PNAS

Supporting Information for
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], Vinod Menon[a,b,c,d,2]

[1]Equal contribution

[2]Vinod Menon
Email: menon@stanford.edu

This PDF file includes:

I. SI Methods

Study cohorts and participants

HCP. We leveraged multisession neuroimaging and behavioral data from the Human Connectome Project (HCP; http://www.humanconnectomeproject.org/) (1). For each individual in the HCP cohort, the resting state functional magnetic resonance imaging (rsfMRI) data were acquired in four runs of approximately 15 minutes each, two runs in one session and two in another session. The two sessions were collected within 1-3 days of each other. Within each session, oblique axial acquisitions alternated between phase encoding in a left-to-right (LR) direction in one run and phase encoding in a right-to-left (RL) direction in the other run. In the current study, we refer to the first run with LR encoding direction as “HCP Session 1”, the second run with LR encoding direction as “HCP Session 2”, the first run with RL encoding direction as “HCP Session 3”, and the second run with RL encoding direction as “HCP Session 4” (see also **Table S2**). Participant selection procedure is illustrated in **Fig. S9.** 499 males and 589 females were included in this study and did not differ in head movement (**Table S17**). **Table S2** shows the demographic information.

NKI-RS. An independent cohort from the Nathan Kline Institute-Rockland Sample (NKI-RS) (2) was used to investigate the replicability and generalizability of our findings from the HCP cohort. Participant selection procedure is illustrated in **Fig. S9.** 97 males and 108 females were included in this study and did not differ in head movement (**Table S17**). **Table S2** shows the demographic information.

MPI Leipzig. An independent cohort from the publicly available Max Planck Institut (MPI) Leipzig Mind-Brain-Body Dataset (https://openneuro.org/datasets/ds000221/versions/1.0.0) (3) was used to investigate the replicability and generalizability of our findings from the HCP cohort. Participant selection procedure is illustrated in **Fig. S9.** 137 males and 78 females were included in this study and did not differ in head movement (**Table S17**). **Table S2** shows the demographic information.

fMRI preprocessing

All functional MRI data were preprocessed by using SPM12 software package, as well as in-house MATLAB scripts. Structural MRI images were segmented into grey matter, white matter (WM), and cerebrospinal fluid (CSF). Prior to preprocessing, QA of functional and structural MRI was performed and subjects with poor quality imaging data were excluded from analysis. Resting-state functional MRI (fMRI) data were realigned to the averaged time frame to correct for head motion, slice-time corrected to the first slice, and co-registered to each participant’s T1-weighted images. The functional images were then normalized to the standard Montreal Neurological Institute (MNI152) template at 2mm^3. A 6-mm Gaussian kernel was used to spatially smooth the functional images and a band-pass filter ranging from 0.01 to 0.1 Hz was applied. Band-pass filtering of fMRI timeseries was used to remove low frequency artifacts such as scanner drifts and high frequency components, which do not contain useful information. Critically, band-pass filtering does not remove non-stationarities in the data, and non-stationarities such as time-varying means and covariances can still exist in a band pass filtered signal. To account for artifacts from motion and nonneural sources, the mean timeseries from each of the CSF and WM masks as well as 6 motion parameters, obtained by rigid body registration, were regressed out from the fMRI data. We used the binarized WM and CSF tissue probability maps provided by FSL.

Data input into the stDNN

We used the Brainnetome Atlas (246 regions) (4) and computed the average resting-state fMRI timeseries across the voxels in a given region of interest (ROI). Each participant's timeseries data was represented by a matrix of size $\mathrm{N_C} \times \mathrm{N_T}$, where $\mathrm{N_C}$ is the number of channels or ROIs and $\mathrm{N_T}$ is the number of time points. We trained our stDNN models on the HCP cohort which has a TR of 0.72 seconds. While the NKI-RS cohort has a TR of 0.645 seconds which is close to that of HCP, the MPI Leipzig cohort has a much slower TR which is 1.4 seconds. Thus, we linearly interpolated timeseries of the MPI Leipzig cohort using interp1d function from the python SciPy package to match the TR of HCP when evaluating performance and generating individual brain fingerprints for the MPI Leipzig cohort. We used Brainnetome as it provides fine-grained brain-wide parcellations of both cortical and subcortical areas with better anatomical and functional interpretability than most other atlases. Critically, the Brainnetome Atlas is one of the most extensively used atlases, with over 1000 studies using it (4), enabling the comparison of our method/findings with those from extant related research work as well as those under development elsewhere.

To demonstrate that our stDNN findings are robust to atlas selection, we additionally examined eight commonly used atlases (covering a broad range of number of cortical and subcortical ROIs), including Automated anatomical labeling (AAL) Atlas (90 regions) (5), Craddock Cameron Atlases (200 regions for CC200 and 392 regions for CC400) (6), Dosenbach Atlas (160 regions) (7), Eickhoff-Zilles Atlas (116 regions) (8), Glasser Atlas (360 regions) (9), Harvard-Oxford Atlas (112 regions) (10-13), and Shen Atlas (268 regions) (14).

Technical innovations of stDNN

Our stDNN model incorporates several technical innovations. First, we used data augmentation (15), which enabled us to increase the size of the training dataset by a factor of 15 that allowed us to train a deeper stDNN model with the potential for more accurate and generalizable models (16). Briefly, for each subject in the HCP training dataset, we divided their fMRI timeseries into multiple segments and then assigned each of the resulting segment the same label (male or female) as the original timeseries, effectively increasing the training dataset from ~800 to 12,000 - a nearly 15-fold increase (see "Data augmentation" section for details). Second, unlike fully connected networks, stDNN has comparatively fewer parameters to train, as it shares parameters across inputs in a given layer. Third, stDNN has a fully convolutional architecture which can predict class labels on test datasets having different lengths, which is typically the case with open-source rsfMRI data. Fourth, one-dimensional convolution used in stDNN exploits both spatial and temporal correlations between brain regions (see "The use of 1D convolutions as the basis for modeling 4D data" section for details). Critically, extant approaches do not exploit the dynamic spatiotemporal characteristics of brain activity which are thought to be more reliable features that distinguish between groups of individuals (17-19). The convolutional architecture of our stDNN model is also particularly well suited to brain imaging applications, which have a limited number of labelled training data of varying lengths.

The use of 1D convolutions as the basis for modeling 4D data

2D and 3D convolutional neural networks (CNNs) have traditionally been used to model spatial correlations in 2D and 3D (volumetric) images, whereas 1D CNNs, LSTMs and, more recently, transformer models have been used to model temporal correlations in timeseries data. fMRI, which is a 4D dataset, has both spatial and temporal correlations that can be modeled using a combination of 2D or 3D CNNs along the spatial dimension and LSTMs, or transformers, or 1D CNNs along the temporal dimension. However, training such models is very challenging due to the large number of model parameters associated with such a model, especially when using a small number of labelled datasets, as is typical with fMRI datasets. Additionally, fMRI datasets have long range spatial correlations that are not modeled by either 2D or 3D CNNs.

To address these challenges, we developed a novel architecture which exploits the spatial smoothness in a 4D fMRI dataset and represents it as a multivariate timeseries thereby reducing the number of free parameters that need to be learned in the model. The resulting multivariate timeseries can be modelled by either LSTMs or transformers or 1D CNN models. However, LSTMs are difficult to train because of vanishing and exploding gradients, whereas transformers require a large number of labelled data due to the large number of model parameters. Furthermore, previous studies have shown that fMRI timeseries exhibit strong short-term temporal correlations but weak long-term temporal correlations. Therefore, we chose to use 1D CNNs because they effectively model short-term temporal correlations and they are very easy to train using a small number of labelled data. As noted earlier, fMRI datasets exhibit long range spatial correlations in addition to strong short-term temporal correlations. 1D CNNs model these long range spatial correlations because, in 1D CNNs, the kernel size of K is a matrix of size $N \times K$ weights, where N is the number of ROIs. Therefore, with multiple 1D CNN filters, the spatial dimension also gets transformed which accounts for spatial correlations. Importantly, the high classification accuracies we obtained in cross-validation as well as in multiple independent cohorts using this model suggest that our parsimonious 1D CNN approach effectively models the spatiotemporal characteristics without losing spatial information in the 4D fMRI data.

Data augmentation

We used a data augmentation strategy that allowed us to increase the number of layers in the stDNN model (see section stDNN model below for details). Deep learning models, in general, require a large number of labelled data to be effectively trained. Here we have about 800 subjects to train the model, which is not sufficient to train a deep model. Data augmentation is typically used to address this issue and increase the number of labelled data. In computer vision applications, data augmentation consists of operations such as rotation, shift, and blurring of images and all of these operations are given the same label, thereby increasing the size of the training dataset. In our stDNN model, the data is a multivariate fMRI timeseries for which we cannot use the same operations as in the case of images. fMRI timeseries exhibit strong local temporal correlations and weak long term correlations. Taking advantage of this, we used a sliding window approach to create additional labelled datasets for training. More specifically, we apply a window size of 256 (184 secs) with an overlap of 64 (46.08 secs) to each of the multivariate timeseries. Data from each of these windows would get the same label (male or female) as the original timeseries. This data augmentation procedure was only applied to participants in the training dataset, not to the participants in the test dataset. Thus the training dataset grew from 800 to 12,000, a nearly 15-fold increase, which is critical for training the deep and generalizable stDNN model used in our study. Furthermore, we were able to train the model with 256 time samples and test it on the full timeseries (1200 samples for HCP) on the left out test and independent cohorts using our fully convolutional model.

stDNN model

We developed an innovative stDNN model to extract informative brain dynamics features that accurately distinguish between males and females. A key advantage of our approach is that it provides a novel technique to capture latent dynamics without the need for explicit feature engineering (20). Our stDNN model consists of two 1D CNN blocks, a "temporal averaging" operation, and then a sigmoid layer for binary classification (**Fig. S2**). The first 1D CNN block consists of two 1D CNN layers and the second 1D CNN block consists of one 1D CNN layer, and each CNN layer followed by a ReLU activation. In the first 1D CNN block, the sizes of the 1D CNN layers were 256× 246 and 256× 256 respectively with a kernel size of 7. In the second 1D CNN block, the size of the 1D CNN layers were 512× 256 with a kernel size of 5. We included a "maxpool" layer with a kernel size of 4 and a stride of 2 after each of the two convolutional block layers.

The input to the stDNN was each subject's $\mathrm{N_C} \times \mathrm{N_T}$ ROI timeseries matrix where $\mathrm{N_C} = 246$ for Brainnetome Atlas. The first CNN block layer transforms the input spatiotemporally to $256 \times \mathrm{N_T}$ with a temporal kernel of size $F = 7$ and with 256 number of filters. In other words, the input was convolved with a filter of size 246×7 and this was repeated 256 times to produce an output of $256 \times \mathrm{N_T}$. The 246 dimensional input was transformed to a 256 dimensional vector by a convolutional weight matrix. Therefore, our CNN not only exploits the temporal correlations in the data but also the spatial correlations across the input ROIs. The convolution operation is mathematically defined as:

$$\tilde{\boldsymbol{x}}_{l+1}(n) = \sum_{k=0}^{K} H_l(k)\boldsymbol{x}_l(n-k)$$

Where, the output of the convolution $\tilde{\boldsymbol{x}}_{l+1}$ at the l-th layer is defined as a linear combination of the convolutional kernel weights $H_l(k)^{\prime}\ s$ and the output of the previous maxpool layer $\boldsymbol{x}_l$. This convolution operation projects the input data into multiple frequency bands. This 1D convolution layer is followed by a ReLU nonlinear operator defined as below:

$$\tilde{\boldsymbol{x}}_{l+1}(n) - \max\,(0, \tilde{\boldsymbol{x}}_{l+1}(n))$$

which results in an output of size $256 \times \mathrm{N_T}$. This nonlinear operation helps in extracting nonlinear features in the data. In modern deep learning architectures, ReLU is preferred over other nonlinear operators such as sigmoid and tanh functions. This is because ReLU, as opposed to the sigmoid and tanh functions, does not saturate the gradients and therefore the backpropagation algorithm can effectively learn the model weights. The ReLU nonlinear operation of the second 1D CNN block layer is followed by a maxpool layer with a kernel size of 4, which produces an output data of size $256 \times \mathrm{N_T}$. The "maxpool" layers help in (a) reducing the temporal dimension of the data, (b) hierarchical representation of the features, and (c) increasing the receptive field of the filter to capture the long-term correlations in the timeseries. The output of this maxpool layer is an input to the second 1D convolution block. The processing of this block is similar to the first block with a difference that the number of filters used in the 1D CNN layers is 512 with a temporal kernel size of 5. The output of this block after ReLU and the maxpool layer is $512 \times \mathrm{N_T}$. The output of the second maxpool layer is to a "temporal averaging layer". Conventionally, after the last convolutional block, the data is flattened and a fully connected layer is connected to the output sigmoid layer. The fully connected layers typically have the maximum number of parameters to be trained compared to the convolutional layers. In our model, instead of the normal flattening operation, we use a "temporal averaging layer" where we average the temporal features for each filter and therefore the number of inputs to the fully connected layer is just the number of output channels of the second convolution block layer. The advantages of averaging layer over the flattening layer are (a) the number of parameters reduced from $N_{C2} \times N_{T2}$ to N_{C2} where N_{C2} is the number of output channels of the second convolutional block layer which is 512 and N_{T2} is the temporal dimension of the output of the second "maxpool" layer, (b) with averaging layer, we can train and test fMRI timeseries with varying time lengths. Temporal averaging layer is a dimensionality reduction step in the latent space and not in the original timeseries space, so is unlikely to cause loss of significant temporal information. Varying time length is common with open-source data where the data is acquired with different data acquisition protocols. We introduce a dropout layer (= 0.5) before the linear layer to avoid overfitting during the model training process. In addition to dropout, we also use a small L2-norm regularization with a weight of 0.0001 for an additional regularization. stDNN classified participants in the two groups by minimizing the binary cross-entropy cost function. We train the model for up to 15 epochs with a stopping criterion and a learning rate of 0.0001 with a batch size of 32. An Adam optimizer starting with a running average between 0.9 and 0.999, and zero weight decay was used to estimate the stDNN model parameters (21).

