Quantifying Dysregulation of fMRI-Derived Control Circuits for Computational Psychiatry

Syed Sultan
State University of New York at Stony Brook
Steve Skiena
Stony Brook University
Lilianne Mujica-Parodi (mujica@lcneuro.org)
State University of New York at Stony Brook

Article

Keywords: brain, circuit, dysregulation, fMRI, control system, trajectory, computational psychiatry, generative model

Posted Date: November 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1413254/v2

Quantifying Dysregulation of fMRI-Derived Control Circuits for Computational Psychiatry

Syed Fahad Sultan[1,], **Steven Skiena**[2,], **and Lilianne R. Mujica-Parodi**[3,4,]✉

[1]Computer Science Department, Furman University, Greenville, SC 29613
[2]Computer Science Department, Stony Brook University, Stony Brook, NY 11794
[3]Biomedical Engineering Department, Stony Brook University, Stony Brook, NY 11794
[4]Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA 02129

Psychiatric disorders are thought to result from *dysregulated* brain circuits, yet human neuroimaging currently lacks standardized methods for quantifying neural dysregulation. Here, we present a scalable framework for extracting fMRI-derived (generative) control circuits, then use circuit trajectories to estimate their control error. Using synthetic circuits, we first demonstrate that our framework accurately identifies each circuit's architecture and models its dynamics by estimation of transfer functions. As a use case, we then apply the framework to human task-based functional MRI data (UK Biobank, N=19,831). In a purely data-driven manner, without priors, our framework identified thalamus-linked *prefrontal-limbic* and *ventral stream* subcircuits, selectively engaged during sensorimotor processing of affective and non-affective stimuli. Finally, we demonstrate that circuit-wide dysregulation, defined by degree of drift from healthy trajectories, tracks symptom severity for neuroticism (*ventral subcircuit*), depression (*prefrontal-limbic subcircuit*), and bipolar disorder (*full circuit*).

brain | circuit | dysregulation | fMRI | control system | trajectory | computational psychiatry| generative model

Correspondence: ***mujica@lcneuro.org***

Introduction

Psychiatric disorders are commonly understood to reflect *dysregulation* of one or more brain circuits. Yet, clinical neuroscience generally conflates the term *circuits* with co-activated brain regions, the latter of which are more accurately described as *networks*. Because neuroimaging-derived networks are normally defined by linear regressions ($y = b_0 + b_1x_1 + ...$), they are capable of reliably modeling only a very narrow range of topologies, in which one or more inputs leads to a single output (1). This limitation excludes the capacity for positive and negative feedback loops, as required for regulation.

To quantify brain circuit dysregulation we exploit the generative aspect of data-derived control circuits, which allows us to predict how a circuit's output time series will evolve over time. In a classic engineering control application, such as autopilot (Figure 1a), a vehicle corrects for deviations from its desired trajectory through negative feedback (e.g., *as the vehicle starts to drift to the right, the control circuit corrects the drift by steering to the left*). As such, the difference between the autopilot's actual versus desired trajectories provides a measure of its *control error*, or dysregulation (Figure 1b). Here, we use trajectory drift as a measure of control error. We calculate circuit-wide dysregulation across fMRI-derived control circuits, and demonstrate its clinical utility as applied to three psychiatric use cases: *neuroticism, depression,* and *bipolar disorder* (See Methods for definitions).

In developing this framework, we started from several desiderata: the ability to test homeostatic regulation in response to driving inputs (perturbation), a fundamental requirement of control theory(2–4); the ability to conduct whole-brain circuit discovery, free of priors; and the ability to scale, thereby leveraging the marked increase in both mega-scale neuroimaging datasets made possible through open-science initiatives, as well as high resolution, fine-granularity parcellations of the brain.

To date, the only standardized method capable of estimating fMRI-derived generative circuits is Dynamic Causal Modeling (DCM) (5). DCM is normally used to estimate circuit architecture, in the form of a directed, weighted graph. However, circuit architecture by itself is not sufficient to provide quantitative estimation of control parameters such as dysregulation. Moreover, DCM's computationally expensive algorithms, even for faster (resting-state only) variants such as spectral DCM (6), result in convergence times so lengthy that they remain impractical for extracting circuits from mega-scale datasets, using purely data-driven approaches (>100 brain regions), with short (e.g., 5 minute) time series. More recent (driving input-compatible) variants such as regression DCM (rDCM) (7) and sparse rDCM (8) replace the hemodynamic forward model with a fixed hemodynamic response function (HRF). As a result, they fail to allow for heterogeneity in the blood oxygenation level dependent (BOLD) signal across brain regions and individuals (9, 10). This biophysical constraint can lead to confounds, particularly when applied to neurodevelopmental, aging, and patient populations (11, 12).

Thus, we introduce a generalized framework based on state space systems that bridges the gap between network theory and control theory, with the scalability required to mine mega-scale datasets such as UK Biobank (N=19,831) (13). First, we confirm that using time-series to estimate systems of differential equations without biophysical constraints still permits recovery of complex causal relationships and feedback loops that characterize brain circuits. Using synthetic data, we generate canonical circuit motifs to simulate circuits

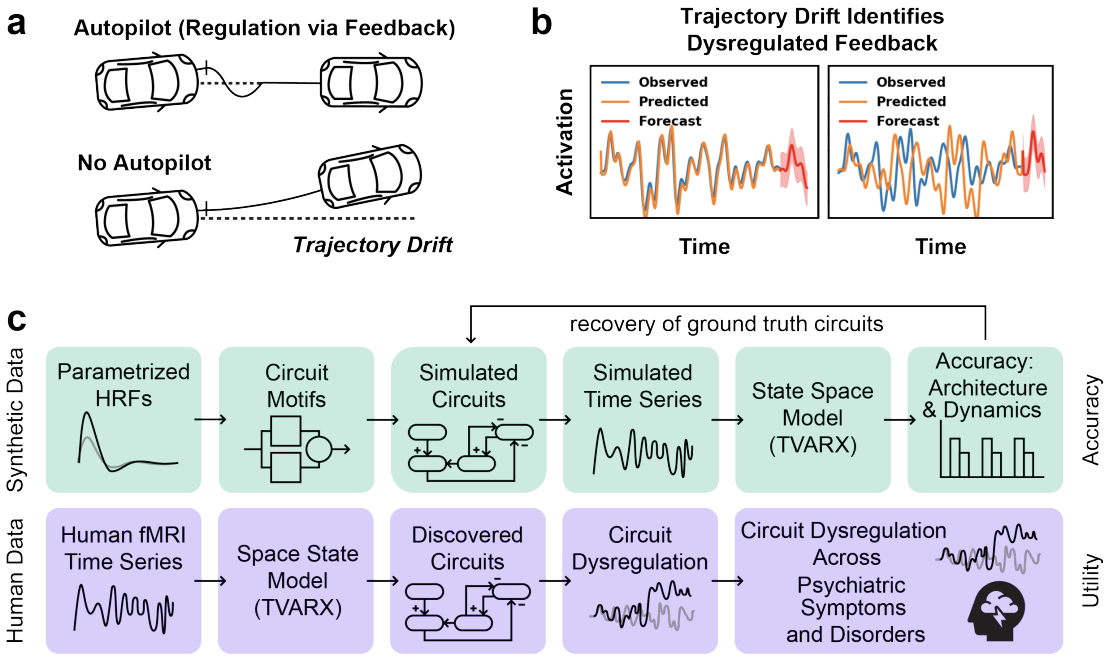

Fig. 1. Trajectory Drift as a Measure of Feedback Control Error, and Thus Circuit Dysregulation. (a) In a classic engineering control application, such as autopilot, a vehicle corrects for deviations from its desired trajectory through negative feedback. As such, the difference between the autopilot's actual versus desired trajectories provides a measure of its *control error*, or dysregulation. (b) As per the autopilot example, we use trajectory drift as control error to calculate dysregulation across fMRI-derived control circuits, and demonstrate its application for three psychiatric use cases: *neuroticism, depression,* and *bipolar disorder* (c) Schematic of the pipeline for the discovery of circuit architecture and dynamics from human fMRI and simulated time series, using Time-Varying Autoregressive Model with Exogenous Inputs (TVARX) and other state space models. We use trajectory drift between predicted and actual trajectories to quantify circuit dysregulation across subjects with varying degrees of severity for psychiatric symptoms.

with varying architectures and dynamics to test our framework's ability to recover both (Figure 1c top row). Second, having validated the framework on synthetic data, we then apply the framework to UK Biobank fMRI data. Using tasks designed to dissociate processing of affective versus non-affective stimuli (14, 15), we extract the control circuit selectively engaged by each. Third, from each individual's circuits we calculate the circuit's trajectory control error, which quantifies its degree of dysregulation. From these control errors, we statistically test the relationship between circuit-wide dysregulation and psychiatric symptoms (Figure 1c bottom row).

