Overlooked features lead to divergent neurobiological interpretations of brain-based machine learning biomarkers

Brendan D. Adkinson[1,*], Matthew Rosenblatt[2], Huili Sun[1], Javid Dadashkarimi[3], Link Tejavibulya[1], Corey Horien[4], Margaret L. Westwater[5], Stephanie Noble[6,7], Dustin Scheinost[1,2,5,8,9,10]

1 Interdepartmental Neuroscience Program, Yale School of Medicine, New Haven, CT, 06510, USA
2 Department of Biomedical Engineering, Yale University, New Haven, CT, 06520, USA
3 Department of Radiology, Perelman School of Medicine, Philadelphia, PA 19104, USA
4 Perelman School of Medicine, Philadelphia, PA 19104, USA
5 Department of Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT, 06510, USA
6 Department of Bioengineering, Northeastern University, Boston, MA, 02120, USA
7 Department of Psychology, Northeastern University, Boston, MA, 02115, USA
8 Department of Statistics & Data Science, Yale University, New Haven, CT, 06520, USA
9 Child Study Center, Yale School of Medicine, New Haven, CT, 06510, USA
10 Wu Tsai Institute, Yale University, New Haven, CT, 06510, USA

*Corresponding author: Brendan Adkinson (brendan.adkinson@yale.edu), Magnetic Resonance Research Center, 300 Cedar Street, New Haven, CT, 06519

Abstract: A central objective in human neuroimaging is to understand the neurobiology underlying cognition and mental health. Machine learning models trained on brain connectivity data are increasingly used as tools for predicting behavioral phenotypes [1,2], enhancing precision medicine [3,4], and improving generalizability compared to traditional MRI studies [5]. However, the high dimensionality of brain connectivity data makes model interpretation challenging [6]. Prevailing practices within the field rely on sparsely selected brain connectivity features, implicitly interpreting identified feature networks as uniquely representative of a given phenotype while overlooking others. Here, we show that commonly overlooked brain connectivity features can achieve similar prediction accuracies while yielding markedly different neurobiological interpretations. Using four large-scale neuroimaging datasets spanning over 12,000 participants and 13 outcomes, we demonstrate that this phenomenon is widespread across cognitive, developmental, and psychiatric phenotypes. It extends to both functional connectivity (fMRI) and structural (DTI) connectomes and remains evident even in external validation. These findings suggest that common practices may lead to feature overinterpretation and a misrepresentation of the neurobiological bases of brain-behavior associations. Such interpretations present only the 'tip of the iceberg' when certain disregarded features may be just as meaningful, potentially contributing to ongoing issues surrounding reproducibility within the field. More broadly, our results point to the possibility that multiple neurobiologically distinct models may exist for the same phenotype, with implications for identifying meaningful subtypes within clinical and research populations.

Keywords: interpretability, subtyping, biotypes, clustering, feature selection, artificial intelligence, connectome based predictive modeling, executive function, language

Introduction

A central goal of human neuroimaging is to uncover the neurobiological underpinnings of cognition and mental health. Aligned with this objective, emerging machine-learning approaches predict individual differences in behavioral phenotypes from brain structure and function [1,2,7,8]. These 'brain-behavior' predictive models reduce overfitting and provide more robust assessments of associations relative to traditional magnetic resonance imaging (MRI) studies by separating datasets into training and test samples [5,9–11]. They also offer opportunities to advance precision medicine strategies, ranging from illness subtyping to predicting individual treatment responses [3,4,12]. Accordingly, substantial efforts have been dedicated to improving model accuracy [13–15], trustworthiness [16], fairness [17,18], reproducibility [19–24], and generalizability [25–28].

Neurobiological interpretability—the degree to which results can be mapped to known neural circuits and processes—is another key metric of model utility [29–31]. An easily interpretable model can uncover the neural circuits responsible for behavior, even with modest prediction performance [32]. However, the high dimensionality of brain connectivity data makes interpretation challenging [6]. To obtain more manageable representations, the field commonly employs methods, such as feature selection and highly penalized models, that favor sparsity [33]. Traditionally, resultant features are interpreted as the neurobiology underlying individual variation in the predicted phenotype and linked to existing neuroscientific knowledge. For instance, models are often named after the phenotype they predict (i.e., "working memory network" or "depression network"). By treating identified networks as definitive, these practices give the impression that a model uniquely represents that phenotype's neural correlates.

Implicitly, alternative features are overlooked as inconsequential to model performance and neurobiology.

Here, we test whether features overlooked by state-of-the-field practices are meaningful for prediction and neurobiological interpretation. Our analyses span 12,200 participants across four large-scale neuroimaging datasets, both functional and structural connectivity, and 13 outcomes, including age, sex, cognitive abilities, developmental measures, and psychiatric phenotypes. We use an original prediction paradigm wherein connectome features are divided into non-overlapping subsets based on their association with a target phenotype. We show that multiple subsets of features overlooked by common methodological approaches can yield similar prediction accuracies but markedly different biological interpretations. These results suggest that prevalent practices that focus on the “top” features may lead to overinterpretation and low replicability of feature sets, potentially misrepresenting the neurobiological bases of brain-behavior associations. They also point to the possibility that multiple neurobiologically-distinct models may exist for a given phenotype, which could have important implications for identifying meaningful subtypes within clinical or research populations.

Results

Overview

We used four developmental neuroimaging datasets: the Healthy Brain Network (HBN, n=1110) dataset [34], the Adolescent Brain Cognitive Development (ABCD, n=9371) Study [35], the Human Connectome Project Development (HCPD, n=428) dataset [36], and the Philadelphia Neurodevelopmental Cohort (PNC, n =1291) dataset [37]. Details about the datasets are presented in Table S1 and in the Methods section. We modified connectome-based predictive modeling [38] such that connectome features were divided into ten non-overlapping deciles based on the strength of their association with a target phenotype (Figure 1A). Specifically, to prevent data leakage, we computed the Pearson correlation coefficient between every connectivity edge and the target phenotype in the training set. Features were then ranked in descending order based on the absolute value of their correlation coefficients. Following ranking, the complete set of connectivity features was partitioned into ten non-overlapping deciles, with the first decile comprising the top 10% of features (*i.e.*, those most likely included in established neuroimaging predictive modeling pipelines) and the last decile comprising the bottom 10% of features. Each decile was subsequently used to predict the phenotype of interest in the respective test set. Models were trained using 10-fold cross-validation. Model performance was evaluated with Pearson's correlation (r), representing the correspondence between predicted and actual behavioral scores, along with the cross-validation coefficient of determination (q^2) and mean square error (MSE) across. Significance was assessed using permutation testing with 1000 iterations of randomly shuffled behavioral data labels.

Prediction accuracy is not exclusive to top-ranked features

We generated language abilities and executive function models in the PNC, HBN, and HCPD datasets using our modified CPM paradigm. As in previous work [25], we used principal component analysis (PCA) to combine individual cognitive measures into latent variables of executive function and language abilities. Models constructed from the first through ninth feature deciles successfully predicted executive function and language abilities across PNC, HBN, and HCPD (Figure 1, Figures S1-3). Surprisingly, lower-ranked feature deciles commonly overlooked during model building continued to demonstrate meaningful predictive power. For example, for PNC executive function, the first decile achieved a prediction performance of r=0.33 (p=0.001, q2=0.09, MSE=1.24), while the second (r=0.32, p=0.001, q2=0.08, MSE=1.24) through sixth (r=0.31, p=0.001, q2=0.07, MSE=1.25) deciles achieved a similar performance. Notably, while these models perform similarly, the underlying edge features are independent. Additionally, the first decile did not always exhibit the best prediction performance. For HCPD executive function, the fifth decile (r=0.16, p=0.001, q2=-0.07, MSE=2.08) numerically outperformed deciles with features ranked as more highly informative (e.g., first decile r=0.14, p=0.002, q2=-0.10, MSE=2.14). Performances for all deciles are presented in Supplementary Table S2.

