nature portfolio

Peer Review File

Brain network dynamics reflect psychiatric illness status and transdiagnostic symptom profiles across health and disease

Corresponding Author: Dr Carrisa Cocuzza

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Cocuzza and colleagues present fMRI and clinical data from 110 individuals. Using fMRI, they integrate resting state, a cognitive control task, and an emotion processing task to look for changes in brain states. Using non-negative matrix factorization, they broadly find that people with clinical diagnoses display less variability across brain states. Follow up analyses largely support those claims with some surprising exceptions. This is an interesting study from a strong research group. I have several concerns with the paper which I detail below.

The final set of analyses (illustrated in Fig 5) raise more questions than answers. It is not clear why there would be such substantial differentiation across effect sizes for all four analyses (case/control, primary diagnosis, strict symptom fingerprints, fuzzy symptom fingerprints). The authors do highlight the consistency of rest & stroop being the 'best' for the symptom fingerprints, but the pattern as a whole is very inconsistent and suggests (to me at least), that much of what the authors are reporting here may be fitting noise.

Staying with this final set of findings, if the overarching argument of the introduction is that the ability of the brain to flexibly enter and exit myriad brain states, it is not clear why the effect size for all states (for the symptom fingerprints and primary diagnosis) would be by far the smallest. This is, in fact, evidence against the theoretical framing of the paper. No where in the paper do the authors truly grapple with this fact, other than to say in passing that, "network dynamics across wider-spanning task contexts may yield suboptimal decision boundaries." This doesn't really say very much other than just rehashing their finding.

Description of what is network efficiency and how to quantify it is lacking. Also, it is not clear why network efficiency would be used as a criteria to focus on the results (i.e., why lower efficient subgraphs would be presented in the supplement whereas subgraph 1 receives the primary focus in the main text).

In the introduction, authors state: "tend to outperform measures that assume stable, temporally-invariant, patterns of connectivity" but can the authors be more specific? In general / on average, how much do they outperform static connectivity metrics?

Because NMF requires positive inputs, how did the authors adjust the connectivity data to adhere to this requirement. That is not described in the methods or results. How stable are these results if the authors use an alternative dimensionality (SVD, PCA as described) reduction approach?

The magnitude of the difference in cosine similarity seems quite small, despite a very 'significant' p-value reported. Can the authors clarify how and over what was this test performed? Was it a linear mixed effects model with pairwise (state-to-state) cosine similarity as the outcome and diagnostic status as the predictor?

The clustering performed for the 'behavioral measures' is odd. In particular, the highest loading on the 'Externalizing Cluster' is the SHAPS anhedonia scale…I don't think any mental health researcher would consider anhedonia part of an externalizing dimension. Moreover, most of the other high loadings on that cluster are poorly accounted for by the construct of 'externalizing'. Seems like that should be re-considered.

Figure 2C is confusing. Why does it start with subject 64?

There is no Figure 5D, E, or F, though they are referenced in the manuscript.

Can the authors clarify this statement: “these states share a structural connectivity backbone”

The description surrounding Figure 2G is not clear. I believe that each cell represents a proportion of total for that patient population, but could the authors describe further?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors provide compelling evidence that brain network dynamics are predictive of transdiagnostic symptom profiles, outperforming case-control and diagnostic category approaches. The methods they employ are robust and well described. While the sample size is modest for the analyses performed, this work is novel, offers a strong proof of concept and addresses a timely issue. Clarification on several points would strengthen this already strong manuscript.
1. The authors compare mean cosine similarities between patients with and without psychiatric diagnoses. Given that cosine similarity values range from -1 to 1, a z-transformation or a non-parametric test may be more appropriate. Please clarify which statistical test was used and the rationale for this choice.
2. Classification was conducted using support vector classifiers (SVCs), which is reasonable. However, tabular data are often better suited to models such as random forests or gradient boosted trees. Please comment on why SVCs were chosen, and discuss the potential trade-offs, particularly regarding the ability of these models to detect important interactions between predictors.
3. The primary performance measure of model performance is limited to accuracy and “fuzzy accuracy”. Could the authors provide standard metrics including sensitivity, specificity, positive and negative predictive values were not evaluated?
4. Simulated lesioning is a reasonable approach for identifying key connectivity measures. However, methods like SHAP or permutation-based importance scores are also widely used. Please briefly discuss the comparative advantages and disadvantages of your approach for this context.
5. In Figure 4B, removing DMN component A seems to improve model performance, though not significantly. This result is unexpected; please comment on possible reasons and implications, even if the difference does not reach significance.
6. Imputation was applied to missing behavioral data, but the overall missingness rate is unclear. Was a threshold for missingness set (e.g., 10–30%) beyond which variables were dropped?
7. Figure 5A compares effect sizes among states for predicting dimensional symptom profiles. Without confidence intervals, however, it is unclear if differences are meaningful. If this is an exploratory analysis, please note this.
8. As MRI data was collected between Yale and McLean, could the authors discuss steps taken to ensure the data was harmonized? Was there a need for ComBat harmonization?
9. There does not appear to be a table outlining cohort characteristics. Relatedly, a summary of the proportion of diagnoses in the transdiagnostic sample would be useful.
10. Dimensional symptom profiles are defined within this modestly sized sample and not with respect to a normative sample; thus, generalizability to other datasets may be limited. Please discuss potential limitations stemming from this.
11. A big-picture consideration is the modest sample size used for the analyses. The study includes 219 participants (with diagnosis n=134 and without n=85) which is admirable for a neuroimaging study but may suggest a need for minor reframing. For example, the identification of symptom dimensions was done using hierarchical clustering of 110 behavioral measures. This is often done using exploratory factor analysis where a rule of thumb requirement is to have 5-10 observations per measure included. Here, however there are about 2 observations per measure which raises some concern about the robustness of estimating these dimensions. This does not detract from the novelty of this valuable paper; however, I suggest that much of it should be framed as exploratory or proof of concept with sample size constraints outlined as limitations.
Minor:
1. Please define the abbreviation TCP dataset at its first use.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed my concerns. I appreciate their attention to details in the revision and response.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Cocuzza et al have provided a comprehensive set of responses to reviewer critiques. Where appropriate they have outlined any deficiencies remaining and future studies to provide clarification. In my opinion the study is ready for publication.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Dear Peer Reviewers:

We are grateful for your close reading of our manuscript and are pleased by your enthusiasm and interest in our study and findings. Each reviewer has raised valuable points, and we have now revised the manuscript to include additional results, new framing- and theory-related text, background literature, and methodological details to improve the clarity of the study. Detailed, point-by-point responses to each comment are included below. Reviewer comments are shown in **bold,** with our responses shown directly beneath each reviewer comment. For ease of reference, quoted text from the manuscript is included when possible and *italicized* and indented, with all newly added manuscript text in blue font, both here and in the revised manuscript.

Thank you again for your time and expertise,
Carrisa V. Cocuzza, PhD, and Coauthors

Reviewer Comments

Reviewer #1 (Remarks to the Author):

Comment 1: Cocuzza and colleagues present fMRI and clinical data from 110 individuals. Using fMRI, they integrate resting state, a cognitive control task, and an emotion processing task to look for changes in brain states. Using non-negative matrix factorization, they broadly find that people with clinical diagnoses display less variability across brain states. Follow up analyses largely support those claims with some surprising exceptions. This is an interesting study from a strong research group. I have several concerns with the paper which I detail below.

Response 1: We thank the reviewer for their interest in our study and their expert feedback. We believe the clarity of the manuscript has notably improved after revising it to address the reviewer's concerns.

Comment 2: The final set of analyses (illustrated in Fig 5) raise more questions than answers. It is not clear why there would be such substantial differentiation across effect sizes for all four analyses (case/control, primary diagnosis, strict symptom fingerprints, fuzzy symptom fingerprints). The authors do highlight the consistency of rest & stroop being the 'best' for the symptom fingerprints, but the pattern as a whole is very inconsistent and suggests (to me at least), that much of what the authors are reporting here may be fitting noise.

Response 2: We thank the reviewer for their expert insights. We are likewise quite interested in the data in Fig. 5 (reproduced at the bottom of this response #2) and did not have strict *a priori* expectations about the pattern of results when comparing model performances for each outcome measure when predictors included varied subsets of across-state connectivity patterns. Thus, these analyses could be considered exploratory, which was brought up by reviewer #2 as well (see: reviewer #2, comment #7). We have now added text to the Results around Fig. 5 that clarify these exploratory model comparisons and further emphasized how the potential sensitivity of varied model groups to underlying shifts in cognitive context is a vital question to be expanded on in future studies in the Discussion.

Revised text now in the relevant **Results** section (lines 510-549):

> *"...Across all models, symptom fingerprints were classified with greater effect size than case/control status and primary diagnostic group (**Fig. 5A**). Further, models with the same outcome measure exhibited a reasonably similar range of effect sizes, suggesting internal consistency across similarly-posed predictive questions. For example, the effect size range for case/control models was 2.32 < D < 3.35, whereas the range for symptom fingerprint models was 17.25 < D < 18.85 (**Fig. 5A**). These ranges were further consistent with each set of comparison models' respective null model (as in **Fig. 4,** with each null model matching the varied predictors in **Fig. 5A**), which is a robust, data-driven way to infer statistical significance with complex, potentially noisy data[109] (see Discussion for further considerations). As in the full model (**Fig. 4**), all model accuracies were submitted to significance testing against empirical-chance*

*levels, which were based on nonparametric null models, with outcome measure labels permuted over 1000 iterations. Empirically-based chance levels were similar to the full model (**Fig. 4**), as follows: classifying case/control status: 49.77% chance; primary diagnosis group: 8.3% chance; symptom fingerprints: 6.1% chance. All variants upon network dynamics were able to classify all outcome measures significantly above empirical chance, but with varying effect size (**Fig. 5A, Table 2**). As expected, there was a broadly consistent pattern of model performances between strict-accuracy and fuzzy-accuracy (**Fig. 5A**) when classifying fingerprints. The all-states model exhibited the lowest performance (strict accuracy: D=17.25; fuzzy accuracy: D=54.73), suggesting that network dynamics across wider-spanning task contexts may yield suboptimal decision boundaries separating clinical symptom structures expressed across individuals. The rest-and-Stroop model performed the best (strict accuracy: D=18.85; fuzzy accuracy: D=62.91), suggesting that connectivity shifts between an intrinsic state and a cognitive control state accounted for individual clinical variability particularly well. Importantly, in symptom fingerprint models, classification accuracy did not simply increase with the amount of brain network data inputted to NMF, but instead increased when the Stroop task was present. This supports the proposal that brain network dynamics underlying shifts into more cognitively demanding task contexts are robust predictors of dimensional symptom structure across psychiatric illnesses. Interestingly, case/control and primary diagnosis were classified with a similar pattern across models. Case/control and primary diagnosis characterizations may involve a shared underlying relationship with symptomatology given that primary diagnosis can be thought of as an expanded version of the binary case/control status. However, the exact extent that each model fully accounts for noisy fluctuations in symptomatology is an important consideration when comparing these two patterns of results (see Discussion). With this consideration, the main difference was that all-states performed worst and best in classifying primary diagnosis and case/control statuses, respectively. These data suggest that separability of case/control status may be sensitive to the number of cognitive states (or, the amount of data) used to uncover network dynamics. Primary diagnoses may be classified best by dynamics across similar cognitive contexts. Conversely, dimensional symptom profiles may be separated best by a specific, foundational shift in cognitive context."*

"Table 2. Significance testing: classification model comparison when various state-to-state brain network reconfiguration dynamics were used as predictors.

Predictors: states	Outcome: case / control status	Outcome: primary diagnosis group	Outcome: fingerprint (strict)	Outcome: fingerprint (fuzzy)
All states	$t=10.8, p=5.8x10^{-14}$	$t=5.2, p=3.1x10^{-6}$	$t=7.0, p=7.4x10^{-9}$	$t=22.2, p=4.2x10^{-25}$
Rest only	$t=8.7, p=3.1x10^{-11}$	$t=5.4, p=1.6x10^{-6}$	$t=7.5, p=1.4x10^{-9}$	$t=23.6, p=3.5x10^{-26}$
Task only	$t=8.2, p=1.5x10^{-10}$	$t=5.3, p=1.9x10^{-6}$	$t=7.4, p=2.1x10^{-9}$	$t=24.4, p=1.1x10^{-26}$
Rest & Stroop	$t=8.1, p=1.8x10^{-10}$	$t=5.3, p=2.0x10^{-6}$	$t=7.6, p=9.3x10^{-10}$	$t=25.5, p=1.9x10^{-26}$
Rest & Emo.	$t=7.5, p=1.6x10^{-9}$	$t=5.3, p=2.2x10^{-6}$	$t=7.2, p=4.0x10^{-9}$	$t=23.0, p=9.5x10^{-27}$
Stroop only	$t=8.0, p=2.7x10^{-10}$	$t=5.2, p=2.5x10^{-6}$	$t=7.3, p=2.4x10^{-9}$	$t=23.9, p=2.4x10^{-26}$
Emo. only	$t=7.9, p=4.6x10^{-10}$	$t=5.2, p=2.4x10^{-6}$	$t=7.1, p=6.0x10^{-9}$	$t=22.6, p=2.1x10^{-25}$

"

Revised text now in the **Results** (lines 498-508):

*"...To test how influential each cognitive state was to the link between brain network dynamics and symptom fingerprints, we implemented an exploratory analysis where NMF was run with varied combinations of input connectivity patterns. This allowed us to differentiate context-specific shifts in connectivity from general rest-to-task shifts in connectivity (as in **Fig. 4**)."*

Revised text now in the **Discussion** (lines 669-688):

*"...To make non-stationarity inferences, we instead opted for a model comparison approach (**Fig. 5**), where connectivity patterns relevant to varied cognitive states were considered to compare reconfiguring network interactions across resting-state-alone, rest-to-Stroop, task-alone, and so forth. The present study advances clinical neuroscience by bypassing undue influence from non-neural sources of variance,*

however, we encourage future work that systematically compares how well other brain network dynamics methods can discriminate the structure of psychiatric symptoms.

Relatedly, while predictors (brain network dynamics) were held constant in all model comparison analyses, the outcome measures studied may exhibit differing levels of inherent noisiness. The strategies we used to account for the possibility that classification models were overly fit to such noise included: comparing effect sizes instead of raw accuracies, given differing chance levels; using common sets of predictors across all models; and a strict scheme of splitting data into discovery, test, and unseen validation sets commonly across all models. Importantly, for all models compared, we also used a non-parametric and data-driven approach to developing null models that has shown to be rigorous for fMRI-based[109], brain connectivity data[21] where the statistical ground truth is not known. While this multi-pronged approach is robust, it is important to note that we cannot precisely quantify how much noise influenced model fit across the differing models. It will be an important methods advancement for future studies to use simulated data – where ground truth properties are known and varied by the researcher – to precisely establish signal and noise influences upon model fit."

