posterior components of the default mode network, bilateral precuneus, right inferior occipital gyrus, left cerebellar regions (Liloia et al. 2023). Significant positive or negative structural saliency correlations revealed in this study can be in part the origin or outcome of functional connectivity changes (over- or under-connectivity) in ASD, together motivating the consideration of brain structural phenotypes of ASD and laying a framework for neuroanatomical heterogeneity in ASD (Shan et al. 2022). Overall, ASD exhibited heterogenous and widespread structural and functional feature prints. The saliency overlay results, showed distribution across areas and hemispheres with high saliency in frontal, limbic and temporal areas. Areas with a role in speech, language, comprehension, and non-verbal communication displayed high saliency. These included pars opercularis, pars orbitalis, pars triangularis, inferior frontal regions (including Broca's area), superior temporal gyrus (Bigler et al. 2007; Foundas et al. 1996; Peeva et al. 2013; Yamasaki et al. 2010), and the horizontal, vertical, and posterior rami of the lateral sulcus, which define the pars triangularis, and pars opercularis regions and also hold and Wernicke's area (Chaichana and Quiñones-Hinojosa 2019; Griffiths et al. 2010). Other areas with high saliency included the insula that has a role in sensory integration and regulation of emotion (Viñas-Guasch and Wu 2017), and the temporal pole, which is commonly associated with visual processing, socio-emotional processing, and other cognitive processes (Herlin, Navarro, and Dupont 2021). Interestingly, left hemisphere regions throughout most of the cortex showed higher saliency values and larger separation across the spectrum of low-high ASD severity whereas regional saliency in the right hemisphere remained lower and with smaller separation across the ASD spectrum.

The saliency-based correlations highlighted involved networks and connectivities and facilitated ranking of regions in terms of affected connectivity. In line with prior reports [reviewed in (Amaral et al. 2008)], our findings also showed the heterogeneous and distributed neuroanatomy of ASD (Amaral et al. 2008), which is associated with functional connectivity and network aberrations (Müller 2007), including widespread reported over- and under-connectivity (Baran et al. 2023; Cherkassky et al. 2006; Just et al. 2007, 2007; Kenet et al. 2012; Khan et al. 2013, 2013; Müller 2007, 2007). Our analysis revealed functional networks implicated in ASD, with the language, default mode (DMN), central executive (CEN or frontoparietal), dorsal/ventral attention, limbic/emotion, and salience (SN), networks at the top of the list. These findings are in line with previous studies by Uddin et al. (2013) that also highlighted the involvement of these and other functional networks, connections, and regions in ASD. Further support for our findings comes from observations of weaker resting state connectivity across distributed networks in ASD,

including frontoamygdala functional connectivity that likely contributes to socioemotional impairments (Odriozola et al. 2019). Resting state underconnectivity has been reported in functional studies using the ABIDE dataset (Duan et al. 2017), indicating underconnectivity in areas within the default mode network (DMN) and within the ventral attention network (VAN), as well as between pairs of networks, including DMN and Visual Network (VN), dorsal attention network (DAN) and VAN, and limbic network (LN) and frontoparietal network (FPN). Schipul et al. (Schipul et al. 2012) showed decreased functional and structural connectivity measures consistent with the corpus collosum size change in ASD compared to TD, which could account for some of the hemispheric differences we observed. Functional and anatomical underconnectivity has also been noted by (Just et al. 2007), using functional measures (synchronization) between parietal and frontal areas in proportion with the smaller size of the genu of corpus callosum. Local and long-range functional connectivity was shown to be reduced within FFA, and between the FFA and three distant cortical regions, the left precuneus, left inferior frontal gyrus (IFG), and left anterior cingulate cortex (Khan et al. 2013), which were also salient in our analysis. The anterior cingulate cortex, in particular, has been a consistent site of ASD pathology in imaging (Thakkar et al. 2008) and histopathological and computational studies (Yazdanbakhsh et al. 2024; Zikopoulos, Liu, et al. 2018; Zikopoulos and Barbas 2010; Zikopoulos, García-Cabezas, and Barbas 2018). Moreover, several resting-state networks may be more loosely connected in ASD, based on lower connectivity of the anterior cingulate region with posterior cingulate and precuneus areas, between anterior and posterior cingulate, and lower functional connectivity between anterior and posterior medial cortices (Cherkassky et al. 2006). Following the same underconnectivity trend, reduced functional connectivity between the frontal eye fields (FEF) and dorsal anterior cingulate cortex, regions critical to volitional ocular motor control, has been reported in ASD (Kenet et al. 2012). In addition, a study that used whole sMRI and SVM for pattern classification, in a similar manner to this work, found distributed networks, limbic, frontal-striatal, fronto-temporal, fronto-parietal and cerebellar systems that distinguished ASD from controls, with better gray matter- compared to white matter-based classification than white matter only (Ecker et al. 2010). On the other hand, regions and networks that were salient in our analyses have also been implicated in reports of hyperconnectivity in the salience network (SN), composed of anterior cingulate cortex ACC, ventral anterior insular cortex, amygdala, hypothalamus, ventral striatum, thalamus, and brainstem nuclei (Uddin et al. 2013), as well as social network-related long- and short- range hyperconnectivity (Supekar et al. 2013).

Methodological considerations and limitations

In this study we used both 3D and 2D neural networks. The 3D ResNet-50 allowed for the whole MRI volume to be used for training whereas 2D neural networks only trained on slices of the MRI. The volumetric nature of the 3D ResNet-50 highlighted patterns that extended across slices, so the 3D ResNet-50 outperformed the 2D ResNet-50 in the sagittal, axial, and coronal planes, demonstrating the importance of data across MRI slices rather than the data on the slice itself.

However, 2D neural networks have become very sophisticated in recent years while 3D neural networks are extremely prone to variability in data. It was therefore essential, for accurate and consistent performance of the 3D neural network, to reduce variability through MRI preprocessing.

In order to use the 3D ResNet-50 as a feature extractor, we modified the model to output feature vectors to a SVM and a kNN. The SVM and kNN essentially compared each test MRI subject with the training subjects. The SVM fitted a hyperplane to the training data and made its prediction of the testing data based on the hyperplane while, the kNN compared the testing data on a *n*-dimensional plane to the *k* nearest training data (Euclidian distance). With a relatively low number of subjects available, the SVM tended to overfit the hyperplane to the training data causing inaccurate predictions of the testing data. However, the kNN was successful in making correct predictions as it merely compared and not fitted the testing to the training data.

Importantly, our study focused on male subjects, due to limited availability of MRI sequences and accompanying diagnostic/phenotypic data for females therefore, reflecting heterogeneity also seen in other studies of male individuals (Chen et al. 2019). Moreover, the amount of publicly available MRI datasets that include phenotypic data such as the SRS Total and its subscores is limited, which naturally could make our ensembled AI prone to overfitting, impacting extracted regions of interest, based on saliency maps and saliency magnitude.

Additionally, the saliency maps depict the contribution of each voxel to making the low to high severity ASD prediction, as a macrostructural pointer to ASD, and the highly significant saliency correlations across pairs of regions can only suggest their involvement in ASD as part of brain functional networks. Nuances such as intra-areal and inter-areal hypo- and hyper-connectivity changes in ASD (Cherkassky et al. 2006; Just et al. 2007; Kenet et al. 2012; Khan et al. 2013; Müller 2007; Supekar et al. 2013) appear as a tip of iceberg in saliency correlation analysis, and neuronal, connectional and neurobiological changes in ASD demands approaches with much higher spatial resolution. Therefore, the current study can at best outline underlying brain region and network heterogeneity in the severity of ASD.

Finally, it is worth noting that we did not train the network on any typically developed (TD) brain MRIs. Instead, we used TD brain MRIs as a test input into our final, developed 3D network ensemble and extracted brain region saliency to compare with ASD population results. This combination revealed the congruency of saliency assignment by the network for TD and low severity ASD subjects, with TD saliency on the low-end tail of the distribution. This key finding suggests a continuous and spectral heterogeneity of structural brain changes across subjects, with low severity ASD and TD individuals lying at one end of the structural spectrum.

