Connectome-based machine learning models are vulnerable to subtle data manipulations

Matthew Rosenblatt¹, Raimundo X. Rodriguez², Margaret L. Westwater³, Corey Horien², Abigail S. Greene², R. Todd Constable^{1,2,3,4}, Stephanie Noble³, Dustin Scheinost^{1,2,3,5,6,7}

1. Dept. of Biomedical Engineering, Yale School of Engineering and Applied Science
2. Interdepartmental Neuroscience Program, Yale School of Medicine
3. Dept. of Radiology & Biomedical Imaging, Yale School of Medicine
4. Dept. of Neurosurgery, Yale School of Medicine
5. Dept. Statistics & Data Science, Yale University
6. Child Study Center, Yale School of Medicine
7. Wu Tsai Institute, Yale University

Correspondence: Matthew Rosenblatt (matthew.rosenblatt@yale.edu) Magnetic Resonance Research Center, 300 Cedar St, PO Box 208043, New Haven, CT 06520-8043, USA.

Dustin Scheinost (dustin.scheinost@yale.edu) Magnetic Resonance Research Center, 300 Cedar St, PO Box 208043, New Haven, CT 06520-8043, USA.

Abstract

Neuroimaging-based predictive models continue to improve in performance, yet a widely overlooked aspect of these models is “trustworthiness,” or robustness to data manipulations. High trustworthiness is imperative for researchers to have confidence in their findings and interpretations. In this work, we used functional connectivity data to explore how minor data manipulations influence machine learning predictions. These manipulations included a novel method to falsely enhance prediction performance, adversarial noise attacks designed to degrade performance, and in-scan changes (i.e., breath-holding) that influence model accuracy. Overall, we found that predictions can be drastically changed with only minor manipulations, which demonstrates the low trustworthiness of current implementations of connectome-based pipelines. These findings highlight the need for counter-measures that improve the trustworthiness of connectome-based analysis pipelines to preserve the integrity of academic research and any potential translational applications.

1. Introduction

Human neuroimaging studies have increasingly used machine learning approaches to improve generalizability with more realistic effect sizes [1,2] and to aggregate weak informative signals occurring throughout the brain [cf. 3,4]. At the forefront of this trend are machine learning models for functional connectomes (“connectome-based models”) [5–7], which have successfully characterized the neural correlates of various clinically-relevant processes [8], including general cognitive ability [9], psychiatric disorders [7,10], affective states [11], and abstinence in individuals with substance use disorder [12]. Although improvements in accuracy [6] and fairness (i.e., race, age, or gender bias) [13,14] of connectome-based models are crucial for improving the quality of academic studies and the potential for clinical translation, accurate

and bias-free models are not enough. Connectome-based models should also have high trustworthiness, which we define as robustness to data manipulations, so researchers can have greater confidence in their findings and ensuing interpretation.

Trustworthiness has been explored from various perspectives in the machine learning literature, including privacy [15] and explainability [16]. In this work, we specifically examine trustworthiness through the lens of robustness to data manipulations [17], given evidence of some researchers in the biosciences [18–20], psychology [21,22], and clinical trials [23] manipulating their data to make results seem more favorable and publishable. A popular form of data manipulation specific to machine learning is adversarial noise (i.e., adversarial attacks), where a pattern (or “noise”) deliberately designed to trick a machine learning model is added to data to cause misclassification [24,25]. These attacks have been investigated in various contexts, including cybersecurity [26,27], image recognition [25,28], and medical imaging or recordings [29–31]. For neuroimaging, adversarial attacks may become problematic in the more distant future (e.g., in the clinic [30,32]). A more immediate concern is the potential for data manipulations to falsely enhance prediction performance in research studies. While the idea of falsely enhancing performance has not yet been explored in machine learning, we expect it to be critical in biomedical research given the history of data manipulations [18–20,23], pressure to publish high-impact works [33], and prevalence of machine learning [10,34]. Determining whether the prediction performance of connectome-based models can be falsely enhanced or diminished via data manipulations is crucial.

In this work, we investigated the trustworthiness of connectome-based predictive models. Specifically, we introduce the “performance enhancement attack” for connectome-based models, where data are injected with small, inconspicuous patterns to falsely improve the prediction performance of a specific phenotype. We also explore the effectiveness of adversarial noise attacks on connectome-based models, and how in-scan manipulations (i.e., breath-holding) may degrade model performance. In all three cases—enhancement attacks, adversarial noise attacks, and in-scan manipulations—we find that only subtle manipulations are required to drastically change predictions in five large datasets. Overall, our findings demonstrate that current implementations of connectome-based models have low trustworthiness, which points toward the need for preventative measures built in to connectome pipelines and data sharing practices.

2. Results

Functional MRI data were obtained from the Adolescent Brain Cognitive Development (ABCD) Study [35], the Human Connectome Project (HCP) [36], the Philadelphia Neurodevelopmental Cohort (PNC) [37], the Southwest University Longitudinal Imaging Multimodal (SLIM) study [38], and the UCLA Consortium for Neuropsychiatric Phenomics (UCLA-CNP) [39] (note that the SLIM and CNP datasets were only used in Sections 2.4 and 2.8, respectively). Unless otherwise specified, all analyses were conducted on resting-state data. For SLIM, we downloaded fully preprocessed functional connectomes. For the ABCD, PNC, and UCLA-CNP datasets, raw data were registered to common space as previously described [40,41]. For the HCP dataset, we

started with the minimally preprocessed data [42]. Next, standard, identical preprocessing steps were performed across all datasets using BiImage Suite [43] (see Methods). In all cases, data were parcellated into 268 nodes with the Shen atlas [44]. After excluding participants for excessive motion ($>0.2\text{mm}$) and missing nodes, 3362 individuals in the ABCD dataset, 506 individuals in the HCP dataset, 562 individuals in the PNC dataset, 445 individuals in the SLIM dataset, and 190 individuals in the UCLA-CNP dataset were retained. In the following sections, we first comprehensively characterize the effects of performance enhancement manipulations, and then evaluate adversarial noise attacks and in-scan manipulations, which decrease model accuracy (Figure S1 contains a summary).

2.1 Baseline model performance

To evaluate trust, we trained baseline regression models of fluid intelligence (IQ) and classification models of self-reported sex. These models provide a good benchmark for trustworthiness due to their wide availability in datasets and prominence in the literature [5,45–48]. For the regression models, we used ridge regression connectome-based predictive modeling (rCPM) [46] with nested 10-fold cross-validation and 10% feature selection. Regression models of IQ were evaluated by Pearson's correlation r and the cross-validated R^2 , called q^2 [49], between the measured and predicted IQ scores. We found near-zero correlations for ABCD and low correlations for HCP and PNC (see Table 1). For classification of self-reported sex, we trained both linear SVM and logistic regression models with all available features using nested 10-fold cross-validation and L_2 regularization. We also evaluated the accuracy of classifiers of self-reported sex and found relatively high success in all three datasets for both SVM and logistic regression, although SVM had higher prediction accuracy (Table 1). In the following sections, we will describe how the performance metrics in Table 1 can be drastically altered through inconspicuous manipulations to the data.

	IQ prediction		Self-reported sex classification					
	rCPM		SVM (Linear kernel)			Logistic regression		
Dataset	r	q^2	Acc	Sens	Spec	Acc	Sens	Spec
ABCD	-0.025 (0.010)	-0.031 (0.003)	0.860 (0.003)	0.850 (0.006)	0.869 (0.007)	0.805 (0.017)	0.800 (0.048)	0.809 (0.049)
HCP	0.177 (0.016)	0.031 (0.006)	0.883 (0.009)	0.859 (0.017)	0.903 (0.011)	0.767 (0.035)	0.736 (0.082)	0.794 (0.072)
PNC	0.243 (0.012)	0.058 (0.005)	0.807 (0.010)	0.743 (0.021)	0.855 (0.013)	0.711 (0.025)	0.629 (0.083)	0.774 (0.069)

Table 1. Baseline accuracies for regression models of IQ and classification models of self-reported sex in ABCD, HCP, and PNC. Prediction performance is evaluated with 10-fold cross-validation, with nested cross-validation to select L_2 regularization. The numbers in parentheses reflect the standard deviation of the metrics across 100 iterations of different random seeds.