We have additionally added this text to the **Discussion,** which is relevant to this comment as well as reviewer #2, comments #10 and #11 below (lines 646-655):

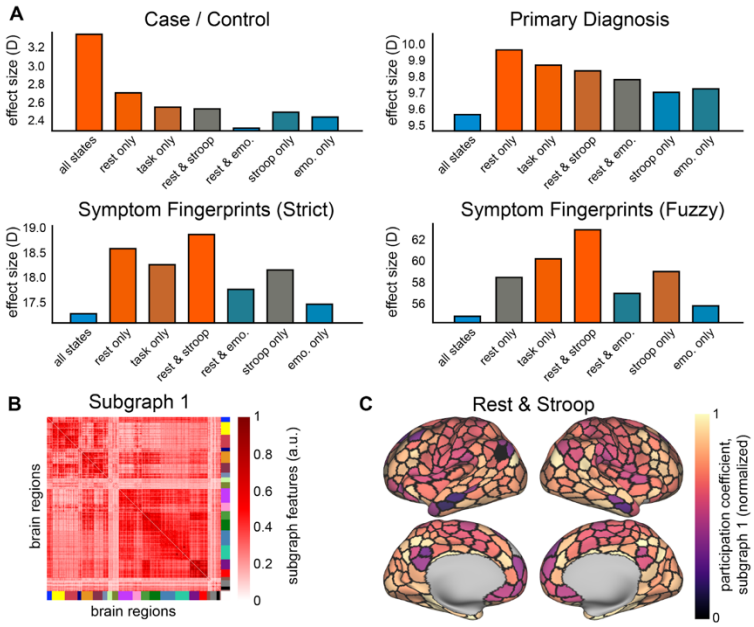
*"...For example, a key hypothesis would be the extent that network dynamics account for symptom structure is determined by a breakdown in the efficiency of information processing across varied task contexts. It will also be essential to leverage the observations of our exploratory analyses (**Fig. 5**) as important constraints on follow-up hypotheses in future work. For example, given the general impact of cognitive control processing, how do dynamic repertoires involving inhibitory control, working memory load, or task switching (to name a select few) specifically impact individual differences in symptom structure? Importantly, future studies should use replication datasets that enable such research questions and allow generalization of the present findings to independently sampled cohorts and task contexts. This would also have the benefit of increased sample sizes, which is particularly important for analyses that may be inherently exploratory."*

In the Results, we now suggest that the consistency of effect size values within each group of models, alongside a comparable range observed in null models, suggests that this is an appropriate way to address the problem of comparing results where outcome data likely has differing fundamental properties. This, alongside results around Fig. 4, supports a central proposal in the study; that dimensional and diagnostic phenotyping are fundamentally different ways of characterizing psychiatric symptom structures (as in Fig. 1A, reproduced below for reference), such that we expect the boundaries of these 2 characterizations will be separated by brain network features with differing levels of accuracy. To make models reasonably comparable, we held predictors (NMF-uncovered brain network dynamics) consistent across all models and applied the same effect size metric to all models. The range of effect sizes are potentially different given different chance levels for the varied groups of models. However, chance levels were handled comparably for each group of models because they were empirically based on nonparametric permutation testing to build null distributions and model relevant significance statistics. We did not originally report null model results for Fig. 5A but thank the reviewer for the opportunity to now add this context to the Results, which was also partly indicated by reviewer #2, comment #7. Additionally, we followed current best practices in data splitting to prevent across-participant data leakage from inflating classification results. This included a strictly held-out validation set used for final model testing and subsampling of a discovery dataset during model training and parameter tuning. Importantly, k-fold cross-validation following this data splitting scheme was applied consistently to all models in the study. With this revision, we suggest that we have controlled for the possibility of models being overfit to noise to a reasonable extent, but have not ruled it out fully (i.e., in the absolute sense), which is now enthusiastically suggested for future strategies in the Discussion.

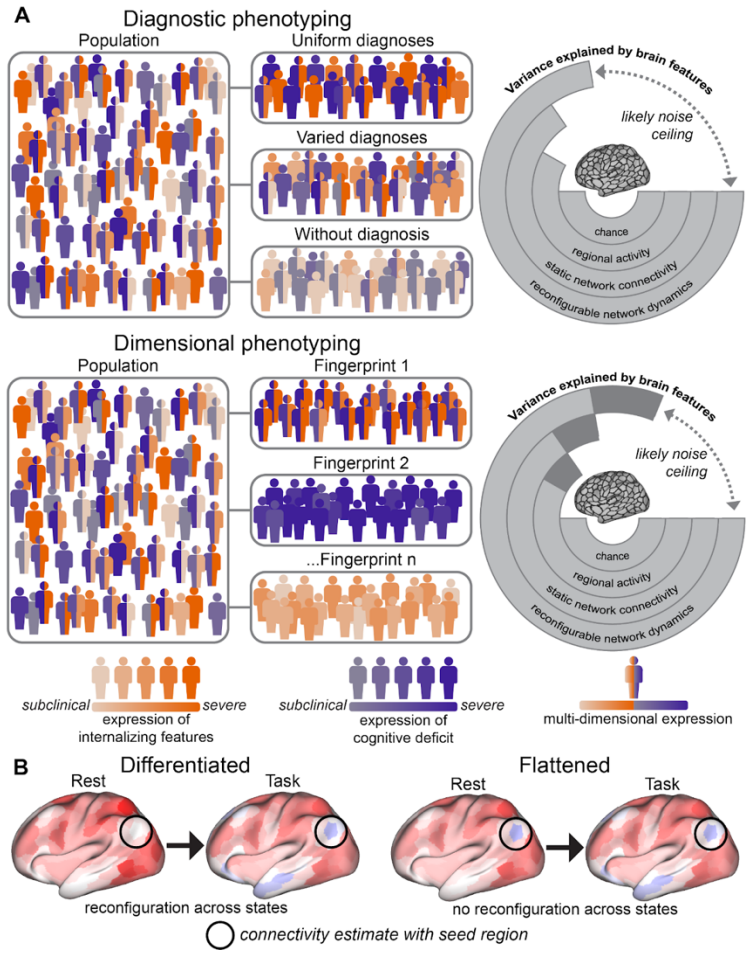
Figures 5 and 1 reproduced for reference:

Please note that we weighed the relative benefits and drawbacks of formatting y-axes in this manner (panel A) and ultimately decided to prioritize visually discerning patterns of model performance. Relatedly, classification

models across all outcome measures exhibited significant performances, just to varied extents, suggesting that these models were not strictly fitting to noise.



"Figure 5. Reconfiguring connectivity patterns from rest-to-Stroop task states prominently influence the classification of dimensional symptom profiles. (A)…*Bars are color-coded from blue-to-orange to indicate worst-to-best model performance for each given model comparison group (different outcome measures). These color mappings are independently represented within each model comparison group.**"*



"Figure 1. Theoretical framework."

Comment 3: Staying with this final set of findings, if the overarching argument of the introduction is that the ability of the brain to flexibly enter and exit myriad brain states, it is not clear why the effect size for all states (for the symptom fingerprints and primary diagnosis) would be by far the smallest. This is, in fact, evidence against the theoretical framing of the paper. No where in the paper do the authors truly grapple with this fact, other than to say in passing that, "network dynamics across wider-spanning task contexts may yield suboptimal decision boundaries." This doesn't really say very much other than just rehashing their finding.

Response 3: We greatly appreciate that the reviewer brought up this miscommunication on our part, as these points are integral to the framework proposed in this study. The overarching argument of the Introduction is that flexibly-instantiated brain network dynamics are more informative with respect to information processing than static brain network features, particularly when considering shifts in cognitive state or context and how these processes may be altered in psychiatric illness. Further, we (along with others; for example: [Yin D, Kaiser M. Understanding neural flexibility from a multifaceted definition. Neuroimage. 2021;235:118027.] and [Safron A, Klimaj V, Hipólito I. On the Importance of Being Flexible: Dynamic Brain Networks and Their Potential Functional Significances. Front Syst Neurosci. 2021;15:688424.]; also see Introduction for further references) propose that this is because brain network dynamics are fundamental to temporally-variant processes and allow for flexible recruitment of goal-relevant resources (i.e., are task linked) to adaptively accomplish behavioral or cognitive goals in real time. One way of examining this through computational modeling is to assume that brain network dynamics can be captured when projected to a lower dimension; what we call hidden brain states (i.e., subgraphs). It is then an empirical question whether or not more data (i.e., more task contexts) would more accurately predict transdiagnostic symptom structure. This was why our exploratory follow-up analyses in Fig. 5 (reproduced in response to comment #2 above) were essential to the present study: to compare classification accuracies for varied shifts in task context, as parameterized by including different subsets of across-state connectivity patterns in NMF. If every model compared in Fig. 5 had accuracies simply going up with more data inputted to NMF, we would not be able to immediately rule out that effects were just an artifact of data quantity. On the contrary, while all fMRI data combinations yielded significant predictions, the model comparison results in Fig. 5 demonstrate that classification results were not solely driven by the amount of data.

Relatedly, we would like to highlight that the outcome measure of symptom fingerprints is not reflective of general functioning, but the extent that participants express deficits in functioning that is clusterized on high-level domains. As far as we know, this research question is posed in a novel way such that we had no strong expectations about which subsets of brain network data would be most predictive of dimensional symptom structures in a highly transdiagnostic cohort. Combined, these are the gaps in understanding we are filling and based on the reviewer's insightful comments, we have clarified the novelty of our findings (quoted text revisions below). Additionally, there is extensive prior literature supporting the importance of cognitive control processing we observed in the present study.

Given that shifts across an intrinsic, resting state to a cognitive control state best model symptom fingerprints (Fig. 5), we have made the following relevant revisions to the **Introduction** to clarify these considerations and framing (lines 79-159):

> *"...This literature provides converging evidence for the key role of network-based mechanisms in the facilitation of flexible information processing across the cortical sheet[7,16,30]. While resting-state connectivity patterns reliably approximate the properties of functional specialization within and between large-scale brain systems[31-33], task-linked changes in connectivity patterns are thought to enable dynamic shifts in context-dependent processing over time[16,34,35]. Moreover, there is mounting evidence that time-varying, flexible shifts in connectivity patterns (here termed reconfigurations[30,36-39]) more optimally account for individual differences in both cognition and behavior than traditional, "static" approaches for studying brain functions[7,40-44]. Importantly, given that flexible processing is key for adaptively responding to changing environmental or cognitive demands, alterations in brain network dynamics are theorized to contribute to the onset and maintenance of psychiatric illness[45-49].*
>
> *…*

*... Here, we adopt a transdiagnostic framework by analyzing data from a sample of individuals with and without psychiatric diagnoses. We used cognitive and behavioral measures to uncover individually-profiled symptom structures across core dimensions of psychiatrically-relevant functioning[62,68,69,83-85] (**Fig. 1**), which we term symptom fingerprints. Given the importance of brain network dynamics to adaptive information processing, we hypothesized that reconfiguring connectivity patterns across a varied set of cognitive states would accurately separate individuals based on dimensionally-organized symptom structure[3,53,58,61] (**Fig. 1**), as well as by their case/control status and primary diagnostic labels. While prior work suggests that impaired cognitive control processing is a transdiagnostic, commonly-exhibited deficit across psychiatric illnesses[46-49,86], the extent that dimensional symptom structures would be separable by specific shifts in cognitive context was unclear. Thus, we implemented follow up analyses that used a systematic model comparison approach to explore the extent symptom structure was accurately separated by network dynamics underlying specific shifts in cognitive state. Specifically, we set out to explore whether classification accuracy would differ in a variety of contexts: (1) with more brain network data included in the predictor set, regardless of cognitive state; (2) in the presence of resting-state network dynamics; (3) task-state dynamics (i.e., without resting-state), regardless of task context; (4) via a specific shift in task context, such as increased cognitive demand, indicative of control processing; and (5) by varying subsets of brain network dynamics included as model predictors. In all analyses, we used non-negative matrix factorization (NMF) to model brain network reconfiguration dynamics, which can reveal underlying constraints upon connectivity patterns as individuals transition across cognitive states[87-93]. We discovered that network reconfiguration dynamics were flattened (i.e., less differentiated) across cognitive states in patients, and these dynamics could be used to accurately predict case/control status, primary diagnoses, and participant-specific symptom fingerprints. The link between network dynamics and dimensionally-organized symptom fingerprints was prominently driven by shifts from intrinsic to cognitive control[94,95] (i.e., Stroop[96,97]) task states. Further, control (frontoparietal), thalamic, and task-linked (i.e., visual and somato/motor) network regions were particularly important for classifying symptom fingerprints. These analyses reflect an essential, proof-of-concept roadmap for framing future experiments that systematically identify the extent that brain network dynamics confer transdiagnostic symptom structures in psychiatric illness."*

Revisions to **Results** text around Fig. 5 (please note that blue-text revisions below were partly motivated by comment #2 above) (lines 498-545):

"...To test how influential each cognitive state was to the link between brain network dynamics and symptom fingerprints, we implemented an exploratory analysis where NMF was run with varied combinations of input connectivity patterns.

…

...The rest-and-Stroop model performed the best (strict accuracy: D=18.85; fuzzy accuracy: D=62.91), suggesting that connectivity shifts between an intrinsic state and a cognitive control state accounted for individual clinical variability particularly well. Importantly, in symptom fingerprint models, classification accuracy did not simply increase with the amount of brain network data inputted to NMF, but instead increased when the Stroop task was present. This supports the proposal that brain network dynamics underlying shifts into more cognitively demanding task contexts are robust predictors of dimensional symptom structure across psychiatric illnesses. Interestingly, case/control and primary diagnosis were classified with a similar pattern across models. Case/control and primary diagnosis characterizations may involve a shared underlying relationship with symptomatology given that primary diagnosis can be thought of as an expanded version of the binary case/control status. However, the exact extent that each model fully accounts for noisy fluctuations in symptomatology is an important consideration when comparing these two patterns of results (see Discussion). With this consideration, the main difference was that all-states performed worst and best in classifying primary diagnosis and case/control statuses, respectively. These data suggest that separability of case/control status may be sensitive to the number of cognitive states (or, the amount of data) used to uncover network dynamics."

Comment 4: Description of what is network efficiency and how to quantify it is lacking. Also, it is not clear why network efficiency would be used as a criteria to focus on the results (i.e., why lower efficient

subgraphs would be presented in the supplement whereas subgraph 1 receives the primary focus in the main text).

Response 4: We have now added details in the revised manuscript's Methods section below, which indicate that brain network efficiency is a measure of how well a given connectomic organization supports information transfer across the system. Thus, we used network efficiency as a targeted test of which subgraph exhibits process-level properties relevant to our research question. However, instead of focusing on the subgraph exhibiting the least efficiency, we focused on the subgraph with the highest efficiency. We did this to err on the conservative side of not overfitting downstream classification models to a network property that is thought to correspond to poor network dynamics. We explain in greater detail below why focusing on one subgraph was important for this work.