Conclusions

Our findings showed that structural MRI contains information that reflects functional network and diagnostic score changes in ASD, which can be identified, using state-of-the-art AI, to pinpoint salient brain regions and their networks underlying ASD heterogeneity. Our analyses provided distinct involvement measures of frontal, temporal, parietal, and occipital lobes, and limbic, insular, and subcortical regions and their constituent areas and subregions in social-awareness, attention, social interactions, and cognitive control scores. Additional correlation and subsequent ranking revealed hierarchically arranged involvement of functional networks in ASD, which, in descending order of impact, included language, default mode (DMN), central executive (CEN or frontoparietal) & dorsal/ventral attention, limbic/emotion networks, salience (SN), sensorimotor, and social, networks. These findings highlight novel, detailed, and systematic anatomical, functional, and phenotypical changes and associations that can be used as biomarkers to parse and frame the heterogeneity and spectral nature of ASD.

References

Aghdam, Maryam Akhavan, Arash Sharifi, and Mir Mohsen Pedram. 2019. 'Diagnosis of Autism Spectrum Disorders in Young Children Based on Resting-State Functional Magnetic Resonance Imaging Data Using Convolutional Neural Networks'. *Journal of Digital Imaging* 32(6):899–918. doi:10.1007/s10278-019-00196-1.

Aglinskas, Aidas, Joshua K. Hartshorne, and Stefano Anzellotti. 2022. 'Contrastive Machine Learning Reveals the Structure of Neuroanatomical Variation within Autism'. *Science* 376(6597):1070–74. doi:10.1126/science.abm2461.

Amaral, David G., Derek Sayre Andrews, and Christine Wu Nordahl. 2024. 'Structural Brain Imaging Biomarkers of Autism Spectrum Disorder'. Pp. 491–509 in.

Amaral, David G., Cynthia Mills Schumann, and Christine Wu Nordahl. 2008. 'Neuroanatomy of Autism'. *Trends in Neurosciences* 31(3):137–45. doi:10.1016/j.tins.2007.12.005.

Baran, Bengi, Quynh Trang Huong Nguyen, Dimitrios Mylonas, Susan L. Santangelo, and Dara S. Manoach. 2023. 'Increased Resting-state Thalamocortical Functional Connectivity in Children and Young Adults with Autism Spectrum Disorder'. *Autism Research* 16(2):271–79. doi:10.1002/aur.2875.

Bigler, Erin D., Sherstin Mortensen, E. Shannon Neeley, Sally Ozonoff, Lori Krasny, Michael Johnson, Jeffrey Lu, Sherri L. Provencal, William McMahon, and Janet E. Lainhart. 2007. 'Superior Temporal Gyrus, Language Function, and Autism'. *Developmental Neuropsychology* 31(2):217–38. doi:10.1080/87565640701190841.

Bruni, Teryn P. 2014a. 'Test Review: Social Responsiveness Scale–Second Edition (SRS-2)'. *Journal of Psychoeducational Assessment* 32(4):365–69. doi:10.1177/0734282913517525.

Bruni, Teryn P. 2014b. 'Test Review: Social Responsiveness Scale–Second Edition (SRS-2)'. *Journal of Psychoeducational Assessment* 32(4):365–69. doi:10.1177/0734282913517525.

Chaichana, Kaisorn, and Alfredo Quiñones-Hinojosa. 2019. *Comprehensive Overview of Modern Surgical Approaches to Intrinsic Brain Tumors*. Elsevier.

Cherkassky, Vladimir L., Rajesh K. Kana, Timothy A. Keller, and Marcel Adam Just. 2006. 'Functional Connectivity in a Baseline Resting-State Network in Autism'. *NeuroReport* 17(16):1687–90. doi:10.1097/01.wnr.0000239956.45448.4c.

Collins, D. L., P. Neelin, T. M. Peters, and A. C. Evans. 1994. 'Automatic 3D Intersubject Registration of MR Volumetric Data in Standardized Talairach Space.' *Journal of Computer Assisted Tomography* 18(2):192–205.

Cortes, Corinna, and Vladimir Vapnik. 1995. 'Support-Vector Networks'. *Machine Learning* 20(3):273–97. doi:10.1007/BF00994018.

Cover, T., and P. Hart. 1967. 'Nearest Neighbor Pattern Classification'. *IEEE Transactions on Information Theory* 13(1):21–27. doi:10.1109/TIT.1967.1053964.

Dale, Anders M., Bruce Fischl, and Martin I. Sereno. 1999. 'Cortical Surface-Based Analysis'. *NeuroImage* 9(2):179–94. doi:10.1006/nimg.1998.0395.

Destrieux, Christophe, Bruce Fischl, Anders Dale, and Eric Halgren. 2010. 'Automatic Parcellation of Human Cortical Gyri and Sulci Using Standard Anatomical Nomenclature'. *NeuroImage* 53(1):1–15. doi:10.1016/j.neuroimage.2010.06.010.

Duan, Xujun, Heng Chen, Changchun He, Zhiliang Long, Xiaonan Guo, Yuanyue Zhou, Lucina Q. Uddin, and Huafu Chen. 2017. 'Resting-State Functional under-Connectivity within and between Large-Scale Cortical Networks across Three Low-Frequency Bands in Adolescents with Autism'. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 79:434–41. doi:10.1016/j.pnpbp.2017.07.027.

Duan, Xujun, Xiaolong Shan, Lucina Q. Uddin, and Huafu Chen. 2025. 'The Future of Disentangling the Heterogeneity of Autism With Neuroimaging Studies'. *Biological Psychiatry* 97(5):428–38. doi:10.1016/j.biopsych.2024.08.008.

Ecker, Christine. 2012. 'Brain Anatomy and Its Relationship to Behavior in Adults With Autism Spectrum Disorder'. *Archives of General Psychiatry* 69(2):195. doi:10.1001/archgenpsychiatry.2011.1251.

Ecker, Christine, Vanessa Rocha-Rego, Patrick Johnston, Janaina Mourao-Miranda, Andre Marquand, Eileen M. Daly, Michael J. Brammer, Clodagh Murphy, and Declan G. Murphy. 2010. 'Investigating the Predictive Value of Whole-Brain Structural MR Scans in Autism: A Pattern Classification Approach'. *NeuroImage* 49(1):44–56. doi:10.1016/j.neuroimage.2009.08.024.

Fischl, Bruce. 2012. 'FreeSurfer'. *NeuroImage* 62(2):774–81. doi:10.1016/j.neuroimage.2012.01.021.

Foundas, A. L., C. M. Leonard, R. L. Gilmore, E. B. Fennell, and K. M. Heilman. 1996. 'Pars Triangularis Asymmetry and Language Dominance.' *Proceedings of the National Academy of Sciences* 93(2):719–22. doi:10.1073/pnas.93.2.719.

Fruchterman, Thomas M. J., and Edward M. Reingold. 1991. 'Graph Drawing by Force-directed Placement'. *Software: Practice and Experience* 21(11):1129–64. doi:10.1002/spe.4380211102.

Griffiths, Paul D., Janet Morris, Jeanne-Claudie Larroche, and Michael Reeves. 2010. *Atlas of Fetal and Postnatal Brain MR*. Elsevier.

Hamaguchi, Hiroki, Naoki Marumo, and Akiko Takeda. 2025. *Initial Placement for Fruchterman-Reingold Force Model with Coordinate Newton Direction*. https://doi.org/10.48550/arXiv.2412.20317.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. 'Deep Residual Learning for Image Recognition'.

Herlin, Bastien, Vincent Navarro, and Sophie Dupont. 2021. 'The Temporal Pole: From Anatomy to Function—A Literature Appraisal'. *Journal of Chemical Neuroanatomy* 113:101925. doi:10.1016/j.jchemneu.2021.101925.

