PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES
NEUROSCIENCE

Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization

Srikanth Ryali[a,1], Yuan Zhang[a,1], Carlo de los Angeles[a], Kaustubh Supekar[a,b,c], and Vinod Menon[a,b,c,d,2]

Edited by Ruben C. Gur, University of Pennsylvania, Philadelphia, PA; received June 23, 2023; accepted December 21, 2023 by Editorial Board Member Terrence J. Sejnowski

Sex plays a crucial role in human brain development, aging, and the manifestation of psychiatric and neurological disorders. However, our understanding of sex differences in human functional brain organization and their behavioral consequences has been hindered by inconsistent findings and a lack of replication. Here, we address these challenges using a spatiotemporal deep neural network (stDNN) model to uncover latent functional brain dynamics that distinguish male and female brains. Our stDNN model accurately differentiated male and female brains, demonstrating consistently high cross-validation accuracy (>90%), replicability, and generalizability across multisession data from the same individuals and three independent cohorts (N ~ 1,500 young adults aged 20 to 35). Explainable AI (XAI) analysis revealed that brain features associated with the default mode network, striatum, and limbic network consistently exhibited significant sex differences (effect sizes > 1.5) across sessions and independent cohorts. Furthermore, XAI-derived brain features accurately predicted sex-specific cognitive profiles, a finding that was also independently replicated. Our results demonstrate that sex differences in functional brain dynamics are not only highly replicable and generalizable but also behaviorally relevant, challenging the notion of a continuum in male-female brain organization. Our findings underscore the crucial role of sex as a biological determinant in human brain organization, have significant implications for developing personalized sex-specific biomarkers in psychiatric and neurological disorders, and provide innovative AI-based computational tools for future research.

explainable AI | sex | brain | human

Significance

Sex is an important biological factor that influences human behavior, impacting brain function and the manifestation of psychiatric and neurological disorders. However, previous research on how brain organization differs between males and females has been inconclusive. Leveraging recent advances in artificial intelligence and large multicohort fMRI (functional MRI) datasets, we identify highly replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization localized to the default mode network, striatum, and limbic network. Our findings advance the understanding of sex-related differences in brain function and behavior. More generally, our approach provides AI–based tools for probing robust, generalizable, and interpretable neurobiological measures of sex differences in psychiatric and neurological disorders.

Sex plays a significant role in early brain development, adolescence, and aging (1), and many aspects of both normal and pathological brain functioning exhibit sex differences (1–5). These differences are particularly evident in the etiology of most psychiatric and neurological disorders (6–9). Research has consistently shown that females are more likely than males to experience depression, anxiety, and eating disorders (10). Disorders such as autism, attention-deficit hyperactivity disorder, and schizophrenia are more prevalent in males compared to females and present sex-specific clinical manifestations and outcomes (11–13). Consequently, knowledge of sex differences in the human brain is critical for understanding both normative behavior and psychopathology.

Most of our understanding of sex differences in the human brain stems from studies of its anatomy and structure (see ref. 14 for a recent review). Postmortem as well as in vivo structural brain imaging studies have demonstrated that males have a larger total brain volume than females (15–18). Furthermore, the percentage of white matter volume in the male brain is found to be higher than the female brain (19). In contrast, female brains have higher gray matter percentage than male brains (19). At the regional level, research has consistently reported sex differences in volumes of the amygdala, hippocampus, and insula (20). Similarly, structural connectivity has been shown to differ by sex. Using diffusion tensor imaging, Inghalikar et al. found that male brains have higher intrahemisphere structural connectivity than female brains, and female brains have higher interhemispheric structural connectivity than male brains (21). Classification analysis has suggested that multivariate structural brain patterns may accurately distinguish between sexes (22–24).

Despite growing evidence of sex differences in structural human brain organization, it is unclear whether and how these structural differences translate to functional brain organization differences. The increasing availability of resting-state functional MRI (rsfMRI) data has led to greater use of connectivity analyses to explore sex differences in brain function. These studies have found sex differences in local and long-range functional connectivity. In particular, females were shown to have higher local functional connectivity density (25)

Author contributions: S.R. and V.M. designed research; S.R., Y.Z., C.d.l.A., K.S., and V.M. performed research; Y.Z. and C.d.l.A. analyzed data; and S.R., Y.Z., K.S., and V.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.C.G. is a guest editor invited by the Editorial Board.

[1]S.R. and Y.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: menon@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2310012121/-/DCSupplemental.

Published February 20, 2024.

as well as stronger functional connectivity in the default mode network (DMN) than males (26–29). Males, on the other hand, have been reported to have stronger functional connectivity in sensorimotor cortices than females (29). There have also been reports of sex differences in the lateralization of functional brain connectivity with males having greater rightward lateralization of short-range connectivity and females having greater leftward lateralization of long-range connectivity (30). Classification analysis has reported that functional brain connectivity patterns can distinguish between sexes with accuracies ranging from 62 to 87% (26, 27, 29–38). However, findings from previous rsfMRI studies have been inconsistent due to wide age ranges spanning childhood through adulthood and the inclusion of individuals with psychopathology (26, 27, 29–38) (see *SI Appendix*, Fig. S1 and Table S1 for a summary). Critically, the replicability and generalizability of findings remain unclear, as few studies have utilized robust predictive models to assess the replicability and stability of sex differences across multiple sessions in the same individual or their generalizability across independent cohorts. One study that used a predictive model to distinguish sex in previously unseen data reported classification accuracy of about 60% (37), raising concerns about the replicability and generalizability of sex differences in human functional brain organization. Moreover, the specific brain regions and networks that underlie sex differences are not well understood. A more rigorous quantitative characterization of brain areas and networks driving sex differences is crucial for understanding normative functional brain organization and for elucidating sex-specific vulnerability to psychiatric and neurological disorders (1).

To address critical gaps in the literature and identify replicable, generalizable, and behaviorally relevant sex differences in functional brain organization, we developed an end-to-end spatiotemporal deep neural network (stDNN) model and an explainable AI (XAI)-based computational framework (Fig. 1). Our stDNN model was trained on a large sample (N ~ 1,000) of rsfMRI data from the Human Connectome Project (HCP) (39). We then assessed the replicability of our predictive models on multiple HCP sessions without additional training. Furthermore, we evaluated the generalizability of the stDNN model to two independent age-matched cohorts from the Nathan Kline Institute–Rockland Sample (NKI-RS) (40) and Max Planck Institute (MPI) Leipzig (41), again without additional training. Our study focuses on young adults ages 20 to 35 y, precluding the use of developmental (e.g., ABCD) and aging (e.g., UK BioBank) cohorts (*SI Appendix*, Table S1).

We had four main goals. Our first goal was to determine whether there are reliable sex differences in the functional organization of the human brain. Recent advances in DNNs have revolutionized the field of machine learning, and there is a growing interest in their use for the classification of normative as well as neuropsychiatric conditions from fMRI data (42–46). DNN models in fMRI research have primarily focused on classification using precomputed functional connectivity between brain regions (33). However, recent studies have shown that fMRI time series are highly nonstationary with significant differences in dynamic brain connectivity within subjects and across groups (47–49). Our stDNN model addresses the limitations of precomputed connectivity features, capturing latent circuit dynamics without stationarity assumptions and feature engineering. This also represents a significant advantage over extant DNN models in fMRI research (50, 51). stDNN directly takes as its input fMRI time series and uses multiple one-dimensional convolutions of time-series segments across brain regions to uncover latent circuit dynamics that distinguish between males and females. Additional details of the technical innovations of our approach are in the *Materials and Methods* section.

Our second goal was to address the reproducibility crisis in sex differences research (52, 53) by investigating the replicability and generalization of sex differences in the functional organization of the human brain. We first examined the performance of the stDNN model trained on HCP data from one session in distinguishing between female and male brains using data from the same individuals acquired in three other HCP sessions (54, 55). Next, we investigated the ability of the stDNN model trained on HCP data to differentiate between female and male brains in independent data from the NKI-RS and MPI-Leipzig cohorts. This approach allowed us to probe the generalization to new (untrained) data acquired on different scanners and data acquisition protocols, thereby addressing the replicability and generalizability of sex differences in the human brain. We hypothesized that our stDNN model, trained on data from one HCP session, would reveal sex differences in the three other HCP sessions and generalize to previously unseen data from entirely different cohorts.

Our third goal was to identify stable neurobiologically interpretable features underlying sex differences. Previous studies using DNNs in brain imaging have almost exclusively focused on classification accuracy and have not paid adequate attention to the neurobiological features that underlie classification. We address this black-box problem associated with DNN-based architectures by using XAI-based techniques, which allowed us to identify brain features or fingerprints (56, 57) that differentiate functional brain organization in females and males (58). We used an integrated gradients (IG) algorithm which estimates the integral of gradients with

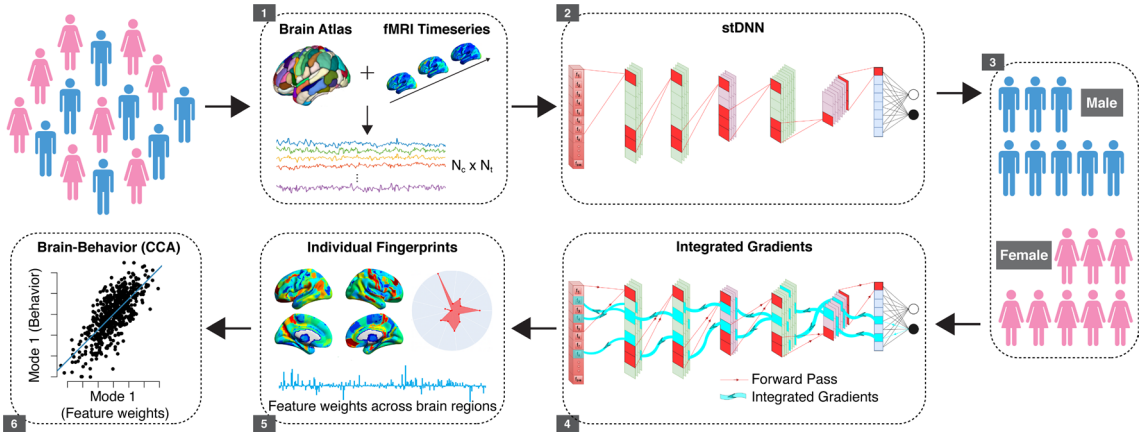

Fig. 1. Schematic overview of the multicomponent XAI framework for identifying individualized brain fingerprints that predict sex and cognitive profiles. Key steps include data extraction (step 1), classification (steps 2 and 3), feature identification, i.e., feature weights ("fingerprints") across brain regions predictive of sex (steps 4 and 5), and prediction of cognitive profiles (step 6). XAI = explainable AI.