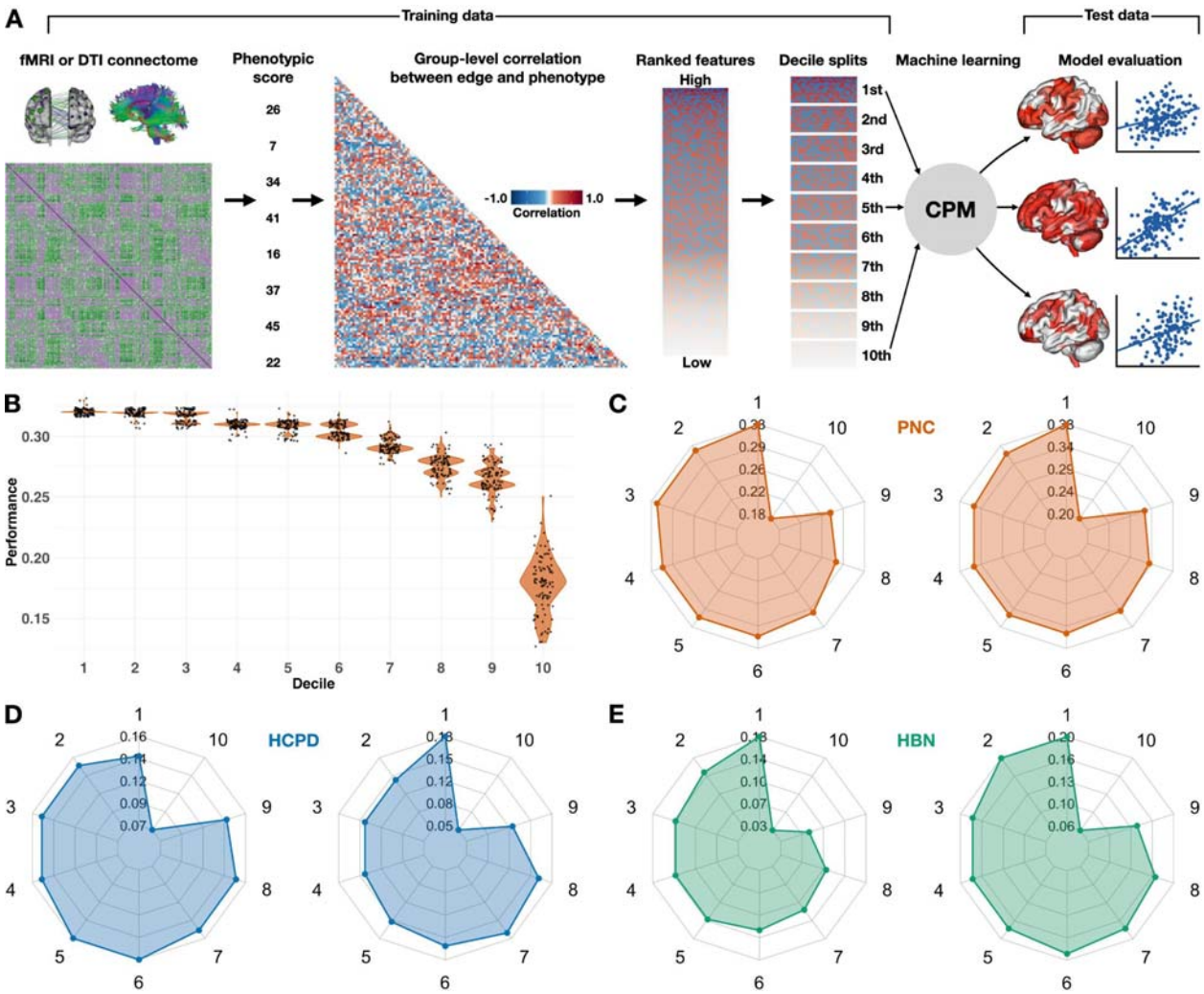

Figure 1. Connectome-based predictive modeling (CPM) across non-overlapping decile-ranked brain connectivity features. (A) Workflow illustrating the decile-based connectome-based predictive modeling (CPM) pipeline, including initial correlation of connectivity features with phenotypic outcome, ranking features based on group-level correlations between edges and phenotype, splitting features into deciles, and evaluating each decile-based model. (B) Violin plot showing the predictive performance of models trained on each decile of features within the PNC dataset for executive function. (C-E) Radar plots depicting predictive performance across deciles for PNC executive function (C, left), PNC language abilities (C, right), HCPD executive function (D, left), HCPD language abilities (D, right), HBN executive function (E, left), and HBN language abilities (E, right).

Overlooked feature sets also pass external validation

As external validation is the gold standard for model evaluation [24], we applied our decile-based models to three independent datasets. Despite the absence of overlapping features between deciles, deciles beyond the first demonstrated successful external validation for executive function and language abilities across all 12 cross-dataset prediction scenarios (Figure 2, Tables S3, S4, S5). For example, PNC executive function models tested in HCPD had no significant differences in external validation between deciles 1 through 9. In other words, the PNC executive function models from the 9th decile (r=0.13, p=0.004) generalized to HCPD just as well as the 1st decile (r=0.14, p=0.002). Similar trends were observed for other models. Models tested in PNC (r's generally >0.25, Figure 2) and HCPD (r's generally >0.10) outperformed those tested in HBN (r's generally <0.10).

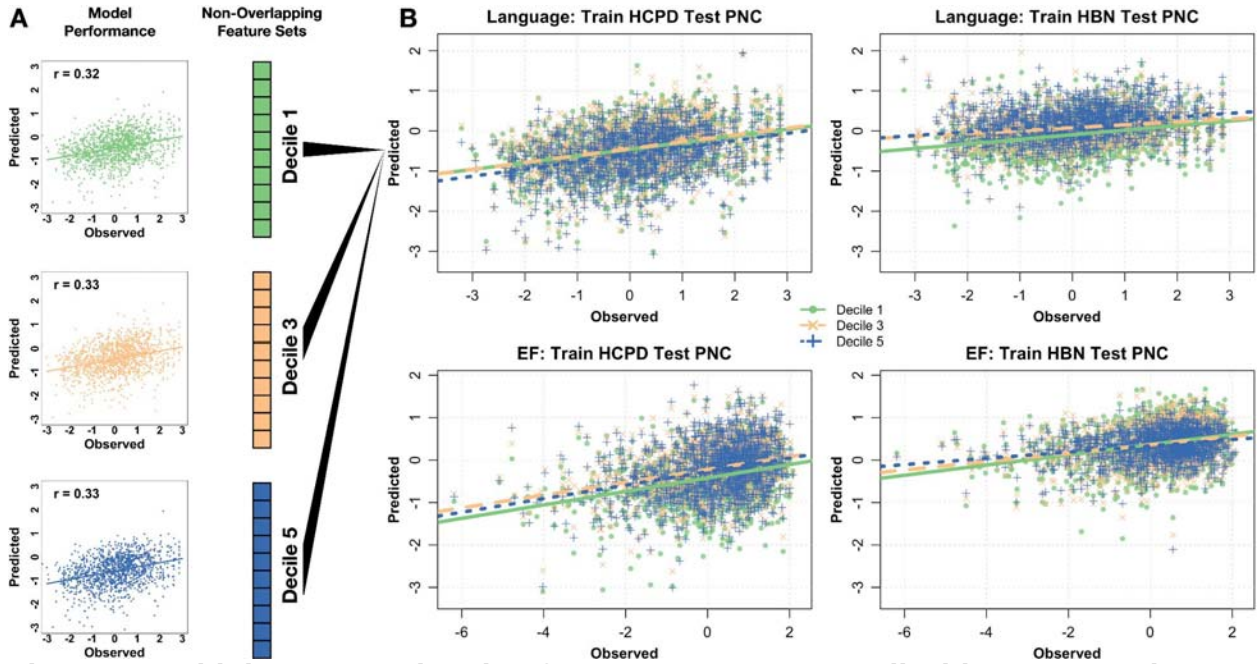

Figure 2. Multiple non-overlapping feature sets are generalizable to external datasets. (A) Scatter plots depicting observed versus predicted values for language abilities predictions when training in HCPD and testing in PNC using 1st (green), 3rd (blue), and 5th (purple) decile feature subsets. (B-E) Scatter plots for 1st (green), 3rd (blue), and 5th (purple) decile feature subsets overlaid to enable direct comparison of predictive accuracy across deciles. (B) Training a language abilities model in HCPD and testing in PNC. (C) Training a language abilities model in HBN and testing in PNC. (D) Training an executive function model in HCPD and testing in PNC. (E) Training an executive function model in HBN and testing in PNC.

Overlooked features offer distinct neurobiological information

Next, because models performed similarly across non-overlapping feature subsets, we sought to determine if similar brain features were implicated in each decile-based model by comparing within- and between-network functional connectivity. Remarkably, the canonical networks underlying predictions differed between deciles for both executive function and language abilities (Figure 3, Figure S4). For example, for PNC executive function, connectivity between the visual association and frontal-parietal networks was prominent in decile 1 but less informative for deciles 2 through 5 (Figure

3B). While networks demonstrated moderate resemblance (*r*'s=0.15-0.81, Figure S5) between neighboring deciles, non-neighboring deciles exhibited marked dissimilarity. For PNC executive function, the correlations were *r*=0.07 between decile 1 and decile 3, *r*=-0.17 between decile 1 and 4, and *r*=-0.21 between decile 1 and 5 (Figure S5). This phenomenon in which networks became increasingly dissimilar with further deciles was consistent across datasets and phenotypes. Trends were also similar when shifting from a network-based comparison to a node-based approach, evaluating features based on the sum of model contributions for each of the 268 nodes (Figure 3C).

