Archival Report

Biological Psychiatry

Robust, Generalizable, and Interpretable Artificial Intelligence–Derived Brain Fingerprints of Autism and Social Communication Symptom Severity

Kaustubh Supekar, Srikanth Ryali, Rui Yuan, Devinder Kumar, Carlo de los Angeles, and Vinod Menon

ABSTRACT

BACKGROUND: Autism spectrum disorder (ASD) is among the most pervasive neurodevelopmental disorders, yet the neurobiology of ASD is still poorly understood because inconsistent findings from underpowered individual studies preclude the identification of robust and interpretable neurobiological markers and predictors of clinical symptoms.

METHODS: We leverage multiple brain imaging cohorts and exciting recent advances in explainable artificial intelligence to develop a novel spatiotemporal deep neural network (stDNN) model, which identifies robust and interpretable dynamic brain markers that distinguish ASD from neurotypical control subjects and predict clinical symptom severity.

RESULTS: stDNN achieved consistently high classification accuracies in cross-validation analysis of data from the multisite ABIDE (Autism Brain Imaging Data Exchange) cohort (n = 834). Crucially, stDNN also accurately classified data from independent Stanford (n = 202) and GENDAAR (Gender Exploration of Neurogenetics and Development to Advanced Autism Research) (n = 90) cohorts without additional training. stDNN could not distinguish attention-deficit/hyperactivity disorder from neurotypical control subjects, highlighting the model's specificity. Explainable artificial intelligence revealed that brain features associated with the posterior cingulate cortex and precuneus, dorsolateral and ventrolateral prefrontal cortex, and superior temporal sulcus, which anchor the default mode network, cognitive control, and human voice processing systems, respectively, most clearly distinguished ASD from neurotypical control subjects in the three cohorts. Furthermore, features associated with the posterior cingulate cortex and precuneus nodes of the default mode network emerged as robust predictors of the severity of core social and communication deficits but not restricted/repetitive behaviors in ASD.

CONCLUSIONS: Our findings, replicated across independent cohorts, reveal robust individualized functional brain fingerprints of ASD psychopathology, which could lead to more objective and precise phenotypic characterization and targeted treatments.

https://doi.org/10.1016/j.biopsych.2022.02.005

Autism spectrum disorder (ASD) is one of the most common neurodevelopmental disorders, characterized by a range of highly debilitating social impairments and communication difficulties (1–3). Despite decades of research, the neurobiology of ASD is still poorly understood. The current paradigm for investigating the neurobiology of ASD has reached a crossroads: inconsistent findings from individual studies fail to address the heterogeneity in ASD, precluding the identification of reliable and interpretable neurobiological markers and predictors of clinical symptoms, which can be used as more accurate diagnostic measures and precise treatment targets (4). Newly available "big data" and powerful explainable artificial intelligence (XAI) techniques offer an unprecedented opportunity to identify individualized and robust markers of ASD (5). Here, we use an end-to-end data-driven XAI-based computational framework for robust identification of the neurobiological markers of ASD by leveraging novel deep neural networks (DNNs) and large-scale open-source brain imaging data acquired from multiple sites, as well as two other independent cohorts.

Over the past few years, DNNs have revolutionized the field of XAI with major successes in applications such as computer vision and natural language processing (6). DNNs, however, have had limited success in ab initio classification and identification of neurobiological features that distinguish psychiatric disorders using functional brain imaging data. This is due to several challenges in applying DNNs to brain imaging data, most notably dealing with the high dimensionality of whole-brain data and noisy measurements with a large degree of individual variability across data acquisition sites (7). A particular challenge here is the application to ASD, a psychiatric disorder characterized by a spectrum of

ISSN: 0006-3223

impairments and high levels of heterogeneity in phenotypic clinical symptoms (4).

A few recent studies have attempted to use DNNs by reducing the dimensionality of brain data through explicit feature engineering (8–16). Typically, in these approaches, precomputed static functional brain connectivities are provided as input to DNN models consisting of multiple fully connected networks followed by a sigmoid layer for classification. This approach has two problems. First, the correlation coefficients, which are used as static functional connectivity features, assume that the functional magnetic resonance imaging (fMRI) time series reflect a stationary process. However, recent studies have shown that fMRI time series are highly nonstationary (17–19). Second, training DNN architectures with fully connected layers is challenging, particularly in neuroimaging applications, because of a large number of free parameters and a limited number of labeled training data. As a result, these architectures tend to overfit the data and exhibit poor out-of-sample prediction (20). Critically, extant approaches do not exploit the dynamic spatiotemporal characteristics of brain activity, which contain more robust features of psychiatric disorders and are thought to inform on interindividual phenotypic differences (21–26).

To address these issues, we developed a novel spatiotemporal DNN (stDNN) model, which directly takes as its input fMRI time-series data from brain regions of interest and models underlying dynamic spatiotemporal characteristics of brain activity (dynamic connectomes) to distinguish between ASD and neurotypical control subjects (Figures 1 and 2). A key idea of our approach is to discover latent spatiotemporal dynamics for classification from brain data without the need for explicit feature engineering. Our stDNN consists of convolutional layers that encode latent information at different temporal resolutions and, unlike fully connected networks, has a comparatively smaller number of parameters to train (Figure 2).

Our study addresses four key challenges. The first challenge we address here is that fMRI data from data sharing consortia are highly heterogenous, due to differences in scanning protocols and MRI scanners used in each site. Thus, the methodology must be robust to, and be able to generalize across, various data acquisition protocols that are not harmonized across sites. We address this challenge by developing an stDNN model with one-hot encoding that incorporates multisite heterogeneity to create a single network that handles heterogeneous data from different imaging protocols while learning robust representations for simultaneously classifying and identifying robust neurobiologically meaningful features that distinguish individuals with ASD. We applied stDNN to resting-state fMRI data from over 800 participants and multiple data acquisition sites that were shared through the Autism Brain Imaging Data Exchange (ABIDE) (27,28).

The second challenge is that the DNN model and methods should be able to accurately classify individuals from independent untrained data. While several previous studies have examined ASD classification accuracy in individual cohorts, to our knowledge, no studies have then gone on to show that the classifiers are generalizable and effective for novel data that the classifier had not seen before (Table S1). This is a crucial step in which most algorithms in other domains are widely known to fail (7). We address this challenge by training our classifier on the ABIDE dataset and testing it on two other independent datasets, one acquired at Stanford and the other by the GENDAAR (Gender Exploration of Neurogenetics and Development to Advanced Autism Research) consortium, which were not used in training of the stDNN model.