Just, Marcel Adam, Vladimir L. Cherkassky, Timothy A. Keller, Rajesh K. Kana, and Nancy J. Minshew. 2007. 'Functional and Anatomical Cortical Underconnectivity in Autism: Evidence from an FMRI Study of an Executive Function Task and Corpus Callosum Morphometry'. *Cerebral Cortex* 17(4):951–61. doi:10.1093/cercor/bhl006.

Kenet, Tal, Elena V. Orekhova, Hari Bharadwaj, Nandita R. Shetty, Emily Israeli, Adrian K. C. Lee, Yigal Agam, Mikael Elam, Robert M. Joseph, Matti S. Hämäläinen, and Dara S. Manoach. 2012. 'Disconnectivity of the Cortical Ocular Motor Control Network in Autism Spectrum Disorders'. *NeuroImage* 61(4):1226–34. doi:10.1016/j.neuroimage.2012.03.010.

Khan, Sheraz, Alexandre Gramfort, Nandita R. Shetty, Manfred G. Kitzbichler, Santosh Ganesan, Joseph M. Moran, Su Mei Lee, John D. E. Gabrieli, Helen B. Tager-Flusberg, Robert M. Joseph, Martha R. Herbert, Matti S. Hämäläinen, and Tal Kenet. 2013. 'Local and Long-Range Functional Connectivity Is Reduced in Concert in Autism Spectrum Disorders'. *Proceedings of the National Academy of Sciences* 110(8):3107–12. doi:10.1073/pnas.1214533110.

Khosla, Meenakshi, Keith Jamison, Amy Kuceyeski, and Mert Sabuncu. 2018. '3D Convolutional Neural Networks for Classification of Functional Connectomes'.

Li, Xiaoxiao, Nicha C. Dvornek, Xenophon Papademetris, Juntang Zhuang, Lawrence H. Staib, Pamela Ventola, and James S. Duncan. 2018. '2-Channel Convolutional 3D Deep Neural Network (2CC3D) for FMRI Analysis: ASD Classification and Feature Learning'. Pp. 1252–55 in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE.

Loomes, Rachel, Laura Hull, and William Polmear Locke Mandy. 2017. 'What Is the Male-to-Female Ratio in Autism Spectrum Disorder? A Systematic Review and Meta-Analysis'. *Journal of the American Academy of Child & Adolescent Psychiatry* 56(6):466–74. doi:10.1016/j.jaac.2017.03.013.

Lord, Catherine, Edwin H. Cook, Bennett L. Leventhal, and David G. Amaral. 2000. 'Autism Spectrum Disorders'. *Neuron* 28(2):355–63. doi:10.1016/S0896-6273(00)00115-X.

Lord, Catherine, Michael Rutter, and Ann Le Couteur. 1994. 'Autism Diagnostic Interview-Revised: A Revised Version of a Diagnostic Interview for Caregivers of Individuals with Possible Pervasive Developmental Disorders'. *Journal of Autism and Developmental Disorders* 24(5):659–85. doi:10.1007/BF02172145.

Di Martino, Adriana, David O'Connor, Bosi Chen, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Joshua H. Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, Laura M. E. Blanken, Susan Y. Bookheimer, B. Blair Braden, Lisa Byrge, F. Xavier Castellanos, Mirella Dapretto, Richard Delorme, Damien A. Fair, Inna Fishman, Jacqueline Fitzgerald, Louise Gallagher, R. Joanne Jao Keehn, Daniel P. Kennedy, Janet E. Lainhart, Beatriz Luna, Stewart H. Mostofsky, Ralph-Axel Müller, Mary Beth Nebel, Joel T. Nigg, Kirsten O'Hearn, Marjorie Solomon, Roberto Toro, Chandan J. Vaidya, Nicole Wenderoth, Tonya White, R. Cameron Craddock, Catherine Lord, Bennett Leventhal, and Michael P. Milham. 2017. 'Enhancing Studies of the Connectome in Autism Using the Autism Brain Imaging Data Exchange II'. *Scientific Data* 4(1):170010. doi:10.1038/sdata.2017.10.

Mikl, Michal, Radek Mareček, Petr Hluštík, Martina Pavlicová, Aleš Drastich, Pavel Chlebus, Milan Brázdil, and Petr Krupa. 2008. 'Effects of Spatial Smoothing on FMRI Group Inferences'. *Magnetic Resonance Imaging* 26(4):490–503. doi:10.1016/j.mri.2007.08.006.

Müller, Ralph-Axel. 2007. 'The Study of Autism as a Distributed Disorder'. *Mental Retardation and Developmental Disabilities Research Reviews* 13(1):85–95. doi:10.1002/mrdd.20141.

NetworkX Developers. 2025. 'Spring_layout in NetworkX, Network Analysis in Python'.

Niu, Ke, Jiayang Guo, Yijie Pan, Xin Gao, Xueping Peng, Ning Li, and Hailong Li. 2020. 'Multichannel Deep Attention Neural Networks for the Classification of Autism Spectrum Disorder Using Neuroimaging and Personal Characteristic Data'. *Complexity* 2020:1–9. doi:10.1155/2020/1357853.

Odriozola, Paola, Dina R. Dajani, Catherine A. Burrows, Laurel J. Gabard-Durnam, Emma Goodman, Adriana C. Baez, Nim Tottenham, Lucina Q. Uddin, and Dylan G. Gee. 2019. 'Atypical Frontoamygdala Functional Connectivity in Youth with Autism'. *Developmental Cognitive Neuroscience* 37:100603. doi:10.1016/j.dcn.2018.12.001.

Peeva, M. G., J. A. Tourville, Y. Agam, B. Holland, D. S. Manoach, and F. H. Guenther. 2013. 'White Matter Impairment in the Speech Network of Individuals with Autism Spectrum Disorder'. *NeuroImage: Clinical* 3:234–41. doi:10.1016/j.nicl.2013.08.011.

Schipul, Sarah E., Diane L. Williams, Timothy A. Keller, Nancy J. Minshew, and Marcel Adam Just. 2012. 'Distinctive Neural Processes during Learning in Autism'. *Cerebral Cortex* 22(4):937–50. doi:10.1093/cercor/bhr162.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. 'Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps'.

Smith, Stephen M., Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E. J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. 2004. 'Advances in Functional and Structural MR Image Analysis and Implementation as FSL'. *NeuroImage* 23:S208–19. doi:10.1016/j.neuroimage.2004.07.051.

Supekar, Kaustubh, Lucina Q. Uddin, Amirah Khouzam, Jennifer Phillips, William D. Gaillard, Lauren E. Kenworthy, Benjamin E. Yerys, Chandan J. Vaidya, and Vinod Menon. 2013. 'Brain Hyperconnectivity in Children with Autism and Its Links to Social Deficits'. *Cell Reports* 5(3):738–47. doi:10.1016/j.celrep.2013.10.001.

Thakkar, Katharine N., Frida E. Polli, Robert M. Joseph, David S. Tuch, Nouchine Hadjikhani, Jason J. S. Barton, and Dara S. Manoach. 2008. 'Response Monitoring, Repetitive Behaviour and Anterior Cingulate Abnormalities in Autism Spectrum Disorders (ASD)'. *Brain* 131(9):2464–78. doi:10.1093/brain/awn099.

Uddin, Lucina Q., Kaustubh Supekar, Charles J. Lynch, Amirah Khouzam, Jennifer Phillips, Carl Feinstein, Srikanth Ryali, and Vinod Menon. 2013. 'Salience Network–Based Classification and Prediction of Symptom Severity in Children With Autism'. *JAMA Psychiatry* 70(8):869. doi:10.1001/jamapsychiatry.2013.104.

Viñas-Guasch, Nestor, and Yan Jing Wu. 2017. 'The Role of the Putamen in Language: A Meta-Analytic Connectivity Modeling Study'. *Brain Structure and Function* 222(9):3991–4004. doi:10.1007/s00429-017-1450-y.

