Information limits of standard task and movie fMRI for mental health evaluation

Cole Korponay, PhD[1,5], Julia E. Cohen-Gilbert, PhD[1,5], Poornima Kumar, PhD[2,5], Nathaniel G. Harnett, PhD[4,5], Adrian A. Medina, MPH, EdM[1], You Cheng, PhD[1,5], Brent P. Forester MD, MSc[6], Kerry J. Ressler MD, PhD[4,5], Jure Demsar[10,11], Blaise B. Frederick[1,5], Christian F. Beckmann, PhD[7,8,9], David G. Harper, PhD[3,5] & Lisa D. Nickerson, PhD[1,5]

[1] *McLean Imaging Center, McLean Hospital, Belmont, MA*
[2] *Center for Depression and Stress Research, McLean Hospital, Belmont, MA*
[3] *Division for Geriatric Psychiatry, McLean Hospital, Belmont, MA*
[4] *Division of Depression and Anxiety Disorders, McLean Hospital, Belmont, MA*
[5] *Department of Psychiatry, Harvard Medical School, Boston, MA*
[6] *Department of Psychiatry, Tufts University School of Medicine, Boston, MA*
[7] *Donders Institute of Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands*
[8] *Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, The Netherlands*
[9] *Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK*
[10] *Faculty of Computer and Information Science, University of Ljubljana, Slovenia*
[11] *Department of Psychology, Faculty of Arts, University of Ljubljana, Slovenia*

Correspondence: Cole Korponay, McLean Imaging Center, McLean Hospital, Belmont, MA 02478, USA ckorponay@mclean.harvard.edu; Lisa Nickerson, McLean Imaging Center, McLean Hospital, Belmont, MA 02478, USA lisa_nickerson@hms.harvard.edu.

Abstract

Robust brain-based mental health biomarkers remain largely elusive. One line of thought attributes this to suboptimal quality and modeling of brain and behavioral data. However, an alternative explanation is that neural activity evoked by common brain imaging paradigms simply reflects little about mental health. Here, we find evidence for this latter explanation by examining multivariate mental health and brain function profiles in hundreds of individuals using high-reliability, latent mental health factors derived from 87 neurocognitive/psychiatric assessments, and tensor independent component analysis of five common functional magnetic resonance imaging (fMRI) task and movie-watching paradigms with putative mental health relevance. Across all tasks and brain networks, individual differences in evoked brain activity poorly reflected variability in negative affect, positive affect, or substance use – despite robustly reflecting variability in cognition. Moreover, across-subject diversity in task-evoked brain function profiles significantly lagged across-subject diversity in mental health profiles. Finally, clustering subjects by their brain function profile versus by their mental health profile produced discordant subtypes, and significantly modulated the findings of group-difference analyses. Findings empirically bound the recoverable information about non-cognitive mental health in normative samples using standard task batteries and motivate use of alternative paradigms with higher sensitivity to individual-specific affective and motivational signals.

Introduction

Our experiences of and responses to cognitive challenges, positive and negative emotions, reward and loss, and social interactions are core determinants and reflections of mental health. In 2009, the Research Domain Criteria (RDoC) framework[1] was introduced to reconceptualize mental health in terms of individual variation in these and other core domains of neurobehavioral functioning, moving away from symptom-based diagnostic categories that increasingly demonstrated questionable neurobiological validity[1,2]. Ultimately, the RDoC framework aims to facilitate more biologically-grounded research and treatment efforts in the field of mental health[1]. As such, a major goal of recent work has been to identify the neurobiological features whose variability between individuals maps to and predicts interindividual variability in behavioural and self-report measures of mental health within RDoC functional domains – both within and beyond normative ranges of functioning. These domains include the cognitive system (CS) encompassing constructs like memory, attention and perception; positive valence system (PVS) comprising reward responsiveness, reward learning and reward valuation; negative valence system (NVS) of fear, threat and loss processing-related constructs, and social processes (SP) including affiliation and attachment, social communication and theory of mind[3].

Among the neuroimaging modalities and imaging derived phenotypes (IDPs) available to quantify structural and functional features of the brain, many studies,[4-7] though not all[8,9], find that functional magnetic resonance imaging (fMRI)-derived IDPs display the strongest relationships with behavioural and neuropsychiatric phenotypes. To date, studies using fMRI have identified links of moderate strength between individual differences in brain activity and individual differences in cognition[10-15]. Group-level fMRI research has robustly mapped neural substrates for behavioural and self-report measures of PVS[16], NVS[17], and SP[18], however, links between individual differences in these units of analysis within the PVS, NVS and SP domains remain largely elusive and well-short of clinical applicability[7,19-21]. Even as many studies achieve statistically significant brain function-mental health associations or predictions, findings are generally of small to modest effect size[7,22,23], and are frequently not reproducible[24,25].

Emerging recognition of these weak associations has motivated several tracks of recent work to improve the effect size and replicability of brain function-mental health associations. These have included application of more rigorous fMRI signal denoising methods[26], development of more sophisticated analytic[27,28] and modelling validation techniques[24,29], the use of larger sample-sizes, and deeply sampled data[30]. These advances have yielded unmistakeable, though still largely marginal, gains in brain-behaviour effect sizes and replicability within individual-differences research. An additional avenue of work has sought to identify whether brain activity evoked under certain conditions might better convey mental health information than others. Indeed, recent work has identified that brain activity recorded during active task and movie-watching conditions has stronger links to individual differences in traits and behaviour than that recorded during resting-state conditions[10,31,32].

As such, task-based and movie-watching fMRI data collected in large-scale neuroimaging cohorts – best powered for individual differences analyses – have held promise for identifying robust brain function-mental health links. While there are dozens of different fMRI tasks, most large-scale neuroimaging studies draw from a similar set of block-design tasks. These include:

the emotional face-matching task (UK Biobank, Human Connectome Project (HCP)-Young Adult, HCP-Development, HCP-Anxious Misery, HCP-Disordered Emotional States, Transdiagnostic Connectome Project); incentive salience/gambling (HCP-Young Adult, HCP-Anxious Misery, HCP-Disordered Emotional States); n-back working memory task (Adolescent Brain Cognitive Development (ABCD), HCP-Young Adult); and Frithe-Happe animations social cognition task (HCP-YA). Several datasets also use naturalistic movie-watching (HCP-YA, Healthy Brain Network, Naturalistic Neuroimaging Database). Yet, even in large-scale cohorts, across resting-state[4,7,23,31,33-36], task-based[4,7,23,37], and movie-watching[31] fMRI paradigms, links between brain activity and non-cognitive mental health domains (e.g., positive and negative valence system functions) remain comparatively weak relative to links with cognitive measures (i.e., CS). Even brain activity during tasks like the emotional face matching task – widely thought to index interindividual variability in the NVS – has recently been found to instead predominantly reflect the CS[37-39]. As such, it remains unclear which – if any – of these common fMRI paradigms evoke brain activity in a way that meaningfully reflects non-cognitive domains of mental health.

An oft-posited[4,31] contributor to this discrepancy is a difference in reliability between behavioural measures of the CS – which tend to be performance-based and relatively stable over time – versus behavioural measures of the PVS and NVS – which tend to be based on self-report or subjective evaluation and can be more variable over time. As a result, the former tend to have higher measurement reliability than the latter[40], and recent work has found that the measurement reliability of target behavioral phenotypes is an important moderator of brain-behavior relationship strengths[41]. One way to increase reliability is to aggregate data across many individual assessments, via techniques such as hierarchical factor analysis (hFA)[42], which can produce composite scores with high reliability even from individual assessment scores with lower reliability[43,44].

A final consideration is the scale of brain activity to be examined. Prior work has demonstrated that properties of brain activity at the level of large-scale distributed networks yield higher subject discriminability[12] and stronger reproducibility[45] than properties of brain activity at the level of individual connections. Moreover, functions spanning CS, PVS, NVS, and SP domains are now understood to be subserved more so by large-scale distributed networks than by individual brain regions or connections[46-49]. In the context of task and movie-watching fMRI, where stimuli presentation is standardized across subjects in time, tensor independent component analysis (tICA) offers a robust, data-driven method to identify and quantify the simultaneous activity of all large-scale brain networks activated by a given task[38,50].

In the present study, to gain insights into the above considerations, we dissected brain activity across both tasks and large-scale brain networks to allow for nuanced probing of where (i.e., in the brain) and when (i.e., during what task) individual differences in brain activity optimally reflect individual differences in key RDoC mental health-related neurobehavioral functional domains. In split-half samples of a large-scale normative cohort (HCP-YA; n=1200), we first applied Schottner and colleague's hFA[42] model to 87 behavioral and neuropsychiatric assessments to generate high reliability mental health factor scores in five latent domains: cognition (CS), positive affect/well-being (PVS), negative affect/internalizing (NVS), processing speed, and substance use. Next, we applied tICA[45] to fMRI data collected across paradigms with

putative relevance to these RDoC domains (i.e., emotional face-matching, incentive salience/gambling, social cognition task, n-back working memory, and movie-watching of socioemotional content) to identify and quantify evoked brain network activity. The subject-level mental health and brain function profiles generated from this integrated multivariate approach were then cross-analyzed and compared in several ways to identify mental health-brain function mappings. In their absence, we sought to determine the contributing factors to and consequences of their non-convergence.

Results

Task-Evoked Brain Networks

tICA identified 77 evoked large-scale network variants across the five fMRI paradigms in the discovery sample (Group 1) (**Fig. 1**). Individual network maps have been made viewable and accessible in a "Task-Evoked Network Atlas" (https://identifiers.org/neurovault.collection:20820). The tICA component spatial maps capture temporal dynamics of brain activity and therefore resemble connectivity states observable using dynamic connectivity analyses. As such, each map often reflects more than one classical static connectivity network (i.e., network averaged over several minutes of resting state). Hierarchical clustering based on the network spatial maps resolved them into a set of nine Network Families (**Figs. 1-2**). Findings are consistent with the notion of a circumscribed set of canonical large-scale networks that manifest in context-dependent spatial variants[51] and undergo network-network coupling reconfigurations[52] to meet particular task demands. tICA components also comprise the temporal course of each spatial pattern and a set of subject loadings for each that reflect the strength of the spatio-temporal pattern of activation in each individual. These loadings can be used to relate network activation to non-imaging measures.

