Title: An interpretable connectivity-based decoding model for classification of chronic marijuana use

Authors: *Kaustubh R. Kulkarni[1], *Matthew Schafer[1], Laura Berner[1], Vincenzo G. Fiore[1], Matt Heflin[1], Gaurav Pandey[2], Kent Hutchison[3], Vince Calhoun[4], Francesca Filbey[5], Daniela Schiller[1], Xiaosi Gu[1]

[1]Center for Computational Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]Department of Genetics and Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[3]Institute for Cognitive Science, University of Colorado, Boulder, CA, USA
[4]Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA
[5]Center for BrainHealth, School of Behavioral and Brain Sciences, University of Texas at Dallas, Texas, USA

*Authors contributed equally

Corresponding author:
Kaustubh R. Kulkarni
kaustubh.kulkarni@icahn.mssm.edu
Center for Computational Psychiatry, Icahn School of Medicine at Mount Sinai
55 W 125th St, New York, NY 10027

ABSTRACT

Background: Psychiatric neuroimaging typically proceeds with one of two approaches: encoding models, which aim to model neural mechanisms, and decoding models, which aim to predict behavioral or clinical features from brain data. In this study, we seek to combine these aims by developing interpretable decoding models that offer both accurate prediction and novel neural insight, using substance use disorder as a test case.

Methods: Chronic marijuana (MJ) users (n=195) and non-using healthy controls (n=128) completed a cue-elicited craving task during functional magnetic resonance imaging. Linear machine learning algorithms were used to classify MJ use based on task-evoked, whole-brain functional connectivity. We then used graph theoretical measures to explore 'predictive functional connectivity' and to elucidate whole-brain regional and network involvement implicated in chronic marijuana use.

Results: We obtained high accuracy (~80% out-of-sample) across four different linear models, demonstrating that task-evoked, whole-brain functional connectivity can successfully differentiate chronic marijuana users from non-users. Subsequent network analysis revealed key predictive regions (e.g., anterior cingulate cortex, dorsolateral prefrontal cortex, precuneus) that are often found in neuroimaging studies of substance use disorders, as well as some key exceptions – such as sensorimotor and visual areas. We also identified a core set of networks of brain regions that contributed to successful classification, comprised of many of the same predictive regions.

Conclusions: Our dual aims of accurate prediction and interpretability were successful, producing a predictive model that also provides interpretability at the neural level. This novel approach may complement other predictive-exploratory approaches for a more complete understanding of neural mechanisms in drug use disorders and other neuropsychiatric disorders.

INTRODUCTION

Psychiatric neuroimaging has two main goals: describing the neural mechanisms of mental dysfunction and predicting clinical characteristics from neural data, both of which have been explored and reviewed in great detail[1]. These goals are typically approached with different statistical and inferential paradigms, with different strengths and weaknesses[2]. Common functional magnetic resonance imaging (fMRI) modeling approaches test hypotheses about how mental processes are represented in brain signals, allowing investigations into the neural mechanisms of psychiatric disorders. Such "encoding" approaches model brain activity as a function of different features [i.e., giving *p(Brain|Features)*, or probability of brain activity conditioned upon features], and do not easily give inferences about processes or clinical categories from brain activity [i.e., *p(Features|Brain)*]. Given the functional diversity of the brain regions implicated in psychiatric disorders, establishing the functional specificity of a brain signal is difficult and limits the ability of encoding models to predict clinical characteristics from brain data[3].

In contrast, "decoding" models provide the opposite type of inference, as in these models, neural data are used to predict features, such as clinical diagnosis [i.e., *p(Diagnosis|Brain)*]. Machine learning (ML) models are often used for this purpose; in psychiatry, and substance use disorders specifically, numerous machine learning approaches have been used, including support vector machines[4–7], logistic regression[8–10], and others[11–13]. However, decoding models do not necessarily give insight into structure-function relationships, or even neurobiological plausibility[2] and are usually considered less interpretable than encoding models.

There have been numerous attempts to unify the descriptive ("encoding") and predictive ("decoding") approaches, especially in recent years, as mounting evidence demonstrates the advantages of both. Linearizing encoding methods such as representational similarity analysis[14–16] explicitly link feature space and brain space, revealing patterns of brain activity that are significantly different during varied both perceptual and non-perceptual task stimuli. Similarly, hyperalignment[17–19] and related approaches[20–22] transform high-dimensional brain spaces such that feature/clinical spaces are as aligned as possible; these representational spaces can subsequently be used for high fidelity decoding. Although each of these is an attempt to link encoding and decoding approaches, none of them explicitly generate predictive models which can subsequently be highly interpretable. In this regard, machine learning approaches, involving separate training and held-out testing sets, provide the strongest decoding constraint for a maximally powerful inference during the interpretation of the model[23]. Wager and others[2,24–26] demonstrate the value of this class of models in elucidating neural signatures, though they largely focus on constructs of physical and emotional pain.

One way to improve decoding models' interpretability is through theory-based modeling decisions: about the types of neural features to train the model on (feature selection), or constraints of the model[27–29]. For example, there may be more information about psychiatric dysfunction in the interactions between regions than in the activities of isolated regions[30]. This generates competing hypotheses in the same data: the predictive performance of models trained on features that capture information about neural interactions (e.g., functional connectivity) can be compared with models trained on the regional activities alone: the features that contain more relevant information should produce better predictions.

Another way to gain insight from decoding models is by performing a systematic evaluation of the fitted model. For example, decoding models trained on functional connectivity can be seen as indicating the features of network activity that are predictive of the outcome. One novel interpretation approach is to combine the trained model weights with the functional correlation values and apply network analysis to the resulting predictive (i.e., weighted) connectivity. In recent years, network neuroscience has emerged as a powerful tool to provide essential metrics and methods to uncover complex brain interactions[31–34]. We employ these network analytic methods to infer brain structures critical for accurate differentiation of chronic marijuana (MJ) users from non-users. Importantly, the inferences we draw about group differences in network features are constrained by the predictive performance of the decoding model [i.e., *p(Network Features|(Diagnosis|Connectivity))*].

In this study, we use a large fMRI dataset[35,36] collected from individuals with and without chronic MJ use (i.e., cannabis use disorder). In recent years, decreased perception of adverse effects of cannabis has coincided with increased usage[37–40] and legalization efforts. Although the adverse clinical effects of cannabis have been well-established[38,41–44], research on them has been hampered by the absence of reliable mechanistic biomarkers of cannabis use disorder. With our predictive and interpretable model, we aim to address this critical gap in knowledge.

We present here a novel modeling approach to balance the dual goals of clinical prediction and mechanistic understanding. To our knowledge, this is, to date, the largest fMRI sample used in the classification of substance use disorders (n=323), the first attempt to classify chronic MJ use with fMRI, and the first utilization of network analysis to interpret a fMRI decoding model. We trained linear decoding models on whole brain functional connectivity from individuals with chronic use and healthy controls during a marijuana cue-induced craving paradigm. The models predicted chronic use of MJ with high accuracy in out-of-sample participants (~80%) and outperformed models that used only regional activities - suggesting that the interactions between brain regions contained more information about the differences between these groups. Network analyses on the predictive connectivity matrices (i.e., functional connectivity weighted by the model coefficients from predictive models) identified brain regions and networks important to successful use classification, demonstrating the utility of interpretable decoding models for neurobiological description.

RESULTS

Model training: classification of chronic marijuana use

We first trained decoding models using two different linear machine learning algorithms (logistic regression and linear support vector classification [SVC]) to classify a clinical label of chronic marijuana use from whole-brain functional connectivity. Two regularization penalty types (L1 and L2 penalty) were chosen to be applied to each learning algorithm, for a total of four candidate learning algorithms. The model inputs consisted of a 4,005-element vector representing pairwise correlation values between every region in the brain as defined by the Stanford 90 region of interest (ROI) atlas (see Methods for more details). The full dataset (two runs each from n=195 chronic marijuana users, n=128 non-users) consisted of 646 total runs, and was divided into training and testing splits: 80% for training (516 samples), and 20% (130 samples) for out-of-sample testing to validate the performance of the best-performing models from training. The 80% training set was further divided into 10 folds that were used to optimize our chosen hyperparameter (regularization penalty strength [alpha]) using cross-validation. The complete pipeline is depicted in **Fig. 1**. Cross-validated accuracies for each combination of hyperparameters are summarized in **Table 1**.

Model training: connectivity- vs. activity-based models' performance

To test our hypothesis that models trained on functional connectivity would have higher prediction accuracy than models trained on regional activities, we performed the same hyperparameter search as described above with models trained on mean regional activation distances: the pairwise absolute value differences between regions' mean estimated activity. Mean distances between the estimates were used rather than the estimates themselves to keep the number of features constant across the models. Three different types of regional activity estimates were tested: mean time courses, mean marijuana cue betas (mj_{cue}) and the contrast of mean marijuana cue betas minus mean control cue betas ($mj_{cue}>ctl_{cue}$). In the cross-validation training set, the highest performing models from each type of estimate had substantially lower accuracy than the best functional connectivity models (time course=60.4%, mj_{cue}=65.1%, $mj_{cue}>ctl_{cu}$=60.5%), supporting our hypothesis that functional correlations are more informative than isolated activities for differentiating chronic MJ users from healthy controls.

Model training: hyperparameter performance

Generally, the L2 penalty was associated with better performance for both logistic regression and linear SVC, and lower alpha (corresponding to lower regularization strength) increased model performance, indicating that widespread information from many region-region correlations contributed to classification success. From the hyperparameter optimization results, we selected 0.0001 as the alpha parameter for following analyses, given its reliably strong cross-validated performance across all penalty types and classification algorithms. Since both L1 and L2 penalties for both algorithms performed well for a range of alpha values, we used both for final evaluation of the logistic regression and linear SVC models on the training and testing sets (2 models x 2 penalties x 1 alpha level = 4 tested models). The receiver operating characteristic (ROC) curves shown in **Fig. 2** demonstrate the classification ability across various decision thresholds within the training set.

Model testing: held-out data performance

After confirming high cross-validated performance in the training dataset, we tested out-of-sample performance on the previously held-out data. As described in **Fig. 1c**, the four models were trained on the full training dataset and tested on the held-out data (20% of total sample). The performance metrics are summarized in **Table 2** for accuracy, AUC, and precision and recall per clinical group for each model. Note that the accuracies for these models are much higher than chance testing set accuracies defined in three different ways: models that simply select the dominant class (60%) or randomly guess (53.85%), or the averages of permutation generated (1000 shuffles of subjects' training labels) distributions (L1 logistic: 53%, L2 logistic: 53.2%, L1 SVC: 52.8%, L2 SVC: 53.6%).

Model predictions and ensemble probabilities as well as true subject labels are compared in **Fig. 3**. Models produced very similar predictions, with high similarity for the pairwise comparisons (Jaccard similarity coefficients: 0.81-0.91). The learned model weights also had high rank similarity (Kendall's tau coefficients: 0.70-0.75), suggesting similar relationships were learned by the different algorithms. All models had good classification accuracy, demonstrating that linear modeling of whole brain functional correlations is effective in classifying chronic marijuana use.

Model interpretation: predictive connectivity

Following confirmation of out-of-label clinical relationships with model metrics, we interpreted the model weights to infer the brain connectivity structure implicated in chronic marijuana use (**Fig. 1d**). One advantage of using linear models is that the input features and learned model weights share the same shape. Further, in our case the model weights are the signed scaling of region-to-region connectivity values that, when added, produce the classification decisions. This enables meaningful interpretability of the brain connectivity patterns most significantly implicated in differentiating clinical cases from controls. First, we constructed model weight matrices for each of our four linear models. Reversing the procedure applied to the initial correlation matrices (see Methods) reverted the 4,005 model weights for each model to a 90x90 region-to-region feature weight matrix. Then we combined the model weights with the region-to-region correlation magnitudes, subject by subject, to produce weighted connectivity matrices - which we refer to as "predictive connectivity." Then we used two approaches: (1) evaluated the regions with the highest predictive importance, per model, and (2) used a graph theoretical analysis was performed on the model weight matrices to examine their network properties.