The third challenge we address here is that previous studies using DNNs have almost exclusively focused on classification

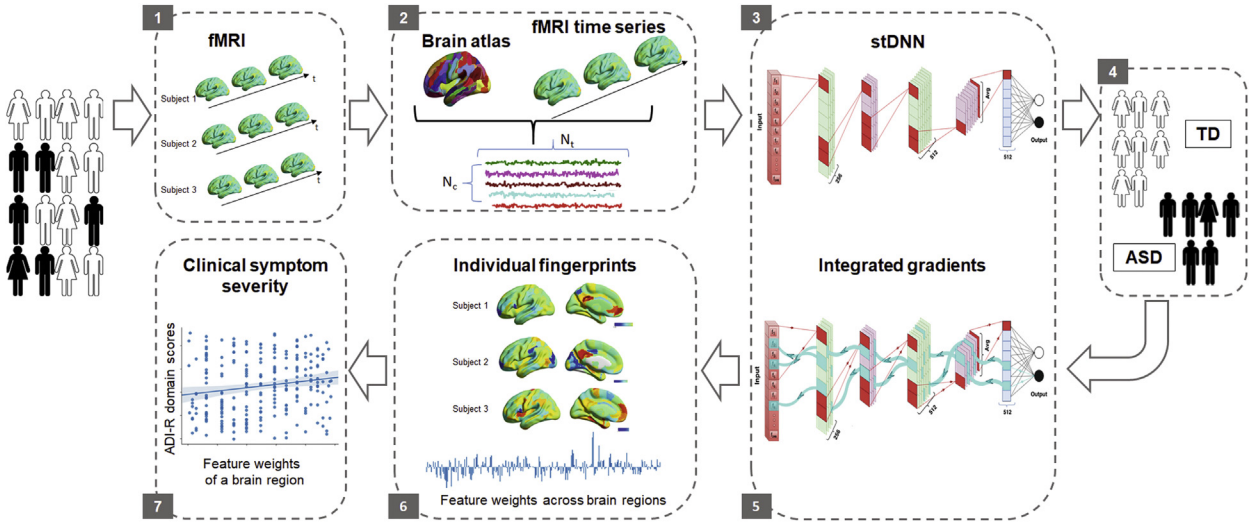

Figure 1. Schematic overview of our multicomponent explainable artificial intelligence framework for discovering individualized neurobiological fingerprints that predict autism spectrum disorder (ASD) psychopathology and severity of clinical symptoms. Key steps include steps 1 and 2, data extraction; steps 3 and 4, classification; steps 5 and 6, feature identification, i.e., predictive feature weights (fingerprints) across brain regions; and step 7, prediction of clinical symptom severity. ADI-R, Autism Diagnostic Interview-Revised; fMRI, functional magnetic resonance imaging; stDNN, spatiotemporal deep neural network; TD, typically developing.

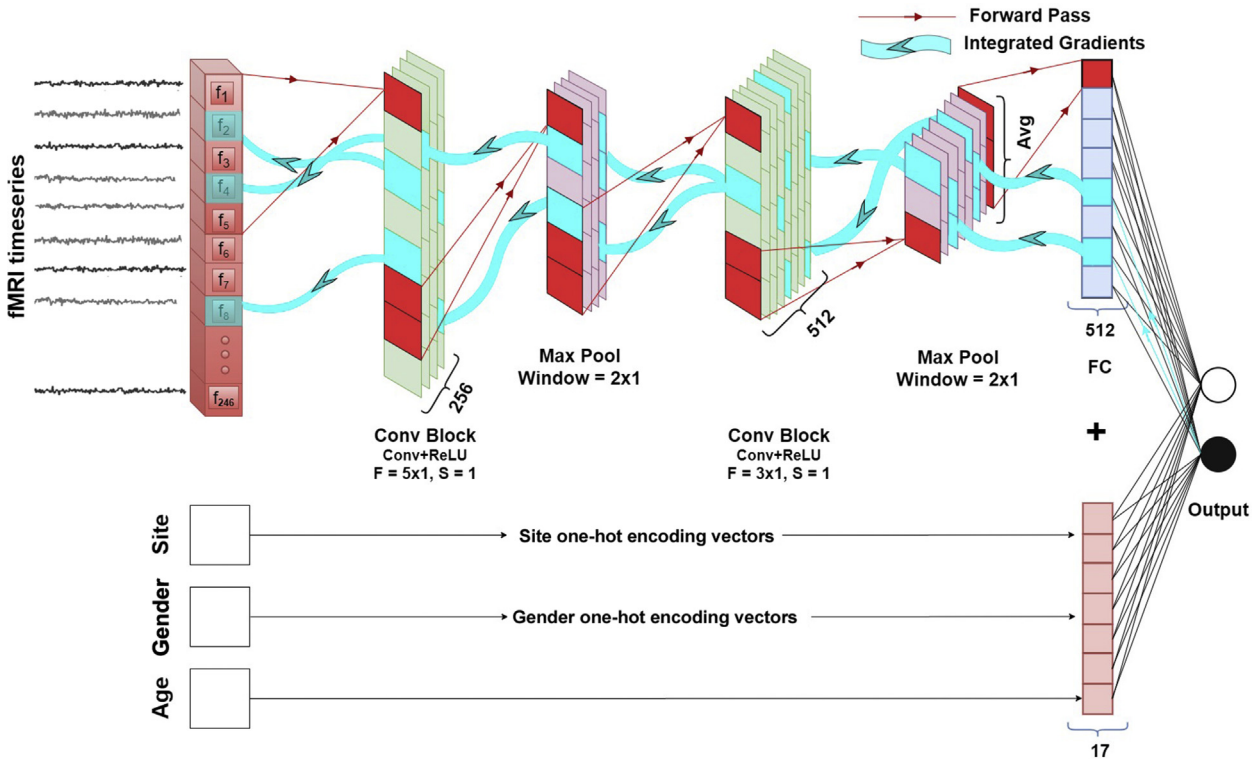

Figure 2. For each subject, the spatiotemporal deep neural network model uses regional functional magnetic resonance imaging (fMRI) time series and covariates, including site, gender, and age, as input. The model predicts the class label (autism spectrum disorder or typically developing) of the subject using spatiotemporal convolutions of the fMRI time series along with the covariate information. The integrated gradient of the model's prediction to its input is used for identifying black box brain features underlying the classification of autism spectrum disorder vs. typically developing control subjects. Avg, average; Conv, convolutional; FC, fully connected; Max, maximum; ReLU, rectified linear unit.

and have not paid adequate attention to identifying the neurobiological features that underlie the classification (Table S1). Therefore, little is currently known about brain features that distinguish ASD without the bias of prior feature engineering. We address this major gap in the field by developing an integrated and automated approach for simultaneously classifying and identifying neurobiologically meaningful features that distinguish individuals with ASD from neurotypical control subjects.

The final challenge we address here is the identification of neurobiologically meaningful features that predict the severity of clinical symptoms associated with the heterogenous manifestations of ASD; to our knowledge, no previous deep learning classification study has investigated brain features that robustly predict clinical symptoms without feature engineering in independent datasets (Table S1).

We demonstrate that our stDNN model 1) accurately distinguishes ASD by modeling spatiotemporal brain dynamics without any predefined assumptions, 2) generalizes to novel data on which it was not trained, and 3) is specific to ASD. We uncover brain features that most clearly distinguish ASD from neurotypical control subjects and robustly predict the severity of social and communication deficits. Based on our theoretical model and empirical findings (29), we predicted that default mode network (DMN) nodes would feature prominently in both classification and symptom prediction.

METHODS AND MATERIALS

Study Cohorts

ABIDE Cohort. We leveraged neuroimaging and phenotypic data from ABIDE (27,28) (Figure S1 and Table S2; see Supplemental Methods for details).

Stanford Cohort. An independent cohort of participants acquired at Stanford (30–32) was used to investigate the generalizability of stDNN to previously unseen data (Figure S2 and Table S2; see Supplemental Methods for details).