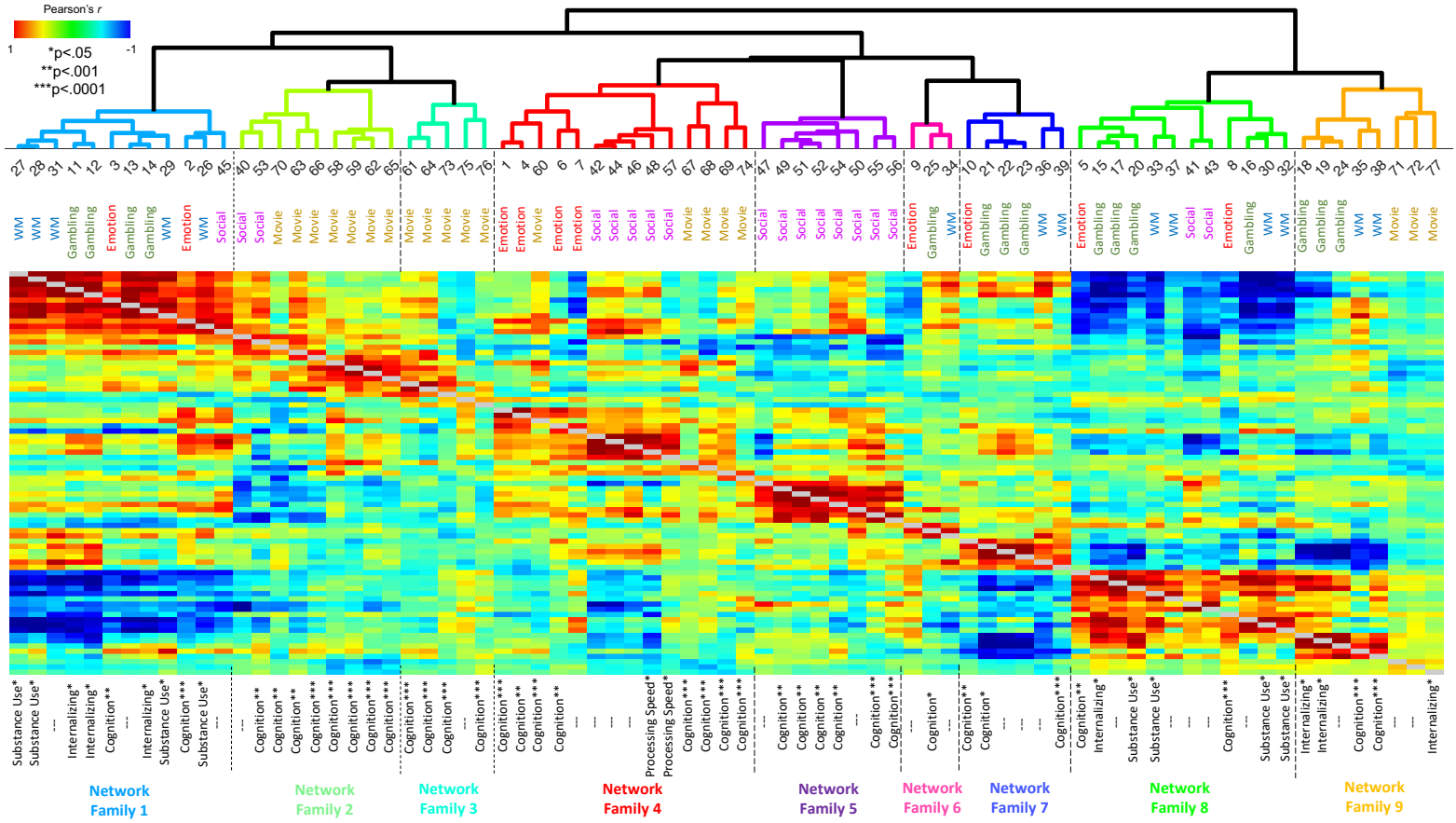

Figure 1. *Hierarchical clustering of task-evoked network spatial maps.* The heat map depicts the correlation matrix of unthresholded network spatial maps, with red indicating positively correlated spatial maps and blue indicating negatively correlated spatial maps. Task labels above each column in the heat map show the task each network was evoked by (i.e., Emotion, Gambling, Working Memory, Social Cognition, or Movie-Watching) and Factor labels below the heat map show whether variability across subjects in a network's activation during the specific task correlated with variability in one of the five mental health domains: Cognition, Internalizing/negative affect, Well-being/ positive affect, Processing Speed, and Substance Use (*p<.05 uncorrected; **p<.001 uncorrected; ***p<.0001 Bonferroni corrected threshold for 385 tests: 77 networks * 5 mental health factors.

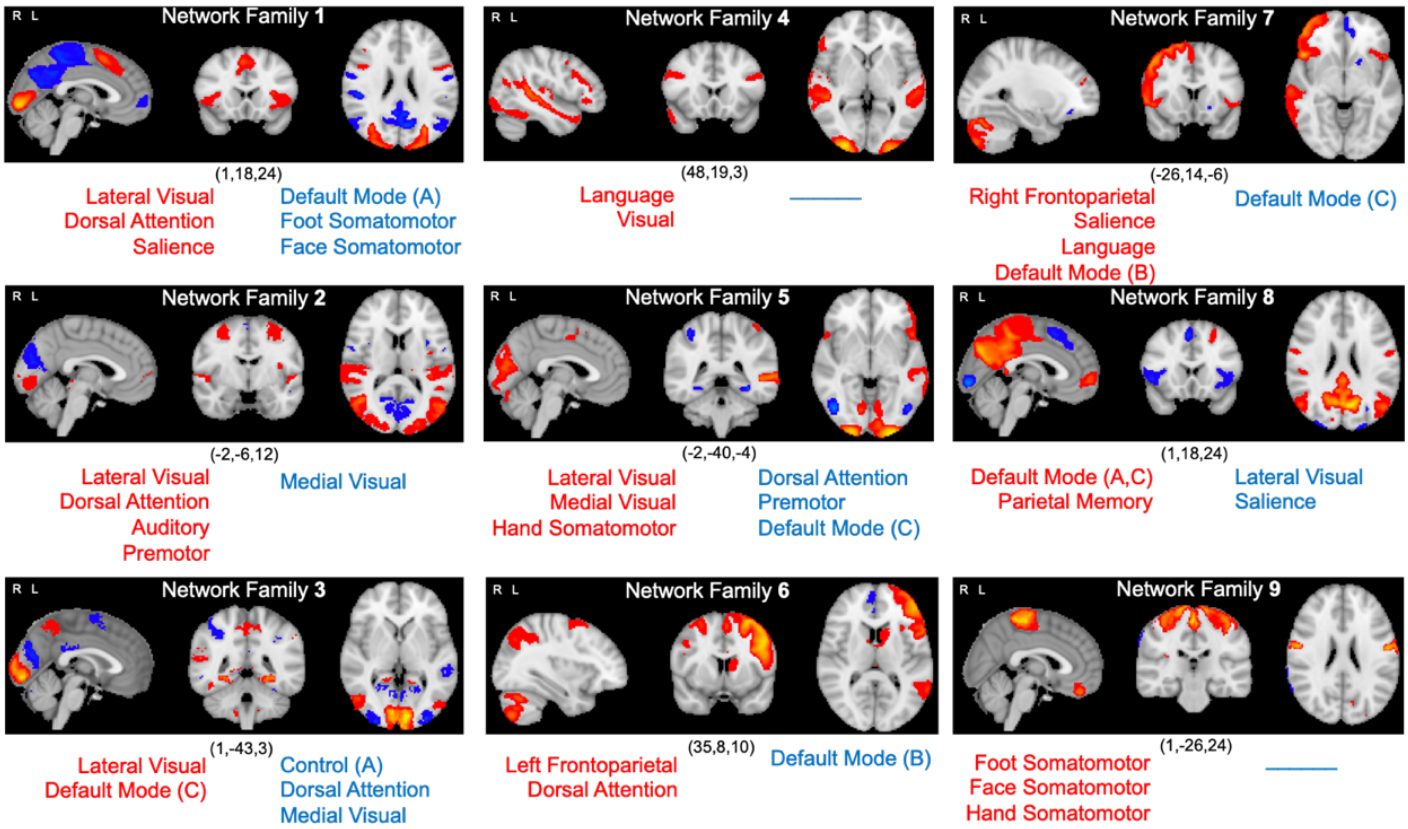

Figure 2. *Task-evoked network family spatial maps.* For each network family, the positive (red) and negative (blue) components of each network represent the average of tICA components in that family (each thresholded at p=0.5 using a Gaussian-Gamma mixture model) to compute a mean network family spatial map. Overlap between the positive and negative components of each network family map and each of 17 canonical large-scale networks defined in the Gordon et al. parcellation[53] was computed using the Network Correspondence Toolbox (NCT)[54]. Networks names listed in red showed significant overlap with positive task network components and networks listed in blue showed significant overlap with negative task network components

Network Activity Predominantly Reflects Cognition Across All Tasks

First, we examined the univariate links between the activity of the 77 evoked large-scale brain network variants (indexed by tICA subject loadings[50]) and factor scores on each of the five

mental health domains. The same analyses were repeated for the 66 networks identified in the replication group (Group 2) to assess reproducibility. Analyses controlled for age and sex and took into account covariance in family structure within the participant sample (i.e., family members with non-independent data) using permutation analysis of linear models (PALM)[55] with multi-level exchangeability blocks[56]. We applied a conservative Bonferroni significance threshold ($p<0.0001$) to correct for multiple comparisons (77 networks * 5 mental health domains = 385 tests). Across both Group 1 (discovery sample) and 2 (replication sample), across all tasks, and across all brain Network Families, only associations between network activity and cognition were statistically significant after correcting for multiple comparisons or replicated across groups even at uncorrected thresholds (**Fig. 3**).