The selected features were more stable across folds in the higher deciles than in lower deciles (Table S6). On average across all models, relative to the first decile, 84.64% of the second decile, 38.82% of the third decile, and 22.19% of the fifth decile edges were consistent in at least five folds. To account for this variation, the means of edgewise regression coefficients across the folds were used for all interpretability analyses. Analyses were repeated using only edges selected consistently across at least five folds, yielding similar results (Figure S6).

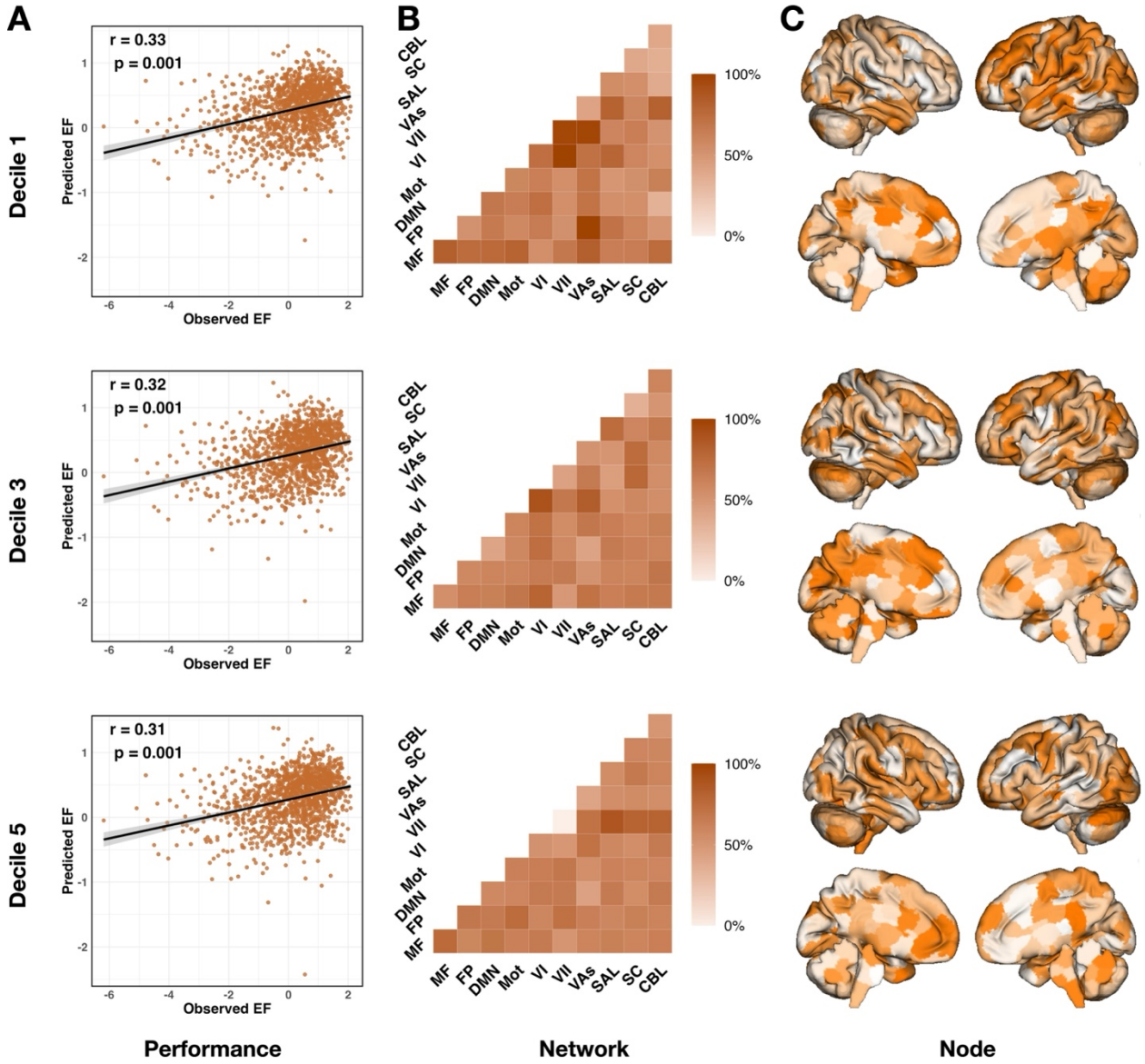

Figure 3. Overlooked feature sets offer similar prediction performance but unique information for neurobiological interpretation. (A) Scatter plots depicting model performances across deciles as measured by the correlation between observed and predicted values. (B) Canonical network contributions to PNC executive function predictions. Contributions of edges within a single network (diagonals) and between networks (off-diagonals) were defined as the means of edgewise regression coefficients averaged across folds, normalized by network size. Darker colors indicate networks with larger contributions. Network Labels: MF, medial frontal; FP, frontoparietal; DMN, default mode; Mot, motor cortex; VI, visual A; VII, visual B; VAs, visual association; SAL, salience; SC, subcortical; CBL, cerebellum. (C) Node-level contributions to PNC executive function predictions. Node contributions show the sum of all edgewise ridge regression coefficients averaged across folds.

Utility of overlooked features generalizes to psychiatric, developmental, and demographic phenotypic domains

Next, we examined whether overlooked features have relevance in domains beyond language and executive function. First, predictions were performed within a diverse selection of developmental and psychiatric phenotypes in the HBN dataset (n=747 participants, see Methods). For every phenotype, deciles beyond the first retained predictive utility (Figure 4A). Notably, for the Social Communication Questionnaire [39], the second through seventh deciles achieved numerically greater prediction performance (*r*'s=0.15-0.16, MSE's=19.61-19.73, q2's=-0.02) than the first decile (*r*=0.14, q2=-0.03, MSE=19.80). Predictive models of cognitive, developmental, and psychiatric phenotypes generally exhibit low effect sizes. As such, we also tested whether our findings extend to predictions of larger effect size phenotypes by predicting age and sex in the PNC, HCPD, and HBN. Once again, traditionally overlooked feature sets retained significant predictive capabilities for sex and age in all three datasets (Figure 4B, C).