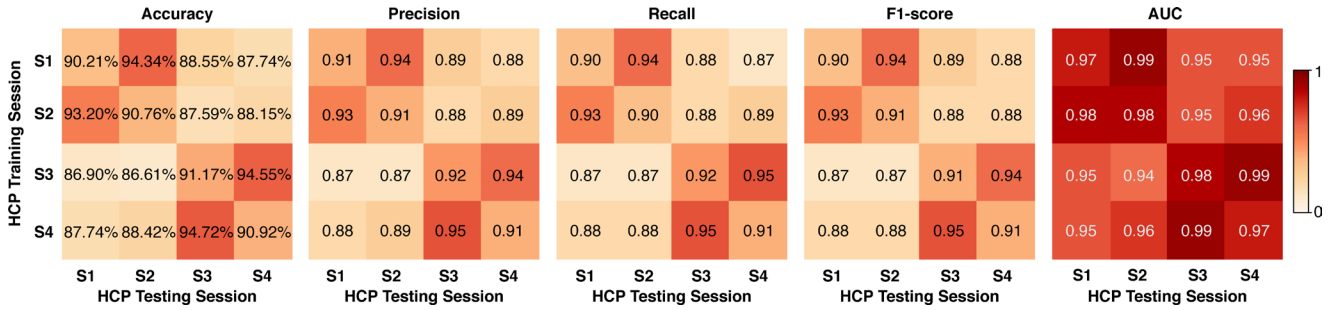

Fig. 2. Fivefold cross-validation classification performance in each HCP session data and its replicability in the other three HCP sessions without any additional training. For each of the five performance metrics (accuracy, macro-precision, macro-recall, macro-F1 score, and AUC), we showed pairwise results of mean performance in a matrix, with rows referring to the HCP training sessions (i.e., which session the stDNN models were trained on) and columns referring to the HCP testing sessions (i.e., which session the stDNN models were tested on).

respect to inputs along the path from a given (or random) baseline to an input, which provides a score of how important each feature contributes to the final prediction (59–62). This XAI algorithm also provides a ranking of brain features (weights) that distinguish between females and males. We then used consensus analysis to identify brain features that are consistent across cross-validation models. We predicted that our XAI-based approach and consensus analysis would allow us to capture interpretable and replicable neurobiological features underlying sex differences in functional brain organization. In addition to the stability of regional brain features underlying sex differentiation, we also examined the consistency of differences in large-scale cortical and subcortical networks.

Our final goal was to relate sex differences in functional brain organization to behavior in females and males. Sex differences in multiple domains of cognitive functioning have been extensively investigated over the past two decades (63). Critically, the relation between sex-specific cognitive profiles and functional brain organization is poorly understood. To address this, we leveraged the deeply phenotyped NIH Toolbox (64) behavioral data and used individual-level brain features derived using stDNN as predictors of cognitive profiles and evaluated the sex-specificity of brain–behavior relations in females and males. We hypothesized that individual-level functional brain features that differ between sexes would predict cognitive profiles, and brain–behavior relationships would differ between sexes.

Our approach using spatiotemporal DNNs and XAI techniques identifies replicable, generalizable, and interpretable sex differences in human functional brain organization across multiple datasets and independent cohorts and, furthermore, reveals that functional brain features that differ between sexes are behaviorally relevant. Finally, we demonstrate the advantages of our approach over conventional machine learning methods.

Results

Classification of Sex Differences within the HCP Cohort. We used stDNN (*SI Appendix*, Fig. S2) to distinguish between females and males using fMRI time series without explicit feature engineering. We first trained stDNN models on each HCP session separately and tested the performance of models within each respective HCP session (*SI Appendix*, Table S2). To assess model performance, 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 testing (*SI Appendix*, Fig. S3*A*). Our stDNN models achieved high average accuracies (mean: 90.21 to 91.17%; SD: 1.21 to 2.85%) across the five folds and high average macro-precision (mean: 0.91 to 0.92; SD: 0.01 to 0.03), macro-recall (mean: 0.90 to 0.92; SD: 0.01 to 0.03), macro-F1 scores (mean: 0.90 to 0.91; SD: 0.01 to 0.03), and AUC (mean: 0.97 to 0.98; SD: 0 to 0.01) (Fig. 2 and *SI Appendix*, Fig. S4). These results demonstrate reliable sex differences across cross-validation folds across sessions.

We then evaluated the replicability of sex differences by applying stDNN models trained on one HCP session to the other three HCP sessions without any additional training. stDNN models achieved high average accuracies across the five folds (mean: 86.61 to 94.72%; SD: 0.35 to 2.85%) and high average macro-precision (mean: 0.87 to 0.95; SD: 0 to 0.03), macro-recall (mean: 0.87 to 0.95; SD: 0.01 to 0.03), macro-F1 scores (mean: 0.87 to 0.95; SD: 0 to 0.03), and AUC (mean: 0.94 to 0.99; SD: 0 to 0.01) (Fig. 2 and *SI Appendix*, Fig. S4). These results demonstrate replicable sex differences across stDNN cross-validation folds and sessions, without the need for additional training.

Distinctiveness of Brain Features Underlying Sex Differences in the HCP Cohort. We then used XAI-based approaches to identify the brain features underlying the classification of female and male brains. We identified individual fingerprints of predictive brain features in each participant using an IG procedure (58) (*SI Appendix*, Fig. S5). Briefly, a "fingerprint" of an individual refers to the unique whole brain pattern of an IG-derived stDNN model feature importance that classifies that individual as either female or male. We evaluated the validity of brain features distinguishing females and males by measuring the similarity between IG-derived dynamic brain features. Based on their fingerprints, individuals of the same sex were clearly grouped into the same cluster (Fig. 3*A*). To further validate our findings, we generated group-level fingerprints for females and males separately. For each individual, we computed the similarity between their fingerprint and the group-level fingerprints as well as the similarity between group-level fingerprints using Pearson correlation. Using Fisher-Z tests, we found that for all males, individual-level fingerprints were significantly more similar to the group-level male fingerprint than to the group-level female fingerprint ($3.35 < Zs < 14.79$, $ps <$ 1e-4; Fig. 3*A*). Similarly, for all females, individual-level fingerprints were significantly more similar to the group-level female fingerprint than to the group-level male fingerprint ($3.22 < Zs < 14.84$, $ps <$ 1e-4; Fig. 3*A*). These results demonstrate that stDNN together with IG procedures reliably identifies discriminating brain features underlying sex differences, without the need for ad hoc feature engineering.

Consensus Analysis of Brain Features Underlying Sex Differences in the HCP Cohort. Next, we sought to identify brain features that most consistently discriminated between female and male

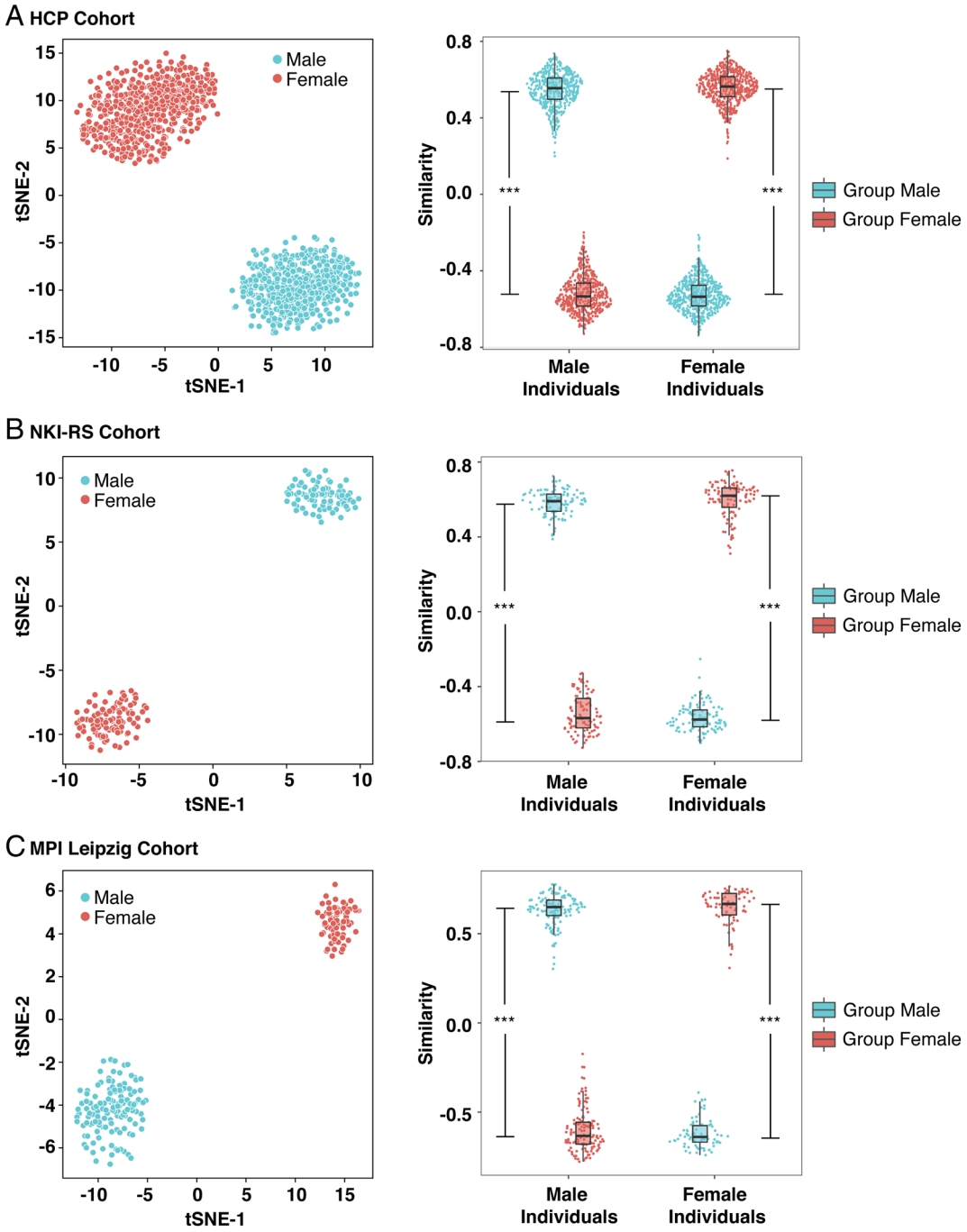

Fig. 3. Distinctiveness of brain fingerprints (feature attribution maps) underlying sex differences in the HCP (*A*), NKI-RS (*B*), and MPI Leipzig (*C*) cohorts. The T-distributed stochastic neighbor embedding (tSNE) plot of individual fingerprints (feature attribution maps) from the trained HCP session 1 stDNN model demonstrates distinct clustering of males and female brain fingerprints across the three cohorts. Violin and box plots of similarity between individual fingerprints and group-level fingerprints from the trained HCP session 1 stDNN model demonstrate that individual fingerprints are more similar to the group-level fingerprints of the same sex across the three cohorts. ***$P < 0.001$.

brains. To address this, we conducted a consensus analysis using multiple fivefold cross-validation iterations in each of the four HCP sessions, which was designed to identify features unbiased by any single cross-validation split of the data. Briefly, for each HCP session, we trained 500 models on different subsets of a specific HCP session (model session), which were used to compute IG-based feature attributions for all subjects in a specific HCP session (testing session), resulting in 500 sets of feature attributions for the testing session (see *Materials and Methods* for details). We then identified the top 20% features for each set, counted their occurrence across all sets, and thresholded them using a binomial distribution. These procedures were repeated for all pairs of HCP sessions, resulting in 16 consensus maps (4 HCP model sessions x 4 HCP testing sessions; Fig. 4). Across all 16 consensus maps, we identified the precuneus, ventromedial prefrontal cortex, ventrolateral prefrontal cortex, dorsolateral prefrontal cortex, and superior temporal gyrus as brain areas that most reliably contributed to sex differences (Fig. 4 and *SI Appendix*, Table S3).