Yamasaki, Syudo, Hidenori Yamasue, Osamu Abe, Motomu Suga, Haruyasu Yamada, Hideyuki Inoue, Hitoshi Kuwabara, Yuki Kawakubo, Noriaki Yahata, Shigeki Aoki, Yukiko Kano, Nobumasa Kato, and Kiyoto Kasai. 2010. 'Reduced Gray Matter Volume of Pars

Opercularis Is Associated with Impaired Social Communication in High-Functioning Autism Spectrum Disorders'. *Biological Psychiatry* 68(12):1141–47. doi:10.1016/j.biopsych.2010.07.012.

Yazdanbakhsh, Arash, Kim Dang, Kelvin Kuang, Tingru Lian, Xuefeng Liu, Songlin Xie, and Basilis Zikopoulos. 2024. 'Artificial Intelligence Networks Combining Histopathology and Machine Learning Can Extract Axon Pathology in Autism Spectrum Disorder (Accepted in Autism Research), DOI: 10.1002/Aur.70135'.

Zhao, Yu, Fangfei Ge, Shu Zhang, and Tianming Liu. 2018a. '3D Deep Convolutional Neural Network Revealed the Value of Brain Network Overlap in Differentiating Autism Spectrum Disorder from Healthy Controls'. Pp. 172–80 in.

Zikopoulos, Basilis, and Helen Barbas. 2010. 'Changes in Prefrontal Axons May Disrupt the Network in Autism'. *The Journal of Neuroscience* 30(44):14595–609. doi:10.1523/JNEUROSCI.2257-10.2010.

Zikopoulos, Basilis, Miguel Ángel García-Cabezas, and Helen Barbas. 2018. 'Parallel Trends in Cortical Gray and White Matter Architecture and Connections in Primates Allow Fine Study of Pathways in Humans and Reveal Network Disruptions in Autism'. *PLOS Biology* 16(2):e2004559. doi:10.1371/journal.pbio.2004559.

Zikopoulos, Basilis, Xuefeng Liu, Justin Tepe, Iris Trutzer, Yohan J. John, and Helen Barbas. 2018. 'Opposite Development of Short- and Long-Range Anterior Cingulate Pathways in Autism'. *Acta Neuropathologica* 136(5):759–78. doi:10.1007/s00401-018-1904-1.

Supplementary Materials

ABIDE-II Site	Total Subjects	Low Severity Subjects	High Severity Subjects	Total Subjects Used
NYU-1	72	15	22	37
NYU-2	26	0	18	18
SDSU	55	6	11	17
TCD	41	2	2	4
OHSU	79	11	7	18
KKI	147	41	20	61

Table S1: Number of subjects acquired from each ABIDE-II site and the split of low and high severity subjects used per site. Total subjects refer to the total number of subjects each site provides to the ABIDE-II database whereas, total subjects used refers to the number of subjects that fit the criteria of male, age 5 to 12 and fit the SRS Total threshold. Note that some subjects are not used in analysis to have a balanced low severity to high severity subject ratio.

Slice Range	Sagittal	Axial	Coronal
24-31	62.90%	52.82%	51.61%
32-39	58.06%	51.61%	53.23%
40-47	60.08%	63.71%	68.15%
48-55	60.08%	60.08%	68.15%
56-63	57.66%	49.19%	64.52%
64-71	70.97%	57.66%	58.47%
72-79	63.71%	71.77%	52.42%
80-87	56.05%	62.50%	64.92%
88-95	58.06%	57.66%	62.90%
96-104	67.74%	51.61%	58.47%

Table S2: Averaged validation of the 2D ResNet-50 on each viable slice range on the sagittal, axial, and coronal axis. Averages were calculated based on three trials of the 2D ResNet-50 on the same slice range and same axis. Slices outside of the range included above were not trained on due to inadequate brain material (skull only and background).

Table S3

τ=0.1 Brain Area	Connections	τ=0.2 Brain Area	Connections	τ=0.3 Brain Area	Connections
RH_S_temporal_sup	126	RH_S_temporal_sup	96	LH_S_circular_insula_sup	55
RH_G_occipital_middle	125	RH_G_occipital_middle	88	LH_G_front_inf-Opercular	50
LH_G_Ins_lg_and_S_cent_ins	125	LH_G_Ins_lg_and_S_cent_ins	87	LH_G_and_S_subcentral	49
RH_S_precentral-sup-part	124	LH_S_circular_insula_sup	82	LH_G_temp_sup-Plan_tempo	47
LH_G_front_inf-Opercular	123	LH_G_front_inf-Opercular	82	RH_G_occipital_middle	46
LH_S_circular_insula_sup	121	LH_G_and_S_subcentral	81	LH_S_front_inf	46
RH_S_central	120	RH_S_occipital_ant	80	RH_G_temporal_middle	44
RH_G_temporal_middle	119	RH_G_temporal_middle	80	RH_S_temporal_sup	43
LH_G_orbital	119	RH_G_temporal_inf	80	RH_G_temporal_inf	43
LH_G_front_inf-Orbital	119	LH_S_circular_insula_inf	80	RH_G_front_sup	43
RH_G_postcentral	118	RH_S_temporal_inf	79	LH_S_precentral-inf-part	42
LH_G_temp_sup-Plan_tempo	118	RH_G_front_sup	79	LH_Lat_Fis-post	42
RH_G_subcallosal	117	LH_S_front_inf	79	LH_G_temp_sup-G_T_transv	42
LH_Pole_temporal	117	LH_G_temp_sup-Plan_tempo	79	LH_G_insular_short	42
RH_S_oc-temp_lat	116	RH_S_precentral-sup-part	78	LH_G_front_middle	42
RH_G_and_S_cingul-Ant	116	RH_S_oc_sup_and_transversal	78	LH_G_front_inf-Triangul	42
LH_S_pericallosal	116	RH_G_and_S_cingul-Ant	77	LH_G_Ins_lg_and_S_cent_ins	42
LH_S_calcarine	116	LH_Lat_Fis-post	76	RH_G_and_S_cingul-Ant	40
LH_G_temp_sup-Lateral	116	LH_Lat_Fis-ant-Vertical	76	LH_G_and_S_cingul-Mid-Ant	40
LH_G_insular_short	116	LH_G_insular_short	76	LH_Lat_Fis-ant-Vertical	39
RH_S_occipital_ant	115	LH_G_front_inf-Triangul	76	RH_S_temporal_inf	38
LH_S_circular_insula_ant	115	LH_G_temp_sup-G_T_transv	75	LH_S_front_sup	37
LH_G_temp_sup-G_T_transv	115	LH_G_pariet_inf-Supramar	75	LH_G_pariet_inf-Supramar	37
RH_S_temporal_inf	114	LH_G_front_middle	75	LH_S_temporal_sup	36
RH_G_temporal_inf	114	RH_S_oc_middle_and_Lunatus	74	LH_G_precentral	35
RH_G_front_sup	114	LH_S_precentral-inf-part	74	RH_S_occipital_ant	34
LH_S_temporal_sup	114	LH_S_pericallosal	74	LH_S_circular_insula_inf	34
LH_G_and_S_subcentral	114	LH_G_orbital	74	LH_G_and_S_cingul-Ant	34
RH_S_postcentral	113	RH_S_oc-temp_lat	72	RH_S_oc_sup_and_transversal	33
RH_G_temp_sup-Lateral	113	RH_Pole_temporal	72	RH_G_orbital	33
RH_S_oc_sup_and_transversal	112	LH_G_front_inf-Orbital	72	LH_S_temporal_transverse	33
RH_S_oc_middle_and_Lunatus	112	RH_S_central	69	LH_S_orbital-H_Shaped	33
RH_Pole_temporal	112	LH_S_circular_insula_ant	69	LH_G_orbital	32
RH_G_front_inf-Triangul	112	RH_S_intrapariet_and_P_trans	68	LH_G_oc-temp_med-Lingual	32
LH_Lat_Fis-ant-Vertical	112	LH_S_temporal_sup	67	RH_S_front_middle	31
LH_S_oc_middle_and_Lunatus	111	LH_S_calcarine	67	RH_Pole_temporal	31
LH_G_oc-temp_med-Lingual	111	LH_G_temp_sup-Lateral	67	LH_S_oc_middle_and_Lunatus	31
RH_G_rectus	110	RH_G_temp_sup-Lateral	66	LH_G_rectus	31
LH_S_orbital-H_Shaped	110	LH_S_oc_middle_and_Lunatus	66	RH_S_intrapariet_and_P_trans	30
LH_G_front_inf-Triangul	110	LH_G_oc-temp_med-Lingual	66	RH_S_central	30
RH_S_circular_insula_inf	109	RH_G_subcallosal	65	LH_Lat_Fis-ant-Horizont	30
LH_G_temp_sup-Plan_polar	109	LH_S_orbital-H_Shaped	65	LH_G_front_sup	30
RH_G_temp_sup-Plan_polar	108	RH_G_temp_sup-Plan_polar	64	LH_G_front_inf-Orbital	30
LH_S_circular_insula_inf	108	RH_G_rectus	64	RH_Pole_occipital	29
RH_S_intrapariet_and_P_trans	107	LH_S_front_sup	63	RH_G_rectus	29
RH_S_collat_transv_ant	106	RH_S_orbital_lateral	62	RH_S_oc_middle_and_Lunatus	28
RH_G_orbital	106	LH_Lat_Fis-ant-Horizont	62	RH_S_collat_transv_ant	28
RH_G_oc-temp_lat-fusifor	106	RH_S_collat_transv_ant	61	LH_S_orbital_lateral	28
LH_S_temporal_transverse	106	LH_S_temporal_transverse	61	LH_S_front_middle	28
RH_S_orbital_lateral	105	LH_Pole_temporal	61	LH_G_occipital_middle	28

