Jeremy R. Manning
Dartmouth College
Department of Psychological & Brain Sciences
HB 6207 Moore Hall
Hanover, NH 03755

December 25, 2023

To the editors of *PNAS*:

We have enclosed our revised manuscript entitled *High-level cognition is supported by information-rich but compressible brain activity patterns* (manuscript tracking number: 2023-07312). We appreciate the reviewers' insightful comments on our original submission. We provide detailed point-by-point responses to each of the reviewers' comments on the following pages. The reviewers' comments are italicized and our responses are in **bold**.

Before providing those point-by-point responses, we wish to first highlight two broad concerns raised by the reviewers. Reviewer 1 raised several concerns about our statistical methods. We have addressed each concern by either providing additional information or by updating our approach, as appropriate. Reviewers 2 and 3 also raised questions about whether our measures of "informativeness" (i.e., peak decoding accuracy) and "compressibility" (i.e., number of components required to achieve at least 5% accuracy) are confounded. We have provided new explanations and analyses to show that these measures are indeed distinguishable, and that they get at different aspects of the data. One illustration of this is provided in our revised Figure 1, where we show four example synthetic datasets that reflect each combination of high and low informativeness, and high and low compressibility. We have also updated the relevant text to better explain and clarify these concepts.

Thank you for considering our revised manuscript.

Sincerely,

Jeremy R. Manning
Jeremy.R.Manning@Dartmouth.edu

Reviewer #1:

In this manuscript, Owen and Manning analyzed an existing fMRI dataset to assess the relationship between information content and compressibility in functional brain networks. Four fMRI conditions were compared; participants had listened to an intact story, a paragraph-scrambled version of the story, a word-scrambled version of the story, or laid quietly in the scanner. By using a dimensionality reduction approach, the authors found that the brain networks of participants in the intact story condition could be represented with fewer components and decoded with higher accuracy than the brain networks of participants in the other three conditions. They also investigated the contributions of known functional subnetworks of the brain to differences in decoding accuracy and compression across conditions.

I found the paper well-written and very interesting to read. Overall, the analyses are comprehensive and clearly presented. The results presented in Figure 2 are particularly striking. There are some areas in which additional details and statistical analysis should be presented to fully justify the conclusions drawn by the authors. Please see my specific comments below:

Major comments:
1. The Methods section is missing information about the statistical tests used to compare decoding accuracy across conditions (as in Fig. 2) and across time (as in Fig. 3). It appears that pairwise t-tests were performed-were these one-tailed or two-tailed tests? Were the reported p-values corrected for multiple comparisons?

We report (p. 20) that, unless otherwise noted, all statistical tests (e.g., pairwise t-tests) are two-sided, all error bars and error ribbons denote bootstrap-estimated 95% confidence intervals (computed across participants), and all reported p-values are corrected for multiple comparisons using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995).

2. The authors state that "moving from lower-order networks to higher-order networks, we found that decoding accuracy tended to increase" (lines 177-178). Though there are clear visual trends present in Fig. 4D and E, it does not appear that this statement is supported by a statistical outcome.

We have added statistical tests of this claim (pp. 10–12):

> **"Moving from low-order networks to higher-order networks, we found that decoding accuracy tended to increase in the higher-level experimental conditions and decrease (slightly) in the lower-level experimental conditions (Fig. 4D, E; Spearman's rank correlation between decoding accuracy and network order: intact: rho = 0.362, p < 0.001; paragraph: rho = 0.441, p < 0.001; word: rho = -0.102, p = 0.007; rest: rho = -0.354, p < 0.001)."**

3. Figure 5 - I appreciate that the authors have condensed a large amount of information in this figure, but it is a bit difficult to read. Increasing the size of the figure would help, but I wonder whether using a broader color spectrum to represent the terms and potentially ordering/grouping the terms by classification (shown in Fig. 6) would also increase readability.

We've made several changes to Figure 5, as suggested. First, we have increased the size of the figure along with the relative size of the legend in the lower right (containing the bulk of the text). We have also incorporated a new color scheme to represent the terms according to the coarser "classification" labels we used in Figure 6, and we have also re-ordered the terms to match the category labels.