Identifying brain features underlying sex classification

We used an integrated gradients (IG)-based feature attribution approach (22-26) to identify brain features that discriminated between males and females. A major problem in developing and evaluating feature attribution methods is that it is difficult to distinguish errors from the DNN model and those from feature attribution procedures. IG solves this problem by taking an approach that satisfies two fundamental axioms – sensitivity and implementation invariance(22-26). Another advantage of IG is that the gradients can be computed easily for any given network architecture. IG estimates the integral of gradients with respect to the *i-th* dimension of the input *x* along the straight-line path from a given (or random) baseline to the input as follows:

$$IG_i = (x - x^{\prime}\ ) \int_0^1 (x_i - x^{\prime}\ {}_i) \frac{\partial F(x' +\ \alpha(x - x'))}{\partial x_i} \partial\alpha$$

where, IG_i is the integrated gradient for the *i-th* component of the input *x* and x' is the baseline input for which the neural network *F* results in a neutral output. IG provides a score of how important each feature contributes to the final prediction. This approach provides insights about important features that predict the sex class label. Conventional gradient-based approaches wrongly assign zero attributions for inputs where the function is flat, even when the output of *F* for such an input is different from the baseline. IG avoids this problem by computing an average gradient along a linear path. Our IG implementation is based on the *"Captum" (https://captum.ai/docs/introduction.html)* module of *Pytorch.* The IG-derived feature importance/weights are computed at an individual level and relative to a baseline that is common across individuals, and therefore were not normalized.

Five-fold cross-validation classification analysis in the HCP cohort

To prevent bias and account for low variance, for each of the four HCP sessions, we conducted a five-fold cross-validation to evaluate the performance of our stDNN model. In the five-fold cross-validation approach, we divided the whole dataset into five different parts. We used four parts for training and validation and the fifth part as the test set. We then rotate through the whole dataset five times to select a different section as the test set during each iteration (**Fig. S3A).** For each of the five subsets, we evaluate the performance of our stDNN model individually and report the mean and standard deviation values of the key performance metrics (accuracy, macro-precision, macro-recall, macro-F1, AUC). Using the five-fold cross-validation approach, the performance for every sample from the HCP data gets accounted, which helps in assessing the effectiveness of the model more robustly instead of just reporting the performance on one-time random split of the data.

Moreover, to test the replicability and generalizability of our stDNN models, we applied each of the five stDNN models trained on different subsets of a specific HCP session to the data from another HCP session without any additional training. We evaluated each model's performance independently (**Fig. S3A**) and reported the mean and standard deviation values of the key performance metrics (accuracy, macro-precision, macro-recall, macro-F1, AUC). We repeated this procedure for all pairs of HCP sessions with one as the model session and the other as the testing session.

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 integrated gradients-derived dynamic brain features in HCP Session 1, which showed the best cross-session replicability. Specifically, we first identified individual fingerprints of predictive brain features in each individual using an integrated gradients (IG) procedure (**Fig. S5**; See **Supplementary Methods** for details). Next, for each individual, we computed the Pearson correlation between their fingerprint and the group-level fingerprint of the same sex (r12), the Pearson correlation between their fingerprint and the group-level fingerprint of the opposite sex (r13), and the Pearson correlation between the group-level male fingerprint and the group-level female fingerprint (r23). Finally, we transformed the correlations into Fisher-Z scores and used the R function diffcor.dep to determine whether the correlation between two variables (r12) differs from the correlation between the first and a third one (r13), given the intercorrelation of the compared constructs (r23).

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 brains. To address this, we conducted a consensus analysis using multiple five-fold cross-validation iterations in each of the four HCP sessions. This analysis was designed to identify features unbiased by any single cross-validation split of the data. Specifically, for each HCP session, we repeated the five-fold cross-validation process 100 times, resulting in 500 stDNN models (5 folds x 100 iterations). For each of the 500 models trained on different subsets of a specific HCP session (model session), we used the IG approach to estimate feature attributions at each brain region and time point for all subjects in a specific HCP session (testing session), then computed the median of feature attributions across time points, averaged the absolute values of medians across subjects for each sex, and thresholded them to get the top 20% features for each sex. Thus, we got 500 sets of top 20% features for each sex in the HCP testing session. Finally, we counted the occurrence of each feature in the 500 sets of top 20% features for each sex, averaged the occurrences, and then thresholded them using a binomial distribution (total number of trials = 500; probability = 0.5) at p = 0.05 (Bonferroni corrected). 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**).

Stability analysis of intra-individual brain features underlying sex differences in the HCP cohort

We further investigated the stability of brain features underlying sex differences within individuals. Specifically, for each individual, we computed the Pearson correlation between their fingerprint in Session 1 and that in Session 2 (i.e., cross-session intra-individual similarity; r12), as well as the Pearson correlations between their fingerprint in Session 1 and all other individuals' fingerprints in Session 2, which were then averaged across individuals to derive cross-session inter-individual similarity (r13). We also computed the Pearson correlations between their fingerprint in Session 2 and all other individuals' fingerprints in Session 2, which were further averaged across individuals to derive within-session inter-individual similarity (r23). Next, we transformed the correlations into Fisher-Z scores and used the R function diffcor.dep to determine whether the correlation between two variables (r12) differs from the correlation between the first and a third one (r13), given the intercorrelation of the compared constructs (r23). Finally, we conducted the same stability analysis using HCP Sessions 3 and 4 to examine the replicability of our findings. We paired HCP Session 1 with Session 2 and HCP Session 3 with Session 4 based on phase encoding direction. The sessions with the same phase encoding direction were paired for analysis.

Classification analysis of independent NKI-RS and MPI Leipzig cohorts using five-fold HCP Session 1 models

As we showed the robustness of classification results within the HCP cohort (**Figs. 2 and S4**), we used only HCP Session 1, which achieved the best cross-session generalizability, to further examine the generalizability of stDNN models to independent cohorts. We applied each of the five stDNN models trained on different subsets of HCP Session 1 to the data from independent NKI-RS and MPI Leipzig cohorts without any additional training. We evaluated each model's performance on each independent cohort independently (**Fig. S3A**) and reported the mean and standard deviation values of the key performance metrics (accuracy, macro-precision, macro-recall, macro-F1, AUC) for each independent cohort.

Generalization of brain features underlying sex differences from the HCP to independent NKI-RS and MPI Leipzig cohorts

We next sought to examine the generalizability of discriminating features identified in HCP data to independent NKI-RS and MPI Leipzig cohorts using consensus analysis. Specifically, for each of the 500 models trained on different subsets of the HCP Session 1 data, we used the IG approach to estimate feature attributions at each brain region and time point for all subjects in NKI-RS and MPI Leipzig cohorts. Next, for each cohort we computed the median of feature attributions across time points, averaged the absolute values of medians across subjects for each sex, thresholded them to get the top 20% features for each sex, counted the occurrence of each feature in the 500 sets of top 20% features for each sex, averaged the occurrences, and finally thresholded them using a binomial distribution (total number of trials = 500; probability = 0.5) at $p = 0.05$ (Bonferroni corrected).

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 by measuring the similarity between integrated gradients-derived dynamic brain features. For each cohort, we conducted the same analysis. Specifically, we first identified individual fingerprints of predictive brain features in each individual using an integrated gradients (IG) procedure (**Fig. S6**; See **Supplementary Methods** for details). Next, for each individual, we computed the Pearson correlation between their fingerprint and the group-level fingerprint of the same sex (r12), the Pearson correlation between their fingerprint and the group-level fingerprint of the opposite sex (r13), and the Pearson correlation between the group-level male fingerprint and the group-level female fingerprint (r23). Finally, we transformed the correlations into Fisher-Z scores and used the R function diffcor.dep to determine whether the correlation between two variables (r12) differs from the correlation between the first and a third one (r13), given the intercorrelation of the compared constructs (r23).

Control analyses with different brain atlases, artifact reduction methods and head movement in the HCP, NKI-RS, and MPI Leipzig cohorts

We used HCP Session 1 based models, which showed the best cross-session generalizability, to further examine if our classification results are robust to the selection of atlases and motion-related artifacts reduction methods, and head movement.

First, to examine the robustness of sex classification with respect to several alternative atlases, including Automated anatomical labeling (AAL) Atlas (5), Craddock Cameron Atlases (CC200 and CC400) (6), Dosenbach Atlas (DOS160) (7), Eickhoff-Zilles Atlas (EZ) (8), Glasser Atlas (9), Harvard-Oxford Atlas (HO) (10-13), and Shen Atlas (Shen268) (14), we extracted resting-state fMRI timeseries based on each atlas and examined the classification accuracy using stDNN and cross-validation analysis.

Next, to test for the effects of nuisance variables such as head motion and sources of physiological noise, we applied an alternative pipeline that, in addition to the steps used in the main pipeline, applied motion scrubbing (27) and aCompCor (28). Motion scrubbing was done using a 0.5mm DVARS threshold, where frames exceeding the threshold were replaced with a linear interpolation of the prior and proceeding time points. To account for additional sources of physiological noise, 5 PCA components from the white matter and CSF timeseries were regressed out from the timeseries during nuisance regression.

Finally, to test for the effect of head motion on feature weights derived from the Brainnetome atlas and main pipeline, we computed the squared distance correlation ($dcor^2$) (29) between the strength of features and the mean framewise displacement (FD) in males and females separately for each of the four HCP sessions and the NKI-RS and MPI Leipzig cohorts. Briefly, $dcor^2$ is a measure of the nonlinear relationship between multidimensional variables, making it a better measure than conventional metrics like Pearson correlation, which only capture univariate linear relationships. $dcor^2$ has a range from 0 to 1, with $dcor^2 = 0$ denoting statistical independence.

Sex-specific neurobiological predictors of cognition and its replicability

We investigated whether stDNN identified brain features (that discriminated females from males) could predict cognitive function in males and females. An evaluation on the behavioral data demonstrates its suitability for factor analysis (i.e., Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.78, and $\chi^2(91) = 3480.04$, $p < 0.001$ for Bartlett's test for sphericity) whose objective is to represent a set of variables in terms of a smaller number of hypothetical variables, which would facilitate the understanding and interpretation of the data.

Principal component analysis (PCA) and exploratory factor analysis (EFA) are often referred to collectively as factor analysis. The key difference between PCA and EFA is that all of the variance (including variance unique to each variable, variance common among variables, and error variance) in the matrix is to be accounted for in PCA whereas only the variance shared with other variables (i.e., excluding variance unique to each variable and error variance) is to be accounted for in EFA. Thus, we determined that PCA is more appropriate in our situation as we do not have a hypothesis regarding the relationships among the 14 cognitive variables.

Specifically, we applied PCA with varimax rotation to the 14 HCP cognition measures, including measures of episodic memory, executive function/flexibility, executive function/inhibition, fluid intelligence, language/reading decoding, language/vocabulary comprehension, processing speed, self-regulation/impulsivity, spatial orientation, sustained attention, verbal episodic memory, and working memory. We identified three principal components and used scores on these three components to derive a cognitive profile in each individual. We then examined sex-specific neurobiological predictors of individual cognitive profiles using canonical correlation analysis (CCA). Specifically, we conducted CCA for males and females separately using brain features (i.e., feature attribution weights) from HCP session 1 as predictors of cognitive profiles to evaluate the multivariate shared relationship between the two variables sets (**Fig. S3B**). To examine the replicability of our findings, we applied the same CCA procedure using brain features from HCP session 3 as predictors of cognitive profiles. To assess the significance of CCA modes, in addition to the use of dimensional reduction analysis, we performed a permutation test by shuffling the rows (subjects) of the behavioral dataset 5000 times and re-run CCA after each permutation.

Finally, we examined whether the CCA model from males could predict cognitive profiles in females, and, conversely, whether the CCA model from females could predict cognitive profiles in males. Specifically, we applied the trained model from males to data from females, calculated the canonical correlation for mode 1 (**Fig. S3B**), and assessed its significance using permutation test by permuting the rows of behavioral canonical variate and re-computing canonical correlation for mode 1 for 5000 times. We then repeated this procedure using model from females and data from males.