Posterior checks/Predictive importance

Model weights are interpretable only within the context of input magnitudes. That is, the strength and sign of the correlation input values determines their importance to prediction. Simply knowing the model weight is not sufficient to determine that the correlation value is important for prediction; it is necessary to examine the magnitude of the weighted correlation value. To this end, we performed a series of posterior checks on weights and correlation values to determine the brain regions most important for prediction in each algorithm.

First, we generated a mean whole-brain correlation matrix across all subjects. This matrix was then multiplied elementwise by the corresponding model weights for each algorithm to generate a whole-brain weighted connectivity matrix for each subject. We then calculated the mean weighted correlation for each region with all other regions. As the magnitude of weighted correlation was important for prediction, the absolute value of this mean weighted correlation was used to rank each region. Across models, we observed high consistency in the predictive importance ranks (Kendall's tau coefficients=0.76-0.83, p-values=$8.45e^{-27}$-$4.71e^{-31}$). Given the similarities in weights, predictions and regions of predictive importance across models, we assumed relative stability across the models and selected one model for subsequent interpretations: the L2 linear SVC (alpha=0.0001), given its high and stable performance across the ranges of alpha tested. For our chosen L2 SVC model, the top twenty regions in terms of mean predictive importance are visualized in **Fig. 4a** and include brain regions such as bilateral anterior cingulate cortex (ACC), left pre/postcentral gyri, right middle frontal gyrus, and right inferior parietal cortex.

As an exemplar region, bilateral ACC showed high mean predictive importance across all models, so its unweighted regional connectivity strengths to every other region were further visualized in **Fig. 4b**. Among the regional connections to ACC, most regions identified as having high weighted connectivity also have high magnitudes of original connectivity strength (r_{mean}=0.478, r_{std}=0.0628). Importantly, however, a number of regions show relatively small magnitudes of correlation strength but high predictive importance (reflected by high model weights). Our interpretable model weights approach allows for the identification of these small fluctuations in connectivity differences that are nevertheless highly implicated in differentiating chronic users from healthy controls.

To examine whether top 20 regions identified by our top weighted connectivities are consistent with those reliably implicated in craving, we compared our region-specific predictive importance scores i.e., ranked weighted connectivities, to uniformity and association maps retrieved from Neurosynth.org using a term-based meta-analysis. The 'craving' keyword yielded aggregated activation maps from 80 published studies thresholded at FDR-corrected $p<0.01$. The most significantly active regions identified by this meta-analytic approach include medial prefrontal cortex, middle cingulate cortex, medial prefrontal cortex, and medial parietal lobule. Each Neurosynth map was projected on an anatomical map and the Stanford functional ROIs with high weighted connectivity (predictive importance) were overlaid on top (**Fig. 5**). Map comparisons were restricted to a qualitative overview due to the highly dissimilar sparsity of the maps, as well as significant differences in the sizes of Stanford ROIs and the activation loci in the Neurosynth maps. Mainly, we aimed to demonstrate that our findings fall in line with current literature, but in summary, regions in the meta-analytic craving map qualitatively show a moderate level of correspondence to regions identified as having high predictive importance, supporting and validating our interpretation pipeline.

Graph theoretical analysis

Our next goal was to more formally investigate the relevant networks that distinguish users and non-users. To this end, a graph theoretical approach was applied to investigate the network properties of regions involved in predicting a label of chronic marijuana use. Graph properties were calculated at local, global, and meso-levels of topological organization (**Fig. 6**). Local properties are the properties of individual nodes in the graph (i.e., individual brain regions), global properties describe properties of the graph as a whole (i.e., the full brain network), and meso-scale properties describe clusters or sub-networks within the full graph that are highly modular (i.e., brain communities).

Just as we had previously generated a mean weighted connectivity matrix, for each subject, we generated a weighted connectivity matrix by performing an element-wise multiplication of the L2 SVC model weight matrix with the subject's connectivity matrix (**Fig. 7a**). Then, each subject's sparse representation of the weighted connectivity matrix was obtained by thresholding at 2% connection density. Matrix thresholding is a commonly used strategy in network neuroscience to remove spurious network connections, and improve stability and modularity of network features[45–48]. The absolute values of the weighted connectivity values were taken, as the magnitudes of weighted connectivity set as the strength of node-to-node connections in the graph. The transformed matrix was used to generate a sparse graph, where nodes represented regions, and the edges represented the strength (i.e., importance) of connectivity values between two regions to prediction.

At the brain (local) level, we calculated a region's importance to prediction by calculating a subject-specific binarized degree centrality score. Degree centrality (DC) of a node is defined as the summation of the number of above threshold connections to a node. Thus, we interpreted the DC as the number of predictively valuable connections a region has with other regions. Note that region-by-region DC is calculated for every subject, providing a measure of each region's importance to every subject's classification. To interpret DC scores, we first assessed *overall* predictive importance of regions. To do this, we performed non-parametric region-by-region significance testing of all subject DC values. Regions were ordered by greatest mean DC score, indicating level of predictive importance. The top twenty regions of highest average DC (along with bottom two lowest for comparison) are shown in **Fig. 7b**. Regions of high overall importance include bilateral ACC, right inferior parietal/angular gyrus, and right middle frontal gyrus. At the network level, regions from numerous resting state networks are represented in the top ten regions, with no particular functional network dominating. Results from this network significance analysis corroborate a widely distributed pattern of connectivity being important for distinguishing individuals who have long-term MJ use from controls.

Next, we discovered properties of network organization at the whole brain (global) level by calculating small-world network efficiency networks. Briefly, network efficiency is defined as the average ability of a network to transmit information in an effective manner, and also quantifies its fault-tolerance to node removal. In our weighted matrices of brain connectivity, efficiency provides a measure of the robustness of the network for

predictive strength. **Fig. 7c** displays global and local efficiencies of whole brain weighted connectivity networks for each diagnostic group. An independent samples t-test between the two groups reveals no difference between groups for global efficiency and local efficiency. These findings indicate that there are no significant differences in information propagation for prediction of users vs. non-users, i.e., both groups have global connectivity structures that are equally robust in prediction of the clinical label.

Finally, at the community (meso) scale, we used community detection algorithms to discover modular sub-networks within the weighted connectivity matrices. First subject-specific unthresholded weighted connectivity matrices were calculated as defined above. Then, matrices were averaged across all subjects and thresholded at 2% sparsity as above, creating a group-averaged thresholded weighted connectivity matrix. The Girvan-Newman community detection algorithm was then applied to discover community structure within this matrix. **Fig. 8a** shows the thresholded weighted connectivity matrix reorganized by the discovered community structure. Each community was then ranked by its average degree centrality (DC) score.

To confirm that this ranking actually reflects the predictive importance of each community, we performed a stepwise prediction analysis to determine the minimal number of communities necessary to produce good predictions. (**Fig. 9**) Starting with the regions in the highest DC ranked community, each region's (non-redundant) pairwise correlations to all other brain regions were used to generate each participant's distance to the hyperplane. With the inclusion of each additional community, the best performing decision threshold was determined in the training data and used to generate testing set predictions. The best testing set prediction came from the first 4 communities with 80% accuracy, outperforming even the overall model - and also performing significantly better than random regions (permutation tested p=0.001).

The highest ranked community included regions from bilateral ACC, bilateral supplementary motor area, right dorsolateral prefrontal cortex, and right inferior parietal/angular gyrus. The second top scoring community included right middle frontal gyrus, left angular gyrus, and bilateral medial precuneus regions. The top 4 modular communities that produced the best prediction are visualized in **Fig. 8b**.

DISCUSSION

Model testing: clinical prediction

In this study, we developed a novel modeling approach to balance accurate clinical prediction and model interpretability. Specifically, this approach classifies chronic marijuana users and healthy controls from task-based fMRI functional connectivity and subsequently identifies the individual regions and networks most important for this distinction. In the largest sample of individuals with long-term MJ use and healthy controls to date, we classified chronic use from functional connectivity during a cue-elicited craving task with nearly 80% out-of-sample accuracy. We used several different linear modeling approaches, all of which produced highly similar model weights, predictions, and regions with high mean predictive connectivity - suggesting they learned similar information. Our accuracies also compare favorably to previous fMRI decoding studies using functional connectivity to classify drug use, in both nicotine smoking[4,7,49] and cocaine use disorder[50] - even though most studies did not test out-of-sample or featured much smaller sample sizes (both of which can inflate prediction performance). Furthermore, this is one of the first fMRI study[51], and the largest to date, to classify chronic MJ use (i.e., cannabis use disorder) - a relatively understudied drug use disorder.

Model interpretation: predictive connectivity

Functional connectivity-based models outperformed models trained on regional activation estimates - suggesting there is more information about chronic MJ use in the interactions between regions than in their isolated activities. Given this, our next goal was to discover brain network patterns that differentiated the groups, starting with the individual regions that are most critical to successful prediction in the best performing model - the L2 linear SVC. Regions with high mean predictive connectivity were distributed across diverse resting state networks, such as the default mode, sensorimotor, salience and executive control networks - suggesting widespread functional differences between the healthy and MJ-using groups. Regions with widespread predictive connections were especially of interest and were judged by the number of functional connections between a region and the rest of the network that helped classify chronic use, so-called ‘predictive degree centrality’ (i.e., predictive DC).

There was high predictive DC in several sensory and motor related regions - including left inferior temporal gyrus, right inferior temporal cortex (both areas along the ventral visual pathway), bilateral primary somatosensory cortex and supplementary motor area. Given that the visual and tactile demands of the task were the same across groups, these regions likely reflect more than the passive reception of sensory information and output of motor commands. For example, these regions may facilitate the recognition of drug cues and retrieval of behavioral associations, such as the initiation of drug seeking/use behaviors[52]. Regions related to attention and its control also ranked highly on this measure - likely reflecting differential recruitment of attention during cue processing between the groups. For example, the right middle frontal gyrus, an important attentional control region and site of convergence for the dorsal and ventral attention networks[53], had the highest predictive DC of any measured region. Bilateral ACC and dorsolateral prefrontal cortex (PFC), areas that feature dense cannabinoid receptors[54] also ranked highly on this measure, corroborating previous reports of dysfunctional attentional and control-like processes during drug cue exposure and craving generally[55,56] and in MJ users specifically[57,58]. High predictive DC was detected in regions associated with cue-reactivity and craving, including the precuneus and posterior cingulate cortex, regions that may work together to process drug cue salience and relevance to the self[59] and in the bilateral medial PFC, which has extensive and recurrent dopaminergic connections with the ventral tegmental area and may direct drug-seeking behavior[60]. These findings suggest our method can recapitulate diverse findings from the literature.

We also discovered sets of brain regions (subnetworks) that were critical to successful prediction. The group-average predictive connectivity matrix was used to discover the community structure - patterns of connectivity that co-occur and likely share some functional basis. Thus, this analysis enables a shift from individual region’s predictive importance to predictive importances of distinct communities of brain regions.

The first four communities produced the best testing set prediction accuracy, even outperforming the inclusion of additional communities. The highest-scoring communities contained regions from different canonical networks, with two network connectivity motifs representing the majority of predictively important regions. The first of these network motifs included regions from bilateral ACC, posterior inferior temporal cortex and superior angular gyrus. The second motif included regions from inferior angular gyrus, middle frontal gyrus, and superior temporal cortex. It is not obvious how these motifs map onto more standard resting-state networks (e.g., salience network), suggesting these results may reflect task-specific network organization. Together, these region-level and community-level properties provide a unique neural signature that differentiates chronic marijuana users and non-users.

The functional diversity of the regions and communities implicated in these analyses suggests widespread functional differences between MJ users and controls - and the need for tasks that measure a wide range of structure-function hypotheses concurrently. It is possible that the relatively high accuracy we achieve in this dataset is due to the task: multiple sensory modalities and motor processes are engaged, allowing for more functional differentiation between individuals with MJ use and controls.

Added value and applications

In general, decoding approaches use whole brain information during model fitting, culminating in a single statistical test, versus more standard encoding approaches (e.g., general linear modelling) that generally perform up to many thousands of tests across the brain and require extensive multiple comparisons correction. To our knowledge, our specific approach represents the first application of network analysis to interpret model weights from predictive models. Further, our model interpretations are constrained on high decoding performance, conditioning our inferences upon the prediction of a real-world label (self-reported behavior).