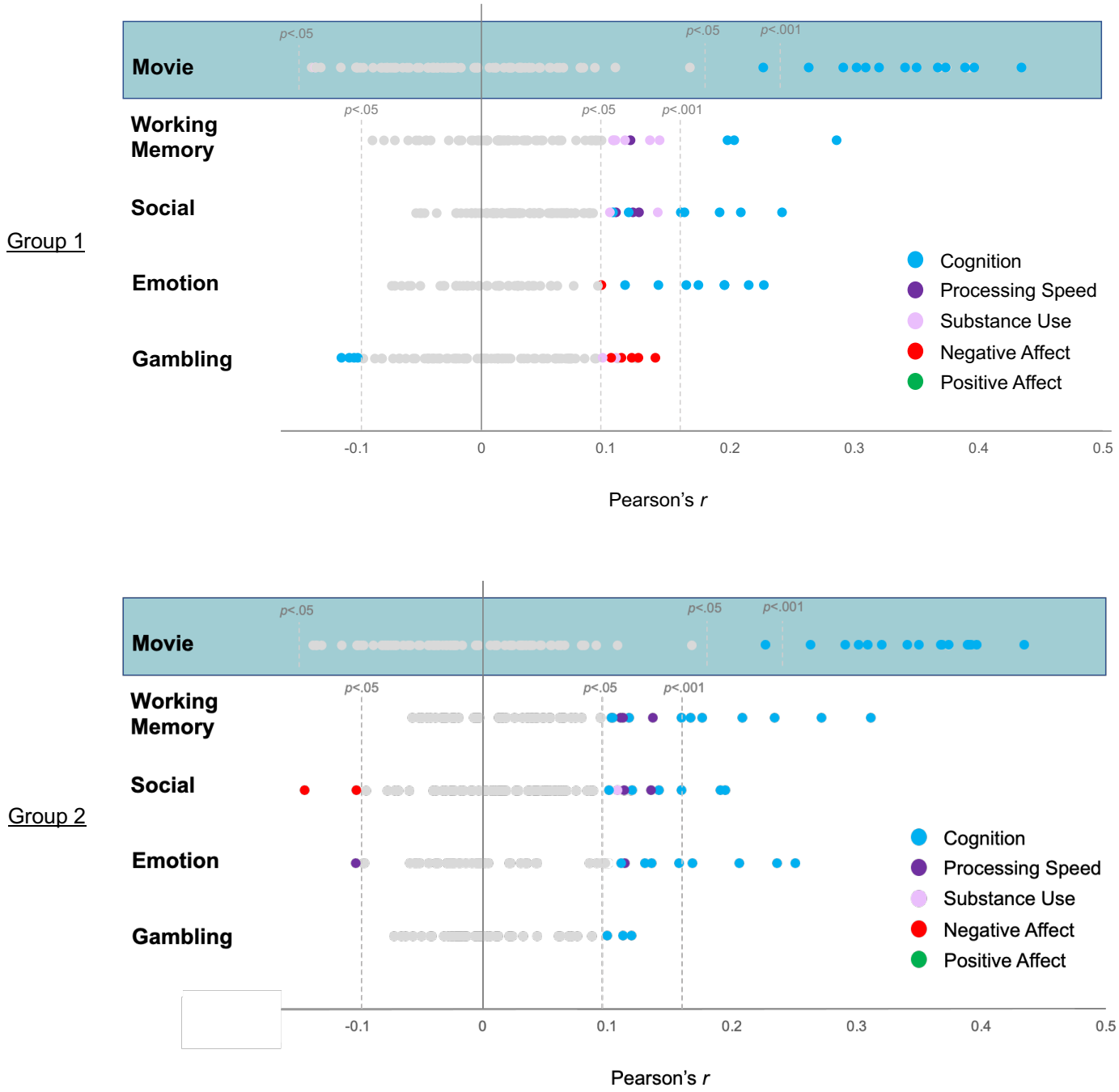

Figure 3. *Univariate links between task-evoked network activation strengths and mental health factor scores.* For each of the five tasks (rows), the Pearson's correlation between the activation strength of each network evoked by the task and each of the five mental health factor scores is shown (dots). Significance was assessed both at the uncorrected p<.05 and p<.001 levels, and Bonferroni-adjusted p<.0001 level to correct for multiple comparisons. Analyses controlled for age and sex and took into account covariance in family structure using PALM. All relationships significant at the p<.05 level or less are color-coded by mental health domain. All non-significant relationships are grayed out. Findings are presented separately for Group 1 and Group 2. Movie-watching data was obtained in only a subset of HCP-YA subject (n=184) and therefore was not split across groups. Data from this subset is visualized on both the Group 1 and Group 2 figures for benchmarking purposes.

Even when considering findings significant at the uncorrected p<.05 or p<.001 levels, activity in five of the nine Network Families (2, 3, 5, 6 and 7) exclusively reflected cognition, regardless of the task (**Fig. 1**). In Group 1, Network Families 1, 8 and 9 were the only ones whose activity reflected negative affect (NVS) or substance use at the uncorrected level. However, none of these non-cognitive relationships replicated in Group 2. Moreover, links with positive affect (PVS) were not observed even at the uncorrected level in either Group 1 or Group 2. Only relationships with cognition replicated across groups. In both Groups 1 and 2, the Movie-Watching task elicited brain activity with the strongest link to individual differences in cognition by a wide margin, followed by the Working Memory task (**Fig. 3**). Conversely, in both groups, brain activity during the Gambling task had the weakest links with individual differences in mental health. While Gambling, Working Memory and Social Cognition task activation linked to substance use in group 1, only brain network activity during the Social Cognition task displayed a somewhat consistent link to individual differences in substance in both groups, albeit at a modest level (p<.05) and in different brain networks.

Measurement Reliability Differences Do Not Account for Predominance of Cognition

Recent work has found that an important moderator of brain-behavior relationship strengths is the measurement reliability of target behavioral phenotypes[41]. Therefore, we examined the possibility that weak brain function links with non-cognitive mental health domains across tasks were an artifact of comparatively lower measurement reliability in these measures relative to the cognition domain. To assess this, we leveraged the HCP-YA's Test-Retest sample (n=46) to compute the intraclass correlation (ICC) of each of the five mental health factors. ICC is the predominant index of measurement reliability in the clinical neuroimaging literature[41,57,58], and quantifies the percent of total variance (i.e., between-subject, within-subject, and error variances) attributable to between-subject variance[59]. Retest subjects conducted their second set of HCP-YA assessments an average of 139.3 days (SD=69.0 days) after conducting their first set of assessments. The retest protocol included 78 of 87 behavioral and neuropsychiatric assessments included in the initial testing protocol, with all nine of the missing variables belonging to the substance abuse factor. While the factor analysis procedure we employed[42] imputes missing values, we elected not to interpret ICC values for the substance abuse factor since only two of the eleven initial substance use measurements were available in the retest data.

ICC values for the remaining four mental health factors were all very high; ($ICC_{COGNITION}$ = 0.991; $ICC_{NEGATIVE\ AFFECT}$ = 0.937; $ICC_{POSITIVE\ AFFECT}$ = 0.988; $ICC_{PROCESSING\ SPEED}$ = 0.999). These high ICC values are consistent with prior work showing that composite measures tend to have much higher reliability than the individual measures that make them up[43,44]. Most importantly, they indicate that the large discrepancy in strength observed here between brain

function relationships with Cognition versus with the other mental health factors cannot be explained by differences in measurement reliability.

Interindividual Variability Differences Do Not Account for Predominance of Cognition

Given that the HCP-YA sample comprised generally healthy young adults, we also examined the possibility that narrower interindividual variability in the non-cognitive mental health domains compared to the cognitive domain accounted for the imbalance in brain function associations. However, interindividual variation, as measured by the average coefficient of variation (CV) of the assessment items in each domain, was actually lowest for cognition ($CV_{COGNITION}$=.16; $CV_{PROCESSING\ SPEED}$=0.18; $CV_{POSITIVE\ AFFECT}$=0.18; $CV_{NEGATIVE\ AFFECT}$=0.47; $CV_{SUBSTANCE\ USE}$=2.21). This shows that variability in brain function data was sensitive to variability in Cognition despite its low variability across subjects compared to the other mental health domains. Conversely, these data underscore the low sensitivity of the brain function measures to the non-cognitive mental health domains despite the larger variability of these domains across subjects. In all, differences in interindividual variability do not likely account for the brain function measures' predominant reflection of cognition.

Multivariate Brain Function Profiles Predict Cognition and Processing Speed but not Positive or Negative Affect

Next, we examined whether considering subjects' multivariate activity across all evoked networks could predict mental health scores. In order to facilitate training and testing across consistent networks in Groups 1 and 2, we restricted this analysis to the 46 Emotion, Gambling, Working Memory, and Social Cognition task networks that were identified across both groups (**Supplementary Table S1**), and to the subjects (Group 1: n=324; Group 2: n=330) who had complete data across all four tasks. Movie-watching networks were not included in this analysis given the smaller subset of subjects with this data (n=184 total). We trained a random forest machine learning regression model on Group 1 to make individualized predictions in Group 2 of each mental health factor score based on subjects' brain function profiles. We subsequently applied SHAP (SHapely Additive exPlanations)[60] to the predictions made by each model to quantify the feature importance of each brain network for predicting scores in each mental health domain.

In the test set (Group 2), a subject's profile of brain activity across the 46 task-evoked networks significantly predicted their Cognition score (correlation between actual and predicted score: r=.15, p=.005, MSE=0.919) and Processing Speed score (r=.11, p=.041, MSE=1.184) to modest degrees. SHAP analysis (**Fig. 4a**) revealed that the network with the highest feature importance for predicting Cognition was a variant from Network Family 9, evoked during the n-back Working Memory task, which involved left-lateralized motor cortex, rostral ventral putamen, and orbitofrontal cortex, as well as bilateral hippocampus and ventromedial prefrontal and ventral visual cortices, negatively correlated with the right frontoparietal control network, the anterior and posterior dorsal midline and cerebellum (**Fig. 4b**). The network with the highest feature importance for predicting processing speed was a variant from Network Family 6 (**Fig. 4a**), which involved bilateral frontoparietal and dorsal attention networks, negatively correlated with

the default mode network (**Fig. 4b**). Multivariate brain network activity profiles did not significantly predict Positive Affect (r=.003, *p*=.956, MSE=1.255), Negative Affect (r=.03, *p*=.630, MSE=1.059) or Substance Use (r=.02, *p*=.681, MSE=1.226).