Control analyses examining sex-specific neurobiological predictors of cognition using static connectivity measures

Using the same CCA procedures described above, we used functional connectivity, which is widely used in resting-state fMRI studies (30-35), as brain variables to examine brain-behavioral relations in each sex and whether the CCA model from one sex could predict the cognitive profile in the opposite sex in HCP cohort. Because the 246 brain regions involve 30,135 functional connectivity pairs, which was far higher than the number of samples, we reduced the dimensionality by applying a PCA to functional connectivity and used the first 246 principal components to keep the number of brain variables comparable to that in CCA using stDNN features.

II. SI Results

Stability analysis of intra-individual brain features underlying sex differences in the HCP cohort

We sought to determine the intra-individual stability of brain features underlying sex differences by leveraging four sessions of the HCP data. We found that for 99.5% of the individuals, brain features derived using HCP Session 1 data were most similar to features of the same individual than other individuals in HCP Session 2 ($3.80 < Zs < 14.01$, $ps < 1e\text{-}4$). Similarly, for 99.6% of the individuals, brain features derived using HCP Session 3 data were most similar to features of the same individual than other individuals in HCP Session 4 ($3.57 < Zs < 12.52$, $ps < 1e\text{-}4$). These results demonstrate that brain features underlying sex differences are stable and replicable at the individual participant level.

Control analyses with different brain atlases, artifact reduction methods and head movement in the HCP cohort

Across multiple atlases (4-14) and motion artifacts reduction techniques (27, 28), we achieved high classification accuracies (89.34 ± 1.88%), macro-precision (0.89 ± 0.02), macro-recall (0.89 ± 0.02), macro-F1 scores (0.89 ± 0.02), and AUC (0.96 ± 0.01) (**Table S4**), demonstrating that our findings are robust to the selection of atlases and motion-related artifacts reduction methods.

Additional analysis confirmed that the our findings were robust against potential confounds such as head motion. Specifically, we computed squared distance correlation (dcor^2) (29) between the strength of features and mean framewise displacement (FD, a measure of head motion) for males and females in each HCP session, and found no significant effect of head motion on the features (HCP Session 1: males: $\text{dcor}^2 = 0.125 \pm 0.004$; female: $\text{dcor}^2 = 0.103 \pm 0.002$; HCP Session 2: males: $\text{dcor}^2 = 0.010 \pm 0.003$; female: $\text{dcor}^2 = 0.097 \pm 0.004$; HCP Session 3: males: $\text{dcor}^2 = 0.107 \pm 0.005$; female: $\text{dcor}^2 = 0.094 \pm 0.004$; HCP Session 4: males: $\text{dcor}^2 = 0.127 \pm 0.008$; female: $\text{dcor}^2 = 0.098 \pm 0.002$). Briefly, dcor^2 captures non-linear relationship between multidimensional variables and ranges between 0 and 1, with 0 indicating statistical independence.

Control analyses with different brain atlases, artifact reduction methods and head movement in independent NKI-RS and MPI Leipzig cohorts

Across all atlases (4-14) and motion artifacts reduction techniques (27, 28), we achieved high classification accuracies (NKI-RS: 76.92 ± 2.70%; Leipzig: 77.78 ± 4.24%), macro-precision (NKI-RS: 0.78 ± 0.03; Leipzig: 0.78 ± 0.04), macro-recall (NKI-RS: 0.76 ± 0.03; Leipzig: 0.79 ± 0.04), macro-F1 scores (NKI-RS: 0.76 ± 0.03; Leipzig: 0.77 ± 0.04), and AUC (NKI-RS: 0.87 ± 0.04; Leipzig: 0.89 ± 0.03) in both NKI-RS (**Table S9**) and MPI Leipzig (**Table S10**) cohorts. These results demonstrated that our findings of generalization are robust to the selection of atlases and motion-related artifacts reduction methods.

Additional analysis confirmed that the our findings were robust against potential confounds such as head motion. Specifically, we computed squared distance correlation (dcor^2) (29) between the strength of features and mean framewise displacement for males and females in the NKI-RS and MPI Leipzig cohorts, and found no significant effect of head motion on the features (NKI-RS: males: $\text{dcor}^2 = 0.206 \pm 0.006$; female: $\text{dcor}^2 = 0.164 \pm 0.005$; MPI Leipzig: males: $\text{dcor}^2 = 0.144 \pm 0.002$; female: $\text{dcor}^2 = 0.174 \pm 0.002$).

Generalization of sex differences to independent cohorts with conventional machine learning methods

We examined the generalizability of conventional functional connectivity approaches using K-Nearest Neighbor, Decision Tree, linear SVM, Logistic Regression, Ridge Classifier, LASSO, and Random Forest (36). Consistent with many prior rsfMRI studies, we used pre-computed functional connectivity between the 246 brain regions as features. We trained and tested models on HCP Session 1 data using a 5-fold cross-validation procedure, and then evaluated generalization on independent NKI-RS and MPI Leipzig cohorts without any additional training.

Within HCP Session 1, conventional approaches on average achieved an accuracy of 77.91 ± 13.72%, macro-precision of 0.78 ± 0.14, macro-recall of 0.78 ± 0.14, macro-F1 score of 0.77 ± 0.14, and AUC of 0.84 ± 0.16 (**Table S11**). With unseen data from an independent NKI-RS cohort, conventional approaches on average achieved an accuracy of 70.79 ± 8.68%, macro-precision of 0.72 ± 0.09, macro-recall of 0.70 ± 0.08, macro-F1 score of 0.70 ± 0.08, and AUC of 0.78 ± 0.12 (**Table S12**). With unseen data from an independent MPI Leipzig cohort, conventional approaches on average achieved an accuracy of 68.60 ± 11.37%, macro-precision of 0.69 ± 0.10, macro-recall of 0.70 ± 0.11, macro-F1 score of 0.68 ± 0.11, and AUC of 0.75 ± 0.14 (**Table S13**).These results suggest that conventional approaches did not generalize as well to untrained data from independent cohorts as our stDNN-based approach, highlighting the novelty of our stDNN approach, which revealed replicable and generalizable sex differences without the need for ad hoc feature engineering procedures.

Conventional approaches fail to uncover sex-specific neurobiological predictors of cognition

We next used static functional connectivity measures as brain variables to examine brain-behavioral relations in each sex and whether CCA model from males or females could predict cognitive profile in the opposite sex in HCP sessions 1 and 3 separately. To reduce the dimensionality of the functional connectivity matrix we used the first 246 principal components to examine brain-behavior relations using procedures similar to the ones described in the Section ***Sex-specific neurobiological predictors of cognition and its replicability***. In both HCP sessions 1 and 3, the first 246 principal components explained about 80% (80.4% and 79.7%, respectively) of the total variance.

In HCP session1, CCA yielded three modes with squared canonical correlations (R_c^2) of 0.68, 0.56, and 0.48 in males (**Fig. S8A**). The CCA model was statistically significant (Pillai's trace = 1.716, p = 4e-4, 95% CI: 1.405 – 1.620, permutation test) and explained about 92% of the variance. We then performed a dimension reduction analysis to determine the significant modes (37). The full model (modes 1 to 3) was statistically significant (F(738, 720.96) = 1.35, p = 2e-5, 95% CI: 0.86 – 1.16) whereas modes 2 to 3 (F(490, 482) = 1.06, p = 0.25, 95% CI: 0.84 – 1.19) and mode 3 (F(244, 242) = 0.90, p = 0.79, 95% CI: 0.78 – 1.29) did not explain significant additional shared variance between brain and cognitive measures, suggesting that only mode 1 was relevant (37). Permutation test with FDR correction further confirmed a significant mode 1 (p < 0.001, permutation test, 95% CI of mode 1 R_c^2: 0.50 – 0.60). In females, CCA yielded three modes with R_c^2 of 0.62, 0.48, and 0.40 (**Fig. S8B**). Collectively, the full model across all modes was statistically significant (Pillai's trace = 1.505, p = 2e-4, 95% CI: 1.186 – 1.385, permutation test) and explained 88% 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) = 1.38, p = 1e-6, 95% CI: 0.87 – 1.14) whereas modes 2 to 3 (F(490, 654) = 1.06, p = 0.23, 95% CI: 0.85 – 1.18) and mode 3 (F(244, 328) = 0.91, p = 0.78, 95% CI: 0.79 – 1.26) did not explain statistically significant shared variance between brain and behavioral measures, suggesting that only mode 1 was relevant (37). Permutation test with FDR correction further confirmed a significant mode 1 (p < 0.001, permutation test, 95% CI of mode 1 R_c^2: 0.43 – 0.52).

We then applied the trained model from males to data from females and found a significant mode 1 with a R_c^2 of 0.06 (p = 2e-5, 95% CI: 1.6e-6 – 9.0e-3, permutation test; **Fig. S8A**). Similarly, when we applied the trained model from females to data from males, we found a significant mode 1 with R_c^2 of 0.02 (p = 8e-4, 95% CI: 1.7e-6 – 1.0e-2; **Fig. S8B**). These results indicate that the CCA model from males or females predicts the cognitive profile in the opposite sex.

To examine the replicability of our findings, we performed similar analyses in HCP session 3. In males, CCA yielded three modes with squared canonical correlations (R_c^2) of 0.69, 0.54, and 0.51 (**Fig. S8C**). The CCA model was statistically significant (Pillai's trace = 1.738, p =2e-4, 95% CI: 1.402 – 1.623, permutation test) and explained about 93% of the variance. We then performed a dimension reduction analysis to determine significant modes (37). The full model (modes 1 to 3) was statistically significant (F(738, 720.96) = 1.39, p = 4e-6, 95% CI: 0.86 – 1.16) whereas modes 2 to 3 (F(490, 482) = 1.08, p = 0.19, 95% CI: 0.84 – 1.19) and mode 3 (F(244, 242) = 1.03, p = 0.41, 95% CI: 0.78 – 1.29) did not explain significant additional shared variance between brain and cognitive measures, suggesting that only mode 1 was relevant (37). Permutation test with FDR correction further confirmed a significant mode 1 (p < 0.001, 95% CI of mode 1 R_c^2: 0.51 – 0.60). In females, CCA yielded three modes with R_c^2 of 0.62, 0.50, and 0.43 (**Fig. S8D**). Collectively, the full model across all modes was statistically significant (Pillai's trace = 1.547, p = 2e-4, 95% CI: 1.203 – 1.398, permutation test) and explained 89% 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) = 1.42, p = 2e-7, 95% CI: 0.87 – 1.14) whereas modes 2 to 3 (F(490, 654) = 1.14, p = 0.07, 95% CI: 0.85 – 1.18) and mode 3 (F(244, 328) = 0.97, p = 0.59, 95% CI: 0.79 – 1.26) did not explain statistically significant shared variance between brain and behavioral measures, suggesting that only mode 1 was relevant (37). Permutation test with FDR correction further confirmed a significant mode 1 (p < 0.001, 95% CI of mode 1 R_c^2: 0.43 – 0.52).

When applying the trained model from males to data from females, we found a significant mode 1 with a R_c^2 of 0.03 (p = 4e-4, permutation test, 95% CI of mode 1 R_c^2: 1.4e-6 – 8.4e-3; **Fig. S8C**). Similarly, when applying the trained model from females to data from males, we found a significant mode 1 with R_c^2 of 0.08 (p = 2e-4, permutation test, 95% CI of mode 1 R_c^2: 2.2e-6 – 1.0e-2; **Fig. S8D**). These results indicate that the CCA model from males or females predicts the cognitive profile in the opposite sex.

Taken together, these results demonstrate that static functional connectivity fails to uncover sex-specific neurobiological predictors of cognition, but instead identifies sex-invariant brain features that are predictive of cognitive profiles in both sexes.

III. SI Discussion

Limitations and directions for future work

While our study presents significant insights, it is important to acknowledge limitations and suggest directions for future research. First, the potential biases in our training data, influenced by factors like ethnicity or socio-economic status, may limit the generalizability of our stDNN model's findings. However, the replication of our main findings across three independent cohorts with diverse demographic profiles and multiple geographical locations provides a degree of confidence in the robustness of our results against these potential biases.

Second, our model does not encompass the entire spectrum of biological sex-related factors, such as hormonal influences, nor does it account for the complexities of gender identity. Future research could expand upon this by integrating these variables to provide a more comprehensive understanding of sex differences in brain organization.

Third, future studies should aim to delineate the precise roles of the sex-specific associations we identified between functional brain organization and cognitive profiles. This should include a focus on cognitive areas exhibiting sex differences as well as those that are consistent across sexes. Moreover, it is crucial to investigate how various factors — such as learning strategies, genetics, developmental processes, hormonal changes, and psychosocial influences — differentially shape brain organization in females and males (38-40). Understanding these aspects is key to unraveling the complex interplay between sex-specific brain organization and behavior, as highlighted in our study.