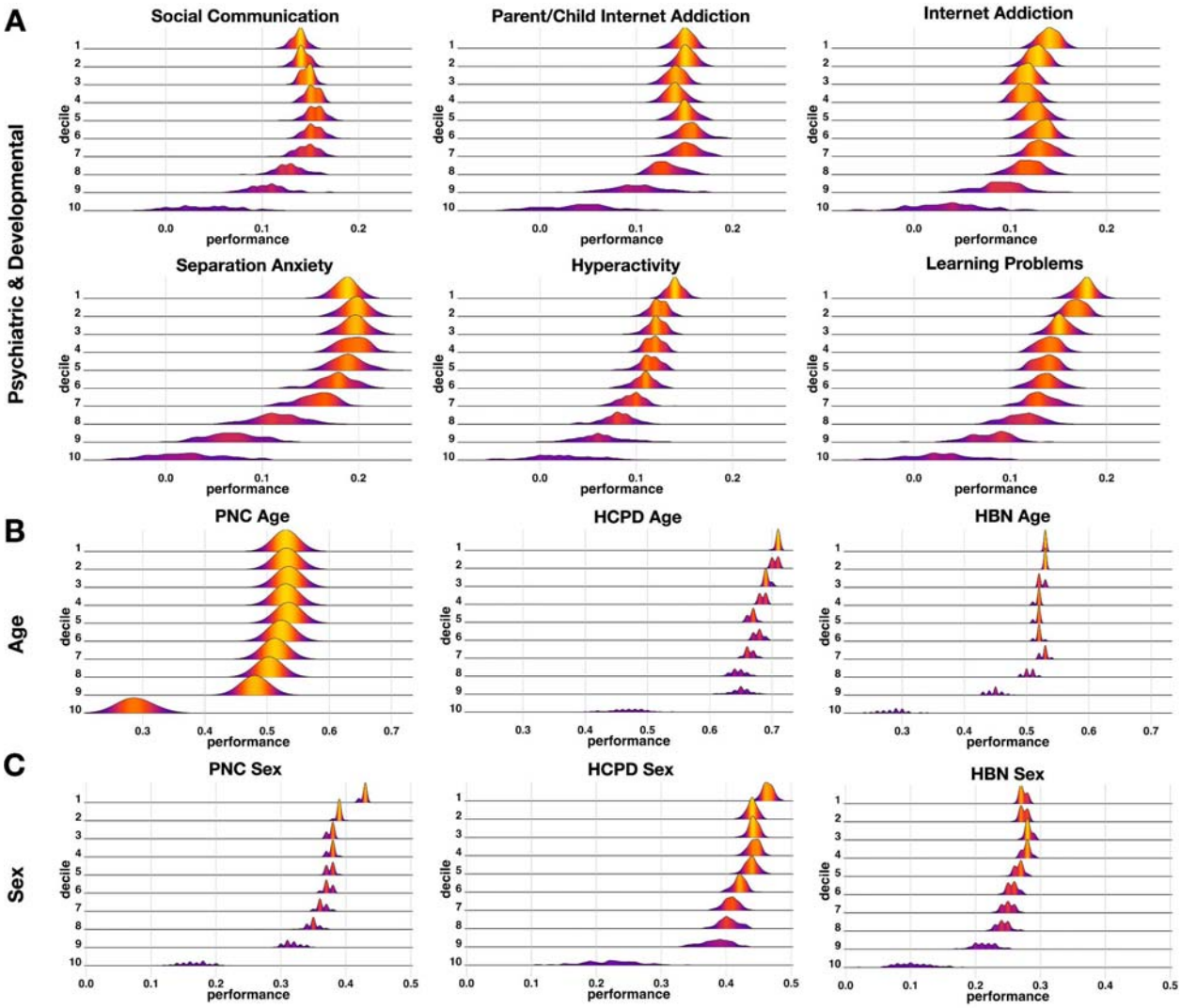

Figure 4. Decile-based predictive modeling performance across psychiatric, developmental, and demographic phenotypes in HBN. (A) Ridgeline plots showing the distribution of prediction performances (Pearson's r) across 100 iterations for the Social Communication Questionnaire (SCQ), Strengths and Difficulties Questionnaire (SDQ) Hyperactivity, Screen for Anxiety Related Disorders (SCARED) Separation Anxiety, Parent-Child Internet Addiction Test (PCIAT), Internet Addiction Test (IAT), and Conners 3 Self-Report (C3SR) Learning Problems scores in HBN. (B) Predictive performance of age for PNC, HCPD, and HBN datasets. (C) Predictive performance of sex in the PNC, HCPD, and HBN datasets.

Overlooked features are also useful in other imaging modalities

We utilized diffusion weighted imaging data from the ABCD study to test whether the predictive utility of lower deciles is conserved across imaging modalities. Structural connectomes were generated based on each fiber's average quantitative anisotropy value connecting two end regions (see Methods). We created functional and structural models for the 6271 ABCD participants with both fMRI and diffusion data to predict NIH Toolbox age-corrected fluid, crystallized, and composite intelligence scores.

Like functional models, structural models retained considerable predictive abilities throughout lower decile feature sets (Figure 5). Diffusion-based models showed a steady, linear decline in predictive performance across deciles, with each decile performing slightly worse than the previous one. This differs from functional data, which maintained relatively stable performance across most deciles, with a sharp drop-off occurring in the lowest deciles. More broadly, functional connectomes generally achieved higher prediction performances than structural models for each phenotype across deciles. DTI-based results were replicable when using the entire ABCD sample (n=9371, Supplementary Figure S7) and when predicting a phenotype of higher effect size in the HBN dataset (age, n=767, Supplementary Figure S8).

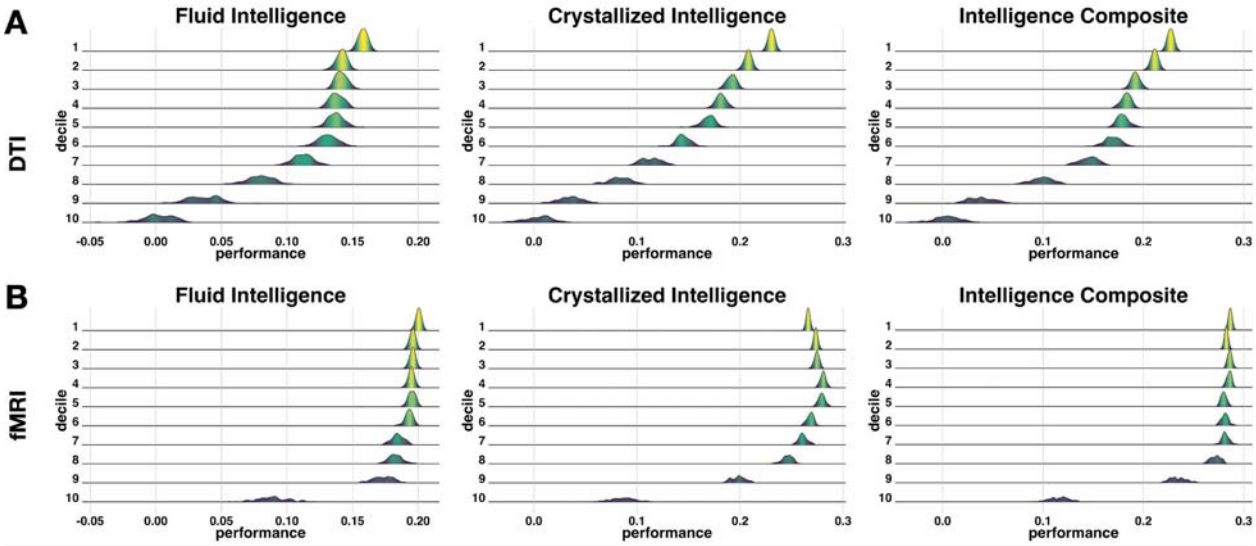

Figure 5. Decile-based predictive modeling performance across DTI and fMRI data in ABCD. Ridgeline plots showing the distribution of prediction performances (Pearson's r) across 100 iterations for NIH Toolbox age-corrected Fluid, Crystallized, and Composite Intelligence scores using diffusion (A) and fMRI (B) data.

Sensitivity Analyses

To evaluate the robustness of our findings, we conducted a series of sensitivity analyses using alternative feature selection and modeling approaches. First, we examined whether the observed predictive performances persisted when using alternative feature groupings. Instead of dividing features into ten deciles, we partitioned them into percentiles (1% subsets), quintiles (20% subsets), and ventiles (5% subsets) based on their correlation with the target phenotype (Supplementary Tables S2, S7, and S8). We also tested whether partitioning features based on alternative statistical metrics influenced results. Features were grouped into non-overlapping bins based on their significance (p-values) or effect size (Pearson's r-values) with the target phenotype (Supplementary Figure S9). We then evaluated prediction performance while controlling for effects of the non-interest variables age, sex, racial/ethnic minority representation,

socioeconomic status, and head motion (Supplementary Table S9). Next, we replaced the connectome-based predictive modeling (CPM) framework with ridge regression, a commonly used alternative that accounts for multicollinearity among predictors (Supplementary Table S10). In each scenario, features subsets beyond those that are most highly ranked demonstrated significant predictions.

Exploratory Analyses

Finally, our results suggest that a given phenotype may have multiple neurobiologically distinct predictive models, possibly reflecting the existence of subtypes—groups of individuals who differ in underlying neurobiology and therefore require different models for accurate prediction. We explored this possibility by investigating which decile feature subsets produced the lowest prediction error for participants (i.e., their 'best fit' model, Supplementary Figure S10). First, we tested whether decile 1 was the feature subset in which the greatest proportion of participants achieved their lowest error term. Decile 1 was not disproportionately associated with participants' lowest error term ($\chi^2(4) = 2.00$, $p = 0.74$). Second, we tested the proportion of participants for whom their lowest error was not achieved in decile 1. The majority of participants exhibited their lowest error in deciles 2 through 5 (74.40% percent) instead of decile 1 (25.60%, $\chi^2(1) = 12.66$, $p < 0.001$). These results highlight the potential presence of subtypes, as evidenced by the top-ranked model not being the best fit for the majority of individuals.

Discussion

Here, we evaluated whether commonly overlooked features are meaningful for neuroimaging predictive model performance and subsequent interpretation. Using a decile-based feature ranking paradigm, we demonstrated that significant predictive capabilities are not exclusive to top-ranked connectivity features. Across functional and structural connectivity data and a wide range of phenotypes and datasets, lower-ranked features achieved comparable prediction accuracy to traditionally emphasized feature sets comprising the most highly ranked features. Notably, lower-ranked features revealed distinct connectivity patterns. Overall, these findings suggest that prevalent practices may lead to overinterpretation and potentially a misrepresentation of the neurobiological bases of brain-behavior associations. Such interpretations present only the 'tip of the iceberg' when certain disregarded features may be just as meaningful.

Our results are aligned with several emerging neuroimaging trends. First, they support the growing appreciation of widely distributed brain networks, as opposed to more localized or simplified networks. Evidence from human [40–43] and animal [44,45] studies demonstrates that the whole brain responds to stimuli, supporting the view that networks—not single nodes—drive behavior and cognition. Because brain-behavior models depend on widely distributed networks, even edges with weak individual correlations to behavior may play a meaningful role when considered as part of a broader functional network. While sparse models are convenient, they may oversimplify the underlying neurobiology of individual variation in behavior.