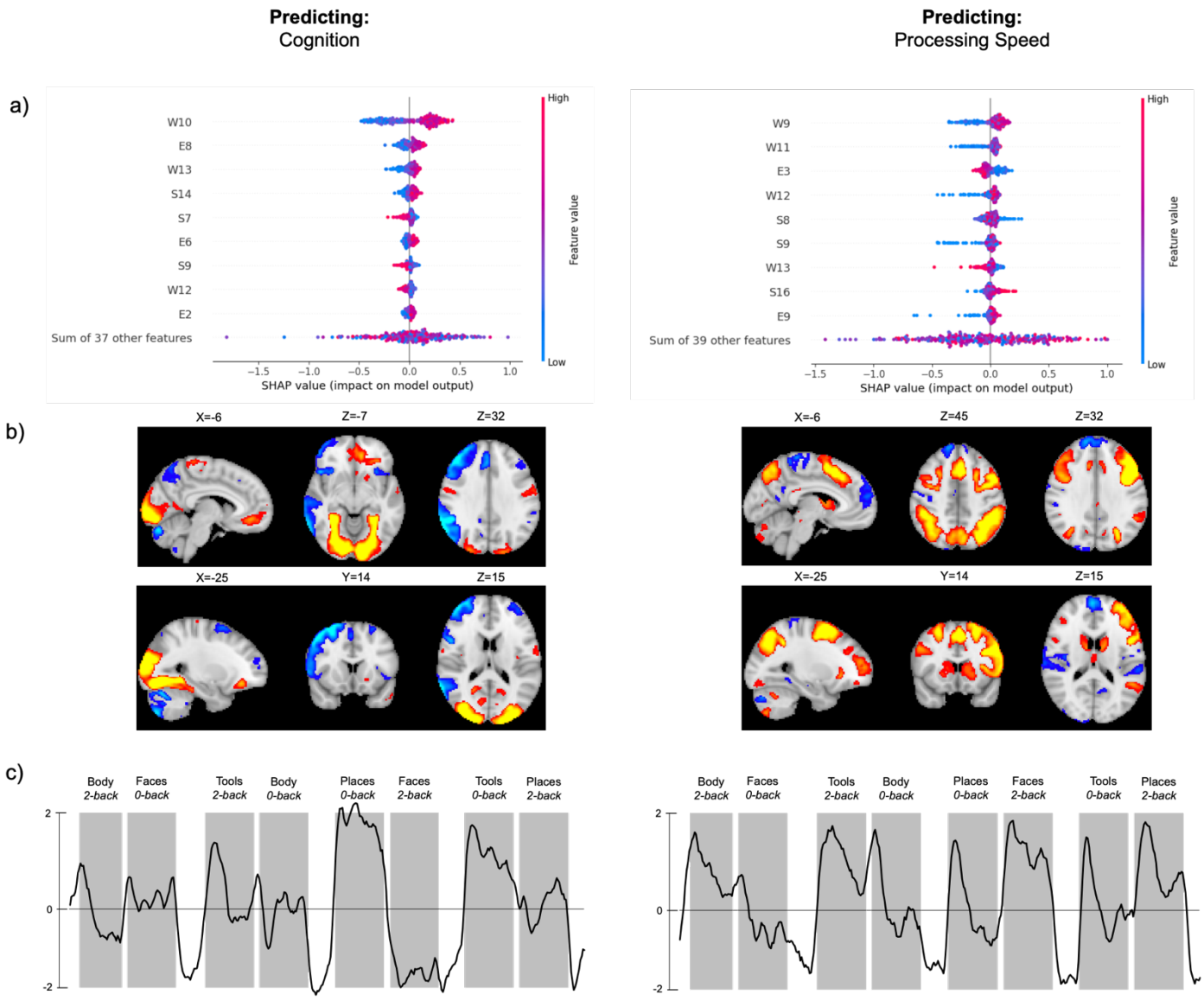

Figure 4. *Task-evoked brain networks predicting Cognition and Processing Speed.* a) Feature importance of task-evoked networks for predicting Cognition (left panel) and Processing Speed (right panel). Networks are ordered from top-to-bottom on average feature importance. Each dot represents a subject's SHAP value for a given network. Dots with positive SHAP values denote data points that increase the prediction for a given network, while dots with negative SHAP values indicate data points that decrease the prediction for that network. Red/blue indicates whether the feature is positive or negative for that data point. b) Spatial map and c) time course of the network with the highest feature importance for predicting Cognition (left panel) and Processing Speed (right panel). Network time courses in c) are overlayed on the Working Memory task stimulation blocks for reference; white spaces indicate inter-block (i.e., rest) intervals. The time course in the left panel of c) (Cognition) suggests strong activation of frontoparietal areas and suppression of medial visual areas during faces 2-back (e.g., activity during the block is negative) vs low activation during most 0- and 2-back conditions, except for places and tools 0-back, which have positive activity of the network (e.g., strong activation of visual areas). The time course in right panel of c)

(Processing Speed) suggests strong activity of the network throughout 2-back conditions and at the start of most of the 0-back conditions.

Brain Function Subtypes and Mental Health Subtypes Have Poor Correspondence

Next, to evaluate the extent to which the multivariate brain function profiles and mental health profiles reflect one another, we conducted a parallel subtyping analysis. To do so, we applied hierarchical clustering, separately, to subjects' brain function profiles and mental health profiles (**Fig. S1**). This clustered each individual into one subtype based on their brain function profile and one subtype based on their mental health profile. Then, we assessed the correspondence between these two sets of subtypes at all levels where the number of clusters was common to both the brain function and mental health subtyping sets (i.e., k=2, 5 and 10; **Fig. S1**). The mental health characteristics of each subtyping solution are presented in **Supplementary Figure S2**.

Brain function subtypes and mental health subtypes were largely non-corresponding, wherein a subject's brain function subtype was not predictive of their mental health subtype and vice-versa (**Fig. 5**). That is, subjects in a given brain function subtype did not predominantly map to a single mental health subtype; rather, their mental health subtypes were randomly distributed. While subtype correspondence was statistically significant at the k=2 clustering level, $X^2(1)=4.49$, $p=.034$, Figure 5 illustrates the low effect size and specificity of the overlap. For instance, of 218 subjects in brain function subtype 1, 42% were in mental health subtype 1 and 58% were in mental health subtype 2. Correspondence did not differ from chance at finer clustering levels (k=5: $X^2(16)=23.81$, $p=.094$; k=10: $X^2(16)=76.91$, $p=.608$).

Mental Health Subtype

Brain Network Subtype	***k=2***	1	2
	1	n=92 *(100.9)*	n=126 *(117.1)*
	2	n=58 *(49.1)*	n=48 *(56.9)*

$X^2(1)=4.49$
$p=.034$

Mental Health Subtype

Brain Network Subtype	***k=5***	1	2	3	4	5
	1	n=25 *(23.5)*	n=6 *(9.4)*	n=11 *(12.9)*	n=12 *(14.0)*	n=17 *(11.2)*
	2	n=24 *(28.4)*	n=11 *(11.4)*	n=17 *(15.7)*	n=21 *(17.0)*	n=13 *(13.5)*
	3	n=20 *(20.2)*	n=6 *(8.1)*	n=13 *(11.1)*	n=8 *(12.1)*	n=14 *(9.6)*
	4	n=16 *(18.8)*	n=13 *(7.6)*	n=11 *(10.4)*	n=13 *(11.3)*	n=4 *(9.0)*
	5	n=22 *(16.2)*	n=7 *(6.5)*	n=7 *(8.9)*	n=10 *(9.7)*	n=3 *(7.7)*

$X^2(16)=23.81$
$p=.094$

Mental Health Subtype

Brain Network Subtype	***k=10***	1	2	3	4	5	6	7	8	9	10
	1	n=3 *(2.4)*	n=3 *(3.5)*	n=3 *(4.4)*	n=5 *(4.1)*	n=2 *(3.1)*	n=4 *(2.5)*	n=4 *(3.4)*	n=0 *(1.2)*	n=1 *(1.6)*	n=6 *(4.9)*
	2	n=3 *(3.1)*	n=3 *(4.5)*	n=10 *(5.7)*	n=1 *(5.3)*	n=4 *(3.9)*	n=1 *(3.2)*	n=4 *(4.3)*	n=2 *(1.5)*	n=1 *(2.1)*	n=11 *(6.3)*
	3	n=3 *(3.6)*	n=4 *(5.2)*	n=6 *(6.7)*	n=4 *(6.3)*	n=4 *(4.7)*	n=5 *(3.8)*	n=6 *(5.1)*	n=1 *(1.7)*	n=5 *(2.5)*	n=9 *(7.4)*
	4	n=1 *(3.0)*	n=3 *(4.3)*	n=7 *(5.6)*	n=7 *(5.2)*	n=5 *(3.9)*	n=3 *(3.1)*	n=5 *(4.2)*	n=3 *(1.4)*	n=1 *(2.1)*	n=4 *(6.2)*
	5	n=2 *(1.8)*	n=1 *(2.6)*	n=4 (3.3)	n=1 *(3.1)*	n=3 *(2.3)*	n=1 *(1.9)*	n=3 *(2.5)*	n=0 *(0.9)*	n=1 *(1.2)*	n=7 *(3.6)*
	6	n=3 *(2.9)*	n=7 *(4.2)*	n=3 (5.4)	n=5 *(5.1)*	n=6 *(3.8)*	n=3 *(3.1)*	n=2 *(4.1)*	n=1 *(1.4)*	n=1 *(2.0)*	n=7 *(6.0)*
	7	n=4 *(2.3)*	n=3 *(3.3)*	n=4 *(4.3)*	n=6 *(4.0)*	n=1 *(3.0)*	n=2 *(2.4)*	n=2 *(3.3)*	n=3 *(1.1)*	n=3 *(1.6)*	n=2 *(4.7)*
	8	n=1 *(2.0)*	n=2 *(2.9)*	n=2 *(3.7)*	n=7 *(3.5)*	n=3 *(2.6)*	n=4 *(2.1)*	n=1 *(2.8)*	n=2 *(1.0)*	n=2 *(1.4)*	n=2 *(4.1)*
	9	n=3 *(2.4)*	n=5 *(3.5)*	n=5 *(4.4)*	n=6 *(4.1)*	n=2 *(3.1)*	n=2 *(2.5)*	n=5 *(3.4)*	n=0 *(1.2)*	n=1 *(1.6)*	n=2 *(4.9)*
	10	n=2 *(1.4)*	n=5 *(2.0)*	n=2 *(2.6)*	n=1 *(2.4)*	n=2 *(1.8)*	n=1 *(1.4)*	n=3 *(2.0)*	n=0 *(0.7)*	n=1 *(0.9)*	n=1 *(2.8)*

$X^2(81)=76.91$
$p=.608$

Figure 5. *Non-alignment of brain function subtypes and mental health subtypes.* At each clustering solution level common to both the brain function and mental health subtyping sets (k=2, 5 and 10), the total count (n) of subjects with each joint brain function subtype–mental health subtype assignment. Grayed values indicate the expected count if joint subtype assignments were random. Chi-squared tests examining the difference between the expected (by random chance) and actual joint assignment counts indicated that subjects in a brain function subtype were randomly distributed across mental health subtypes, and vice versa, underscoring the non-alignment of the subtypes.

Subtyping Misalignment Yields Groups with Discordant Brain Function-Mental Health Associations

We next examined the consequence of this subtyping misalignment for group-level analyses of brain function-mental health associations. We found that the starting point of analysis (i.e., whether to subtype subjects by mental health data or by brain activation data) meaningfully altered conclusions about brain function-mental health group-level associations, in sometimes opposite directions (**Fig. 6**).