IV. SI Figures

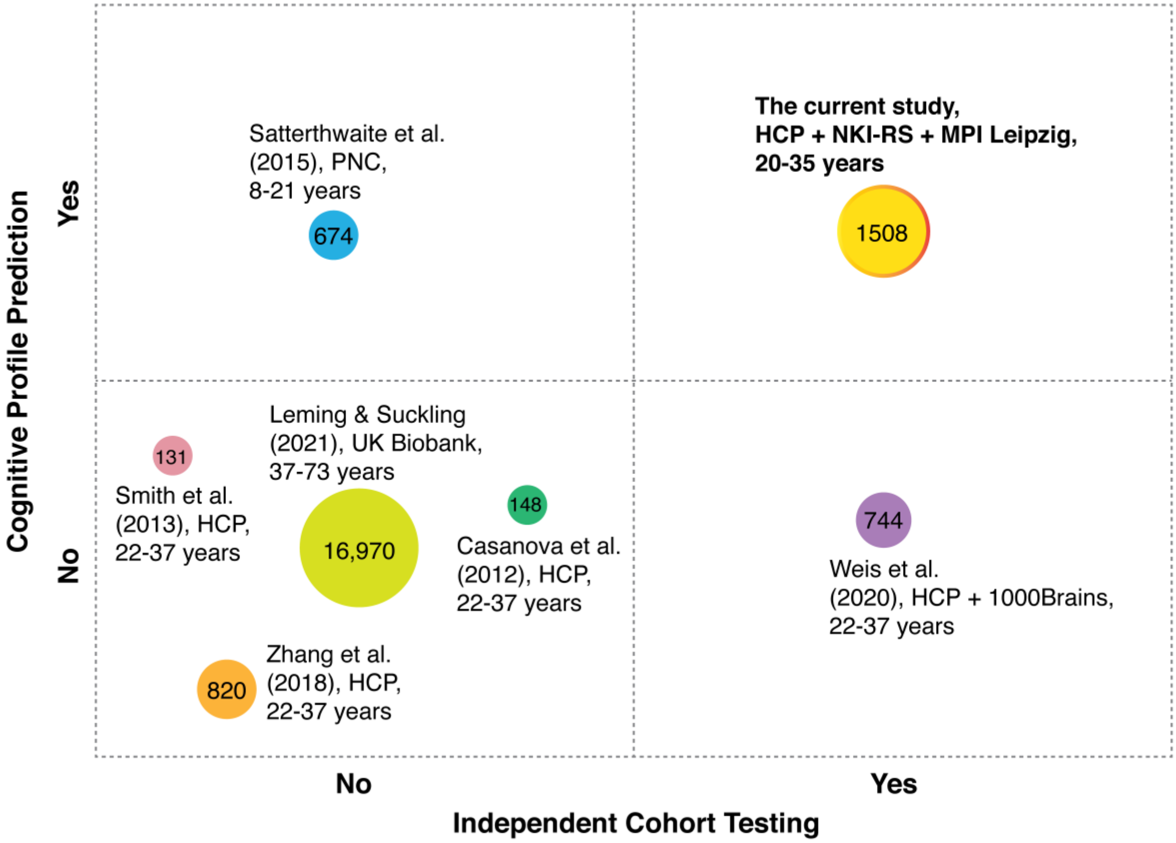

Fig. S1. Current study contrasted with previous studies (30-35) which have used machine learning and functional brain imaging data to distinguish functional organization in male and female brains. Each colored circle refers to an individual study, with size proportional to sample size. Information about each the study (i.e., authors, published year, cohort used in the study) is shown beside each circle. Studies are arranged in terms of whether independent cohort testing and whether cognitive profile prediction were examined. HCP = Human Connectome Project; PNC = Philadelphia Neurodevelopmental Cohort; NKI-RS = Nathan Kline Institute-Rockland Sample; Leipzig refers to Max Planck Institut Leipzig Mind-Brain-Body Dataset. Note that the PNC and UK Biobank datasets span wide age ranges across development, aging and psychiatric and neurological disorders (ages 8-21 and 37-73, respectively). The HCP, NKI-RS and Leipzig datasets used in our study focus on a narrow age range of young adults (ages 20-35). The current study is depicted on the top right.

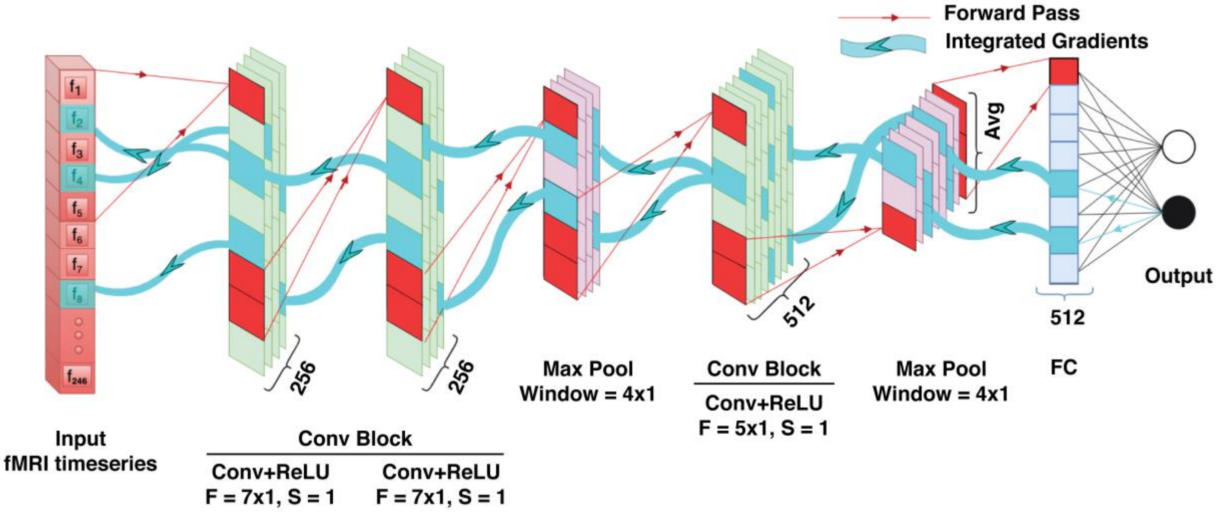

Fig. S2. Spatiotemporal deep neural network (stDNN) model for classification and integrated gradients method for feature identification. A spatiotemporal stDNN model uses raw fMRI timeseries from 246 brain regions as input to predict sex (male or female). The model predicts the class label (male or female) of each individual using spatiotemporal convolutions of the fMRI timeseries. An integrated gradients method is used to identify "black-box" brain features underlying sex classification.

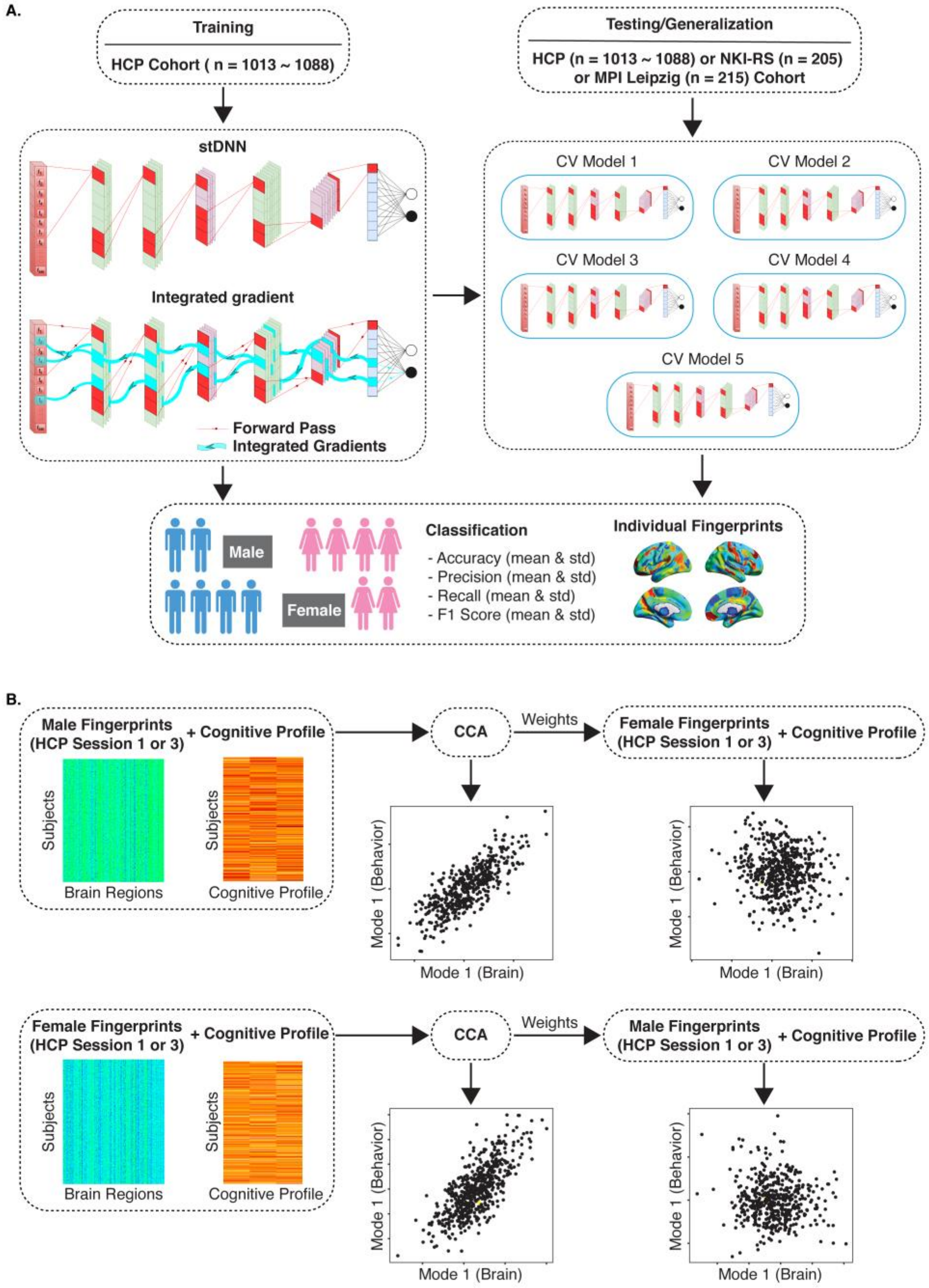

Fig. S3. Data analysis strategy. (A) Five-fold cross-validation procedures for testing and validation of sex classification using data from the HCP, NKI-RS and MPI Leipzig cohorts. The five models from a specific HCP session were then used to independently test male vs. female classification in HCP, NKI-RS and MPI Leipzig data without additional training. **(B)** The behavioral relevance of individual-specific brain features was examined using canonical correlation analysis (CCA), with separate models in males and females. Brain measures consisted of fingerprints (feature attribution maps) reflecting individual contributions to sex classification based on functional brain organization. Behavioral measures were derived from 14 cognitive tests from the NIH Toolbox.

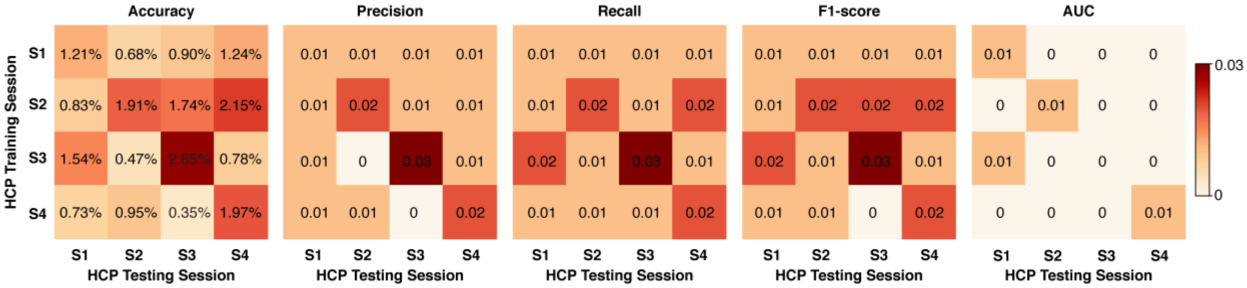

Fig. S4. Standard deviation of five-fold cross-validation classification performance within the HCP cohort. For each of the five performance metrics (accuracy, macro-precision, macro-recall, macro-F1 score, and AUC), we showed all pairwise results of standard deviation across the five folds in a matrix, 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).