τ=0.4 Brain Area	Connections	τ=0.5 Brain Area	Connections	τ=0.6 Brain Area	Connections
LH_S_front_inf	27	LH_S_front_inf	11	LH_S_temporal_inf	4
LH_S_circular_insula_sup	27	LH_S_circular_insula_sup	11	LH_G_temporal_inf	4
LH_G_front_middle	26	LH_G_precentral	11	LH_G_front_middle	4
LH_G_pariet_inf-Supramar	23	LH_S_front_sup	10	LH_S_suborbital	3
RH_G_and_S_cingul-Ant	21	LH_G_front_middle	10	LH_S_precentral-sup-part	3
RH_G_occipital_middle	20	LH_S_precentral-inf-part	9	LH_S_oc-temp_lat	3
LH_S_orbital-H_Shaped	20	LH_S_orbital-H_Shaped	9	LH_S_front_sup	3
LH_G_front_inf-Opercular	19	LH_G_temp_sup-Plan_tempo	9	LH_S_front_inf	3

LH_Lat_Fis-post	18	LH_G_and_S_cingul-Mid-Ant	9	LH_S_central	3
LH_G_temp_sup-Plan_tempo	18	LH_G_pariet_inf-Supramar	8	LH_G_precentral	3
LH_G_precentral	18	LH_G_and_S_cingul-Mid-Post	7	LH_G_postcentral	3
LH_G_front_inf-Triangul	18	RH_G_temporal_middle	6	LH_G_oc-temp_lat-fusifor	3
LH_G_and_S_cingul-Ant	18	RH_G_occipital_middle	6	RH_S_temporal_inf	2
RH_S_occipital_ant	17	LH_S_orbital_lateral	6	RH_G_temp_sup-Plan_polar	2
LH_S_precentral-inf-part	17	LH_S_front_middle	6	RH_G_rectus	2
LH_S_front_sup	17	LH_S_central	6	RH_G_occipital_middle	2
LH_G_and_S_subcentral	17	LH_G_temporal_middle	6	LH_S_orbital-H_Shaped	2
LH_S_orbital_lateral	16	LH_G_temp_sup-G_T_transv	6	LH_Lat_Fis-post	2
LH_S_oc-temp_med_and_Lingual	16	LH_G_rectus	6	LH_G_temporal_middle	2
LH_Lat_Fis-ant-Vertical	16	LH_G_front_sup	6	LH_G_temp_sup-Plan_tempo	2
LH_S_suborbital	15	LH_G_front_inf-Triangul	6	LH_G_temp_sup-G_T_transv	2
LH_G_temp_sup-G_T_transv	15	LH_G_and_S_cingul-Ant	6	LH_G_rectus	2
LH_G_front_sup	15	RH_G_and_S_cingul-Mid-Ant	5	LH_G_pariet_inf-Supramar	2
LH_Lat_Fis-ant-Horizont	14	LH_S_suborbital	5	LH_G_cuneus	2
LH_G_and_S_cingul-Mid-Ant	14	LH_S_oc-temp_lat	5	LH_G_and_S_subcentral	2
RH_S_temporal_inf	13	LH_Lat_Fis-post	5	LH_G_and_S_cingul-Ant	2
RH_G_temporal_middle	13	LH_G_temporal_inf	5	RH_S_temporal_transverse	1
LH_S_temporal_sup	13	LH_G_postcentral	5	RH_S_suborbital	1
LH_G_orbital	13	LH_G_oc-temp_lat-fusifor	5	RH_S_occipital_ant	1
RH_S_front_middle	12	LH_G_cuneus	5	RH_S_oc_middle_and_Lunatus	1
RH_G_temporal_inf	12	LH_G_and_S_subcentral	5	RH_S_oc-temp_med_and_Lingual	1
RH_G_Ins_lg_and_S_cent_ins	12	LH_G_and_S_frontomargin	5	RH_Pole_temporal	1
LH_G_temp_sup-Lateral	12	RH_S_suborbital	4	RH_G_temporal_middle	1
LH_G_rectus	12	RH_G_temp_sup-Plan_polar	4	RH_G_temporal_inf	1
LH_G_and_S_frontomargin	12	RH_G_temp_sup-Lateral	4	RH_G_temp_sup-Lateral	1
LH_G_and_S_cingul-Mid-Post	12	RH_G_front_sup	4	RH_G_temp_sup-G_T_transv	1
RH_G_orbital	11	RH_G_and_S_cingul-Ant	4	RH_G_precuneus	1
LH_S_oc_middle_and_Lunatus	11	RH_G_Ins_lg_and_S_cent_ins	4	RH_G_oc-temp_lat-fusifor	1
LH_S_front_middle	11	LH_S_temporal_inf	4	RH_G_cuneus	1
LH_G_temporal_inf	11	LH_S_precentral-sup-part	4	RH_G_cingul-Post-dorsal	1
LH_G_oc-temp_med-Lingual	11	LH_Lat_Fis-ant-Vertical	4	RH_G_and_S_cingul-Ant	1
LH_G_insular_short	11	LH_Lat_Fis-ant-Horizont	4	LH_S_temporal_transverse	1
LH_G_and_S_occipital_inf	11	LH_G_occipital_sup	4	LH_S_precentral-inf-part	1
LH_G_Ins_lg_and_S_cent_ins	11	LH_G_front_inf-Orbital	4	LH_S_postcentral	1
RH_G_rectus	10	LH_G_front_inf-Opercular	4	LH_S_orbital_lateral	1
RH_G_front_sup	10	RH_S_temporal_sup	3	LH_S_oc_sup_and_transversal	1
RH_G_and_S_cingul-Mid-Ant	10	RH_S_occipital_ant	3	LH_S_oc_middle_and_Lunatus	1
LH_S_temporal_inf	10	RH_S_oc_middle_and_Lunatus	3	LH_S_intrapariet_and_P_trans	1
LH_S_precentral-sup-part	10	RH_S_central	3	LH_S_front_middle	1
LH_G_temporal_middle	10	RH_Pole_temporal	3	LH_S_collat_transv_post	1

Table S3: Ordered list of cortical regions ranked by their number of significant saliency correlations with other brain regions passing each of 6 correlation coefficient thresholds $\boldsymbol{\tau} = 0.1, 0.2, \ldots, 0.6$.