We have added the following text to the **Methods**; detailing the computation of network efficiency and apologize that this was originally missing (lines 960-981):

> *"**Focus on one subgraph: network efficiency***
> *In the present study we opted to report on one subgraph in the main text and all other subgraphs in the Supplement (see Discussion for considerations). Unlike other dimensionality reduction techniques such as PCA, NMF does not inherently output subgraphs that are ranked by explained variance in descending order. Instead, properties of subgraphs can be assessed based on research-question- or discipline-specific priorities. In the current application, we quantified the network efficiency of each NMF-uncovered subgraph, which measures the extent that a given network organization supports the flow of information across the system. Efficiency was calculated locally, or for each brain region, which can then be averaged to estimate global efficiency across all brain networks. Following Wang et al.[165] local efficiency of weighted, undirected networks is given by:*
> $$efficiency_{local}(i) = \frac{1}{k_i(k_i-1)} \sum_{j,h;j\neq h} a_{ij}\, a_{ih} \left[d_{jh}(N_i)\right]^{-1}$$ eq. (1)
> *In a network with N nodes, or brain regions, we focus on three regions (triangles) at a time, i, j, and h, with k_i being the number of other regions connected to region i, or its degree. A subset of the larger network that contains all the regions neighboring i (but not i) is N_i. The term d_{jh} is the shortest path length between regions j and h considering only N_i, or neighbors of i; this is set to infinity if there is no path with these neighbors. Here, a path between neighbors that have a strong connection with i will impact the efficiency score more than neighbors with a weak connection to i. Lastly, a terms are set by the edge weight (connectivity estimate) between relevant regions. Wang et al.[161] refer to efficiency as a measure of "…fault tolerance of the system…", which indicates how resilient communication pathways are to removal of their commonly connected neighbor-of-interest. In the present study we focused on the subgraph that exhibited the highest network efficiency (all other subgraph results are reported in the Supplement)."*

We also added a straightforward estimate of *system segregation*, or the extent that network boundaries are distinct and large-scale functional systems are less specialized. The inverse of segregation can be used as a proxy measure for de-differentiation or network flatness; i.e., segregation of 0 suggests no differentiation of within- versus between-network boundaries; segregation of 0.5 suggests within-network connectivity exhibits twice the weight of between-network connectivity; segregation of 0.99 or greater suggests that within-network weights are much larger than between-network. This measure tightly correlates with modularity, but is simpler to compute and interpret. The following analysis of network segregation (flatness) has been added to the **Results**, which provide further support that subgraph 1 exhibits the properties relevant, but not overfit, to our research questions (lines 233-247):

> *"...Given that across-state subgraph coefficient patterns were expressed consistently in each of the k=5 subgraphs for both groups (**Fig. 2D**), we first restructured each subgraphs' features (**Fig. 2C**, middle matrix; **Supplemental Fig. S1**) into standard region-by-region networks (**Fig. 2A-B**) and then applied a network efficiency metric[105-107] (**equation 1**). Subgraph 1 exhibited the greatest network efficiency (**Fig. 2E**) both locally (brain regions) and globally (functional brain networks), suggesting that information flows across this subgraph's functional architecture more readily than subgraphs 2 through 5. We also quantified system segregation (**equation 2**) and found the following: subgraph 1 S=0.31, subgraph 2*

S=0.25, subgraph 3 S=0.23, subgraph 4 S=0.27, and subgraph 5 S=0.3. Subgraph 1 exhibited the greatest system segregation, suggesting that these features are organized in a manner that supports differentiated information processing across shifts in cognitive state. Lastly, Mantel correlations[108] between each pair of NMF-uncovered subgraphs ranged from 0.968 to 0.996, suggesting that all subgraphs exhibit largely similar topology. Altogether, these results support focus on subgraph 1 for the scope of the present research questions (see Discussion for considerations). Therefore main-text ***Results*** *refer to subgraph 1 and subgraphs 2 through 5 are reported in the* ***Supplement****."*

And the corresponding **Methods** for system segregation (lines 983-993):

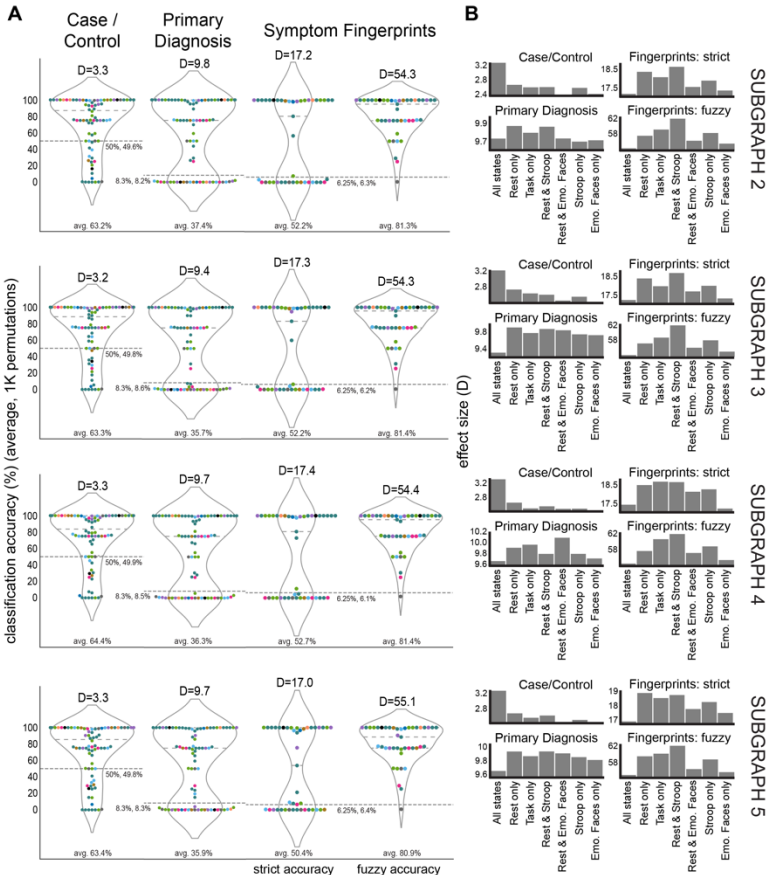
> *"**Focus on one subgraph: system segregation***
> *In addition to network efficiency, we measured system segregation[162,163,170] of each NMF-uncovered subgraph. Following Chan et al.[162], system segregation is a global network metric given by:*
> $S = \frac{\bar{x}_{within} - \bar{x}_{between}}{\bar{x}_{within}}$ *eq. (2)*
> *Where $\bar{x}$ is the mean of the edge weights: within-network ($\bar{x}_{within}$) and between-network ($\bar{x}_{between}$). Segregation is a measure of de-differentiation, or flattened network structure, where S=0 indicates $\bar{x}_{within} = \bar{x}_{between}$; S=0.5 indicates that $\bar{x}_{within}$ exhibits twice the average connectivity weight of $\bar{x}_{between}$; and S values reaching 0.99 indicate $\bar{x}_{within}$ is much larger than $\bar{x}_{between}$. A perfect score of S=1.0 is possible in sparse networks where average between-network connections are weighted 0. At higher S values, network organization is more modular. At lower S values, network organization is flatter and linked with suboptimal recruitment of task relevant resources[163]."*

Additionally, unlike with PCA components, the extent that NMF subgraphs explain variance is not guaranteed to be ranked in descending order (i.e., the first NMF subgraph may or may not explain the most variance). While explicit tests of explained variance are not straightforward with NMF, we have tested the extent that NMF subgraphs overlap, and briefly the range of Mantel r values across subgraph-to-subgraph pairs was 0.968 - 0.996, suggesting high similarity of all subgraphs. We have added this to the Results text (quoted above). Further, if we adapt the Frobenius norm (which is measured between the real across-state connectivity patterns and the NMF-reconstituted connectivity patterns) for each subgraph separately, the values are non-zero (ranging from 745.9 - 2635.2) suggesting non-orthogonality was captured. This is adapted from a penalty term in the objective function for "orthogonal NMF" (oNMF), which is a variant on the standard version of NMF that we used, which explicitly enforces across-subgraph Frobenius norms that are near 0. Thus, given that vanilla NMF is flexible with respect to non-orthogonality (an important model specification that we expand upon in response to comment #6 below), it is expected for NMF to uncover recurring patterns of features that constrain the input matrix. When applied to brain connectivity data, these recurring patterns could be considered a repertoire of (at least partially) repeating brain states [see here: Khambhati AN, Sizemore AE, Betzel RF, Bassett DS. Modeling and interpreting mesoscale network dynamics. Neuroimage. Published online June 20, 2017. doi:10.1016/j.neuroimage.2017.06.029.].

There have been studies that explicitly test the extent that subgraphs differ, and given corresponding expression levels (in the coefficient matrix), how these differences are associated with participants, tasks, etc. (i.e., samples from the input matrix). To the best of our knowledge, such prior studies have assessed data from healthy participants. In the current study, we adapted NMF techniques to study a psychiatric cohort, which also included extensive consideration and model development around dimensional symptom structures via clinical, cognitive, and behavioral data. Therefore the present study has two concurrent pipelines of model development: brain network dynamics and dimensional symptoms. We aimed to provide a robust, albeit proof-of-concept-style, analysis testing the initial hypothesis that brain network dynamics are linked with dimensional symptom structures transdiagnostically. Toward this goal we intentionally performed follow up analyses to explore the extent that: (1) large scale functional systems contribute to this link (Fig. 4B-C), and (2) whether a subset of state-to-state shifts in network interaction patterns are more predictive of symptom structure than others (Fig. 5). It was essential to answer these questions before examining interactions between subgraph expressions and task structure across symptom profiles – i.e., where it would become necessary to report on all subgraphs equally. In short, analyses that include all possible subgraph solutions were out of scope for the goals of the present study. However, we have reported the classification model results for each subgraph in the Supplement (Fig. S2 reproduced below for reference). Coauthors of the present study share the reviewer's interest and are currently preparing follow up analyses that systematically examine the relationships between brain network

dynamics supporting varied task contexts (and cognitive demand levels) with symptom structure. For these combined reasons we have added the considerations to the **Discussion** (lines 710-726):

> *"...The current approach can be readily expanded on. For example, we used state-general estimates of FC – that is, connectivity estimates for the entire duration of the Stroop task, Emotional Faces task, and resting-state. This could be refined in future work by estimating FC based on task context, condition, or other temporally-linked factors. Relatedly, connectivity can be estimated by alternatives to product-moment correlations, such as the improved validity given by regularized regression[101]. Additionally, we focused on results for subgraph 1, which exhibited the most network efficiency[161] and system segregation[162,163]. We reasoned that, given all subgraphs 1-5 had highly similar topology (measured via Mantel r), one subgraph is likely a reasonable approximation of the other subgraphs, particularly as this study was partly an exploratory proof-of-concept (e.g., results in **Fig. 5**). We thus chose subgraph 1 given that it exhibited two network properties known to efficiently support shifts in information processing. Importantly, we expected lower network efficiency and segregation to correspond with transdiagnostic symptomatology, so the focus on subgraph 1 (with high efficiency and segregation) suggests that classification model results were not overly driven by features linked with generally suboptimal network dynamics. However, given recent work suggesting that subgraph organization covaries with cognitive demand, future work should build upon model comparisons in the present study by characterizing how brain network dynamics supporting specific shifts in task context confer individual differences in symptom structure."*



"Figure S2. Brain network dynamics can accurately separate symptom fingerprints, subgraphs 2 through 5 (paneled rows)."

Comment 5: In the introduction, authors state: "tend to outperform measures that assume stable, temporally-invariant, patterns of connectivity" but can the authors be more specific? In general / on average, how much do they outperform static connectivity metrics?

Response 5: We have revised the Introduction to include more specific studies demonstrating this. To the best of our knowledge, there are no specific meta-analyses or large-scale methodological studies that extensively cross-compared all/majority variants of static connectivity versus all/majority variants of dynamic connectivity, with respect to diagnostic group differences. The **Introduction** text was revised as follows (lines 99-114):

> *"...Importantly, with respect to predicting diagnostic status, network dynamics (**Fig. 1A**, e.g., inter- or across-state changes in connectivity, as in **Fig. 1B**) tend to outperform measures that assume stable, temporally-invariant, patterns of connectivity[53,58,61]. Specifically, de Lacy and Calhoun [58] found that adolescents diagnosed with attention-deficit/hyperactivity disorder (ADHD) were not differentiated from typically developing (TD) youth on the basis of time-invariant (i.e., static) functional connectivity. However, they found significant differences in how fluid global brain network dynamics were between ADHD and TD youth. These findings were replicated in analyses that varied the number of states in the state-space solution and varied temporal window size. Further, Rashid et al. [61] explicitly compared static FC against dynamic FC, with respect to how well each classified the diagnosis groups of schizophrenia, bipolar disorder, and no diagnosis. Classification accuracy in the static FC model was approximately 59% and approximately 84% in the dynamic FC model. In a hybrid model using both static and dynamic FC as predictors, classification accuracy nominally increased to 89%, suggesting that the bulk of the variance was carried by dynamic FC. However, despite growing evidence for the clinical relevance of altered network dynamics in patient populations [also see: Reinen et al.[52]], the extent that shifts in connectivity patterns across cognitive states accounts for individual-specific symptom structure remains to be established."*

Comment 6: Because NMF requires positive inputs, how did the authors adjust the connectivity data to adhere to this requirement. That is not described in the methods or results. How stable are these results if the authors use an alternative dimensionality (SVD, PCA as described) reduction approach?

Response 6: We thank the reviewer for pointing out this oversight. We have added the following text to the appropriate **Methods** section (lines 893-897) to clarify how we positively thresholded NMF inputs, following standard practices from recent literature:

> *"...Standardly, all matrices and/or tensors are non-negative. We accomplished this by following prior literature[87] and having the model only consider positively-weighted connectivity estimates in NMF input data; negative FC estimates were set to 0. As noted by Anderson et al.[89], positive thresholding of connectivity estimates yields a sparser network, which is a constraint that is thought to impose a more neurobiologically plausible information processing architecture (also see [101])."*

We opted for NMF instead of alternative dimensionality reduction techniques given the non-orthogonality of resulting subgraphs, as well as the relative ease of interpreting NMF results given that it is a parts-based decomposition. In a 2015 study, Sotiras and colleagues [Sotiras A, Resnick SM, Davatzikos C. Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. *Neuroimage*. 2015;108:1-16.] systematically compared PCA, ICA, and NMF and found that NMF results most closely aligned with known functional organization properties of brain connectivity and captured pathology better than alternatives. However, we are likewise interested in the stability of these results in a more methods-focused study cross-comparing NMF with other dimensionality reduction techniques. We strongly encourage this for future work in the **Discussion** (lines 731-741) as follows:

> *"...NMF has also been successful with multivariate data inputs, such as structural and functional connectivity[89], which may be an important extension of the present work given evidence that properties of the structural connectome constrain the spread of psychopathological brain processes[137]. Further, while recent studies suggest that NMF enables discovery of more biologically realistic brain network dynamics than other dimensionality reduction techniques[87-89], this should be explicitly tested in follow up, methods-focused developments that build upon the present study. Related to this question, Sotiras et al. [93] analyzed multivariate patterns in structural neuroimaging data by comparing NMF against PCA and ICA. They found that NMF had improved ability to organize structural brain images in a manner that closely aligned with known functional systems. While this supports our use of NMF over PCA or ICA (or*

related techniques), the stability of our findings over such methodological variants is an important empirical question that future work should address."

Additionally, we expanded our prior considerations in the **Results** to emphasize the utility of non-orthogonality (lines 897-903):

"...NMF-uncovered features are optimized for a linear approximation of the original input data and are not constrained by independence or orthogonality (as in principal components analysis)[87], which is likely more neurobiologically appropriate because flexibly co-occurring and/or overlapping network structures are permitted. Importantly, this relaxed constraint of orthogonality is thought to allow for the possibility that brain regions participate in more than one network repertoire, which likely varies over time and cognitive processing context[88]."