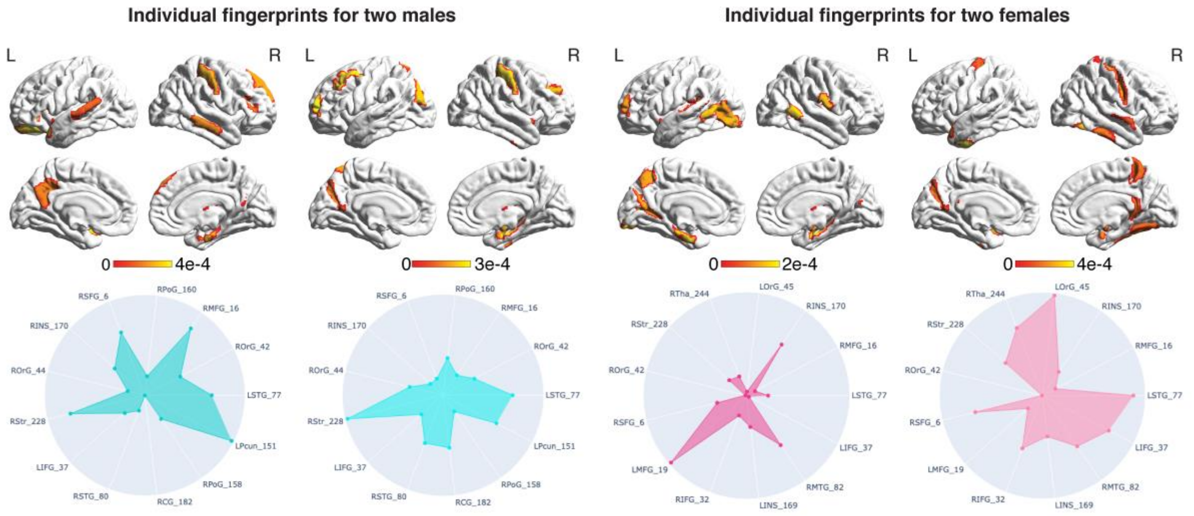

Fig. S5. Individual brain fingerprints (feature attribution maps) in the HCP cohort. stDNN-derived individual brain fingerprints in two randomly selected males and two females from HCP Sessions 1 data.

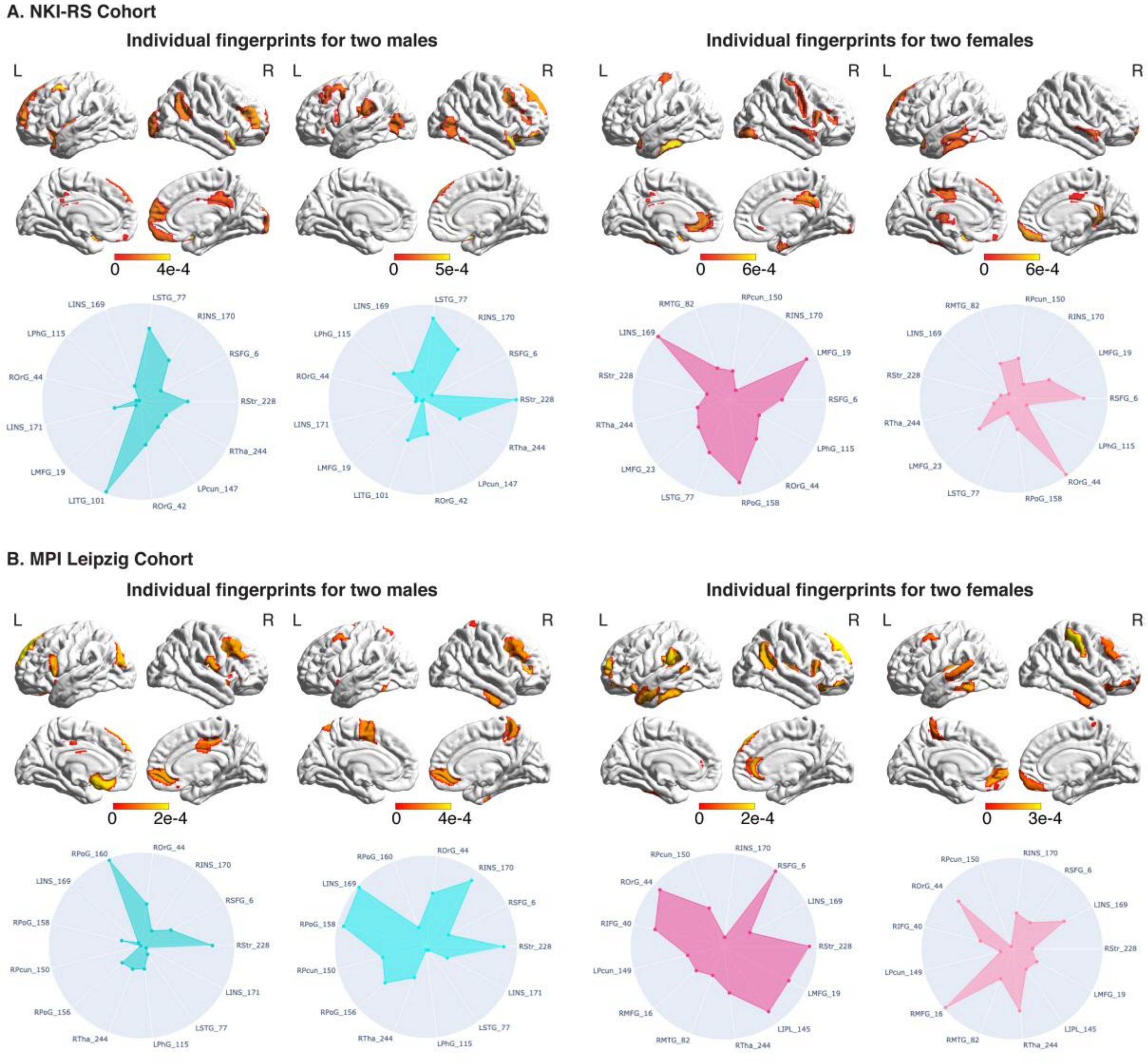

Fig. S6. **Individual brain fingerprints (feature attribution maps) in the NKI-RS and MPI Leipzig cohorts. (A)** stDNN-derived individual fingerprints from NKI-RS data. stDNN was trained with HCP Session 1 data, which generalized to NKI-RS data without any additional training. **(B)** stDNN-derived individual fingerprints from MPI Leipzig data. stDNN was trained with HCP Session 1 data, which generalized to MPI Leipzig data without any additional training.

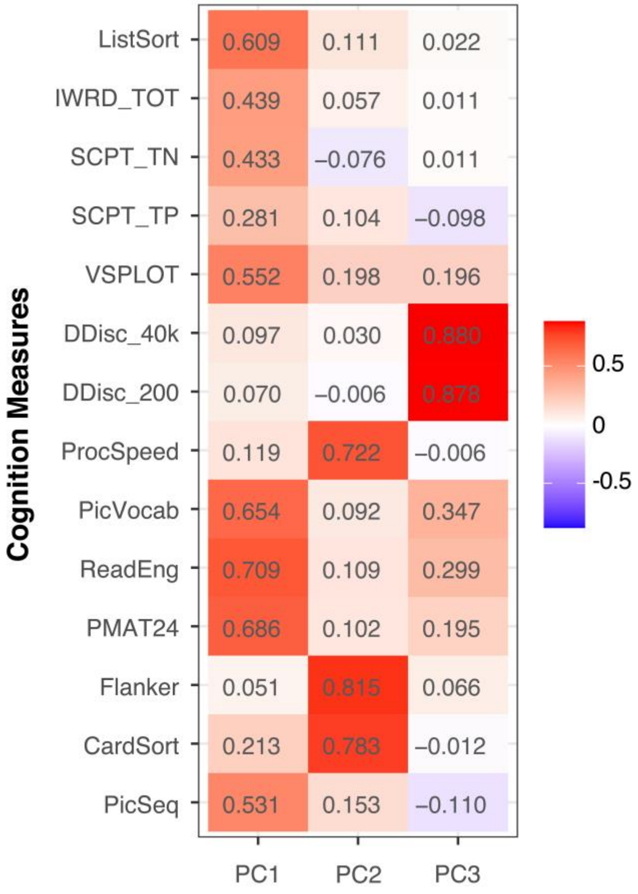

Fig. S7. **Principal component analysis of cognition measures.** Loadings of the three principal components on each of the 14 cognition measures from HCP are shown, which demonstrates that the first component is primarily associated with IQ, the second component is primarily associated with behavioral inhibition, and the third component is primarily associated with reward-related self-regulation. See also **Table S14** for details of these cognition measures.

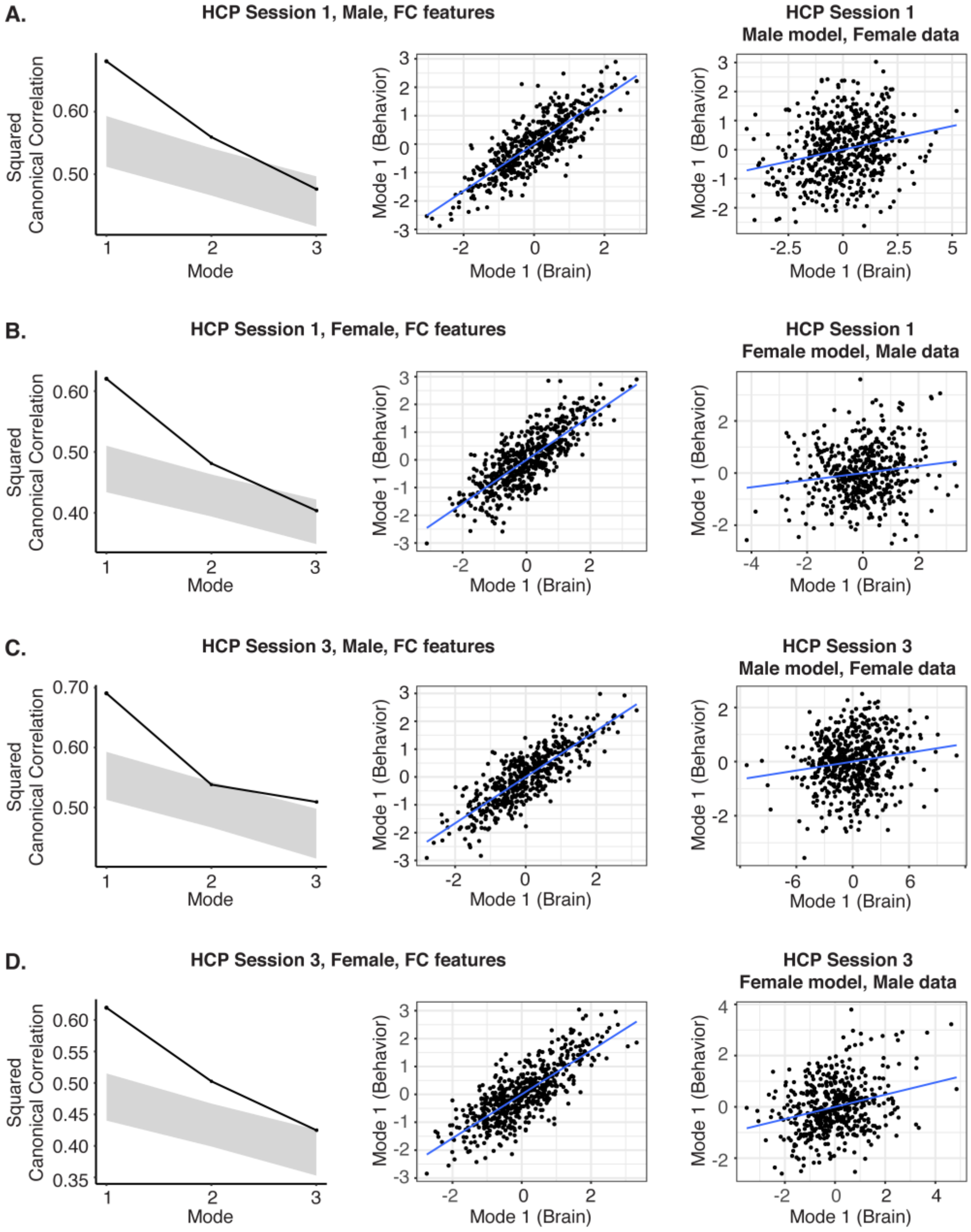

Fig. S8. Canonical correlation analysis (CCA) reveals significant sex-invariant associations between functional connectivity features and behavior. (A) CCA model from males in HCP Session 1 data predicted cognitive profiles in males and females. **(B)** CCA model from females in HCP Session 1 data predicted cognitive profiles in females and males. **(C)** CCA model from males in HCP Session 3 data predicted cognitive profiles in males and females. **(D)** CCA model from females in HCP Session 3 data predicted cognitive profiles in females and males. Line plots show squared canonical correlations, indicating the variance explained by each CCA mode. Grey area displays the 5th and 95th percentiles of the null distribution estimated via permutation testing.