Many extensions to this joint predictive/explanatory approach are possible. The network analysis may be refined at the spatial scale, by generating voxel-level connectivity matrices and recalculating network properties. Another possibility would be to build predictive models from regions of interest, in a more hypothesis-driven manner (e.g., derived from areas of significant activation in an encoding model). Additionally, a regression-based predictive model would be an improvement over the classifiers outlined here: such approaches can make stronger inferences about the neural patterns of clinical features directly (e.g., symptom severity), rather than indirect conclusions about patterns that differentiate clinical groups (i.e., chronic use or not)[2,24,28].

Limitations and next steps

There are several limitations to this study. First, these accuracies are not high enough for direct clinical deployment: any clinical useful tool would likely require accuracy greater than 90%, depending on the relative clinical burden of false positives and false negatives. Further, as mentioned above, the sample was divided into chronic cannabis users versus non-users, not allowing us to disentangle continuous effects related to use. We also predicted a categorical label based upon self-report and thus are bound by the accuracy of that label - not by real world behavior or underlying functional dimensions. This study also precludes most inferences about the specificity of the effect of marijuana use. Future work should compare marijuana users to chronic users of other drugs, as well as non-drug using individuals with other psychiatric dysfunction, in order to establish marijuana-specific neural signatures. Additionally, more data-driven parcellation approaches, using ICA, gradient-based methods, or multimodal data, may elucidate more robust, replicable, or task-evoked neural signatures associated with chronic MJ use[61–63]. Another important direction for this approach clinically would be a longitudinal study predicting future risk of chronic use, especially in adolescents or young adults. A model that can reliably predict maladaptive use at or beyond the ability of a physician would be highly valuable and might enable preemptive or preventative care for high-risk individuals prior to onset of severe symptoms.

This study is a first step towards building accurate and interpretable predictive models that have both theoretical and clinical significance. The models performed well in out-of-sample data, with high predictive

accuracies. Further, we interpreted the best performing model to both corroborate prior findings and discover novel network level properties in the context of drug use disorders. Future work can build on this approach of using joint predictive/explanatory models to constrain neurobiological inferences.

MATERIALS AND METHODS

Preprocessing of fMRI data

Subjects

This study combines data from two pre-existing fMRI datasets (n=125, n=198 respectively) measuring cue-elicited drug craving in participants recruited from the community (i.e. not treatment-seeking or inpatient) in Albuquerque, NM, with and without chronic marijuana use (CD n=195, HC n=128 respectively)[35,36]. The combined data set has a mean age of 30, with 65% male participants.

Scanner specifications

The two samples had different scan specifications, and as such are described separately below. 2009 sample: MRI images in this sample were collected in a 3T Siemens Trio scanner over two runs, for approximately 9 minutes and 22 seconds of scan time. T2* images were collected with a gradient echo, echo planar imaging protocol, with the following specifications: time to repetition (TR) of 2,000ms, time to echo (TE) of 27 ms, α: 70°, matrix size: 64 x 64, 32 slices, voxel size 3x3x4 mm3). High resolution T1-weighted images were collected with a multiecho magnetization prepared gradient echo (MPRAGE) sequence, TR=2,300ms, TE=2.74ms, flip angle = 8 deg, matrix = 256x256x176 mm, voxel size = 1x1x1mm. 2016 sample: MRI images in this sample were collected using a 3T Philips scanner, over two runs for total scan time of 7 minutes 54 seconds. T2*-weighted images were collected using a gradient echo, echo-planar sequence (TR: 2,000 ms, TE: 29 ms, flip angle: 75deg, matrix size: 64 x 64 x 39, voxel size: 3.44 x 3.44 x 3.5mm^3). High resolution T1-weighted images were collected with a MPRAGE sequence with the following parameters: TR/TE = 29/2,000 ms, flip angle=12 deg, matrix=256x256x160 mm, voxel size =1x1x1mm.

Task design

For the Filbey 2009 dataset[35], the task consists of two runs of a pseudo randomized order of 12 tactile/visual stimulus presentations. Two types of stimuli are presented: (1) a marijuana cue (pipe, bong, blunt, joint), and (2) a neutral cue (pencil). Cues are presented for 20 s, followed by a 5 s rating period, during which craving ratings are measured on an 11-point scale. This is followed by a 20 s fixation period. The full task consists of a total of 12 pseudorandomized cue presentations. The task structure for the Filbey 2016 dataset[36] is largely similar, but also includes a naturalistic cue (participant's chosen fruit) for a total of 3 cue types and 18 presentations per run. Craving ratings are measured just as described above.

Preprocessing

Results included in this manuscript come from preprocessing performed using FMRIPREP version stable[64], a Nipype[65] based tool. Each T1w (T1-weighted) volume was corrected for INU (intensity non-uniformity) using N4BiasFieldCorrection v2.1.0[66] and skull-stripped using antsBrainExtraction.sh v2.1.0 (using the OASIS template). Brain surfaces were reconstructed using recon-all from FreeSurfer v6.0.1[67], and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle[68]. Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c[69] was performed through nonlinear registration with the antsRegistration tool of ANTs v2.1.0[70], using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast[71] (FSL v5.0.9). Functional data was slice time corrected using 3dTshift from AFNI v16.2.07[72] and motion corrected using mcflirt[73] (FSL v5.0.9). This was followed by co-registration to the corresponding T1w using boundary-based registration[74] with six degrees of freedom, using bbregister (FreeSurfer v6.0.1). Motion correcting transformations, BOLD-to-T1w transformation and T1w-to-template (MNI) warp were concatenated and applied in a single step using antsApplyTransforms (ANTs v2.1.0) using Lanczos interpolation. Physiological noise regressors were extracted applying CompCor[75]. Principal components were estimated for the two CompCor variants: temporal (tCompCor) and anatomical (aCompCor). A mask to exclude signals with cortical origin was obtained by eroding the brain mask, ensuring it only

contained subcortical structures. Six tCompCor components were then calculated including only the top 5% variable voxels within that subcortical mask. For aCompCor, six components were calculated within the intersection of the subcortical mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run. Framewise displacement[76] was calculated for each functional run using the implementation of Nipype. Combined task/nuisance regression was then performed on the minimally preprocessed data using SPM12 (Wellcome Trust Centre for Neuroimaging). The nuisance regressor set consisted of the six realignment parameters, aCompCor regressors, discrete cosine-basis regressors, and a framewise displacement regressor. The task regressor set included onsets for marijuana cue presentation, marijuana cue rating period, control cue presentation, control cue rating period, and washouts for each cue. In addition, the Filbey 2016 dataset included regressors for fruit cue presentation and fruit cue rating period.

Parcellation

The noise-regressed voxelwise data was then parcellated using the Stanford functional ROIs for volumetric regions and networks, a highly validated scheme that is widely used for ROI-based and connectivity-based analyses[77]. The mean time series of each parcellated region was then computed. This procedure served a dual purpose: first, it increased signal-to-noise ratio for relevant brain areas compared to voxel-based analyses. Second, it reduced the dimensionality of the data for subsequent analysis, which allowed for faster and more accurate machine learning. The Stanford ROI atlas contains 90 regions, so the parcellation results in a 90 x (# of time points) matrix of whole brain activity for each subject.

Data preparation for classifiers

All 323 subjects (195 subjects with clinical label of chronic use) had two runs of data. Run length varied by the dataset from which the subject was taken. The subjects from the 2009 dataset had 281 TRs, and the subjects from the 2016 dataset had 405 TRs. For every subject, these TRs represented the totality of the run, including cue stimulus presentation periods, rating periods, and inter-trial intervals.

Each region's preprocessed time series was then correlated (Pearson) to all other regions' time series. Pearson correlation automatically standardizes each region's mean time series, so it is insensitive to differences in activation magnitude (i.e., scale) between the regions. Instead, it gives estimates of the pairwise timeseries activation similarities.

The decision to use parcellated functional connectivities was to 1) reduce the data dimensionality and the number of features relative to the number of observations, which is important in model fitting; 2) test the ability of network information to predict clinical label; and 3) improve our ability to subsequently interpret the fitted models, by using network analysis approaches. Further, functional connectivity has shown promise in other predictive modeling studies[30].

This approach each yielded a 646x90x90 matrix. To eliminate redundancy, only the upper triangles of the symmetric correlation matrices were retained (diagonal is each region's correlation with itself), leading to a final vector input size of $(90^2 - 90)/2 = 4{,}005$ features.

To allow for out-of-sample validation, the full sample was then divided into training and testing sets, using an 80/20 split: the training set included 516 samples (0.80 * 646) and the testing set included 130 (0.20 * 646). The training set was used for the 10-fold cross validated model fitting for algorithm selection and hyperparameter optimization. The testing set was set aside until the very end to test the out-of-sample fit of the four best performing models. All splits were constructed to balance the overall clinical label (CD or HC) proportions and include both runs of any subject completely in either the training or testing set.

Linear Classifiers

Four types of linear classifiers (L1 Logistic Regression, L2 Logistic Regression, L1 SVC, and L2 SVC) were used to predict clinical label and compared on their performance, including prediction accuracy, precision-

recall, and AUC scores. These linear classifiers were implemented using the scikit-learn package in Python[78,79]. Generally, to separate classes, linear classifiers learn a decision boundary that is a linear function (in greater than two dimensions, a hyperplane) in the feature space that then can be used to make class label predictions in new, out-of-sample data. In other words, the classification prediction (i.e., clinical label) is made based on the linear combination of the weighted input features - in our case, the whole-brain pairwise correlation values. How logistic regression and SVC learn linear boundaries varies; a brief description of each approach is given below.

Logistic regression learns the logistic function that best fits the observations: the resulting sigmoidal curve gives the probabilities that each observation is in either class, which are thresholded at 50% to produce the predictions. In contrast, SVC produces predictions by learning a hyperplane that separates the two classes by the largest distance (i.e., margin). The distances of the observations (each subject's brain-wide pairwise functional correlations) to the hyperplane were then converted to probabilities using Platt's method in the CalibratedClassifierCV class in Python's scikit-learn, to allow comparison to the logistic regression algorithm.

L1 and L2 regularization were used with both logistic regression and SVC to penalize different kinds of information. L1 ("Lasso") penalizes the magnitude of feature weights and in doing so produces a "sparse" feature space: only those features (e.g., region-region correlations) most informative to successful prediction will have a non-zero weight. Thus, L1 penalization reduces the number of features. This reduction may be important for two reasons: 1) when the number of features greatly outnumbers the number of observations, reducing the feature number can improve the fit and prediction and 2) feature interpretation should be improved: only the most important regions 'survive'. In contrast, L2 ("Ridge") penalizes the squares of feature weights and minimizes the feature weights, reducing their variance while retaining all of the features. This can improve prediction accuracy. Elastic net, a combination of L1 and L2 regularization, was also tested in the hyperparameter optimization (see Supplementary), but did not outperform lasso or ridge regression, and was subsequently removed from further analysis.

Various regularization strengths were tested in the training data, with larger strengths reflecting stronger penalization. In our Python-implemented machine learning pipeline, this regularization strength is represented by the alpha parameter, where higher alpha values reflected higher regularization. A range of alpha values were tested, from low to high regularization (alpha=1e-10, 1e-7, 1e-4, 0.1, 1, 10, 100, 1000). In general, low regularization was found to have the highest cross-validated training performance. The modeling parameters (i.e., hyperparameters) resulting in the highest overall accuracy in the training set were selected for the following analysis (alpha=1e-4, L1 and L2 penalties).

All models were cross-validated in the same way, with each fold stratified by class label to ensure the proportion of class labels was the same as in the larger dataset. In training, each fold's class labels were predicted by the model being trained on the other 9 folds. Further, each model had the same exact subject partitions to ensure maximal comparability on their training set performance.

For each model's cross-validated performance, a receiver operating characteristic curve (ROC) was generated, displaying fold-by-fold area under the curve (AUC) scores, as compared to chance. A separate ROC was generated to display AUC scores for the testing dataset.