4. It's unclear to me how Figure 6E illustrates the finding described in the Discussion (lines 233-237). Across all four conditions, the slopes appear very close to 0. As far as I can tell, the analysis used to produce this portion of the figure is not described in the Methods section. The authors should provide details about the procedure used to generate this figure and the statistical analysis that justifies their conclusions.

It's true that the slopes are all quite close to 0. We have added a note to this effect to the paper (pp. 14–15), along with details of how this analysis was carried out (pp. 26–27), and a statistical test to justify our claim (p. 14). We have also added bootstrap-estimated error bars to the figure to convey the reliability of the effect.

To summarize our interpretation and claims, we see a (small) effect of top neural components in the "higher-order" conditions weighing more heavily on higher-level cognitive functions. To characterize this, we computed (separately for each of the top 5 components shown in Fig. 5) the average weights assigned to each of the 11 categories of cognitive function (cognitive control, language processing, memory, emotion, etc.; see Figs. 5 and 6, Tab. S1). We also used ChatGPT to rank the cognitive functions from low-order to high-order (Tab. S2). We then fit a line separately for each experimental condition (x-values: rank; y-values: weights). We found that in the intact and paragraph conditions, higher-order cognitive functions tended to be weighted more heavily, whereas the opposite was true for the lower-order word condition. We used *t*-tests to compare bootstrap-estimated distributions of slopes across the experimental conditions. We verified that the slopes were greater (i.e., more positive) for the intact and paragraph conditions than for either the word or rest conditions.

Given that the overall size of these effects are small, we chose not to emphasize them strongly. Rather, we see the effects as "potentially interesting" and perhaps worthy of further exploration in future work. It also makes some intuitive sense that when participants were engaged more strongly (in the more engaging experimental conditions), their dominant neural patterns reflected higher-level cognitive functions. In contrast, when participants were engaged less strongly (in the less engaging experimental conditions), their dominant neural patterns reflected lower-level cognitive functions.

Minor comments:
5. The Abstract should begin with a sentence or two that introduce the problem to be addressed and provide context for the work.

We have updated our abstract as suggested.

6. In the Results section, the justification for carrying out some of the analyses is "we wondered about x". Stating specific hypotheses for each set of analyses would be more compelling.

We have reworded the relevant text to instead motivate those analyses by more specific questions about the data.

7. Fig. 2 C and D would benefit from a scale bar so that the height of the peaks can be appreciated.

We have added scale bars to Figure 2, Panels C and D.

8. Figure 6A, C - suggest including a color scale bar for the correlations.

We have added color bars for the correlation matrices in Figures 6A, 6C, S5A, and S5C

Reviewer #2:

I did not see a significance statement; the concluding remarks read like a significance statement, but I think were a bit over-extended relative to the findings in the actual work (e.g., talk of the neural code doesn't seem very well connected with the analyses).

We have added a significance statement (p. 1) and re-worded the "Concluding remarks" section to better relate to our actual work.

In this manuscript, Owen & Manning analyze a previously collected fMRI dataset from Simony and colleagues, where 36 individuals listened to a podcast recording, listened to a paragraph or word scrambled version of the recording, or rested in the scanner. The authors began by reducing the whole brain voxel-level data from each participant and condition into a smaller set of 700 basis functions using a technique they had previously developed. They then examined to what extent timepoints of fMRI activity could be predicted in held out participants (which they equated with "informativeness"), and how the accuracy of this prediction differed with the number of principle components included ("compressibility"). The authors also explored how these factors varied over the course of the story, and the extent to which principle components were shared across conditions and correlated with cognitive processes as inferred from a Neurosynth analysis of the PCs. The authors demonstrate that prediction accuracy was highest for the intact narrative and lowest for the resting condition, and that the intact data reached 5% accuracy with fewer components, suggesting that both the informativeness

and compressibility of brain activity data is highest for conditions that allow for the deepest encoding. They show that these patterns differ somewhat over time, and are related to differences in the activation of PC networks connected to diverse cognitive processes.

This manuscript was well written, and does a nice job of merging together theories from different disciplines to provide a fresh interpretation of fMRI narrative decoding. However, I had some concerns regarding the extent to which informativeness and compressibility could really be separated in the current work (which I saw as the major novel aspect of the project). I discuss this and some other confound concerns in detail below.