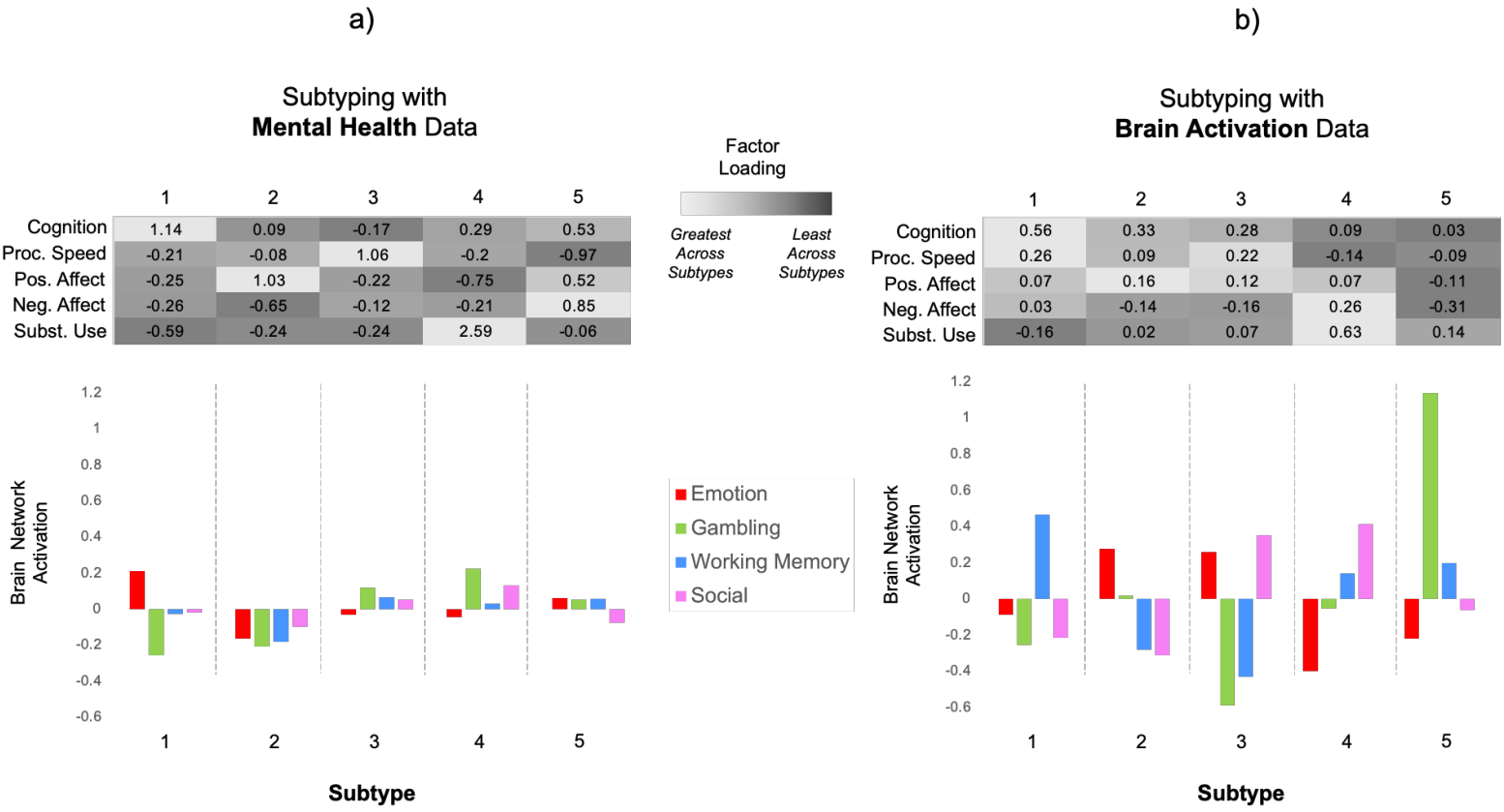

Figure 6. *Comparing brain function-mental health associations across subtyping solutions.* Average mental health profile (top panel; *Z*-scores) and average brain function profile (bottom panel; *Z*-scores) of each subtype derived from clustering of a) mental health data and b) brain function data at the k=5 clustering solution level. Given that within-subject loadings on networks activated by the same task were strongly correlated such that participants with high loadings on one pattern generally showed high loadings on other patterns (**Fig. S4**), for visualization purposes we operationalized brain network activation on each task as the *z*-scored mean subject loading across all networks activated by the task.

For instance, in the k=5 mental health subtyping set (**Fig. 6a**), the high cognition, low substance use group (subtype 1) was characterized at the brain level by high network activation to the Emotion task and low network activation to the Gambling task, with average activation on the Working Memory and Social Cognition tasks (**Fig. 6a**). But in the k=5 brain function subtyping set (**Fig. 6b**), the analogous high cognition, low substance use group (subtype 1) was characterized at the brain level by very high network activation to the Working Memory task and average activation on the Emotion task. Similarly, in the mental health subtyping set, the high substance use group (subtype 4) was characterized at the brain level by high network activation on the Gambling task and average network activation on the Emotion task, whereas in the brain function subtype set, the high substance use group (subtype 4) was characterized at the brain level by average network activation on the Gambling task and very low network activation on the Emotion task. Overall, due to the discordance of the mental health and brain function data, the starting point of analysis (i.e., whether to subtype subjects by mental health or brain function) meaningfully alters group-level associations between brain function and mental health.

Brain Function Subtypes Delineate Mental Health Differences Better than Mental Health Subtypes Delineate Brain Function Differences

Finally, given the observed discordance between subtyping sets, we compared the relative ability of subtypes generated by one modality to yield groups that significantly differed in facets of the other modality. Independent samples t-tests at the k=2 clustering level (controlling for age and sex, and accounting for covariance in family structure using PALM) demonstrated that subtypes generated using brain function data had significantly different means of cognition ($t(322)=3.18$, $p<.001$), processing speed ($t(322)=2.27$, $p=.014$), and substance use ($t(322)=2.77$, $p=.002$), but not of internalizing ($t(322)=0.71$, $p=.790$) or positive affect ($t(322)=1.06$, $p=.220$). On the other hand, subtypes generated using mental health data only significantly differed in brain activation on the Gambling task ($t(322)=2.84$, $p=.002$), and not in brain activation on the Emotion ($t(322)=0.72$, $p=.258$), Working Memory ($t(322)=1.13$, $p=.110$) or Social Cognition ($t(322)=1.56$, $p=.064$) tasks. Similarly, one-way ANOVAs at the k=5 clustering level demonstrated that subtypes generated using brain function data had significantly different means of cognition ($F(6,316)=3.84$, $p=.004$), internalizing ($F(6,316)=2.67$, $p=.030$), and substance use ($F(6,316)=3.20$, $p=.014$), but not of positive affect ($F(6,316)=0.53$, $p=.733$) or processing speed ($F(6,316)=1.49$, $p=.240$). On the other hand, subtypes generated using mental health data only significantly differed in brain activation on the Gambling task ($F(6,316)=3.03$, $p=.016$), and not in brain activation during the Emotion ($F(6,316)=2.02$, $p=.110$), Working Memory ($F(6,316)=1.16$, $p=.328$) or Social Cognition ($F(6,316)=0.77$, $p=.565$) tasks. Overall, these findings indicate brain function data were more sensitive to mental health differences than mental health data were sensitive to brain function differences.

Comparing Interindividual Variability and Assessment Diversity in Brain Function and Mental Health Data

To investigate what may account for the above findings, we compared the brain function and mental health data in terms of interindividual variability and assessment diversity. First, within brain function data, significantly greater ($p<.05$) interindividual variation in brain activity was elicited by the Emotion task (mean coefficient of variation (CV) of 10 evoked networks=0.55)

than by the Social Cognition (mean CV of 18 evoked networks=0.28), Working Memory (mean CV of 14 evoked networks=0.31), Gambling (mean CV of 15 evoked networks=0.33) or Movie-Watching (mean CV of 20 evoked networks=0.36) tasks (**Fig. S3**). Within mental health data, interindividual variation was significantly greater ($p<.05$) for measures of substance use (mean CV of 12 assessments=2.21) and negative affect (mean CV of 21 assessments=0.47) than for measures of cognition (mean CV of 27 assessments=0.16), processing speed (mean CV of 16 assessments=0.18), and well-being/positive affect (mean CV of 9 assessments=0.18). Overall, the interindividual variability of individual assessments did not differ significantly between brain function and mental health data, $t(151)=1.34$, $p=.18$.

However, the 46 reproducible brain function assessments (mean absolute assessment-assessment correlation, $r=.278$) were significantly more redundant with one another than the 87 mental health assessments (mean absolute assessment-assessment correlation, $r=.106$), $t(4603)=42.73$, $p<.001$ (**Fig. S4**). This reflects that within an individual, the activity of any given brain network on any given task moderately predicted the activity of the individual's other brain networks in the same and other tasks. Moreover, individuals exhibited almost twice as much similarity to one another in brain function (mean absolute subject-subject correlation amongst 46-network activity profiles: $r=.224$) as they did in mental health (mean absolute subject-subject correlation amongst 87-assessment profiles: $r=.133$), $t(101779)=109.64$, $p<.001$, indicating poorer overall subject discriminability in brain function data relative to mental health data.

Discussion

Hedge and colleagues describe a "reliability paradox"[61] built into many fMRI tasks, in which tasks designed to elicit robust group-level brain activations to map task-relevant brain regions must inherently have low between-subject variability in brain activity – causing low reliability for individual differences research. Yet, we demonstrate here that the magnitude of between-subject variability in task fMRI-evoked brain activity in generally healthy adults can be sufficient – with large enough sample sizes and optimized methodology – to robustly parse individual differences even in low-variability target phenotypes (i.e., Cognition). What, then, could explain why the same fMRI data does not parse individual differences in other phenotypes (i.e., PVS and NVS) with comparable measurement reliability and even larger interindividual variability? We posit that it is because the set of fMRI tasks commonly used in large-scale neuroimaging studies simply do not elicit individual differences in brain activity that are relevant to variability in non-cognitive aspects of mental health – likely because these tasks are not emotionally evocative enough to map to the range of affective function in healthy adults. If this is the case, continuing advances in neuronal signal purification, modelling optimization and spatiotemporal resolution may not meaningfully strengthen our ability to extract this information. Rather, the current findings point to a need to leverage different types of tasks or paradigm designs in future large-scale neuroimaging studies if our goal is to evoke brain activity that more holistically represents mental health.