Each linear model returned a set of trained model weights sharing the same architecture as the input features (i.e., the connectivity matrices), and which, when considered along with the values of the input features, are interpretable as the importances assigned to each pairwise connectivity in determining class label. These model weights were further explored using a correlation difference analysis and a network analysis, and results obtained were directly compared to a meta-analytic activation map examining brain activations implicated in drug craving.

Mean distance classification control

To test our assumption that functional correlations are more informative than more standard measures of activation magnitude, we also ran classification models with pairwise mean distances as model inputs. Three different mean distance controls were performed. In the first, the absolute value of the mean time series differences (i.e., mean distance) for all of the regional pairwise comparisons were used. In the second, the pairwise absolute value differences of the mean marijuana cue beta values (from the task regression) were used. In the third, the pairwise absolute value differences of the mean marijuana cue betas minus the mean control cue betas were used. These different mean distance controls were subjected to the same hyperparameter search as the functional connectivity inputs.

Predictive importance

Predictive importance analysis started by first generating an average connectivity matrix across the full data sample. To do this, the 90x90 connectivity matrix generated for each subject to be used as the input for the machine learning algorithm was averaged over all subjects. Next, the model weights were obtained for each linear model after training the full data sample. Note that this step differs from the model performance evaluation step above, which only used 4/5 of the data sample to train, and the other 1/5 to test. Finally, the element-wise multiplication (i.e. Hadamard product) was computed between each model weight matrix and the group-averaged connectivity matrix to generate the weighted connectivity matrix.

For each row in this matrix, corresponding to the weighted connectivity pattern associated with a particular region, the mean of the absolute value the pattern was taken to represent the average importance of the region's weighted connectivity on the prediction of the clinical label. Four such scores were generated for each of the 90 regions, and ranked by their average importance across all four algorithms. Algorithm rankings were statistically compared using Kendall's tau to assess correspondence between each pair of algorithms.

The top twenty regions of highest importance were selected to visually examine their individual connectivity patterns and corresponding weights. For these regions, group averaged connectivity patterns were generated for users and non-users separately. These per-group connectivities, along with the corresponding model weights were plotted together. This allowed for determination of the strength of regional connectivity and its impact on driving model performance. Most regions identified by this approach had relatively high connectivity values and significant differences between users and non-users; however, a few of the identified regions showed low connectivity values, but were nevertheless highly weighted, indicating that even small differences in connectivity across the groups in these regions were important for differentiating groups.

Region-specific predictive importance scores were validated by comparison to uniformity and association maps retrieved from Neurosynth[80]. The keyword 'craving' was used to yield aggregated activation maps from 80 published studies. The uniformity and association maps each provide unique information; the uniformity map displays regions of consistent activation across all studies, while the association map displays regions that are preferentially active in relation to keywords chosen over and above other keywords. Neurosynth provides activation maps for each thresholded at $p<0.01$ with FDR correction. Each activation map was then projected on the Stanford functional ROIs. For each Stanford ROI, proportion of non-zero voxels, and average non-zero signal was calculated. Finally, the scores were thresholded to limit reporting of voxel activity in regions that contained too few active voxels. Given the relative sparsity of the association map compared to the uniformity map, association was thresholded at 5% voxel participation and uniformity at 25% participation. Subsequent interpretation was performed by reporting the regions which survive thresholding with the top average activation scores.

Network Analysis

In the network analysis, the classifier-specific weighted connectivity matrices are treated as adjacency matrices to an undirected weighted graph. We first perform a pairwise multiplication of the 4,005-element model weight vector from each model with every subject's upper vector representation of connectivity (used as the input

value for predictive model) to generate a weighted connectivity vector for each subject. Then, we generate weighted connectivity matrices for each subject by reverting the 4005-element upper triangle representations into the native space 90x90 symmetric representation. For each subject, we take the magnitude of each element in the matrix as a measure of importance to model prediction. We threshold the dense weighted connectivity matrix at 2% density to improve signal-to-noise ratio and remove spurious connectivity strengths. The node-level and graph-level properties will be calculated on a binarized representation of the weighted connectivity matrix, so it is important to remove low-magnitude connections between nodes. Finally, we generate the graph structure by using the transformed weighted connectivity matrix as an adjacency matrix using the 'networkx' module in Python[81].

With a unique graph structure for each subject, we calculate subject-specific degree centrality (DC), a node-level graph property, which refers to a normalized summation of binarized connections to a node. In this graph, each node represents a brain region and connection edges between two nodes represent the importance of the connectivity between those two nodes for the classifier. Thus, nodes with high degree centrality can be considered to be brain regions whose connectivities to other regions help the classifier distinguish chronic users from non-users. Conversely, graph isolates are defined as nodes with lowest degree centrality across participants. In other words, they are brain regions whose connectivities to other regions do not help the classifier distinguish between chronic users and healthy controls. DC calculation is performed using built-in networkx 'degree_centrality' function, which accepts a graph structure and automatically calculates and normalizes the degree centrality of each node.

For each brain level (i.e., node in the network), the distribution of degree centrality for that region was aggregated, first across all participants, then across participants within each clinical group. First, a Mann-Whitney U test was performed to test for significance of each region's DC across all participants. Regional DC scores were ordered and reported by highest median score, and corresponding p-values also reported. Additionally, DC score ranking was compared across algorithms using Kendall's tau to assess correspondence of rankings across every pair of algorithms.

Graph-level metrics of graph structure were calculated next by deriving global and local efficiency scores at a subject-specific level. The efficiency between two nodes is defined as the multiplicative inverse of the shortest path between them. Global efficiency is an averaged measure of efficiency over all nodes of the graph, whereas local efficiency is the averaged measure of efficiency limited to the subgraph of the local neighbors of each node. Global efficiency provides an overall measure of the ability of a network to propagate information effectively, and local efficiency measures this in local subgraphs. Efficiency metrics were calculated using the built-in networkx functions 'global_efficiency' and 'local_efficiency'. Two-sample Mann-Whitney U tests were performed to test for differences in median efficiency scores between users and non-users.

Finally, meso-level properties were calculated to characterize connectivity motifs with high predictive importance in classifying chronic marijuana use. First, a group-average weighted correlation matrix was calculated by taking the mean of all un-thresholded subject-specific weighted connectivity matrices calculated above. This mean correlation matrix was then thresholded at 2% density and used to generate a group-average graph structure. Then, the Girvan-Newman hierarchical community-detection algorithm was used to detect clusters of high modularity within the graph. Briefly, the Girvan-Newman algorithm iterates as follows: (1) edge betweenness, defined as number of paths between all nodes that use a particular edge, is calculated for each edge; (2) edges with the highest betweenness are removed; (3) betweenness of all edges is recalculated. The final communities are defined as the node clusters that continue to share edges after high-betweenness edges are removed. The original weighted correlation matrix was then reorganized by discovered community structure. Each community (connectivity motif) was also ranked by its average degree centrality score, and most important motifs were defined as those with the highest average degree centrality score.

The predictive importance of the communities was corroborated using the following stepwise prediction approach. Starting with the highest ranked community, the correlations of all regions in the community to all

other regions (only non-redundant values) were used to generate distances to the hyperplane for each subject (i.e., by taking dot product of subset of weights and correlations and adding the intercept from the whole brain trained model). At each step, the decision threshold that maximized prediction accuracy in the training data was applied to the testing data to produce test set predictions. The best performing subset of communities was determined by the testing accuracy. To determine whether these accuracies were a function of the unique, included communities or just the number of pairwise functional correlations, a permutation approach was used. 1000 permutations were computed using the same approach as described above, except randomly shuffling the regions included in each community while preserving the number of pairwise correlations included at each step. The permutation p-value was calculated as the percentile of the best performing non-permuted accuracy in the distribution of the 1000 permuted accuracies at that same step.

DATA AND CODE AVAILABILITY

All code related to analyses in this study will be publicly released on GitHub at https://github.com/kulkarnik/craving-classifier. To request access to data, please contact the corresponding author.

AUTHOR CONTRIBUTIONS

K.K. and M.S. conceptualized and designed the predictive-explanatory modeling framework, carried out the implementation, and analyzed the data. G.P. provided feedback on modeling framework. V.C., F.F. and K.H. contributed to the interpretation of the results. K.K. and M.S. wrote the manuscript with critical feedback from all authors. F.F. and K.H. collected and organized the data. X.G. and D.S. supervised the project.

ACKNOWLEDGEMENTS

The authors acknowledge support by the US National Institutes on Drug Abuse under awards R01 DA043695 and R21 DA0492243. The authors also acknowledge the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse.

REFERENCES

1. Huys, Q. J. M., Maia, T. V. & Frank, M. J. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nat. Neurosci.* **19**, 404–413 (2016).
2. 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).
3. Poldrack, R. A. Inferring Mental States from Neuroimaging Data: From Reverse Inference to Large-Scale Decoding. *Neuron* **72**, 692–697 (2011).
4. Ding, X., Yang, Y., Stein, E. A. & Ross, T. J. Combining Multiple Resting-State fMRI Features during Classification: Optimized Frameworks and Their Application to Nicotine Addiction. *Front. Hum. Neurosci.* **11**, (2017).
5. Mete, M. *et al.* Successful classification of cocaine dependence using brain imaging: a generalizable machine learning approach. *BMC Bioinformatics* **17**, 357 (2016).
6. Rish, I., Bashivan, P., Cecchi, G. A. & Goldstein, R. Z. Evaluating effects of methylphenidate on brain activity in cocaine addiction: a machine-learning approach. in *Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging* vol. 9788 97880O (International Society for Optics and Photonics, 2016).
7. Vergara, V. M., Mayer, A. R., Damaraju, E., Hutchison, K. & Calhoun, V. D. The effect of preprocessing pipelines in subject classification and detection of abnormal resting state functional network connectivity using group ICA. *NeuroImage* **145**, 365–376 (2017).
8. Acion, L. *et al.* Use of a machine learning framework to predict substance use disorder treatment success. *PLOS ONE* **12**, e0175383 (2017).
9. Afzali, M. H. *et al.* Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation. *Addiction* **114**, 662–671 (2019).
10. Liu, J., Weitzman, E. R. & Chunara, R. Assessing Behavior Stage Progression From Social Media Data. in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 1320–1333 (Association for Computing Machinery, 2017). doi:10.1145/2998181.2998336.
11. Dumortier, A., Beckjord, E., Shiffman, S. & Sejdić, E. Classifying smoking urges via machine learning. *Comput. Methods Programs Biomed.* **137**, 203–213 (2016).

12. Rho, M. J. *et al.* Predictors and patterns of problematic Internet game use using a decision tree model. *J. Behav. Addict.* **5**, 500–509 (2016).

13. Mak, K. K., Lee, K. & Park, C. Applications of machine learning in addiction studies: A systematic review. *Psychiatry Res.* **275**, 53–60 (2019).

14. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).

15. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, (2008).

16. Diedrichsen, J. & Kriegeskorte, N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Comput. Biol.* **13**, e1005508 (2017).

17. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).

18. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).

19. Haxby, J. V. *et al.* A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron* **72**, 404–416 (2011).

20. Baldassano, C. *et al.* Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron* **95**, 709-721.e5 (2017).

21. Chen, J. *et al.* Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* **20**, 115–125 (2017).

22. Chen, P.-H. (Cameron) *et al.* A Reduced-Dimension fMRI Shared Response Model. in *Advances in Neural Information Processing Systems 28* (eds. Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 460–468 (Curran Associates, Inc., 2015).

23. Arbabshirani, M. R., Plis, S., Sui, J. & Calhoun, V. D. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage* **145**, 137–165 (2017).

24. Wager, T. D. *et al.* An fMRI-Based Neurologic Signature of Physical Pain. *N. Engl. J. Med.* **368**, 1388–1397 (2013).

25. Rashid, B. & Calhoun, V. Towards a brain-based predictome of mental illness. *Hum. Brain Mapp.* **n/a**,.

26. 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).

27. Bzdok, D. & Ioannidis, J. P. A. Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends Neurosci.* **42**, 251–262 (2019).

28. Kohoutová, L. *et al.* Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nat. Protoc.* **15**, 1399–1435 (2020).

29. Paulus, M. P. Pragmatism Instead of Mechanism: A Call for Impactful Biological Psychiatry. *JAMA Psychiatry* **72**, 631–632 (2015).