Next, while predictive models are more replicable than brain-wide association studies [7,10], replicability challenges still exist [5,46,47]. It is not uncommon for multiple of the

same phenotype to implicate different neurobiological processes [19]. Our findings may help explain a portion of this variability. In certain instances of replication failure, the different networks identified across studies may represent select snapshots of a more complex neurobiological landscape rather than true disagreement. This concept aligns with the growing trend toward transparent visualizations, which emphasizes retaining and displaying all data rather than discarding sub-threshold results [6,48]. Thresholded representations that enforce sparsity tend to emphasize differences while obscuring underlying similarities.

Our results also align with recognized differences between multivariate and univariate analyses in neuroimaging [1,49]. For example, univariate features in the connectome (i.e., edges) tend to have poor reliability [50,51]. However, multivariate reliability methods (e.g., fingerprinting, discriminability) show comparably high reliability of connectomes [52–57]. This increased reliability reflects the high-dimensional structure of connectomes. Similarly, selecting a single feature at a time without considering the other features ignores this high-dimensional variance structure [58]. Given selected features, the next best feature to choose would be one with a large effect size and independent of other selected features. When multiple datasets are available, selecting features that are stable across datasets may enhance generalizability of features. Alternatively, approaches that create and select the predictive features, like deep learning, may be paths forward [59,60]. Overall, the limitations of univariate feature selection should be kept in mind when considering the strengths of multivariate methods.

On the surface, it may appear contradictory that higher and lower-ranked features can predict outcomes similarly. This result is likely explained by the high degree of spatial and temporal autocorrelation among brain features [61,62]. When adding new features to a model, accuracy will increase only if the new feature has information that is unique from the existing features. Two features that are highly correlated with a phenotype likely will have a high correlation between them. Thus, pooling lower-ranked (but uncorrelated) features may perform as well or better than pooling highly ranked (but highly correlated) features. Alternatively, predictive models can also inadvertently capture information that is not of interest, such as confounds [18,63], and they are known to perform more poorly for individuals who defy sample stereotypes [17]. Lower-ranked features may be more susceptible to learning from such information. As such, the predictions from lower-ranked features may reflect these properties rather than the underlying neurobiology.

Results from our structural models largely aligned with those generated from functional connectomes; however, while functional model performances were relatively stable until the lowest deciles, diffusion-based model performances showed a more steady, linear decline across deciles. Given that functional connectomes achieved higher prediction performances, one possibility is that the utility of lower deciles persists longer for higher effect size predictions. This pattern could suggest that high effect size phenotypes are distributed more broadly across the brain, or alternatively, that machine learning is more effective at capturing phenotypes with a larger number of contributing features. However, our initial investigation using age prediction with structural connectomes suggests that this explanation does not hold. An

alternative explanation could relate to the low signal-to-noise ratio and high temporal autocorrelation of fMRI time series data, which might yield signal redundancies across deciles and account for why performance remains similar across several deciles. Nevertheless, considerations discussed for functional models likely hold for structural models.

Identifying and defining biologically meaningful clusters (e.g., subtypes) of individuals within a dataset is of great interest within the field [64–67]. Our results suggest that multiple neurobiologically distinct models may capture a given phenotype, each requiring different models [17,68], which has important implications for these ongoing efforts [69–78]. For example, exploratory analyses indicate that one subtype may be best predicted by decile 1, while another subtype may be best predicted by decile 2. If this is the case, the top-ranked feature set identified by a particular analysis could be influenced by the relative proportions of subtypes within the sample. As the distribution of subtypes shifts, the features deemed most predictive could also change. While our preliminary results point to this possibility, further work is needed to validate the extent to which subtype composition affects which features are selected.

Similarly, overlooked, lower-ranked features may be new targets for emerging interventions and strategies for individualized medicine. Traditionally, the highest ranked features from a predictive model are selected as potential targets [32], though these features may not match a specific therapy. For example, it is difficult to stimulate subcortical regions with transcranial magnetic stimulation (TMS) [79,80]. If a model consists of these regions, it would be a poor fit to inform TMS targets. However, overlooked, lower-ranked features may be more anatomically accessible (e.g., cortical

versus subcortical) and still predict the phenotype of interest with the same accuracy as the high-ranked features. Looking beyond the traditional way of selecting features may better align predictive models with therapeutic approaches. Still, it remains unclear which features are the best targets for intervention. Brain features that predict inter-individual variation in symptoms or treatment outcomes are not necessarily the same as those causally involved in symptom manifestation [81,82].

A strength of this study lies in the comprehensive validation of our models across diverse datasets and methodological frameworks. We leveraged four large-scale neuroimaging datasets to enhance statistical power and generalizability. The PNC, HBN, HCPD, and ABCD datasets exhibit notable variability in key aspects of study design including recruitment geography, behavioral assessment, imaging acquisition, participant demographics, and clinical symptom burden [25]. Demonstrating the resilience of our models to differences in participant demographics and study design, the utility of commonly disregarded features survived external validation. Neuroimaging predictive models will achieve real-world utility only if they can overcome these dataset-specific idiosyncrasies (i.e., “dataset shifts”) [25,46,83–85]. Finally, we also performed an extensive set of sensitivity analyses. Our results were robust to both functional and structural imaging modalities and across a range of cognitive, psychiatric, and developmental phenotypic domains.

There were several limitations to our study. First, we pursued a network-centric approach focusing on the connectivity strengths (e.g., edges) between 268 brain regions (e.g., nodes) as features. Node-based models, which may capture local brain dynamics or region-specific activity patterns that are obscured in connectome-based

approaches, could reveal different results. However, connectome-based models are more frequently employed in the field, in part due to superior predictive abilities [86]. Next, it is unclear whether overlooked features maintain their predictive utility across phenotypes that are hypothesized to rely on more circumscribed functional networks, such as reaction time. Finally, this investigation relies on developmental datasets exclusively from the United States. Future work should investigate the utility of commonly overlooked features in adult populations and those from non-Western countries [87,88].

Collectively, our findings highlight that commonly overlooked features may possess comparable predictive and neurobiological significance. They support the growing appreciation of widely distributed brain networks, as opposed to more localized or simplified networks. Our findings also point to the potential presence of subtypes in the connectivity data, wherein different sets of features represent the best model for different groups of individuals. Overall, better understanding and characterization of overlooked features will help improve the generalizability of predictive models.

Methods

Datasets

PNC participants were 1291 individuals ages 8-21 recruited from the greater Philadelphia, Pennsylvania area [37]. All imaging was performed on a single scanner at the Hospital of the University of Pennsylvania. Participants completed rest, emotion task, and n-back task fMRI runs [89]. Measures of language abilities were the Penn Verbal Reasoning Task from the Penn Computerized Neurocognitive Battery (CNB) and the total standard score from the Wide Range Assessment Test (WRAT) Reading Subscale [90,91]. Executive function measures were the Letter N-Back, Conditional Exclusion, and Continuous Performance tasks from the CNB.

HBN participants were 1110 individuals ages 6-17 recruited from the New York City, New York region [34]. Imaging for this multi-site study was performed via the HBN mobile MRI scanner in Staten Island and at the Rutgers University Brain Imaging Center, the CitiGroup Cornell Brain Imaging Center, and the CUNY Advanced Science Research Center. Participants completed two rest fMRI runs as well as 'Despicable Me' and 'The Present' movie-watching scan sessions. Measures of language abilities were the Elision, Blending Words, Nonword Repetition, Rapid Digit Naming, and Rapid Letter Naming scaled scores from the Comprehensive Test of Phonological Processing (CTOPP-2) and the Phonemic Decoding Efficiency, Sight Word Efficiency, and Total Word Reading Efficiency scaled scores from the Test of Word Reading Efficiency (TOWRE-2) [92,93]. Executive function measures were the Flanker Inhibitory Control and Attention, List Sorting Working Memory, Pattern Comparison Processing Speed, and Dimensional Change Card Sort age-corrected standard scores from the NIH Toolbox [94].

To select developmental and psychiatric phenotypes of meaningful effect size, we first performed predictions using the original CPM framework. We identified scales and subscales with prediction performance $r>0.10$ by testing predictions on the 575 HBN participants who had complete data for all 71 measures. This resulted in 13 scales and subscales. We repeated the predictions using the 747 participants who had data available for all 13 scales. Of these, 6 scales achieved prediction performances of $r>=0.15$, which were used for our decile-based prediction paradigm: the Social Communication Questionnaire (SCQ) [39] score, the Hyperactivity Scale score from the Strengths and Difficulties Questionnaire (SDQ) [95,96], the Separation Anxiety score from the Screen for Anxiety Related Disorders (SCARED) [97], the Learning Problems score from the Conners 3 Self-Report (C3SR) [98], and the Internet Addiction Test (IAT) and Parent-Child Internet Addiction Test (PCIAT) [99] scores.