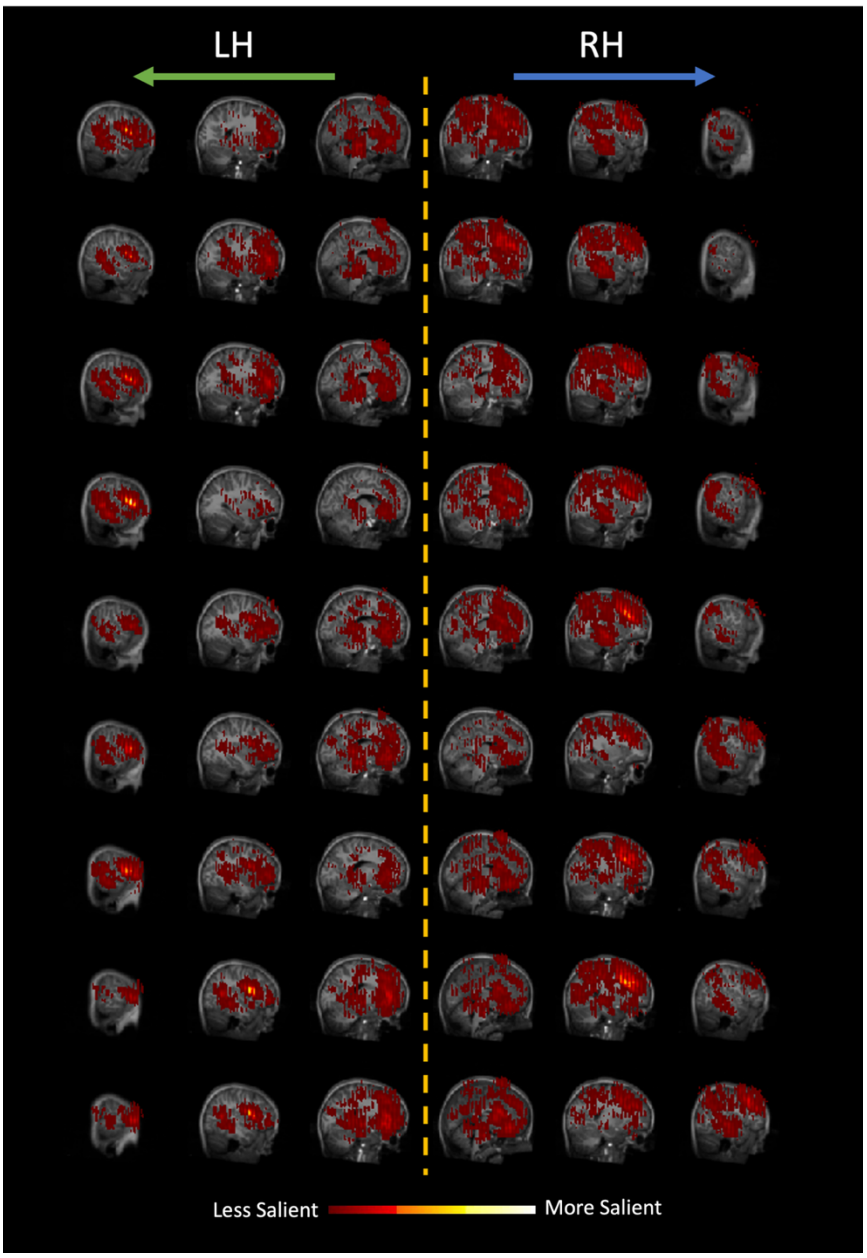

Figure S1: Saliency over sagittal sections for subject [002] visualized through FSLeyes. Read bottom to top, left to right (left hemisphere to right hemisphere).

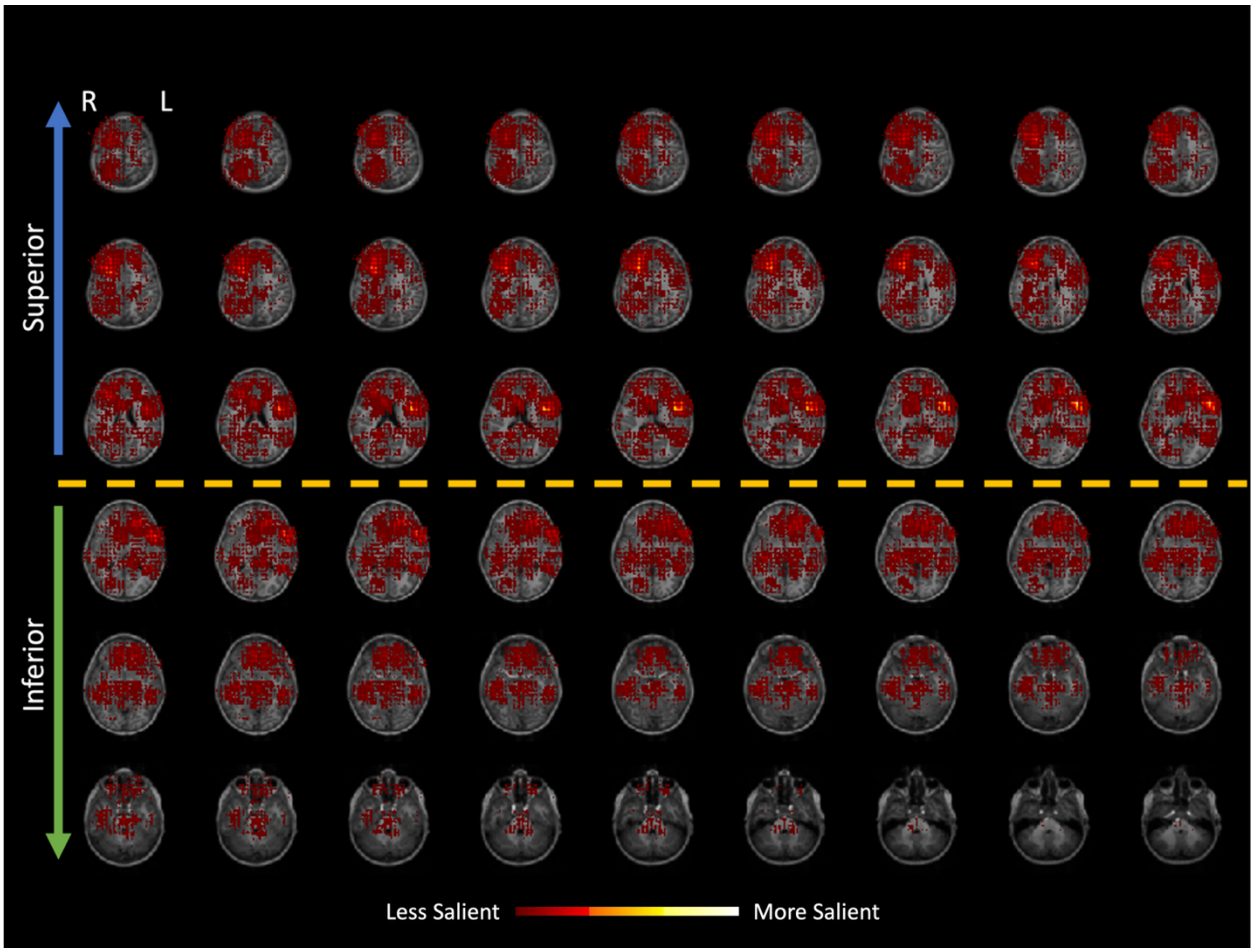

Figure S2: Saliency over axial sections for subject [002] visualized through FSLeyes. Read left to right, top to bottom (superior to inferior).