We have also quoted relevant text from the **Results** section corresponding to initial network dynamics findings (lines 208-226); please note that this was revised mostly in response to comment 11 (see below), but is also relevant here:

"...Consistent with other dimensionality reduction techniques, NMF extracts a hidden structure that best accounts for a given input dataset. Compared to other techniques, such as PCA, NMF has the benefit of being parts-based, allowing for straightforward inferences regarding reconstruction accuracy. NMF is also less constrained with respect to orthogonality. Non-orthogonality is neurobiologically relevant for multi-state or multi-modal data, as features underlying time-varying neural data are likely expressed with some overlap. For example, while we explicitly examined reconfigurations across intrinsic and task-evoked FC patterns, these states share a structural connectivity backbone[99-102] and their underlying dynamics are not strictly orthogonal. That is, structural connectivity – or estimates of physical connections between brain regions, typically via white matter tracts[103] – is not thought to be modified across short timescales (e.g., seconds, minutes) in response to information processing shifts. Therefore, structural connectivity is thought to be constant during resting-state and varied task states. However, decades of functional connectivity studies suggest that factors that change across short timescales, such as cognitive context and task goals, are linked with transient reconfigurations in FC[21,30,35,52,101,104]. We aimed to examine such dynamics in the present study and reasoned that the relaxed orthogonality constraint of NMF[88,89] allows for reconfiguring network interaction patterns that otherwise share an underlying structural connectome. Altogether, NMF-uncovered subgraph features, and corresponding expressions of those subgraphs, allowed us to infer key constraints upon the dynamics that underlie across-state brain network reconfigurations."

Comment 7: The magnitude of the difference in cosine similarity seems quite small, despite a very 'significant' p-value reported. Can the authors clarify how and over what was this test performed? Was it a linear mixed effects model with pairwise (state-to-state) cosine similarity as the outcome and diagnostic status as the predictor?

Response 7: We greatly appreciate the reviewer's feedback on the cosine similarity analysis, and apologize that it was not initially clear. We agree and believe that the opportunity to revise the manuscript to explain this analysis in greater detail will improve reader understanding of the findings around flattened network dynamics. We also note that reviewer #2 recommended we update these analyses to use a nonparametric approach to significance testing (their comment #2 below). Therefore, quoted text below partly considers a response to reviewer #2 comment #1. Lastly, these two comments are relevant with this reviewer's (reviewer #1) comment #12 (i.e., clarify Fig. 2G), which we will note in response to comment #12 below as well.

We have added the following text to the Methods section on brain network dynamics, which now fully outlines analyses around Fig. 2 and corresponding **Results** text, which demonstrate flattened brain network dynamics in the TCP dataset's transdiagnostic sample of psychiatric patients (lines 923-958):

"...We reported the expression patterns of subgraph coefficients exhibited across cognitive states in terms of relative percent. This was accomplished via min-max normalization, with zero and one as the lower

and upper limits, respectively. Normalized subgraph coefficients were then multiplied by 100 to quantify an expression of network dynamics out of an empirically-constrained 100% upper limit. This was reported for both case/control groupings, where we conducted a cosine similarity analysis with significance testing on group differences (see below), and primary diagnosis groupings, which were exploratory and only relative differences were summarized.

Flattened constraints on dynamics: cosine similarly of subgraph expression patterns

We quantified the similarity of each state-to-state pair of brain network dynamic measures in order to test the hypothesis that dynamics are flattened in participants with psychiatric diagnoses relative to those without diagnoses. As in other analyses in this study, the measure capturing network dynamics was the across-state expression patterns of NMF-uncovered subgraph coefficients (focusing on the first subgraph in the main Results). We quantified the similarity of each state-to-state pair of subgraph coefficient vectors (i.e., cross-participant values comprising a vector) with cosine similarity, which is a measure of the extent that two data vectors are oriented similarly in high-dimensional space. Consistent with all other analyses, there were 6 cognitive states considered (rest 1 AP, rest 2 AP, rest 1 PA, Stroop AP, Stroop PA, and emotional faces AP), which meant 15 comparisons via cosine similarity (i.e., combinations without repeats). These state-to-state comparisons were conducted separately for participants with psychiatric diagnoses and those without diagnoses, which were then compared statistically. To compare with-diagnosis versus without-diagnosis, we first did a one sample t-test on the true data, testing the alternative hypothesis that state-to-state cosine similarities of with-diagnosis participants minus cosine similarities of without-diagnosis participants is significantly greater than zero. This test indicated whether the extent of flatness exhibited across state-to-state pairs of network dynamics (as quantified with cosine similarity) is significantly greater across the transdiagnostic cohort of psychiatric patients, relative to participants with no psychiatric diagnosis. Then, we randomly shuffled case/control labels (i.e., shuffling with-diagnosis and without-diagnosis labels) over 10,000 permutations and built a nonparametric null distribution of t-statistics by testing the same alternative hypothesis, just with null group labels[109]. Here, significance was determined by a true t-statistic that was greater than the t-statistic at the 95th percentile of the null distribution, indicating a p<0.0001.

We also applied a more exploratory version of this cosine similarity analysis to data stratified by primary diagnosis grouping. Here, we did not subtract any group from any other group, but instead reported cosine similarity of each state-to-state pair of network dynamic measures (NMF subgraph coefficients) for each primary diagnosis group. Higher similarities indicated relative flatness (or de-differentiation) of network dynamics exhibited by that group; with a maximum score of 1.0."

Additionally, we have added the following text to the **Results** section, which summarizes results from the nonparametric null modeling approach (lines 254-264):

*"...Thus, a positive difference in cosine similarity indicated relative flattening of dynamics (see Methods for full details), which we expected for non-diagnosed participants relative to diagnosed participants (**Fig. 1B**). In the majority of state-to-state pairings, the transdiagnostic psychiatric sample exhibited more flattened subgraph expression coefficients. Across state-to-state pairs, the average cosine similarity of those without diagnosis was 0.91 (standard deviation (SD)=0.01) and 0.95 with diagnoses (SD=0.03; cosine similarity can range from -1 to 1). We built a nonparametric null model[109] over 10,000 shuffled case/control labels (see Methods) and found that the difference of with-diagnosis minus without-diagnosis cosine similarities across all state-to-state pairs of network dynamics were significantly greater than zero with a p<0.0001 (true t(14)=5.1, null t-threshold=4.2). This suggests that the dynamics constraining across-state brain network reconfigurations were comparatively flattened in those with psychiatric diagnoses versus those without diagnoses."*

Comment 8: The clustering performed for the 'behavioral measures' is odd. In particular, the highest loading on the 'Externalizing Cluster' is the SHAPS anhedonia scale…I don't think any mental health researcher would consider anhedonia part of an externalizing dimension. Moreover, most of the other high loadings on that cluster are poorly accounted for by the construct of 'externalizing'. Seems like that should be re-considered.

Response 8: We thank the reviewer for providing their valuable expertise and agree that the clusterized domain we termed *externalizing* maintains some unexpected edge cases. To empirically derive names for these symptom clusters, we took a weighted winner-take-all approach so that cluster naming was not overly driven by a limited subset of the included measures (unless a measure's factor loading score was exceptionally high, which we did not observe). We note that the anhedonia SHAPS measure only accounted for ~16% of the overall loadings in the externalizing cluster, which also included 23 other measures. This approach is outlined in the **Methods** as follows (lines 1053-1060):

> *"…Following prior work[80,118,119], clusters were named for interpretability, however we caution against overly interpreting this naming system (as well as priorly used naming systems). To partially account for potential researcher bias in naming clusters, one researcher pre-labeled each of the 110 behavioral measures with broad dimensions of cognitive or behavioral functioning (referred to as dimensions of functioning in the main text), which was corroborated by two other researchers (**Supplemental Table S1**). Then, following clustering and PCA, the percentages of each label in each cluster – weighted by normalized factor loadings – were summarized programmatically."*

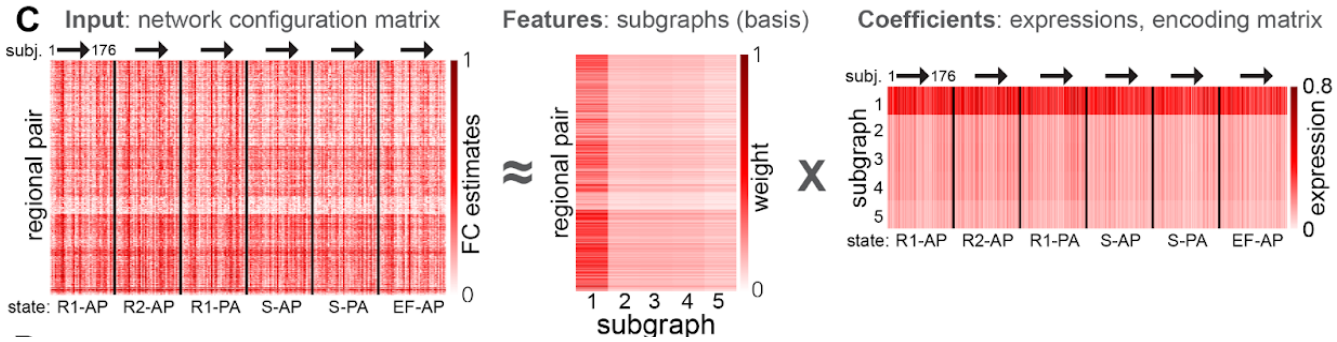
Supplemental Table S1 also provides detailed information for each measure, *a priori* labels, cluster assignment, and factor loading for transparency. Importantly, to minimize researcher bias (including our potential bias in the present work and the influence of others' biases from prior work), we included *a priori* labels that covered all applicable domains of functioning. This meant that some measures had more than 1 *a priori* label. SHAPS had three: internalizing, reward, and externalizing. The *a priori* inclusion of externalizing as a possible label was based on the measure partly considering social factors, such as getting pleasure from helping others. Given that the construct of externalizing is framed with respect to outward, often anti-social, behaviors that negatively impact one's functioning, there appears to be a face-value link between social anhedonia and externalizing symptoms. However, we agree that aspects of the SHAPS are more suited to the internalizing label, and have thus added this as an important consideration in the Discussion. Moreover, we emphasize that our dataset includes a dense, widely-spanning set of behavioral measures that allow for minor discrepancies in how select measures are allocated. This is important for the sample size of the TCP dataset, but also to explain enough variance in each domain of functioning that a data-driven approach remains valid.

Revisions to **Discussion** (lines 741-753):

> *"...Lastly, we are enthusiastic about future work that expands upon transdiagnostic symptom fingerprinting, for example, probing if the extent of differentiation in brain network dynamics scales with the extent that dimensions of functioning are expressed in individual symptom profiles. Relatedly, it will be important to validate and extend our symptom fingerprinting approach (**Fig. 3; Supplemental Table S1**) with future data collection efforts that likewise collect a widely-spanning set of behavioral, cognitive, and clinically-relevant measures. While dimensionally-based characterizations of psychiatric symptoms benefit from the emphasis on data-driven methods that capture individual differences, there are some inherent limitations. For example, select measures may be assigned to a cluster that is unexpected given prior literature. In the present study, we minimized this kind of noise by first collecting as many measures as is feasible without fatiguing participants, and second by implementing a weighted, winner-take-all approach that is not overly driven by a limited subset of included measures (see Methods). Additionally, future work could further mitigate cluster assignment noise via larger sample sizes and data collections usable for validation analyses."*

Comment 9: Figure 2C is confusing. Why does it start with subject 64?

Response 9: We are grateful to the reviewer for their attentive read of the manuscript. Our figure generation tools automatically started labeling at subject 64 on the top x-axis, but we have now manually modified it to start at 1 and end at 176 (the sample size for the discovery subset of data), with repeating arrows indicating that each subsequent block is repeating this pattern of participants 1-176. We hope this formatting is more intuitive and thank the reviewer for pointing this out. **Figure** 2C has been revised as follows (cropped for space; see comment #12 below as well for Fig. 2):



"Figure 2. Across-state brain network dynamics in health and psychiatric illness."

Comment 10: There is no Figure 5D, E, or F, though they are referenced in the manuscript.

Response 10: We greatly thank the reviewer for catching this error in the text around Fig. 5. We have now double checked the text and updated it to accurately reference panels in Fig. 5.

Comment 11: Can the authors clarify this statement: "these states share a structural connectivity backbone"

Response 11: We have now expanded upon the original statement and thank the reviewer for the opportunity to clarify. The **Results** now include the following text (lines 212-226):

> *"...Non-orthogonality is neurobiologically relevant for multi-state or multi-modal data, as features underlying time-varying neural data are likely expressed with some overlap. For example, while we explicitly examined reconfigurations across intrinsic and task-evoked FC patterns, these states share a structural connectivity backbone[99-102] and their underlying dynamics are not strictly orthogonal. That is, structural connectivity – or estimates of physical connections between brain regions, typically via white matter tracts[103] – is not thought to be modified across short timescales (e.g., seconds, minutes) in response to information processing shifts. Therefore, structural connectivity is thought to be constant during resting-state and varied task states. However, decades of functional connectivity studies suggest that factors that change across short timescales, such as cognitive context and task goals, are linked with transient reconfigurations in FC[21,30,35,52,101,104]. We aimed to examine such dynamics in the present study and reasoned that the relaxed orthogonality constraint of NMF[88,89] allows for reconfiguring network interaction patterns that otherwise share an underlying structural connectome. Altogether, NMF-uncovered subgraph features, and corresponding expressions of those subgraphs, allowed us to infer key constraints upon the dynamics that underlie across-state brain network reconfigurations."*

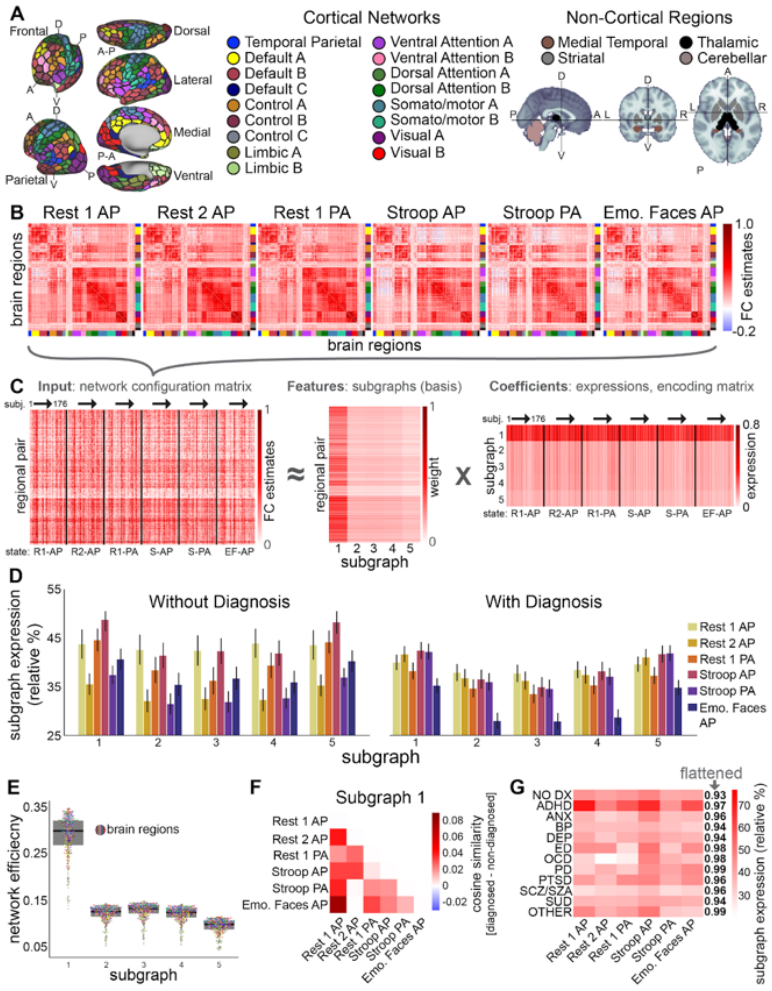
Comment 12: The description surrounding Figure 2G is not clear. I believe that each cell represents a proportion of total for that patient population, but could the authors describe further?