Results

Recovering circuit motifs from dynamic outputs

We first evaluate our framework using circuit motifs. Five thousand synthetic circuits are constructed by connecting nodes, each with its own transfer function, according to three basic motifs: *series*, *parallel* and *feedback*. These motifs are then combined in a modular fashion, to create larger circuits of varying levels of complexity (See Methods, Figure 4a).

The transfer functions used in our experiments were designed to resemble the hemodynamic response function (HRF) (16) extensively used to model blood oxygen level dependent (BOLD) (17) signals measured using functional magnetic resonance imaging (fMRI) (Figure 4b). The HRF function is parameterized by response height, time-to-peak and full-width at half-max. In our simulations, each node had different parameter values for the HRF, as previously shown for human data (9, 12).

The transfer function for each motif is an algebraic combination of node transfer functions (Figure 4c) Each motif also had an inverse variant. Serial and parallel connections each had both excitatory and inhibitory variants, while feedback loops had both positive and negative variants. The inverse variants are obtained by inverting the sign of their corresponding algebraic expressions (See Methods, Figure 4c). Note that although the node transfer functions are parameterized HRFs, successive connections and their corresponding algebraic expressions applied to HRFs can result in complex transfer functions.

We evaluate the ability to recover the canonical circuits both in terms of architecture and trajectory dynamics using a range of models with varying levels of complexity (Figure 2a). All of these models, with the exception of DCM, can be generalized by our state space system equations (See Methods, equation 3, Table 2). A detailed discussion regarding comparison with DCM is presented in a subsequent subsection.

The architecture is estimated through classifying relationship between each pair of nodes. For each node, parameters of its state space model are learned against time series of all other nodes (details in Methods). Based on the learned feed-forward matrix, a time-varying causality graph is established

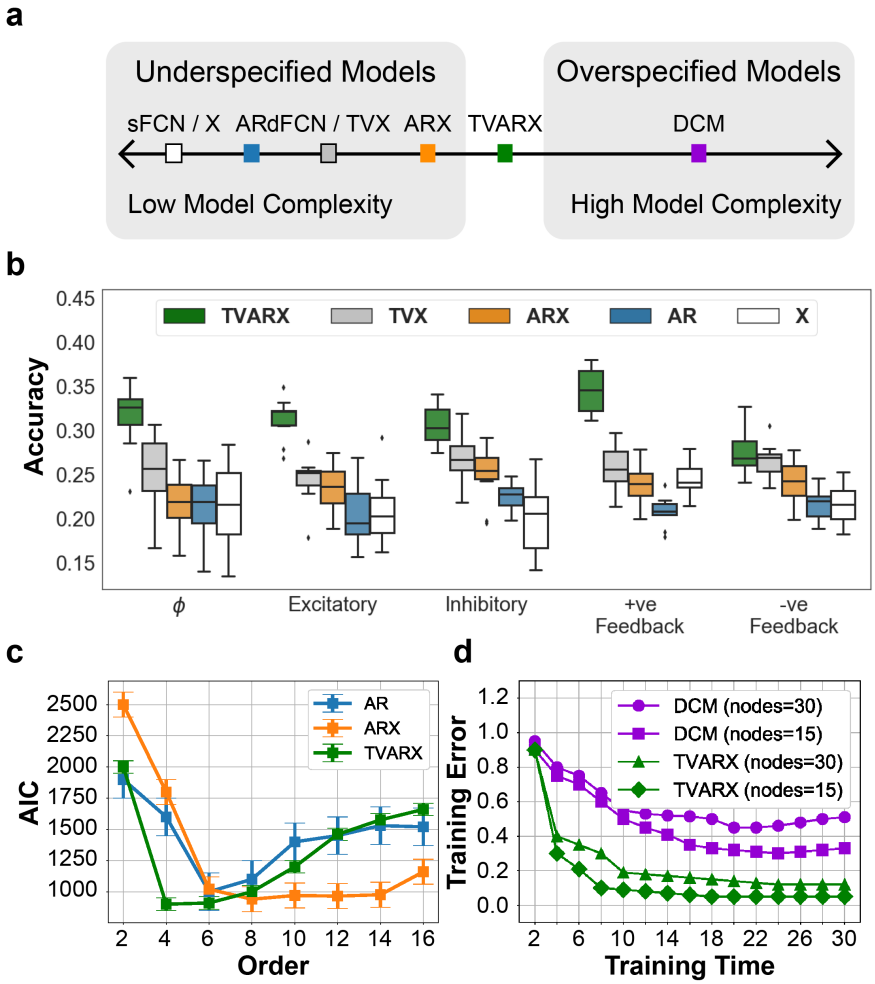

Fig. 2. Using Synthetic Data, we Compare the Performance (Accuracy and Speed) of System Identification Algorithms in Recovering Control Circuit Architecture and Dynamics (a) In this work, we fill the gap in literature between under-specified models that model networks, rather than circuits, and over-specified models that cannot scale computationally and therefore fail to extract circuits for large number of regions using shorter (∼5 minute) time series. (b) Accuracy scores for classification of 5000 simulated circuits and their components, including correct classification of no connection: ϕ) using different state space models discussed in Table 2. TVARX performed best at recovering the original circuit topology (c) AIC scores across models with respect to order of autoregressive component, which accounts for increasing complexity. TVARX performs best even when penalized for having the larger number of parameters. (d) Comparison of TVARX with (stochastic) Dynamic Causal Modeling (DCM) on human task-based fMRI. DCM fails to converge for shorter (∼5 minute) fMRI time series as well as for larger number of nodes. *sFCN=static Functional Connectivity Networks, X=eXogneous inputs, AR=AutoRegressive, dFCN=dynamic Functional Connectivity Networks, TVX=Time Varying with eXogenous inputs, ARX=AutoRegressive with eXogenous inputs, TVARX=Time Varying AutoRegressive with eXogenous inputs, DCM=Dynamic Causal Modeling.*

(Figure 4e) and Eulerian/elementary circuits are identified. The connections of these circuits are further classified into one of the four connection types.

Figure 2b provides the accuracy scores for each model with respect to classification of each connection type. The absence of DCM in Figure 2b reflects the fact that it failed to converge and thus did not yield meaningful results for our synthetic circuits. The Time-Varying autoRegressive with eXogenous inputs (TVARX) model outperforms other, simpler, models in identifying each connection type and thus recovering the overall circuit architecture.

To account for varying model complexities, in our evaluations for predicting trajectories we compute the Akaike Information Criterion (AIC) for different models as we increase the autoregressive order (Figure 2c). Only models that include past states were included in these comparisons, as majority of parameters are part of the autoregressive component and only these models are capable of generating future predicted non-linear trajectories. Here again, the TVARX model outperforms other models we evaluated in our experiments, even after accounting for the larger number of parameters.

Data-Driven Circuit Discovery Using Human fMRI

Participants from UK-Biobank (N=19,831) (13) were scanned while engaged in a task designed to elicit affective and non-affective sensorimotor processing. These were used to identify circuits selectively activated for each type of processing.