GENDAAR Cohort. An independent cohort of participants acquired by GENDAAR was used to further investigate the generalizability of stDNN to previously unseen data (Figure S3 and Table S2; see Supplemental Methods for details).

ADHD200-NYU Cohort. An independent cohort of participants acquired at New York University was used to investigate the specificity of the stDNN (Table S22; see Supplemental Methods for details).

stDNN Model

We developed an innovative stDNN model (Figure 2) to extract informative brain dynamics features that accurately distinguish

between ASD and neurotypical control subjects. A key advantage of our approach is that it provides a novel technique to capture latent dynamics without the need for explicit feature engineering (33). Our stDNN model consists of two one-dimensional convolutional block layers, a temporal averaging operation, a dropout layer, and a sigmoid layer (Figure 2; see Supplemental Methods for details). Preprocessed regional fMRI time series from 246 brain regions defined in the Brainnetome atlas (34) were given as input to the first one-dimensional convolutional block layer (see Supplemental Methods for details). To account for site-, gender-, and age-related heterogeneity, site and gender information encoded with a one-hot encoding scheme, which is the most commonly used method for encoding categorial variables, and age were given as an input to the final fully connected layer. stDNN classified participants in the two groups by minimizing the binary cross-entropy cost function.

Fivefold Cross-validation Classification Analysis of ABIDE Cohort Data

To prevent bias and account for low variance, we conducted a fivefold cross-validation to evaluate the performance (accuracy, precision, recall, F1) of our stDNN model (Figure 3; see Supplemental Methods for details).

Classification Analysis of Stanford, GENDAAR, and ADHD-NYU Cohort Data Using Fivefold ABIDE Cohort Models

Similar to the fivefold cross-validation process used for ABIDE, for reporting the performance of our stDNN for Stanford, GENDAAR, and ADHD-NYU cohorts, we used each of the five stDNN models trained on different subsets of ABIDE. Using the five different models, we evaluate each model's performance on the cohort data independently (Figure 3; see Supplemental Methods for details) and report the mean and standard deviation values of the key performance metrics for each of the three cohorts.

Identifying Brain Features Underlying Classification

We used an integrated gradients (IG)–based feature attribution approach (35–39) (see Supplemental Methods for details) to identify brain features that discriminated between the ASD and typically developing (TD) groups.

Fingerprint Analysis of Brain Features Underlying ASD Classification

We examined whether IG-based individual fingerprints of predictive brain features cluster differently in individuals with ASD than in TD control subjects. The Pearson correlation between individuals' IG-derived brain feature maps was used to calculate the distance between them, and these distances were compared between individuals in the ASD and TD groups (see Supplemental Methods for details).

Clinical Symptom Prediction in ASD

We investigated the relationship between stDNN-identified neurobiological features with the severity of clinical symptoms in individuals with ASD. Spearman correlations between the Autism Diagnostic Interview-Revised (ADI-R) domain scores and the brain features derived from each of the five

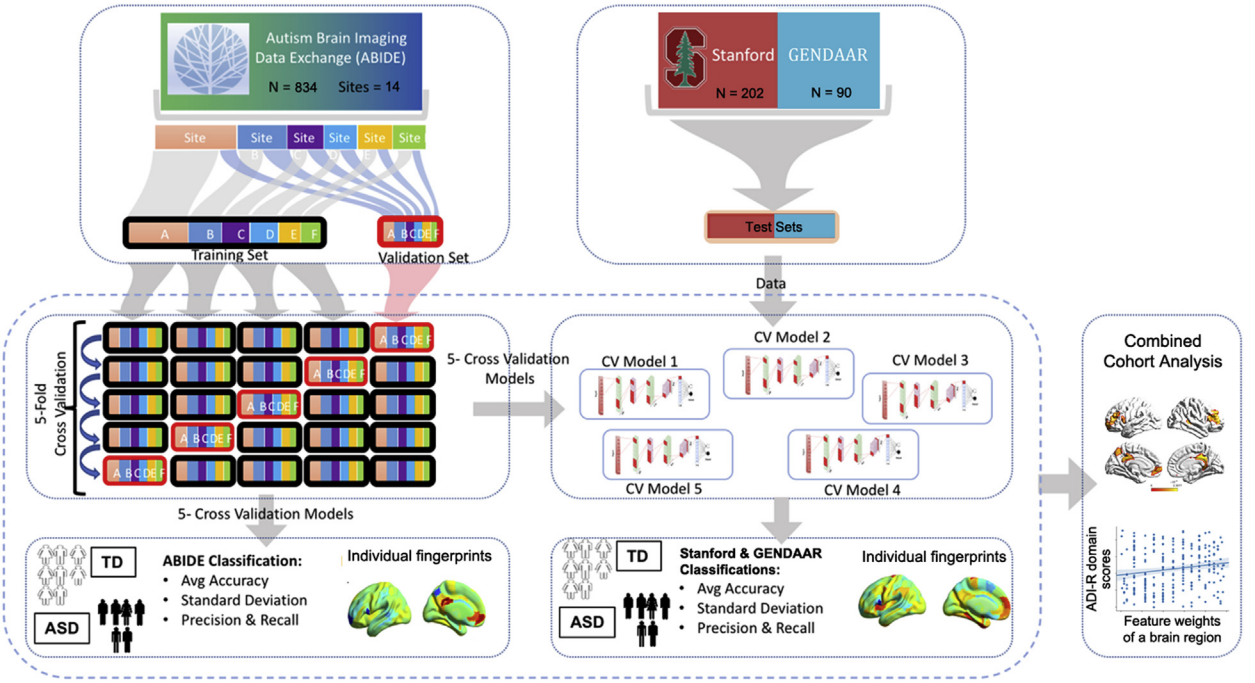

Figure 3. Fivefold cross-validation (CV) procedures for testing and validation of autism spectrum disorder (ASD) vs. typically developing (TD) classification in the ABIDE (Autism Brain Imaging Data Exchange) cohort. The five models are then used for independently testing ASD vs. TD classification in the Stanford cohort and the GENDAAR (Gender Exploration of Neurogenetics and Development to Advanced Autism Research) cohort. Note that these models are not trained on the Stanford cohort and the GENDAAR cohort. ADI-R, Autism Diagnostic Interview-Revised; Avg, average.

stDNN models were computed (see Supplemental Methods for details).

Control Analyses

We performed several control analyses to demonstrate that our findings are robust to brain atlas selection and head motion–related confounds and to show that stDNN outperforms extant classification algorithms and approaches (see Supplemental Methods for details).