Response 12: We apologize for the lack of clarity in the description of Figure 2G analyses and thank the reviewer for providing the opportunity to improve the comprehension of these results. We have revised the Methods to include a section detailing this result, as well as the corresponding summary of Figure 2, which are quoted below. Please note that there is overlap with responses to comment #7 above and reviewer #2, comment #1 below.

Revisions to text in **Methods** section on brain network dynamics (lines 923-929):

> "...*We reported the expression patterns of subgraph coefficients exhibited across cognitive states in terms of relative percent. This was accomplished via min-max normalization, with zero and one as the lower and upper limits, respectively. Normalized subgraph coefficients were then multiplied by 100 to quantify an expression of network dynamics out of an empirically-constrained 100% upper limit. This was reported for both case/control groupings, where we conducted a cosine similarity analysis with significance testing on group differences (see below), and primary diagnosis groupings, which were exploratory and only relative differences were summarized.*"

Revisions to Figure 2G summary (and copy of Fig. 2 for reference) (lines 302-308):



*"**Figure 2. Across-state brain network dynamics in health and psychiatric illness.** … (**G**) Same as panel D (subgraph coefficients, which indicate expression of network dynamics), but participants grouped by primary diagnosis (rows). NO DX: no diagnosis; ANX: anxiety; BP: bipolar disorder; DEP: depression; ED: eating disorder; PD: panic disorder or phobia; SCZ/SZA: schizophrenia/schizoaffective; SUD: substance use disorder. Rightmost column: average cosine similarity (here, an index flattened dynamics) of each diagnostic group (i.e., averaging the lower triangle as in panel F, without group-based subtraction). Higher indices (max of 1.0) suggest across-state brain network dynamics exhibit more flatness in that diagnosis group."*

Reviewer #2 (Remarks to the Author):

Summary: The authors provide compelling evidence that brain network dynamics are predictive of transdiagnostic symptom profiles, outperforming case-control and diagnostic category approaches. The methods they employ are robust and well described. While the sample size is modest for the analyses performed, this work is novel, offers a strong proof of concept and addresses a timely issue. Clarification on several points would strengthen this already strong manuscript.

Response: We thank the reviewer for their enthusiasm about our manuscript. We have incorporated revisions following each of the reviewer's expert points and believe this has demonstrably strengthened the manuscript.

Comment 1: The authors compare mean cosine similarities between patients with and without psychiatric diagnoses. Given that cosine similarity values range from -1 to 1, a z-transformation or a

non-parametric test may be more appropriate. Please clarify which statistical test was used and the rationale for this choice.

Response 1: We are thankful to the reviewer for providing their expert insight on this analysis and agree that nonparametric significance testing would be more robust in these analyses, which quantify extent of flattened network dynamics in the transdiagnostic sample of psychiatric patients. We would like to note that two comments by reviewer #1 above (comments #7, #12) partly overlap with this comment, thus this represents a significant improvement to the clarity of our manuscript. We have quoted all relevant revisions below, which include considerations for reviewer #1's related feedback, which we believe all together represents one cohesive revision.

Revisions to **Methods** section (Brain network dynamics) on flattened network dynamics, cosine similarity analyses, and added nonparametric null model (lines 923-958):

> "...*We reported the expression patterns of subgraph coefficients exhibited across cognitive states in terms of relative percent. This was accomplished via min-max normalization, with zero and one as the lower and upper limits, respectively. Normalized subgraph coefficients were then multiplied by 100 to quantify an expression of network dynamics out of an empirically-constrained 100% upper limit. This was reported for both case/control groupings, where we conducted a cosine similarity analysis with significance testing on group differences (see below), and primary diagnosis groupings, which were exploratory and only relative differences were summarized.*
>
> *We quantified the similarity of each state-to-state pair of brain network dynamic measures in order to test the hypothesis that dynamics are flattened in participants with psychiatric diagnoses relative to those without diagnoses. As in other analyses in this study, the measure capturing network dynamics was the across-state expression patterns of NMF-uncovered subgraph coefficients (focusing on the first subgraph in the main Results). We quantified similarity of each state-to-state pair of subgraph coefficient vectors (i.e., cross-participant values comprising a vector) with cosine similarity, which is a measure of the extent that two data vectors are oriented similarly in high-dimensional space. Consistent with all other analyses, there were 6 cognitive states considered (rest 1 AP, rest 2 AP, rest 1 PA, Stroop AP, Stroop PA, and emotional faces AP), which meant 15 comparisons via cosine similarity (i.e., combinations without repeats). These state-to-state comparisons were conducted separately for participants with psychiatric diagnoses and those without diagnoses, which were then compared statistically. To compare with-diagnosis versus without-diagnosis, we first did a one sample t-test on the true data, testing the alternative hypothesis that state-to-state cosine similarities of with-diagnosis participants minus cosine similarities of without-diagnosis participants is significantly greater than zero. This test indicated whether the extent of flatness exhibited across state-to-state pairs of network dynamics (as quantified with cosine similarity) is significantly greater across the transdiagnostic cohort of psychiatric patients, relative to participants with no psychiatric diagnosis. Then, we randomly shuffled case/control labels (i.e., shuffling with-diagnosis and without-diagnosis labels) over 10,000 permutations and built a nonparametric null distribution of t-statistics by testing the same alternative hypothesis, just with null group labels[107]. Here, significance was determined by a true t-statistic that was greater than the t-statistic at the 95th percentile of the null distribution, indicating a p<0.0001.*
>
> *We also applied a more exploratory version of this cosine similarity analysis to data stratified by primary diagnosis grouping. Here, we did not subtract any group from any other group, but instead reported cosine similarity of each state-to-state pair of network dynamic measures (NMF subgraph coefficients) for each primary diagnosis group. Higher similarities indicated relative flatness (or de-differentiation) of network dynamics exhibited by that group; with a maximum score of 1.0.*"

Revisions to **Results** section on brain network dynamics, which summarizes results from the nonparametric null modeling approach. These results robustly demonstrate that even when compared against an empirically-based and nonparametric null distribution (with large set of 10,000 permutations), cosine similarity was significantly greater in the psychiatric cohort (lines 254-264):

*"...Thus, a positive difference in cosine similarity indicated relative flattening of dynamics (see Methods for full details), which we expected for non-diagnosed participants relative to diagnosed participants (**Fig. 1B**). In the majority of state-to-state pairings, the transdiagnostic psychiatric sample exhibited more flattened subgraph expression coefficients. Across state-to-state pairs, the average cosine similarity of those without diagnosis was 0.91 (standard deviation (SD)=0.01) and 0.95 with diagnoses (SD=0.03; cosine similarity can range from -1 to 1). We built a nonparametric null model[109] over 10,000 shuffled case/control labels (see Methods) and found that the difference of with-diagnosis minus without-diagnosis cosine similarities across all state-to-state pairs of network dynamics were significantly greater than zero with a p<0.0001 (true t(14)=5.1, null t-threshold=4.2). This suggests that the dynamics constraining across-state brain network reconfigurations were comparatively flattened in those with psychiatric diagnoses versus those without diagnoses."*

Comment 2: Classification was conducted using support vector classifiers (SVCs), which is reasonable. However, tabular data are often better suited to models such as random forests or gradient boosted trees. Please comment on why SVCs were chosen, and discuss the potential trade-offs, particularly regarding the ability of these models to detect important interactions between predictors.

Response 2: We greatly appreciate the reviewer providing their expertise and agree that classification model is an important methodological consideration. We chose an SVM over other methods given that they are a well-founded approach to finding decision boundaries in complex data, particularly with the NuSVC implementation, which accounts for potential sources of noise more directly than the standard SVC. Additionally, the NuSVC kernel does not have strictly linear constraints, as in vanilla SVC, and is thus more flexibly able to capture nonlinear interactions between model features. Other methods may potentially provide beneficial trade-offs to an SVM – such as random forest (RF) classification – which we support future, methods-forward work in testing. RF classification is quite robust, particularly with larger sample sizes, which will become relevant in future work aimed to generalize the present findings in replication datasets, as they become available. While it is an empirical question for the present data, NuSVC and RF likely handle nonlinear interactions between predictors relatively equally. However, interpretation of feature importance is more inherent to RF classification, given that the structure of decision trees is conditional on feature interactions. Characterizing the importance of interactions between classification model predictors was not the focus of the present study, but we do agree that future work would greatly benefit from these methodological considerations, particularly when predictor interactions are central to the research question and thank the reviewer for this important feedback. We have added the following text to the **Discussion** which details a future line of research that fits these criteria (note this includes revisions partly in response to reviewer #1 comment #4 above, lines 723-729):

"...*However, given recent work suggesting that subgraph organization covaries with cognitive demand, future work should build upon model comparisons in the present study by characterizing how brain network dynamics supporting specific shifts in task context confer individual differences in symptom structure. This important continuation of the present findings would emphasize interactions between classification model predictors (i.e., network dynamics across varied task contexts), which may benefit from alternative classification approaches such as random forest, which more explicitly capture feature importance*."

Comment 3: The primary performance measure of model performance is limited to accuracy and "fuzzy accuracy". Could the authors provide standard metrics including sensitivity, specificity, positive and negative predictive values were not evaluated?

Response 3: We thank the reviewer for the opportunity to remark on reporting model accuracies, but not other metrics of model performance. Our classification problems were mostly multi-class, which means that even with multiple labels, each label can only be used once during classification (i.e., versus multi-label, where multiple labels can be applied simultaneously to each data point in the to-be-classified outcome measure). In both cases of primary diagnosis and symptom fingerprints, classes are not distributed evenly across the dataset, which we would not expect nor want in a transdiagnostic sample because incidence rates of varied psychiatric illnesses are not evenly distributed in the larger population. Also, there are 12 and 16 possible labels in diagnosis and

fingerprint models, respectively, which is more complex than typical multi-class problems. These combined reasons make interpreting measures of precision, recall, etc. challenging.

Therefore, we focused on building a data-driven null distribution around the metric of accuracy – this null distribution is the basis for what we termed empirical chance. While accuracy has similar interpretational pitfalls as precision and recall in inherently unbalanced and complex multi-class problems, it is the most straightforward and widely used. This combined with the nonparametrically structured chance level makes above-chance-accuracy the most robust inference out of these possibilities. Note that our use of effect size after significance testing of model accuracies against empirical chance was another step toward an unbiased estimate that can be used in model comparison. While it is possible to build a null distribution of the non-accuracy performance measures (precision, recall, F1, etc.), most of these require the "one versus all" method given multi-class outcome data; or otherwise require choosing a method for summarization across classes including whether to summarize before or after aggregating true positives, false positives, etc. (i.e., micro-averaging and macro-averaging). It is not immediately clear what the optimal decisions would be with respect to these considerations around multi-class models and building a null distribution of precision, recall, F1 score, etc., – as we did for accuracy, which does not have those added considerations. We have summarized these considerations in the revised **Discussion** near relevant considerations around model fit and controlling for sources of noise and thank the reviewer again for their expertise. For reference, these revisions are as follows (lines 677-703):

> *"...Relatedly, while predictors (brain network dynamics) were held constant in all model comparison analyses, the outcome measures studied may exhibit differing levels of inherent noisiness. The strategies we used to account for the possibility that classification models were overly fit to such noise included: comparing effect sizes instead of raw accuracies, given differing chance levels; using common sets of predictors across all models; and a strict scheme of splitting data into discovery, test, and unseen validation sets commonly across all models. Importantly, for all models compared, we also used a non-parametric and data-driven approach to developing null models that has shown to be rigorous for fMRI-based[109], brain connectivity data[21] where the statistical ground truth is not known. While this multi-pronged approach is robust, it is important to note that we cannot precisely quantify how much noise influenced model fit across the differing models. It will be an important methods advancement for future studies to use simulated data – where ground truth properties are known and varied by the researcher – to precisely establish signal and noise influences upon model fit. We also emphasize that future methods-forward studies should implement alternative metrics of model performance, such as precision and recall. The complex, multi-class predictive models assessed in the present study are most accessibly interpreted in terms of accuracy – in particular when chance accuracy levels were based on a nonparametric null distribution – however, adapting these considerations to alternative metrics will likely provide fuller insight with respect to model comparison. Similarly, we do not know the ground truth on precisely how much collection site differences impact model performance. We ensured that data collection was equivalent across Yale and McLean sites, particularly with standardized neuroimaging acquisition and processing protocols[116]. We likewise held the proportion of participants recruited at each site consistent when splitting data into discovery and validation sets (**Supplemental Fig. S7**). While these steps appeared to reasonably control for the influence of collection site given that representative proportions of Yale and McLean participants were classified above-chance (**Supplemental Fig. S5**), there is still a possibility that hidden sources of variability between sites influenced model fit. In future work where collection site differences are more likely to contaminate model inferences, or otherwise multiple datasets are being assessed for generalizability, we encourage the use of additional harmonization techniques[157], such as ComBat[158] or DeepResBat[159]."*

Comment 4: Simulated lesioning is a reasonable approach for identifying key connectivity measures. However, methods like SHAP or permutation-based importance scores are also widely used. Please briefly discuss the comparative advantages and disadvantages of your approach for this context.

Response 4: We thank the reviewer for their feedback and share their enthusiasm about methods considerations in computational modeling. Shapley additive explanations (SHAP) and permutation-based feature importance (PFI) techniques are quite similar to the simulated network lesioning approach in that all aim to identify important contributions to high dimensional models or data, such as experiments based on functional connectivity. Ultimately, we chose the simulated lesioning approach given its simplicity, which makes results more

straightforward to interpret. This was particularly important in the present study given that the research questions were inherently complex, and this study is partly a proof of concept given the novelty. Thus, reducing inferential complexity where possible became a higher priority than usual. Our approach was also computationally fast/inexpensive, which tends to make analytic pipelines more reproducible or otherwise more accessible to future researchers.

SHAP values are highly informative for feature selection and/or importance in predictive models, particularly when the underlying properties of the data are known, causal inference is well supported, or otherwise if the data structure is tightly controlled. In these contexts, SHAP is well suited to explain individual prediction instances. Further, SHAP serves best in problems where there is feature independence, and while our fMRI processing steps aim to reduce confounds such as autocorrelation, functional connectivity-based features will exhibit a degree of dependency between networks; in the manuscript we have termed this *network interaction patterns* where appropriate. SHAP is nonetheless a powerful approach and may be better suited for analyses focusing on sparse or regularized brain network organization.

PFI is an interesting alternative to simulated lesioning. As far as we know, the largest difference between these methods is how the feature-of-interest is handled when the model is re-run. In our approach, features were temporarily removed from the model, and in PFI they would be temporarily permuted within the model. It is likely that the results of these two approaches would converge, given that both involve a broken relationship between the feature and the model/system. It is not clear whether re-running NMF with permuted, instead of lesioned, network interaction patterns (on a network-by-network basis) would potentially introduce artificial statistical structure to the NMF configuration matrix, instead of removing that network from consideration altogether. While this is an empirical question (that we fully support future investigation into), most null network models applied to fMRI data aim to preserve properties exhibited across the connectome, while only breaking the single property of interest [see: Váša F, Mišić B. Null models in network neuroscience. *Nat Rev Neurosci*. 2022;23(8):493-504.].