HCPD participants were 428 individuals ages 8-22 recruited from St. Louis, Missouri, Twin Cities, Minnesota, Boston, Massachusetts, and Los Angeles, California [36]. Imaging for this multi-site study was performed at Harvard University, the University of California-Los Angeles, the University of Minnesota, and Washington University in St. Louis. Participants completed rest fMRI runs [100]. Measures of language abilities were the Picture Vocabulary and Oral Reading Recognition age-corrected standard scores from the NIH Toolbox. Executive function measures were the Flanker Inhibitory Control and Attention, List Sorting Working Memory, Pattern Comparison Processing Speed, Dimensional Change Card Sort, and Picture Sequence Memory age-corrected standard scores from the NIH Toolbox.

ABCD participants were 9371 individuals ages 9-10 recruited from 21 sites across the United States: Children's Hospital Los Angeles, Florida International University, Laureate Institute for Brain Research, Medical University of South Carolina, Oregon Health & Science University, SRI International, UC San Diego, UCLA, University of Colorado Boulder, University of Florida, University of Maryland at Baltimore, University of Michigan, University of Minnesota, University of Pittsburgh, University of Rochester, University of Utah, University of Vermont, University of Wisconsin-Milwaukee, Virginia Commonwealth University, Washington University in St. Louis, and Yale University [35,101]. Cognitive measures were the Fluid Intelligence, Crystalized Intelligence, and Intelligence Composite age-corrected scores from the NIH Toolbox.

fMRI data preprocessing

In all datasets, data were motion-corrected. Additional preprocessing steps were performed using BioImage Suite [102]. This included regression of covariates of no interest from the functional data, including linear and quadratic drifts, mean cerebrospinal fluid signal, mean white matter signal, and mean global signal. Additional motion control was applied by regressing a 24-parameter motion model, which included six rigid body motion parameters, six temporal derivatives, and the square of these terms, from the data. Subsequently, we applied temporal smoothing with a Gaussian filter (approximate cutoff frequency=0.12 Hz) and gray matter masking, as defined in common space. Then, the Shen 268-node atlas was applied to parcellate the denoised

data into 268 nodes [103]. Finally, we generated functional connectivity matrices by correlating each node time series data pair and applying the Fisher transform.

Data were excluded for poor data quality, missing nodes due to lack of full brain coverage, high motion (>0.2 mm mean frame-wise motion), or missing behavioral/phenotypic data. Each participant's connectome included all available resting-state and task fMRI data with low motion (<0.2 mm). Combining connectomes across fMRI data improves reliability and predictive power [15,104]. Participants without one low-motion fMRI run were excluded. For PNC, 246 participants were excluded due to image quality or motion, and 61 participants were excluded due to incomplete phenotypic data. For HBN, 1387 participants were excluded due to image quality or motion, and 829 participants were excluded due to incomplete phenotypic data. For HCPD, 57 participants were excluded due to image quality or motion, and 167 participants were excluded due to incomplete phenotypic data. For ABCD, 342 participants were excluded due to incomplete phenotypic data.

ABCD diffusion data preprocessing

FIB files were downloaded from https://brain.labsolver.org. As described on https://brain.labsolver.org/, a multishell diffusion scheme was used, and the b-values were 500, 1000, 2000, and 3000 s/mm². The number of diffusion sampling directions were 6, 15, 15, and 60, respectively. The in-plane resolution was 1.7 mm. The slice thickness was 1.7 mm. The diffusion MRI data were rotated to align with the AC-PC line at an isotropic resolution of 1.7 (mm). The restricted diffusion was quantified using restricted diffusion imaging [105]. The diffusion data were reconstructed using generalized

q-sampling imaging [106] with a diffusion sampling length ratio of 1.25. The tensor metrics were calculated using DWI with b-value lower than 1750 s/mm².

As stated in previous works and duplicated here for consistency [107,108], whole-brain fibre tracking was conducted with DSI-Studio with quantitative anisotropy (QA) as the termination threshold. QA values were computed in each voxel in their native space for each subject and were then used to warp the brain to the template in Montreal Neurological Institute (MNI) space using the statistical parametric mapping nonlinear registration algorithm. Once in MNI space, spin density functions were again reconstructed with a mean diffusion distance of 1.25 mm using three fibre orientations per voxel. Fibre tracking was performed in DSI-Studio with an angular cut-off of 60 degrees, step size of 1.0 mm, minimum length of 30 mm, spin density function smoothing of 0.0, maximum length of 300 mm and a QA threshold determined by the diffusion-weighted imaging signal. Deterministic fibre tracking using a modified FACT algorithm [109] was performed until 10,000,000 streamlines were reconstructed for each individual. We used the Shen atlas[103] in MNI space with 268 nodes to construct individual structural connectomes: the pairwise connectivity strength was calculated as the average QA value of each fibre connecting the two end regions and thresholded at 0.001, which results in a 268 × 268 adjacency matrix for each participant.

Creating latent factors of language abilities and executive function

Whether machine learning models are used for real-world prediction or advancing our understanding of neurobiology [110,111], it is important to overcome dataset-specific idiosyncrasies [1]. The PNC, HCPD, and HBN datasets employed disparate

measures to assess executive function and language abilities. To harmonize these measures and facilitate direct comparisons across datasets, we used principal component analysis (PCA) to derive 'latent' factors of executive function and language abilities within each dataset. Briefly, as in previous work [25], PCA was applied to behavioral measures to reduce measurement noise and derive a single composite factor per domain. For PNC and HBN, PCA was conducted using participants without usable neuroimaging data to maintain independent train and test groups, while for HCPD, PCA was performed within each fold of cross-validation, using only training data. Behavioral data from participants with imaging data were projected onto the first principal component, which explained 70%, 55%, and 77% of the variance in language measures and 53%, 48%, and 40% in EF measures for PNC, HBN, and HCPD, respectively.

Decile-based connectome-based predictive modeling

We implemented an original adaptation of the original CPM framework (Shen et al., 2017) to evaluate the predictive utility of connectivity features across varying levels of association with the target phenotype. For within-dataset predictions, we performed 100 iterations of a 10-fold cross validation scheme. Within each fold, the model was trained in approximately 90% of the data and tested in the remaining 10% of the data. For each target phenotype, within each fold, we computed the Pearson correlation coefficient between every connectivity edge and the target phenotype across all participants in the training subset. This univariate analysis provided a measure of association strength for each feature. Features were then ranked in descending order

based on the absolute value of their correlation coefficients, ensuring that both positive and negative associations were considered equally in the ranking process. Following ranking, the complete set of connectivity features was partitioned into ten non-overlapping deciles. Each decile represented 10% of the total features, with decile 1 comprising the top 10% of features exhibiting the strongest associations with the target phenotype, and decile 10 containing the bottom 10%. Predictions were performed using each decile and tested on the fold's test set. For sensitivity analyses, independent predictive models were created to account for non-interest variables (age, sex, racial/ethnic minority representation, socioeconomic status, head motion, and clinical symptom burden) via partial correlation at the step in which brain edges were related to each phenotype [112].

Ridge regression connectome-based predictive modeling

We repeated analyses using ridge regression CPM to better suit the high-dimensional nature of connectivity data [15]. Specifically, due to the positive semi-definite nature of a functional connectivity matrix, the edges are not independent. Ridge regression is more robust than OLS in this case. Instead of summing selected edges and fitting a one-dimensional OLS model, we directly fit a ridge regression model with training individuals using the selected edges from all the tasks and apply the model to testing individuals in the cross-validation framework. The L2 regularization λ parameter was chosen by an inner 10-fold cross-validation which uses only the training individuals. The largest λ value with a mean squared error (MSE) within one standard error of the minimum MSE is chosen.

Model performance

Predictive models were trained and tested within each dataset using 100 iterations of 10-fold cross-validation. Model performance was evaluated with Pearson's correlation (r), representing the correspondence between predicted and actual behavioral scores. We report mean Pearson's r across the 100 iterations, along with the cross-validation coefficient of determination (q2) and mean square error (MSE) [5]. Within-dataset plots show data for the median-performing models. To generate null distributions for significance testing, we randomly shuffled the correspondence between behavioral variables and connectivity matrices 1000 times and re-ran the CPM analysis with the shuffled data. Based on these null distributions, the p-values for predictions were calculated as in prior work [113]. Only a positive association between predicted and actual values indicates prediction above chance (with negative associations indicating a failure to predict), so one-tailed p-values are reported. Pearson's correlation was tested between actual and predicted values to evaluate cross-dataset predictions.

Model contribution

Predictive networks identified using CPM are complex and composed of multiple brain regions and networks. To quantify the contribution of each edge to a given predictive model, we calculated the kth edge's weight (labeled $W_{k,}$) to the model as: $W_k = abs(\beta^k)/std(E_k)$, where $std(E_k)$ represents the standard deviation of the kth edge, and β^k represents the weight learned by CPM for the kth edge. To quantify the contribution of each node to a given predictive model, we calculated the nth node's

weight summed across all edges (labeled W_n) to the model as: $W_n = \sum_{k=1}^{35,778} W_k$, for all k edges connected to the nth node. Next, for the network level, W_k was averaged over each edge within or between canonical functional networks.