Stability Analysis of Intraindividual Brain Features Underlying Sex Differences in the HCP Cohort. Results from stability analysis confirmed that brain features underlying sex differences are stable at the individual participant level (*SI Appendix, Supplementary Results*).

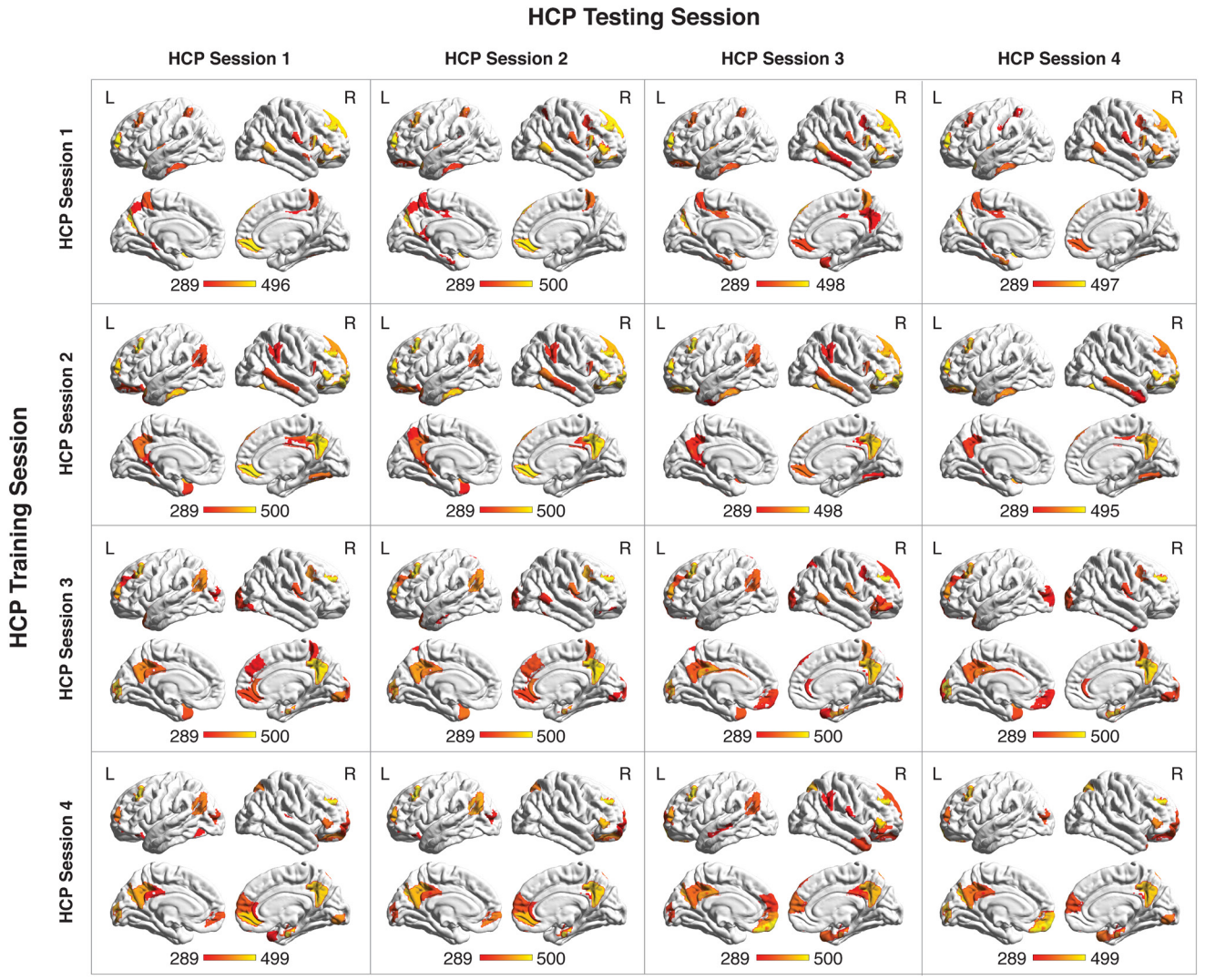

Fig. 4. Consensus maps of discriminating brain features in the HCP cohort. Consensus maps showing robust discriminating features underlying males vs. females classification for each pair of HCP sessions (one as the training session and the other as the testing session), including precuneus, ventromedial prefrontal cortex, ventrolateral prefrontal cortex, dorsolateral prefrontal cortex, and superior temporal gyrus (see *SI Appendix*, Table S4 for detailed listing of brain areas and the total count of occurrence across all 16 consensus maps).

Control Analyses with Different Brain Atlases, Artifact Reduction Methods, and Head Movement in the HCP Cohort. Results from several control analyses confirmed that our findings were robust with respect to brain atlases and artifacts reduction methods (*SI Appendix*, Table S4) and head motion (*SI Appendix*, *Supplementary Results*).

Generalization of Sex Classification Models Trained on the HCP Cohort to Independent NKI-RS and MPI Leipzig Cohorts. Next, we examined whether stDNN models trained on HCP data could distinguish, without any additional training, between females and males using rsfMRI from the NKI-RS and MPI Leipzig cohorts. We first applied the stDNN models trained on HCP session 1 data to NKI-RS cohort data (N = 205) consisting of 108 females and 97 males who were age matched to the HCP cohort. Among the four HCP sessions, we chose the stDNN model trained on HCP session 1 data to assess generalizability and subsequent analyses as it achieved the best cross-session generalizability among the four sessions (Fig. 2 and *SI Appendix*, Fig. S4). We found that stDNN models trained on HCP session 1 rsfMRI data achieved an average accuracy of 81.84 ± 1.43%, across the five folds, and an average macro-precision of 0.83 ± 0.01, macro-recall of 0.82 ± 0.02, macro-F1 score of 0.81 ± 0.02, and AUC of 0.90 ± 0.01 (*SI Appendix*, Table S5) in the NKI-RS cohort data.

We then applied the stDNN models trained on HCP session 1 data to MPI Leipzig cohort rsfMRI data (N = 215) consisting of 78 females and 137 males who were age-matched to the HCP cohort. We found that stDNN models trained on HCP session 1 rsfMRI data achieved an average accuracy of 82.60 ± 1.68%, across the five folds, and an average macro-precision of 0.82 ± 0.02, macro-recall of 0.82 ± 0.01, macro-F1 score of 0.81 ± 0.01, and AUC of 0.89 ± 0.01 (*SI Appendix*, Table S5). These results demonstrate generalizable sex differences in human functional brain organization in new cohorts without any additional training.

Generalization of Brain Features Underlying Sex Differences from the HCP to Independent NKI-RS and MPI Leipzig Cohorts. We examined the generalizability of discriminating features identified in HCP data to independent NKI-RS and MPI Leipzig cohorts. We trained 500 stDNN models (5 folds × 100 iterations) on HCP session 1 data and determined brain feature attributions in each participant from the NKI-RS and MPI Leipzig cohorts. Consensus analyses identified precuneus, ventromedial prefrontal cortex, ventrolateral prefrontal cortex, dorsolateral prefrontal

cortex, and middle and superior temporal gyri as brain areas that most consistently predicted sex (Fig. 5 *A* and *B* and *SI Appendix*, Tables S6 and S7). Additional consensus analysis across all three cohorts further confirmed our findings (Fig. 5*C* and *SI Appendix*, Table S8). These results demonstrate that brain features that discriminate between females and males generalized well from the HCP cohort to two independent cohorts.

Distinctiveness of Brain Features Underlying Sex Differences in NKI-RS and MPI Leipzig Cohorts. We evaluated the distinctiveness of brain features distinguishing females and males by measuring the similarity between IG-derived dynamic brain features. Individual fingerprints were computed for each participant in the NKI-RS cohort and MPI Leipzig cohort using stDNN models trained on HCP session 1 data (*SI Appendix*, Fig. S6). Individuals of the same sex were clearly grouped into the same cluster in both cohorts (Fig. 3 *B* and *C*). We further evaluated the distinctiveness using group-level fingerprints for females and males separately. For each individual, we computed the similarity between their fingerprint and the group-level fingerprints as well as the similarity between the group-level fingerprints using Pearson correlation. Using Fisher Z tests, we found that for all males, individual-level fingerprints were significantly more similar to the group-level male fingerprint than to the group-level female fingerprint (NKI-RS: $6.27 < Zs < 14.18$, $ps <$ 1e-4, Fig. 3*B*; MPI Leipzig: $3.86 < Zs < 16.19$, $ps <$ 1e-4, Fig. 3*C*). Similarly, for all females, individual-level fingerprints were significantly more similar to the group-level female fingerprint than to the group-level male fingerprint (NKI-RS: $4.61 < Zs < 14.53$, $ps <$ 1e-4, Fig. 3*B*; MPI Leipzig: $5.78 < Zs < 15.20$, $ps <$ 1e-4, Fig. 3*C*). These results demonstrate the distinctiveness of brain features underlying sex differences in two independent cohorts.