Cardiac stress tests provide an apt analogy. Holistic assessments of cardiovascular health require not only measures of heart function at the resting or low-level activity state, but at the vigorous activity state as well. We posit that the fMRI tasks common to large-scale studies do not take the brain particularly far away from its resting-state, and that more demanding tasks that facilitate

measurement of brain function during vigorous activity states will need to be more broadly implemented and incorporated into large-scale studies. fMRI tasks implemented in large-scale neuroimaging studies tend to be selected for brevity, ease of administration, and established consistency of results. This skews paradigm selection towards block designs that provide maximized signal-to-noise ratios and require less individualized analytic processing pipelines, but which often mask important nuances in brain reactivity to specific stimuli (e.g., correct versus incorrect trials) and average out significant individual differences in brain activation. While more challenging to administer consistently in a large-scale study, tasks that model subject responses and brain activity at higher detail (i.e., event-related designs) and that are more affectively provocative and challenging (e.g., social exclusion task, cold pressor task, monetary incentive delay task) may prove more powerful in extracting individual differences in non-cognitive mental health. Importantly, though, paradigms involving highly salient real-life consequences must be carefully considered in the context of research ethics and subject beneficence.

Many basic science investigations, clinical trials, grant proposals and large-scale study designs rely on explicit or implicit notions that particular fMRI tasks differentially probe individual differences in distinct RDoC domains. For instance, the Emotion task and related variants are commonly used as probes of NVS; the Gambling task and related variants are commonly used as probes of PVS; and the Working Memory task and related variants are commonly used as probes of CS[62]. These tasks are indeed validated "functional localizers"[63] of the neural substrates relevant to distinct domains. However, as the present findings underscore, just because a task activates neural substrates underlying a domain of function, does not necessarily mean that it elicits interindividual variability in the activity of these substrates that reflects variability in that domain of function. This likely reflects the accumulating evidence that brain regions are multifunctional and have roles in diverse processing streams depending on task needs. For instance, while the Emotion task robustly activated the amygdala – a structure with known roles in NVS – neither variability in networks involving the amygdala, nor amygdala connectivity or activity itself indexed interindividual variability in negative affect/internalizing[38]. Instead, it appears that individual variability in amygdala activity on this task primarily reflects variation in the amygdala's roles in cognition. In this way, despite these tasks' distinct names, stimuli and instructions, individual differences in brain activity evoked by the Emotion, Gambling, Working Memory, Social Cognition, and Movie-Watching paradigms all predominantly reflected variability in cognition. Therefore, these tasks may all be more appropriately categorized as paradigms in the CS domain of RDoC – at least when administered to generally healthy young adults. For this population, these tasks may be similarly unchallenging, unprovocative, and easily navigable with minimal recruitment of non-cognitive neural resources. In addition, some of the tasks may map to general cognition rather than a specific cognitive domain (such as working memory). Certainly, these tasks' ability to elicit brain activity reflecting cognition is valuable, given that cognitive deficits represent a core feature of numerous forms of mental illness. However, most disorders also involve a broader constellation of affective and other deficits. Eliciting and identifying brain activity that reflects these non-cognitive domains will remain a critical task of clinical neuroimaging moving forward.

It remains possible that variability in non-cognitive mental health domains may be better detected in brain activity variability modelled at other scales of brain organization than large-

scale networks. However, recent work has demonstrated that region-of-interest activation and edge-level connectivity were no better than large-scale network activation in the Emotion task in representing non-cognitive mental health domains[38]. Collectively, we posit that the data implicates as the key contributing factor a fundamentally poor ability of fMRI tasks currently included in most large-scale studies to elicit brain activity that reflects individual differences in non-cognitive mental health domains.

Our parallel subtyping findings point to another important set of issues. Many studies begin with mental health- or behaviour-based subject groupings (e.g., a clinical sample and a control sample) and seek to identify brain differences between the groups. Yet, we show here that groups constructed to have maximally dissimilar mental health profiles from one another nonetheless have largely similar brain function profiles during these tasks. This converges with findings from the clinical realm wherein patients classified to distinct DSM diagnostic categories have minimal neurobiological differentiation from one another[64]. Our findings support the movement toward a primacy of neurobiologically-based clinical taxonomy instead[65]. Here, subtype groups constructed to have maximally dissimilar brain function profiles from one another were modestly successful in separating out subjects with distinct mental health profiles. For instance, brain function subtypes were able to separate out subjects with distinct levels of cognition, internalizing and substance use. Meanwhile, mental health subtypes generally were not able to parse out differences in task-evoked brain activity. These findings indicate that beginning analyses with brain-based subtypes rather than mental health/behaviour-based subtypes may provide a better means of identifying fundamentally similar individuals and may be more successful at targeting biologically based treatment approaches to the individuals who will best respond to them.

A potential limitation is the present study's use of tICA to quantify brain function phenotypes. While tICA is a purely data-driven technique, its use involves some subjective decision points as described in the Methods, including the choice of model order and the classification of noise and signal components. The present study is also limited in examining generally healthy young adults, constraining the generalization of findings to clinical populations. Though, it should be noted that HCP-YA contains several hundred subjects with alcohol and cannabis use disorders, and subjects with subclinical depression and anxiety, and we found that HCP-YA subjects are more heterogeneous in negative affect/internalizing and positive affect/well-being than they are in cognition. An important aim of future work will be to explore whether the examined fMRI paradigms, and those similar to them, are more effective at eliciting brain activity reflecting non-cognitive mental health in clinical cohorts. Nonetheless, given that RDoC casts mental illness as a set of deviations from normative biobehavioral functioning, it is important to map variability in large-scale normative populations like HCP-YA. If the task batteries in these normative large-scale studies are ineffective at eliciting mental health-relevant brain activity, this creates a challenge for identifying the deviations that define mental illness.

Methods

Subjects

Data used in these analyses include fMRI scans from individuals collected as part of the Human Connectome Project (HCP) 1200 subject release. A detailed description of the recruitment for the HCP is provided by others[66,67]. Briefly, individuals were excluded by the HCP if they reported a history of major psychiatric disorder, neurological disorder, or medical disorder known to influence brain function. For reproducibility assessment, we randomly separated eligible subjects into two groups (n1=607; n2=599) with non-overlapping family structures. Given the oversampling of siblings and twins in the HCP-YA cohort, consolidating subjects from the same family within the same group prevented family structure-related data leakage across the two groups. To maximize the generalizability of replication analyses, we did not rebalance the family-separated groups by other demographic variables.

Subjects had variable data availability and quality across the different fMRI tasks examined in the present study. To maximize the quantity of analyzable data for detection of task-evoked brain networks and brain function-mental health univariate relationships, we first established task-specific groups by excluding only subjects from each task group who had missing data, HCP quality assessment flags, or excessive head motion (i.e., 1 or more frames with framewise displacement > 2mm) for that task. This produced the following sample sizes for each set of task groups: Emotion (n1=413; n2=416), Gambling (n1=444; n2=444), Social Cognition (n1=406; n2=435), Working Memory (n1=404; n2=395). Then, for multi-network predictive modeling across groups, we winnowed the task-specific groups into a set of task-general groups (n1=324; n2=330) that comprised subjects with available and high-quality data across all four tasks. Separately, we also examined the subset of subjects (n=184) who had 7T movie-watching data. Given the appreciably smaller sample size of subjects with movie-watching data, we did not split this sample or include movie-watching data in multi-network predictive modeling for reproducibility analysis.

fMRI Acquisition

HCP neuroimaging data were acquired with a standard 32-channel head coil on a Siemens 3T Skyra modified to achieve a maximum gradient strength of 100 mT/m[66]. T1-weighted high resolution structural images, acquired using a 3D MPRAGE sequence with 0.7mm isotropic resolution (FOV=224mm, matrix=320, 256 sagittal slices, TR=2400ms, TE=2.14ms, TI=1000ms, FA=8°), were used in the HCP minimal pre-processing pipelines to register functional MRI data to a standard brain space. Blood oxygenation level dependent (BOLD) functional magnetic resonance imaging (fMRI) data were acquired using a gradient-echo EPI sequence with the following parameters: TR = 720 ms, TE = 33.1 ms, flip angle = 52°, FOV = 280 × 180 mm, Matrix = 140 × 90, Echo spacing = 0.58 ms, BW = 2290 Hz/Px, 72 slices, 2.0 mm isotropic voxels, with a multiband acceleration factor of 8. HCP neuroimaging data were acquired in two sessions, each occurring on a different day.

HCP movie-watching fMRI data were acquired during a separate, third session on a 7T Siemens Magnetom scanner at the Center for Magnetic Resonance Research at the University of

Minnesota. Data were collected using a gradient-echo-planar imaging (EPI) sequence with the following parameters: repetition time (TR) = 1000 ms, echo time (TE) = 22.2 ms, flip angle = 45 deg, field of view (FOV) = 208 × 208 mm, matrix = 130 × 130, spatial resolution = 1.6 mm^3, number of slices = 85, multiband factor = 5, image acceleration factor (iPAT) = 2, partial Fourier sampling = 7/8, echo spacing = 0.64 ms, bandwidth = 1924 Hz/Px. The direction of phase encoding alternated between anterior-to-posterior (AP: MOVIE1, MOVIE4) and posterior-to-anterior (PA: MOVIE2, MOVIE3).

fMRI Tasks

Detailed descriptions of the HCP-YA fMRI task paradigms can be found here[31,68]. Brief summaries of each task paradigm are provided below.

The Emotion task (i.e., emotional face-matching task) was adapted from Hariri and colleagues[69]. Subjects were presented with blocks of trials that asked them to decide either which of two faces presented on the bottom of the screen matched the face at the top of the screen, or which of two shapes presented at the bottom of the screen matched the shape at the top of the screen. The faces had either angry or fearful expressions. Each run included blocks of 6 trials of the same task (face or shape), with the stimulus presented for 2 s and a 1 s ITI. Each block was preceded by a 3 s task cue ("shape" or "face"), so that each block was 21 s including the cue. Each of the two runs included 3 face blocks and 3 shape blocks.

The Gambling task (i.e., incentive processing task) was adapted from Delgado and Fiez[70]. Subjects guessed the number on a mystery card in order to win or lose money. Card numbers ranged from 1–9, and they were told to indicate if they thought the mystery card number was more or less than 5 by pressing one of two buttons on the response box. Feedback consisted of the number on the card (generated by the program as a function of whether the trial was a reward, loss or neutral trial) and either: 1) a green up arrow with "$1" for reward trials, 2) a red down arrow next to −$0.50 for loss trials; or 3) the number 5 and a gray double headed arrow for neutral trials. The task was presented in blocks of 8 trials that were either mostly reward (6 reward trials pseudo randomly interleaved with either 1 neutral and 1 loss trial, 2 neutral trials, or 2 loss trials) or mostly loss (6 loss trials interleaved with either 1 neutral and 1 reward trial, 2 neutral trials, or 2 reward trials). All participants were provided with money for completing the task, though it was a standard amount across subjects.