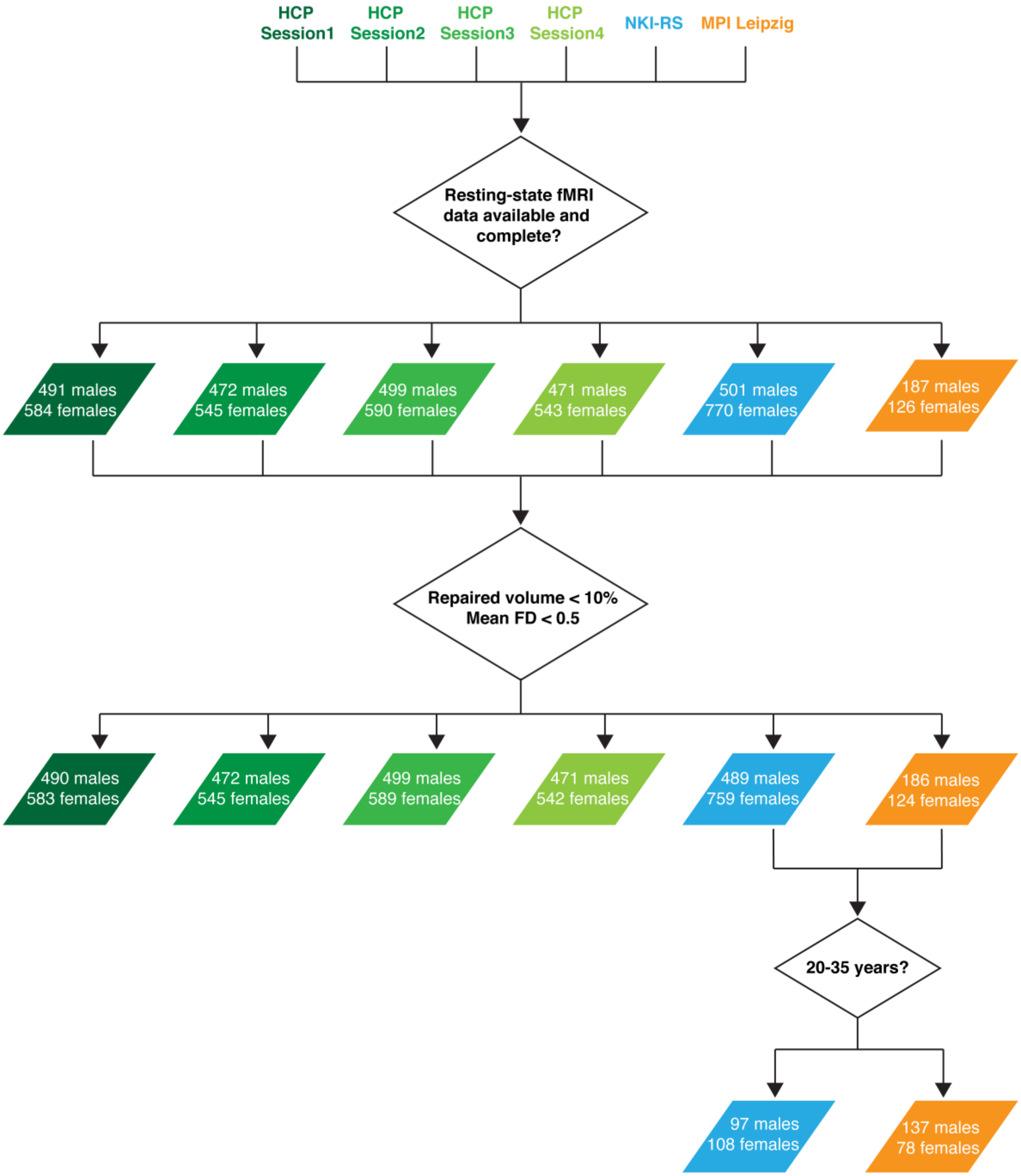

Fig. S9. Participant selection procedure for the HCP, NKI-RS, and MPI Leipzig cohorts.

V. SI Tables

Table S1. Summary of previous studies that have used machine learning and functional brain imaging data to distinguish between males from females. HCP = Human Connectome Project ; PNC = Philadelphia Neurodevelopmental Cohort.

Study	Method	Primary Cohort	Primary Cohort Result	Independent Cohort Testing	Cognitive Profile Prediction	Sex-specific relation to behavior
Weis et al. (2020) (34)	Precomputed features + SVM	HCP (ages: 22-37 years; n1 = 434, n2 = 310)	75.1%	60% (dataset: 1000Brains)	No	No
Satterthwaite et al. (2015) (32)*	Precomputed features + SVM	PNC (ages: 8-21 years; n = 674)	71%	No	Masculinity of cognitive profile associated with masculinity of brain connectivity	No
Leming & Suckling (2021) (31)*	Precomputed features + CNN	UK Biobank (ages: 37-73 years; n =16,970)	84.78%	No	No	No
Zhang et al. (2018) (35)	Precomputed features + Partial least square regression	HCP (ages: 22-37 years; n = 820)	85%	No	No	No
Smith et al. (2013) (33)	Precomputed features + multivariate classifier	HCP (ages: 22-37 years; n = 131)	87%	No	No	No
Casanova et al. (2012) (30)	Precomputed features + ensemble classifier	HCP (ages: 22-37 years; n = 148)	62.3%	No	No	No

*PNC and UK Biobank include samples of psychiatric and neurological disorders.

Table S2. Demographic information for females and males in the HCP, NKI-RS, and MPI Leipzig cohorts. LR = Left-Right; RL = Right-Left; AP = Anterior-Posterior.

Cohort	fMRI encoding direction	Sample Size	Age	Sex (female/male)
HCP Session 1	LR	1073	22-35 years	583/490
HCP Session 2	LR	1017	22-35 years	545/472
HCP Session 3	RL	1088	22-35 years	589/499
HCP Session 4	RL	1013	22-35 years	542/471
NKI-RS	AP	205	22-35 years	108/97
MPI Leipzig	AP	215	20-35 years	78/137

Note that statistical comparisons of age differences between sessions or cohorts cannot be conducted because only the age range was provided for each participant in the three cohorts.

Table S3. Brain regions whose dynamic brain connectivity features most consistently contributed to sex differences within the HCP cohort. Results from consensus analysis across 16 pairs of HCP Sessions (4 training sessions x 4 testing sessions) and 100 five-folds per paired session. Top 20% of the features whose total occurrence across all 16 paired sessions is above 4158 (i.e., Bonferroni corrected *p* < 0.05) are shown. See also **Fig. 4**. ATC = Anterior temporal cortex.

Brain Regions	Subdivision	(ID) Region Label	Count
R DLPFC	A9l	(6), SFG_R_7_3	4282.5
R DLPFC	A9/46d	(16) MFG_R_7_1	7709.5
L DLPFC	A46	(19) MFG_L_7_3	7335
L VLPFC	A8vl	(23), MFG_L_7_5	7413
R VLPFC	A45r	(36), IFG_R_6_4	5781
vmPFC	A14m	(42), OrG_R_6_1	4665
vmPFC	A12/47o	(44), OrG_R_6_2	4487
L ATC (STG)	A38l	(77), STG_L_6_5	4614.5
PCC, Precuneus	A31 (Lc1)	(153), Pcun_L_4_4	4620.5
PCC, Precuneus	A31 (Lc1)	(154), Pcun_R_4_4	5959.5
L Insula	vId/vIg	(169), INS_L_6_4	4979.5
Thalamus	cTtha	(244), Tha_R_8_7	4452

Table S4. Classification accuracy, macro-precision, macro-recall, macro-F1 score, and AUC averaged across five-folds (mean ± standardized deviation) in HCP Session 1 data for each combination of motion correction pipeline and brain atlas that were used for fMRI timeseries extraction.

Atlas	Pipeline	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	AUC
AAL	default	88.90 ± 1.18%	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
	acompcor	88.01 ± 2.29%	0.88 ± 0.02	0.88 ± 0.03	0.88 ± 0.02	0.96 ± 0.01
Brainnetome	default	90.21 ± 1.21%	0.91 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.97 ± 0.01
	acompcor	88.91 ± 1.56%	0.89 ± 0.02	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.01
CC200	default	89.64 ± 1.74%	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.97 ± 0.01
	acompcor	90.30 ± 2.03%	0.90 ± 0.02	0.90 ± 0.02	0.90 ± 0.02	0.97 ± 0.01
CC400	default	91.71 ± 1.48%	0.92 ± 0.02	0.92 ± 0.02	0.92 ± 0.02	0.98 ± 0.00
	acompcor	91.98 ± 1.28%	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.98 ± 0.01
DOS160	default	86.82 ± 1.71%	0.87 ± 0.02	0.87 ± 0.02	0.87 ± 0.02	0.94 ± 0.01
	acompcor	87.51 ± 2.00%	0.88 ± 0.02	0.88 ± 0.02	0.87 ± 0.02	0.94 ± 0.02
EZ	default	88.72 ± 1.68%	0.89 ± 0.01	0.89 ± 0.02	0.89 ± 0.02	0.96 ± 0.00
	acompcor	87.98 ± 2.66%	0.88 ± 0.03	0.88 ± 0.03	0.88 ± 0.03	0.96 ± 0.01
Glasser	default	91.24 ± 0.34%	0.91 ± 0.00	0.91 ± 0.01	0.91 ± 0.00	0.97 ± 0.00
	acompcor	90.12 ± 1.25%	0.90 ± 0.01	0.90 ± 0.01	0.90 ± 0.01	0.97 ± 0.01
HO	default	86.39 ± 3.13%	0.86 ± 0.03	0.87 ± 0.03	0.86 ± 0.03	0.94 ± 0.01
	acompcor	86.54 ± 3.71%	0.86 ± 0.04	0.87 ± 0.04	0.86 ± 0.04	0.94 ± 0.02
Shen268	default	92.17 ± 0.80%	0.92 ± 0.01	0.92 ± 0.01	0.92 ± 0.01	0.98 ± 0.00
	acompcor	90.95 ± 1.71%	0.91 ± 0.02	0.91 ± 0.02	0.91 ± 0.02	0.97 ± 0.01

Table S5. Five-fold cross-validation classification accuracy, macro-precision, macro-recall, macro-F1 score, and AUC in NKI-RS and MPI Leipzig cohorts, showing generalization of HCP Session 1 models to independent cohorts without any additional training.

Test data	Fold Number	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	AUC
NKI-RS	1	83.01%	0.83	0.83	0.82	0.90
	2	79.13%	0.81	0.79	0.78	0.88
	3	83.01%	0.83	0.83	0.83	0.91
	4	82.04%	0.82	0.82	0.82	0.89
	5	82.04%	0.83	0.81	0.81	0.90
	Avg ± Std	**81.84 ± 1.43%**	**0.83 ± 0.01**	**0.82 ± 0.02**	**0.81 ± 0.02**	**0.90 ± 0.01**
MPI Leipzig	1	84.65%	0.83	0.83	0.83	0.89
	2	80.00%	0.79	0.81	0.79	0.87
	3	84.19%	0.84	0.81	0.82	0.89
	4	82.33%	0.81	0.81	0.81	0.88
	5	81.86%	0.81	0.83	0.81	0.90
	Avg ± Std	**82.60 ± 1.68%**	**0.82 ± 0.02**	**0.82 ± 0.01**	**0.81 ± 0.01**	**0.89 ± 0.01**

Table S6. Brain regions whose dynamic brain connectivity features most consistently contributed to sex differences in NKI-RS cohort. Results from consensus analysis of NKI-RS cohort using 500 HCP Session 1 models. Top 20% of the features whose occurrence is above 289 (i.e., Bonferroni corrected $p < 0.05$) are shown. See also **Fig. 5A**.

Brain Regions	Subdivision	(ID) Region Label	Count
R DLPFC	A9l	(6), SFG_R_7_3	479.5
R DLPFC	A9/46d	(16), MFG_R_7_1	487.5
R VLPFC	IFJ	(18), MFG_R_7_2	325
L DLPFC	A46	(19), MFG_L_7_3	500
L VLPFC	A8vl	(23), MFG_L_7_5	487.5
R VLPFC	A45r	(36), IFG_R_6_4	456
R VLPFC	A44v	(40), IFG_R_6_6	488.5
vmPFC	A14m	(42), OrG_R_6_1	443.5
vmPFC	A12/47o	(44), OrG_R_6_2	489
L STG	A41/42	(71), STG_L_6_2	379
L STG	A38l	(77), STG_L_6_5	370.5
R MTG	A21c	(82), MTG_R_4_1	394.5
R MTG	A37dl	(86), MTG_R_4_3	390
L ITG	A20cv	(101), ITG_L_7_7	480.5
L PhG	A28/34 (EC)	(115), PhG_L_6_4	473.5
R PhG	TI (temporal agranular insular cortex)	(118), PhG_R_6_5	314.5
L PhG	TH (medial PPHC)	(119), PhG_L_6_6	359
PCC, Precuneus	A7m (PEp)	(147), Pcun_L_4_1	350
PCC, Precuneus	A5m (PEm)	(150), Pcun_R_4_2	453
PCC, Precuneus	dmPOS (PEr)	(151), Pcun_L_4_3	496.5
Postcentral Gyrus	A1/2/3	(158), PoG_R_4_2	312.5
L Insula	vId/vIg	(169), INS_L_6_4	445.5
R Insula	vId/vIg	(170), INS_R_6_4	327.5
L Insula	dIg	(171), INS_L_6_5	460.5
PCC	A23v	(181), CG_L_7_4	347.5
Occipital Gyrus	lsOccG	(209), sOcG_L_2_2	301
Striatum	dCa, dorsal caudate	(228), Str_R_6_5	464
Thalamus	cTtha	(244), Tha_R_8_7	357

Table S7. Brain regions whose dynamic brain connectivity features most consistently contributed to sex differences in MPI Leipzig cohort. Results from consensus analysis of MPI Leipzig cohort using 500 HCP Session 1 models. Top 20% of the features whose occurrence is above 289 (i.e., Bonferroni corrected $p < 0.05$) are shown. See also **Fig. 5B**. ATC = Anterior temporal cortex.