Control Analyses with Different Brain Atlases, Artifact Reduction Methods, and Head Movement in the NKI-RS and MPI Leipzig Cohorts. Results from several control analyses confirmed that our findings were robust with respect to brain atlases and artifact reduction methods (*SI Appendix*, Tables S9 and S10) and head motion (*SI Appendix*, *Supplementary Results*).

Network-Level Differences in Brain Features Underlying Sex Differences. Extending our analysis of regional brain features, we then examined sex differences in 20 brain networks, including the 17 cortical networks (65) and three additional subcortical networks encompassing the amygdala–hippocampus, striatum, and thalamus. We computed the effect size of weighted brain features in each network and rank-ordered them based on the consistency of the effect size across six datasets, including four HCP sessions and the NKI-RS and MPI Leipzig cohorts. We found that the DMN most consistently showed the largest effect size (Cohen's $d > 2$), followed by the striatum and limbic network ($d > 1.5$) (Fig. 6). These results converge on and extend a regional-level consensus analysis of brain features that differentiate female and male brains.

Generalization of Sex Differences Using Conventional Machine Learning Methods. We examined the generalizability of seven conventional machine learning approaches (66). Consistent with many prior rsfMRI studies (31, 33, 34, 36–38), we used precomputed functional connectivity between the 246 brain regions as brain features in the classification analysis. We first trained and tested models on HCP session 1 data using a fivefold cross-validation procedure and then evaluated generalization on independent NKI-RS and MPI Leipzig cohorts without any additional training. These analyses reveal that, unlike our stDNN models, conventional approaches do not generalize well to untrained data from independent cohorts (*SI Appendix*, Tables S11–S13 and *Supplementary Results*).

Sex-Specific Neurobiological Predictors of Cognition. We examined a comprehensive battery of 14 cognitive measures from the NIH toolbox in the HCP cohort, including episodic memory, cognitive flexibility, response inhibition, fluid intelligence, reading, vocabulary comprehension, processing speed, and delay discounting (*SI Appendix*, Table S14). Principal component analysis with varimax rotation identified three components that together explained 47.7% of the total variance (*SI Appendix*, Fig. S7). The first component was aligned with general intelligence, the second with response inhibition and processing speed, and the third with delay discounting and reward sensitivity. Scores on these three components were used to derive a cognitive profile for each individual. We then examined sex-specific neurobiological predictors of cognitive function using canonical correlation analysis (CCA; *SI Appendix*, Fig. S3*B*), with the three principal components as behavioral variables and the feature importance of the 246 brain regions as brain variables.

We first conducted CCA using brain features from HCP session 1, as described above, to determine sex differences in predictors of the relationship between brain and cognitive measures. In males, CCA yielded three modes with squared canonical correlations (R_c^2) of 0.62, 0.53, and 0.48 (Fig. 7*A*). The CCA model was statistically significant (Pillai's trace = 1.624, P = 0.024, 95% CI: 1.406 to 1.621, permutation test) and explained over 90% of the variance. We then performed a dimension reduction analysis to determine significant modes (67). The full model (modes 1 to 3) was statistically significant [$F(738, 720.96)$ = 1.17, P = 0.016, 95% CI: 0.86 to 1.16] whereas modes 2 to 3 [$F(490, 482)$ = 0.99, P = 0.54, 95% CI: 0.84 to 1.19] and mode 3 [$F(244, 242)$ = 0.88, P = 0.84, 95% CI: 0.78 to 1.29] did not explain significant additional shared variance between brain and cognitive measures, suggesting that only mode 1 was relevant (67). Permutation test with FDR correction further confirmed a significant mode 1 (P = 0.009, 95% CI for mode 1 R_c^2: 0.50 to 0.60). Brain features associated with the dorsolateral prefrontal cortex, posterior cingulate cortex, precuneus, and postcentral gyrus predicted component three scores, which are associated with delay discounting and reward sensitivity, in males (*SI Appendix*, Table S15).

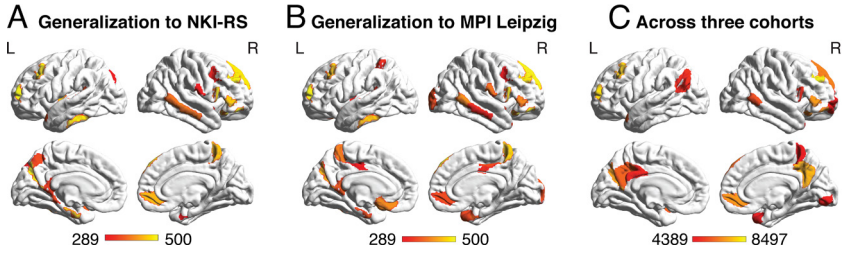

Fig. 5. Consensus maps of discriminating brain features in the independent (*A*) NKI-RS and (*B*) MPI Leipzig cohorts and (*C*) across all three cohorts. Consensus maps showing robust discriminating features underlying males vs. females classification for NKI-RS and MPI Leipzig cohorts as well as across the three cohorts (HCP, NKI-RS, and MPI Leipzig), including precuneus, ventromedial prefrontal cortex, ventrolateral prefrontal cortex, dorsolateral prefrontal cortex, and superior temporal gyrus (see *SI Appendix*, Tables S6–S8 for detailed listing of brain areas and the count of occurrence).

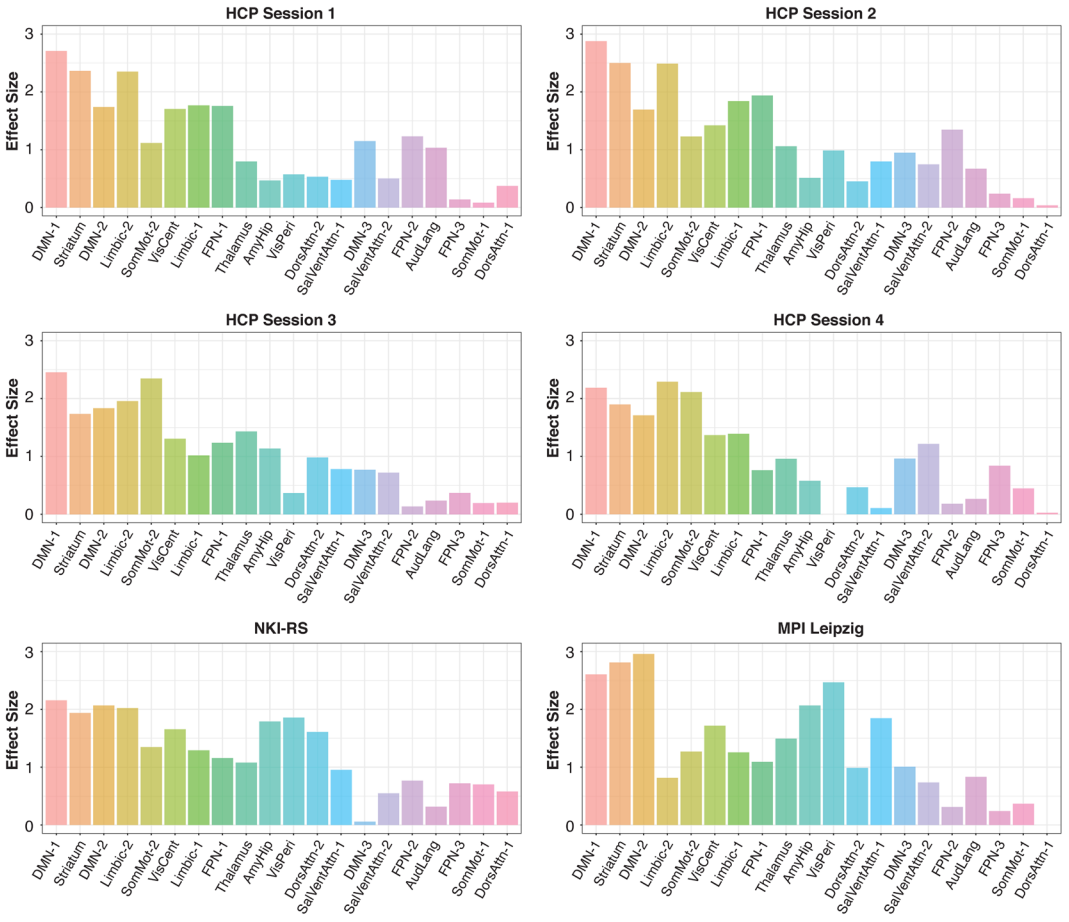

Fig. 6. Effect size of sex differences in 20 networks across the HCP, NKI-RS, and MPI Leipzig cohorts. Brain networks were ordered based on the ranking of effect sizes across the 4 HCP sessions and the two independent NKI-RS, and MPI Leipzig cohorts. The 20 networks consist of 17 cortical networks (65) and three subcortical networks encompassing the striatum, amygdala–hippocampus, and thalamus (*SI Appendix*, Table S18). The dorsal default mode network (DMN-1) showed the largest effect size (Cohen's $d > 2$) across all networks and cohorts, followed by the striatum and limbic networks ($d > 1.5$). Rank order: DMN-1 = dorsal default mode network; Striatum; DMN-2 = ventral default mode network; Limbic-2 = limbic network; SomMot-2 = somatomotor network; VisCent = visual central network; Limbic-1 = limbic network; FPN-1 = frontoparietal network; Thalamus; Amy-Hip = amygdala–hippocampus network; VisPeri = visual peripheral network; DorsAttn-2 = dorsal attention network; SalVentAttn-1 = salience/ventral attention network; DMN-3 = default mode network; SAlVentAttn-2 = salience/ventral attention network; FPN-2 = frontoparietal network; AudLang = auditory language network; FPN-3 = frontoparietal network; SomMot-1 = somatomotor network; DorsAttn-1 = dorsal attention network.

In females, CCA yielded three modes with R_c^2 of 0.55, 0.49, and 0.42 (Fig. 7*B*). Collectively, the full model across all modes was statistically significant (Pillai's trace = 1.453, $P = 0.001$, 95% CI: 1.190 to 1.381, permutation test) and explained 86% of the variance shared between the variable sets. Dimension reduction analysis showed that the full model (modes 1 to 3) was statistically significant [$F(738, 978.95) = 1.26$, $P = 4e\text{-}4$, 95% CI: 0.87 to 1.14] whereas modes 2 to 3 [$F(490, 654) = 1.11$, $P = 0.10$, 95% CI: 0.85 to 1.18] and mode 3 [$F(244, 328) = 0.97$, $P = 0.59$, 95% CI: 0.79 to 1.26] did not explain statistically significant shared variance between brain and behavioral measures, suggesting that only mode 1 was relevant (67). Permutation test with FDR correction further confirmed a significant mode 1 ($P = 0.002$, 95% CI for mode 1 R_c^2: 0.43 to 0.52). Brain features associated with the ventromedial prefrontal cortex, middle temporal gyrus, posterior cingulate cortex, precuneus, and postcentral gyrus predicted component one scores, which are associated with general intelligence, in females (*SI Appendix*, Table S16).