Major concerns
- At face value, the analysis of compressibility seems like the most novel aspect of the current work, relative to past analyses in this and other similar naturalistic datasets using decoding based methods. The joint examination of informativeness and compressibility is also intriguing. However, the way compressibility was analyzed in this dataset (equated to the # of components needed to reach 5% accuracy in decoding) makes the results of this finding appear redundant. This method will be likely to confound peak decoding performance with compressibility - conditions that have higher decoding at peak will also be more likely to reach the 5% mark more quickly, given typical profiles to decoding performance. Indeed, this is demonstrated in plots like Fig. 2E which show that informativeness and compressibility are strongly associated with one another (especially if one ignores the resting condition, which will essentially be noise for this sort of time-point level analysis). So, it's not clear what we gain by adding in this compressibility analysis. One possibility would be to define compressibility as a normalized measure, in terms of the % information (of the total per condition) represented with each component. This would provide a less trivial relationship and might produce more interesting findings - as suggested by Fig. S1.

The relation between informativeness and compressibility is a core aspect of our paper. Although informativeness and compressibility *can* be related (e.g., as the reviewer notes, appears to be the case to some extent for our dataset), these factors are not *always* related. We have updated Figure 1, including an analysis of four new synthetic datasets, to unpack and further illustrate this point (see *Synthetic data*, Fig. 1, and new introductory text on pages 3–5).

In brief, informativeness ends up being related to how timepoint-specific each observed pattern is. In our simulations, we construct data with "high informativeness" by drawing each sample independently from the other timepoints' samples. We construct data with "low informativeness" by introducing strong autocorrelations into the data, which results in higher similarities (i.e., less "distinctiveness") across observations. In other words, informativeness relates to how observations change over *time*.

Compressibility is related to relations between *features*. We can construct compressible data (with high *or* low informativeness) by introducing covariance between different features. For example, if a given pair of features tends to exhibit similar values, then those features can

often be "compressed" into the same factor when we project the data into a lower-dimensional embedding space. We can also construct data with *low* compressibility, by drawing data from distributions with 0 off-diagonal covariance (i.e., where the covariance between any two different features is 0). In other words, compressibility relates to how *features* relate to each other.

As we show in Figure 1C and 1D, these two aspects of data (informativeness and compressibility) can be varied independently of each other. In particular, we show that we can construct datasets with all four combinations of high/low informativeness and high/low compressibility.

Our simulations are instructive in that they show how informativeness and compressibility can vary when their connection to the underlying representations we "care about" (i.e., stimulus-driven activity) is assumed. In other words, although we don't explicitly model a "stimulus" as part of our synthetic data generation procedure, our procedure results in activity patterns that are shared (by design) across participants.

In our real data, there is an additional layer of abstraction between the observed activity patterns and the underlying neural representations that we are most interested in. Rather than analyzing activity patterns directly (e.g., measuring autocorrelations across time or features), we use inter-subject functional correlations to specifically home in on *stimulus-driven* aspects of the data. Essentially, following the logic outlined in Simony et al., 2016 (Nature Comms), aspects of neural activity that are similar across people are very likely to be stimulus driven, since non-stimulus driven activity would not be expected to be time-locked to the stimulus in similar ways across people. Our use of across-participants temporal decoding as our measure of "informativeness" enables us to specifically characterize *stimulus-driven* informativeness, as opposed to (potentially ambiguous) timepoint "uniqueness" or other non stimulus-related aspects of the data.

We have also added some discussion about *why* the informativeness and compressibility of our brain patterns may change over time (pp. 15–17):