2.2 Performance enhancement attacks are effective and nearly unnoticeable

To date, most research on data manipulations for machine learning has focused on corrupting data to decrease model accuracy [50]. However, here, we investigated the feasibility of data manipulations designed to *increase* the accuracy (which we label “enhancement attacks”) in ways that cannot be readily detected by the human eye or by changes in downstream analyses. Current open science standards would not offer protection and trustworthiness if prediction performance is enhanced but connectomes are not statistically altered.

First, we trained a model to predict IQ with resting-state connectomes in ABCD (n=3262), HCP (first session, n=506), and PNC (n=562) with rCPM [46], using 10-fold cross-validation and 10% feature selection performed by a two-sample t-test (Figure 1, top). To enhance the data for IQ prediction, we randomly selected 20% of all edges and then added an individual-specific pattern that was correlated (or anti-correlated) with each participant’s fluid intelligence score. We varied the magnitude of this pattern and repeated model training and evaluation with 10-fold cross-validation, recording changes in the correlation between measured and predicted IQ (Figure 1). Results were easily manipulated even with a low-magnitude enhancement pattern, achieving a near-perfect correlation between measured and predicted IQ scores ($r > 0.9$) for corrupted connectomes that still maintained an extremely high edge-wise correlation ($r \approx 0.99$) with their original counterparts. For the results in Figures 1-2, we selected among large regularization parameters with nested cross-validation, but using a smaller regularization parameter made enhancement attacks even more effective (Figure S2). These results suggest that minor changes to a functional connectivity matrix can undermine the trustworthiness of predictive models, even in the context of open science practices.

Figure 1. Main pipeline of performance enhancement attacks. This example is shown for prediction of IQ in the HCP dataset with resting-state connectomes and rCPM. The original dataset results in a prediction performance of $r=0.18$ between measured and predicted IQ. Enhancement patterns (mean enhancement pattern shown) are added to the original connectome proportional to each participant's z-scored IQ. For the sake of visualization, we multiplied the enhancement patterns by 120, 80, and 40, or else they would be too small to see. The corresponding enhanced connectomes maintain average correlations of $r \approx 0.99$ with the original connectomes, but the prediction performance is greatly enhanced. The networks labeled on the connectomes are: MF: Medial-Frontal, FP: Fronto-parietal, DMN: Default Mode, MOT: Motor, VI: Visual I, VII: Visual II, VAs: Visual Association, SAL: Salience, SC: Subcortical, CBL: Cerebellum [53,54].

2.3 Performance enhancement attacks preserve individuality

While the enhanced connectomes appeared almost visually identical to the original connectomes (Figure 1), we also investigated if the enhancement patterns affected other downstream analyses. If common, downstream analyses were not affected, this would make it difficult to determine if connectomes have been manipulated. First, we varied the mean absolute value of the enhancement pattern and trained rCPM models to predict IQ scores in ABCD, HCP, and PNC. As we increased the scale of the enhancement pattern, prediction performance greatly increased, with correlations between measured and predicted IQ achieving $r > 0.9$ (Figure 1, Figure 2a). Corroborating visual inspection, the correlations between edges of original and enhanced connectomes remained very high ($r \approx 0.99$) (Figure 2a, top row). In addition, a participant-wise Kolmogorov-Smirnov test [51] suggested no significant differences in edge distributions between original and enhanced data (median $p > 0.9999$).

Using the enhanced IQ data, we next trained SVM classifiers for self-reported sex and found that sex classification accuracy stayed essentially constant (Figure 2a, bottom row), even when the prediction performance for IQ was drastically different. Moreover, we compared functional connectome fingerprinting with the original and enhanced data in HCP. Functional connectomes have previously been used as participant-specific “fingerprints” [52,53] that can accurately identify participants across different fMRI sessions or tasks. To perform fingerprinting we calculated the edge-wise correlation between the first (Rest1) and second (Rest2) resting-state connectomes in HCP. The predicted identity of the participant was the connectome from the other session with the highest edge-wise correlation [53]. We performed this identification process in each fold of our 10-fold cross-validation, so each identification procedure included 10% of the 506 participants in the HCP dataset. Whether using the original or the enhanced Rest1 connectomes, the identification rate remained the same ($p \approx 1$, Figure 2a, bottom row, HCP). Furthermore, we performed fingerprinting in HCP following the procedure used in Figure 2a, except we only used edges corresponding to various subnetworks, as previously defined with the Shen 268 atlas [53,54] (Figure S3). Even when only using a single subnetwork to fingerprint, there was essentially no difference in accuracy for original and enhanced connectomes (median $p \approx 0.5$ for each subnetwork; median overall $p \approx 1$). The fingerprinting results indicate that enhancement attacks preserved the individuality of connectomes at both the whole-brain and subnetwork levels, despite having a large effect on IQ prediction.

Finally, we evaluated several graph properties, including strength, assortativity, and clustering coefficient, in the original and enhanced connectomes [55]. Despite the sensitivity of graph theory metrics to minor changes [56], the correlation between these node-level metrics in the original and enhanced connectomes was very high ($r \approx 0.99$) (Figure 2c). As such, enhancement attacks appear to uniquely strengthen brain-behavior associations with the phenotype of interest, making them difficult to detect even with other analyses.

Figure 2. Performance enhancement attacks only cause minor changes to connectomes. **a)** Data are enhanced to predict IQ measurements in ABCD, HCP, and PNC for 100 iterations of different enhancement patterns (all 100 iterations are shown as points; there is a lot of overlap between iterations). The x-axis reflects the mean absolute value of the enhancement pattern added at the edge level (i.e., the absolute mean of the enhancement pattern across all participants for the 20% of edges we altered). At $x=0$, there is no enhancement. As a larger enhancement pattern is added, the prediction performance (prediction correlation) increases to $r>0.9$, although the edge-wise correlation between original and enhanced connectomes is still $r\approx 0.99$. In the second row of **a)**, enhancement attacks are shown to not affect downstream analyses, which included a sex classification model and participant identification (“fingerprinting”) for HCP. **b)** Identification rates by subnetwork between Rest1 original/enhanced and Rest2 connectomes in HCP. **c)** Several graph metrics, including strength, assortativity, and clustering coefficient, were calculated for the original connectomes and enhanced connectomes, using the largest scale of enhancement presented in **a)**. The correlation between these metrics for original and enhanced connectomes is presented in **c)**, with error bars representing the standard deviation of the correlation across participants.

2.4 Performance enhancement attacks can be used to alter interpretations

In addition to falsely improving predictive ability, enhancement manipulations can be leveraged to reinforce a particular brain-behavior relationship. For this example, we used rCPM with 10-fold cross-validation to predict state anxiety (State-Trait Anxiety Inventory) [57] in the SLIM dataset, which is an open-source dataset of preprocessed connectomes. We first explored the prediction performance in unaltered connectomes and found essentially no predictive power

(Figure 3, top row). As in the previous section, we successfully manipulated random edges to increase prediction performance to $r=0.93$ (Figure 3, middle row). Then, we altered only edges involved in the salience network to enhance prediction to $r=0.9$ (Figure 3, bottom row). This influenced model coefficients to be dominated by edges of the salience network and thus would suggest that the salience network can predict anxiety scores in this dataset. The same targeted pattern injection could be similarly performed for other subnetworks to enforce a different interpretation. These results highlight the potential power and importance of enhancement attacks, which could not only alter performance, but also support an unfounded neuroscientific interpretation of a clinically-relevant phenotype.

Figure 3. Performance enhancement attacks applied to an open-source dataset of preprocessed connectomes. This example is shown for prediction of state anxiety in the SLIM dataset with resting-state connectomes and rCPM. In the top row, prediction with the original dataset shows poor performance ($r \approx 0$). In the second row, as in Figure 1, an enhancement pattern proportional to the state anxiety measure can be added to random edges to enhance performance while maintaining very high correlations between the original and enhanced connectomes ($r \approx 0.99$). In the bottom row, an enhancement pattern can be added to specific subnetworks to alter interpretation. Here, we targeted the enhancement pattern to the salience subnetwork, and the resulting coefficients reflect that edges in the salience network dominate the prediction outcome.