We then examined whether the CCA model from males could predict cognitive profiles in females and whether the CCA model from females could predict cognitive profiles in males. Applying the trained model from males to data from females revealed a mode 1 with R_c^2 of 0.008, which was not significant in terms of the permutation test ($P > 0.99$; Fig. 7*A*). Similarly, applying the trained model from females to data from males revealed a mode 1 with R_c^2 of 0.005, which was not significant in terms of the permutation test ($P > 0.93$; Fig. 7*B*).

These results demonstrate that the CCA model from males does not predict cognitive profiles in females, and conversely, the CCA model from females does not predict cognitive profiles in males.

Replication of Sex-Specific Neurobiological Predictors of Cognition. To examine the replicability of our findings, we conducted CCA on HCP session 3 data. In males, CCA yielded three modes with R_c^2 of 0.60, 0.56, and 0.50 for each successive function (Fig. 7*C*). Collectively, the full model across all modes was statistically significant (Pillai's trace = 1.659, $P = 0.004$, 95% CI: 1.403 to 1.620, permutation test) and explained a substantial portion, about 91%, of the variance shared between the variable sets. Dimension reduction analysis showed that the full model (modes 1 to 3) was statistically significant [$F(738, 720.96) = 1.22$, $P = 0.004$, 95% CI: 0.86 to 1.16] whereas modes 2 to 3 [$F(490, 482) = 1.11$, $P = 0.13$, 95% CI: 0.84 to 1.19] and mode 3 [$F(244, 242) = 0.99$, $P = 0.54$, 95% CI: 0.78 to 1.29] did not explain a statistically significant amount of shared variance between the variable sets, suggesting that only mode 1 was relevant (67). Permutation test with FDR correction further confirmed a significant mode 1 ($P = 0.034$, 95% CI for mode 1 R_c^2: 0.50 to 0.60) (Fig. 7*C*). Brain features associated with the dorsolateral prefrontal cortex, posterior cingulate cortex, precuneus, and postcentral gyrus again predicted component three scores, which are associated with delay discounting and reward sensitivity, in males (*SI Appendix*, Table S15).

In females, CCA yielded three modes with R_c^2 of 0.56, 0.46, and 0.40 for each successive function (Fig. 7*D*). Collectively, the full model across all modes was statistically significant (Pillai's trace = 1.427, $P = 0.008$, 95% CI: 1.201 to 1.398, permutation test) and explained a substantial portion, about 86%, of the variance shared between the variable sets. The dimension reduction analysis showed that the full model (modes 1 to 3) was statistically significant [$F(738, 957.95) = 1.20$, $P = 0.004$, 95% CI: 0.87 to 1.14] whereas modes 2 to 3 [$F(490, 640) = 1.00$, P

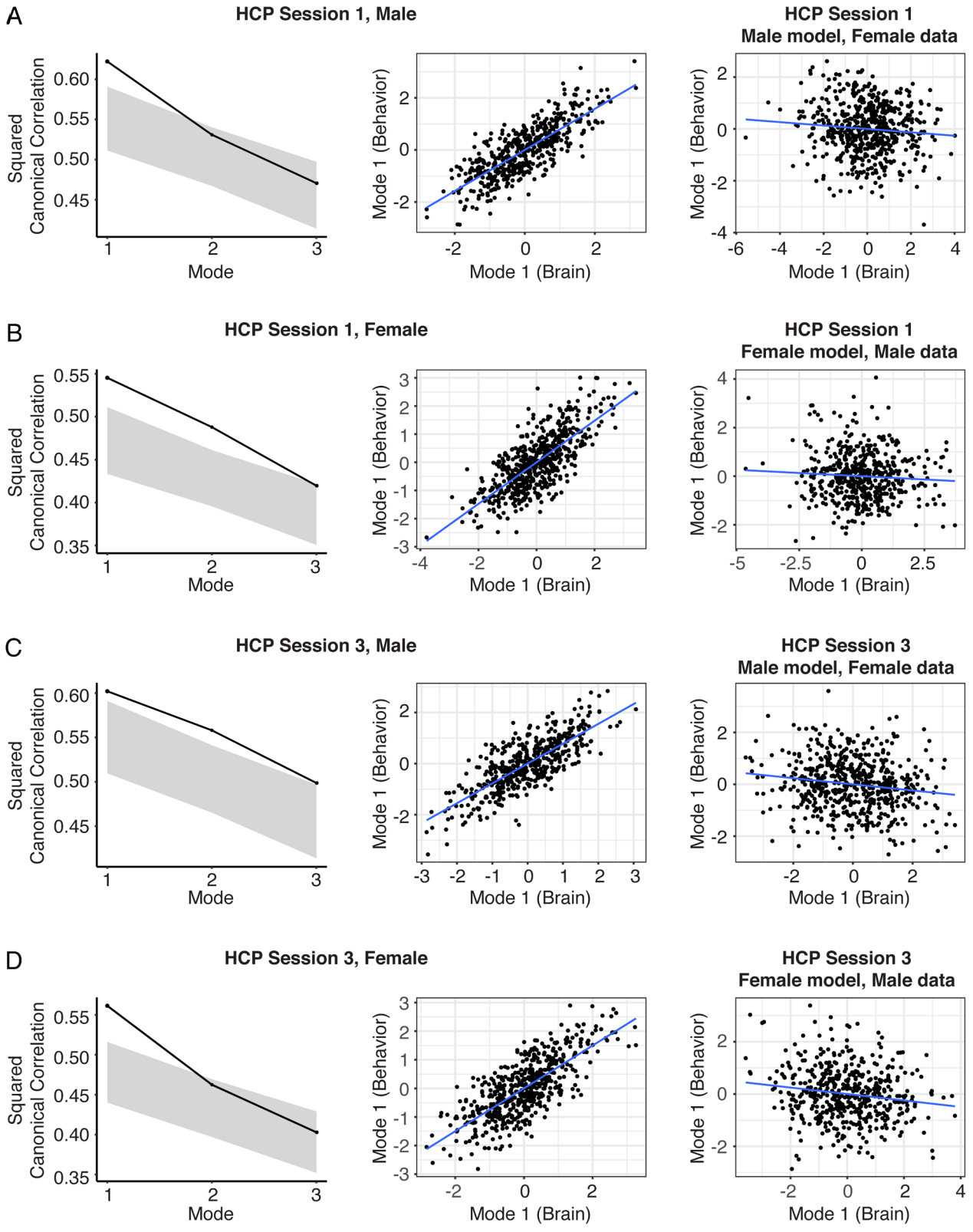

Fig. 7. CCA reveals significant sex-specific associations between stDNN-derived brain features and cognitive profiles. (*A*) CCA model from males in HCP session 1 data predicted cognitive profiles in males but not females. (*B*) CCA model from females in HCP session 1 data predicted cognitive profiles in females but not males. (*C*) CCA model from males in HCP session 3 data predicted cognitive profiles in males but not females. (*D*) CCA model from females in HCP session 3 data predicted cognitive profiles in females but not males. Line plots show squared canonical correlations, indicating the variance explained by each CCA mode. The gray area displays the 5th and 95th percentiles of the null distribution estimated via permutation testing.

= 0.50, 95% CI: 0.85 to 1.18] and mode 3 [$F(244, 321) = 0.89$, $P = 0.84$, 95% CI: 0.79 to 1.26] did not explain a statistically significant amount of shared variance between behavioral and brain measures, suggesting that only mode 1 was relevant (67). Permutation test with FDR correction further confirmed a significant mode 1 ($P = 0.001$, 95% CI for mode 1 R_c^2: 0.43 to 0.52) (Fig. 7*D*). Brain features associated with the ventromedial prefrontal cortex, middle temporal gyrus, posterior cingulate

cortex, precuneus, and postcentral gyrus predicted component one scores, which are associated with general intelligence, in females (*SI Appendix*, Table S16).

We further examined whether the CCA model from males could predict cognitive profiles in females and whether the CCA model from females could predict the cognitive profiles in males. Applying the trained model from males to data from females revealed a mode 1 with R_c^2 of 0.020, which was not significant in terms of the permutation test ($P > 0.99$; *SI Appendix*). Similarly, applying the trained model from females to data from males revealed a mode 1 with R_c^2 of 0.025, which was not significant in terms of the permutation test ($P > 0.99$; Fig. 7*D*). These results demonstrate that the CCA model from males cannot predict cognitive profiles in females, and the CCA model from females did not predict cognitive profiles in males.

A similar analysis could not be performed on the NKI-RS and MPI Leipzig cohorts because NIH toolbox behavioral data were not collected from participants in these cohorts.

In sum, these results demonstrate that stDNN together with IG procedures, which capture dynamic brain characteristics and their importance to sex differences classification, identifies sex-specific brain features that are differentially predictive of cognitive profiles in females and males.

Conventional Approaches Fail to Uncover Sex-Specific Neurobiological Predictors of Cognition. Finally, we found that conventional approaches using static functional connectivity measures as brain features failed to uncover sex-specific neurobiological predictors of cognition but instead revealed sex-invariant brain features underlying individual differences in cognition (*SI Appendix*, Fig. S8 and *Supplementary Results*).

Discussion

We examined sex differences in functional brain organization leveraging a deep neural network applied to rsfMRI data. Our approach marks a significant departure from traditional methods by directly learning latent brain dynamics from raw rsfMRI time-series data, bypassing the need for pre-engineered features like interregional functional connectivity. An innovative data augmentation strategy allowed us to train a deeper neural network model (55, 68) (*SI Appendix*) which distinguished between female and male brains with high accuracy, replicability, and generalizability across multiple sessions within the same individuals and three independent cohorts of young adults. Our findings provide strong reproducible evidence for differences in how female and male brains are intrinsically organized. Furthermore, our analysis uncovered both sex-independent and sex-specific differences in the relationship between functional brain organization and cognition. Our study advances understanding of sex differences in human brain functioning and their relation to behavior.

The first goal of our study was to investigate whether there are reliable sex differences in the functional organization of the human brain using a stDNN model. Our stDNN model uncovered reliable sex differences with over 90% cross-validation classification accuracies, outperforming previous studies (31–34, 36–38) (*SI Appendix*, Table S1). Additionally, the narrow SD bounds observed in cross-validation classification accuracy across folds underscore the reliability of our classification. These results demonstrate that AI techniques based on latent spatiotemporal dynamic representations in DNNs can reliably uncover sex differences in the human brain.