30. Du, Y., Fu, Z. & Calhoun, V. D. Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. *Front. Neurosci.* **12**, (2018).

31. Bassett, D. S. & Sporns, O. Network neuroscience. *Nat. Neurosci.* **20**, 353–364 (2017).

32. Bertolero, M. A. & Bassett, D. S. On the Nature of Explanations Offered by Network Science: A Perspective From and for Practicing Neuroscientists. *Top. Cogn. Sci.* **n/a**,.

33. Betzel, R. F. & Bassett, D. S. Multi-scale brain networks. *NeuroImage* **160**, 73–83 (2017).

34. Gosak, M. *et al.* Network science of biological systems at different scales: A review. *Phys. Life Rev.* **24**, 118–135 (2018).

35. Filbey, F. M., Schacht, J. P., Myers, U. S., Chavez, R. S. & Hutchison, K. E. Marijuana craving in the brain. *Proc. Natl. Acad. Sci.* **106**, 13016–13021 (2009).

36. Filbey, F. M. *et al.* fMRI study of neural sensitization to hedonic stimuli in long-term, daily cannabis users. *Hum. Brain Mapp.* **37**, 3431–3443 (2016).

37. Hasin, D. S. *et al.* Prevalence and Correlates of DSM-5 Cannabis Use Disorder, 2012-2013: Findings from the National Epidemiologic Survey on Alcohol and Related Conditions–III. *Am. J. Psychiatry* **173**, 588–599 (2016).

38. Hasin, D. S. *et al.* Cannabis withdrawal in the United States: a general population study. *J. Clin. Psychiatry* **69**, 1354–1363 (2008).

39. Carliner, H., Brown, Q. L., Sarvet, A. L. & Hasin, D. S. Cannabis use, attitudes, and legal status in the U.S.: A review. *Prev. Med.* **104**, 13–23 (2017).

40. Zehra, A. *et al.* Cannabis Addiction and the Brain: a Review. *J. Neuroimmune Pharmacol.* **13**, 438–452 (2018).

41. Koob, G. F. & Volkow, N. D. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry* **3**, 760–773 (2016).

42. Lynskey, M. & Hall, W. The effects of adolescent cannabis use on educational attainment: a review. *Addiction* **95**, 1621–1630 (2000).

43. Compton, W. M., Gfroerer, J., Conway, K. P. & Finger, M. S. Unemployment and Substance Outcomes in the United States 2002-2010. *Drug Alcohol Depend.* **0**, 350–353 (2014).

44. Meier, M. H. *et al.* Persistent cannabis users show neuropsychological decline from childhood to midlife. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2657–E2664 (2012).

45. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**, 186–198 (2009).

46. Bullmore, E. T. & Bassett, D. S. Brain Graphs: Graphical Models of the Human Brain Connectome. *Annu. Rev. Clin. Psychol.* **7**, 113–140 (2011).

47. Langer, N., Pedroni, A. & Jäncke, L. The Problem of Thresholding in Small-World Network Analysis. *PLOS ONE* **8**, e53199 (2013).

48. Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52**, 1059–1069 (2010).

49. Pariyadath, V., Stein, E. A. & Ross, T. J. Machine learning classification of resting state functional connectivity predicts smoking status. *Front. Hum. Neurosci.* **8**, (2014).

50. Sakoglu, U. *et al.* Classification of cocaine-dependent participants with dynamic functional connectivity from functional magnetic resonance imaging data. *J. Neurosci. Res.* **97**, 790–803 (2019).

51. Cheng, H. *et al.* Resting state functional magnetic resonance imaging reveals distinct brain activity in heavy cannabis users – a multi-voxel pattern analysis. *J. Psychopharmacol. (Oxf.)* **28**, 1030–1040 (2014).

52. Yalachkov, Y., Kaiser, J. & Naumer, M. J. Sensory and motor aspects of addiction. *Behav. Brain Res.* **207**, 215–222 (2010).

53. Japee, S., Holiday, K., Satyshur, M. D., Mukai, I. & Ungerleider, L. G. A role of right middle frontal gyrus in reorienting of attention: a case study. *Front. Syst. Neurosci.* **9**, (2015).

54. Tao, R. *et al.* Cannabinoid receptor CNR1 expression and DNA methylation in human prefrontal cortex, hippocampus and caudate in brain development and schizophrenia. *Transl. Psychiatry* **10**, 1–13 (2020).

55. Goldstein, R. Z. *et al.* Anterior cingulate cortex hypoactivations to an emotionally salient task in cocaine addiction. *Proc. Natl. Acad. Sci.* **106**, 9453–9458 (2009).

56. Kober, H. *et al.* Prefrontal–striatal pathway underlies cognitive regulation of craving. *Proc. Natl. Acad. Sci.* **107**, 14811–14816 (2010).

57. Gruber, S. A., Rogowska, J. & Yurgelun-Todd, D. A. Altered affective response in marijuana smokers: An FMRI study. *Drug Alcohol Depend.* **105**, 139–153 (2009).

58. Schweinsburg, A. D. *et al.* Abstinent adolescent marijuana users show altered fMRI response during spatial working memory. *Psychiatry Res. Neuroimaging* **163**, 40–51 (2008).

59. DeWitt, S. J., Ketcherside, A., McQueeny, T. M., Dunlop, J. P. & Filbey, F. M. The hyper-sentient addict: an exteroception model of addiction. *Am. J. Drug Alcohol Abuse* **41**, 374–381 (2015).

60. Moorman, D. E., James, M. H., McGlinchey, E. M. & Aston-Jones, G. Differential roles of medial prefrontal subregions in the regulation of drug seeking. *Brain Res.* **1628**, 130–146 (2015).

61. Du, Y. *et al.* NeuroMark: An automated and adaptive ICA based pipeline to identify reproducible fMRI markers of brain disorders. *NeuroImage Clin.* **28**, 102375 (2020).

62. Laumann, T. O. *et al.* Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron* **87**, 657–70 (2015).

63. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

64. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).

65. Gorgolewski, K. *et al.* Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Front. Neuroinformatics* **5**, (2011).

66. Tustison, N. J. *et al.* N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).

67. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction. *NeuroImage* **9**, 179–194 (1999).

68. Klein, A. *et al.* Mindboggling morphometry of human brains. *PLOS Comput. Biol.* **13**, e1005350 (2017).

69. Fonov, V., Evans, A., McKinstry, R., Almli, C. & Collins, D. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009).

70. Avants, B. B., Epstein, C. L., Grossman, M. & Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* **12**, 26–41 (2008).

71. Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* **20**, 45–57 (2001).

72. Cox, R. W. AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).

73. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* **17**, 825–841 (2002).

74. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).

75. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).

76. Power, J. D. *et al.* Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).

77. Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V. & Greicius, M. D. Decoding Subject-Driven Cognitive States with Whole-Brain Connectivity Patterns. *Cereb. Cortex* **22**, 158–165 (2012).

78. Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Front. Neuroinformatics* **8**, (2014).

79. Pedregosa, F. & Varoquaux, G. Scikit-learn: Machine Learning in Python. *J. Mach. …* **12**, 2825–2830 (2011).

80. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).

81. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. 5 (2008).

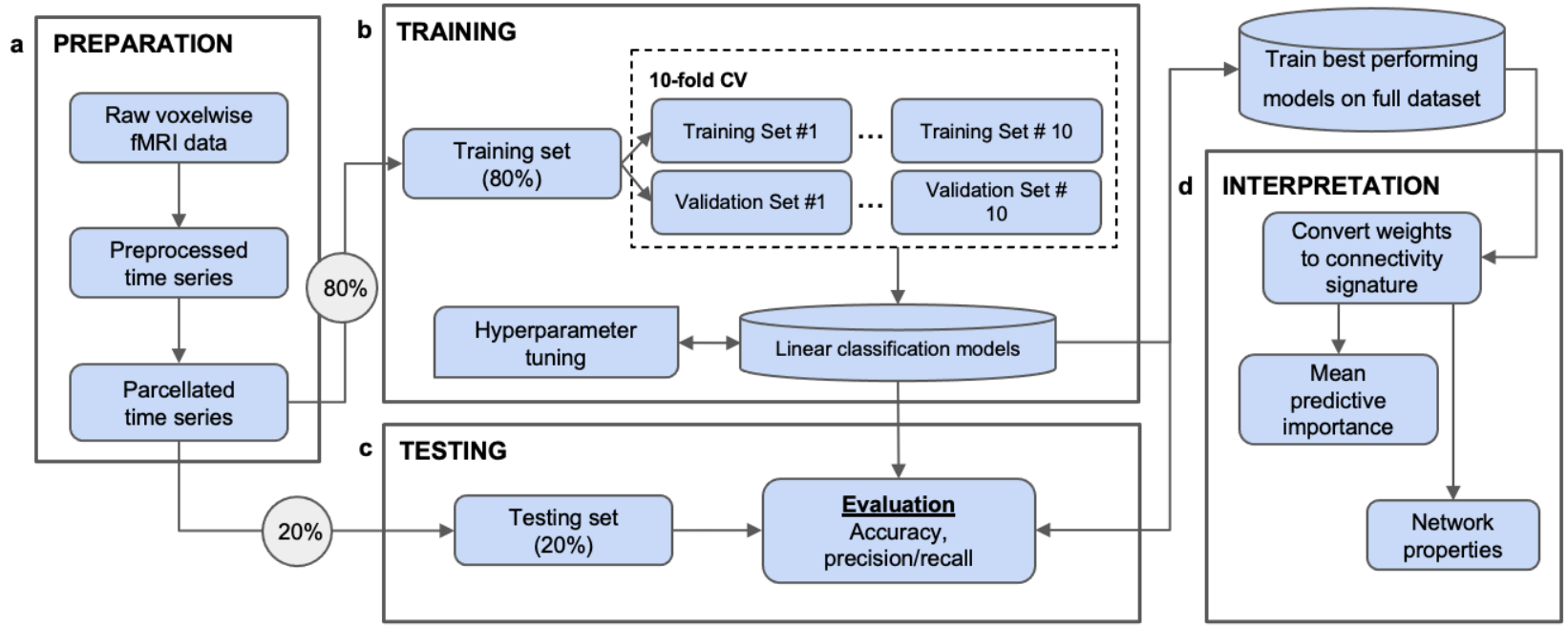

Fig 1. Machine learning pipeline. (a) Raw voxelwise time series are preprocessed using the fmriprep preprocessing pipeline. Minimally preprocessed files are brain-masked and smoothed with a 4mm FWHM Gaussian kernel. Nuisance/task regression is performed (see Methods for list of regressors used). Clean voxelwise time series is parcellated into 90 functional ROIs using Stanford functional atlas. **(b)** Parcellated data are divided into 2 sets; the training set is used for training and cross-validation, the testing set is used to evaluate the optimized classification models (shown in the cylinders). The optimization set is further divided into 10 subsets for cross-validation. Four linear classification algorithms are selected for hyperparameter tuning (L1, L2 penalized logistic regression and linear support vector classification). An alpha hyperparameter, corresponding to regularization strength is selected cross-validated accuracy as a metric. **(c)** The optimized hyperparameter tuned model is re-trained with the full training dataset and evaluated using the testing dataset. Evaluation parameters include accuracy, and precision/recall scores. (d) The best performing model (shown in the cylinder) is then trained on the full dataset (training + testing) to prepare for interpretation analysis. The weights derived from the linear models are converted to a connectivity signature and used to characterize brain connectivity structures important for prediction of chronic cannabis use. This analysis includes a regional mean predictive importance metric, as well as network characterization of subject-specific connectivity matrices weighted by the model weights.