We hope that these comparisons provide sufficient detail for the reviewer and sincerely thank them again for providing the opportunity to articulate our rationale around these choices. At this time it is out of scope to compare and contrast these methodological variants to the present study, but we will strongly consider them in future work that builds on the present study. In particular, we have a study that is in-preparation that uses a variant on permutation-based importance, where surrogate, but non-shuffled, data is used in place of the network of interest. That is, we are replacing within- and between-network portions of task-FC data with rest-FC data, on a network-by-network basis. This allows a similar inference to network importance, but largely maintains connectome-level properties given the orthogonality of rest- and task-FC.

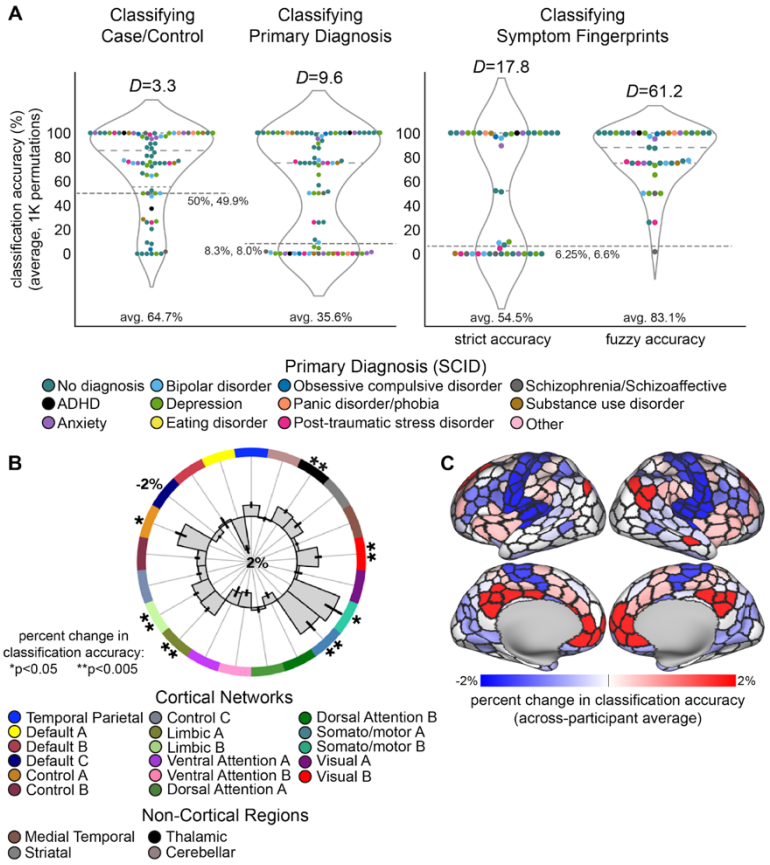
Comment 5: In Figure 4B, removing DMN component A seems to improve model performance, though not significantly. This result is unexpected; please comment on possible reasons and implications, even if the difference does not reach significance.

Response 5: We sincerely share the reviewer's interest in the default network's impact on the ability for network dynamics to classify symptom fingerprints. We have reproduced Fig. 4 below for reference, and point the reviewer to Fig. 4B-C. In particular, accuracy improves when DMN-A is removed, which suggests that, although this effect does not reach significance, the presence of DMN-A may be impeding the separability of decision boundaries with respect to classifying symptom fingerprints. It is likely that the default network has subfunctions depending on which component is being studied, with component A largely comprised of posterior cingulate and orbitofrontal regions. It is also noteworthy that research focused on the DMN has often observed unexpected results across studies [see here for an overview and a compelling synthesis of these disparate findings: Menon V. 20 years of the default network: A review and synthesis. Neuron. 2023;111(16):2469-2487.].

With that in mind, we would like to highlight an emerging line of evidence suggesting that the structural hub property of the DMN (i.e., dense connectedness across the system) allows it to prime the system such that it can more easily process information in cognitively non-demanding task contexts [see: Gu S, Pasqualetti F, Cieslak M, et al. Controllability of structural brain networks. Nat Commun. 2015;6:8414.]. This suggests a complementary (and possibly temporally ordered) push/pull between the operating principles of the DMN and cognitive control systems (e.g., frontoparietal network or FPN) with respect to moving brain network organization into and out of easy and difficult to reach states, depending on task context. Thus, while flexible processing across diverse task contexts may be clearly facilitated by cognitive control systems (supported by Fig. 4B-C), it is also possible that default systems promote flexibility indirectly (and in the opposing direction). Moreover, even if DMN processes are indirectly linked with cognitive control, these processes may not explain enough variance

in psychiatric symptom structure and appear to impede the model. We are hopeful that our work alongside other recent studies motivates future work to constrain this hypothesis space further by specifying the mechanisms by which various brain systems contribute to the altered brain network dynamics that underlie psychopathology. We have added the following text to the **Results** surrounding Fig. 4 to suggest a possible interpretation of DMN-A effects (lines 450-458):

> *"...This suggests that brain systems known to enable cognitive control processing (frontoparietal and thalamus) and networks linked with task-relevant resource processing (visual and somato/motor) exhibit dynamics that are prominently linked with dimensionally-organized cognitive and behavioral functioning across health and psychiatric disease. Interestingly, lesioning of default network A improved classification of symptom fingerprints (**Fig. 4C**), however this effect did not reach statistical significance. A possible interpretation that should be systematically tested in future studies is that default network processes are complementary to frontoparietal processes, with the default network promoting transitions into easier to reach neural states and frontoparietal processes switching the brain into cognitively demanding states[21,130]."*

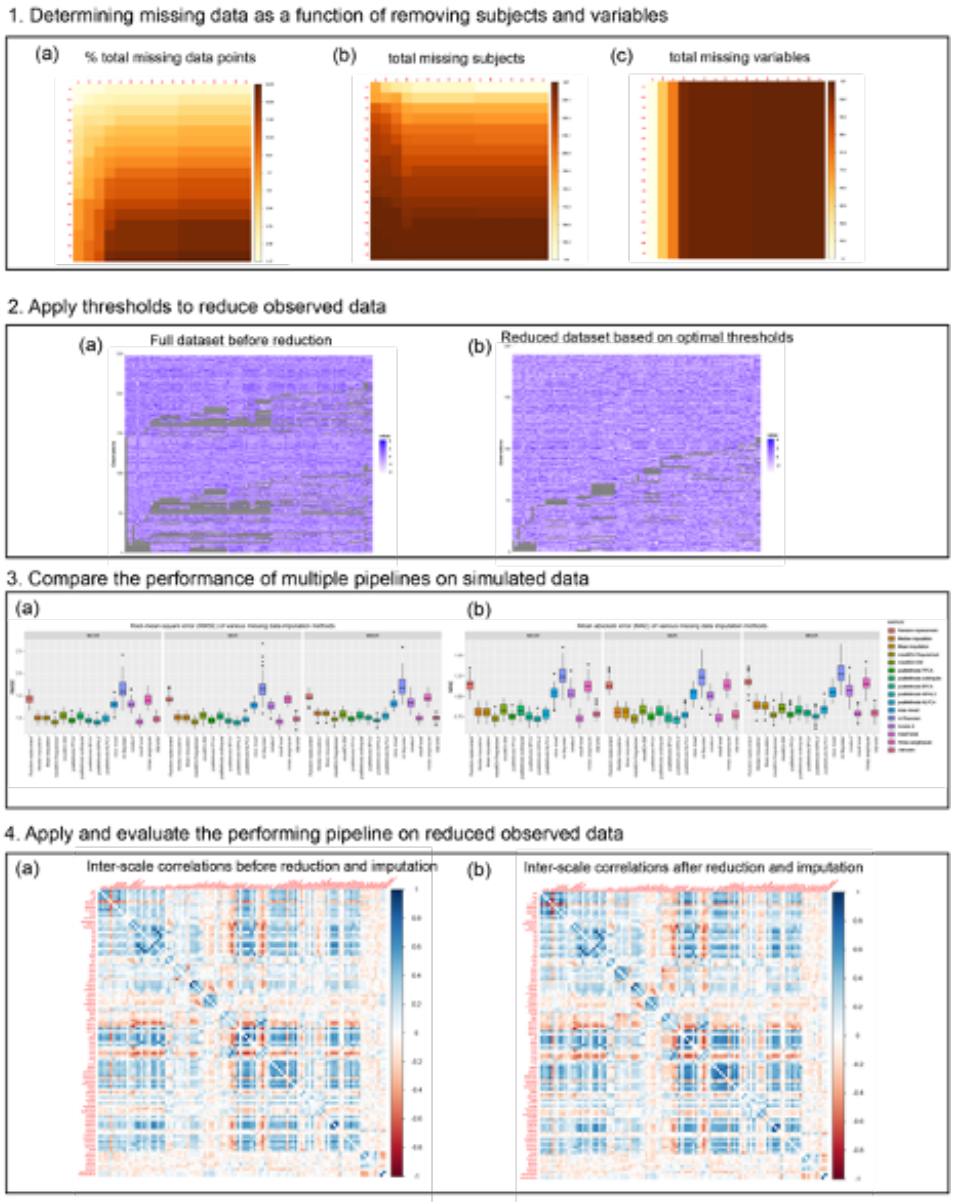


"Figure 4. Brain network dynamics classified dimensional symptom fingerprints better than case/control status and primary diagnoses."

Comment 6: Imputation was applied to missing behavioral data, but the overall missingness rate is unclear. Was a threshold for missingness set (e.g., 10–30%) beyond which variables were dropped?

Response 6: We apologize for the oversight in not including this information in the Methods and thank the reviewer for pointing this out so it can be added to the present manuscript. We recently compared how well different imputation approaches account for missing data in the TCP dataset using the R package missCompare, which was part of our publication here: [Chopra S*, Cocuzza CV*, Lawhead C, et al. The Transdiagnostic Connectome Project: an open dataset for studying brain-behavior relationships in psychiatry. Sci Data. 2025;12(1):923.]. The missCompare toolkit uses the underlying statistics of a given real dataset to guide comparisons on simulated data with matched properties, and then finds the optimal imputation algorithm. In conjunction, we also re-ran missCompare with varied missingness thresholds. Altogether we found the optimal

algorithm (missForest) and missingness threshold (between 20-25%), which we similarly applied in the present study and used a 25% threshold. We have reproduced the relevant Figure from that prior publication (Chopra, Cocuzza, et al., 2025, *Sci Data*) below for reference:



"Imputation pipeline for behavioral data prior to dimensionality reduction. *1) Optimal thresholds for removing participants and variables with excessive missing data prior to imputation were computed based on inflection points that retained the most data across total data points, participants, and variables, while reducing overall missingness. 2) This resulted removing participants with >20% and variables with >25% missing data. This resulted in 191 participants and 104 measures entering the imputation process. 3) Using the missCompare package, we compared the performance of 16 different imputation pipelines on 50 simulated version of the observed data under three conditions: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random or Non-Ignorable (MNAR). 4) The missForrest pipeline showed the best performance and was used to impute data in the reduced observed dataset. Distributions and correlations between scales were compared before and after imputation to ensure imputation quality." (Chopra, Cocuzza, et al., 2025)*

Relatedly, we followed best practices suggested in the missCompare documentation and pre-processed the behavioral data to reduce missingness present in the data downstream during imputation. We outlined these

steps in the **Methods** section on *Dimensional symptom analysis*, which also details the preprocessing steps necessary to fully curate and quality control TCP behavioral data.

We revised the **Methods** as follows (lines 1020-1027):

> *"...Following normalizing and binarizing transformations, select behavioral measures were unavailable in some participants. To account for this, we used random forest imputation (also termed "miss forest")[173] based on prior evidence that it is robustly applicable to nonparametric and/or mixed data types. Random forest imputation is a supervised learning algorithm that we performed on the discovery dataset; where 80% of participants were randomly allocated to the training set and 20% to the testing set. Additionally, following prior quality control analyses on the TCP dataset (see [116]), we set a missingness threshold of 25%, where if 25% of data points were missing from a given measure, then it was not considered further in analyses.* ***Supplemental Table S1*** *lists the final set of behavioral measures."*

Comment 7: Figure 5A compares effect sizes among states for predicting dimensional symptom profiles. Without confidence intervals, however, it is unclear if differences are meaningful. If this is an exploratory analysis, please note this.

Response 7: We are thankful to the reviewer for bringing this to our attention and agree that further statistics are needed for clarifying how to interpret these results. This was also partially brought up by reviewer #1 (comment #2), who highlighted the need for further context on how the pattern of results in Fig. 5 were not fit to noise. In that response (see above, reviewer #1, comment #2), we detail how the nonparametric null modeling approach (with revisions including relevant statistics, as quoted below, including new Table 2), plus various steps in the analysis pipeline to make models as comparable as possible, account for outcome measure noisiness to a reasonable extent. We also added text to emphasize that this analysis in particular was exploratory and thank the reviewer for their helpful feedback to support this framing for relevant portions of the paper. All revisions relevant to both comments are quoted below for ease of reference. Altogether we believe these related lines of feedback have greatly clarified the model-comparison analyses suggesting brain network dynamics classify case/control status versus diagnosis group versus dimensional symptom profiles with differential sensitivity on the basis of cognitive context and/or amount of data included (Fig. 5). We also suggest future strategies to disentangle these possibilities further in the Discussion (see quoted revisions below).

Revised text in the relevant **Results** section (lines 498-549):

> *"...To test how influential each cognitive state was to the link between brain network dynamics and symptom fingerprints, we implemented an exploratory analysis where NMF was run with varied combinations of input connectivity patterns. This allowed us to differentiate context-specific shifts in connectivity from general rest-to-task shifts in connectivity (as in* ***Fig. 4****)…*
> *…*
> *... Across all models, symptom fingerprints were classified with greater effect size than case/control status and primary diagnostic group (****Fig. 5A****). Further, models with the same outcome measure exhibited a reasonably similar range of effect sizes, suggesting internal consistency across similarly-posed predictive questions. For example, the effect size range for case/control models was 2.32 < D < 3.35, whereas the range for symptom fingerprint models was 17.25 < D < 18.85 (****Fig. 5A****). These ranges were further consistent with each set of comparison models' respective null model (as in* ***Fig. 4,*** *with each null model matching the varied predictors in* ***Fig. 5A****), which is a robust, data-driven way to infer statistical significance with complex, potentially noisy data[109] (see Discussion for further considerations). As in the full model (****Fig. 4****), all model accuracies were submitted to significance testing against empirical-chance levels, which were based on nonparametric null models, with outcome measure labels permuted over 1000 iterations. Empirically-based chance levels were similar to the full model (****Fig. 4****), as follows: classifying case/control status: 49.77% chance; primary diagnosis group: 8.3% chance; symptom fingerprints: 6.1% chance. All variants upon network dynamics were able to classify all outcome measures significantly above empirical chance, but with varying effect size (****Fig. 5A, Table 2****). As expected, there was a broadly consistent pattern of model performances between strict-accuracy and fuzzy-accuracy (****Fig. 5A****) when classifying fingerprints. The all-states model exhibited the lowest*

performance (strict accuracy: D=17.25; fuzzy accuracy: D=54.73), suggesting that network dynamics across wider-spanning task contexts may yield suboptimal decision boundaries separating clinical symptom structures expressed across individuals. The rest-and-Stroop model performed the best (strict accuracy: D=18.85; fuzzy accuracy: D=62.91), suggesting that connectivity shifts between an intrinsic state and a cognitive control state accounted for individual clinical variability particularly well. Importantly, in symptom fingerprint models, classification accuracy did not simply increase with the amount of brain network data inputted to NMF, but instead increased when the Stroop task was present. This supports the proposal that brain network dynamics underlying shifts into more cognitively demanding task contexts are robust predictors of dimensional symptom structure across psychiatric illnesses. Interestingly, case/control and primary diagnosis were classified with a similar pattern across models. Case/control and primary diagnosis characterizations may involve a shared underlying relationship with symptomatology given that primary diagnosis can be thought of as an expanded version of the binary case/control status. However, the exact extent that each model fully accounts for noisy fluctuations in symptomatology is an important consideration when comparing these two patterns of results (see Discussion). With this consideration, the main difference was that all-states performed worst and best in classifying primary diagnosis and case/control statuses, respectively. These data suggest that separability of case/control status may be sensitive to the number of cognitive states (or, the amount of data) used to uncover network dynamics. Primary diagnoses may be classified best by dynamics across similar cognitive contexts. Conversely, dimensional symptom profiles may be separated best by a specific, foundational shift in cognitive context."