Our second goal was to address the replicability crisis in neuroscience in the context of establishing consistent sex differences in brain organization. We sought to determine whether such differences could be replicated across multisession data from the same individuals, and then generalized to independent cohorts. We found that our stDNN model not only uncovered replicable sex differences in the human brain in multisession HCP data from the same individuals but also generalized to new data from the NKI-RS and MPI-Leipzig cohorts without any additional training of the model. To our knowledge, replication and generalization of sex differences in functional brain organization across sessions and independent cohorts have not been demonstrated before. Critically, our model outperformed previous studies in both test and independent datasets (31–34, 36–38) (see *SI Appendix*, Fig. S1 and Table S1 for a summary). It is noteworthy that the use of weaker algorithms has led to the erroneous conclusion that poor classification reflects a continuum of functional brain organization in females and males (69). Our results provide the most compelling and generalizable evidence to date, refuting this continuum hypothesis and firmly demonstrating sex differences in the functional organization of the human brain.

Our third goal was to identify neurobiologically interpretable features underlying sex differences in brain organization, assessing their stability, replicability across sessions, and generalizability across independent cohorts. Traditional DNN models, especially those applied to time-series data operate as black box models (70) which do not provide insights into the neural features driving classification. To address this, we employed an XAI approach which allowed us to pinpoint brain features linked to sex differences (*SI Appendix*, Fig. S2). This technique not only identified individualized brain features associated with sex differences but also, through consensus and cross-validation analyses, confirmed their stability, replicability, and generalizability across HCP sessions and independent NKI-RS and MPI-Leipzig cohorts.

Significantly, we found that brain features associated with the DMN most reliably distinguished between female and male brains, a finding consistent at both regional and network levels with large effect sizes ($d > 2.0$). This finding resolves previously inconsistent reports of sex differences (26, 27, 29, 37, 38, 71). Through consensus analysis, we further identified the posterior cingulate cortex, precuneus, and ventromedial prefrontal cortex nodes of the DMN as the most consistent discriminators between sexes. The DMN plays a critical role in integrating self-referential information processing and monitoring of the internal mental landscape (72, 73), including introspection, mind-wandering, and autobiographical memory retrieval (71, 72, 74). These cognitive processes may differ between females and males, potentially influencing self-regulation, beliefs, and social interactions. Sex-specific differences in the DMN may also influence how females and males recall past experiences, form self-concepts, or engage in perspective-taking. Our findings underscore the pivotal role of the DMN in elucidating sex differences in brain functionality and advance our understanding of how these differences influence various cognitive and social behaviors.

Notably, network analysis also revealed large differences in the striatum and limbic networks ($d > 1.5$). While the striatum has not been a primary focus of investigations into sex-specific differences in the functional organization of the human brain, there is a considerable evidence for sexual dimorphism in its anatomy (20, 29). The striatum is important for learning cue associations, habit formation, reinforcement learning, and reward sensitivity (75). In parallel, we also observed significant differences in the limbic network which includes, most prominently, the orbitofrontal cortex (65). The orbitofrontal cortex is involved in learning and reversal of stimulus-reinforcement

associations, and correction of behavioral responses when they are no longer appropriate because previous reinforcement contingencies have changed (76). The human orbitofrontal cortex is also implicated in representing the reward value, expected reward value, and subjective pleasantness of reinforcers (77). This link to subjective pleasantness could provide a basis for investigating the limbic network's role in sex differences in hedonic experiences.

Collectively, our findings suggest that females and males differ in how they engage dynamic functional circuits involved in both self-referential and internal mental processes, reward sensitivity, reinforcement learning, and subjective experiences of pleasure. Notably, the DMN, striatum, and limbic network are also loci of dysfunction in psychiatric disorders with female or male bias in prevalence rates, including autism, attention deficit disorders, depression, addiction, schizophrenia, and Parkinson's disease all of which have sex-specific sequelae and outcomes (78–86). Our findings may therefore offer a template for investigations of sex differences in vulnerability to individual psychiatric and neurological disorders.

The final goal of our study was to determine whether sex differences in functional brain organization predict cognitive profiles differently in females and males. Despite extensive research on the anatomical and functional basis of sex differences, the behavioral significance of brain features that differentiate between sexes has remained unclear, reflecting ongoing debates regarding sex differences in brain and behavioral measures (63, 87–92). Critically, the brain features identified by XAI that reliably distinguished functional brain organization between sexes also predicted unique cognitive profiles in females and males. These profiles were derived from a principal component analysis of a comprehensive cognitive assessment using the widely used NIH toolbox (64), revealing three key components: general intelligence, response inhibition and processing speed, and delay discounting and reward sensitivity. Although the reliability of sex differences in neurotypical behavior has been contentious (63, 87, 88, 90), clinical studies in neurodevelopmental and psychiatric disorders have consistently pointed out that males display more externalizing problems while females tend to exhibit internalizing problems (6, 7, 86). Finally, it is noteworthy that in contrast to our stDNN-based sex-specific findings, static functional connectivity identified sex-invariant, but not sex-specific, brain features predictive of cognitive profiles in both sexes. These results suggest that dynamic and static functional connectivity approaches may serve as complementary tools for the identification of sex-specific and sex-invariant brain features underlying individual differences in cognition.

Conclusions

Our study provides compelling evidence for replicable and generalizable sex differences in the functional organization of the human brain. We identified replicable and generalizable brain features within the DMN, striatum, and limbic network that differentiate between sexes. Critically, these brain features predict unique patterns of cognitive profiles in females and males, demonstrating their behavioral significance. The finding of robust functional brain features underlying sex differences has the potential to inform quantitatively precise models for investigating sex differences in psychiatric and neurological disorders. This work paves the way for more targeted and personalized approaches in both cognitive neuroscience research and clinical applications.

Materials and Methods

Study Cohorts and Participants. Given the large sample size of the HCP cohort, we used multisession resting-state fMRI and phenotypic data from the HCP as our primary cohort. We used data from two independent cohorts: the NKI-RS (40) and the MPI Leipzig Mind-Brain-Body (41) cohorts to examine the replicability and generalizability of our findings from the HCP cohort. *SI Appendix*, Table S2 shows demographic information, *SI Appendix*, Table S17 shows head motion statistics, and *SI Appendix*, Fig. S9 shows the participation selection procedure. See *SI Appendix*, *Supplementary Methods* for details.

Data Augmentation. We used a data augmentation strategy that allowed us to train the deep and generalizable stDNN model used in our study (see *SI Appendix*, *Supplementary Methods* for details). Briefly, we applied a window size of 256 with an overlap of 64 to each of the multivariate time series in the training HCP dataset. As a result, the training dataset grew from 800 to 12,000, a nearly 15-fold increase.

stDNN Model. We developed an innovative stDNN model that takes as input resting-state fMRI time series and extracts latent brain dynamics features that accurately distinguish between young adult females and males (93) (see *SI Appendix*, *Supplementary Methods* for details). Briefly, our stDNN model consists of two 1D CNN blocks for spatiotemporal input transformation, coupled with ReLu and max pool layers for feature extraction and dimensionality reduction (*SI Appendix*, Fig. S2). It also includes a "temporal averaging" operation and then a sigmoid layer for binary classification. The input to the stDNN is each subject's $N_C \times N_T$ ROI fMRI time-series matrix where $N_C = 246$ for Brainnetome Atlas, processed through layers with varying filter counts and sizes. We used a dropout layer and L2-norm regularization to prevent overfitting and used binary cross-entropy optimization, a 15-epoch training cycle, and an Adam optimizer to fine-tune the parameters.

Fivefold Cross-Validation Classification Analysis in the HCP Cohort. To prevent bias and account for low variance, we conducted a fivefold cross-validation to evaluate the performance (accuracy, macro-precision, macro-recall, macro-F1, AUC) of our stDNN model in distinguishing females from males (*SI Appendix*, Fig. S3*A*; see *SI Appendix*, *Supplementary Methods* for details). We used a stratified split procedure to ensure that our training and test samples were equally divided by sex.

Identifying Brain Features Underlying Sex Classification/Differences. We used an IG-based feature attribution approach to identify brain features that distinguished between females and males (see *SI Appendix*, *Supplementary Methods* for details).

Distinctiveness of Brain Features Underlying Sex Differences in the HCP Cohort. We evaluated the validity of brain features distinguishing females and males by measuring the similarity between IG-derived dynamic brain features in HCP session 1, which showed the best cross-session replicability. Briefly, for each individual, we computed Pearson correlations between their fingerprint and the group-level fingerprint of the same sex (r12) and opposite sex (r13), as well as between the group-level male and female fingerprints (r23). We transformed the correlations into Fisher-Z scores and used the R function diffcor.dep to determine whether r12 differs from r13, given their intercorrelation (r23) (see *SI Appendix*, *Supplementary Methods* for details).

Consensus Analysis of Brain Features Underlying Sex Differences in the HCP Cohort. Next, we performed a consensus analysis to identify brain features consistently distinguishing female from male brains, using multiple fivefold cross-validation iterations over four HCP sessions (see *SI Appendix*, *Supplementary Methods* for details). Briefly, we trained 500 stDNN models from 100 cross-validation iterations per session, applied the IG method to estimate feature attribution per brain region and time point, and then identified the top 20% features. We aggregated these features across subjects and sessions and applied a binomial test to determine the most consistent discriminators, resulting in 16 consensus maps (4 HCP model sessions × 4 HCP testing sessions; Fig. 4).

Stability Analysis of Intraindividual Brain Features Underlying Sex Differences in the HCP Cohort. We investigated the individual-level stability of brain features distinguishing females and males. Briefly, for each individual, we computed the Pearson correlation between their fingerprint in session 1 and session 2 (cross-session intraindividual similarity; r12), the average Pearson correlation between their fingerprint in session 1 and all other individuals' fingerprints in session 2 (cross-session interindividual similarity; r13), as well as the average Pearson correlation between their fingerprint in session 2 and all other individuals' fingerprints in session 2 (within-session interindividual similarity; r23). After transforming the correlations into Fisher-Z scores, we used the R function

diffcor.dep to determine whether r12 differs from r13, given their intercorrelation r23. We repeated this analysis with HCP sessions 3 and 4 for validation (see *SI Appendix, Supplementary Methods* for details).