The Social Cognition task (i.e., theory of mind task) was adapted from Castelli and colleagues[71] and Martin and colleagues[72]. Subjects watched short video clips (20 s) of objects (squares, circles, triangles) either interacting in some way, or moving randomly. After each video clip, subjects chose between 3 possibilities: whether the objects had a social interaction (an interaction that appeared as if the shapes were considering each other's feelings and thoughts), Not Sure, or No interaction (i.e., there was no obvious interaction between the shapes and the movement appeared random). The task included 5 video blocks (2 Mental and 3 Random in one run, 3 Mental and 2 Random in the other run) and 5 fixation blocks (15 s each).

The Working Memory task (i.e., N-back task) involved blocks of trials that consisted of pictures of faces, places, tools and body parts. Within each run, the 4 different stimulus types were

presented in separate blocks within the run. Within each run, ½ of the blocks consisted of a 2-back working memory task (respond 'target' whenever the current stimulus is the same as the one two back) and ½ presented a 0-back working memory task (a target cue is presented at the start of each block, and the person must respond 'target' to any presentation of that stimulus during the block). A 2.5 s cue indicated the task type (and target for 0-back) at the start of the block. Each of the two runs contains 8 task blocks (10 trials of 2.5 s each, for 25 s) and 4 fixation blocks (15 s each). On each trial, the stimulus was presented for 2 s, followed by a 500 ms ITI. Each block contained 10 trials, of which 2 were targets, and 2–3 were non-target lures (e.g., repeated items in the wrong n-back position, either 1-back or 3-back). The inclusion of lures was critical to ensure that participants were using an active memory approach to the task and allowed one to assess conflict related activity as well as error related activity.

The Movie-Watching task involved passively viewing a series of video clips with audiovisual content. Each MOVIE run consisted of 4 or 5 clips, separated by 20 s of rest (indicated by the word "REST" in white text on a black background). The MOVIE1 run examined here contained clips from independent films (both fiction and documentary) made freely available under Creative Commons license on Vimeo, as well as a final clip comprising a montage of brief (1.5 s) videos that was identical across each of the four Movie-Watching runs (to facilitate test-retest and/or validation analyses). We focused on MOVIE1 because it was the only Movie-Watching run previously found to elicit functional brain organization predictive of both cognition and emotion[31]. Audio was delivered via Sensimetric earbuds.

fMRI Preprocessing

For the Emotion, Gambling, Social Cognition and Working Memory task fMRI data, we began with the HCP's minimally preprocessed data. The HCP minimal preprocessing pipeline for fMRI data removes spatial distortions, realigns volumes to compensate for subject motion, registers the echo planar functional data to the structural data, reduces the bias field, normalizes the 4D image to a global mean, and masks the data with a final FreeSurfer-generated brain mask[73]. We additionally applied 4mm spatial smoothing, 150 sec temporal smoothing, removal of non-steady-state frames (first 15 frames), and ICA-FIX denoising[74], which denoises residual subject motion, physiological motions, and artifacts, to the HCP minimally preprocessed task fMRI data. ICA-FIX was implemented via the Quantitative Neuroimaging Environment & Toolbox (QuNex)[75]. For movie-watching fMRI data, we used the FIX-denoised data already provided by the HCP and removed non-steady frames (i.e., the first 15 frames).

Tensor Independent Component Analysis (tICA)

tICA is a well-validated[38,50] unsupervised learning technique that detects distributed sets of brain activity whose time courses align with one or more aspects of a task stimulation time course. It can thereby identify, trace the moment-by-moment activity of, and quantify individual differences in task-evoked brain network activity. We applied tICA separately to the denoised fMRI data from each task group using MELODIC[50,76] in FSL (version 6.0). This identified task-recruited brain networks, including their voxel-wise spatial maps, their frame-wise time courses, and subject loadings (i.e., the extent to which a subject's data loads on the spatial and temporal pattern of the group-level network) on each network, for each task in both groups. For each task,

we ran tICA at varying model orders to inspect the output for the optimal model order for analysis. Because the goal was to investigate large-scale canonical brain networks, the model order of the group ICA needed to be relatively low to result in a coarse spatial scale parcellation of the brain into networks. Model orders that are too high produce finer spatial scale parcellations that reflect excessive segregation of known large-scale networks into individual or bilateral brain areas rather than networks. At the same time, model orders that are too low fail to fully disaggregate known canonical brain networks. Therefore, we tested a range of model orders from 5 to 25 to identify which decomposition parcellated the brain into canonical network patterns while effectively unmixing any noise/artifacts into separate components. For each model order, we examined the time course, power spectrum and spatial map of each component to identify those most likely to reflect task-related brain activity (e.g., components with spatial patterns resembling large-scale brain networks and with temporal courses reflecting one or more aspects of the task time course). Ultimately, we identified to d = 20 to be the optimal model order in our data that best achieved the balance between disaggregating large-scale canonical networks and not producing excessive segregation; this model order is also in line with previous reports examining the correspondence between task activation networks and canonical resting state brain networks[77,78]. Among the 20 components, those identified as noise (e.g., physiological, motion, scanner artifacts, drift) via examination of their time course, power spectrum and spatial distribution were removed from analysis, in line with best practices for identification of ICA noise components[79]. The identified network components were assessed for both spatial and temporal reproducibility across Group 1 and Group 2. For each group, before input into hierarchical clustering for subject subtyping, subject loadings were normalized by subtracting network loading means and then dividing by network loading standard deviations.

Assessing Overlap with Canonical Networks

To further characterize the spatial maps of the tICA-identified network families, we used the network correspondence toolbox (NCT)[54] to compute overlap (Dice coefficients) between the positive and negative components of each task network and each of 17 canonical large-scale networks defined in the Gordon et al. parcellation[53]. Positive and negative components were analyzed separately because we observed visually that the positive and negative parts of each component typically represented distinct networks. More generally, the tensor ICA can identify states comprised of anti-correlated network patterns that have been reported by others during both resting state and task performance (see for example Li et al. 2021[80]) as an important characteristic of brain organization. As such, the positive and negative parts were analyzed separately to determine their distinct network makeups. In order to assess statistical significance, a spin test implemented in Python was performed with 1000 permutations for each reference atlas network[54]. Significant overlap was thresholded at $p < .05$.

Hierarchical Factor Analysis (hFA)

hFA identified sets of highly correlated variables from the 87 measured by HCP assessments. These variables included behavioral measures of motor and sensory functioning, alertness, and cognition, including in-scanner task performance measures, as well as self-report measures of emotion, personality, substance use, psychiatric symptoms and general life function[42]. Full details of the procedure are outlined in Schottner et al., 2023[42]. Briefly, prior to the application of

factor analysis, missing assessment values were first imputed using Multiple Imputation by Chained Equations. Next, age and gender were regressed out by entering them as predictors in a linear regression. The residuals of this regression served as the inputs to the factor analysis. Reaction times and other error measures were inverted, and data were z-scored across subjects. We used the five-factor solution which included well-being/positive affect, internalizing/negative affect, processing speed, cognition and substance use[42].

Machine Learning Prediction of Mental Health Scores

For the machine learning regression analyses, the number of trees in the random forest was set to 1000. No maximum depth was specified for trees, the minimum number of samples required at each leaf was set to 1, the minimum number of samples required to split an internal node was set to 2, and the maximum number of features to consider when looking for the best split was set to 1. This analysis was carried out using scikit-learn[81].

Hierarchical Clustering

Hierarchical clustering was used for two applications. First, to cluster spatially similar families of brain networks identified from the application of tICA to the five fMRI paradigms (**Fig. 1**). Second, to identify subject subtypes based on brain function and mental health data (**Fig. S1**; **Figs. 5-6**). Hierarchical clustering was implemented by the nets_hierarchy function in the FSLNets package (fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets).

Data Availability

The HCP-YA dataset is publicly available on an open access repository (https://db.humanconnectome.org/app/template/Login.vm), which can be accessed after signing a data use agreement. Individual network maps identified by tICA for the Emotion, Gambling, Working Memory, Social, and Movie-Watching paradigms have been made viewable and accessible in a "Task-Evoked Network Atlas" at https://identifiers.org/neurovault.collection:20820.

Code Availability

All code used in this paper is freely and publicly available. Code to run tensor independent component analysis (tICA) is available in Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library (FSL). Code to run the Network Correspondence Toolbox (NCT) is available here: https://github.com/rubykong/cbig_network_correspondence. Code to run factor analysis on the HCP-YA dataset to generate latent mental health factor scores is available here: https://github.com/connectomicslab/hcp-behavioral-domains. Code to run SHapely Additive exPlanations (SHAP) is available here: https://shap.readthedocs.io/en/latest/. Code to run random forest machine learning regression is available here: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

References

1 Cuthbert, B. N. The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology. *World psychiatry* **13**, 28-35 (2014).

2 Hyman, S. E. Can neuroscience be integrated into the DSM-V? *Nature Reviews Neuroscience* **8**, 725-732 (2007).

3 Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC medicine* **11**, 1-8 (2013).

4 Ooi, L. Q. R. *et al.* Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *Neuroimage* **263**, 119636, doi:10.1016/j.neuroimage.2022.119636 (2022).

5 Dhamala, E., Jamison, K. W., Jaywant, A., Dennis, S. & Kuceyeski, A. Distinct functional and structural connections predict crystallised and fluid cognition in healthy adults. *Human brain mapping* **42**, 3102-3118 (2021).

6 Tian, Y., Yeo, B. T., Cropley, V. & Zalesky, A. High-resolution connectomic fingerprints: Mapping neural identity and behavior. *NeuroImage* **229**, 117695 (2021).

7 Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654-660, doi:10.1038/s41586-022-04492-9 (2022).

8 Rasero, J., Sentis, A. I., Yeh, F.-C. & Verstynen, T. Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLoS computational biology* **17**, e1008347 (2021).

9 Xiao, Y. *et al.* Predicting visual working memory with multimodal magnetic resonance imaging. *Human brain mapping* **42**, 1446-1462 (2021).

10 Greene, A. S., Gao, S., Scheinost, D. & Constable, R. T. Task-induced brain state manipulation improves prediction of individual traits. *Nature communications* **9**, 2807 (2018).