2.5 Performance enhancement extends beyond preprocessed connectomes

Because the previous sections demonstrated that performance enhancement attacks are highly effective against connectomes, a potential solution would be to always release raw or time-series data. Therefore, we investigated whether these attacks could be implemented earlier in the processing pipeline, such as on node time-series data. In this example (Figure 4), we manipulated time-series data from HCP (rest session 1, $n=506$) that was parcellated into 268 nodes with the Shen atlas [44] to falsely enhance prediction of IQ. We selected a pattern—in this case we arbitrarily selected a low-frequency sinusoid—to add to or subtract from each node's time course, scaled by a factor proportional to each participant's IQ, to increase or decrease correlations between nodes. The resulting time-series data (median $r=0.994$) and the calculated connectomes (mean $r=0.991$) were very similar, but the prediction of IQ drastically increased using the enhanced data (Figure 4). More representative node time-series traces are shown in the Figure S4. The efficacy of performance enhancement on node time-series data showed data manipulations are not exclusive to preprocessed connectomes and that more complex algorithms may be able to manipulate raw data to achieve desired results.

Figure 4. Time series performance enhancement attacks. Node time-series data can be manipulated by adding a pattern with amplitude proportional to the IQ of each participant to increase/decrease the calculated functional connectivity between specific nodes. In this case, we chose a sinusoid pattern to add to the time-series data. A representative node is shown in this figure. The correlations between original and enhanced time-series ($r=0.988$) and resulting connectome ($r=0.985$) data are very high, despite large differences in prediction performance ($r=0.15$ vs. $r=0.77$).

2.6 Adversarial noise degrades connectome-based model accuracy

In contrast to the previous sections where data were manipulated to falsely enhance performance, other manipulations are designed to decrease prediction accuracy, most notably adversarial noise attacks [24]. Adversarial noise attacks have been effectively implemented in numerous fields [25–31], where only minor perturbations to the data are required to change the prediction. We set out to determine the extent to which connectome-based models are susceptible to adversarial attacks as a measure of trustworthiness.

Figure 5. Adversarial attack accuracy as a function of magnitude of attack for our three datasets and SVM classifiers of self-reported sex. The x-axis reflects an increase in the size of the attacks, represented as the mean absolute value of the added noise pattern, while the y-axis shows accuracy on the manipulated data. The experiment is repeated for 100 different random seeds and standard deviations across the 100 iterations are shown (very small standard deviations). At three points for the HCP line, representative connectomes are shown, as well as histograms with edge values for the original connectomes, adversarial connectomes, and adversarial noise pattern. Above each representative connectome is the edge-wise correlation with the original connectome.

For a SVM classifier of self-reported sex in ABCD, HCP, and PNC, we used a gradient-based method [26] to create adversarial noise at the time of model testing based on the model parameters. For each fold of our 10-fold cross-validation, a single, sex-specific adversarial pattern was updated and added to each test connectome until all connectomes were classified incorrectly (i.e., accuracy=0; Figure 5). As the mean absolute value (on the edge level) of the attack increased, more connectomes were classified incorrectly (Figure 5). Even when manipulating the data to achieve 0% accuracy, the original and adversarial connectomes showed very strong edge-wise correlations ($r \sim 0.99$). We repeated this analysis for logistic regression and found a similar trend (Figure S5). Hence, even very subtle adversarial attacks can completely degrade connectome-based modeling pipelines.

2.7 Adversarial noise is small and does not significantly change a connectome

All analyses in this section were performed using the minimum adversarial noise magnitude required for 0% accuracy. Crucially, we found that the mean absolute value of adversarial noise (whole-connectome level) required to trick connectome-based classifiers was small (between 0.01-0.03 to achieve 0% accuracy). As a result of the small magnitude, addition of the adversarial noise caused no apparent visual differences between the real and corrupted connectomes, and the distribution of edge values was nearly identical (Figure 5). Moreover, since adversarial noise was small in magnitude, real and corrupted connectomes maintained a high edge-wise correlation ($r \sim 0.99$). We also investigated the mean absolute value of the adversarial noise across 10 canonical resting-state networks, previously defined with the Shen 268 atlas [53,54]. The scale of the adversarial attacks was small across each subnetwork (Figure 6a), though notably the within-network noise values (diagonal elements of matrices in Figure 6a) were significantly larger than between networks ($p < 0.004$ for all three datasets).

Next, we investigated individual differences in both real and adversarial connectomes through functional connectome fingerprinting. We performed the fingerprinting identification process in each fold of our 10-fold cross-validation, so each identification procedure included 10% of the 506 participants in the HCP dataset. The identification rate between Rest1 and Rest2 connectomes was 96.4% when using unmodified Rest1 data and 96.3% when using adversarial Rest1 data. There was no significant difference in the identification rate for real and adversarial Rest1 scans when using any of the available tasks to perform fingerprinting (median p 's > 0.14 for each subnetwork; median overall $p = 0.5$; Figure 6b). Furthermore, we performed identification between Rest1 original/adversarial and Rest2 original connectomes using only edges corresponding to each subnetwork and found very similar identification rates across every subnetwork (median p 's > 0.38 for each subnetwork; median overall $p = 0.58$; Figure 6c). Overall, adversarial noise attacks preserved the individual uniqueness of connectomes in a fingerprinting paradigm at both the whole-brain level and the subnetwork level.

a.
Distribution of Adv. Noise (SVM)

b.

c.
ID Rates by Subnetworks

Figure 6. a) Breakdown of SVM adversarial noise into subnetworks. Brighter colors reflect higher mean absolute value of noise in that subnetwork. **b)** Identification rates in original and adversarial connectomes in the HCP dataset. The original or adversarial Rest1 scans were compared to connectomes in another session (Rest2) or task. The connectome with the highest edge-wise correlation was selected as the predicted identity. **c)** Using original or adversarial Rest1 scans, we identified participants based on their correlations with the original Rest2 scans. For this portion, we only used a specific subset of edges corresponding to each subnetwork to predict the identity.

2.8 In-scan data manipulations

Previous results with performance enhancement and adversarial noise attacks required someone intentionally manipulating the data to affect prediction performance. Unlike attacks that required malicious actions, we also explored manipulations that may occur naturally at different points of a study, including 1) the participant holding their breath in the scanner and 2) changes in the participant's brain state (i.e., task-based activation).

Figure 7. Several in-scan manipulations may decrease model accuracy. All models here are SVM classifiers for self-reported sex, trained on resting-state connectomes with 10-fold cross-validation. Error bars represent standard deviation across 100 iterations of different 10-fold cross-validation splits. **a)** In the CNP dataset, using the breath-hold task at test time instead of resting-state in a resting-state model causes a large decrease in accuracy for female connectomes. **b)** In HCP, using a different task at test time causes large changes in accuracy, with most task-based activation resulting in decrease in accuracy for female connectomes.

First, we investigated whether participants can make a simple physical change (i.e. holding their breath) to alter classification accuracy using the CNP dataset. A linear SVM model for classification of self-reported sex was trained with 10-fold cross-validation to predict sex on resting-state connectomes. We tested the model on both resting-state connectomes and breath-hold task connectomes. When participants held their breath, they were classified more often as male (84.1% accuracy at rest, 92.5% accuracy with breath-hold; median $p=0.049$) than female (62.3% accuracy at rest, 35.2% accuracy with breath-hold; median $p=0.0002$; Figure 7a; Figure S6a for logistic regression).