Task Design

During fMRI scans, participants were administered the Hariri faces/shapes "emotion" task (14, 15), as also implemented in the Human Connectome Project (HCP) (18) but with shorter overall duration and hence fewer total stimulus block repeats. Participants were presented with alternating blocks of trials with visual stimuli consisting of human faces and geometric shapes (circles, ellipses), with brief periods of rest in between. For facial expressions, 12 different images were used: six of each gender and affect (angry or fearful), all from a standardized set of pictures of facial affect (19). In each trial, either three faces or three shapes were presented in a triangular configuration: one centered above the other two. The participants were asked to indicate, by pressing a button, which stimulus on the bottom row matched the stimulus on the top row. The response triggered either the next trial or an eight second period of rest. The total length of the scan for each subject was 4 minutes; we obtained data for N=19,831 participants. The data were acquired on harmonized Siemens 3T Skyra scanners. The scans are 2.4mm isotropic with TR of 0.735s and 332 frames per run. Each subject had one run. The parcellation used in our experiments was provided by UK Biobank and included 139 regions of interest (ROIs). These ROIs are defined in MNI152 space, combining parcellations from several atlases: the Harvard-Oxford cortical and subcortical atlases (20, 21) and the Diedrichsen cerebellar atlas (22). Further information for the dataset is provided in Methods.

Comparison of TVARX with Dynamic Causal Modeling

In Figure 2d, we present a comparison between TVARX and DCM with respect to training time and corresponding training error. Even for small circuits (nodes ≤ 30), DCM fails to converge for time series of 4 minutes (332 timepoints) as available in UK-Biobank. As shown, not only does TVARX converge considerably faster, but training error for DCM does not decrease monotonically, indicating failure to converge.

This is not surprising, since DCM is designed to be hypothesis-driven, testing competing models from a prespecified set of nodes(6). Several competing hypotheses that constitute a model space are specified in the form of subgraphs, which are then compared using Bayesian model selection. Increasing the number of nodes is challenging because the number of extrinsic (between-node) connections or edges increases with the square of the number of nodes. This can lead to models with an enormous number of free parameters and profound conditional dependencies among the parameters. Furthermore, the computational time required to invert these models grows exponentially with the number of free parameters. More recent variants have been developed to successfully address this issue, but were not compared to TVARX because of other limitations: spectral DCM is not appropriate for measuring homeostatic regulation in response to driving inputs, and regression/sparse DCM (7, 8) constrains the HRF in ways that can introduce confounds in clinical populations (9–12).

Prefrontal-Limbic and Ventral Stream Subcircuits

In spite of the fact that no regions or connections were prespecified, our data-driven system identification methods were highly successful in accurately identifying linked subcircuits (Figure 3), each of which was consistent with independently validated experiments in the rodent, macaque, and human. Each subcircuit was uniquely specified with respect to its showing the largest absolute $\mathbf{D}$ value in response to its respective stimulus-type (See Methods 5).

As per the translationally-established *prefrontal-limbic subcircuit* (PFLC) for processing of affective stimuli, in which the thalamus provides a hub connecting a "low road" pathway to the amygdala and a "high road" pathway to the orbitofrontal cortex (OFC) and ventromedial prefrontal cortex (vmPFC)(23–26), our model recovered all key components and their relationships (Figure 3, *red*).

Likewise, for processing object form and recognition of non-affective stimuli, the *ventral stream subcircuit* has been shown to originate in the thalamus, project to V1-V2-V4, terminating in the inferior temporal gyrus (ITG), then progress to the inferior frontal gyrus (ITG, specifically the ventrolateral prefrontal cortex) and then to the orbitofrontal cortex/ventromedial prefrontal cortex (vmPFC)(27–29). Our model was able to recover nearly all of ventral stream (V1-V2-V4 were implicit in connecting the thalamus to the ITG), including "top-down" feedback (30) from the inferior frontal gyrus (IFG) to the inferior temporal gyrus (ITG) (Figure 3, *blue*).

In the context of psychiatry it is important to note that, while in human neuroimaging the ventral stream's (VS) input to the IFG has most been most often studied in the context of disambiguation of semantic meaning(31, 32), this same subcircuit also disambiguates perceptual meaning in the context of ambiguous threat in assessing risk ("threat generalization"). Indeed, it is dysregulation of the VS subcircuit–in response to *non-affective*, rather than affective stimuli–that we have previously shown (in four independent datasets, totaling N=226) to track the spectrum of trait to clinical anxiety (33), which most closely relates to the UK Biobank variable "neuroticism."

The two subcircuits were found to be mutually interacting with each other, with the pivot point centered at the thalamus, a known neuroanatomical hub shared by both circuits(29). Our system identification methods identified the thalamus to have two inputs providing negative feedback, one from hippocampus for the prefrontal-limbic subcircuit and one from inferior frontal gyrus for the ventral stream subcircuit; i.e.

$$Hippocampus \xrightarrow[PFLC]{-} Thalamus \xleftarrow[VS]{-} IFG^{1}$$

[1]Notation: Region A $\xrightarrow[circuit]{connectiontype}$ Region B. Excitatory connections are denoted by $+$ and inhibitory connections by $-$. Direction of the arrow head represents indicates directionality of causation.