Brain Regions	Subdivision	(ID) Region Label	Count
R DLPFC	A9l	(6), SFG_R_7_3	488.5
R DLPFC	A9/46d	(16), MFG_R_7_1	479
R VLPFC	IFJ	(18), MFG_R_7_2	296.5
L DLPFC	A46	(19), MFG_L_7_3	498.5
L VLPFC	A8vl	(23), MFG_L_7_5	461
R VLPFC	A45r	(36), IFG_R_6_4	403.5
R VLPFC	A44v	(40), IFG_R_6_6	488
vmPFC	A14m	(42), OrG_R_6_1	351.5
vmPFC	A12/47o	(44), OrG_R_6_2	500
vmPFC	A13	(49), OrG_L_6_5	408
R ATC (STG)	A38m	(70), STG_R_6_1	355
L STG	A41/42	(71), STG_L_6_2	347.5
R MTG	A21c	(82), MTG_R_4_1	326.5
R MTG	A37dl	(86), MTG_R_4_3	411.5
L ITG	A20cv	(101), ITG_L_7_7	453
L PhG	A28/34 (EC)	(115), PhG_L_6_4	378
R PhG	TI (temporal agranular insular cortex)	(118), PhG_R_6_5	381.5
L SPL	A7pc	(129), SPL_L_5_3	332.5
PCC, Precuneus	A5m (PEm)	(149), Pcun_L_4_2	397.5
PCC, Precuneus	A5m, medial area 5(PEm)	(150), Pcun_R_4_2	466
PCC, Precuneus	dmPOS (PEr)	(151), Pcun_L_4_3	445.5
Postcentral Gyrus	A1/2/3	(158), PoG_R_4_2	390.5
L Insula	vId/vIg	(169), INS_L_6_4	422
R Insula	vId/vIg	(170), INS_R_6_4	366
L Insula	dIg	(171), INS_L_6_5	304
PCC	A23v	(181), CG_L_7_4	374
PCC	A23c	(185), CG_L_7_6	293.5
PCC	A23c	(186), CG_R_7_6	352
Occipital Gyrus	OPC	(204), OcG_R_4_3	363.5
Striatum	dCa, dorsal caudate	(228), Str_R_6_5	472

Table S8. Brain regions whose dynamic brain connectivity features most consistently contributed to sex differences across the HCP, NKI-RS, and MPI Leipzig cohorts. Top 20% of the features whose occurrence is above 4332 (i.e., Bonferroni corrected $p < 0.05$) are shown. See also **Fig. 5C**. ATC = Anterior temporal cortex.

Brain Regions	Subdivision	(ID) Region Label	Count
R DLPFC	A9l	(6), SFG_R_7_3	6493
R DLPFC	A9/46d	(16), MFG_R_7_1	8497
L DLPFC	A46	(19), MFG_L_7_3	7990
L VLPFC	A8vl	(23), MFG_L_7_5	7869
R VLPFC	A10l	(28), MFG_R_7_7	4824
R VLPFC	A45r	(36), IFG_R_6_4	6939
R VLPFC	A44v	(40), IFG_R_6_6	5315
vmPFC	A14m	(42), OrG_R_6_1	6861
vmPFC	A12/47o	(44), OrG_R_6_2	6998
R STG	A38m	(70), STG_R_6_1	4878
L STG	A38l	(77), STG_L_6_5	5979
R MTG	A37dl	(86), MTG_R_4_3	5747
L IPL	A39rv	(143), IPL_L_6_5	5183
PCC, Precuneus	A5m (PEm)	(150), Pcun_R_4_2	4612
PCC, Precuneus	dmPOS (PEr)	(151), Pcun_L_4_3	7162
PCC, Precuneus	A31 (Lc1)	(153), Pcun_L_4_4	6017
PCC, Precuneus	A31 (Lc1)	(154), Pcun_R_4_4	7376
Postcentral Gyrus	A1/2/3	(158), PoG_R_4_2	4389
L Insula	vId/vIg	(169), INS_L_6_4	5614
PCC	A23d	(175), CG_L_7_1	5071
PCC, Precuneus	cLinG	(190), Cun_R_5_1	4847
Striatum	dCa	(228), Str_R_6_5	4629
Thalamus	cTtha	(244), Tha_R_8_7	4881

Table S9. Generalization of HCP Session 1 models to NKI-RS data for each combination of motion correction pipeline and brain atlas that were used for fMRI timeseries extraction. Performance metrics include classification accuracy, macro-precision, macro-recall, macro-F1 scores, and AUC across five-folds (mean ± standardized deviation).

Atlas	Pipeline	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	AUC
AAL	default	73.84 ± 2.00%	0.74 ± 0.02	0.74 ± 0.02	0.74 ± 0.02	0.82 ± 0.01
	acompcor	76.75 ± 1.10%	0.78 ± 0.01	0.77 ± 0.01	0.76 ± 0.01	0.86 ± 0.01
Brainnetome	default	81.84 ± 1.43%	0.83 ± 0.01	0.82 ± 0.02	0.81 ± 0.02	0.90 ± 0.01
	acompcor	79.13 ± 1.47%	0.79 ± 0.01	0.79 ± 0.02	0.79 ± 0.01	0.88 ± 0.01
CC200	default	79.22 ± 3.23%	0.82 ± 0.02	0.78 ± 0.03	0.78 ± 0.04	0.91 ± 0.01
	acompcor	78.06 ± 3.35%	0.80 ± 0.03	0.77 ± 0.04	0.77 ± 0.04	0.89 ± 0.01
CC400	default	78.74 ± 3.25%	0.83 ± 0.02	0.78 ± 0.04	0.78 ± 0.04	0.92 ± 0.01
	acompcor	78.34 ± 2.42%	0.81 ± 0.02	0.78 ± 0.03	0.78 ± 0.03	0.89 ± 0.01
DOS160	default	77.38 ± 2.05%	0.78 ± 0.02	0.77 ± 0.02	0.77 ± 0.02	0.84 ± 0.01
	acompcor	74.15 ± 2.41%	0.74 ± 0.02	0.74 ± 0.02	0.74 ± 0.02	0.83 ± 0.01
EZ	default	72.62 ± 3.17%	0.75 ± 0.02	0.72 ± 0.04	0.71 ± 0.04	0.85 ± 0.01
	acompcor	78.54 ± 2.29%	0.79 ± 0.02	0.78 ± 0.02	0.78 ± 0.02	0.87 ± 0.02
Glasser	default	77.38 ± 1.78%	0.80 ± 0.02	0.76 ± 0.02	0.77 ± 0.02	0.90 ± 0.02
	acompcor	79.71 ± 1.33%	0.80 ± 0.02	0.79 ± 0.01	0.79 ± 0.01	0.88 ± 0.02
HO	default	73.01 ± 3.37%	0.73 ± 0.03	0.73 ± 0.03	0.73 ± 0.03	0.81 ± 0.02
	acompcor	72.88 ± 3.05%	0.73 ± 0.03	0.73 ± 0.03	0.73 ± 0.03	0.80 ± 0.02
Shen268	default	77.96 ± 4.81%	0.82 ± 0.02	0.77 ± 0.05	0.77 ± 0.06	0.91 ± 0.00
	acompcor	74.93 ± 1.18%	0.76 ± 0.02	0.74 ± 0.01	0.74 ± 0.01	0.86 ± 0.01

Table S10. Generalization of HCP Session 1 models to MPI Leipzig data for each combination of motion correction pipeline and brain atlas that were used for fMRI timeseries extraction. Performance metrics include classification accuracy, macro-precision, macro-recall, macro-F1 scores, and AUC across five-folds (mean ± standardized deviation).

Atlas	Pipeline	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	AUC
AAL	default	76.09 ± 3.48%	0.75 ± 0.03	0.77 ± 0.03	0.75 ± 0.03	0.87 ± 0.01
	acompcor	73.49 ± 4.97%	0.75 ± 0.03	0.76 ± 0.03	0.73 ± 0.04	0.88 ± 0.01
Brainnetome	default	82.60 ± 1.68%	0.82 ± 0.02	0.82 ± 0.01	0.81 ± 0.01	0.89 ± 0.01
	acompcor	86.23 ± 0.63%	0.86 ± 0.01	0.85 ± 0.01	0.85 ± 0.01	0.92 ± 0.01
CC200	default	75.44 ± 2.80%	0.78 ± 0.02	0.79 ± 0.03	0.75 ± 0.03	0.92 ± 0.01
	acompcor	81.67 ± 3.18%	0.82 ± 0.02	0.84 ± 0.02	0.81 ± 0.03	0.93 ± 0.00
CC400	default	73.30 ± 3.02%	0.76 ± 0.02	0.77 ± 0.02	0.73 ± 0.03	0.91 ± 0.00
	acompcor	81.49 ± 4.77%	0.82 ± 0.03	0.84 ± 0.03	0.81 ± 0.04	0.94 ± 0.00
DOS160	default	76.00 ± 1.80%	0.76 ± 0.01	0.77 ± 0.01	0.75 ± 0.01	0.85 ± 0.01
	acompcor	77.02 ± 3.95%	0.76 ± 0.03	0.78 ± 0.03	0.76 ± 0.03	0.86 ± 0.02
EZ	default	72.37 ± 6.63%	0.75 ± 0.03	0.75 ± 0.04	0.72 ± 0.06	0.88 ± 0.01
	acompcor	76.47 ± 4.55%	0.77 ± 0.02	0.79 ± 0.03	0.76 ± 0.04	0.89 ± 0.01
Glasser	default	72.28 ± 6.10%	0.75 ± 0.03	0.76 ± 0.04	0.72 ± 0.06	0.88 ± 0.01
	acompcor	82.14 ± 2.36%	0.81 ± 0.03	0.82 ± 0.02	0.81 ± 0.02	0.90 ± 0.01
HO	default	76.37 ± 4.77%	0.75 ± 0.04	0.74 ± 0.05	0.74 ± 0.05	0.83 ± 0.03
	acompcor	75.72 ± 1.77%	0.76 ± 0.02	0.71 ± 0.03	0.71 ± 0.03	0.83 ± 0.02
Shen268	default	76.93 ± 3.70%	0.78 ± 0.02	0.80 ± 0.02	0.77 ± 0.03	0.91 ± 0.01
	acompcor	84.47 ± 1.99%	0.84 ± 0.02	0.85 ± 0.02	0.84 ± 0.02	0.92 ± 0.01

Table S11. Five-fold cross-validation classification accuracies, macro-precision, macro-recall, macro-F1 scores, and AUC in HCP Session 1 data using conventional approaches with functional connectivity as features.

Classifier	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	AUC
K-Nearest Neighbor	64.86 ± 3.41%	0.64 ± 0.03	0.64 ± 0.03	0.64 ± 0.03	0.68 ± 0.03
Decision Tree	56.85 ± 2.18%	0.57 ± 0.02	0.57 ± 0.02	0.57 ± 0.02	0.57 ± 0.02
Linear SVM	88.91 ± 1.06%	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.96 ± 0.01
Logistic Regression	89.66 ± 1.69%	0.90 ± 0.02	0.90 ± 0.02	0.89 ± 0.02	0.96 ± 0.01
Ridge Classifier	89.47 ± 1.51%	0.90 ± 0.02	0.89 ± 0.01	0.89 ± 0.02	0.96 ± 0.01
LASSO	85.65 ± 1.72%	0.86 ± 0.02	0.85 ± 0.01	0.85 ± 0.02	0.94 ± 0.01
Random Forest	69.99 ± 2.46%	0.70 ± 0.02	0.69 ± 0.02	0.69 ± 0.03	0.78 ± 0.03
Average across conventional methods	**77.91 ± 13.72%**	**0.78 ± 0.14**	**0.78 ± 0.14**	**0.77 ± 0.14**	**0.84 ± 0.16**
Our stDNN model Average (see Figs. 2 and S4)	**90.21 ± 1.21%**	**0.91 ± 0.01**	**0.90 ± 0.01**	**0.90 ± 0.01**	**0.97 ± 0.01**

Table S12. Generalizability of conventional models trained on HCP Session 1 data to an independent NKI-RS cohort, with functional connectivity as features.