RESULTS

Classification of ASD Versus Control Subjects in the ABIDE Cohort

We first trained our stDNN on the multisite ABIDE cohort (n = 834) of individuals with ASD and TD individuals as control subjects (Table S2). To assess the performance of our stDNN model, we used a fivefold cross-validation procedure in which 80% of the sample was used for training while the other 20% of the sample was used for validation (Figure 3). stDNN achieved an average accuracy of 78.2 ± 2.84% across the five folds and an average precision of 0.76 ± 0.03, recall of 0.82 ± 0.03, and F1 score of 0.79 ± 0.03 (Table S3). stDNN achieved comparable accuracies using other commonly used brain atlases instead of the Brainnetome atlas (Table S4), as well as across sites and genders (see Supplemental Results for details). Notably, stDNN outperformed conventional approaches that used 1) static functional connectivity, 2) amplitude of low-frequency fluctuation, 3) blood oxygen level–dependent signal variability, 4) sliding-window functional connectivity variability, and 5) fMRI time series as features, which achieved average accuracies ranging from 50% to 63% (Tables S5–S9; see Supplemental Results for details), further highlighting the strength of our approach. Additional analyses confirmed that the observed results were robust to potential confounds such as head motion (see Supplemental Results for details). These results demonstrate that stDNN can accurately distinguish individuals with ASD from TD control subjects in a multisite cohort and, furthermore, does so in a robust and consistent manner across cross-validation folds.

Classification of ASD Versus Control Subjects in Independent Stanford and GENDAAR Cohorts

We then evaluated the performance of our stDNN model on two independent cohorts of individuals with ASD and TD control subjects, one acquired at Stanford (n = 202) and the other from the GENDAAR consortium (n = 90) (Table S2). stDNN was not trained on the Stanford and GENDAAR cohort data. We evaluated five models corresponding to each of the folds in the cross-validation as described above (Figure 3). In the Stanford cohort, stDNN achieved an average accuracy of 72.6 ± 2.11% across the five folds and an average precision of 0.70 ± 0.03, recall of 0.81 ± 0.06, and F1 score of 0.75 ± 0.02 (Table S10). In the GENDAAR cohort, stDNN achieved an average accuracy of 75.11 ± 2.86% across the five folds and an average precision of 0.76 ± 0.03, recall of 0.80 ± 0.03, and F1 score of 0.78 ± 0.02 (Table S16). Notably, in both cohorts, stDNN outperformed conventional approaches, which achieved average accuracies ranging from 55% to 63% in the Stanford cohort (Tables S11–S15; see Supplemental Results for details) and from 49% to 60% in the GENDAAR cohort (Tables S17–S21; see Supplemental Results for details), further highlighting the strength of our approach. These results demonstrate that stDNN can accurately classify individuals with ASD from TD control subjects in a robust and consistent manner across cross-validation folds in two independent cohorts without additional training.

Classification of Attention-Deficit/Hyperactivity Disorder Versus Control Subjects

To examine the specificity of our stDNN ASD classification model, we investigated whether the stDNN model trained to distinguish between ASD and TD participants can distinguish between attention-deficit/hyperactivity disorder (ADHD) and TD participants (n = 221) (Table S22). For ADHD versus TD classification, the stDNN model trained on the ABIDE data achieved around chance level accuracy of 56.0 ± 1.84% across the five folds and an average precision of 0.69 ± 0.01, recall of 0.39 ± 0.07, and F1 score of 0.49 ± 0.06 (Table S23), highlighting the specificity of our stDNN model for ASD classification.

Identification of Brain Features Underlying ASD Classification in the ABIDE Cohort

We then used an IG procedure (35–39) to compute the feature attributes underlying the ASD class label in the ABIDE cohort. This analysis yields a measure of feature strength associated with ASD versus TD classification in each brain region and at each time point. This procedure also identifies an individual fingerprint of predictive features in each participant (Figure 4). We first examined whether these fingerprints cluster differently in individuals with ASD than TD control subjects and found that intra-ASD group distance metrics were significantly shorter than distances with the TD group ($p < .0001$). These results demonstrate that individualized brain fingerprints mirrored the broader diagnostic discrimination of ASD.

To identify brain areas that contributed the most to classification, we computed the median of feature scores across the five folds and thresholded them—top 5% of features—based on the feature scores distribution across all time points and regions. This resulted in the identification of a distributed set of brain areas including the posterior cingulate cortex (PCC), precuneus, ventromedial prefrontal cortex, dorsolateral and ventrolateral prefrontal cortex, and superior temporal sulcus, that contributed most significantly to predicting the ASD class label (Figure 5 and Table S24). Additional analyses confirmed that the observed results were robust to potential confounds such as head motion (see Supplemental Results for details). These results demonstrate that stDNN, together with IG procedures, automatically identifies discriminating features without the need for ad hoc feature engineering.

Identification of Brain Features Underlying ASD Classification in Independent Stanford and GENDAAR Cohorts

We then used the same procedures as described in the previous section to determine predictive feature attributes in each ASD participant in the Stanford and GENDAAR cohorts. These