> **"Our explorations of informativeness and compressibility are related to a much broader literature on the correlational and causal structure of brain activity patterns and networks (Adachi et al., 2012; Bassett & Sporns, 2017; Brovelli et al., 2004; Bullmore & Sporns, 2009; Dhamala et al., 2008; Korzeniewska et al., 2008; Lynn & Bassett, 2021; Owen et al., 2021; Preti et al., 2017; Rogers et al., 2007; Rubinov & Sporns, 2010; Sizemore et al., 2018; Smith, Beckmann, et al., 2013; Smith, Vidaurre, et al., 2013; Sporns & Betzel, 2016; Sporns & Honey, 2006; Sporns & Zwi, 2004; Srinivasan et al., 2007; Tomasi & Volkow, 2011; Yeo et al., 2011). Correlations or causal associations between different brain regions simultaneously imply that full-brain activity patterns will be compressible and also that those activity patterns will contain redundancies. For example, the extent to which activity patterns at one**

brain area can be inferred or predicted from activity patterns at other areas (e.g., Owen et al., 2020; Scangos et al., 2021), reflects overlap in the information available in or represented by those brain areas. If brain patterns in one area are recoverable using brain patterns in another area, then a "signal" used to convey the activity patterns could be compressed by removing the recoverable activity. Predictable (and therefore redundant) brain activity patterns are also more robust to signal corruption. For example, even if the activity patterns at one region are unreadable or unreliable at a given moment, that unreliability could be compensated for by other regions' activity patterns that were predictive of the unreliable region. Whereas compressible brain patterns are robust to spatial signal corruption, high versus low informativeness reflects a similar (though dissociable; e.g., Fig. 1) tradeoff between expressiveness and robustness of *temporal* patterns. Highly informative brain patterns (by our measure; i.e., patterns that yield greater temporal decoding accuracy) are expressive about ongoing experiences or cognitive states, since each moment's pattern is reliably distinguishable from other moments' patterns. However, when each moment's pattern is unique, brain activity becomes less robust to temporal signal corruption. Our finding that brain activity patterns becomes more informative (i.e., less robust to temporal signal corruption) and compressible (i.e., more robust to spatial signal corruption) when cognitive engagement is higher suggests that our brain may optimize its activity patterns to prioritize either temporal or spatial robustness, according to task demands."

- The temporal analyses over the course of the story are interesting. But I think they are challenging to strongly interpret in terms of cognitive processes, given that only one story/one randomization order was tested. It seems likely to me that at least some of these time-varying effects will be confounded by specific aspects of the stimulus (e.g., informativeness of particular paragraphs/words, interest/arousal/attention evoked by particular portions of the story). A strong interpretation of these findings would require showing a replication with a new set of stories to be convincing.

This is a fair point. We (as the reviewer suggests) agree that at least some aspects of how decoding accuracy and compression change over time are likely to reflect stimulus-specific attributes. On the other hand we also show our own "replications" of sorts. Both cognitively rich conditions ("intact" and "paragraph"-- blue and green curves in Fig. 3) tend to show increases in decoding accuracy as the story progresses, whereas both less cognitively rich conditions ("word" and "rest"-- yellow and purple curves in Fig. 3) tend to show decreases in decoding accuracy as the story progresses. So there is some consistency across at least two stimuli of each type, and across the different participants who participated in those conditions.

These results also make some intuitive sense. As the contextual information available to participants increases (i.e., over time in the cognitively rich conditions), it makes sense that this might constrain neural responses to a greater extent. While this would not necessarily be the case for *every* story or stimulus, we suspect that it is generally the case that our

knowledge about what is happening in a story tends to increase as we experience more of it. And similarly, as participants are left to "mind wander" or as they experience mental fatigue (i.e., over time in the less cognitively rich conditions), it makes sense that this might lead to greater *variability* in neural responses across people, resulting in lower decoding accuracy. Again, it's not necessarily the case that *every* possible "unengaging" stimulus will lead to greater neural variability as time progresses, but it seems reasonably likely to happen for a variety of such stimuli.

To address this point, we have added a note to the text (p. 10) and we have also somewhat toned down our claims to make it clearer what we can vs. can't conclude from our work.

- The neurosynth analyses of the PCs in this dataset did not seem especially new relative to other works that have done extensive PCA/ICA analyses of task and rest fMRI data merged with tools like neurosynth and brainmap (e.g., Smith et al., 2009; Laird et al., 2011 and many others). What was striking to me is how correlated the top PCs were across conditions (e.g., Fig. 6C - I'm assuming that sign isn't very meaningful in considering these components), but the authors did not discuss this similarity in any detail. This said, the differences in activations across PCs was interesting (and I also liked the unique use of chatgpt in this context). It would be helpful to have some more text explaining how this neurosynth + PC approach is providing fundamentally similar or different information from the analysis across Yeo networks, aside from representing a different parcellation of the data.