11 Jiang, R. *et al.* Task-induced brain connectivity promotes the detection of individual differences in brain-behavior relationships. *NeuroImage* **207**, 116370 (2020).

12 Finn, E. S. *et al.* Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience* **18**, 1664-1671 (2015).

13 Thiele, J. A., Faskowitz, J., Sporns, O. & Hilger, K. Multitask brain network reconfiguration is inversely associated with human intelligence. *Cerebral Cortex* **32**, 4172-4182 (2022).

14 Tetereva, A., Li, J., Deng, J. D., Stringaris, A. & Pat, N. Capturing brain-cognition relationship: Integrating task-based fMRI across tasks markedly boosts prediction and test-retest reliability. *Neuroimage* **263**, 119588, doi:10.1016/j.neuroimage.2022.119588 (2022).

15 Gal, S., Coldham, Y., Tik, N., Bernstein-Eliav, M. & Tavor, I. Act natural: Functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity. *NeuroImage* **258**, 119359 (2022).

16 Haruno, M. & Kawato, M. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of neurophysiology* **95**, 948-959 (2006).

17 Kim, S. H. & Hamann, S. Neural correlates of positive and negative emotion regulation. *Journal of cognitive neuroscience* **19**, 776-798 (2007).

18 Wolf, I., Dziobek, I. & Heekeren, H. R. Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. *Neuroimage* **49**, 894-904 (2010).
19 Parkes, L., Satterthwaite, T. D. & Bassett, D. S. Towards precise resting-state fMRI biomarkers in psychiatry: synthesizing developments in transdiagnostic research, dimensional models of psychopathology, and normative neurodevelopment. *Current Opinion in Neurobiology* **65**, 120-128 (2020).
20 Laumann, T. O., Zorumski, C. F. & Dosenbach, N. U. Precision neuroimaging for localization-related psychiatry. *JAMA psychiatry* **80**, 763-764 (2023).
21 Hill, E. D. *et al.* Prediction of mental health risk in adolescents. *Nature Medicine*, 1-7 (2025).
22 Cui, Z. *et al.* Linking individual differences in personalized functional network topography to psychopathology in youth. *Biological psychiatry* **92**, 973-983 (2022).
23 Chen, J. *et al.* Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nature communications* **13**, 1-17 (2022).
24 Yip, S. W., Kiluk, B. & Scheinost, D. Toward addiction prediction: an overview of cross-validated predictive modeling findings and considerations for future neuroimaging research. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **5**, 748-758 (2020).
25 Yarkoni, T. & Westfall, J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* **12**, 1100-1122 (2017).
26 Korponay, C., Janes, A. C. & Frederick, B. B. Brain-wide functional connectivity artifactually inflates throughout functional magnetic resonance imaging scans. *Nature human behaviour*.
27 Kashyap, R. *et al.* Individual-specific fMRI-Subspaces improve functional connectivity prediction of behavior. *NeuroImage* **189**, 804-812 (2019).
28 Korponay, C., Stein, E. A. & Ross, T. J. Misconfigured striatal connectivity profiles in smokers. *Neuropsychopharmacology* **47**, 2081-2089 (2022).
29 Scheinost, D. *et al.* Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* **193**, 35-45 (2019).
30 Gordon, E. M. *et al.* Precision functional mapping of individual human brains. *Neuron* **95**, 791-807. e797 (2017).
31 Finn, E. S. & Bandettini, P. A. Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *Neuroimage* **235**, 117963, doi:10.1016/j.neuroimage.2021.117963 (2021).
32 Zhao, W. *et al.* Task fMRI paradigms may capture more behaviorally relevant information than resting-state functional connectivity. *NeuroImage* **270**, 119946 (2023).
33 Kong, R. *et al.* Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. *Cereb Cortex* **29**, 2533-2551, doi:10.1093/cercor/bhy123 (2019).
34 Li, J. *et al.* Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage* **196**, 126-141 (2019).
35 Liégeois, R. *et al.* Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nature communications* **10**, 2317 (2019).
36 Kong, R. *et al.* Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral cortex* **31**, 4477-4500 (2021).

37 West, H. V., Burgess, G. C., Dust, J., Kandala, S. & Barch, D. M. Amygdala activation in cognitive task fMRI varies with individual differences in cognitive traits. *Cognitive, Affective, & Behavioral Neuroscience* **21**, 254-264 (2021).
38 Korponay, C. *et al.* Concurrent large-scale brain dynamics during the emotional face matching task and their relation to behavior and mental health. *bioRxiv*, 2024.2009. 2021.613739 (2024).
39 Savage, H. *et al.* Do you match what I match? A systematic review of 20 years of the emotional face matching task reveals vast inconsistencies in design and implementation. (2023).
40 Nikolaidis, A. *et al.* Suboptimal phenotypic reliability impedes reproducible human neuroscience. *BioRxiv*, 2022.2007. 2022.501193 (2022).
41 Gell, M. *et al.* How measurement noise limits the accuracy of brain-behaviour predictions. *Nature Communications* **15**, 1-12 (2024).
42 Schottner, M. *et al.* Exploring the latent structure of behavior using the Human Connectome Project's data. *Sci Rep* **13**, 713, doi:10.1038/s41598-022-27101-1 (2023).
43 Eisenberg, I. W. *et al.* Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature communications* **10**, 2319 (2019).
44 Enkavi, A. Z. *et al.* Large-scale analysis of test–retest reliabilities of self-regulation measures. *Proceedings of the National Academy of Sciences* **116**, 5472-5477 (2019).
45 Peng, S. *et al.* Heterogenous brain activations across individuals localize to a common network. *Commun Biol* **7**, 1270, doi:10.1038/s42003-024-06969-x (2024).
46 Pessoa, L. A network model of the emotional brain. *Trends in cognitive sciences* **21**, 357-371 (2017).
47 Sesack, S. R. & Grace, A. A. Cortico-basal ganglia reward network: microcircuitry. *Neuropsychopharmacology* **35**, 27-47 (2010).
48 Barrett, L. F. & Satpute, A. B. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology* **23**, 361-372 (2013).
49 Bressler, S. L. & Menon, V. Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences* **14**, 277-290 (2010).
50 Beckmann, C. F. & Smith, S. M. Tensorial extensions of independent component analysis for multisubject FMRI analysis. *Neuroimage* **25**, 294-311 (2005).
51 Cole, M. W. *et al.* Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature neuroscience* **16**, 1348-1355 (2013).
52 Mohr, H. *et al.* Integration and segregation of large-scale brain networks during short-term task automatization. *Nat Commun* **7**, 13217, doi:10.1038/ncomms13217 (2016).
53 Gordon, E. M. *et al.* Precision Functional Mapping of Individual Human Brains. *Neuron* **95**, 791-807 e797, doi:10.1016/j.neuron.2017.07.011 (2017).
54 Kong, R. Q. *et al.* A network correspondence toolbox for quantitative evaluation of novel neuroimaging results. *bioRxiv*, 2024.2006. 2017.599426 (2024).
55 Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. Permutation inference for the general linear model. *Neuroimage* **92**, 381-397 (2014).
56 Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E. & Smith, S. M. Multi-level block permutation. *Neuroimage* **123**, 253-268 (2015).

57 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, doi:10.1016/j.neuroimage.2019.116157 (2019).
58 Noble, S., Scheinost, D. & Constable, R. T. A guide to the measurement and interpretation of fMRI test-retest reliability. *Curr Opin Behav Sci* **40**, 27-32, doi:10.1016/j.cobeha.2020.12.012 (2021).
59 McGraw, K. O. & Wong, S. P. Forming inferences about some intraclass correlation coefficients. *Psychological methods* **1**, 30 (1996).
60 Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **30** (2017).
61 Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods* **50**, 1166-1186 (2018).
62 Sadraee, A., Paulus, M. & Ekhtiari, H. fMRI as an outcome measure in clinical trials: A systematic review in clinicaltrials. gov. *Brain and Behavior* **11**, e02089 (2021).
63 Drobyshevsky, A., Baumann, S. B. & Schneider, W. A rapid fMRI task battery for mapping of visual, motor, cognitive, and emotional function. *Neuroimage* **31**, 732-744 (2006).
64 Kapur, S., Phillips, A. G. & Insel, T. R. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular psychiatry* **17**, 1174-1179 (2012).
65 Brucar, L. R., Feczko, E., Fair, D. A. & Zilverstand, A. Current approaches in computational psychiatry for the data-driven identification of brain-based subtypes. *Biological psychiatry* **93**, 704-716 (2023).
66 Glasser, M. F. *et al.* The Human Connectome Project's neuroimaging approach. *Nat Neurosci* **19**, 1175-1187, doi:10.1038/nn.4361 (2016).
67 Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: an overview. *Neuroimage* **80**, 62-79, doi:10.1016/j.neuroimage.2013.05.041 (2013).
68 Barch, D. M. *et al.* Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage* **80**, 169-189 (2013).
69 Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F. & Weinberger, D. R. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage* **17**, 317-323 (2002).
70 Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. & Fiez, J. A. Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of neurophysiology* **84**, 3072-3077 (2000).
71 Castelli, F., Happé, F., Frith, U. & Frith, C. in *Social neuroscience* 155-169 (Psychology Press, 2013).
72 Wheatley, T., Milleville, S. C. & Martin, A. Understanding animate agents: distinct roles for the social network and mirror system. *Psychological science* **18**, 469-474 (2007).
73 Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage* **80**, 105-124, doi:10.1016/j.neuroimage.2013.04.127 (2013).
74 Salimi-Khorshidi, G. *et al.* Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage* **90**, 449-468, doi:10.1016/j.neuroimage.2013.11.046 (2014).

75 Ji, J. L. *et al.* QuNex—An integrative platform for reproducible neuroimaging analytics. *Frontiers in Neuroinformatics* **17**, 1104508 (2023).
76 Beckmann, C. F. & Smith, S. M. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging* **23**, 137-152, doi:10.1109/TMI.2003.822821 (2004).
77 Smith, S. M. *et al.* Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the national academy of sciences* **106**, 13040-13045 (2009).
78 Nickerson, L. D. Replication of resting state-task network correspondence and novel findings on brain network activation during task fMRI in the human connectome project study. *Scientific reports* **8**, 17543 (2018).
79 Griffanti, L. *et al.* Hand classification of fMRI ICA noise components. *Neuroimage* **154**, 188-205 (2017).
80 Li, M. *et al.* Co-activation patterns across multiple tasks reveal robust anti-correlated functional networks. *Neuroimage* **227**, 117680 (2021).
81 Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics* **8**, 14 (2014).