Data availability

Data are available through the Philadelphia Neurodevelopmental Cohort, Human Connectome Project in Development, Healthy Brain Network, and Adolescent Brain Cognitive Development Study datasets.

References

1. Genon, S., Eickhoff, S. B. & Kharabian, S. Linking interindividual variability in brain structure to behaviour. *Nat. Rev. Neurosci.* **23**, 307–318 (2022).
2. Sui, J., Jiang, R., Bustillo, J. & Calhoun, V. Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biol. Psychiatry* **88**, 818–828 (2020).
3. Williams, L. M. & Whitfield Gabrieli, S. Neuroimaging for precision medicine in psychiatry. *Neuropsychopharmacology* **50**, 246–257 (2024).
4. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* **20**, 365–377 (2017).
5. Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry* **77**, 534–540 (2020).
6. Allen, E. A., Erhardt, E. B. & Calhoun, V. D. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron* **74**, 603–608 (2012).
7. 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).
8. Scheinost, D. *et al.* Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* **193**, 35–45 (2019).
9. Goltermann, J. *et al.* Cross-validation for the estimation of effect size generalizability in mass-univariate brain-wide association studies. 2023.03.29.534696 Preprint at https://doi.org/10.1101/2023.03.29.534696 (2023).
10. Spisak, T., Bingel, U. & Wager, T. D. Multivariate BWAS can be replicable with moderate sample sizes. *Nature* **615**, E4–E7 (2023).
11. Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci. J. Assoc. Psychol. Sci.* **12**, 1100–1122

(2017).

12. Greene, A. S. & Constable, R. T. Clinical Promise of Brain-Phenotype Modeling: A Review. *JAMA Psychiatry* **80**, 848 (2023).
13. 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).
14. Dhamala, E., Yeo, B. T. T. & Holmes, A. J. One Size Does Not Fit All: Methodological Considerations for Brain-Based Predictive Modeling in Psychiatry. *Biol. Psychiatry* **93**, 717–728 (2023).
15. Gao, S., Greene, A. S., Constable, R. T. & Scheinost, D. Combining multiple connectomes improves predictive modeling of phenotypic measures. *NeuroImage* **201**, 116038 (2019).
16. Rosenblatt, M. *et al.* Connectome-based machine learning models are vulnerable to subtle data manipulations. *Patterns* **4**, (2023).
17. Greene, A. S. *et al.* Brain-phenotype models fail for individuals who defy sample stereotypes. *Nature* **609**, 109–118 (2022).
18. Li, J. *et al.* Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* **8**, eabj1812 (2022).
19. Botvinik-Nezer, R. *et al.* Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* **582**, 84–88 (2020).
20. Kang, K. *et al.* Study design features increase replicability in brain-wide association studies. *Nature* **636**, 719–727 (2024).
21. Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M. & Mills, K. L. Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Dev. Cogn. Neurosci.* **47**, 100902 (2020).
22. Liu, S., Abdellaoui, A., Verweij, K. J. H. & van Wingen, G. A. Replicable brain–phenotype

associations require large-scale neuroimaging data. *Nat. Hum. Behav.* **7**, 1344–1356 (2023).

23. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
24. Rosenblatt, M. *et al.* Power and reproducibility in the external validation of brain-phenotype predictions. *Nat. Hum. Behav.* 1–16 (2024) doi:10.1038/s41562-024-01931-7.
25. Adkinson, B. D. *et al.* Brain-phenotype predictions of language and executive function can survive across diverse real-world data: Dataset shifts in developmental populations. *Dev. Cogn. Neurosci.* **70**, 101464 (2024).
26. Marek, S. & Laumann, T. O. Replicability and generalizability in population psychiatric neuroimaging. *Neuropsychopharmacology* **50**, 52–57 (2024).
27. Rosenberg, M. D. & Finn, E. S. How to establish robust brain–behavior relationships without thousands of individuals. *Nat. Neurosci.* **25**, 835–837 (2022).
28. Wu, J. *et al.* Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *NeuroImage* **262**, 119569 (2022).
29. Jiang, R., Woo, C.-W., Qi, S., Wu, J. & Sui, J. Interpreting Brain Biomarkers: Challenges and solutions in interpreting machine learning-based predictive neuroimaging. *IEEE Signal Process. Mag.* **39**, 107–118 (2022).
30. Kohoutová, L. *et al.* Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* **15**, 1399–1435 (2020).
31. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
32. Tejavibulya, L. *et al.* Predicting the future of neuroimaging predictive models in mental health. *Mol. Psychiatry* **27**, 3129–3137 (2022).
33. Mwangi, B., Tian, T. S. & Soares, J. C. A review of feature reduction techniques in

neuroimaging. *Neuroinformatics* **12**, 229–244 (2014).

34. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* **4**, 170181 (2017).
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. Somerville, L. H. *et al.* The Lifespan Human Connectome Project in Development: A large-scale study of brain connectivity development in 5–21 year olds. *NeuroImage* **183**, 456–468 (2018).
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. Shen, X. *et al.* Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518 (2017).
39. Chandler, S. *et al.* Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders. *J. Am. Acad. Child Adolesc. Psychiatry* **46**, 1324–1332 (2007).
40. Dubois, J., Galdi, P., Paul, L. K. & Adolphs, R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170284 (2018).
41. Gonzalez-Castillo, J. *et al.* Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl. Acad. Sci.* **109**, 5487–5492 (2012).
42. Noble, S., Curtiss, J., Pessoa, L. & Scheinost, D. The tip of the iceberg: A call to embrace anti-localizationism in human neuroscience research. *Imaging Neurosci.* **2**, 1–10 (2024).
43. Westlin, C. *et al.* Improving the study of brain-behavior relationships by revisiting basic

assumptions. *Trends Cogn. Sci.* **27**, 246–257 (2023).

44. Machado, T. A., Kauvar, I. V. & Deisseroth, K. Multiregion neuronal activity: the forest and the trees. *Nat. Rev. Neurosci.* **23**, 683–704 (2022).
45. Tremblay, S., Testard, C., Inchauspé, J. & Petrides, M. Epiphenomenal neural activity in the primate cortex. 2022.09.12.506984 Preprint at https://doi.org/10.1101/2022.09.12.506984 (2023).
46. Dockès, J., Varoquaux, G. & Poline, J.-B. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience* **10**, giab055 (2021).
47. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2020).
48. Taylor, P. A. *et al.* Highlight results, don't hide them: Enhance interpretation, reduce biases and improve reproducibility. *NeuroImage* **274**, 120138 (2023).
49. Makowski, C. *et al.* Leveraging the adolescent brain cognitive development study to improve behavioral prediction from neuroimaging in smaller replication samples. *Cereb. Cortex* **34**, bhae223 (2024).
50. Tozzi, L., Fleming, S. L., Taylor, Z. D., Raterink, C. D. & Williams, L. M. Test-retest reliability of the human functional connectome over consecutive days: identifying highly reliable portions and assessing the impact of methodological choices. *Netw. Neurosci.* **4**, 925–945 (2020).
51. Wang, Y. *et al.* Intra-session test-retest reliability of functional connectivity in infants. *NeuroImage* **239**, 118284 (2021).
52. Amico, E. & Goñi, J. The quest for identifiability in human functional connectomes. *Sci. Rep.* **8**, 8254 (2018).
53. Graff, K. *et al.* Benchmarking common preprocessing strategies in early childhood functional connectivity and intersubject correlation fMRI. *Dev. Cogn. Neurosci.* **54**, 101087 (2022).

54. Horien, C., Shen, X., Scheinost, D. & Constable, R. T. The individual functional connectome is unique and stable over months to years. *NeuroImage* **189**, 676–687 (2019).

55. Kaufmann, T. *et al.* Delayed stabilization and individualization in connectome development are related to psychiatric disorders. *Nat. Neurosci.* **20**, 513–515 (2017).

56. Noble, S., Scheinost, D. & Constable, R. T. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage* **203**, 116157 (2019).

57. Vanderwal, T. *et al.* Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage* **157**, 521–530 (2017).

58. Sizemore, A. E. *et al.* Cliques and cavities in the human connectome. *J. Comput. Neurosci.* **44**, 115–145 (2018).

59. Anbarasi, J., Kumari, R., Ganesh, M. & Agrawal, R. Translational Connectomics: overview of machine learning in macroscale Connectomics for clinical insights. *BMC Neurol.* **24**, 364 (2024).

60. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

61. Burt, J. B., Helmer, M., Shinn, M., Anticevic, A. & Murray, J. D. Generative modeling of brain maps with spatial autocorrelation. *NeuroImage* **220**, 117038 (2020).

62. Shinn, M. *et al.* Functional brain networks reflect spatial and temporal autocorrelation. *Nat. Neurosci.* **26**, 867–878 (2023).

63. Snoek, L., Miletić, S. & Scholte, H. S. How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage* **184**, 741–760 (2019).