We appreciate the pointers to prior work with neurosynth and brainmap, and we have added references to the suggested papers. To our reading, the Smith et al. (2009) and Laird et al. (2011) papers are focused on characterizing whether task "specific" networks are also identifiable and/or active at "rest." In that sense, we replicate one core aspect of those findings, in that we found that the top principal components from all four of the experimental conditions were highly similar (as the reviewer also notes). In other words, the dominant dimension of variation in neural activity was similar across all four conditions.

We also show that, across conditions varying in the degree of "cognitive engagement," we see some systematic differences in which networks are carrying temporally specific information (Fig. 4E) and/or how dominant components weigh on high-level vs. low-level cognitive functions (Figs. 5 and 6). (We note that we see the "Yeo network analyses" and the "neurosynth analyses" as fundamentally similar– we think they provide two different approaches to showing essentially the same basic phenomenon.) We have added a brief discussion of these points on pages 19–20.

- Regarding the analyses based on the Yeo networks: to what extent are the results shown in Fig. 5E confounded by differing numbers of features associated with each network? Is it fair to interpret comparisons across networks directly? It would be helpful to use a null model with randomized feature selection across networks to confirm this finding.

The results shown in Figure 5E do not seem to be confounded by differing numbers of features. To confirm this, we designed a permutation-based "control" analysis as the reviewer suggested.

Specifically, for each of *n* = 10 iterations, we randomly shuffled (without replacement) the network labels of the HTFA nodes, and then we re-ran our entire decoding analysis pipeline, including applying PCA with 3…m features for each condition (where m is the number of nodes in the given network), and then running 100 cross-validation runs of the decoding procedure for each condition and number of components. This resulted in 10 sets of shuffled data where each network had the same numbers of nodes, but where the decoding results no longer maintained the fidelity of each individual network. The correlation between network "order" and decoding accuracy broke down in the shuffled data, suggesting that the numbers of nodes in each network cannot account for the results shown in Fig. 5E (comparing bootstrap-estimated distributions of observed vs. shuffled correlations: Intact: $t(1998) = 276.431$, $p < 0.001$; Paragraph: $t(1998) = 330.334$, $p < 0.001$; Word: $t(1998) = -16.386$, $p < 0.001$; Rest: $t(1998) = -318.631$, $p < 0.001$).

We report these new results (p. 12) and describe the network permutation analysis in our revised methods section (p. 23).

- The HTFA analysis seems interesting, but relatively under-motivated in the current work. Why use this approach, rather than other approaches to compress the data (e.g., public parcellations like the Schaeffer or Gordon)? To what extent does this data reduction influence the compressibility and informativeness results?

HTFA is designed to provide a compressed representation of multi-subject brain data that preserves the original brain activations as closely as possible using a given number of spatially focal (Gaussian sphere) nodes. In our 2018 NeuroImage paper (Manning et al., 2018, NeuroImage), we showed that we can use HTFA to describe full-brain activity from the same dataset used in our current paper within a maximum of 0.25 standard deviations of each voxel's actual activity, taken across all voxels, images, and participants, using 700 nodes (the same representation we used in the current paper). Some of the explanatory power of HTFA comes from the fact that each node's explanatory power falls off *smoothly* with distance to its center. Intuitively, the result is a representation that looks like a lightly spatially smoothed version of the original data, but where the degree of smoothing varies across the brain according to how spatially autocorrelated the local activity patterns are.

Network parcellation approaches, like the Schaeffer or Gordon parcellations, are typically designed to segment "functional connectivity" (i.e., correlation) matrices, e.g., collected from large numbers of participants (often using resting state data). The goal is to split the correlation matrix into cohesive parcels that optimizes some objective function that varies across approaches. Some approaches to designing these objective functions, like that taken by Schaefer et al., 2018 (Cerebral Cortex), share some important features with HTFA. For

example, Schaefer et al. include three terms in their objective function; one term is intended to encourage global similarity in the activity between voxels in the same parcel, the second term encourages voxels in the same parcel to have strong resting state functional connectivity, and the third term encourages spatial contiguity. These factors each have analogs in HTFA (activity and functional connectivity based similarity is enforced using a hierarchical model, and spatial contiguity is enforced by requiring each "node" in the final representation to be a Gaussian sphere). In that sense, the approaches end up being conceptually similar, although the specific implementations differ substantially.