Second, we considered whether changes in participants' brain states could affect prediction accuracy. This may be particularly relevant to the interpretation of resting state data, where, unlike task-based fMRI, brain states are not experimentally manipulated. To assess the importance of presumed brain state on predictive models, we asked an extreme version of this question: what if the brain state is entirely different from what the model is expecting? As such, we investigated whether using task-based connectomes altered classification accuracy of

models that were built on resting-state connectomes [58] in the HCP dataset. We trained a linear SVM model to predict sex on Rest1 connectomes. Next, we varied the test connectome with each of the nine connectomes available in HCP, two of which are resting-state scans and seven of which are task-based scans. The model classified resting-state connectomes with high accuracy for both males and females, whether using Rest1 or Rest2 connectomes. However, using task-based connectomes greatly affected accuracy, which was often reduced for females (Figure 7b; Figure S6b for logistic regression). For example, when using the working memory task, only 23.0% of female participants were classified correctly compared to 88.3% with the Rest1 scan (median $p=7e-44$). Manipulation of resting-state activity is a subtle attack that is difficult to control and can be induced by previous tasks [59], or by the participant's tendency to potentially mimic a task during resting-state. These manipulations add to growing criticism of the interpretation and utility of functional activity "at rest" [60–62].

Together, these examples indicate that connectome-based predictive models are susceptible to both manipulations designed to enhance/corrupt performance and in-scan, unintentional manipulations of data.

3. Discussion

In this study, we demonstrated that three types of connectome-based models (rCPM, SVM, logistic regression) were fooled by small and simple data manipulations, thus suggesting low trustworthiness. We introduced novel "enhancement attacks", which falsely increased prediction performance from $r=0-0.2$ to $r>0.9$, and we also applied adversarial noise attacks to reduce model accuracy from $\sim 80\%$ to 0% . Despite the large differences in performance between the original and manipulated data, the edges were highly correlated ($r\approx 0.99$) and downstream analyses (identification rate, sex classification, graph topology) were unaffected. Model performance could also be decreased by in-scan changes, including breath-holding and task-based activation. Overall, these three data manipulations techniques present a concern for a wide range of settings from scientific integrity to potential down-the-line clinical applications.

Enhancement attacks are a novel method to falsely increase performance of machine learning models via data manipulation, and they are distinct from adversarial noise attacks in both implementation and motivation. While adversarial noise attacks alter only the test data to ruin model performance, enhancement attacks alter the entire dataset (i.e., train and test data) to falsely improve performance. In academic settings, a researcher might use enhancement attacks to make prediction performance higher and more publishable, or to support an unfounded neuroscientific claim. Similarly, in a commercial setting, a start-up could use enhancement attacks to deceive investors and increase the valuation of their company. In contrast, one might use adversarial noise to evade a model in a real-world application of machine learning, such as to bypass computer virus detection software.

With enhancement attacks, we demonstrated that both connectome and node time-series data can be manipulated to falsely enhance performance or provide evidence for a baseless interpretation. Although sharing data, especially processed data, has many benefits [63–65],

data-sharing is not a universal safeguard against data manipulations. Beyond enhancement attacks affecting individual studies, if enhanced data are shared on openly available repositories, independent researchers could unknowingly publish results with enhanced data and never be aware of any manipulation. An “enhanced” dataset circulating within the neuroimaging community would cause wasted resources and possibly harmful neuroscientific conclusions. We still advocate for sharing data [66], including preprocessed connectomes (when appropriate) to lower barriers to entry and minimize duplication of effort. But the potential for sharing data manipulated in undetectable ways should be noted.

In addition, we showed that in-scan manipulations (breath-holding and changes in brain state) may degrade accuracy in fMRI studies. For other data types, such as facial recognition, self-driving cars, and skin cancer detection [30], data manipulations that alter prediction outcomes [67] would generally be easier to detect: a strange pair of 3D-printed glasses [68], a pattern on a stop sign [69], or a physical mark near a mole on the skin. Yet, fMRI presents a unique challenge because we cannot “see” changes in brain state, medication use, or other factors that may affect the BOLD signal [70,71]. Manipulations that occur before or during acquisition, including breath-holding and task-based activation, may affect current studies, and subsequent work should comprehensively characterize the effects of different manipulations on various model predictions.

Furthermore, adversarial noise attacks degrade the accuracy of classification models of self-reported sex. These attacks would primarily occur in clinical applications of fMRI (e.g., misclassification of an individual into a diagnostic category), which are currently limited by other existing roadblocks [72,73]. Still, adversarial noise attacks illustrate the fragility of connectome-based predictive models by finding the minimum perturbation required to change classification outcomes.

An important remaining question is how to make connectome-based pipelines more trustworthy. In the machine learning literature, defenses to adversarial attacks, called “adversarial defenses” [74–76], have been widely studied. However, the same strategies for defending against enhancement attacks may not apply because enhancement attacks alter the entire dataset (i.e., not just the test data), which makes distinguishing between true signal and false manipulations more difficult. Future work should explore how to best alleviate the effects of enhancement attacks through changes to the analysis pipeline, emphasis on cross-dataset predictions, and maintenance of a “transaction history” to track all operations on data, such as blockchains [77]. For now, adherence to ethical principles, rigorous study designs, and awareness of the low trustworthiness of connectome-based models are the best strategies to prevent data manipulation.

There are several final methodological considerations of our work. Enhancement attacks could be applied to “trick” any statistical analysis, but the high dimensionality of connectome-based machine learning models makes them a prime target for inconspicuous enhancement manipulations. Since specific patterns must be added to the data to cause performance enhancement, enhancement attacks do not extend to cross-dataset predictions. For adversarial

attacks on linear models, the noise generation method is proportional to model coefficients and can be directly calculated. In addition, SVM is more easily attacked than logistic regression; however, logistic regression suffers from a lower baseline accuracy (Figure S5). Finally, this study only examined trust in connectome-based models and is not an exhaustive test of all neuroimaging modalities and models.

The neuroimaging community is beginning to recognize and explore the issues concerning ethics in machine learning, with a particular focus on bias in datasets and models [13,14]. Trust is distinct from bias and represents an equally important, yet widely overlooked, facet of ethics in neuroimaging models. Whereas bias describes performance discrepancies due to a static trait, trust involves manipulating the data to cause a different outcome. This work has highlighted the low trustworthiness of connectome-based models, where data can be easily manipulated to produce different results. Although trust is just one aspect of ethics in machine learning, it can complement ethical benchmarks [78–80] that have been designed to ameliorate other rampant ethical issues in machine learning models [81,82]. Future efforts to improve trustworthiness will be necessary to ensure fair and ethical machine learning practices in neuroimaging.

4. Methods

4.1 Datasets

We used classification and regression models in five open-source datasets—the Adolescent Brain Cognitive Development (ABCD) Study [35], the Human Connectome Project (HCP) [36], the Philadelphia Neurodevelopmental Cohort (PNC) [37], the Southwest University Longitudinal Imaging Multimodal study (SLIM) [38], and the UCLA Consortium for Neuropsychiatric Phenomics (UCLA-CNP) [39]—to evaluate the robustness of connectome-based models to several styles of adversarial attacks. The SLIM dataset was only used to illustrate an example of open-source preprocessed data being enhanced, and the UCLA-CNP dataset was only used to demonstrate in-scan manipulations.

The ABCD dataset consists of participants ranging from 9-10 years old in the United States from 21 different acquisition sites. The HCP dataset has imaging data from healthy adults ages 22-35 in the United States. The PNC dataset contains data from youth ages 8-21 in the greater Philadelphia area. The SLIM dataset has healthy young adults ages 17-27 in China. The UCLA-CNP consists of healthy adults and participants with ADHD, bipolar disorder, or schizophrenia, ranging from ages 21-50, in the Los Angeles area.

4.2 Processing

For the ABCD, PNC, and UCLA-CNP, fMRI data were motion corrected and the Shen atlas was warped to single participant space as previously described [40,41]. For the HCP dataset, we started with the minimally preprocessed HCP data [42]. Further preprocessing steps were performed using *BiImage Suite* [43] and are the same across studies. Several covariates of no interest were regressed from participants' functional data including linear and quadratic drifts, mean cerebrospinal fluid signal, mean white matter signal, and mean global signal. For additional control of possible motion-related confounds, a 24-parameter motion model (including six rigid body motion parameters, six temporal derivatives, and these terms squared) was regressed from the data. The data were temporally smoothed with a Gaussian filter (approximate cutoff frequency=0.12 Hz). We then applied a canonical gray matter mask defined in common space, so only voxels in the gray matter were used in further calculations. Denoised data were parcellated into 268 nodes using the Shen atlas [44]. Next, the mean time courses of each node pair were correlated, and correlation coefficients were Fisher transformed, generating a connectome for each participant. For the HCP dataset, connectomes for each phase encoding (i.e., RL and LR) were calculated independently and then averaged together. For the SLIM dataset, preprocessed connectomes were downloaded, with preprocessing steps described in [38]. After excluding participants for excessive motion (>0.2mm) and missing nodes, 3362 individuals in the ABCD dataset, 506 in the HCP dataset, 561 in the PNC dataset, 445 in the SLIM dataset, and 190 in the UCLA-CNP dataset were retained.