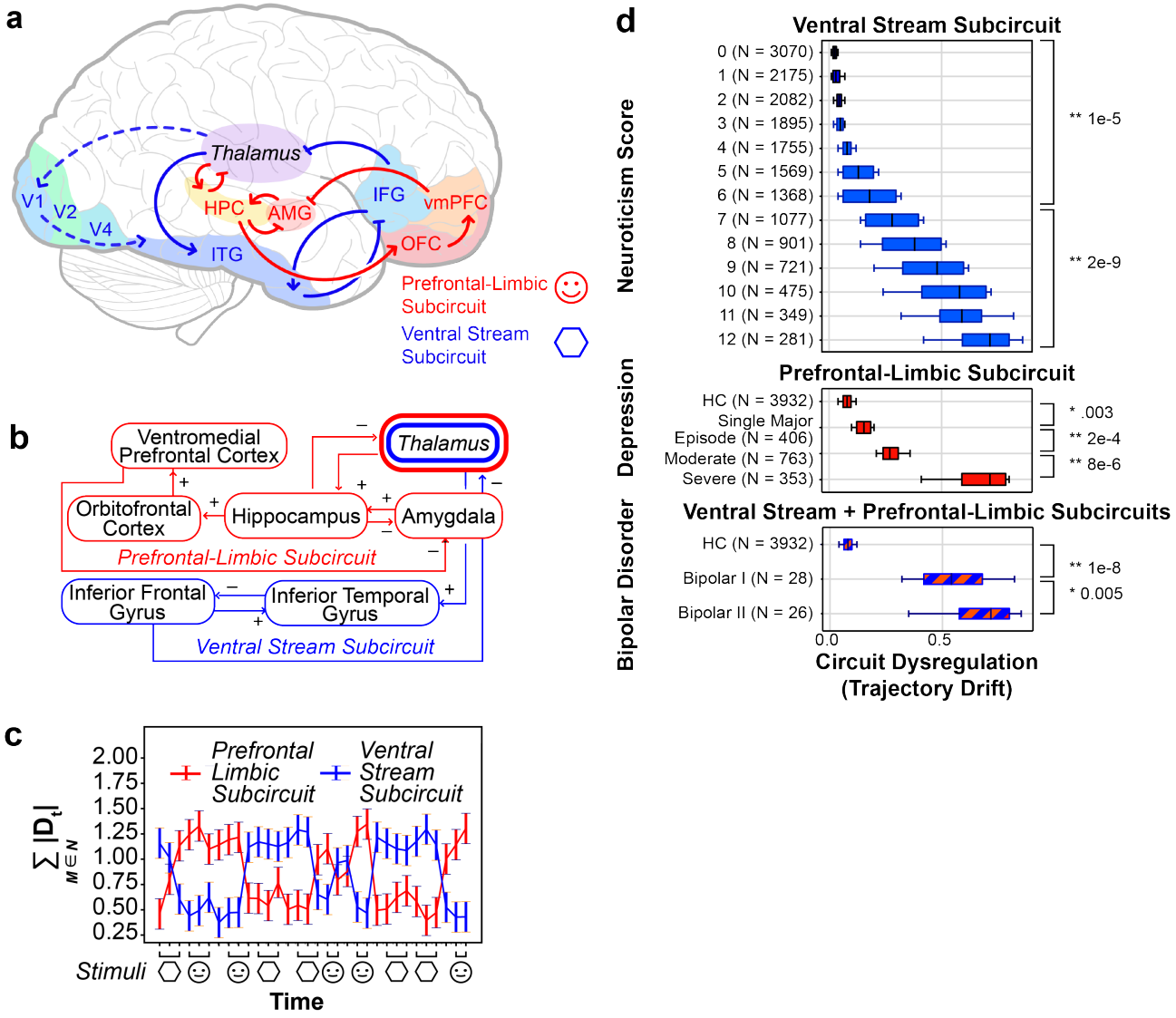

Fig. 3. Purely Data-Driven System Identification Accurately Recovers Known Subcircuits, Whose Degree of Dysregulation is Linked to Psychiatric Symptom Severity.(a-b) Without including any pre-specified information regarding regions or connections, our data-driven methods were highly successful in recovering two known subcircuits, interacting and linked via the thalamus, that have been independently validated in the rodent, macaque, and human (details in Discussion). (c) Results show selective dominance of each subcircuit with changing stimuli: the *prefrontal-limbic* subcircuit selectively engaged during processing of affective (faces) stimuli while the *ventral stream* subcircuit selectively engaged during processing of non-affective (shapes) stimuli (the y-axis is the relative dominance of one circuit versus the other, defined as absolute sum of relevant entries of the feedforward matrix D in our state space equation (Eqn. 3). For prefrontal-limbic subcircuit $M = \{thalamus, hippocampus, OFC, vmPFC\}$ and for ventral stream subcircuit $M = \{thalamus, ITF, IFG\}$ (d) We use trajectory drift (measured as the mean squared error between actual and predicted trajectories) as a measure of feedback control error, or dysregulation, of each subcircuit. Dysregulation of the prefrontal-limbic subcircuit was measured as the error in the thalamus trajectory as predicted from negative feedback by the hippocampus (hippocampus $\dashrightarrow$ thalamus). Dysregulation of the ventral stream was measured as error in the thalamus trajectory as predicted from negative feedback by the inferior frontal gyrus (inferior frontal gyrus $\dashrightarrow$ thalamus). Trajectory drift of the ventral stream subcircuit tracks severity of neuroticism, while trajectory drift of the prefrontal-limbic circuit tracks severity of depression. For bipolar disorder, the thalamic trajectory could not be predicted from either hippocampal or IFG trajectories, and thus reflects dysregulation of the full circuit. This could be due to either more systemic problems with feedback across both circuits, or that the full circuit is receiving dysregulated inputs from another, different, circuit not identified by these tasks. Bonferroni corrected $^{\star}P < 0.05$; $^{\star\star}P < 0.01$

Similarly, the thalamus has two outputs, both excitatory connections, one to hippocampus in the PFLC and the other to ITG in VS i.e.

$$Hippocampus \xleftarrow[PFLC]{+} Thalamus \xrightarrow[VS]{+} ITG$$

In this way, the thalamus *completes* one of two competing negative feedback loops, one for each of the two identified subcircuits.

We measure the relative dominance of one subcircuit versus the other at any point in time as the absolute sum of relevant entries in the time-varying feedforward matrix D_t in our state space equation (Methods, Equation 3). Note that each entry in $D_t^{(a,b)}$ represents the causal dependence of trajectory of b on trajectory of a (i.e. effective connectivity $a \rightarrow b$).

Our results show these two competing feedback loops to be alternatively dominating in strength based on the stimuli during the scan for each subject (Figure 3b). The prefrontal-limbic subcircuit was found to be the dominant loop at points in time when the subjects were tasked with matching facial stimuli of angry or fearful affect. The opposite was observed for ventral stream subcircuit, which was the dominant feedback loop when subjects were tasked with matching geometric shapes.

Trajectory Drift as Dysregulation

We use trajectory drift as measure of circuit-wide control error, and therefore dysregulation. This drift is measured as the mean squared error between the actual trajectories and the predicted trajectories. We further compare these varying levels of dysregulation with the severity and type of psychiatric symptoms and diagnoses. These include scored degrees of neuroticism (a measure of stress vulnerability, anxiety), depression, and diagnosis of Type 1 and 2 bipolar disorders (definitions in Methods).

To estimate dysregulation of the PFLC and VS subcircuits identified for N=19,831 subjects, we use the same task fMRI scans. However, unlike our identification of circuits in the previous section, here we measure dysregulation across the entire scan, independent of the design matrix.

Our results show marked association between greater dysregulation of specific subcircuits and the severity of the psychiatric symptoms (Figure 3c).

Since the thalamus was identified as a pivot point for switching between the two subcircuits, in determining which of the two competing feedback loops dominates the system, we specifically focused on regulation of the thalamus; i.e. prediction of thalamus's trajectory as a function of negative feedback by either the hippocampus (for PFLC) or the inferior frontal gyrus (for VS).

Our results show more severe neuroticism to be associated with greater trajectory drift (control error) in the ventral stream subcircuit (level 0 [N=3070] vs. 6 [N=1368] $^{**}p \leq$ 1e–5; level 6 [N=1368] vs. 12 [N=281] $^{**}p \leq$ 2e–9) (Figure 3c top), specific to weakened negative feedback from the inferior frontal gyrus (IFG) to the thalamus (IFG $\xrightarrow{-}$ Thalamus). This inhibitory connection is critical to stable regulation of the ventral stream and was observed to in turn result in greater dysregulation downstream with respect to thalamic outputs to the inferior temporal gyrus (ITG) (Thalamus $\xrightarrow{+}$ ITG).

In contrast, our results show more severe depression to be associated with greater trajectory drift (control error) in the prefrontal-limbic subcircuit (HC [N=3932] vs. Single Major Episode [N=406] $^{*}p \leq 0.003$; Single Major Episode [N=406] vs. Moderate [N=763] $^{**}p \leq$ 2e-4; Moderate [N=763] vs. Severe [N=353] $^{**}p \leq$ 8e-6) (Figure 3c center), specific to weakened negative feedback from the hippocampus to the thalamus (Hippocampus $\xrightarrow{-}$ Thalamus). Note that this relationship is itself dependent on the excitatory inputs from the thalamus to the hippocampus (Thalamus $\xrightarrow{+}$ Hippocampus) completing the negative feedback loop.

In the case of bipolar disorder, the thalamus was observed to be dysregulated with respect to both of its inhibitory inputs (Hippocampus $\xrightarrow[PFLC]{-}$ Thalamus $\xleftarrow[VS]{-}$ Inferior Temporal Gyrus) with greater dysregulation observed for subjects with Biopolar I Disorder compared to subjects with Bipolar II Disorder (Figure 3 bottom) (HC [N=3932] vs. Bipolar I [N=28] $^{*}p \leq$ 1e-8; Bipolar I [N=28] vs. Bipolar II [N=26] $^{*}p \leq 0.0005$). In the case of Bipolar Disorder I & II, the thalamic trajectory was observed to drift significantly from its predicted trajectory, but the system was not dominated by either of the two competing feedback loops. This could be due either to more systemic problems with feedback across both circuits, or that the full circuit is receiving dysregulated inputs from another, different, circuit not identified by these tasks.

Finally, we compare trajectory drifts of discovered circuits with more conventional methods currently prevalent in clinical neuroscience (Table 1). These standard methods include Stochastic DCM, correlation-based functional connectivity (34), and activation based Generalized Linear Models (GLM) (35). Note, however, that different methods answer fundamentally different (if complementary) questions: DCM and Trajectory Drift capture circuit-wide dynamics, functional connectivity provides the strength of (undirected) signaling across pairs of regions, and GLM provides activation of individual regions. To allow for a more direct comparison of our circuit-wide measure to activation and networks, in Table 1, we report results for the subcircuit regions and connections that we identified as tracking symptom severity. Our results show that modeling psychiatric disorders in terms of circuit dysregulation achieves markedly greater detection sensitivity across all three sets of psychiatric symptoms. Beyond identification of differences, however, the most important advantage of our method is that it uses data-driven methods to construct generative computational neuroscience models that explicitly consider homeostatic regulation across negative feedback loops. This has the potential to allow hypotheses regarding dysregulation across psychiatrically relevant circuits to be rigorously specified and empirically tested.