Despite these similarities, we see several advantages to HTFA given our paper's specific goals. First, HTFA was fit specifically to this dataset and these participants. We would not have had enough data to reliably carry out something more like the Schaefer et al. (2018) parcellation approach on this dataset. Of course we could have instead used the already published parcellations, but we felt this might miss important idiosyncrasies in our particular participants or the particular experiment/task we studied here. Second, the "result" of applying HTFA to a dataset is a reduced representation of the data (in this case, a number-of-timepoints by 700 matrix for each condition/participant). We could have mapped network parcellations onto this new dataset, but in order to construct "reduced" representations we would then still need to average activity across voxels within a parcel, or to use some other approach to aggregating data. This would have implicitly built in an assumption about spatial uniformity within each parcel, which could miss important aspects of the data. In the extreme, a representation or computation localized to a given parcel could be missed by spatial averaging if it were expressed through spatial, rather than temporal, changes in activity. HTFA can retain these subtle spatial patterns because each node is influenced to some extent by activity *everywhere* in the brain (though with more influence from nearby regions). Third, whereas applying a preexisting parcellation model to our dataset would result in using the same model for all participants, HTFA provides a participant-specific model. For example, even if two participants share the same fundamental representation, the activity pattern in one participant may be spatially warped relative to the other (e.g., as described by Haxby et al., 2011, Neuron, among others). HTFA attempts to account for spatial warping across participants by finding network nodes that behave similarly across people, even if they are not in exactly the same locations across people.

Of course, parcellation approaches also have some benefits over HTFA. For example, they have been more extensively studied and more widely applied. From large datasets and cross-validation studies, we can also have some confidence that parcellation approaches yield (at least somewhat, and for many regions) consistent results across participants.

Ultimately, although we felt HTFA was more appropriate in this specific setting, we do not see either parcellation approaches or HTFA as fundamentally "better" in all applications or cases. Since we had done extensive evaluations of HTFA on this dataset in our prior work (Manning et al., 2018, NeuroImage), we decided to utilize the same approach in our current

study rather than undertaking a new series of quality control checks over and above what we already report.

Although much of this discussion is beyond the intended scope of our paper, we have added some additional text motivating our use of HTFA on pages 22–23.

- The authors state on pg. 6 " the ordering implies that cognitively richer conditions evoke more stable brain activity patterns across people". To what extent is this driven by the fact that the intact condition is also the one that will constrain people to stay on task the most, and therefore to evoke the most similar pattern of brain activity across people (necessary for decoding)? Is there any task required in the other conditions to check for this concern?

Our point here was just what the reviewer is suggesting– that the intact condition is most cognitively and neurally engaging and constraining, the paragraph condition less so, the word condition still less, and the rest condition least of all. We think this is why we see higher decoding accuracy in the more engaging conditions.

We note that the main "task" (passive listening) in the experiment is consistent across all of the experimental conditions (p. 21). In all of the listening conditions (intact, paragraph, and word), the stimulus (i.e., the total set of moment-by-moment auditory patterns) is also held constant, aside from temporal orderings (which vary across conditions); pages 21–22. When participants (via their neural responses) are sensitive to these temporal ordering manipulations, it tells us that contextual information—i.e., participants' experiences prior to a given moment—is driving those consistencies across people (e.g., see Hasson et al., 2008, J. Neurosci.).

Minor comment:
- The authors primarily reference functional connectivity background in motivating their work, and use "network" terminology, even though their analyses focus on timepoint level activation analyses. I found this confusing, as activation findings are likely to have distinct properties from connectivity based findings. Most functional connectivity studies, for example, would not attempt to decode timepoint level data (especially at rest), and instead would focus on analyzing spatial properties of the network maps and how they differed across conditions. It would help if the authors were clearer about the distinctions regarding these types of analyses.

Although our analyses focus on timepoint level activity, what we think is most interesting about our findings is that informativeness and compressibility change systematically across the different experimental conditions. We see this as, fundamentally, a network phenomenon: changing the covariance structure of neural activity patterns would seem to entail some sort of coordination (either directly, or indirectly via common influences) *across* individual brain areas (or "nodes") in the broader network.