4.3 Baseline regression models

For all baseline regression models, we trained ridge-regression connectome-based predictive models (rCPM) [46] in MATLAB (Mathworks, Natick, US) with 10-fold cross-validation and a nested 10-fold cross-validation to select the L_2 regularization parameter. For feature selection, we correlated each edge with the phenotype of interest and picked the top 10% of edges with the lowest p-values.

For ABCD, HCP, and PNC, the phenotype of interest was a fluid intelligence (IQ) measurement. For ABCD, Raven's progressive matrices [83] were used, scaled by age (mean 6.31, s.d. 2.56, range 1-17, median 6), and the same measure, though not scaled by age, was used for HCP (mean 17.54, s.d. 4.45, range 5-24, median 19). For PNC, IQ was assessed with the Penn Matrix Reasoning test (mean 12.28, s.d. 4.04, range 0-23, median 12) [84]. For SLIM, the phenotype of interest was the state anxiety score, as assessed by the State-Trait Anxiety Inventory [57] (mean 35.66, s.d. 8.28, range 20-65, median 35).

The main metric we used to determine prediction performance was Pearson's r between original and predicted phenotypes. We also reported the cross-validation R^2 , called q^2 [49], which is defined as:

$$q^2 = 1 - \frac{MSE(\hat{y}, y)}{MSE(y, y)} = 1 - NMSE$$

4.4 Baseline classification models

We trained both SVM (linear kernel) and logistic regression models in MATLAB (Mathworks, Natick, US) to predict self-reported sex in ABCD, HCP, PNC, and UCLA-CNP (in-scan manipulations only). The self-reported sex of participants by dataset was: ABCD (1653 female, 1609 male), HCP (270 female, 236 male), PNC (318 female, 244 male), and UCLA-CNP (117 female, 155 male). Models were trained with 10-fold cross-validation, with nested 5-fold cross-validation to select an L_2 regularization parameter. We used accuracy as our primary evaluation metric, though we also reported sensitivity and specificity in Table 1.

4.5 Connectome enhancement

We enhanced IQ prediction in ABCD, HCP, and PNC. To enhance connectome data, we first z-scored the phenotypic measurements.

$$y_z \leftarrow \frac{y - \bar{y}}{s_y}$$

Then, we randomly selected 20% of all edges e to perturb. For each of the selected edges, we added a value with magnitude proportional to each participant's z-scored IQ.

$$e_{i,j} \leftarrow e_{i,j} \pm k * y_{z,j}$$

where $e_{i,j}$ represents edge i for participant j and $y_{z,j}$ is the z-scored phenotype for participant j .

Whether we added or subtracted each value was randomly determined for each edge. The results presented in Figure 2 used $k=\{0, 0.004, 0.007, 0.01, 0.014, 0.02, 0.03\}$. $k=0$ was used as a reference and means that no enhancement was performed.

After injecting the enhancement pattern into each connectome, rCPM models were re-trained to predict each phenotype. Because we injected patterns proportional to the phenotype of interest, we would expect performance to be falsely enhanced. We repeated enhancement for each value of k for 100 different random seeds and recorded r and q^2 at each iteration.

4.6 Enhancement downstream analyses

In addition to r and q^2 , we evaluated enhanced connectomes in several other analyses—self-reported sex classification, functional connectome fingerprinting, and investigation of graph properties—to determine their similarity with original connectomes.

For self-reported sex classification, we used the data that was enhanced for IQ prediction to train a SVM classifier as described in *4.4 Baseline classification models*. At each different scale of the enhancement k , we compared the sex classification performance to that of the unaltered dataset, using accuracy as the main metric.

We also performed functional connectome fingerprinting [52,53] following the method of [53]. We performed fingerprinting only in the HCP dataset because it has two resting-state scans. A total of 50-51 participants were used at a time for fingerprinting evaluation, corresponding to all the connectomes in a single fold of the 10-fold cross-validation enhancement pipeline (10% of 506 HCP participants). With 50 participants, the identification rate of random guessing is 2%. To predict identity, we first calculated the edge-wise correlation of each participant's Rest1 connectome with their Rest2 connectome, and the predicted identity of each participant corresponded to the Rest2 connectome with the highest edge-wise correlation with their Rest1 connectome. We repeated the fingerprinting process using both original Rest1 connectomes and Rest1 connectomes that had been enhanced for IQ prediction.

In a further implementation of fingerprinting, we identified participants using only edges that are part of selected subnetworks, as previously defined with the Shen 268 atlas [53,54]. 10 networks were defined based on nodes (Figure S3): Medial-Frontal (MF), Fronto-parietal (FP), Default Mode (DMN), Motor (MOT), Visual I (VI), Visual II (VII), Visual Association (VAs), Salience (SAL), Subcortical (SC), Cerebellum (CBL). Based on these 10 networks, we defined 55 subnetworks, where each subnetwork was defined as edges belonging to each pair of networks. For example, 10 subnetworks involved the MF network: MF-MF, MF-FP, MF-DMN, MF-MOT, MF-VI, MF-VII, MF-VAs, MF-SAL, MF-SC, MF-CBL. The fingerprinting procedure followed the same process as above, except edge-wise correlations were calculated only with edges belonging to one of the 55 subnetworks.

Moreover, we evaluated several graph properties for the positive edges only, including strength, assortativity, and clustering coefficient, in the original and enhanced connectomes [55]. All three

of these properties are node-level metrics, and we evaluated the similarity between original and enhanced connectomes by correlating the node-level metrics for each of these measures. Strength is the sum of edge weights for each node. Assortativity measures how similar the strengths are between connected nodes. The clustering coefficient is the mean weight of triangles for each node [55].

4.7 Targeted enhancement

In an extension of the above attack, we constrained the injected enhancement pattern to a specific resting-state network instead of a random subset of edges. We performed this enhancement in the SLIM [38] dataset (N=445) to predict state anxiety scores from the State-Trait Anxiety Inventory [57]. After perturbing edges of one network, we repeated model training with rCPM. We also evaluated the model coefficients (averaged over the 10 folds of cross-validation) to assess how changing a specific network altered the distribution of coefficients across networks.

4.8 Time-series enhancement

We used HCP node time-series data (268-node Shen atlas; 1200 time points) and prediction of IQ as an example of time-series enhancement. We arbitrarily selected an enhancement pattern as a sinusoid with four periods across the 1200 time points. For each participant, the enhancement pattern (sinusoid) was scaled by a factor proportional to their IQ score and the original correlation of that edge with the IQ scores. Then, for each edge, we found the two nodes corresponding to that particular edge. If the edge was positively correlated with IQ, we added the participant-specific enhancement pattern to both node time-series. If the edge was negatively correlated with IQ, we added the enhancement pattern to one node and subtracted it from the other node.

After enhancing the time-series data, we computed the connectomes by taking the correlation between each pair of nodes and applying the Fisher transform. Then, rCPM models were trained with 10-fold cross-validation to predict IQ. We recorded the similarity (Pearson's r) between the original and enhanced time-series data along with similarity between the resulting original and enhanced connectomes. While there may be other more effective strategies to enhance time-series data, we performed this simple attack as a proof-of-concept.