	LOGISTIC REGRESSION		LINEAR SUPPORT VECTOR CLASSIFICATION	
Regularization strength (alpha)	**L1 penalty**	**L2 penalty**	**L1 penalty**	**L2 penalty**
1e-10	0.680228	0.682170	0.689955	0.689955
1e-7	0.697704	0.670612	0.682207	0.682188
1e-4	0.682095	0.689881	0.693764	0.707356
0.1	0.604649	0.693801	0.604649	0.707319
1	0.604649	0.658925	0.604649	0.670650
10	0.519212	0.608532	0.519212	0.604649
100	0.519212	0.651251	0.519212	0.660960
1000	0.519212	0.519212	0.519212	0.519212

Table 1. Hyperparameter optimization. Algorithm performance is compared across two hyperparameter domains: penalty type, and regularization strength. Higher alpha values correspond to higher regularization. Results show that low regularization strength works most effectively across all penalty types. Generally, L1 and L2 penalties work equally well at low regularization and L2 outperforms L1 at high regularization. A regularization value of alpha=0.0001 was chosen for subsequent analyses. Both classification methods and penalty types were retained as well.

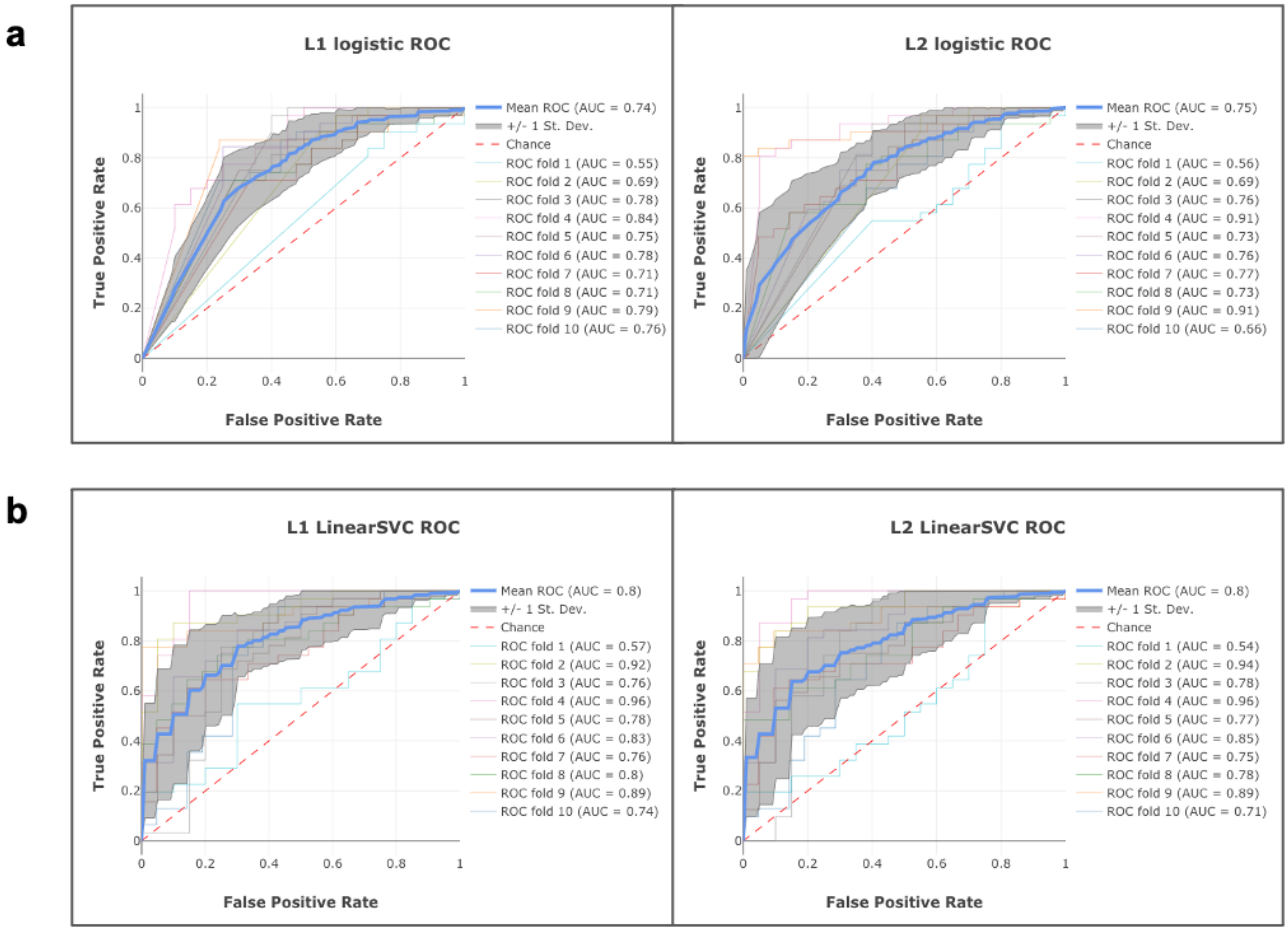

Fig 2. Performance metrics for linear classification algorithms. Performance was measured with 10-fold cross-validation of the training set (516 subjects). Performance metrics are summarized in Table 2. All four models performed well in cross-validation metrics with the mean receiver operating characteristic curve well above chance (red dotted line). **(a)** The logistic regression algorithm returns class probabilities which can be directly mapped to the ROC. **(b)** The linear support vector classification algorithm returns only a decision function, corresponding to the signed distances to the hyperplane. These distances are converted to probabilities using Platt's method.

	Modeling Algorithm	L1 Logistic	L2 Logistic	L1 SVC	L2 SVC
a	10-fold CV Accuracy	72.87%	73.06%	75.0%	74.61%
	10-fold CV AUC	0.74	0.75	0.80	0.80
	10-fold CV Precision	62.75%	65.69%	57.84%	58.33%
	10-fold CV Recall	66.67%	66.01%	73.29%	72.12%
b	OOS (Out-of-sample) Accuracy	82.31%	74.62%	80.00%	79.23%
	OOS AUC	0.83	0.82	0.87	0.86
	OOS Precision	71.15%	61.54%	78.85%	73.08%
	OOS Recall	82.22%	71.11%	73.21%	74.51%

Table 2. Cross-validation and out-of-sample performance metrics.
(a) The logistic regression models had relatively high 5-fold cross-validated performance, high AUC, and high precision-recall scores. Similarly, linear support vector classification (SVC) demonstrated high performance metrics for both penalty types. **(b)** Out-of-sample (OOS) performance metrics are summarized after re-training each model on the full training set (379 subjects). Note that 10-fold CV performance metrics are significantly lower than OOS accuracy, mainly due to sample size for training/testing. 10-fold CV divides the training set (516 samples) further into 10 folds, where only 9 of the 10 folds (464 samples) are used for training. For the OOS testing, the full training set is used to train. This explanation was tested post-hoc using leave-one-out cross validation in the training set, which yielded performance very similar to OOS testing.

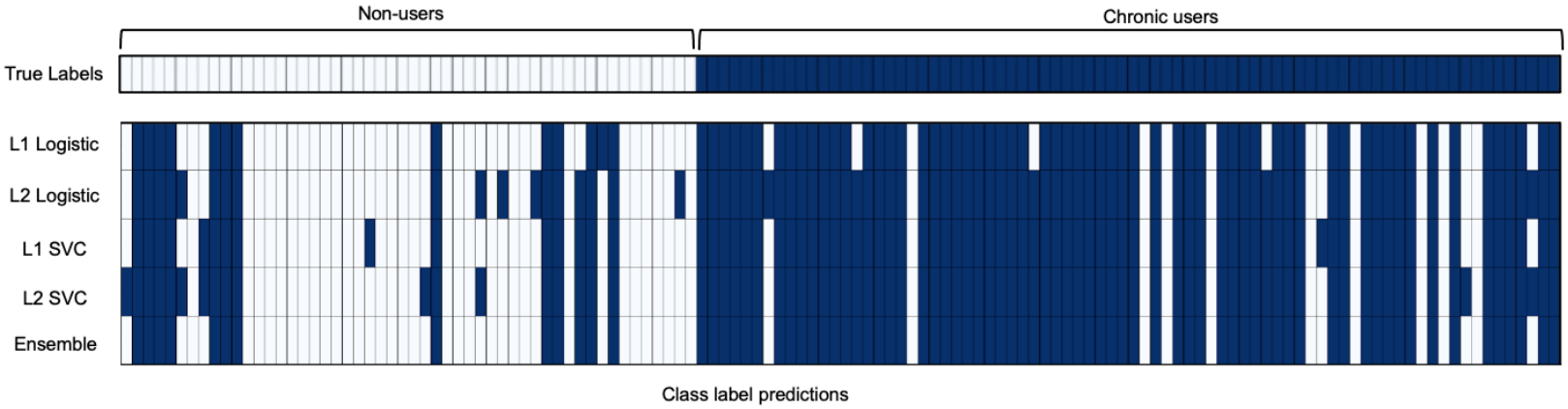

Fig. 3. Algorithm-specific out-of-sample predictions by class. The top row defines the true labels for each subject in the testing sample, sorted by label. Individual subjects are on the x-axis. The next 4 rows define the predicted labels for each subject from each model. The final row displays the ensemble prediction across all models, where the prediction is made by averaging across all other algorithm predictions and thresholding at 0.5. Performance metrics on the testing sample are summarized in Table 2.

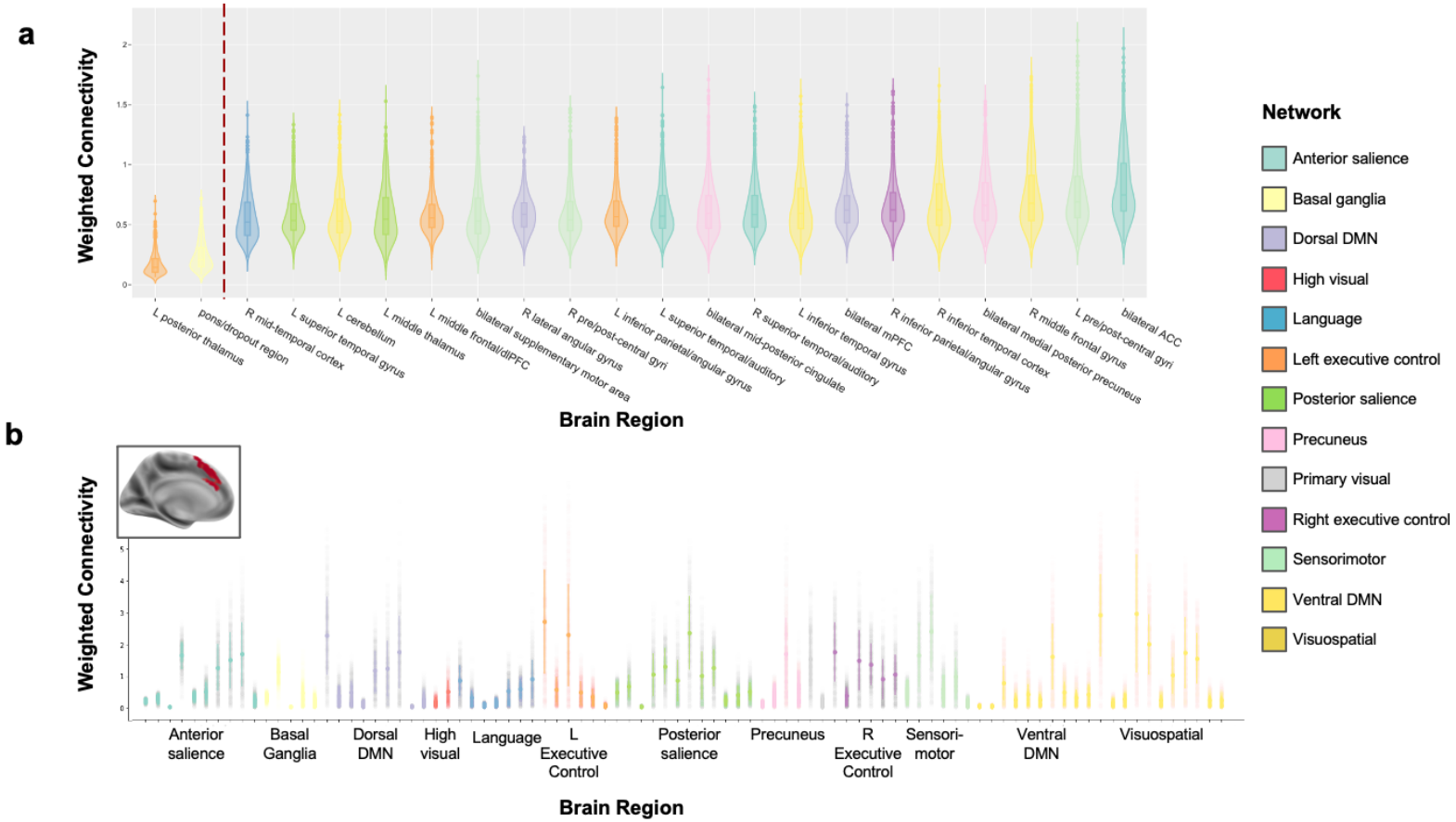

Fig. 4. Top weighted averaged parcel connectivities for L2 LinearSVC classifier. (a) Functional connectivity matrices were averaged across all subjects, and element-wise multiplication was performed with weights generated after model fitting with the L2 LinearSVC algorithm. The resulting matrix was a weighted region-to-region connectivity matrix. The mean of absolute weighted connectivity was calculated for each region for each algorithm. The distributions (mean, quartiles and outliers) of the absolute weighted connectivities across all subjects is shown above. The top twenty regions with highest means of weighted absolute connectivity are shown on the right side of the graph, while the two lowest are shown on the left for comparison. Regions with the highest weighted connectivity include bilateral ACC, left sensorimotor cortex, middle frontal gyrus and bilateral angular gyrus. **(b)** For the regions identified as having high weighted connectivities, region-specific connectivity patterns were assessed at a group level. Here, the connectivity strength and direction are shown from bilateral ACC, the region with the highest weighted connectivity across participants, to every other region. ACC appears to have have high connectivity specifically to inferior, middle, and superior frontal cortical areas across multiple functional networks (executive control, ventral default mode, visuospatial) as well as precuneus/angular gyrus regions. This suggest the presence of a ACC + frontal cortex + lateral parietal cortex task network, later supported by our community detection analysis (see Fig. 10).