We also draw direct inspiration from some of the prior functional connectivity work mentioned in the introduction. For example, we started thinking about the ideas we present in this paper after our 2021 Cerebral Cortex paper (Owen et al., 2021), where we show that activity patterns at one region may be predicted by activity patterns in other (sparsely sampled) regions. This seemed to suggest our neural activity patterns, at least under some circumstances, are highly redundant. For example, why might our brains *need* a region X, if that region's activity could be accurately inferred from other areas? In our current paper we framed these ideas around the notion of compressibility.

We were also inspired by work like Simony et al., 2016 (Nature Comms) showing that the level of similarity in functional correlations across participants varies systematically with cognitive engagement. Prior work on inter-subject correlations (at the level of corresponding voxels, across people), e.g., from Uri Hasson's group, had shown that contextual factors modulate the similarity in neural responses at the *same* brain areas across people. The Simony et al., 2016 paper (also from the same group) suggested to us that these changes might also reflect a larger "network" phenomenon. In our paper, we framed ideas about timepoint-specific activity patterns around the notion of informativeness.

The above notwithstanding, we of course agree that it is important to be clear about what we are claiming and what we are measuring. Throughout our revised text, we have added notes or revised our wording to clarify that we are specifically measuring *activity* and not functional connectivity (correlations).

- The abstract was a bit abrupt at its start; no statement was provided on the motivation of the study or why it was being conducted.

We have added a few sentences to the beginning of the abstract to help motivate the study and our core questions.

- Fig 1 is nice, but it's not completely clear to me how A/B are separate from one another. Aren't they just representing the same data at slightly different points on the curve? It would also be helpful to include a few sentences in the figure legend to explain how this representation would apply to brain activity data.

We have overhauled Figure 1. First, though, the reviewer is correct the panels A and B are essentially representing the same thing, but in two different ways (we have updated the caption to clarify this). The difference between the two panels is in which specific "points along the curve" are sampled. In Panel A, we ask: how many components are needed to explain a fixed proportion of the variance in the two images? We also provide a sense of how each image "looks" when they are reconstructed only up to the given fidelities. In Panel B, we ask: given that we're using just *k* components to describe the images, what proportion of variance can we explain? Again, we provide images to give some intuitions about how the reconstructed images "look" using different numbers of components. We have also added a

new Panel C and D to the figure, along with refactoring the caption, to make it clearer how the concepts in the figure relate to the rest of the paper (e.g., to brain data, etc.).

- I appreciate that the dataset has already been described in prior work, but I think the authors should include at least a brief summary of imaging details relevant to the current manuscript (e.g., TR, TE, # of minutes in each condition, how preprocessing was conducted, how subject motion was addressed) to help readers with interpretation.

We have added the requested information (pp. 21–23).

- Note that the limbic network in Fig. 4 typically overlaps with very low signal regions of the brain, and results tend to be difficult to interpret because of this confound.

We have added a note to this effect (p. 12).

Reviewer #3:

Owen and Manning present a compelling analysis of compressibility and information in fMRI activity patterns. In viewing the brain as a large-scale network, the authors outline a space of brain states defined by dimensions of compressibility, in which activity patterns from brain regions/voxels exhibit varying degrees of similarity, and informativeness, in which activity patterns reflect some quantity of perceptual/cognitive relevance. The key question is where brain states reside in this space under different task demands requiring varying levels of processing. The authors analyzed data from a prior study in which participants listened to an intact story, the story with rearranged paragraphs, randomly shuffled words from the story, or were at rest. PCA-based compression of brain patterns showed higher informativeness (as assessed by intra-subject temporal decoding) with fewer components for intact stories relative to all other conditions. Activity within established resting state networks showed greater informativeness for intact stories overall, but with notably fewer components in default mode and frontoparietal networks. Further, reverse inference of PC brain maps for the top 5 components with Neurosynth topic maps showed a correspondence between higher-level cognitive topics and the intact and paragraph conditions whereas the word condition associated more with lower-level topics. The authors interpret these findings as evidence that cognitively richer stimuli leads to more compressible yet also informative neural codes.