4.9 Adversarial noise attacks

In both classifiers (SVM, logistic regression), we used a gradient-based attack following the method of Biggio [26]. The attacks occurred on the test data at the time of model testing and are “white-box” attacks, meaning that we had access to the model parameters. Let our decision function be represented by $g(x)$ with input features x . A participant was classified as female if $g(x) < 0$ and male if $g(x) > 0$. The goal of the adversarial noise attack was to perturb all the true female (male) connectomes such that the adversarial $g(x) > 0$ (< 0). The noise for female and male connectomes were optimized separately.

The loss functions were

$$\begin{aligned} L_f &= -g(x + n_f) \text{ (all female connectomes)} \\ L_m &= g(x + n_m) \text{ (all male connectomes)} \end{aligned}$$

Adversarial noise was initialized to zeros and iteratively updated based on the gradient:

$$n_{f/m} \leftarrow n_{f/m} - \lambda \frac{dL_{f/m}}{dn_{f/m}}$$

where λ is the step size.

For linear SVM and logistic regression, the derivative term was given by:

$$\frac{dL_{f/m}}{dn_{f/m}} = \mp \beta$$

for coefficients β . For these linear models, this process just simplifies to adding adversarial noise proportional to the coefficients.

We also explored how the adversarial noise was distributed across subnetworks by taking the mean absolute value of the adversarial noise in each subnetwork.

4.10 Adversarial noise downstream analyses

We compared functional connectome fingerprinting rates between original and adversarial Rest1 connectomes, as previously described. Notably, we performed fingerprinting with the Rest1 connectomes and each of the eight other tasks in HCP (Rest2, Gambling, Language, Motor, Relational, Social, Working Memory, Emotion). We also performed subnetwork-specific fingerprinting.

4.11 In-scan manipulations

We explored two in-scan manipulations: breath-holding and task-based activation. In both cases, classifiers were trained with 10-fold cross-validation (nested loop for L_2 regularization parameter selection) on unaltered resting-state data to predict self-reported sex.

For the breath-hold manipulation, we applied a resting-state model to breath-hold task data in UCLA-CNP (n=190). For the task-based activation manipulation, we applied a resting-state model to one of the nine available connectomes (Rest1, Rest2, Gambling, Language, Motor, Relational, Social, Working Memory, Emotion) in HCP (n=434 after removing participants with missing nodes in any task).

The classifiers were tested with both resting-state data (i.e., expected input) and manipulated data (breath-hold or different task). Accuracies for male and female connectomes were then compared between resting-state connectomes and connectomes that underwent in-scan manipulations.

4.12 Statistics

As previously described, performance of regression models was quantified with Pearson's r or cross-validation R^2 , called q^2 , between measured and predicted phenotypes. Classifiers were assessed by prediction accuracy, sensitivity, and specificity. For comparisons of real and manipulated connectomes, we computed Pearson's r between original and manipulated edges, to which we referred as "edge-wise correlation" throughout this work. To assess the significance of the differences in ID rate and accuracy between original and manipulated data, we used McNemar's test [85] and reported the median p value across the 100 iterations.

4.13 Data and Code Availability Statement

All five datasets used in this study are open-source: ABCD (NIMH Data Archive, <https://nda.nih.gov/abcd>), HCP (ConnectomeDB database, <https://db.humanconnectome.org>), PNC (dbGaP Study, accession code: phs000607.v3.p2, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000607.v3.p2), SLIM (INDI, http://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html), and UCLA-CNP (OpenNeuro, <https://openneuro.org/datasets/ds000030>). BioImage Suite tools used for processing can be accessed at (<https://bioimagesuiteweb.github.io/>). MATLAB scripts for trust analyses are available on GitHub (https://github.com/mattrosenblatt7/trust_connectomes).

5. Citation Diversity Statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field [86–90]. Here we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman [90,91]. By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 12.81% woman(first)/woman(last), 22.74% man/woman, 22.74% woman/man, and 41.72% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, non-binary, or transgender people. Second, we obtained the predicted racial/ethnic category of the first and last author of each reference by databases that store the probability of a first and last name being carried by an author of color [92,93]. By this measure (and excluding self-citations), our references contain 16.91% author of color (first)/author of color(last), 19.0% white author/author of color, 18.95% author of color/white author, and 45.13% white author/white author. This method is limited in that a) names and Florida Voter Data to make the predictions may not be indicative of racial/ethnic identity, and b) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases due to the ambiguous racialization or ethnicization of their names. Please note that 4 references were excluded from this analysis for difficulties with the code. We look

forward to future work that could help us to better understand how to support equitable practices in science.

Acknowledgements

This study was supported by R01MH121095 and the Wellcome Leap The First 1000 Days. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Data were provided by the Philadelphia Neurodevelopmental Cohort (Principal Investigators: Hakon Hakonarson and Raquel Gur; phs000607.v1.p1). Support for the collection of the data sets was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson. Data were provided by the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580, RL1MH083268, RL1MH083269, RL1DA024853, RL1MH083270, RL1LM009833, PL1MH083271, and PL1NS062410). Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, and U24DA041147. A full list of supporters is available at abcdstudy.org/nih-collaborators. A listing of participating sites and a complete listing of the study investigators can be found at abcdstudy.org/principal-investigators.html. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from NIMH Data Archive Digital Object Identifier 10.15154/1504041. DOIs can be found at nda.nih.gov/study.html?id=721.

References

1. Whelan, R. & Garavan, H. When optimism hurts: inflated predictions in psychiatric neuroimaging. *Biol. Psychiatry* **75**, 746–748 (2014).
2. Gabrieli, J. D. E., Ghosh, S. S. & Whitfield-Gabrieli, S. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron* **85**, 11–26 (2015).
3. Cremers, H. R., Wager, T. D. & Yarkoni, T. The relation between statistical power and inference in fMRI. *PLoS One* **12**, e0184923 (2017).
4. Noble, S., Mejia, A. F., Zalesky, A. & Scheinost, D. Leveling up: improving power in fMRI by moving beyond cluster-level inference. Preprint at <https://www.biorxiv.org/content/10.1101/2021.09.23.461354v1> (2021).
5. Shen, X. *et al.* Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518 (2017).
6. Cui, Z. & Gong, G. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage* **178**, 622–637 (2018).
7. Du, Y., Fu, Z. & Calhoun, V. D. Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. *Front. Neurosci.* **12**, 525 (2018).
8. Rosenberg, M. D., Casey, B. J. & Holmes, A. J. Prediction complements explanation in understanding the developing brain. *Nat. Commun.* **9**, 589 (2018).
9. Song, H., Finn, E. S. & Rosenberg, M. D. Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proc. Natl. Acad. Sci. USA* **118**, (2021).
10. Nielsen, A. N., Barch, D. M., Petersen, S. E., Schlaggar, B. L. & Greene, D. J. Machine Learning With Neuroimaging: Evaluating Its Applications in Psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging* **5**, 791–798 (2020).

11. Goldfarb, E. V., Rosenberg, M. D., Seo, D., Constable, R. T. & Sinha, R. Hippocampal seed connectome-based modeling predicts the feeling of stress. *Nat. Commun.* **11**, 2650 (2020).
12. Yip, S. W., Scheinost, D., Potenza, M. N. & Carroll, K. M. Connectome-Based Prediction of Cocaine Abstinence. *Am. J. Psychiatry* **176**, 156–164 (2019).
13. Benkarim, O. *et al.* The Cost of Untracked Diversity in Brain-Imaging Prediction. Preprint at <https://www.biorxiv.org/content/10.1101/2021.06.16.448764v1> (2021).
14. Li, J. *et al.* Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci Adv* **8**, eabj1812 (2022).
15. Brundage, M. *et al.* Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. Preprint at <https://arxiv.org/abs/2004.07213> (2020).
16. Rawal, Mccoy, Rawat, Sadler & Amant. Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives. *IEEE Transactions on Artificial Intelligence* **1**, 1–1 (2021).
17. Eshete, B. Making machine learning trustworthy. *Science* **373**, 743–744 (2021).
18. Acuna, D. E., Brookes, P. S. & Kording, K. P. Bioscience-scale automated detection of figure element reuse. Preprint at <https://www.biorxiv.org/content/10.1101/269415v3> (2018).
19. Bucci, E. M. Automatic detection of image manipulations in the biomedical literature. *Cell Death Dis.* **9**, 400 (2018).
20. Cicconet, M., Elliott, H., Richmond, D. L., Wainstock, D. & Walsh, M. Image Forensics: Detecting duplication of scientific images with manipulation-invariant image similarity. Preprint at <https://arxiv.org/abs/1802.06515> (2018).
21. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359–1366 (2011).
22. Wicherts, J. How misconduct helped psychological science to thrive. *Nature* vol. **597**, 153 (2021).