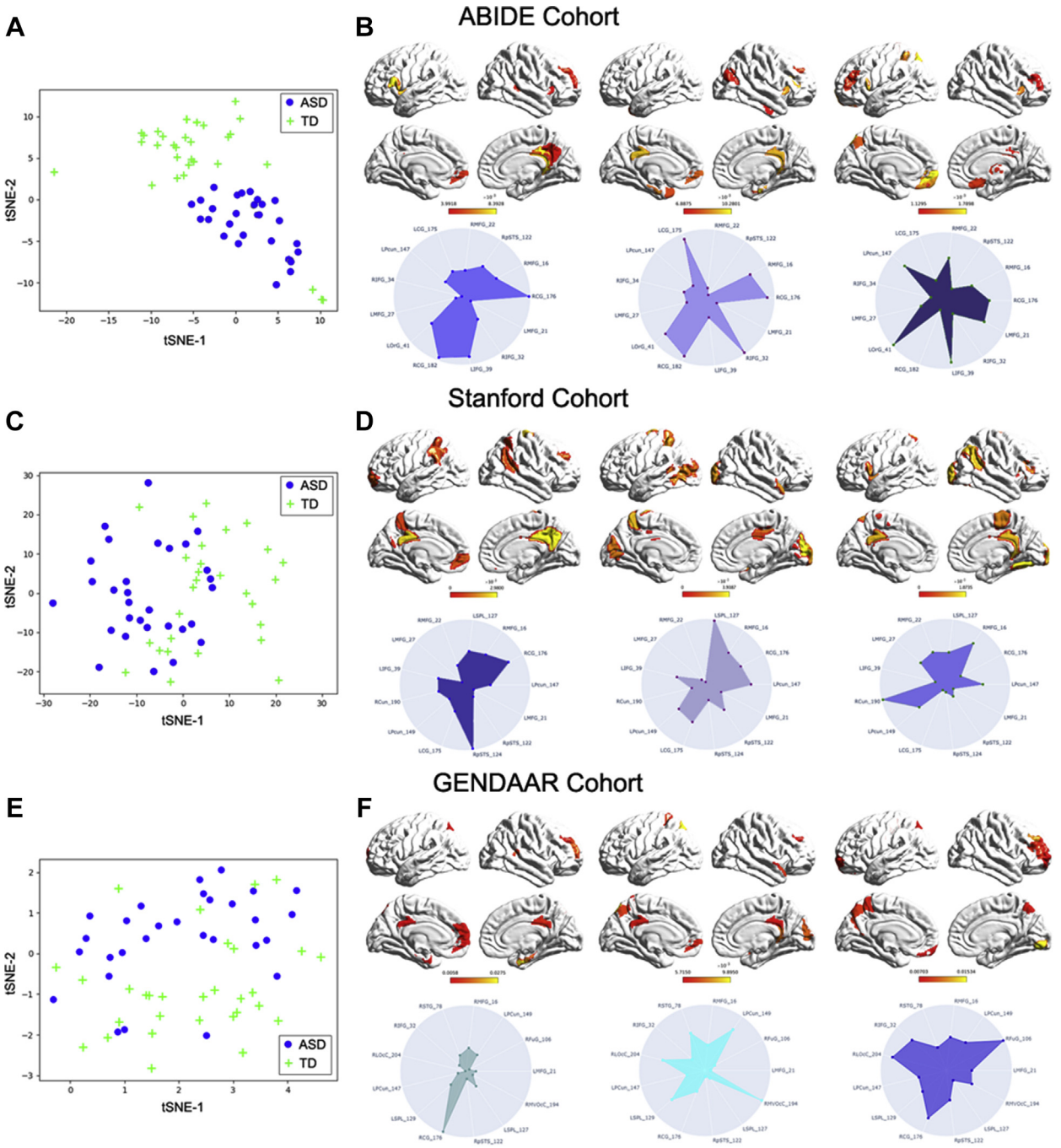

Figure 4. **(A)** t-distributed stochastic neighbor embedding (tSNE) plot of spatiotemporal deep neural network (stDNN)–derived individual feature attribution maps/fingerprints of 60 representative individuals with autism spectrum disorder (ASD) and typically developing (TD) participants randomly selected from the ABIDE (Autism Brain Imaging Data Exchange) cohort, demonstrating clustering of ASD and TD individuals. **(B)** stDNN-derived individual feature attribution maps/fingerprints in three ASD participants from the ABIDE cohort. **(C)** tSNE plot of stDNN-derived individual feature attribution maps/fingerprints of 60 representative individuals with ASD and TD participants randomly selected from the Stanford cohort, demonstrating clustering of ASD and TD individuals. **(D)** stDNN-derived individual feature attribution maps/fingerprints in 3 ASD participants randomly selected from the Stanford cohort. **(E)** tSNE plot of stDNN-derived individual feature attribution maps/fingerprints of 60 representative individuals with ASD and TD participants randomly selected from the GENDAAR (Gender Exploration of Neurogenetics and Development to Advanced Autism Research) cohort, demonstrating clustering of ASD and TD individuals. **(F)** stDNN-derived individual feature attribution maps/fingerprints in 3 ASD participants randomly selected from the GENDAAR cohort.

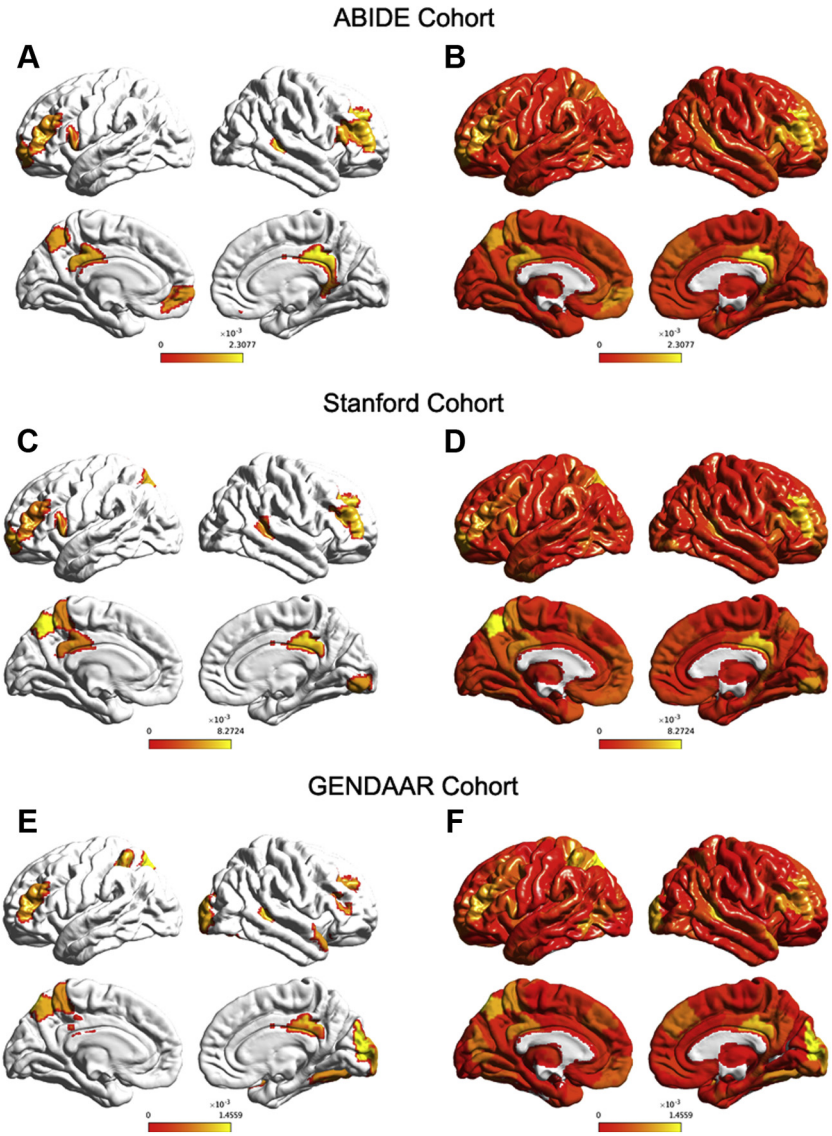

Figure 5. **(A)** Feature attribution map showing the top 5% of features that underlie autism spectrum disorder vs. typically developing classification in the ABIDE (Autism Brain Imaging Data Exchange) cohort. The spatiotemporal deep neural network with integrated gradients identified brain features that distinguish individuals with autism spectrum disorder from typically developing control subjects. The algorithm automatically identified distinguishing features in the posterior cingulate cortex, precuneus, ventromedial prefrontal cortex, dorsolateral and ventrolateral prefrontal cortex, and superior temporal sulcus, which anchor the default mode network, cognitive control, and human voice processing systems, respectively (Table S24 for a detailed listing of brain areas and predictive feature weights). **(B)** Visualization of (unthresholded) feature weights across the whole brain in the ABIDE cohort. **(C)** Feature attribution maps showing the top 5% of features showing replication of the predictive default mode network, dorsolateral and ventrolateral prefrontal cortex, and superior temporal sulcus features in the Stanford cohort (Table S25 for a detailed listing of brain areas and predictive feature weights). **(D)** Visualization of (unthresholded) feature weights across the whole brain in the Stanford cohort. **(E)** Feature attribution maps showing the top 5% of features showing replication of the predictive default mode network, dorsolateral and ventrolateral prefrontal cortex, and superior temporal sulcus features in the GENDAAR (Gender Exploration of Neurogenetics and Development to Advanced Autism Research) cohort (see Table S26 for a detailed listing of brain areas and predictive feature weights). **(F)** Visualization of (unthresholded) feature weights across the whole brain in the GENDAAR cohort.

cohortwise analyses revealed individualized brain fingerprints that mirrored the broader diagnostic discrimination of ASD ($p < .0001$) (Figure 4) and identified the PCC, precuneus, dorsolateral and ventrolateral prefrontal cortex, and superior temporal sulcus as the brain areas that contributed most significantly to predicting the ASD class label (Figure 5, Tables S25–S26) in both cohorts. These results demonstrate that stDNN, together with IG procedures, automatically identifies similar discriminating features as in the ABIDE cohort, again without the need for ad hoc feature engineering.