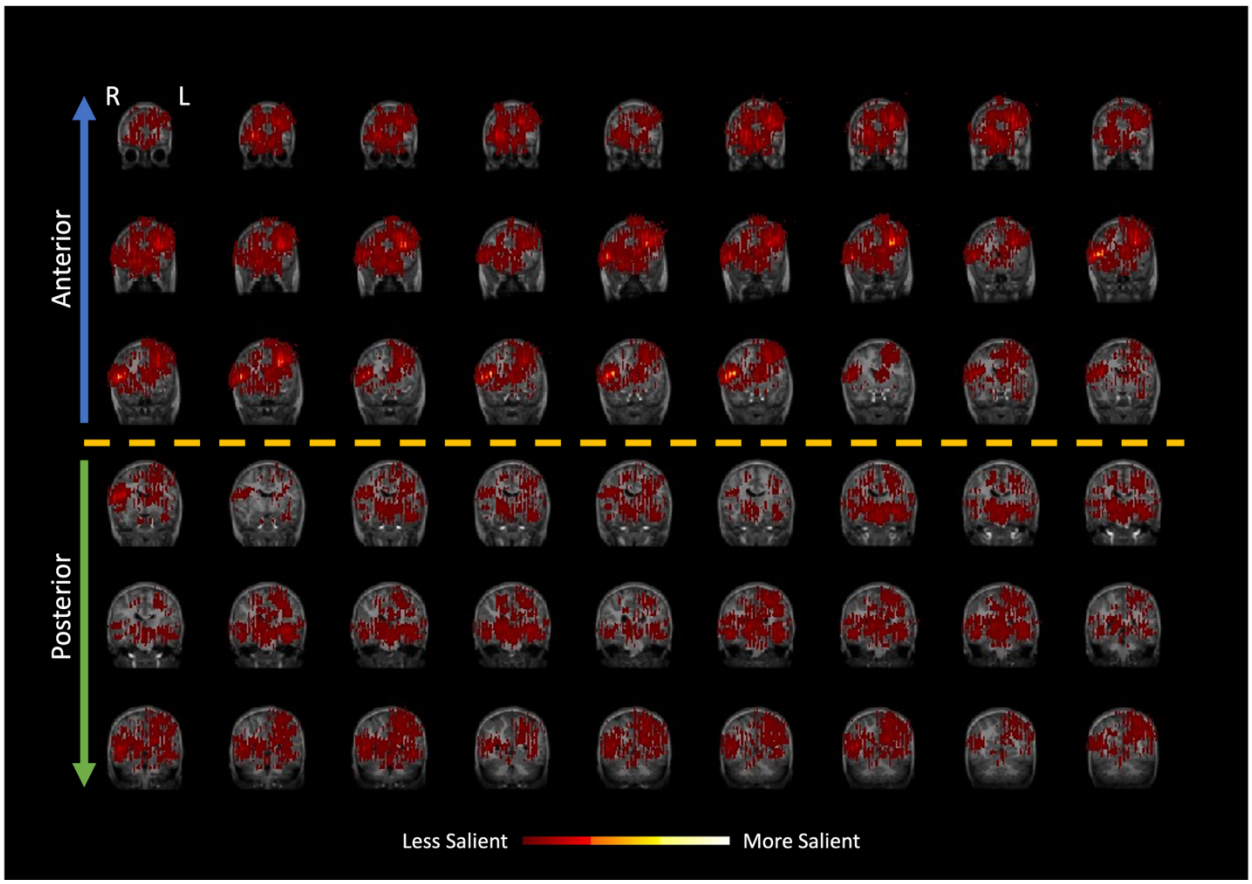

Figure S3: Saliency over coronal sections for subject [002] visualized through FSLeyes. Read left to right, top to bottom (anterior to posterior).

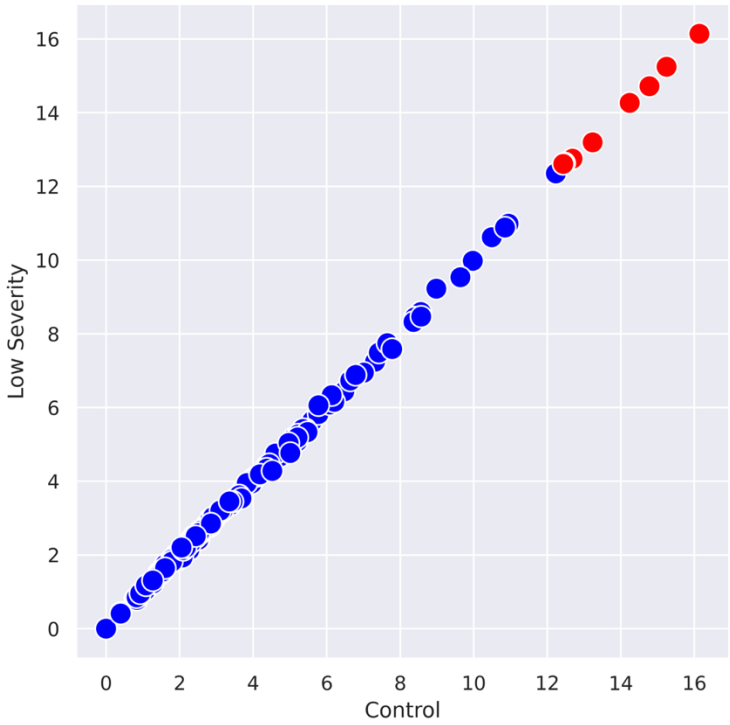

Figure S4: Scatterplot depicting average saliency values of all regions in the Destrieux atlas of the low severity subjects and control subjects. The control subjects were predicted as low severity subjects 86.47% of the time on the 3D ResNet50 trained on low/high severity data. Linear relationship suggests the control group and low severity share similar defining characteristics. Red dots signify regions in the 95% percentile of highest saliency in the control and low severity subjects.

.

Materials and Methods

Data Acquisition

We used sMRI sequences from the Autism Brain Imaging Data Exchange II (ABIDE-II), which is a collection of 1114 subjects (521 diagnosed with ASD and 593 control) gathered from 19 institutions. ABIDE-II consists of male and female subjects aging from 5 years old to 64 years old. Compared to its previous version ABIDE I, ABIDE-II has a greater phenotypic characterization through including additional measures such as different ASD assessment and phenotypes related to ASD research (Di Martino et al. 2017).

We only considered sequences from subjects that were assessed with the Social Response Scale 2 (SRS-2), which is an assessment for quantifying the severity of five ASD related phenotypes (Awareness, Cognition, Communication, Motivation, and Mannerisms). This assessment can be administered to subjects ranging from ages 2 years and 5 months to 18 years old and by gender. By quantifying the severity of these phenotypes, it is possible to compare the severity of ASD in different subjects. Each phenotype is scored by summing a questionnaire with a 4-point Likert-

scale, constituting the raw score for each respective phenotype. The SRS-2 Total T-score is generated by scaling the raw score for each phenotype and combining them together (Bruni 2014a).

We considered subjects' data from the KKI, NYU-1, NYU-2, OHSU, SDSU, and TCD sites (only subjects in these sites were assessed with SRS Total T-Score) from the ABIDE-II dataset, and subsequently uploaded to the Boston University Shared Computing Cluster (BU SCC). The total number of the subjects with SRS-T Total Scores we accessed from the sites totaled to 420 subjects. Since ASD is more prevalent in the male population than in the female population with a ratio of 3:1 (Loomes, Hull, and Mandy 2017), only male subjects were used in this study. Additionally, subjects with the age 12 years or younger were considered to be in the age range of 6-12 years old. For AI supervised training subjects were labeled based on their SRS Total T-Score; low severity subjects were labeled with an SRS Total T-Score of 45 or less and high severity subjects were labeled with an SRS Total T-Score of 70 or higher (Bruni 2014b). These score criteria yielded 149 subjects for our AI training and analyses (Fig. S5).

A control group was also established with control labeled subjects. These subjects were not associated with SRS Total scores. 74 control cases were preprocessed and predicted by the neural network. The network was able to distinguish the control subjects as low severity subjects with 86.47% accuracy.