"Table 2. Significance testing: classification model comparison when various state-to-state brain network reconfiguration dynamics were used as predictors.

Predictors: states	***Outcome: case / control status***	***Outcome: primary diagnosis group***	***Outcome: fingerprint (strict)***	***Outcome: fingerprint (fuzzy)***
All states	$t=10.8, p=5.8x10^{-14}$	$t=5.2, p=3.1x10^{-6}$	$t=7.0, p=7.4x10^{-9}$	$t=22.2, p=4.2x10^{-25}$
Rest only	$t=8.7, p=3.1x10^{-11}$	$t=5.4, p=1.6x10^{-6}$	$t=7.5, p=1.4x10^{-9}$	$t=23.6, p=3.5x10^{-26}$
Task only	$t=8.2, p=1.5x10^{-10}$	$t=5.3, p=1.9x10^{-6}$	$t=7.4, p=2.1x10^{-9}$	$t=24.4, p=1.1x10^{-26}$
Rest & Stroop	$t=8.1, p=1.8x10^{-10}$	$t=5.3, p=2.0x10^{-6}$	$t=7.6, p=9.3x10^{-10}$	$t=25.5, p=1.9x10^{-26}$
Rest & Emo.	$t=7.5, p=1.6x10^{-9}$	$t=5.3, p=2.2x10^{-6}$	$t=7.2, p=4.0x10^{-9}$	$t=23.0, p=9.5x10^{-27}$
Stroop only	$t=8.0, p=2.7x10^{-10}$	$t=5.2, p=2.5x10^{-6}$	$t=7.3, p=2.4x10^{-9}$	$t=23.9, p=2.4x10^{-26}$
Emo. only	$t=7.9, p=4.6x10^{-10}$	$t=5.2, p=2.4x10^{-6}$	$t=7.1, p=6.0x10^{-9}$	$t=22.6, p=2.1x10^{-25}$

"

Revised text in the **Discussion** (lines 669-688):

*"...To make non-stationarity inferences, we instead opted for a model comparison approach (**Fig. 5**), where connectivity patterns relevant to varied cognitive states were considered to compare reconfiguring network interactions across resting-state-alone, rest-to-Stroop, task-alone, and so forth. The present study advances clinical neuroscience by bypassing undue influence from non-neural sources of variance, however, we encourage future work that systematically compares how well other brain network dynamics methods can discriminate the structure of psychiatric symptoms.*

Relatedly, while predictors (brain network dynamics) were held constant in all model comparison analyses, the outcome measures studied may exhibit differing levels of inherent noisiness. The strategies we used to account for the possibility that classification models were overly fit to such noise included: comparing effect sizes instead of raw accuracies, given differing chance levels; using common sets of predictors across all models; and a strict scheme of splitting data into discovery, test, and unseen validation sets commonly across all models. Importantly, for all models compared, we also used a non-parametric and data-driven approach to developing null models that has shown to be rigorous for fMRI-based[109], brain connectivity data[21] where the statistical ground truth is not known. While this multi-pronged approach is robust, it is important to note that we cannot precisely quantify how much noise influenced model fit across the differing models. It will be an important methods advancement for future studies to use simulated data – where ground truth properties are known and varied by the researcher – to precisely establish signal and noise influences upon model fit."

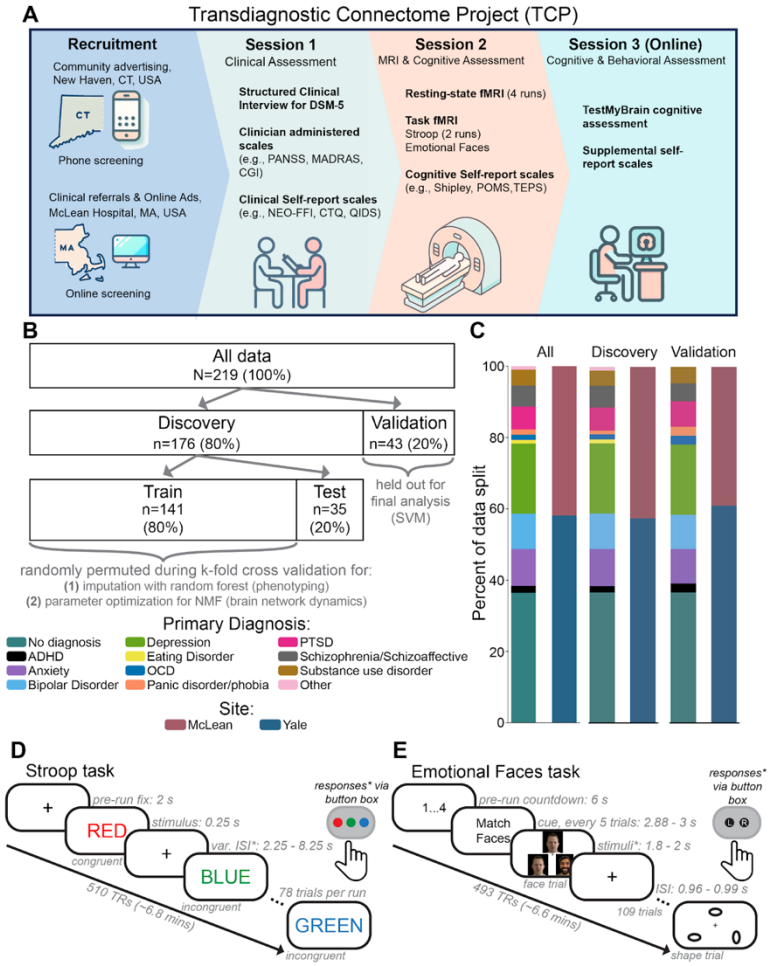
We would additionally like to remark on confidence intervals (CIs) for the purposes of these analyses. Adding this statistic has the potential to be misleading, so we have erred on the side of emphasizing nonparametric null modeling and emphasizing that model comparisons were exploratory. To explain further: with low sample sizes, there is inherently more sampling error uncertainty around the true effect, and we expect the effect size to be less precise. However, Cohen's D is still useful as it is an unbiased estimate. In response to the reviewer's insightful comments #10 #11 below, we also have added text to the **Discussion** on the need for future work to use expanded data or validation datasets, which is relevant insofar as relatively low sample sizes are a constraint upon interpreting CIs (in Fig. 5 analyses) CIs around Cohen's D.

Relevant **Discussion** revisions (lines 646-655):

> *"…For example, a key hypothesis would be the extent that network dynamics account for symptom structure is determined by a breakdown in the efficiency of information processing across varied task contexts. It will also be essential to leverage the observations of our exploratory analyses (**Fig. 5**) as important constraints on follow-up hypotheses in future work. For example, given the general impact of cognitive control processing, how do dynamic repertoires involving inhibitory control, working memory load, or task switching (to name a select few) specifically impact individual differences in symptom structure? Importantly, future studies should use replication datasets that enable such research questions and allow generalization of the present findings to independently sampled cohorts and task contexts. This would also have the benefit of increased sample sizes, which is particularly important for analyses that may be inherently exploratory."*

Comment 8: As MRI data was collected between Yale and McLean, could the authors discuss steps taken to ensure the data was harmonized? Was there a need for ComBat harmonization?

Response 8: We thank the reviewer for this question as it is an essential consideration for studies with multi-site data collection. We have done the following to promote harmonization throughout the study: (1) Detailed coordination between sites and personnel, where all standard operating procedures and protocols (i.e., everything from recruitment through scanning) were designed to be as closely matched as possible. In particular, our neuroimaging acquisition protocols were precisely matched between scanners. (2) Processing of both behavioral and neuroimaging data was done identically and with consistent versioning within computational environments. (3) When considering data leakage [see: Rosenblatt M, Tejavibulya L, Jiang R, Noble S, Scheinost D. Data leakage inflates prediction performance in connectome-based machine learning models. Nat Commun. 2024;15(1):1829.], we split data into discovery (train, test), and validation sets (this is detailed in the manuscript in the Methods section *Data splitting scheme*). When doing this, we ensured that two aspects of the full dataset were reflected in the discovery and validation sets; this included the proportions of participants recruited at Yale and Mclean sites, as well as the proportion of participants in each primary diagnosis group. This is depicted in **Supplemental Figure** S7C, reproduced below:



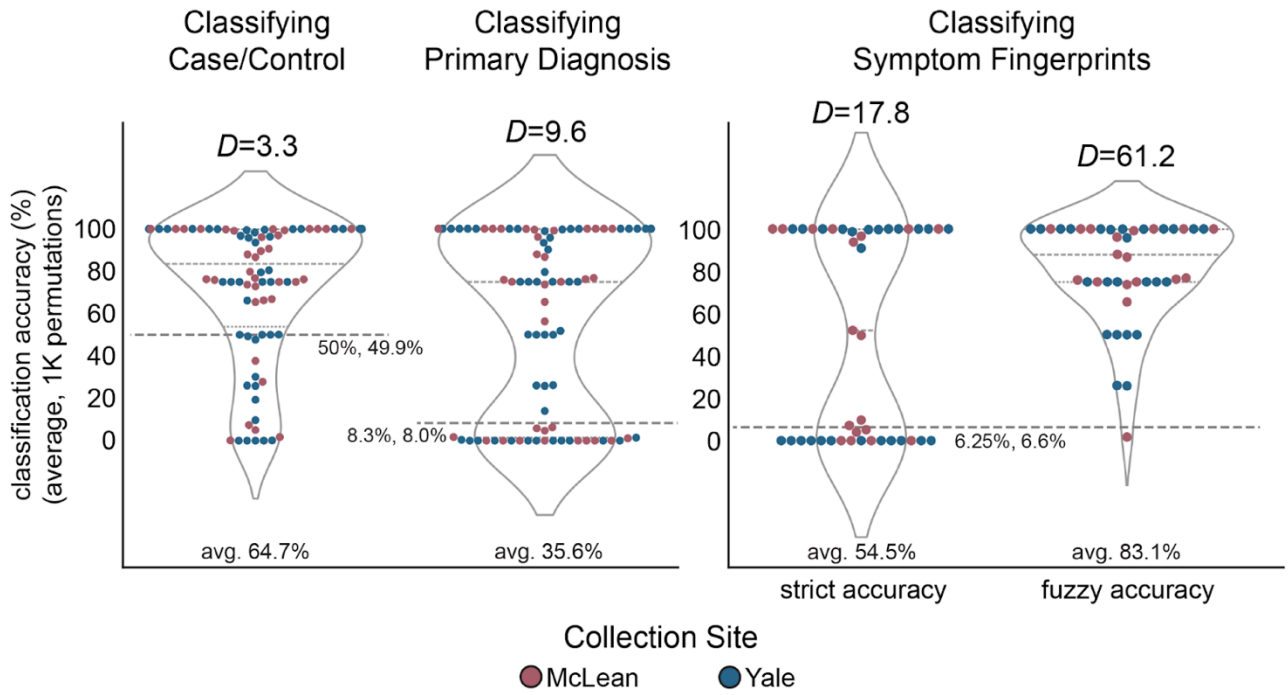
"Figure S7. The Transdiagnostic Connectome Project (TCP) dataset."

Additionally, we checked the extent that participants from each collection site were classified above chance in each of our main models. This has been added to the Results with a corresponding Supplemental Figure, reproduced below. All models exhibited approximately equal or representative percentages of participants from each collection site who were classified above chance, which suggests that model performance was not driven by site. However, there is still a possibility that collection site differences that were infeasible to control for (e.g., duration of illness), particularly with a smaller sample size, will have hidden impacts on model fit. We have revised the Discussion to include this consideration and suggest the use of harmonization methods, such as ComBat, in future work; particularly future studies that will involve replication datasets (or otherwise increased sample sizes).

Results now report that equal proportions of Yale and McLean participants were classified above chance (lines 430-440):

> *"...This suggests that the boundaries between groups of individuals exhibiting symptom fingerprints are separated more precisely by brain network dynamics than case/control status or primary diagnosis categories. Further, in all models, above-chance accuracy was exhibited in representative or otherwise equal proportions by participants recruited at Yale and those recruited at McLean (**Supplemental Fig. S5**). Brain network dynamics classified symptom fingerprints as follows: 50% of participants with above-chance strict accuracy from Yale and 50% from McLean; fuzzy accuracy was split 58.1% Yale and 41.9% McLean. Similarly, 54.1% from Yale and 45.9% from McLean were classified above-chance in the case/control model; and 58.3% Yale and 41.7% McLean in the primary diagnosis model. Given that the TCP dataset studied here included 57.1% participants recruited at Yale and 42.9% from McLean, the similarly exhibited percentages across models suggest that classification was not solely driven by collection site differences."*

Supplemental Figure S5 to support **Results** text above, where main text Fig. 4 – which shows the primary result of the present study: how well the full model of brain network dynamics classifies a transdiagnostic psychiatric cohort – is reproduced, but now color-coded by collection site:



"Figure S5. Model fit was proportional across data collection sites. *Quality control analyses demonstrating that collection site did not influence the brain-behavior relationship between brain network dynamics and psychiatric grouping. Corresponding to Results Fig. 4, the primary classification model results were re-colored by collection site (red: McLean, blue: Yale). Given no apparent pattern or clustered distribution of classification accuracy of McLean versus Yale participants, this suggests that collection site did not drive model performances (see main Results text for further details)."*

Discussion revisions to suggest that future work consider further harmonization techniques and/or tools (lines 693-703):

*"...Similarly, we do not know the ground truth on precisely how much collection site differences impact model performance. We ensured that data collection was equivalent across Yale and McLean sites, particularly with standardized neuroimaging acquisition and processing protocols[116]. We likewise held the proportion of participants recruited at each site consistent when splitting data into discovery and validation sets (**Supplemental Fig. S7**). While these steps appeared to reasonably control for the influence of collection site given that representative proportions of Yale and McLean participants were classified above-chance (**Supplemental Fig. S5**), there is still a possibility that hidden sources of variability between sites influenced model fit. In future work where collection site differences are more likely to contaminate model inferences, or otherwise multiple datasets are being assessed for generalizability, we encourage the use of additional harmonization techniques[157], such as ComBat[158] or DeepResBat[159]."*

Comment 9: There does not appear to be a table outlining cohort characteristics. Relatedly, a summary of the proportion of diagnoses in the transdiagnostic sample would be useful.