Predicting Clinical Symptoms Using Brain Features

We investigated whether our stDNN-identified brain features could predict the severity of clinical symptoms in individuals with ASD in the combined cohort (n_ASD = 417). The PCC, precuneus, and ventrolateral prefrontal cortex were the only brain regions whose features predicted ADI-R social scores ($p < .001$; FDR-corrected) in each of the five cross-validation folds. The PCC, precuneus, and dorsolateral prefrontal cortex were the only brain regions whose features predicted ADI-R communication scores ($p < .001$; FDR-corrected) in each of the five cross-validation folds. These features did not predict ADI-R restricted and repetitive behavior scores (all p values $> .30$), demonstrating the specificity of the findings related to core social and communication phenotypic features associated with ASD.

DISCUSSION

We identified dynamic brain features that distinguish individuals with ASD from neurotypical control subjects and predict clinical symptom severity using a novel stDNN. Our

model learned latent dynamic interactions among distributed brain areas without ad hoc feature engineering, achieving high classification accuracies in cross-validation analysis of data from the multisite ABIDE cohort. Crucially, the stDNN model also accurately classified data from two independent cohorts without any additional training. stDNN could not accurately distinguish individuals with ADHD from neurotypical control subjects, highlighting the specificity of the model. Feature identification using an IG approach revealed that brain features associated with the key nodes of the DMN, human voice processing, and cognitive control systems (31,40–43) most clearly distinguished ASD from neurotypical control subjects in the three cohorts, and the posterior DMN nodes predicted core social and communication, but not restricted and repetitive behavior, phenotypic features associated with ASD.

stDNN overcomes the key limitations of commonly used DNN-based methods for classification of brain imaging data (see Supplemental Discussion for details) and addresses several challenges associated with developing robust neurobiological markers of ASD using large-scale open-source brain imaging data.

The first challenge we addressed was to develop a robust classifier that distinguishes individuals with ASD from neurotypical control subjects. stDNN with site harmonization achieved a cross-validation classification accuracy of 78.2 ± 2.84% in the multisite ABIDE cohort, outperforming previously published studies (Table S1). The cross-validation sensitivity and specificity values achieved by our stDNN model are comparable to the gold standard Autism Diagnostic Observation Schedule evaluation for the diagnosis of verbally fluent individuals with ASD (44,45). Furthermore, stDNN significantly outperformed conventional approaches that use static functional connectivity features and regional time series features for ASD classification (Tables S5–S9). Our findings suggest that atypical intrinsic spatiotemporal brain dynamics, captured by our stDNN model, is a more robust feature of ASD than conventional functional brain connectivity and regional brain activity.

The second challenge we addressed was whether the stDNN model trained on the ABIDE data could accurately classify participants from a completely left-out dataset. stDNN achieved a high classification accuracy in both the independent Stanford and GENDAAR cohorts, demonstrating that stDNN can accurately classify individuals with ASD from TD control subjects in a consistent manner in independent cohorts without additional training. Notably, the stDNN model achieved high accuracy in spite of differences in age and severity of ASD symptoms between the cohorts, and the accuracy levels were considerably higher than those obtained using conventional methods (Tables S11–S15 and S17–S21). Another aspect that has not been addressed in previous ASD classification studies is determining the specificity of the classification model, which is crucial for demonstrating the usefulness of such models in clinical settings. The fact that our stDNN model, which was trained to distinguish between ASD and TD groups, was unable to distinguish between ADHD—one of the most common co-occurring disorders with ASD (46)—and TD groups, illustrates the model's specificity.

The third challenge we addressed was to uncover neurobiologically interpretable features associated with ASD. Conventional DNN approaches, especially those applied to time-series data, are black box models that lack interpretability in terms of the underlying neurobiological features (7). A model might achieve high levels of classification accuracy but provide no insight into which features are important for classification or whether the features are neurobiologically interpretable in the context of previous systems neuroscience models of ASD. Here, our deep learning model allowed us to identify and rank features that distinguish ASD from control subjects using an IG approach (Figure 2). Furthermore, our predictive features identify the brain fingerprints, which index the differential contribution of different brain areas to the broader clinical diagnostic discrimination of ASD and is an important feature of our model.

Our stDNN-based feature detection analysis identified the PCC and precuneus, which anchor the DMN, as brain areas whose dynamic properties most clearly distinguished the ASD group from control subjects. Crucially, these features not only were observed in the ABIDE cohort but also were replicated in the independent Stanford and GENDAAR cohorts, further attesting to the robustness and generalizability of our findings. Aberrancies in DMN nodes and their dynamic functional interactions contribute to the atypical integration of information about the self in relation to other, as well as impairments in the ability to flexibly attend to socially relevant stimuli (47–51). Altered structural and functional organization of the DMN and its atypical developmental trajectory are prominent neurobiological features of ASD (29,32,52–57). Together, our findings point to evidence for DMN dysfunction in the context of specific components of social cognitive dysfunction in ASD: self-referential processing and theory of mind.

Our stDNN-based feature detection analysis also identified the superior temporal sulcus and the dorsolateral and ventrolateral prefrontal cortex, which anchor the human voice processing and cognitive control systems, respectively, as brain areas whose dynamic properties distinguished the ASD and control groups in three different cohorts. The human voice is a key social stimulus, and engaging with it is important for language and social-emotional learning during typical development (30,31,58–60). Notably, individuals with ASD are often not responsive to human voice (1,61), and core deficits in processing biologically and emotionally salient vocal cues contribute to pronounced social communication difficulties in ASD (2,30,31,58). Impairments in prefrontal cortex areas associated with cognitive control are a prominent defining feature of ASD. Cognitive control systems anchored in the dorsolateral and ventrolateral prefrontal cortex facilitate complex goal-directed behaviors, including attention allocation, task-switching, and other adaptive flexible behaviors, processes known to be impaired in ASD (62–64).

Given the heterogeneity of ASD, the fourth and final challenge we addressed here was to uncover neurobiologically interpretable features associated with the severity of social and communication deficits, a core defining characteristic of the disorder. Across the five folds of a cross-validation analysis, the PCC and precuneus were the only brain areas whose features predicted the heterogeneity of both social and communication deficits, but not restricted and repetitive behaviors, in individuals with ASD. Postmortem histological studies of brain tissue have demonstrated altered laminar

patterns and increased density of white matter neurons in the superficial layers of the PCC and precuneus, owing to disruption of neuronal migration from the germinal zone to the cortical plate in affected individuals (65). Furthermore, disruptions to the PCC and precuneus occur between the 16th and 20th weeks of gestation (65), and the presence of such early developmental aberrations may underpin the common profile of deficits observed in the three cohorts across a wide age range. Thus, an early and variable profile of insults to the cellular organization of the PCC and precuneus may adversely impact brain development and contribute to the ASD phenotype by virtue of focused and early disruption of core brain hubs.