Discussion

In this work, we present a scalable fMRI data-driven technique that allows for construction of generative circuits in the human brain, and provides a quantitative measure–trajectory drift–of their control error, or circuit-wide dysregulation. We demonstrate the effectiveness of our technique in recovering artificially generated circuits of varied architectures and transfer functions. To demonstrate the applicability of the technique to computational psychiatry, we use large-scale fMRI data to identify two subcircuits and demonstrate that their dysregulation tracks with symptom severity with

		Neuroticism Score		Depression			Bipolar Disorder	
		0 (N=3070) vs. 6 (N=1368)	6 (N=1368) vs. 12 (N=281)	HC (N=3932) vs. Single Major Episode (N=406)	Single Major Episode (N=406) vs. Moderate (N=763)	Moderate (N=763) vs. Severe (N=353)	HC (N=3932) vs. Bipolar I (N=28)	Bipolar I (N=28) vs. Bipolar II (N=26)
DCM	*Failed to Converge*							
Trajectory Drift	Hippocampus $\overline{\longrightarrow}$ Thalamus			*0.003	**2e-4	**8e-6	**1e-8	*0.005
	Inferior Frontal Gyrus $\overline{\longrightarrow}$ Thalamus	**1e-5	**2e-9					
Functional Connectivity	Hippocampus $\leftrightarrow$ Thalamus			0.06	0.1	*0.001	**1e-4	0.3
	Inferior Frontal Gyrus $\leftrightarrow$ Thalamus	*0.002	**1e-4				0.12	0.26
GLM	Thalamus	0.08	*0.01	0.13	*0.012	0.16	*0.01	0.2

Table 1. Comparison of Trajectory Drift with fMRI Analytical Methods: Dynamic Causal Modeling (DCM), Functional Connectivity, and Activation Based Generalized Linear Models (GLM). We took a whole-brain purely data-driven approach in identifying circuits for the two circuit-based methods: DCM and Trajectory Drift. Of these, only Trajectory Drift was able to converge for the parcellation (139 regions of interest) and sample size (UK Biobank N=19,831). For the two subcircuits identified by Trajectory Drift: Prefrontal-Limbic and Ventral Stream, we then tested how the key regulatory components for Prefrontal-Limbic (negative feedback by the hippocampus) and Ventral Stream (negative feedback by the inferior frontal gyrus) were interpreted by non-circuit-based methods: GLM and Functional connectivity. For each comparison of psychiatric variables, we report p-values from statistical significance testing using Welch's t-test for unequal variances and sample sizes. Bonferroni corrected $^{\star}P < 0.05$; $^{\star\star}P < 0.01$. Cells that were not applicable are greyed out.

markedly greater detection sensitivity than standard analytic methods.

FMRI has conventionally been used to either compute brain activation maps, as areas of differential hemodynamic response, or to quantify pairwise connectivity between brain regions using Pearson correlation(36). More recent developments in fMRI analyses consider graph-theoretic measures (37) and a shift towards dynamic patterns of connectivity using time varying connections (38, 39). What all of these methods lack, however, is a conceptual and mathematical framework for considering the implications of closed feedback loops. Without these, activation maps and connectivity-derived networks can *suggest* the presence of neural circuits, but can neither define nor simulate their behavior, which includes their regulation. Given the assumption that psychiatric disorders are grounded in the failure of circuits to maintain homeostatic regulation, the ability to identify trajectories–including drift from normative trajectories–is thus an important step in the development of computational psychiatry and its characterization of dynamical disease(40, 41).

ACKNOWLEDGEMENTS

The research described in this article was funded by the Baszucki Brain Research Fund (L.R.M.-P.)

DATA AVAILABILITY

The neuroimaging and phenotypic data used in this work was obtained from UK-Biobank under Data Access Application 37462 and is available upon application at http://www.ukbiobank.ac.uk/register-apply/.

CODE AVAILABILITY

All source code is available at https://github.com/fahadsultan/QuantifyingDysregulation.

Bibliography

1. Lilianne Mujica-Parodi and Helmut Strey. Making sense of computational psychiatry. *The international journal of neuropsychopharmacology*, 23, 03 2020. doi: 10.1093/ijnp/pyaa013.
2. E. J. Hinch. *Perturbation Methods*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 1991. doi: 10.1017/CBO9781139172189.
3. Michael CK Khoo. *Physiological control systems: analysis, simulation, and estimation*. John Wiley & Sons, 2018.
4. Petar Kokotović, Hassan K. Khalil, and John O'Reilly. *Singular Perturbation Methods in Control: Analysis and Design*. Society for Industrial and Applied Mathematics, 1999. doi: 10.1137/1.9781611971118.
5. Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19 (4):1273–1302, 2003.
6. Karl J Friston, Joshua Kahan, Bharat Biswal, and Adeel Razi. A dcm for resting state fmri. *Neuroimage*, 94:396–407, 2014.
7. Stefan Frässle, Ekaterina I Lomakina, Lars Kasper, Zina M Manjaly, Alex Leff, Klaas P Pruessmann, Joachim M Buhmann, and Klaas E Stephan. A generative model of whole-brain effective connectivity. *Neuroimage*, 179:505–529, 2018.
8. Stefan Frässle, Ekaterina I Lomakina, Adeel Razi, Karl J Friston, Joachim M Buhmann, and Klaas E Stephan. Regression dcm for fmri. *Neuroimage*, 155:406–421, 2017.
9. Daniel A Handwerker, John M Ollinger, and Mark D'Esposito. Variation of bold hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *Neuroimage*, 21(4):1639–1651, 2004.
10. Geoffrey Karl Aguirre, Eric Zarahn, and Mark D'Esposito. The variability of human, bold hemodynamic responses. *Neuroimage*, 8(4):360–369, 1998.
11. Olivier David, Isabelle Guillemain, Sandrine Saillet, Sebastien Reyt, Colin Deransart, Christoph Segebarth, and Antoine Depaulis. Identifying neural drivers with functional mri: an electrophysiological validation. *PLoS biology*, 6(12):e315, 2008.
12. Pedro A Valdes-Sosa, Alard Roebroeck, Jean Daunizeau, and Karl Friston. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361, 2011.
13. Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.
14. Ahmad R Hariri, Alessandro Tessitore, Venkata S Mattay, Francesco Fera, and Daniel R Weinberger. The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323, 2002.
15. Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.
16. Richard B Buxton, Kâmil Uludağ, David J Dubowitz, and Thomas T Liu. Modeling the hemodynamic response to brain activation. *Neuroimage*, 23:S220–S233, 2004.
17. Seiji Ogawa, RS Menon, David W Tank, SG Kim, H Merkle, JM Ellermann, and K Ugurbil. Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging. a comparison of signal characteristics with a biophysical model. *Biophysical journal*, 64(3):803–812, 1993.
18. David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn human connectome project: an overview. *Neuroimage*, 80:62–79, 2013.
19. Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.
20. Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson,

Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.
21. Jean A Frazier, Sufen Chiu, Janis L Breeze, Nikos Makris, Nicholas Lange, David N Kennedy, Martha R Herbert, Eileen K Bent, Vamsi K Koneru, Megan E Dieterich, et al. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, 162(7):1256–1265, 2005.
22. Jörn Diedrichsen, Joshua H Balsters, Jonathan Flavell, Emma Cussans, and Narender Ramnani. A probabilistic mr atlas of the human cerebellum. *Neuroimage*, 46(1):39–46, 2009.
23. Elizabeth A. Phelps and Joseph E. LeDoux. Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48, 2005. ISSN 08966273. doi: 10.1016/j.neuron.2005.09.025.
24. M. Alexandra Kredlow, Robert J. Fenster, Emma S. Laurent, Kerry J. Ressler, and Elizabeth A. Phelps. Prefrontal cortex, amygdala, and threat processing: implications for ptsd. *Neuropsychopharmacology*, 47, 2022. ISSN 1740634X. doi: 10.1038/s41386-021-01155-7.
25. Joseph E. LeDoux. Emotion circuits in the brain. *Annual Review of Neuroscience*, 23(1): 155–184, 2000.
26. MR Bennett. The prefrontal–limbic network in depression: Modulation by hypothalamus, basal ganglia and midbrain. *Progress in neurobiology*, 93(4):468–487, 2011.
27. Dwight J. Kravitz, Kadharbatcha S. Saleem, Chris I. Baker, Leslie G. Ungerleider, and Mortimer Mishkin. The ventral visual pathway: An expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17, 2013. ISSN 13646613. doi: 10.1016/j.tics.2012.10.011.
28. Cornelius Weiller, Marco Reisert, Ivo Peto, Jºrgen Hennig, Nikos Makris, Michael Petrides, Michel Rijntjes, and Karl Egger. The ventral pathway of the human brain: A continuous association tract system. *NeuroImage*, 234, 2021. ISSN 10959572. doi: 10.1016/j.neuroimage.2021.117977.
29. Luiz Pessoa and Ralph Adolphs. Emotion processing and the amygdala: From a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, 11, 2010. ISSN 1471003X. doi: 10.1038/nrn2920.
30. Charles D. Gilbert and Mariano Sigman. Brain states: Top-down influences in sensory processing. *Neuron*, 54, 2007. ISSN 08966273. doi: 10.1016/j.neuron.2007.05.019.
31. Mirjana Bozic, Lorraine K. Tyler, David T. Ives, Billi Randall, and William D. Marslen-Wilson. Bihemispheric foundations for human speech comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 2010. ISSN 00278424. doi: 10.1073/pnas.1000531107.
32. Jennifer M. Rodd, Ingrid S. Johnsrude, and Matthew H. Davis. Dissociating frontotemporal contributions to semantic ambiguity resolution in spoken sentences. *Cerebral Cortex*, 22, 2012. ISSN 10473211. doi: 10.1093/cercor/bhr252.
33. Lilianne R Mujica-Parodi, Jiook Cha, and Jonathan Gao. From anxious to reckless: a control systems approach unifies prefrontal-limbic regulation across the spectrum of threat detection. *Frontiers in systems neuroscience*, 11:18, 2017.
34. Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78, 1994.
35. Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.
36. Jean-Baptiste Poline and Matthew Brett. The general linear model and fmri: does love last forever? *Neuroimage*, 62(2):871–880, 2012.
37. Danielle Smith Bassett and ED Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.
38. Daniel J Lurie, Daniel Kessler, Danielle S Bassett, Richard F Betzel, Michael Breakspear, Shella Kheilholz, Aaron Kucyi, Raphaël Liégeois, Martin A Lindquist, Anthony Randal McIntosh, et al. Questions and controversies in the study of time-varying functional connectivity in resting fmri. *Network Neuroscience*, 4(1):30–69, 2020.
39. Leonardo L Gollo and Michael Breakspear. The frustrated brain: from dynamics on motifs to communities and networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653):20130532, 2014.
40. Quentin JM Huys, Michael Browning, Martin P Paulus, and Michael J Frank. Advances in the computational understanding of mental illness. *Neuropsychopharmacology*, 46(1):3–19, 2021.
41. Daniel Durstewitz, Quentin JM Huys, and Georgia Koppe. Psychiatric illnesses as disorders of network dynamics. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 6 (9):865–876, 2021.
42. Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.
43. Steen Moeller, Essa Yacoub, Cheryl A Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic resonance in medicine*, 63(5):1144–1153, 2010.
44. Christian F Beckmann and Stephen M Smith. Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2): 137–152, 2004.
45. Gholamreza Salimi-Khorshidi, Gwenaëlle Douaud, Christian F Beckmann, Matthew F Glasser, Ludovica Griffanti, and Stephen M Smith. Automatic denoising of functional mri data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.
46. Mark W Woolrich, Brian D Ripley, Michael Brady, and Stephen M Smith. Temporal autocorrelation in univariate linear modeling of fmri data. *Neuroimage*, 14(6):1370–1386, 2001.
47. Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.
48. Stephen B Manuck, Sarah M Brown, Erika E Forbes, and Ahmad R Hariri. Temporal stability of individual differences in amygdala reactivity. *American Journal of Psychiatry*, 164(10): 1613–1614, 2007.
49. Daniel J Smith, Barbara I Nicholl, Breda Cullen, Daniel Martin, Zia Ul-Haq, Jonathan Evans, Jason MR Gill, Beverly Roberts, John Gallacher, Daniel Mackay, et al. Prevalence and characteristics of probable major depression and bipolar disorder within uk biobank: cross-sectional study of 172,751 participants. *PloS one*, 8(11):e75362, 2013.

Methods

Time Varying AutoRegressive eXogenous (TVARX) Model

Functional connectivity network for N regions-of-interest is traditionally defined as an $N \times N$ adjacency matrix A where

$$A_{x,y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}, A \in \mathbb{R}^{(\mathbb{N}\times\mathbb{N})}, x, y \leq N \tag{1}$$

When time series are normalized (zero mean and unit standard deviation), a common fMRI preprocessing step, the Pearson correlation coefficient is equal to the slope of the regression. Thus resulting in a linear regression model of the form

$$y_t = A_{x,y}x + b_t \tag{2}$$

Most dynamic variants simply extend this definition by adding an additional temporal dimension resulting in a time-varying adjacency matrix $A \in \mathbb{R}^{N\times N\times T}$ where T is either the length of the time series or the number of sliding time-windows.

In this work, we extend this simple prevalent linear model by modeling BOLD time series observed for a brain region using a state space model of the form:

$$\begin{aligned} y_t &= Z_t\alpha_t + D_t u_t + d_t + \epsilon \\ \alpha_{t+1} &= T_t\alpha_t + B_t u_t + c_t + R_t\eta_t \end{aligned} \tag{3}$$

where y_t refers to the observation vector at time t, u_t refers to the input (or control) vector from other regions of the brain, α_t refers to the (unobserved) state vector at time t, and where the irregular components are defined as $\epsilon_t \sim N(0, H_t)$ and $\eta_t \sim N(0, Q_t)$.

The remaining variables in the equations are matrices describing the process. The total length of the time series being T, the number of ROIs being N and K being the number of states, their variable names and dimensions are as follows: design $Z \in \mathbb{R}^{N\times K\times T}$, input $B \in \mathbb{R}^{N\times K\times T}$, observation intercept $d \in \mathbb{R}^{N\times T}$, observation covariance $H \in \mathbb{R}^{N\times N\times T}$, transition $T \in \mathbb{R}^{K\times K\times T}$, state intercept $c \in \mathbb{R}^{K\times T}$, selection $R \in \mathbb{R}^{K\times K\times T}$, state covariance $Q \in \mathbb{R}^{K\times K\times T}$

Note that this formulation is a generalized framework with prevailing definitions of static functional connectivity as correlations ($D_t = D_{t+1}\ \forall t, Z_t = O\ \forall t$) and dynamic functional connectivity as time-varying correlations($Z_t = O\ \forall t$) as special cases. Table 2 breaks down existing models and presents

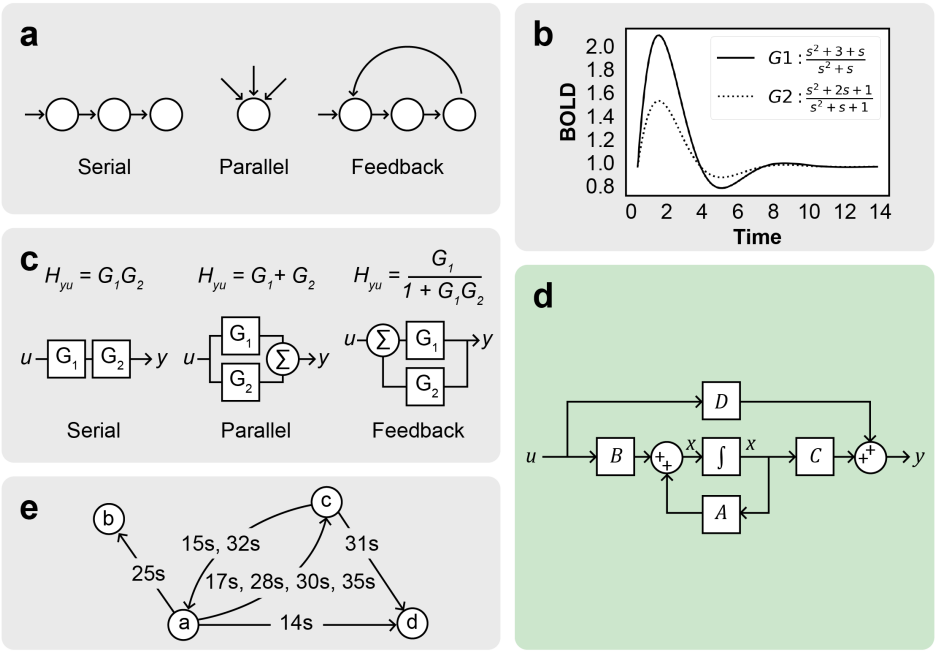

Fig. 4. Inferring Closed-Loop Circuits(a) Standard connectivity ("network") analyses depend upon linear regressions, which are only capable of modeling a very specific topology: parallel inputs. In contrast to parallel inputs *A*, most neurobiological circuits of relevance to psychiatry also require serial *B* and feedback *C* components, structures that could lead to an explosion of error propagation using standard statistical methods. (b) Impulse response for two hemodynamic response functions (HRFs) with different relaxation times and transfer functions. (c) To illustrate how transfer function structure changes with different circuit topologies, we show three transfer functions, each of which corresponds to a different kind of "motif," with series, parallel and feedback connections. By using pairs of inputs *u* and outputs *y* to obtain their transfer function, we systematically infer circuit topology. (d) Block diagram representation of state-space equations (e) Dynamic effective connectivity as a time-stamped temporal graph.