Control Analyses with Different Brain Atlases, Artifact Reduction Methods, and Head Movement in the HCP Cohort. To validate the robustness of classification results, we tested HCP session 1 models, which showed the best cross-session replicability, against different atlases, motion-related artifact reduction methods, and head movement (see *SI Appendix, Supplementary Methods* for details). Briefly, we extracted resting-state fMRI time series based on several alternative atlases and examined the classification accuracy using stDNN and cross-validation analysis. We then examined the influence of motion and physiological noise by including motion scrubbing (94) and aCompCor (95) in our analysis. Finally, we computed the squared distance correlation ($dcor^2$) (96) between the strength of features and the mean framewise displacement in females and males separately to evaluate the impact of motion on our results.

Generalization of Sex Classification Models Trained on the HCP Cohort to Independent NKI-RS and MPI Leipzig Cohorts. We used HCP session 1–based models, which showed the best cross-session replicability, to examine the generalizability to independent cohorts. For assessing the performance of our stDNN model for independent NKI-RS and MPI Leipzig cohorts, we used each of the five stDNN models trained on different subsets of HCP session 1 data (*SI Appendix*, Fig. S3*A*; see *SI Appendix, Supplementary Methods* for details). Note that in this analysis, the stDNN models were not trained on NKI-RS or MPI Leipzig data.

Generalization of Brain Features Underlying Sex Differences from the HCP to Independent NKI-RS and MPI Leipzig Cohorts. We next examined the generalizability of discriminating features identified in HCP data to independent NKI-RS and MPI Leipzig cohorts using consensus analysis (see *SI Appendix, Supplementary Methods* for details). Briefly, for each cohort, we used 500 stDNN models trained on HCP session 1 data, applied the IG method to estimate feature attribution per brain region and time point, and then identified the top 20% features. Within each cohort, we aggregated these features across subjects and applied a binomial test to determine the most consistent discriminators.

Distinctiveness of Brain Features Underlying Sex Differences in NKI-RS and MPI Leipzig Cohorts. We evaluated the validity of brain features distinguishing females and males in NKI-RS and MPI Leipzig cohorts using the same distinctiveness analysis approach described for the HCP cohort (see *SI Appendix, Supplementary Methods* for details).

Control Analyses with Different Brain Atlases, Artifact Reduction Methods, and Head Movement in the NKI-RS and MPI Leipzig Cohorts. We used HCP session 1–based models to examine whether our classification results in the two independent cohorts are robust to the selection of atlases and motion-related artifacts reduction methods and head movement (see *SI Appendix, Supplementary Methods* for details).

Network-Level Differences in Brain Features Underlying Sex Differences. Extending our analysis of regional brain features, we examined sex differences in 20 brain networks including the 17 cortical networks (65) and three additional subcortical networks encompassing the amygdala–hippocampus, striatum, and thalamus (*SI Appendix*, Table S18). Specifically, for each of the 20 networks, we computed network attribution by averaging weighted feature attributions across all regions within the same network, and then assessed sex differences in network attribution for each network using two-sample *t* tests. We computed the effect size of sex differences in each network and ranked them based on the consistency of effect size across six datasets, including four HCP sessions and the NKI-RS and MPI Leipzig cohorts.

Generalization of Sex Differences Using Conventional Machine Learning Methods. To examine the generalizability of conventional functional connectivity approaches, we used K-Nearest Neighbor, Decision Tree, linear SVM, Logistic Regression, Ridge Classifier, LASSO, and Random Forest (66). Consistent with many prior rsfMRI studies, we used precomputed functional connectivity between the 246 brain regions as features. We trained and tested models on HCP session 1 data using a fivefold cross-validation procedure and then evaluated generalization on independent NKI-RS and MPI Leipzig cohorts without any additional training.

Sex-Specific Neurobiological Predictors of Cognition and Its Replicability. We investigated whether stDNN-identified brain features could predict cognitive profiles in females and males (see *SI Appendix, Supplementary Methods* for details). Briefly, using principal component analysis, we distilled 14 HCP cognitive measures into three components to create individual cognitive profiles. We then examined sex-specific neurobiological predictors of individual cognitive profiles for HCP session 1 using CCA and also applied the same CCA procedure for HCP session 3 to examine replicability (*SI Appendix*, Fig. S3*B*). The significance of CCA modes was assessed using dimensional reduction and nonparametric analyses. Finally, we examined whether the CCA model from one sex could predict the cognitive profile in the opposite sex.

Control Analyses Examining Sex-Specific Neurobiological Predictors of Cognition Using Static Connectivity Measures. We used the same CCA procedures and static functional connectivity as brain variables to examine brain–behavior relations in each sex and whether the CCA model from one sex could predict the cognitive profile in the opposite sex in the HCP cohort.

Data, Materials, and Software Availability. Data used in this study are available from the HCP (http://www.humanconnectomeproject.org/) (97), the Nathan Kline Institute-Rockland Sample (http://fcon_1000.projects.nitrc.org/indi/enhanced/data.html) (98), and the MPI Leipzig Mind-Brain-Body dataset (https://openneuro.org/datasets/ds000221/versions/1.0.0) (99). Code used in the analyses can be found at https://github.com/scsnl/YZ_HCP_DNN_Gender_2023 (100).

ACKNOWLEDGMENTS. This work was supported by NIH grants MH084164 (V.M.), EB022907 (V.M.), MH121069 (V.M.), K25HD074652 (S.R.), and AG072114 (K.S.); Transdisciplinary Initiative and Uytengsu-Hamilton 22q11 Programs, Stanford Maternal and Child Health Research Institute (V.M. and K.S.); and NARSAD Young Investigator Award (K.S.).

Author affiliations: [a]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305; [b]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305; [c]Stanford Institute for Human-Centered Artificial Intelligence, Stanford University, Stanford, CA 94305; and [d]Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA 94305

1. L. Cahill, Why sex matters for neuroscience. *Nat. Rev. Neurosci.* **7**, 477–484 (2006).
2. M. M. McCarthy, Multifaceted origins of sex differences in the brain. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150106 (2016).
3. R. M. Shansky, C. S. Woolley, Considering sex as a biological variable will be valuable for neuroscience research. *J. Neurosci.* **36**, 11817–11822 (2016).
4. J. A. Clayton, Studying both sexes: A guiding principle for biomedicine. *FASEB J.* **30**, 519–524 (2016).
5. M. M. McCarthy, A. P. Arnold, G. F. Ball, J. D. Blaustein, G. J. De Vries, Sex differences in the brain: The not so inconvenient truth. *J. Neurosci.* **32**, 2241–2247 (2012).
6. A. Riecher-Rossler, Sex and gender differences in mental disorders. *Lancet Psychiatry* **4**, 8–9 (2017).
7. M. Rutter, A. Caspi, T. E. Moffitt, Using sex differences in psychopathology to study causal mechanisms: Unifying issues and research strategies. *J. Child Psychol. Psychiatry* **44**, 1092–1115 (2003).
8. K. Supekar *et al.*, Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism. *Br. J. Psychiatry*, 10.1192/bjp.2022.13 (2022).
9. S. Baron-Cohen *et al.*, Why are autism spectrum conditions more prevalent in males? *PLoS Biol.* **9**, e1001081 (2011).
10. A. R. Gobinath, E. Choleris, L. A. Galea, Sex, hormones, and genotype interact to influence psychiatric disease, treatment, and behavioral research. *J. Neurosci. Res.* **95**, 50–64 (2017).
11. A. B. Arnett, B. F. Pennington, E. G. Willcutt, J. C. DeFries, R. K. Olson, Sex differences in ADHD symptom severity. *J. Child Psychol. Psychiatry* **56**, 632–639 (2015).
12. D. M. Werling, D. H. Geschwind, Sex differences in autism spectrum disorders. *Curr. Opin. Neurol.* **26**, 146–153 (2013).
13. R. E. Gur, R. G. Petty, B. I. Turetsky, R. C. Gur, Schizophrenia throughout life: Sex differences in severity and profile of symptoms. *Schizophr. Res.* **21**, 1–12 (1996).
14. E. Luders, F. Kurth, Structural differences between male and female brains. *Handb. Clin. Neurol.* **175**, 3–11 (2020).
15. C. D. Good *et al.*, Cerebral asymmetry and the effects of sex and handedness on brain structure: A voxel-based morphometric analysis of 465 normal adult human brains. *Neuroimage* **14**, 685–700 (2001).
16. S. Liu, J. Seidlitz, J. D. Blumenthal, L. S. Clasen, A. Raznahan, Integrative structural, functional, and transcriptomic analyses of sex-biased brain organization in humans. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 18788–18798 (2020).
17. E. Luders, A. W. Toga, Sex differences in brain anatomy. *Prog. Brain Res.* **186**, 3–12 (2010).