Conclusions

Despite the considerable heterogeneity of phenotypic features in ASD and study limitations (see Supplemental Discussion for details), our stDNN model distinguished individuals with ASD from neurotypical controls with a high accuracy in multisite data. Our stDNN ASD model could not accurately distinguish individuals with ADHD from neurotypical controls, highlighting the specificity of the model. Our model together with an integrated feature identification approach successfully uncovered the neurobiological features associated with ASD. Notwithstanding the wide range in symptom profiles, age, and data acquisition protocols, our model accurately distinguished ASD from control subjects and uncovered similar distinguishing neurobiological features associated with ASD in two other independent cohorts. Our findings also yielded a unique predictive fingerprint in each individual that robustly predicted the severity of social and communication deficits in ASD. Our discovery of robust individualized functional brain biomarkers of ASD psychopathology could transform our understanding of the etiology, diagnosis, and treatment of the disorder. More generally, our approach provides new XAI-based tools for probing the robust and interpretable neurobiological bases of psychiatric disorders and the underlying clinical symptoms, with the potential to inform precision psychiatry.

ACKNOWLEDGMENTS AND DISCLOSURES

This research was supported by grants from the National Institutes of Health (Grant Nos. MH084164, EB022907, and MH121069 [to VM]; Grant No. K25HD074652 [to SR]; Grant No. AG072114 [to KS]), the Stanford Maternal and Child Health Research Institute through the Transdisciplinary Initiatives and Uytengsu-Hamilton 22q11 Programs (to VM and KS), a NARSAD Young Investigator Award (to KS), a Stanford Innovator Award (to KS), and the Taube Maternal and Child Health Research Fund (to KS). KS is a Taube Family Endowed Transdisciplinary Investigator for Maternal Child Health.

We greatly appreciate the contributions of the study participants and the Autism Brain Imaging Data Exchange, ADHD200, and GENDAAR initiatives, without which this work would not be possible.

The authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Department of Psychiatry and Behavioral Sciences (KS, SR, RY, DK, CdlA, VM), Department of Neurology & Neurological Sciences (VM), and the Stanford Neurosciences Institute (VM), Stanford University School of Medicine, Stanford, California.

KS and SR contributed equally to this work.

Address correspondence to Kaustubh Supekar, Ph.D., at ksupekar@stanford.edu, or Vinod Menon, Ph.D., at menon@stanford.edu.

Received Jun 1, 2021; revised and accepted Feb 4, 2022.

Supplementary material cited in this article is available online at https://doi.org/10.1016/j.biopsych.2022.02.005.

REFERENCES

1. Kanner L (1943): Autistic disturbances of affective contact. Nervous Child 2:217–250.
2. Klin A (1991): Young autistic children's listening preferences in regard to speech: A possible characterization of the symptom of social withdrawal. J Autism Dev Disord 21:29–42.
3. Maenner MJ, Shaw KA, Baio J, Washington A, Patrick M, DiRienzo M, *et al.* (2020): Prevalence of autism spectrum disorder among children aged 8 years—Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2016 [published correction appears in MMWR Surveill Summ 2020; 69:503]. MMWR Surveill Summ 69: 1–12.
4. Lord C, Brugha TS, Charman T, Cusack J, Dumas G, Frazier T, *et al.* (2020): Autism spectrum disorder. Nat Rev Dis Primers 6:5.
5. Lombardo MV, Lai MC, Baron-Cohen S (2019): Big data approaches to decomposing heterogeneity across the autism spectrum. Mol Psychiatry 24:1435–1450.
6. Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017): A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26.
7. Durstewitz D, Koppe G, Meyer-Lindenberg A (2019): Deep neural networks in psychiatry. Mol Psychiatry 24:1583–1598.
8. Aghdam MA, Sharifi A, Pedram MM (2019): Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks. J Digit Imaging 32:899–918.
9. Dvornek NC, Ventola P, Duncan JS (2018): Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks. Proc IEEE Int Symp Biomed Imaging 2018:725–728.
10. Dvornek NC, Ventola P, Pelphrey KA, Duncan JS (2017): Identifying autism from resting-state fMRI using long short-term memory networks. Mach Learn Med Imaging 10541:362–370.
11. Quaak M, van de Mortel L, Thomas RM, van Wingen G (2021): Deep learning applications for the classification of psychiatric disorders using neuroimaging data: Systematic review and meta-analysis. Neuroimage Clin 30:102584.
12. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F (2017): Identification of autism spectrum disorder using deep learning and the ABIDE dataset. Neuroimage Clin 17:16–23.
13. Sherkatghanad Z, Akhondzadeh M, Salari S, Zomorodi-Moghadam M, Abdar M, Acharya UR, *et al.* (2020): Automated detection of autism spectrum disorder using a convolutional neural network. Front Neurosci 13:1325.
14. Rakić M, Cabezas M, Kushibar K, Oliver A, Lladó X (2020): Improving the detection of autism spectrum disorder by combining structural and functional MRI information. Neuroimage Clin 25:102181.
15. Zhuang J, Dvornek NC, Li X, Ventola P, Duncan JS (2019): Invertible network for classification and biomarker selection for ASD. Med Image Comput Comput Assist Interv 11766:700–708.
16. Rathore A, Palande S, Anderson JS, Zielinski BA, Fletcher PT, Wang B (2019): Autism classification using topological features and deep learning: A cautionary tale. Med Image Comput Comput Assist Interv 11766:736–744.
17. Ryali S, Supekar K, Chen T, Kochalka J, Cai W, Nicholas J, *et al.* (2016): Temporal dynamics and developmental maturation of salience, default and central-executive network interactions revealed by variational Bayes hidden Markov modeling. PLoS Comput Biol 12: e1005138.
18. Taghia J, Cai W, Ryali S, Kochalka J, Nicholas J, Chen T, Menon V (2018): Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. Nat Commun 9:2505.