64. Brucar, L. R., Feczko, E., Fair, D. A. & Zilverstand, A. Current Approaches in Computational Psychiatry for the Data-Driven Identification of Brain-Based Subtypes. *Biol. Psychiatry* **93**, 704–716 (2023).

65. Hong, S.-J. *et al.* Toward Neurosubtypes in Autism. *Biol. Psychiatry* **88**, 111–128 (2020).

66. Pettorruso, M. *et al.* Addiction biotypes: a paradigm shift for future treatment strategies?

Mol. Psychiatry **29**, 1450–1452 (2024).

67. Williams, L. M. Defining biotypes for depression and anxiety based on large-scale circuit dysfunction: a theoretical review of the evidence and future directions for clinical translation. *Depress. Anxiety* **34**, 9–24 (2017).
68. Zhao, Q. *et al.* The Transition From Homogeneous to Heterogeneous Machine Learning in Neuropsychiatric Research. *Biol. Psychiatry Glob. Open Sci.* **5**, 100397 (2024).
69. Buch, A. M. *et al.* Molecular and network-level mechanisms explaining individual differences in autism spectrum disorder. *Nat. Neurosci.* **26**, 650–663 (2023).
70. Chen, P. *et al.* Four Distinct Subtypes of Alzheimer's Disease Based on Resting-State Connectivity Biomarkers. *Biol. Psychiatry* **93**, 759–769 (2023).
71. Dinga, R. *et al.* Evaluating the evidence for biotypes of depression: Methodological replication and extension of. *NeuroImage Clin.* **22**, 101796 (2019).
72. Drysdale, A. T. *et al.* Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* **23**, 28–38 (2017).
73. Esterman, M. *et al.* Evaluating the evidence for a neuroimaging subtype of posttraumatic stress disorder. *Sci. Transl. Med.* **12**, eaaz9343 (2020).
74. Etkin, A. *et al.* Using fMRI connectivity to define a treatment-resistant form of post-traumatic stress disorder. *Sci. Transl. Med.* **11**, eaal3236 (2019).
75. Jiang, Y. *et al.* Neuroimaging biomarkers define neurophysiological subtypes with distinct trajectories in schizophrenia. *Nat. Ment. Health* **1**, 186–199 (2023).
76. Murray, L., Frederick, B. B. & Janes, A. C. Data-driven connectivity profiles relate to smoking cessation outcomes. *Neuropsychopharmacology* **49**, 1007–1013 (2024).
77. Pagani, M. *et al.* Biological subtyping of autism via cross-species fMRI. 2025.03.04.641400 Preprint at https://doi.org/10.1101/2025.03.04.641400 (2025).
78. Wen, Z. *et al.* Neuroimaging-based variability in subtyping biomarkers for psychiatric

heterogeneity. *Mol. Psychiatry* 1–10 (2024) doi:10.1038/s41380-024-02807-y.

79. Bolognini, N. & Ro, T. Transcranial Magnetic Stimulation: Disrupting Neural Activity to Alter and Assess Brain Function. *J. Neurosci.* **30**, 9647–9650 (2010).
80. Luber, B. *et al.* Using diffusion tensor imaging to effectively target TMS to deep brain structures. *NeuroImage* **249**, 118863 (2022).
81. McPartland, J. C. Developing Clinically Practicable Biomarkers for Autism Spectrum Disorder. *J. Autism Dev. Disord.* **47**, 2935–2937 (2017).
82. Yip, S. W., Kiluk, B. & Scheinost, D. Toward Addiction Prediction: An Overview of Cross-Validated Predictive Modeling Findings and Considerations for Future Neuroimaging Research. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 748–758 (2020).
83. Chekroud, A. M. *et al.* Illusory generalizability of clinical prediction models. *Science* **383**, 164–167 (2024).
84. Kopal, J., Uddin, L. Q. & Bzdok, D. The end game: respecting major sources of population diversity. *Nat. Methods* **20**, 1122–1128 (2023).
85. Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D. *Dataset Shift in Machine Learning*. (MIT Press, 2022).
86. Horien, C., Greene, A. S., Constable, R. T. & Scheinost, D. Regions and Connections: Complementary Approaches to Characterize Brain Organization and Function. *The Neuroscientist* **26**, 117–133 (2020).
87. Bzdok, D., Wolf, G. & Kopal, J. Harnessing population diversity: in search of tools of the trade. *GigaScience* **13**, giae068 (2024).
88. Ge, J. *et al.* Increasing diversity in connectomics with the Chinese Human Connectome Project. *Nat. Neurosci.* **26**, 163–172 (2023).
89. Satterthwaite, T. D. *et al.* Neuroimaging of the Philadelphia neurodevelopmental cohort. *NeuroImage* **86**, 544–553 (2014).

90. Gur, R. C. *et al.* A cognitive neuroscience based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. *J. Neurosci. Methods* **187**, 254–262 (2010).

91. Wilkinson, G. S. & Robertson, G. J. Wide Range Achievement Test 4. (2006) doi:10.1037/t27160-000.

92. Dickens, R. H., Meisinger, E. B. & Tarar, J. M. Test Review: Comprehensive Test of Phonological Processing–2nd ed. (CTOPP-2) by Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. *Can. J. Sch. Psychol.* **30**, 155–162 (2015).

93. Tarar, J. M., Meisinger, E. B. & Dickens, R. H. Test Review: Test of Word Reading Efficiency–Second Edition (TOWRE-2) by Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. *Can. J. Sch. Psychol.* **30**, 320–326 (2015).

94. Weintraub, S. *et al.* Cognition assessment using the NIH Toolbox. *Neurology* **80**, S54–S64 (2013).

95. Goodman, A., Lamping, D. L. & Ploubidis, G. B. When to Use Broader Internalising and Externalising Subscales Instead of the Hypothesised Five Subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British Parents, Teachers and Children. *J. Abnorm. Child Psychol.* **38**, 1179–1191 (2010).

96. Goodman, R. The Strengths and Difficulties Questionnaire: a research note. *J. Child Psychol. Psychiatry* **38**, 581–586 (1997).

97. Birmaher, B. *et al.* Psychometric properties of the Screen for Child Anxiety Related Emotional Disorders (SCARED): a replication study. *J. Am. Acad. Child Adolesc. Psychiatry* **38**, 1230–1236 (1999).

98. Conners, C.K. *Conners' Rating Scales-Revised*. (North Multi-Health Systems, Inc., Tonawanda, NY, 2001).

99. Young, K. S. Internet addiction: The emergence of a new clinical disorder. *Cyberpsychol.*

Behav. **1**, 237–244 (1998).

100. Harms, M. P. *et al.* Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *NeuroImage* **183**, 972–984 (2018).
101. Garavan, H. *et al.* Recruiting the ABCD sample: Design considerations and procedures. *Dev. Cogn. Neurosci.* **32**, 16–22 (2018).
102. Papademetris, X. *et al.* BioImage Suite: An integrated medical image analysis suite: An update. *Insight J.* **2006**, 209 (2006).
103. 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).
104. Elliott, M. L. *et al.* General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *Neuroimage* **189**, 516–532 (2019).
105. Yeh, F.-C., Liu, L., Hitchens, T. K. & Wu, Y. L. Mapping immune cell infiltration using restricted diffusion MRI. *Magn. Reson. Med.* **77**, 603–612 (2017).
106. Yeh, F.-C., Wedeen, V. J. & Tseng, W.-Y. I. Generalized q-sampling imaging. *IEEE Trans. Med. Imaging* **29**, 1626–1635 (2010).
107. Gu, S. *et al.* The Energy Landscape of Neurophysiological Activity Implicit in Brain Network Structure. *Sci. Rep.* **8**, 2507 (2018).
108. Sun, H. *et al.* Network controllability of structural connectomes in the neonatal brain. *Nat. Commun.* **14**, 5820 (2023).
109. Yeh, F.-C., Verstynen, T. D., Wang, Y., Fernández-Miranda, J. C. & Tseng, W.-Y. I. Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PloS One* **8**, e80713 (2013).

110. Adkinson, B. D. & Chekroud, A. M. Broadening the Use of Machine Learning in Psychiatry. *Biol. Psychiatry* **93**, 4–5 (2023).

111. Chen, J., Patil, K. R., Yeo, B. T. T. & Eickhoff, S. B. Leveraging Machine Learning for Gaining Neurobiological and Nosological Insights in Psychiatric Research. *Biol. Psychiatry* **93**, 18–28 (2023).

112. Hsu, W.-T., Rosenberg, M. D., Scheinost, D., Constable, R. T. & Chun, M. M. Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals. *Soc. Cogn. Affect. Neurosci.* **13**, 224–232 (2018).

113. Garrison, K. A. *et al.* Transdiagnostic Connectome-Based Prediction of Craving. *Am. J. Psychiatry* **180**, 445–453 (2023).