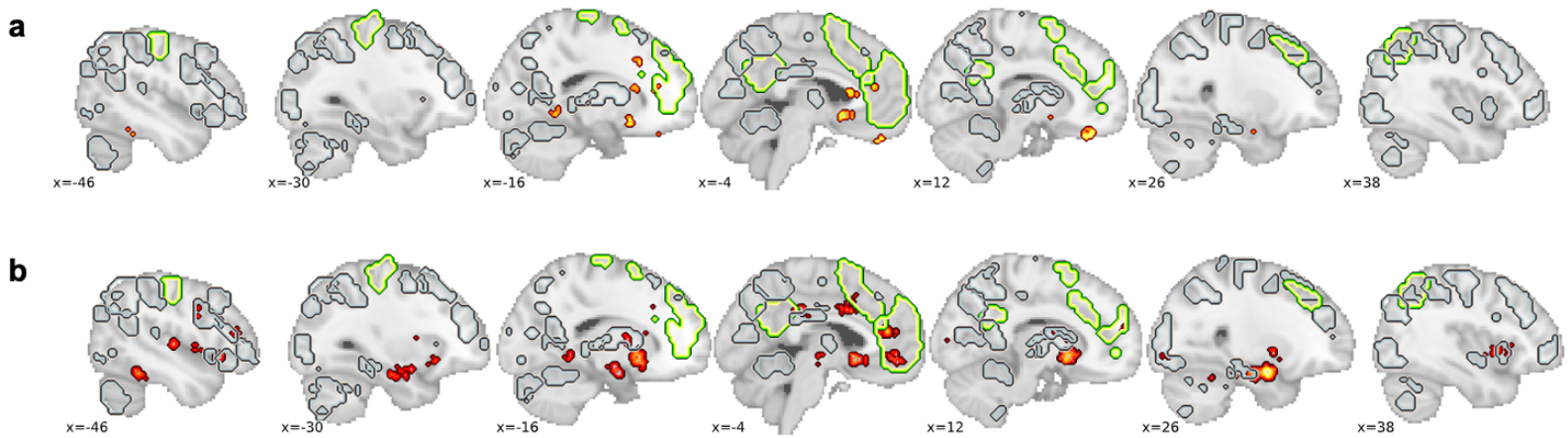

Fig 5. Meta-analytic comparison with Neurosynth craving maps. To compare our top regions of predictive importance to existing literature, we performed a direct comparison to association and uniformity maps retrieved from Neurosynth, a meta-analytic database. We used the 'craving' keyword to identify activations corresponding to all activations (uniformity) and unique activations (association) related to craving in the meta-analytic database. The average signal and proportion of voxels activated within-region was calculated and thresholded. Given the relative sparsity of the association map compared to the uniformity map, association was thresholded at 5% voxel participation and uniformity at 25% participation. These activations are shown in red, with (a) showing the association map and (b) showing the uniformity map. All the Stanford ROIs are overlaid on this map, with green regions corresponding to ROIs identified as having high predictive importance in our analysis. There is a moderate level of overlap between the craving maps and our predictively important regions, demonstrating the utility of our approach in identifying regions grounded in previous literature, but also being able to generate new hypotheses for regions involved in distinguishing cannabis users from non-users.

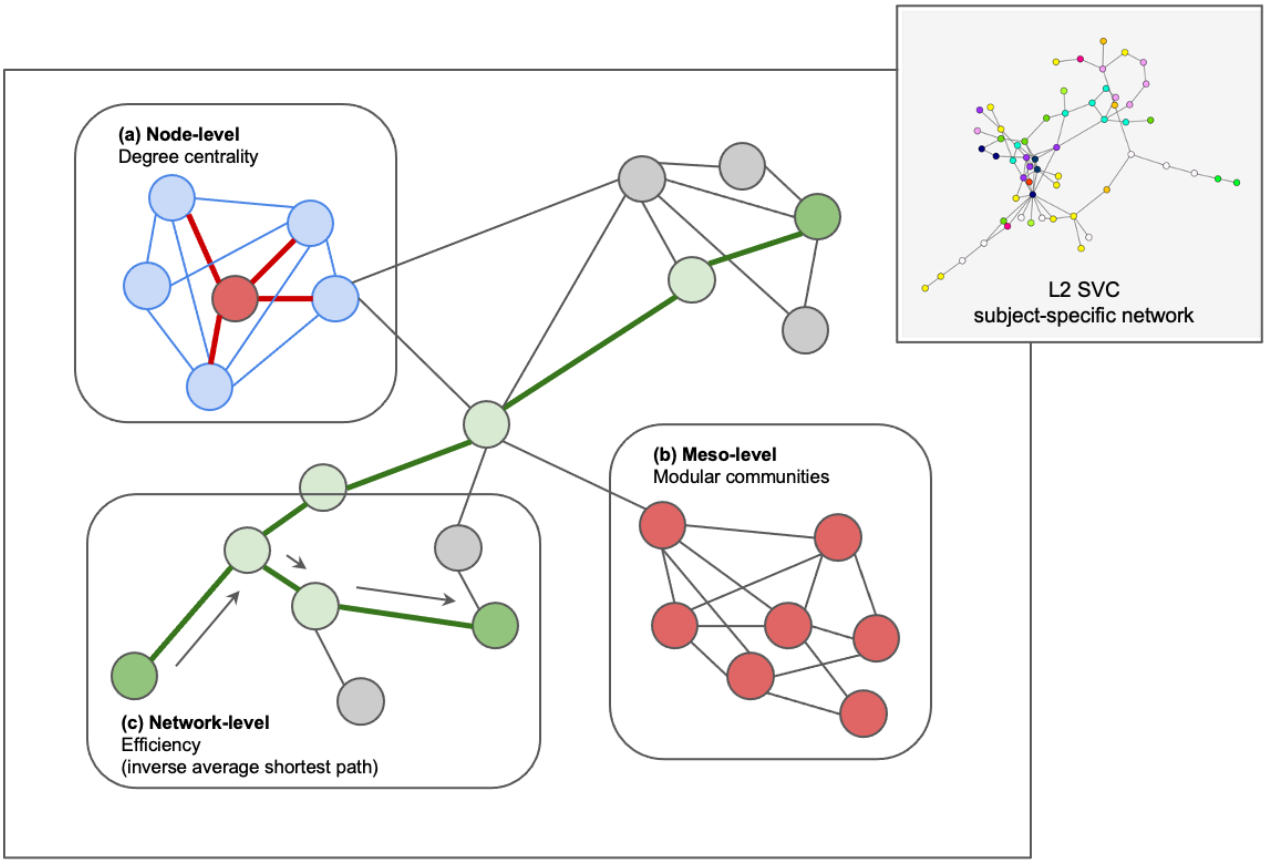

Fig 6. Network properties workflow. For each subject, a weighted connectivity matrix is generated by performing an element-wise multiplication of the original subject connectivity matrix and the model weights. The resulting matrices are thresholded to 2% sparsity to restrict to only highly informative connections. Network properties are then calculated at three different levels to characterize the subject-specific networks. (a) The degree centrality of each node of the network, i.e. a brain ROI, is obtained by calculating a normalized sum of surviving links to other nodes. In principle, this provides a measure of the importance of a region's connections to other regions for prediction. (b) At the meso-level, community detection algorithms are used to divide the full network into modular sub-networks that are highly connected to each other. These communities correspond to brain patterns that together are highly important for prediction of chronic cannabis use. (c) At the network-level, global efficiency of the network is calculated by determining the inverse average shortest path. For each node, the distance to every other node is calculated and averaged. The process is repeated for every node and averaged across nodes. The inverse of this averaged shortest path length is the efficiency of the network. High efficiency networks exchange information well because they are densely connected, and thus have fairly low average path lengths.