Model	D (feedforward)	T (transition)	B (input)	Z (design)
Correlations (sFCN / X)	Time invariant	O	O	O
Autoregressive (AR)	O	Time invariant	O	I
Time-varying Correlations (dFCN / TVX)	Time varying	O	O	O
Autoregressive w/ eXogenous inputs (ARX)	Time invariant	Time invariant	O	O
Time-varying Autoregressive w/ eXogenous inputs (TVARX)	Time varying	Time invariant	O	I

Table 2. Bridging the gap between network theory and control theory: extending existing correlation based models to our TVARX model. The breakdown and comparison in terms of state space parameters elucidates how our model is a generalized version of existing definitions of static (sFCN) and dynamic (dFCN) functional connectivity and extends networks to circuits. O: zero matrix, I: identity matrix

a comparison with our extended model (TVARX) in terms of parameters in our state space equations.

The time-varying feedthrough matrix D is used as dynamic effective connectivity between nodes. Note that effective connectivity defined this way is akin to a general form of Granger causality or transfer entropy (42).
U granger-causes $(\dot{Y})$ if

$$Y = D\dot{U} + T\dot{Y}$$

where $\dot{U}$ and $\dot{Y}$ represent lagged values of U and V.

$$y_t = \sum_{i=1}^{m} d_i u_{t-i} + \sum_{j=1}^{n} \tau_j y_{t-c} + c$$

null hypothesis: D = 0 (lagged values of U do not explain variance in Y)

This dynamic effective connectivity graph is a temporal graph as shown in 4e and can be formally defined as a set of time-stamped edges, each with its own connectivity strength $\{(a,b,t_1,D_{a,b}),(c,d,t_2,D_{c,d}),....,(x,y,T,D_{x,y})\}$.

Linear time invariant systems represented in state space form can be converted into input/output transfer functions by

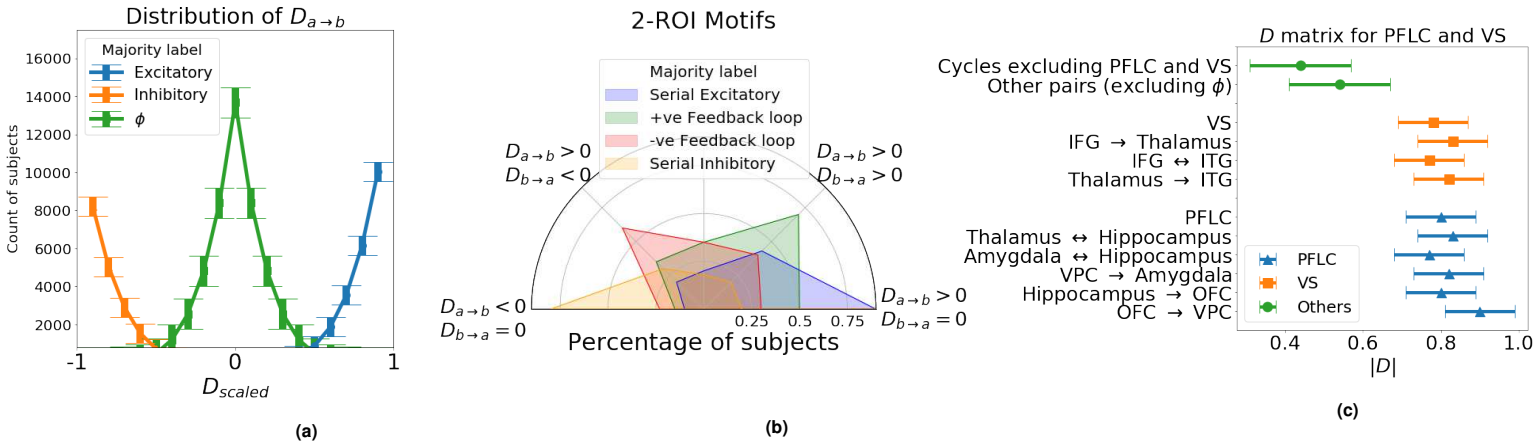

Fig. 5. Three steps of Circuit Discovery (a) As step 1, based on D in equation 3 after fitting on task fMRI, directed relationship between pairs of ROIs is labeled as either Excitatory, Inhibitory or as having no relationship (ϕ). Label for $a \to b$ is assigned through majority count between subjects. (b) In step 2, based on pairs of directed relationship labels $a \to b$ from step 1, an undirected label is assigned to each pair of regions, based on individual directed relationships in terms of corresponding values in the D matrix: $D_{a\to b}$ and $D_{b\to a}$. These labels are assigned also using majority count across subjects(c) In step 3, once labels of pairs of regions are determined in step 2, cycles are identified via depth-first traversal and the cycle with the largest D value is picked as the active circuit. In our experiments, PreFrontal Limbic Circuit (PFLC) and Ventral Stream (VS) were found to be two cycles with largest total D values for the two stimuli (faces and shapes) respectively.

applying Laplace transform

$$G(s) = \frac{\text{num}(s)}{\text{den}(s)} = \frac{a_0 s^m + a_1 s^{m-1} + \cdots + a_m}{b_0 s^n + b_1 s^{n-1} + \cdots + b_n} \quad \textbf{(4)}$$

where n is generally greater than or equal to m (for a proper transfer function).

State space systems can be manipulated using standard arithmetic operations as well as the feedback, parallel, and series. Vice versa, each of the connection types: feedback, parallel, and series represent arithmetic operations over state space systems and/or transfer functions as given in Figure 4 Panel c.

The parameters of the TVARX model are learned by maximizing loglikelihood via Kalman filter. The method for calculating the covariance matrix of parameter estimates uses outer product of gradient estimator using Broyden-Fletcher-Goldfarb-Shanno (BFGS) solver. The method by which the Hessian is numerically approximated is outer product of gradients.

The model is fit on seventy-five percent of the time series for each subject and dysregulation is measured as the error in prediction of the remaining twenty-five percent of the time series from the actual signal.

After fitting the TVARX model to data, trajectory drift is measured as error between actual trajectories and trajectories predicted by fitted model. This error is measured using mean squared error. Trajectory drift is used as a measure of feedback control error, or dysregulation, of each subcircuit.

The statistical significance testing between error distributions between different cohort of individuals is carried out using Welsch's t-test to account for skewed distributions between healthy and diseased populations.

Circuit Discovery

A circuit in our framework is defined as a set of connections such that there exists an elementary/Eulerian circuit (simple cycles) of length > 2. Each connection is defined for a pair of regions and of the following four types: excitatory, inhibitory, negative feedback loop and positive feedback loop defined over elements of the feedthrough matrix D. Excitatory and inhibitory connections between x and y are defined simply as effective connections where $D_{x,y} > 0$ and $D_{x,y} < 0$. Feedback loops are defined as Eulerian/elementary circuits of length = 2. Positive feedback loops are ones where both connections are excitatory. Inversely, negative feedback loops are ones with at least one inhibitory connection i.e.

$$\begin{aligned} &Serial: && Y = f(U), |D| > 0 \\ &&& Inhibitory: D < 0 \\ &&& Excitatory: D > 0 \\ &Parallel: && Y = f(U,V), |DU| > 0 \wedge |DV| > 0 \\ &Feedback: && Y = f(U) \text{ and } U = f(Y) \end{aligned}$$

where f(U) is represented by Eq. (3)

Parallel connections/inputs in the circuit are implicit as BOLD signal for a region y at time t is fitted against multiple inputs. All excitatory and inhibitory connections are series by default.