23. Al-Marzouki, S., Evans, S., Marshall, T. & Roberts, I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *BMJ* **331**, 267–270 (2005).
24. Szegedy, C. *et al.* Intriguing properties of neural networks. Preprint at <https://arxiv.org/abs/1312.6199> (2013).
25. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. Preprint at <https://arxiv.org/abs/1412.6572> (2014).
26. Biggio, B. *et al.* Evasion Attacks against Machine Learning at Test Time. in *Machine Learning and Knowledge Discovery in Databases* 387–402 (2013).
27. Demontis, A. *et al.* Yes, machine learning can be more secure! a case study on android malware detection. *IEEE Trans. Dependable Secure Comput.* **16**, 711–724 (2017).
28. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J. & Song, D. Natural adversarial examples. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15262–15271 (2021).
29. Paschali, M., Conjeti, S., Navarro, F. & Navab, N. Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018* 493–501 (2018).
30. Finlayson, S. G. *et al.* Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
31. Han, X. *et al.* Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nat. Med.* **26**, 360–363 (2020).
32. Finlayson, S. G., Chung, H. W., Kohane, I. S. & Beam, A. L. Adversarial Attacks Against Medical Deep Learning Systems. Preprint at <https://arxiv.org/abs/1804.05296> (2018).
33. Fanelli, D. Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One* **5**, e10271 (2010).
34. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How Machine Learning Will Transform Biomedicine. *Cell* **181**, 92–101 (2020).

35. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* **32**, 43–54 (2018).
36. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62–79 (2013).
37. Satterthwaite, T. D. *et al.* The Philadelphia Neurodevelopmental Cohort: A publicly available resource for the study of normal and abnormal brain development in youth. *Neuroimage* **124**, 1115–1119 (2016).
38. Liu, W. *et al.* Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. *Sci Data* **4**, 170017 (2017).
39. Poldrack, R. A. *et al.* A phenome-wide examination of neural and cognitive function. *Sci Data* **3**, 160110 (2016).
40. Greene, A. S., Gao, S., Scheinost, D. & Constable, R. T. Task-induced brain state manipulation improves prediction of individual traits. *Nat. Commun.* **9**, 2807 (2018).
41. Rapuano, K. M. *et al.* Behavioral and brain signatures of substance use vulnerability in childhood. *Dev. Cogn. Neurosci.* **46**, 100878 (2020).
42. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105–124 (2013).
43. Joshi, A. *et al.* Unified framework for development, deployment and robust testing of neuroimaging algorithms. *Neuroinformatics* **9**, 69–84 (2011).
44. Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *Neuroimage* **82**, 403–415 (2013).
45. Biswal, B. B. *et al.* Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. USA* **107**, 4734–4739 (2010).
46. Gao, S., Greene, A. S., Constable, R. T. & Scheinost, D. Combining multiple connectomes improves predictive modeling of phenotypic measures. *Neuroimage* **201**, 116038 (2019).

47. Weis, S. *et al.* Sex Classification by Resting State Brain Connectivity. *Cereb. Cortex* **30**, 824–835 (2020).
48. Eliot, L., Ahmed, A., Khan, H. & Patel, J. Dump the ‘dimorphism’: Comprehensive synthesis of human brain studies reveals few male-female differences beyond size. *Neurosci. Biobehav. Rev.* **125**, 667–697 (2021).
49. Scheinost, D. *et al.* Ten simple rules for predictive modeling of individual differences in neuroimaging. *Neuroimage* **193**, 35–45 (2019).
50. Biggio, B., Nelson, B. & Laskov, P. Poisoning Attacks against Support Vector Machines. Preprint at <https://arxiv.org/abs/1206.6389> (2012).
51. Finn, E. S. *et al.* Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18**, 1664–1671 (2015).
52. Noble, S. *et al.* Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cereb. Cortex* **27**, 5415–5429 (2017).
53. Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).
54. Miranda-Dominguez, O. *et al.* Connectotyping: model based fingerprinting of the functional connectome. *PLoS One* **9**, e111048 (2014).
55. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**, 1059–1069 (2010).
56. Luo, W., Greene, A. S. & Constable, R. T. Within node connectivity changes, not simply edge changes, influence graph theory measures in functional connectivity studies of the brain. *Neuroimage* **240**, 118332 (2021).
57. Spielberger, C. D. *Manual for the State-trait Anxiety Inventory (form Y) ('Self-evaluation Questionnaire')*. Consulting Psychologists Press (1983).
58. Buckner, R. L., Krienen, F. M. & Yeo, B. T. T. Opportunities and limitations of intrinsic functional connectivity MRI. *Nat. Neurosci.* **16**, 832–837 (2013).

59. Hasson, U., Nusbaum, H. C. & Small, S. L. Task-dependent organization of brain regions active during rest. *Proc. Natl. Acad. Sci. USA* **106**, 10841–10846 (2009).
60. Morcom, A. M. & Fletcher, P. C. Does the brain have a baseline? Why we should be resisting a rest. *Neuroimage* **37**, 1073–1082 (2007).
61. Lurie, D. J. *et al.* Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Network Neuroscience* vol. 4 30–69 (2020).
62. Finn, E. S. Is it time to put rest to rest? *Trends Cogn. Sci.* **25**, 1021-1032 (2021).
63. Craddock, C. *et al.* The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* **7**, (2013).
64. Mennes, M., Biswal, B. B., Castellanos, F. X. & Milham, M. P. Making data sharing work: the FCP/INDI experience. *Neuroimage* **82**, 683–691 (2013).
65. Markiewicz, C. J. *et al.* OpenNeuro: An open resource for sharing of neuroimaging data. Preprint at <https://www.biorxiv.org/content/10.1101/2021.06.28.450168v3> (2021).
66. Horien, C. *et al.* A hitchhiker’s guide to working with large, open-source neuroimaging datasets. *Nat Hum Behav* **5**, 185–193 (2021).
67. Kurakin, A., Goodfellow, I. & Bengio, S. Adversarial examples in the physical world. Preprint at <https://arxiv.org/abs/1607.02533> (2016).
68. Sharif, M., Bhagavatula, S., Bauer, L. & Reiter, M. K. A General Framework for Adversarial Examples with Objectives. *ACM Trans. Priv. Secur.* **22**, 1–30 (2019).
69. Eykholt, K. *et al.* Robust physical-world attacks on deep learning visual classification. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 1625–1634 (2018).
70. Magalhães, R. *et al.* Habitual coffee drinkers display a distinct pattern of brain functional connectivity. *Mol. Psychiatry* **26**, 6589-6598 (2021).
71. Pereira-Sanchez, V. *et al.* Systematic Review: Medication Effects on Brain Intrinsic Functional Connectivity in Patients With Attention-Deficit/Hyperactivity Disorder. *J. Am.*

- Acad. Child Adolesc. Psychiatry* **60**, 222–235 (2021).
72. Dadi, K. *et al.* Benchmarking functional connectome-based predictive models for resting-state fMRI. *Neuroimage* **192**, 115–134 (2019).
 73. Specht, K. Current Challenges in Translational and Clinical fMRI and Future Directions. *Front. Psychiatry* **10**, 924 (2019).
 74. Meng, D. & Chen, H. MagNet: A Two-Pronged Defense against Adversarial Examples. in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* 135–147 (Association for Computing Machinery, 2017).
 75. Qiu, S., Liu, Q., Zhou, S. & Wu, C. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.* **9**, 909 (2019).
 76. Zhang, Y. & Liang, P. Defending against Whitebox Adversarial Attacks via Randomized Discretization. in *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (eds. Chaudhuri, K. & Sugiyama, M.) **89**, 684–693 (2019).
 77. Bell, J., LaToza, T. D., Baldmitsi, F. & Stavrou, A. Advancing Open Science with Version Control and Blockchains. in *2017 IEEE/ACM 12th International Workshop on Software Engineering for Science (SE4Science)* 13–14 (2017).
 78. Mitchell, M. *et al.* Model Cards for Model Reporting. in *Proceedings of the Conference on Fairness, Accountability, and Transparency* 220–229 (2019).
 79. Raji, I. D. & Yang, J. ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles. Preprint at <https://arxiv.org/abs/1912.06166> (2019).
 80. Jiang, H., Kim, B., Guan, M. & Gupta, M. To Trust Or Not To Trust A Classifier. in *Advances in Neural Information Processing Systems* (eds. Bengio, S. *et al.*) **31** (2018).
 81. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (eds. Friedler, S. A. & Wilson, C.) **81**, 77–91 (2018).