19. Ryali S, Supekar K, Chen T, Menon V (2011): Multivariate dynamical systems models for estimating causal interactions in fMRI. Neuroimage 54:807–823.
20. Koppe G, Meyer-Lindenberg A, Durstewitz D (2021): Deep learning for small and big data in psychiatry. Neuropsychopharmacology 46:176–190.
21. Calhoun VD, Miller R, Pearlson G, Adalı T (2014): The chronnectome: Time-varying connectivity networks as the next frontier in fMRI data discovery. Neuron 84:262–274.
22. Cai W, Chen T, Szegletes L, Supekar K, Menon V (2018): Aberrant time-varying cross-network interactions in children with attention-deficit/hyperactivity disorder and the relation to attention deficits. Biol Psychiatry Cogn Neurosci Neuroimaging 3:263–273.
23. Supekar K, Cai W, Krishnadas R, Palaniyappan L, Menon V (2019): Dysregulated brain dynamics in a triple-network saliency model of schizophrenia and its relation to psychosis. Biol Psychiatry 85:60–69.
24. Bolton TAW, Morgenroth E, Preti MG, Van De Ville D (2020): Tapping into multi-faceted human behavior and psychopathology using fMRI brain dynamics. Trends Neurosci 43:667–680.
25. Preti MG, Bolton TA, Van De Ville D (2017): The dynamic functional connectome: State-of-the-art and perspectives. Neuroimage 160:41–54.
26. Christoff K, Irving ZC, Fox KCR, Spreng RN, Andrews-Hanna JR (2016): Mind-wandering as spontaneous thought: A dynamic framework. Nat Rev Neurosci 17:718–731.
27. Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, *et al.* (2017): Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data 4:170010.
28. Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, *et al.* (2014): The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19:659–667.
29. Padmanabhan A, Lynch CJ, Schaer M, Menon V (2017): The default mode network in autism. Biol Psychiatry Cogn Neurosci Neuroimaging 2:476–486.
30. Abrams DA, Padmanabhan A, Chen T, Odriozola P, Baker AE, Kochalka J, *et al.* (2019): Impaired voice processing in reward and salience circuits predicts social communication in children with autism. Elife 8:e39906.
31. Abrams DA, Lynch CJ, Cheng KM, Phillips J, Supekar K, Ryali S, *et al.* (2013): Underconnectivity between voice-selective cortex and reward circuitry in children with autism. Proc Natl Acad Sci U S A 110:12060–12065.
32. Lynch CJ, Uddin LQ, Supekar K, Khouzam A, Phillips J, Menon V (2013): Default mode network in childhood autism: Posteromedial cortex heterogeneity and relationship with social deficits. Biol Psychiatry 74:212–219.
33. Davatzikos C (2019): Machine learning in neuroimaging: Progress and challenges. Neuroimage 197:652–656.
34. Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, *et al.* (2016): The human Brainnetome atlas: A new brain atlas based on connectional architecture. Cereb Cortex 26:3508–3526.
35. Lundberg SM, Lee SI (2017): A unified approach to interpreting model predictions. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, 4768–4777.
36. Simonyan K, Vedaldi A, Zisserman A (2013): Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv. https://arxiv.org/abs/1312.6034.
37. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M (2014): Striving for simplicity: The all convolutional net. arXiv. https://arxiv.org/abs/1412.6806.
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017): Grad-CAM: Visual explanations from deep networks via gradient-based localization. 2017 IEEE International Conference on Computer Vision (ICCV) 618–626.
39. Sundararajan M, Taly A, Yan Q (2017): Axiomatic attribution for deep networks. In: Precup D, Teh YW, editors. ICML '17: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR.org, 3319–3328.
40. Abrams DA, Kochalka J, Bhide S, Ryali S, Menon V (2020): Intrinsic functional architecture of the human speech processing network. Cortex 129:41–56.
41. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000): Voice-selective areas in human auditory cortex. Nature 403:309–312.
42. Kanwisher N, McDermott J, Chun MM (1997): The fusiform face area: A module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.
43. Gervais H, Belin P, Boddaert N, Leboyer M, Coez A, Sfaello I, *et al.* (2004): Abnormal cortical voice processing in autism. Nat Neurosci 7:801–802.
44. Hus V, Lord C (2014): The autism diagnostic observation schedule, module 4: Revised algorithm and standardized severity scores. J Autism Dev Disord 44:1996–2012.
45. Gotham K, Pickles A, Lord C (2009): Standardizing ADOS scores for a measure of severity in autism spectrum disorders. J Autism Dev Disord 39:693–705.
46. Di Martino A, Zuo XN, Kelly C, Grzadzinski R, Mennes M, Schvarcz A, *et al.* (2013): Shared and distinct intrinsic functional network centrality in autism and attention-deficit/hyperactivity disorder. Biol Psychiatry 74:623–632.
47. Schilbach L, Eickhoff SB, Rotarska-Jagiela A, Fink GR, Vogeley K (2008): Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. Conscious Cogn 17:457–467.
48. Mars RB, Neubert FX, Noonan MP, Sallet J, Toni I, Rushworth MFS (2012): On the relationship between the "default mode network" and the "social brain.". Front Hum Neurosci 6:189.
49. Spreng RN, Mar RA, Kim ASN (2009): The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. J Cogn Neurosci 21:489–510.
50. Laird AR, Fox PM, Eickhoff SB, Turner JA, Ray KL, McKay DR, *et al.* (2011): Behavioral interpretations of intrinsic connectivity networks. J Cogn Neurosci 23:4022–4037.
51. Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, *et al.* (2009): Correspondence of the brain's functional architecture during activation and rest. Proc Natl Acad Sci U S A 106:13040–13045.
52. Haar S, Berman S, Behrmann M, Dinstein I (2016): Anatomical abnormalities in autism? Cereb Cortex 26:1440–1452.
53. Valk SL, Di Martino A, Milham MP, Bernhardt BC (2015): Multicenter mapping of structural network alterations in autism. Hum Brain Mapp 36:2364–2373.
54. Uddin LQ, Menon V, Young CB, Ryali S, Chen T, Khouzam A, *et al.* (2011): Multivariate searchlight classification of structural magnetic resonance imaging in children and adolescents with autism. Biol Psychiatry 70:833–841.
55. Moseley RL, Ypma RJF, Holt RJ, Floris D, Chura LR, Spencer MD, *et al.* (2015): Whole-brain functional hypoconnectivity as an endophenotype of autism in adolescents. Neuroimage Clin 9:140–152.
56. Glerean E, Pan RK, Salmi J, Kujala R, Lahnakoski JM, Roine U, *et al.* (2016): Reorganization of functionally connected brain subnetworks in high-functioning autism. Hum Brain Mapp 37:1066–1079.
57. Yerys BE, Gordon EM, Abrams DN, Satterthwaite TD, Weinblatt R, Jankowski KF, *et al.* (2015): Default mode network segregation and social deficits in autism spectrum disorder: Evidence from non-medicated children. Neuroimage Clin 9:223–232.
58. Kuhl PK, Conboy BT, Padden D, Nelson T, Pruitt J (2005): Early speech perception and later language development: Implications for the "critical period.". Lang Learn Dev 1:237–264.
59. Christophe A, Dupoux E, Bertoncini J, Mehler J (1994): Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. J Acoust Soc Am 95:1570–1580.
60. DeCasper AJ, Fifer WP (1980): Of human bonding: Newborns prefer their mothers' voices. Science 208:1174–1176.
61. Harstad L, Baum C, Yatchmink Y: Early warning signs of autism spectrum disorder. National Center on Birth Defects and Developmental

Disabilities, Centers for Disease Control and Prevention Available at: https://www.cdc.gov/ncbddd/actearly/autism/curriculum/documents/early-warning-signs-autism_508.pdf. Accessed March 17, 2022.

62. Demetriou EA, Lampit A, Quintana DS, Naismith SL, Song YJC, Pye JE, *et al.* (2018): Autism spectrum disorders: A meta-analysis of executive function. Mol Psychiatry 23:1198–1204.
63. Hill EL (2004): Executive dysfunction in autism. Trends Cogn Sci 8:26–32.
64. Lai CLE, Lau Z, Lui SSY, Lok E, Tam V, Chan Q, *et al.* (2017): Meta-analysis of neuropsychological measures of executive functioning in children and adolescents with high-functioning autism spectrum disorder. Autism Res 10:911–939.
65. Oblak AL, Rosene DL, Kemper TL, Bauman ML, Blatt GJ (2011): Altered posterior cingulate cortical cytoarchitecture, but normal density of neurons and interneurons in the posterior cingulate cortex and fusiform gyrus in autism. Autism Res 4:200–211.