Response 9: We are grateful to the reviewer for pointing this out and believe the transdiagnostic representation in the TCP cohort is much clearer after adding these details. We have added the following text and table (Table 1) to the relevant **Methods** section (lines 776-793):

*"...Of this N=219 in the present study (mean age=36.2, SD=12.9 years; n=93 (42.5%) identified as male, n=123 (56.2%) as female, and n=3 (1.4%) as nonbinary), there were n=134 (61.2%) with psychiatric diagnoses, n=85 (38.8%) without a history of psychiatric diagnoses or treatment (**Supplemental Fig.***

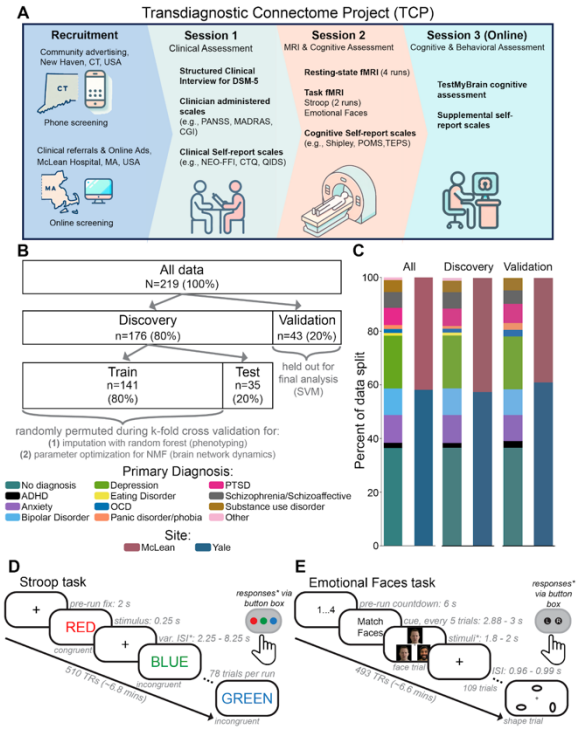
***S7C**; and see sections below for further details on MRI data acquisition and brain network dynamics analyses). In addition to the details available in [116], **Table 1** includes an overview of important demographic characteristics of the TCP dataset used in the present study. **Supplemental Fig. S7C** shows the proportion of participants with each primary diagnosis (as given by the SCID-V-RV; as well as those without a psychiatric diagnosis) and from each collection site in the total, discovery, and validation datasets (see section below on data splitting scheme for details on discovery and validation sets). Across the entire dataset in the present study, primary diagnoses were distributed as follows: no diagnosis: n=85 (38.9%); ADHD: n=4 (1.8%); anxiety: n=21 (9.6%); bipolar disorder: n=22 (10.0%); depression: n=42 (19.2%); eating disorder: n=2 (0.9%); obsessive compulsive disorder: n=3 (1.4%); panic disorder/phobia: n=3 (1.4%); posttraumatic stress disorder: n=14 (6.4%); schizophrenia/schizoaffective: n=12 (5.5%); substance use disorder: n=9 (4.1%); other/rare: n=2 (0.9%).”*

“Table 1. Demographics of TCP dataset.

Characteristic	***Yale (n=125)*** *n (% of site)*	***McLean (n=94)*** *n (% of site)*	***With diagnosis (n=134)*** *n (% of group)*	***Without diagnosis (n=85)*** *n (% of group)*
Ethnicity				
Hispanic or Latino	*16 (12.8%)*	*10 (10.6%)*	*16 (11.9%)*	*10 (11.8%)*
Not Hispanic or Latino	*109 (87.2%)*	*84 (89.4%)*	*118 (88.1%)*	*75 (88.2%)*
Race				
Asian	*19 (15.2%)*	*9 (9.6%)*	*12 (9.0%)*	*16 (20.3%)*
Black or African American	*24 (19.2%)*	*11 (11.7%)*	*14 (10.4%)*	*20 (25.3%)*
More than one race	*7 (5.6%)*	*6 (6.4%)*	*10 (7.5%)*	*3 (3.8%)*
Prefer not to answer	*6 (4.8%)*	*0*	*4 (3.0%)*	*2 (2.5%)*
White or Caucasian	*69 (55.2%)*	*68 (72.3%)*	*94 (70.1%)*	*38 (48.1%)*
Age				
18-24	*14 (11.2%)*	*8 (8.5%)*	*11 (8.2%)*	*11 (12.9%)*
25-34	*56 (44.8%)*	*41 (43.6%)*	*61 (45.6%)*	*36 (42.4%)*
35-44	*22 (17.6%)*	*15 (16%)*	*20 (14.9%)*	*17 (20%)*
45-54	*13 (10.4%)*	*14 (14.9%)*	*18 (13.4%)*	*9 (10.6%)*
55-64	*15 (12%)*	*10 (10.6%)*	*18 (13.4%)*	*7 (8.2%)*
65+	*4 (3.2%)*	*6 (6.4%)*	*5 (3.7%)*	*5 (5.9%)*
Prefer not to answer	*1 (0.8%)*	*0*	*1 (0.7%)*	*0*
Sex				
Female	*72 (57.6%)*	*51 (54.3%)*	*76 (56.7%)*	*47 (55.3%)*
Male	*53 (42.4%)*	*40 (42.6%)*	*55 (41.0%)*	*38 (44.7%)*
Nonbinary	*0*	*3 (3.2%)*	*3 (2.2%)*	*0*

”

Relatedly we are copying **Supplemental Figure 7** below for reference, and refer the reviewer to panel C, which complements the information added to the Methods by showing the proportion of participants in each diagnostic group, when split into discovery and validation subsets of data; as well as breaking this down by collection site.



"Figure S7. The Transdiagnostic Connectome Project (TCP) dataset"

Comment 10: Dimensional symptom profiles are defined within this modestly sized sample and not with respect to a normative sample; thus, generalizability to other datasets may be limited. Please discuss potential limitations stemming from this.

Response 10: We agree with the reviewer that this is a key consideration for the present study, and equally as important, for studies that build on these findings in the future. We apologize these points were not fully considered and thank the reviewer for the opportunity to add this important context. Indeed, normative modeling is a powerful approach that we support but is challenging without similar data collected for a large sample of individuals without psychiatric diagnoses; in particular, the densely sampled behavioral, cognitive, and clinical measures in the TCP dataset. We of course hope these data types become available in the future, but in the meantime, we have suggested future work use a replication approach with similar-but-different transdiagnostic datasets in the **Discussion**, as follows (lines 609-655) (note: combined with considerations from comment #11 below):

> *"...This supports our hypothesis that brain dynamics contain more process-relevant information than static network features and thus confer an enhanced separation of the boundaries between diagnostic groups (**Fig. 1**). These findings also serve as a critical proof-of-concept that parameterizing brain network dynamics with respect to across-state shifts in connectivity patterns is a robust, readily built-upon approach to developing clinically-relevant brain-behavior predictive models.*
>
> *…*
>
> *… For example, a key hypothesis would be the extent that network dynamics account for symptom structure is determined by a breakdown in the efficiency of information processing across varied task contexts. It will also be essential to leverage the observations of our exploratory analyses (**Fig. 5**) as important constraints on follow-up hypotheses in future work. For example, given the general impact of cognitive control processing, how do dynamic repertoires involving inhibitory control, working memory load, or task switching (to name a select few) specifically impact individual differences in symptom structure? Importantly, future studies should use replication datasets that enable such research questions and allow generalization of the present findings to independently sampled cohorts and task contexts. This would also have the benefit of increased sample sizes, which is particularly important for analyses that may be inherently exploratory."*

Comment 11: A big-picture consideration is the modest sample size used for the analyses. The study includes 219 participants (with diagnosis n=134 and without n=85) which is admirable for a neuroimaging study but may suggest a need for minor reframing. For example, the identification of symptom dimensions was done using hierarchical clustering of 110 behavioral measures. This is often done using exploratory factor analysis where a rule of thumb requirement is to have 5-10 observations per measure included. Here, however there are about 2 observations per measure which raises some concern about the robustness of estimating these dimensions. This does not detract from the novelty of this valuable paper; however, I suggest that much of it should be framed as exploratory or proof of concept with sample size constraints outlined as limitations.

Response 11: We thank the reviewer for their insights, and we agree that our inferences were greatly clarified when further emphasizing which analyses were exploratory. In particular, our main model findings in Fig. 4A were hypothesis based and follow-up analyses in Fig. 4B-C and Fig. 5 were exploratory. We have added text (quoted below) to clarify this and thank the reviewer for the opportunity to improve communications around our findings. Please note that these revisions partly overlap with responses to reviewer #1.

Introduction revisions to clarify which analyses were hypothesis driven and which were exploratory (lines 138-148):

> *"...Given the importance of brain network dynamics to adaptive information processing, we hypothesized that reconfiguring connectivity patterns across a varied set of cognitive states would accurately separate individuals based on dimensionally-organized symptom structure[3,53,58,61] (**Fig. 1**), as well as by their case/control status and primary diagnostic labels. While prior work suggests that impaired cognitive control processing is a transdiagnostic, commonly-exhibited deficit across psychiatric illnesses[46-49,86], the extent that dimensional symptom structures would be separable by specific shifts in cognitive context was unclear. Thus, we implemented follow up analyses that used a systematic model comparison approach to explore the extent symptom structure was accurately separated by network dynamics underlying specific shifts in cognitive state. Specifically, we set out to explore whether classification accuracy would differ in a variety of contexts: (1) with more brain network data included in the predictor set, regardless of cognitive state; (2) in the presence of resting-state network dynamics; (3) task-state dynamics (i.e., without resting-state), regardless of task context; (4) via a specific shift in task context, such as increased cognitive demand, indicative of control processing; and (5) by varying subsets of brain network dynamics included as model predictors."*

Further, we have added this sentence to the end of the **Introduction** (lines 157-159):

> *"...These analyses reflect an essential, proof-of-concept roadmap for framing future experiments that systematically identify the extent that brain network dynamics confer transdiagnostic symptom structures in psychiatric illness."*

Relevant **Results** revisions (lines 498-508):

> *"...To test how influential each cognitive state was to the link between brain network dynamics and symptom fingerprints, we implemented an exploratory analysis where NMF was run with varied combinations of input connectivity patterns. This allowed us to differentiate context-specific shifts in connectivity from general rest-to-task shifts in connectivity (as in **Fig. 4**)."*

Additionally, we thank the reviewer for pointing out the typical rule of thumb for exploratory factor analyses and agree that hierarchical clustering was thus a more appropriate choice for our dataset. In hierarchical clustering, a more serious concern is having more variables than observations, which is not applicable to the TCP dataset. Relatedly, we used an implementation of hierarchical clustering that assesses distances between data points or vectors (i.e., a distance matrix). We used correlations (rho) between all pairs of behavioral measures, where correlations were across vectors of participant data (i.e., processed behavioral data). Thus our input to hierarchical clustering was a 110 x 110 distance matrix, what we term individual differences correlations. In the step where this distance matrix is estimated, it is important that there are more observations than variables

(although we are not aware of any peer-reviewed studies that indicate a specific ratio for hierarchical clustering, as the reviewer indicted for factor analysis) to avoid spurious or otherwise noisy correlations. We have updated the **Methods** to emphasize these details and apologize that it was not clear (lines 1048-1051).

> *"...Additionally, while select prior work on dimensionally-organized symptomatology has used similar methods such as factor analysis[80], we opted for hierarchical clustering on a distance matrix then applying PCA[176] because it has less strict requirements about number of observations per variable. These methods have convergent inferences, which importantly includes data-driven symptom structures."*

Lastly, we fully agree that the paper can reasonably be discussed as a proof of concept, and N=219 is strong for an fMRI study, particularly when transdiagnostic datasets are limited (i.e., complexities with respect to motivating funding bodies). The derived symptom fingerprints were data-driven, which is a core principle to many proposals advocating for dimensional phenotyping in psychiatry (see: [Forbes MK, Baillie A, Batterham PJ, et al. Reconstructing psychopathology: A data-driven reorganization of the symptoms in the *Diagnostic and Statistical Manual of Mental Disorders*. *Clin Psychol Sci*. Published online October 17, 2024.], [Ruggero CJ, Kotov R, Hopwood CJ, et al. Integrating the Hierarchical Taxonomy of Psychopathology (HiTOP) into clinical practice. *J Consult Clin Psychol*. 2019;87(12):1069-1084.], and [Sanislow CA, Ferrante M, Pacheco J, Rudorfer MV, Morris SE. Advancing Translational Research Using NIMH Research Domain Criteria and Computational Methods. *Neuron*. 2019;101(5):779-782].). However, we agree that sampling variability is a key consideration and necessitates follow up studies to replicate these findings in larger cohorts or with replication datasets. We have added this context and related considerations to the **Discussion** in the appropriate places, as follows:

Discussion revisions (lines 715-720); this revision was partly motivated by a response to reviewer #1, comment #4 above:

> *"...Additionally, we focused on results for subgraph 1, which exhibited the most network efficiency[161] and system segregation[162,163]. We reasoned that, given all subgraphs 1-5 had highly similar topology (measured via Mantel r), one subgraph is likely a reasonable approximation of the other subgraphs, particularly as this study was partly an exploratory proof-of-concept (e.g., results in **Fig. 5**). We thus chose subgraph 1 given that it exhibited two network properties known to efficiently support shifts in information processing."*

Further **Discussion** revisions (lines 609-655) (note: combined with considerations from comment #10):

> *"...This supports our hypothesis that brain dynamics contain more process-relevant information than static network features and thus confer an enhanced separation of the boundaries between diagnostic groups (**Fig. 1**). These findings also serve as a critical proof-of-concept that parameterizing brain network dynamics with respect to across-state shifts in connectivity patterns is a robust, readily built-upon approach to developing clinically-relevant brain-behavior predictive models. …*
> *…*
> *…For example, a key hypothesis would be the extent that network dynamics account for symptom structure is determined by a breakdown in the efficiency of information processing across varied task contexts. It will also be essential to leverage the observations of our exploratory analyses (**Fig. 5**) as important constraints on follow-up hypotheses in future work. For example, given the general impact of cognitive control processing, how do dynamic repertoires involving inhibitory control, working memory load, or task switching (to name a select few) specifically impact individual differences in symptom structure? Importantly, future studies should use replication datasets that enable such research questions and allow generalization of the present findings to independently sampled cohorts and task contexts. This would also have the benefit of increased sample sizes, which is particularly important for analyses that may be inherently exploratory."*

Reviewer #2, Minor Comments:

Minor Comment 1: Please define the abbreviation TCP dataset at its first use.

Response Minor 1: We thank the reviewer for catching this error. The TCP acronym is now defined as follows on line 324:

> *"...First, by using the Transdiagnostic Connectome Project (TCP) dataset[110], which includes participants with and without varied mental health concerns and diagnoses."*

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response: We would like to extend our gratitude to this reviewer for providing vital assistance to another reviewer, and also to Nature Communications for providing such a valuable training opportunity to early career researchers.