Classifier	Accuracy	Macro-precision	Macro-recall	Macro-F1 score	AUC
K-Nearest Neighbor	65.05 ± 1.47%	0.65 ± 0.01	0.65 ± 0.02	0.65 ± 0.02	0.69 ± 0.02
Decision Tree	55.53 ± 4.57%	0.56 ± 0.05	0.55 ± 0.04	0.55 ± 0.05	0.56 ± 0.05
Linear SVM	76.12 ± 1.32%	0.78 ± 0.01	0.75 ± 0.02	0.75 ± 0.02	0.85 ± 0.01
Logistic Regression	77.77 ± 0.84%	0.79 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.87 ± 0.01
Ridge Classifier	77.86 ± 0.95%	0.79 ± 0.01	0.77 ± 0.01	0.77 ± 0.01	0.87 ± 0.01
LASSO	76.99 ± 1.36%	0.79 ± 0.01	0.76 ± 0.01	0.76 ± 0.02	0.86 ± 0.01
Random Forest	66.21 ± 0.85%	0.66 ± 0.01	0.66 ± 0.01	0.66 ± 0.01	0.73 ± 0.02
Average across conventional methods	**70.79 ± 8.68%**	**0.72 ± 0.09**	**0.70 ± 0.08**	**0.70 ± 0.08**	**0.78 ± 0.12**
Our stDNN model Average (see Supplementary Table S5)	**81.84 ± 1.43%**	**0.83 ± 0.01**	**0.82 ± 0.02**	**0.81 ± 0.02**	**0.90 ± 0.01**

Table S13. Generalizability of conventional models trained on HCP Session 1 data to an independent MPI Leipzig cohort, with functional connectivity as features.

Classifier	**Accuracy**	**Macro-precision**	**Macro-recall**	**Macro-F1 score**	**AUC**
K-Nearest Neighbor	56.56 ± 1.37%	0.57 ± 0.01	0.58 ± 0.01	0.56 ± 0.01	0.61 ± 0.01
Decision Tree	56.56 ± 2.85%	0.56 ± 0.02	0.56 ± 0.02	0.55 ± 0.02	0.56 ± 0.02
Linear SVM	77.86 ± 1.73%	0.77 ± 0.02	0.79 ± 0.02	0.77 ± 0.02	0.86 ± 0.01
Logistic Regression	78.60 ± 1.21%	0.77 ± 0.01	0.79 ± 0.01	0.78 ± 0.01	0.86 ± 0.01
Ridge Classifier	78.79 ± 1.12%	0.78 ± 0.01	0.80 ± 0.01	0.78 ± 0.01	0.86 ± 0.01
LASSO	75.44 ± 1.70%	0.75 ± 0.02	0.77 ± 0.02	0.75 ± 0.02	0.85 ± 0.01
Random Forest	56.37 ± 3.68%	0.60 ± 0.03	0.60 ± 0.03	0.56 ± 0.04	0.65 ± 0.01
Average across conventional methods	**68.60 ± 11.37%**	**0.69 ± 0.10**	**0.70 ± 0.11**	**0.68 ± 0.11**	**0.75 ± 0.14**
Our stDNN model Average (see Supplementary Table S5)	**82.60 ± 1.68%**	**0.82 ± 0.02**	**0.82 ± 0.01**	**0.81 ± 0.01**	**0.89 ± 0.01**

Table S14. HCP cognitive function measures.

Cognition	Test	Description
Episodic Memory	Picture Sequence Memory	Participants are asked to recall increasingly lengthy series of illustrated objects and activities that are presented in a particular order on the computer screen, and are given credit for each adjacent pair of pictures they correctly place up to the maximum value for the sequence.
Executive Function/Cognitive Flexibility	Dimensional Change Card Sort	Two target pictures are presented that vary along two dimensions (e.g., shape and color). Participants are asked to match a series of bivalent test pictures (e.g., yellow balls and blue trucks) to the target pictures, first according to one dimension (e.g., color) and then, after a number of trials, according to the other dimension (e.g., shape). "Switch" trials are also employed, in which the participant must change the dimension being matched. Scoring is based on a combination of accuracy and reaction time.
Executive Function/Inhibition	Flanker Task	Participants are required to focus on a given stimulus while inhibiting attention to stimuli (fish for ages 3-7 or arrows for ages 8-85) flanking it. Sometimes the middle stimulus is pointing in the same direction as the "flankers" (congruent) and sometimes in the opposite direction (incongruent). Scoring is based on a combination of accuracy and reaction time.
Fluid Intelligence	Penn Progressive Matrices	Participants are presented with patterns made up of 2x2, 3x3 or 1x5 arrangements of squares, with one of the squares missing. The participant must pick one of five response choices that best fits the missing square on the pattern. The task has 24 items and 3 bonus items, arranged in order of increasing difficulty. However, the task discontinues if the participant makes 5 incorrect responses in a row.
Language/Reading Decoding	Oral Reading Recognition	Participants are asked to read and pronounce letters and words as accurately as possible.
Language/Vocabulary Comprehension	Picture Vocabulary	Participants are presented with an audio recording of a word and four photographic images on the computer screen and are asked to select the picture that most closely matches the meaning of the word.
Processing Speed	Pattern Completion Processing Speed	Participants are asked to discern whether two side-by-side pictures are the same or not, and their raw score is the number of items correct in a 90-second period.
Self-regulation/Impulsivity	Delay Discounting	Delay discounting describes the undervaluing of rewards that are delayed in time. It is

		illustrated by the fact that humans (and other animals) will often choose a smaller immediate reward over an objectively larger, but delayed reward. We use a version of the discounting task that identifies ‘indifference points’ at which a person is equally likely to choose a smaller reward (e.g., $100) sooner versus a larger reward later (e.g., $200 in 3 years). An adjusting-amount approach is used, in which delays are fixed and reward amounts are adjusted on a trial-by-trial basis based on participants’ choices, to rapidly hone in on indifference points.
Spatial Orientation	Variable Short Penn Line Orientation Test	Participants are shown two lines with different orientations. They have to rotate one of the lines (a moveable blue one) so that is parallel to the other line (a fixed red line).
Sustained Attention	Short Penn Continuous Performance Test	Participants see vertical and horizontal red lines flash on the computer screen. In one block, they must press the spacebar when the lines form a number and in the other block they press the spacebar when the lines form a letter.
Verbal Episodic Memory	Penn Word Memory Test	Participants are shown 20 words and asked to remember them for a subsequent memory test. They are then shown 40 words (the 20 previously presented words and 20 new words matched on memory related characteristics). They decide whether they have seen the word previously by choosing among “definitely yes,” “probably yes,” “probably no,” and “definitely no.”
Working Memory	List Sorting	In the 1-List condition, Participants are required to order a series of objects (either food or animals displayed with both a sound clip and written text that name the item) in size order from smallest to largest. In the 2-List condition, participants are presented both food and animals and are asked to report the food in size order, followed by the animals in size order.

Table S15. Overlap in brain regions whose dynamic functional circuit features predicted cognitive profiles in males in HCP Sessions 1 and 3. Only regions with sign consistent across the two sessions are included here.

Brain Regions	Subdivision	(ID) Region Label	Standardized Coefficients	
			Session 1	Session 3
R DLPFC	A9/46d	(16), MFG_R_7_1	0.253	0.172
PCC, Precuneus	A31	(153), Pcun_L_4_4	0.261	0.254
R Postcentral Gyrus	A1/2/3	(156), PoG_R_4_1	-0.309	-0.211

Table S16. Overlap in brain regions whose dynamic functional circuit features predicted cognitive profiles in females in HCP Sessions 1 and 3. Only regions with sign consistent across the two sessions are included here.

Brain Regions	**Subdivision**	**(ID) Region Label**	**Standardized Coefficients**	
			Session 1	**Session 3**
vmPFC	A12/47o	(43), OrG_L_6_2	0.198	0.202
MTG	A35/36r	(109), PhG_L_6_1	0.174	0.196
PCC, Precuneus	A31	(153), Pcun_L_4_4	-0.208	-0.207
R Postcentral Gyrus	A1/2/3	(158), PoG_R_4_2	-0.217	-0.208

Table S17. Head movement (mean scan-to-scan movement) did not differ between males and females. CI = Confidence Interval

Cohort/Session	**Mean**		*t*	*df*	*p*	*95% CI*
	Male	**Female**				
HCP Session 1	0.081	0.083	-1.239	1071	0.216	-0.002 – 0.007
HCP Session 2	0.079	0.080	-0.340	1015	0.734	-0.003 – 0.004
HCP Session 3	0.078	0.080	-1.329	1086	0.184	-0.001 – 0.006
HCP Session 4	0.081	0.083	-0.623	1011	0.533	-0.003 – 0.005
NKI-RS	0.068	0.073	-1.141	203	0.255	-0.003 – 0.012
MPI Leipzig	0.109	0.118	1.94	213	0.053	-1e-4 – 0.018

Table S18. Network Names. See also Fig. 6.

Network #	Our Terms	Yeo Terms	Full Names of Our Terms
1	VisPeri	VisPeri	Visual Peripheral
2	VisCent	VisCent	Visual Central
3	SomMot-1	SomMotA	Somato-Motor 1
4	SomMot-2	SomMotB	Somato-Motor 2
5	DorsAttn-1	DorsAttnA	Dorsal Attention 1
6	DorsAttn-2	DorsAttnB	Dorsal Attention 2
7	SalVentAttn-1	SalVentAttnA	Salience/Ventral Attention 1
8	SalVentAttn-2	SalVentAttnB	Salience/Ventral Attention 2
9	Limbic-2	LimbicB	Limbic 2
10	Limbic-1	LimbicA	Limbic 1
11	FPN-1	FPA	Frontoparietal Network 1
12	FPN-2	FPB	Frontoparietal Network 2
13	FPN-3	FPC	Frontoparietal Network 3
14	AudLang	DefaultA	Auditory Language
15	DMN-3	DefaultB	Default Mode Network 3
16	DMN-1	DefaultC	Default Mode Network 1
17	DMN-2	TempPar	Default Mode Network 2
18	AmyHip		Amygdala Hippocampus
19	Striatum		Striatum
20	Thalamus		Thalamus

VI. SI References

1. D. C. Van Essen *et al.*, The Human Connectome Project: a data acquisition perspective. *Neuroimage* **62**, 2222-2231 (2012).
2. 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).
3. 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).
4. L. Fan *et al.*, The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cereb Cortex* **26**, 3508-3526 (2016).
5. N. Tzourio-Mazoyer *et al.*, Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273-289 (2002).
6. R. C. Craddock, G. A. James, P. E. Holtzheimer, 3rd, X. P. Hu, H. S. Mayberg, A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum Brain Mapp* **33**, 1914-1928 (2012).
7. N. U. Dosenbach *et al.*, Prediction of individual brain maturity using fMRI. *Science* **329**, 1358-1361 (2010).
8. S. B. Eickhoff *et al.*, A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage* **25**, 1325-1335 (2005).
9. M. F. Glasser *et al.*, A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171-178 (2016).
10. R. S. Desikan *et al.*, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968-980 (2006).
11. J. A. Frazier *et al.*, Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *Am J Psychiatry* **162**, 1256-1265 (2005).
12. J. M. Goldstein *et al.*, Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability. *Biol Psychiatry* **61**, 935-945 (2007).
13. N. Makris *et al.*, Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophr Res* **83**, 155-171 (2006).
14. X. Shen, F. Tokoglu, X. Papademetris, R. T. Constable, Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* **82**, 403-415 (2013).
15. C. Shorten, T. M. Khoshgoftaar, A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* **6**, 1-48 (2019).
16. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436-444 (2015).
17. V. D. Calhoun, R. Miller, G. Pearlson, T. Adali, The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron* **84**, 262-274 (2014).
18. 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).
19. 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).
20. C. Davatzikos, Machine learning in neuroimaging: Progress and challenges. *Neuroimage* **197**, 652-656 (2019).
21. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* **1412** (2015).

22. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint* **arXiv:1312.6034** (2013).
23. J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net. *arXiv preprint* **arVix:1412.6806** (2014).
24. S. M. Lundberg, S. I. Lee, A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
25. R. R. Selvaraju *et al.*, Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*, 618-626 (2017).
26. M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319-3328 (2017).
27. 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).
28. 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).
29. G. Székely, M. Rizzo, N. Bakirov, Measuring and testing dependence by correlation of distances. *The annals of statistics* **35**, 2769-2794 (2007).
30. 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).
31. 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).
32. T. D. Satterthwaite *et al.*, Linked Sex Differences in Cognition and Functional Connectivity in Youth. *Cereb Cortex* **25**, 2383-2394 (2015).
33. S. M. Smith *et al.*, Functional connectomics from resting-state fMRI. *Trends Cogn Sci* **17**, 666-682 (2013).
34. S. Weis *et al.*, Sex Classification by Resting State Brain Connectivity. *Cereb Cortex* **30**, 824-835 (2020).
35. 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).
36. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *The Journal of machine Learning research* **12**, 2825-2830 (2011).
37. 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).
38. M. J. Lefner, M. I. Dejeux, M. J. Wanat, Sex Differences in Behavioral Responding and Dopamine Release during Pavlovian Learning. *eNeuro* **9** (2022).
39. C. S. Chen, E. Knep, A. Han, R. B. Ebitz, N. M. Grissom, Sex differences in learning from exploration. *Elife* **10** (2021).
40. C. S. Chen *et al.*, Divergent Strategies for Learning in Males and Females. *Curr Biol* **31**, 39-50 e34 (2021).