82. Nicol, T. L. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society* **16**, 252–260 (2018).
83. Bilker, W. B. *et al.* Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. *Assessment* **19**, 354–369 (2012).
84. Moore, T. M., Reise, S. P., Gur, R. E., Hakonarson, H. & Gur, R. C. Psychometric properties of the Penn Computerized Neurocognitive Battery. *Neuropsychology* **29**, 235–246 (2015).
85. McNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**, 153–157 (1947).
86. Mitchell, S. M., Lange, S. & Brus, H. Gendered citation patterns in international relations journals. *International Studies Perspectives* **14**, 485–492 (2013).
87. Maliniak, D., Powers, R. & Walter, B. F. The Gender Citation Gap in International Relations. *International Organization* **67**, 889–922 (2013).
88. Caplar, N., Tacchella, S. & Birrer, S. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy* **1**, 1–5 (2017).
89. Dion, M. L., Sumner, J. L. & Mitchell, S. M. Gendered Citation Patterns across Political Science and Social Science Methodology Fields. *Political Analysis* **26**, 312–327 (2018).
90. Dworkin, J. D. *et al.* The extent and drivers of gender imbalance in neuroscience reference lists. *Nat. Neurosci.* **23**, 918–926 (2020).
91. Zhou, D., Cornblath, E. J., Stiso, J., Teich, E. G. & Dworkin, J. D. Gender diversity statement and code notebook v1. 0. *Zenodo* doi:10.5281/zenodo.3672110 (2020).
92. Ambekar, A., Ward, C., Mohammed, J., Male, S. & Skiena, S. Name-ethnicity classification from open sources. in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* 49–58 (2009).
93. Sood, G. & Laohaprapanon, S. Predicting Race and Ethnicity From the Sequence of Characters in a Name. Preprint at <https://arxiv.org/abs/1805.02109> (2018).

S1. Sensitivity of performance enhancement attacks to regularization parameter

In this work, we implemented enhancement attacks, adversarial noise attacks, and in-scan manipulations on connectome-based models. A summary of a typical connectome-based modeling pipeline and where each manipulation may occur is presented in Figure S1. In-scan manipulations may occur naturally during data acquisition, and thus could affect research studies or real-world applications. Since enhancement attacks (time-series or connectome enhancement) alter the entire dataset prior to model training, they are most relevant for research settings. For example, a researcher could enhance their dataset, publish this result, and release the dataset publicly. Adversarial noise attacks occur after model training, at the time of testing. Adversarial noise attacks are most applicable to real-world applications (i.e., using a model in a clinical setting) because they are designed to trick or bypass the model for a particular test instance.

Figure S1. Summary of the manipulations investigated in this study. The left half shows a typical connectome-based pipeline. The right half shows where each manipulation can be applied in the pipeline. Red text indicates attacks that degrade performance, while green text indicates attacks that falsely enhance performance. **In-scan manipulations** naturally occur and may affect academic studies or translational applications. **Enhancement attacks** are applied to all data. These attacks are relevant for false enhancement of academic studies or open-source data. They can be applied at multiple points in the processing pipeline (**time-series enhancement** or **connectome enhancement**) to falsely enhance performance or alter neuroscientific interpretations. **Adversarial noise attacks** are applied to only the test data, based on the model coefficients. These attacks have implications in potential translational applications.

S2. Sensitivity of performance enhancement attacks to regularization parameter

There is a strong relationship between the L_2 regularization parameter, λ , and the effectiveness of the enhancement attack. In the main results, we used a λ grid-search consistent with MATLAB's default setting: it searches over a geometric sequence with a maximum being the largest λ that gives a nonnull model and the minimum being $\lambda_{\max} * 10^{-4}$. λ is then selected as the largest λ for which the mean-squared error (MSE) is within 1 standard error of the minimum MSE. With this method, the λ is generally very high, often $\lambda \approx 100$ -150 ($\log(\lambda) \approx 4.6$ -5). However, if choosing from a less restricted range of λ , enhancement becomes even more effective. In Figure S1, we performed an enhancement attack on HCP data with a scale of 0.01 (the mean absolute value of the enhancement pattern added at the edge level) for numerous λ values (natural log of λ reported on x-axis). If λ is small, the ridge penalty is small, and coefficients corresponding to enhanced edges will dominate.

Figure S2. Enhancement performance for an attack scale of 0.01 and a variety of λ . The plots show the enhanced **a)** Pearson's r and **b)** q^2 as a function of the regularization parameter.

S3. Network definitions

Figure S3. Network definitions with the Shen 268 atlas.

S4. Examples of enhanced node time-series data

Figure S4 extends upon the example presented in Figure 4, where the enhanced predictions had a correlation of $r=0.77$ with the measured phenotypes. We explored the correlations between original and enhanced node time-series data across all participants and all nodes; the vast majority of these correlations are above 0.9 (Figure S4a). For the lowest correlation among all participants and nodes, there is a noticeable difference between the original and enhanced data (Figure S4b). For the 25th percentile, median, and 75th percentile of correlations between original and enhanced data, the time-series appear nearly identical (Figure S4c-e).

Figure S4. Examples of original and enhanced node time-series data. **a)** Histogram of correlation values between original and enhanced time-series data across all nodes (268) and participants (506). **b)** Original and enhanced time-series data with the lowest correlation across all nodes and participants ($r=0.690$). **c)** Data with the 25th percentile of correlations ($r=0.986$). **d)** Data with the median correlation ($r=0.995$). **e)** Data with the 75th percentile of correlations ($r=0.999$).

S5. Adversarial noise attacks for logistic regression

In addition, we evaluated the adversarial robustness of logistic regression, as compared to SVM. We found that logistic regression was more adversarially robust than SVM for these particular predictions (Figure S5). In other words, the accuracy of logistic regression models decreased more gradually than SVM as the scale of the attack increased. Although it showed higher adversarial robustness, logistic regression also had a lower baseline accuracy than SVM (Figure S5; Attack Scale=0 for baseline).

Figure S5. Comparison of adversarial robustness of SVM and logistic regression. The logistic regression models had higher robustness to manipulations, meaning that a larger attack scale was required to decrease the accuracy. However, the baseline accuracy of logistic regression models was lower than that of SVM.

S6. In-scan manipulations using logistic regression

We re-created Figure 7 using logistic regression instead of SVM. When participants held their breath, they were classified more often as male (77.4% accuracy at rest, 89.9% accuracy with breath-hold; median $p=0.0025$) than female (54.0% accuracy at rest, 31.7% accuracy with breath-hold; median $p=0.0005$; Figure 7a; Figure S6a for logistic regression). Notably, while logistic regression showed the same trend as SVM, the differences in male accuracies for rest vs. breath-hold were much more significant ($p=0.0025$ vs $p=0.049$).

For the task-swap data, we saw slightly less drastic but still significant results compared to SVM (Figure S6b). For example, when using SVM (Figure 7), only 23.0% of female participants were classified correctly with the working memory task compared to 88.3% with the Rest1 scan (median $p=7e-44$). However, with logistic regression, 38.2% of female participants were classified correctly with the working memory task compared to 76.0% with the Rest1 scan (median $p=7e-24$).

Figure S6. Analogue to Figure 7 for logistic regression.