Discovery of circuits in our experiments is done in the following three steps:

1. Once TVARX model is fitted on human fMRI data, based on fitted values of D matrix in equation 3, directed relationships between pairs of ROIs is labeled

as either excitatory, inhibitory or as having no relationship (ϕ). The fitted value of $D_{a\to b}$ varied for different subjects. Since, we use a 139-ROI parcellation, the number of directed relationships equal 139^2. For each of these directed relationships $D_{a\to b}$, we have a count for each of $count_{D<0}$, $count_{D=0}$ and $count_{D>0}$, where $count_{D>0} + count_{D>0} + count_{D<0} = N(19,831)$. We assign a final label for $a \to b$ through majority count between subjects i.e. the label that was observed for most subjects was used for all subjects. In our results (Figure 5 panel a), we observed exponential curves, with a majority of labels being ϕ (no relationship).

2. Based on pairs of directed relationship labels $a \to b$ from step 1, an undirected label is assigned to each pair of ROIs based on individual directed relationships as determined by corresponding values in the D matrix: $D_{a\to b}$ and $D_{b\to a}$. Just as in step 1, different subjects have different $D_{a\to b}$ and $D_{b\to a}$ values. These conflicts are resolved by assigning a final label based on majority count across subjects (Figure 5 panel b).

3. Once labels of pairs of regions are determined in step 2, cycles are identified via depth-first traversal and the cycle with the largest D value is picked as the active circuit. In our experiments, PreFrontal Limbic Circuit (PFLC) and Ventral Stream (VS) subcircuit were found to be two cycles with largest absolute cumulative D values for the two stimuli (faces and shapes) respectively. The absolute D values for PFLC and VS were found to be significantly larger than for other circuits and relationships between regions not part of PFLC and VS (Figure 5 panel c).

Image Acquisition

Task fMRI data (tfMRI) were acquired on harmonized Siemens 3T Skyra scanners at four UK Biobank imaging centres (Cheadle, Manchester, Newcastle, and Reading). The scans were 2.4mm isotropic with TR of 0.735s and 332 frames per run (4 mins). Each subject had one run. The resolution of the images is 2.4x2.4x2.4 mm with a field-of-view of 88x88x64 matrix. The duration was four minutes (332 timepoints) with TR of 0.735 s and TE of 39ms, GE-EPI with x8 multislice acceleration, no iPAT, flip angle 52 degrees, and fat saturation.

A separate "single-band reference scan" was also acquired, as implemented in the Center for Magnetic Resonance Research (CMRR) multiband acquisition (43). This has the same geometry (including echo-planar imaging distortion) as the timeseries data, but has higher between-tissue contrast to noise, and is used as the reference scan in head motion correction and alignment to other modalities.

Data Preprocessing

Spatial smoothing, using a Gaussian kernel of FWHM 5 mm, was applied before the intensity normalisation, and neither Independent Component Analysis (ICA) (44) nor FMRIB's ICA-based X-noiseifier (FIX) (45) artefact removal was performed, both decisions being largely driven by the shorter timeseries in the tfMRI and because of the greater general reliance in tfMRI analysis on voxelwise timeseries modeling. All time series signal are standardized to z-scores (shifted to zero mean and scaled to unit variance) and the global signal is regressed out.

Pre-processing and task-induced activation modeling was carried out using FEAT (FMRI Expert Analysis Tool); timeseries statistical analysis was carried out using FMRIB's Improved Linear Model (FILM) with local autocorrelation correction (46). The timings of the blocks of the two task conditions (shapes and faces) are defined in 2 text files. Display of the task video and logging of participant responses is carried out by ePrime software. The timings of the task blocks are fixed and already known as well as the correctness of subject responses. For more details on data collection, processing of collected images and quality control, please see (47).

Regions of Interest

The parcellation used in our experiments was provided by UK Biobank and included 139 regions of interest (ROIs). These ROIs are defined in MNI152 space, combining parcellations from the following atlases: the Harvard-Oxford cortical and subcortical atlases (20, 21) and the Diedrichsen cerebellar atlas (22).

Task

This task was adapted from the one developed by Hariri and colleagues which had shown evidence as a functional localizer (14) with moderate reliability across time (48). Participants are presented with blocks of trials that either ask them to decide which of two faces presented on the bottom of the screen match the face at the top of the screen, or which of two shapes presented at the bottom of the screen match the shape at the top of the screen. The faces have either angry or fearful expressions. Trials are presented in blocks of 6 trials of the same task (face or shape), with the stimulus presented for 2 s and a 1 s inter trial interval. Each block is preceded by a 3 s task cue ("shape" or "face"), so that each block is 21 s including the cue. Each of the two runs include 3 face blocks and 3 shape blocks.

For facial expressions, 12 different images were used, 6 of each gender and affect (angry or afraid), all derived from a standard set of pictures of facial affect (19). Simple geometric shapes (circles, vertical, and horizontal ellipses) were used as control stimuli.

Subjects were asked to match one of two simultaneously presented images with an identical target image. As a sensorimotor control task, the subjects were asked to match ge-

ometric shapes. For each face block, three images of each gender and target affect (angry or fearful) were presented. For each control block, six different geometric shapes were presented as targets. During imaging, subjects responded by pressing one of two buttons with their dominant hand, allowing for the determination of accuracy and reaction time.

Clinical Variables

Neuroticism

Participants were assessed for twelve domains of neurotic behaviours via the touchscreen questionnaire. Neuroticism summarises the number of Yes answers across these twelve questions into a single integer score for each participant. Participants could answer Yes, No, Do not know or Prefer not to answer. Questions included:

1. Does your mood often go up and down?
2. Do you ever feel 'just miserable' for no reason?
3. Are you an irritable person?
4. Are your feelings easily hurt?
5. Do you often feel 'fed-up'?
6. Would you call yourself a nervous person?
7. Are you a worrier?
8. Would you call yourself tense or 'highly strung'?
9. Do you worry too long after an embarrassing experience?
10. Do you suffer from 'nerves'?
11. Do you often feel lonely?
12. Are you often troubled by feelings of guilt?

This derived data field has come from Professor Jill Pell from the Institute of Health and Wellbeing, University of Glasgow (49).

Depression

Depression status of participants is defined from the touchscreen questionnaire at baseline. Each of the three depression states were defined based on a number of criteria:

1. Ever felt depressed for a whole week
2. Ever disinterested or unenthusiastic for a whole week
3. Only 1 episode
4. $\geq$ 2 episodes
5. Episode lasted $\geq$ 2 weeks
6. Ever seen a GP for nerves, anxiety, tension or depression
7. Ever seen a psychiatrist for nerves, anxiety, tension or depression

Definitions Single Probable Major Depressive Episode: (1) AND (3) AND (5) AND [(6) OR (7)] OR (1) AND (3) AND (5) AND [(6) OR (7)]
Probable Recurrent Major Depression (Moderate): [(1) OR (2)] AND (4) AND (5) AND (6)
Probable Recurrent Major Depression (Severe): [(1) OR (2)] AND (4) AND (5) AND (7)

Bipolar Disorder

UKB data-fields from the touchscreen (which were based on the Structured Clinical Interview for DSM IV Axis I Disorders1) were classified into criteria groups to define a probable case of Bipolar I or II.
Bipolar I (probable mania) was classified as (1) ever manic or hyper for $\geq$2 days OR ever irritable or argumentative for $\geq$2 days AND (2) manic episodes characterised by at least 3 of 'more talkative', 'more active', 'needed less sleep', 'more creative/more ideas' AND (3) longest manic episode $\geq$ one week duration AND (4) episode needed treatment or caused problems at work.
Bipolar II (probable hypomania) classified as fulfilling criteria (1), (2) and (3) of the Bipolar I definition, NOT criteria (4).