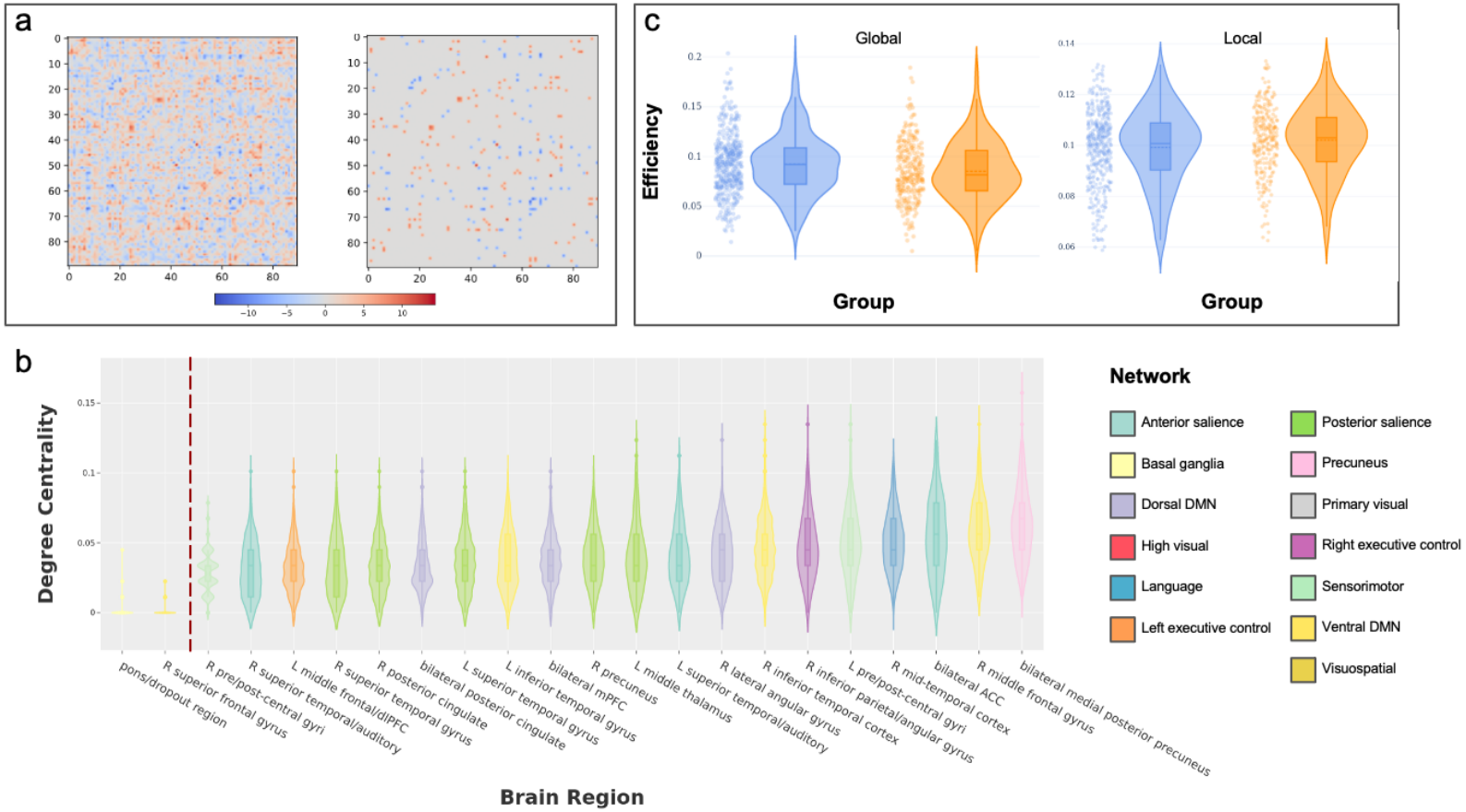

Fig. 7. Subject-level degree centrality and global-level efficiency. (a) Model weights followed the 4,005-element shape of the input vectors. These vectors are reverted to a symmetric 90x90 matrix and shown here. This matrix is thresholded at 2% of top weights to remove spurious connections. **(b)** Degree centrality represents the normalized number of weighted connections for each brain region that survive thresholding. In other words, it provides a measure of the level of distributed connectivity displayed by a brain region. In the plot above, degree centrality is calculated for each region independently, for each subject. The top twenty regions of highest mean degree centrality are shown, in addition to the lowest two for comparison. Regions identified as having high degree centrality across participants include middle frontal gyrus, bilateral ACC, and bilateral medial PFC. Note that there is a significant overlap here with regions identified as having highest absolute weighted connectivity (Fig. 5) but there are significant differences as well. **(c)** Global efficiency is defined as the average inverse path length between pairs of nodes across the full network. Local efficiency is the same but restricts paths to the local neighborhood of each node. For each participant, the global efficiency and local efficiency scores are calculated on the participant-specific weighted connectivity network. There is no significant difference in either global (p=0.0944) or local (p=0.0803) efficiency between chronic cannabis users and non-users, indicating that these network-level properties do not distinguish cannabis usage.

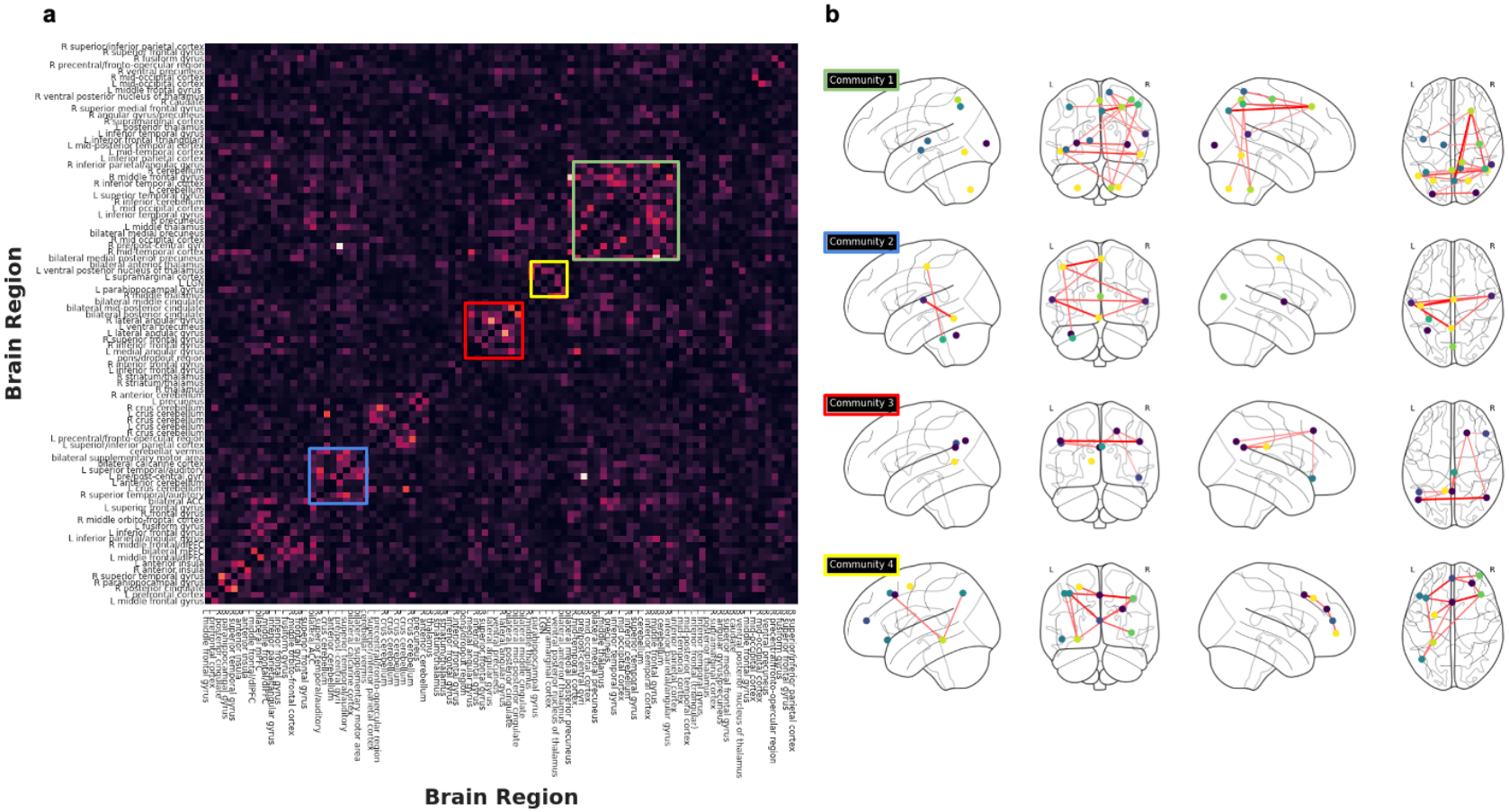

Fig. 8. Predictive motifs in group-average weighted connectivity matrix. The Girvan-Newman community detection algorithm was applied to the group-average weighted connectivity matrix. Girvan-Newman segregates communities within a group by iteratively removing edges with the highest betweenness centrality until a target modularity score is reached. Each disconnected set of nodes is then characterized as a community. **(a)** The group-averaged thresholded weighted connectivity is sorted by community assignment. Each colored square represents one of the top 4 communities by average degree centrality within community. **(b)** The color-corresponding communities are projected onto the brain and colored by resting-state network assignment as determined by the Stanford functional parcellation. The top 4 networks are largely bilateral. Community 1 is distributed mainly over posterior aspects of the brain and includes right pre/post-central gyrus, mid/superior temporal gyrus, precuneus, middle frontal gyrus and inferior parietal cortex. Community 2 is contains cerebellar regions, superior temporal cortex, and left pre/post-central gyrus. Community 3 includes inferior and superior frontal gyri, angular gyri, and bilateral posterior cingulate. Finally, Community 4 includes bilateral medial PFC, bilateral dorsolateral PFC, bilateral anterior cingulate and right orbitofrontal cortex.

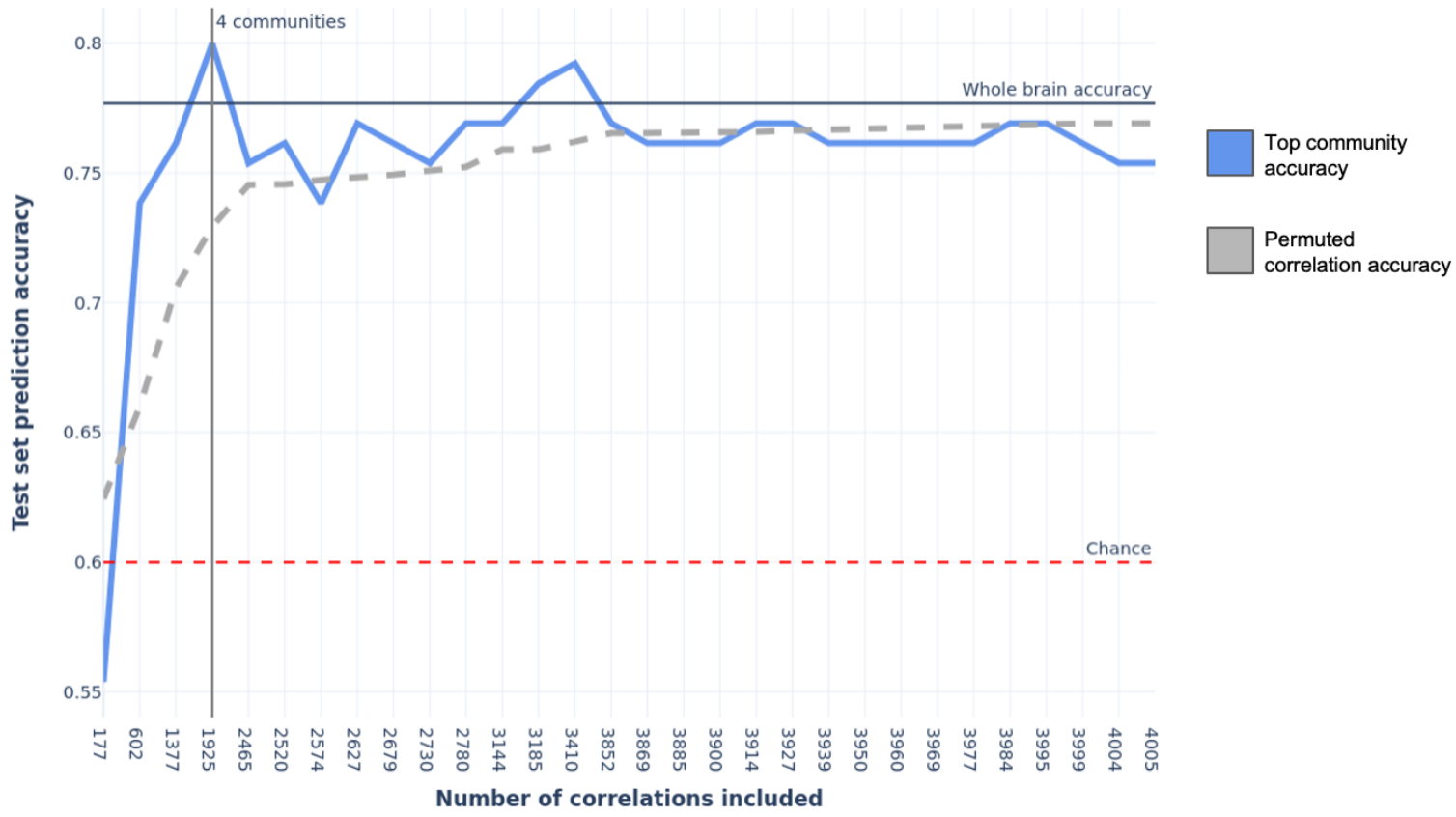

Fig. 9. Predictive accuracies of top communities. A stepwise prediction analysis was performed to confirm the predictive importance of the top-ranked communities discovered in the community detection analysis. Starting with the highest ranked community, the correlations of all regions in the community to all other regions were used to generate distances to the hyperplane for each subject. Then, a search was performed for the optimal decision threshold that maximized prediction accuracy in the training data. Finally, this threshold was applied to the testing data to produce test set predictions. The best performing subset of communities was determined by the testing accuracy. Permutation testing (1000 permutations) was performed to judge the relative increase in performance using the top communities vs. using a random set of correlations while preserving the number of pairwise correlations included at each step. The permutation p-value was calculated as the percentile of the best performing non-permuted accuracy in the distribution of the 1000 permuted accuracies at that same step. Chance was defined as a naive classifier that always picks the dominant class (chance=0.60). The best testing set prediction came from the first 4 communities with 80% accuracy, performing significantly better than random regions (permutation tested p=0.001) and above chance.