Figure S5: SRS-T Total Score and Age (in years) of each subject in our dataset. Scatter of data points demonstrates a relatively uniform distribution within the age range and therefore doesn't confound the neural network's ability to predict the severity.

Preprocessing Methods

We preprocessed MRI data to reduce within and across subject groups (low vs. high severity) variabilities unrelated to pathology, in order to minimize confounding neural network performance for the classification of ASD severity. We utilized FreeSurfer (Fischl 2012) and FSL (Smith et al. 2004) to preprocess the MRI volumes. We used Recon-all, a preprocessing pipeline that performs the Freesurfer cortical reconstruction process. These preprocessing steps included:

Motion Correction

Human motion that may have been recorded during scanning procedures was corrected for by constructing the mean of multiple scans. Additionally, subject volumes were transformed from their unique native space to anatomical space (1 mm voxels in a 256⨯256⨯256 space) (Collins et al. 1994).

Intensity Normalization

Four iterations of non-parametric non-uniform intensity normalization was performed to reduce fluctuations in intensity through the nu_correct algorithm in recon-all. We used the mri_normalize algorithm to normalize the intensity value so that the white matter peak normalized to a value of 110 (Dale, Fischl, and Sereno 1999).

Talairach Transformation

We used MINC program mritotal to calculate affine transformation for the original volume to be aligned with the MNI305 atlas. The subject MRI volume was aligned to the MNI305 template (at 1 mm resolution) in anatomical space. This allowed similar regions in each subject brain to share similar voxel regions within the template (Collins et al. 1994).

Smoothing

A multidimensional (3D) gaussian filter was applied to smooth MRI volumes through applying the standard deviation of the gaussian kernel σ, which can be computed from the width at half maximum *FWHM* with this equation:

$$FWHM = \sqrt{8ln2}\,\sigma,$$

which can be rearranged to solve for σ:

$$\sigma = \frac{FWHM}{\sqrt{8\ln 2}}.$$

A *FWHM* value of 3 times the voxel size was suggested (Mikl et al. 2008), so a *FWHM* value of 3 ($\sigma = 1.2739$) was used in this study as the MRIs have voxel size of 1 mm, and smoothing was performed before resizing the MRI volumes to smaller dimensions.

Neural Network Architecture

3D ResNet-50

A ResNet-50 was used for the base model as the ResNet family of neural networks are known for their depth and ability to be easily optimized (He et al. 2015). We expanded the 2D convolutional, max pooling, and average pooling layers in the original 2D ResNet-50 to 3D to develop a 3D ResNet-50. The input layer of the 3D ResNet-50 was modified to pass a 128x128x128 grayscale MRI volume and the identity and convolution blocks of the ResNet-50 remained unchanged except for the 2D layers being replaced with 3D layers. A simple fully connected layer with 2 nodes (one for each class, i.e., low SRS-T Total Score and high SRS-T Total Score), softmax activation, and he_normal weight initializer served as the classification layer of the neural network.

kNN

After training the 3D ResNet-50, the softmax classification layer of the 3D ResNet-50 was replaced with a fully connected (dense) layer with *n* nodes and leaky-relu activation. This allowed for an output of length *n* single dimensional feature vector for the kNN algorithm. The 3D ResNet-50 served as a feature extractor of the MRI volumes to be fed into the kNN algorithm (Fig. 1A). Both the initial training and testing datasets of MRI subjects were passed through the modified 3D ResNet-50 to create respective training and testing dataset of feature vectors for the kNN. The training feature vectors were fitted to a kNN algorithm (Fig. 1A) with uniform weights (all points were weighed equally) and *k* neighbors. The kNN algorithm predicted each feature vector in the testing dataset for all values of *k* (1 to the amount of total training subjects available) to predict the severity of ASD in the testing subject. The *k* value that yielded the highest accuracy was chosen.

SVM

We also similarly utilized SVM as kNN where the feature vectors were fed to a SVM rather than a kNN algorithm. The SVM classified each feature vector in the testing dataset by a n-dimensional hyperplane fitted on the training data.

ML algorithms including kNN and Support Vector Machines (SVM) predict new data based on patterns observed from the training data. The kNN algorithm compares new data points to the closest *k* training points on a n-dimensional plane to base its prediction (Cover and Hart 1967) whereas the SVM algorithm creates a n-dimensional hyperplane based on the training points to base its predictions (Cortes and Vapnik 1995).

2D ResNet-50

MRIs were divided into groups of 8 slices deviating from the center slice. This process was applied through the sagittal, axial, and coronal axes, and then the 2D ResNet-50 model was trained by each group at a time.

Training

The 3D ResNet-50 + kNN hybrid was trained on the BU SCC with Adams optimizer and a learning rate of 3e-3. Early stopping was used to prevent the model from overfitting by stopping the 3D ResNet-50 training after 10 epochs of the validation loss not improving. The model was reverted to the weights with the lowest loss. The model was cross validated with 10% of the training dataset and was tested with the never-before-seen testing dataset after training had completed. The 3D ResNet-50 + kNN hybrid was compared to a 3D ResNet-50 (softmax loss, hinge loss, tanh), 3D ResNet-50 + SVM hybrid, 3D ResNet-50 + PCA + SVM hybrid, 2D ResNet on MRI slices (sagittal, axial, and coronal), and a 2D ResNet ensemble.

Saliency Maps

To visualize key areas contributing in the CNN prediction of MRI volumes, we extracted saliency maps from the last fully connected (dense) layer of the 3D ResNet-50. These maps highlight the importance (w) of each voxel in the MRI volume V for the 3D ResNet-50 to make an accurate prediction, highlighting relative importance of each voxel to nearby voxels through calculating a first-order Taylor expansion:

$$S_C(V) \approx w^T V + b$$

The w of each voxel was calculated by changes in class score S_C for V, generating a saliency map for V, and b is the bias (Simonyan, Vedaldi, and Zisserman 2013).

Training 3D CNNs or 2D CNNs by MRI data each have their own drawbacks. Training fMRI slices rather than through a volume hinders the networks ability to extract features across all axes. On the other hand, training 3D CNN by MRI data can result in noisy feature extraction as 3D convolutional filters have an increased variability. The saliency maps were globally normalized and scaled to a 0-255 range. Four saliency maps were overlayed on their respective FreeSurfer pial surface, to visualize the saliency spread across the cortex (Fig. 2).

<u>Average Saliency and Analysis</u>

Masks of 75 regions (per hemisphere) in the Destriuex atlas and 28 total regions in the FreeSurfer subcortical atlas were generated through FreeSurfer's mri_binarize and mri_mask commands. These masks were aligned to the saliency maps, and the average saliency for each region was calculated by dividing the sum of the regional saliency by the area of the mask. After average saliencies for all regions in all subjects were obtained, the average saliency per region and severity (low v. high) was obtained. Average left hemisphere saliencies in low severity subjects, average left hemisphere saliencies in high severity subjects, average right hemisphere saliencies in low severity subjects, and average right hemisphere saliencies in high severity subjects were compared to determine cortical and network patterns.

Additionally, rank-based Kendall Tau correlation was conducted with primary focus cross-cortical correlations as well as cortical and subcortical correlations. This allowed us to find potential

regions that displayed saliencies at a similar frequency. Correlations were visualized using the circular_layout function in the MNE-Python library, and the regions were organized by lobe. Correlations were filtered by alpha value of 0.01, and strength of the correlations was determined by the τ values.

Cortical regions with saliencies that were significantly correlated with SRS subscores were obtained by conducting rank-based correlation between each cortical region saliency and the five subscores. Results were filtered to only include results with a *p*-value less that 0.05.

Furthermore, we graphed scatterplots of the average and standard deviation regional saliency of the left hemisphere low severity subjects, left hemisphere high severity subjects, right hemisphere low severity subjects, and right hemisphere high severity subjects to compare the consistency of saliency across the different groups. Each point in the scatterplot represents a region and is organized by lobe to determine any lobular biases the saliency might have (Fig. 8).