18. S. F. Witelson, H. Beresh, D. L. Kigar, Intelligence and brain size in 100 postmortem brains: Sex, lateralization and age factors. *Brain* **129**, 386–398 (2006).
19. R. C. Gur *et al.*, Sex differences in brain gray and white matter in healthy young adults: Correlations with cognitive performance. *J. Neurosci.* **19**, 4065–4072 (1999).
20. A. N. Ruigrok *et al.*, A meta-analysis of sex differences in human brain structure. *Neurosci. Biobehav. Rev.* **39**, 34–50 (2014).
21. M. Ingalhalikar *et al.*, Sex differences in the structural connectome of the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 823–828 (2014).
22. A. M. Chekroud, E. J. Ward, M. D. Rosenberg, A. J. Holmes, Patterns in the human brain mosaic discriminate males from females. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1968 (2016).
23. D. L. Feis, K. H. Brodersen, D. Y. von Cramon, E. Luders, M. Tittgemeyer, Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. *Neuroimage* **70**, 250–257 (2013).
24. J. D. Rosenblatt, Multivariate revisit to "sex beyond the genitalia". *Proc. Natl. Acad. Sci. U.S.A.* **113**, E1966–E1967 (2016).
25. D. Tomasi, N. D. Volkow, Gender differences in brain functional connectivity density. *Hum. Brain Mapp.* **33**, 849–860 (2012).
26. E. A. Allen *et al.*, A baseline for the multivariate comparison of resting-state networks. *Front. Syst. Neurosci.* **5**, 2 (2011).
27. B. B. Biswal *et al.*, Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4734–4739 (2010).
28. R. L. Bluhm *et al.*, Default mode network connectivity: Effects of age, sex, and analytic approach. *Neuroreport* **19**, 887–891 (2008).
29. S. J. Ritchie *et al.*, Sex differences in the adult human brain: Evidence from 5216 UK biobank participants. *Cereb. Cortex* **28**, 2959–2975 (2018).
30. D. Tomasi, N. D. Volkow, Laterality patterns of brain functional connectivity: Gender effects. *Cereb. Cortex* **22**, 1455–1462 (2012).
31. R. Casanova, C. T. Whitlow, B. Wagner, M. A. Espeland, J. A. Maldjian, Combining graph and machine learning methods to analyze differences in functional connectivity across sex. *Open Neuroimag. J.* **6**, 1–9 (2012).
32. R. E. Gur, R. C. Gur, Sex differences in brain and behavior in adolescence: Findings from the Philadelphia Neurodevelopmental Cohort. *Neurosci. Biobehav. Rev.* **70**, 159–170 (2016).
33. M. Leming, J. Suckling, Deep learning for sex classification in resting-state and task functional brain networks from the UK Biobank. *Neuroimage* **241**, 118409 (2021).
34. T. D. Satterthwaite *et al.*, Linked sex differences in cognition and functional connectivity in youth. *Cereb. Cortex* **25**, 2383–2394 (2015).
35. S. Shanmugan *et al.*, Sex differences in the functional topography of association networks in youth. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2110416119 (2022).
36. S. M. Smith *et al.*, Functional connectomics from resting-state fMRI. *Trends Cogn. Sci.* **17**, 666–682 (2013).
37. S. Weis *et al.*, Sex classification by resting state brain connectivity. *Cereb. Cortex* **30**, 824–835 (2020).
38. C. Zhang, C. C. Dougherty, S. A. Baum, T. White, A. M. Michael, Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. *Hum. Brain Mapp.* **39**, 1765–1776 (2018).
39. D. C. Van Essen *et al.*, The WU-Minn human connectome project: An overview. *Neuroimage* **80**, 62–79 (2013).
40. K. B. Nooner *et al.*, The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Front. Neurosci.* **6**, 152 (2012).
41. A. Babayan *et al.*, A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Sci. Data* **6**, 180308 (2019).
42. T. Brosch, R. Tam; Alzheimer's Disease Neuroimaging Initiative, "Manifold learning of brain MRIs by deep learning" in *International Conference on Medical Image Computing and Computer-Assisted Intervention,* K. Mori, I. Sakuma, Y. Sato, C. Barillot, N. Navab, Eds. (Springer, Berlin, Germany, 2013), pp. 633–640.
43. R. D. Hjelm *et al.*, Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks. *Neuroimage* **96**, 245–260 (2014).
44. S. M. Plis *et al.*, Deep learning for neuroimaging: A validation study. *Front. Neurosci.* **8**, 229 (2014).
45. H.-I. Suk, S.-W. Lee, D. Shen; Alzheimer's Disease Neuroimaging Initiative, Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *Neuroimage* **101**, 569–582 (2014).
46. A. El Gazzar, L. Cerliani, G. van Wingen, R. M. Thomas, "Simple 1-D convolutional networks for resting-state fMRI based classification in autism" in *2019 International Joint Conference on Neural Networks (IJCNN)* (IEEE, New York, USA, 2019), pp. 1–6.
47. S. Ryali *et al.*, 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 (2016).
48. S. Ryali, K. Supekar, T. Chen, V. Menon, Multivariate dynamical systems models for estimating causal interactions in fMRI. *Neuroimage* **54**, 807–823 (2011).
49. J. Taghia *et al.*, Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nat. Commun.* **9**, 1–19 (2018).
50. A. Ghorbani *et al.*, Deep learning interpretation of echocardiograms. *NPJ Digit. Med.* **3**, 1–10 (2020).
51. D. Ouyang *et al.*, Video-based AI for beat-to-beat assessment of cardiac function. *Nature* **580**, 252–256 (2020).
52. J. P. Ioannidis, Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
53. D. Szucs, J. P. Ioannidis, Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol.* **15**, e2000797 (2017).
54. G. Koppe, A. Meyer-Lindenberg, D. Durstewitz, Deep learning for small and big data in psychiatry. *Neuropsychopharmacology* **46**, 176–190 (2021).
55. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
56. D. Van De Ville, Y. Farouj, M. G. Preti, R. Liegeois, E. Amico, When makes you unique: Temporality of the human brain fingerprint. *Sci. Adv.* **7**, eabj0751 (2021).
57. E. S. Finn *et al.*, Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18**, 1664–1671 (2015).
58. M. Sundararajan, A. Taly, Q. Yan, "Axiomatic attribution for deep networks" in *International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 2017), pp. 3319–3328.
59. S. M. Lundberg, S.-I. Lee, "A unified approach to interpreting model predictions" in *Advances in Neural Information Processing Systems,* I. Guyon *et al.,* Eds. (NeurIPS Foundation, USA, 2017), pp. 4765–4774.
60. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv [Preprint] (2013). https://doi.org/10.48550/arXiv.1312.6034 (Accessed 17 January 2024).
61. J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net. arXiv [Preprint] (2014). https://doi.org/10.48550/arXiv.1412.6806 (Accessed 17 January 2024).
62. R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization" in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE Computer Society, Conference Publishing Services (CPS), New York, USA, 2017), pp. 618–626.
63. C. Gurvich, N. Thomas, J. Kulkarni, Sex differences in cognition and aging and the influence of sex hormones. *Handb. Clin. Neurol.* **175**, 103–115 (2020).
64. R. J. Hodes, T. R. Insel, S. C. Landis; NIH Blueprint for Neuroscience Research, The NIH toolbox: Setting a standard for biomedical research. *Neurology* **80**, S1 (2013).
65. B. T. Yeo *et al.*, The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
66. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
67. A. Sherry, R. K. Henson, Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *J. Pers. Assess.* **84**, 37–48 (2005).
68. C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48 (2019).
69. Y. Zhang *et al.*, The human brain is best described as being on a female/male continuum: Evidence from a neuroimaging connectivity study. *Cereb. Cortex* **31**, 3021–3033 (2021).
70. D. Durstewitz, G. Koppe, A. Meyer-Lindenberg, Deep neural networks in psychiatry. *Mol. Psychiatry* **24**, 1583–1598 (2019).
71. I. Weissman-Fogel, M. Moayedi, K. S. Taylor, G. Pope, K. D. Davis, Cognitive and default-mode resting state networks: Do male and female brains "rest" differently? *Hum. Brain Mapp.* **31**, 1713–1726 (2010).
72. M. D. Greicius, B. Krasnow, A. L. Reiss, V. Menon, Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 253–258 (2003).
73. P. Qin, G. Northoff, How is our self related to midline regions and the default-mode network? *Neuroimage* **57**, 1221–1233 (2011).
74. V. Menon, 20 years of the default mode network: A review and synthesis. *Neuron* **111**, 2469–2487 (2023).
75. A. M. Graybiel, S. T. Grafton, The striatum: Where skills and habits meet. *Cold Spring Harb. Perspect. Biol.* **7**, a021691 (2015).
76. E. T. Rolls, The functions of the orbitofrontal cortex. *Brain Cogn.* **55**, 11–29 (2004).
77. M. L. Kringelbach, The human orbitofrontal cortex: Linking reward to hedonic experience. *Nat. Rev. Neurosci.* **6**, 691–702 (2005).
78. R. H. Kaiser, J. R. Andrews-Hanna, T. D. Wager, D. A. Pizzagalli, Large-scale network dysfunction in major depressive disorder: A meta-analysis of resting-state functional connectivity. *JAMA Psychiatry* **72**, 603–611 (2015).
79. V. Menon, Large-scale brain networks and psychopathology: A unifying triple network model. *Trends Cogn. Sci.* **15**, 483–506 (2011).
80. V. Menon, The triple network model, insight, and large-scale brain organization in Autism. *Biol. Psychiatry* **84**, 236–238 (2018).
81. A. Padmanabhan, C. J. Lynch, M. Schaer, V. Menon, The default mode network in Autism. *Biol. Psychiatry Cogn. Neurosci. Neuroimag.* **2**, 476–486 (2017).
82. Y. I. Sheline *et al.*, The default mode network and self-referential processes in depression. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 1942–1947 (2009).
83. T. J. Shors, E. M. Millon, H. Y. Chang, R. L. Olson, B. L. Alderman, Do sex differences in rumination explain sex differences in depression? *J. Neurosci. Res.* **95**, 711–718 (2017).
84. K. Supekar, W. Cai, R. Krishnadas, L. Palaniyappan, V. Menon, Dysregulated brain dynamics in a triple-network saliency model of schizophrenia and its relation to psychosis. *Biol. Psychiatry* **85**, 60–69 (2019).
85. S. Whitfield-Gabrieli, J. M. Ford, Default mode network activity and connectivity in psychopathology. *Annu. Rev. Clin. Psychol.* **8**, 49–76 (2012).
86. T. L. Bale, Sex matters. *Neuropsychopharmacology* **44**, 1–3 (2019).
87. L. Eliot, A. Ahmed, H. Khan, J. Patel, Dump the "dimorphism": Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neurosci. Biobehav. Rev.* **125**, 667–697 (2021).
88. A. Gaillard, D. J. Fehring, S. L. Rossell, A systematic review and meta-analysis of behavioural sex differences in executive control. *Eur. J. Neurosci.* **53**, 519–542 (2021).
89. E. D. Gennatas *et al.*, Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood. *J. Neurosci.* **37**, 5065–5073 (2017).
90. N. M. Grissom, T. M. Reyes, Let's call the whole thing off: Evaluating gender and sex differences in executive function. *Neuropsychopharmacology* **44**, 86–96 (2019).
91. R. C. Gur, R. E. Gur, Complementarity of sex differences in brain and behavior: From laterality to multimodal neuroimaging. *J. Neurosci. Res.* **95**, 189–199 (2017).
92. D. F. Halpern *et al.*, The science of sex differences in science and mathematics. *Psychol. Sci. Public Interest* **8**, 1–51 (2007).
93. C. Davatzikos, Machine learning in neuroimaging: Progress and challenges. *Neuroimage* **197**, 652–656 (2019).
94. J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
95. Y. Behzadi, K. Restom, J. Liau, T. T. Liu, A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
96. G. Székely, M. Rizzo, N. Bakirov, Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794 (2007).
97. D. C. Van Essen *et al.*, Data from "The Human Connectome Project: a data acquisition perspective." *Neuroimage* **62**, 2222–2231 (2012).
98. K. B. Nooner *et al.*, Data from "The NKI-Rockland Sample: A Model for Accelerating the Pace of Discovery Science in Psychiatry." *Front Neurosci.* **6**, 152 (2012).
99. A. Babayan *et al.*, Data from "A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults." *Sci. Data* **6**, 180308 (2019).
100. S. Ryali, Y. Zhang, C. de los Angeles, K. Supekar, V. Menon, Code from "Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization". GitHub. https://github.com/scsnl/YZ_HCP_DNN_Gender_2023. Deposited 1 March 2023.