This a compelling paper with many sophisticated and interesting approaches to investigating neural codes. On the whole, I think this paper provides an important contribution to the field. However, I have several questions about the relationship between the measures of compressibility and informativeness, the generalizability of the findings, and how much the current work extends prior publications that use the same dataset:

- The authors motivate well the conceptual differences between compressibility and informativeness. It is an intuitive difference and, at least in conceptual terms, the two can be thought of as independent (as presented in Fig. 1). However, I'm not sure if the same can be said of the measures used in the

analysis. Compressibility is based on PCA of HTFA factor loadings across all participants. The resulting components thus reflect primary dimensions of temporal variability in neural activation. Informativeness is based on temporal decoding of activity patterns across participants. From my understanding, the decoding approach assesses the degree that PCA components capture distinct activity patterns across time. Higher informativeness, therefore, depends on 1) a task that provides temporally distinct stimuli presented at a rate amenable to fMRI sampling and 2) perceptual/cognitive processing that engages with this stimuli. The intact condition seems to satisfy these two constraints more than the others, which makes sense and is reflected in the much higher decoding accuracy. However, what I'm unsure of is, with these two measures, is it possible to have high compressibility with low decoding accuracy? Given that they are both measuring group-level temporal regularity, the measures seem to be intertwined in important ways that should be more fully considered in the manuscript. Are components with higher temporal decoding more likely have higher compressibility?

This is an important question, also raised by Reviewer 2 (though with a slightly different framing). As the reviewer notes, compressibility and informativeness are conceptually distinct: informativeness (in the context of our paper) refers to the distinctness and across-subjects reliability of different *timepoints*, whereas compressibility is related to correlations across neural *features* (i.e., across space).

In our updated manuscript, we have also added some new simulations to show that, indeed, informativeness and compressibility, as we define and measure them, can be varied independently. (E.g., as the reviewer asks, it is possible to have a dataset with high compressibility but low decoding accuracy– we show an example of this in Fig. 1C and D (lower left sub-panels).

We've also copied our response to Reviewer 2's related question here for convenience:

> **The relation between informativeness and compressibility is a core aspect of our paper. Although informativeness and compressibility can be related (e.g., as the reviewer notes, appears to be the case to some extent for our dataset), these factors are not always related. We have updated Figure 1, including an analysis of four new synthetic datasets, to unpack and further illustrate this point (see Synthetic data, Fig. 1, and new introductory text on pages 3–5).**
>
> **In brief, informativeness ends up being related to how timepoint-specific each observed pattern is. In our simulations, we construct data with "high informativeness" by drawing each sample independently from the other timepoints' samples. We construct data with "low informativeness" by introducing strong autocorrelations into the data, which results in higher similarities (i.e., less "distinctiveness") across observations. In other words, informativeness relates to how observations change over time.**

Compressibility is related to relations between features. We can construct compressible data (with high or low informativeness) by introducing covariance between different features. For example, if a given pair of features tends to exhibit similar values, then those features can often be "compressed" into the same factor when we project the data into a lower-dimensional embedding space. We can also construct data with low compressibility, by drawing data from distributions with 0 off-diagonal covariance (i.e., where the covariance between any two different features is 0). In other words, compressibility relates to how features relate to each other.

As we show in Figure 1C and 1D, these two aspects of data (informativeness and compressibility) can be varied independently of each other. In particular, we show that we can construct datasets with all four combinations of high/low informativeness and high/low compressibility.

Our simulations are instructive in that they show how informativeness and compressibility can vary when their connection to the underlying representations we "care about" (i.e., stimulus-driven activity) is assumed. In other words, although we don't explicitly model a "stimulus" as part of our synthetic data generation procedure, our procedure results in activity patterns that are shared (by design) across participants.

In our real data, there is an additional layer of abstraction between the observed activity patterns and the underlying neural representations that we are most interested in. Rather than analyzing activity patterns directly (e.g., measuring autocorrelations across time or features), we use inter-subject functional correlations to specifically home in on stimulus-driven aspects of the data. Essentially, following the logic outlined in Simony et al., 2016 (Nature Comms), aspects of neural activity that are similar across people are very likely to be stimulus driven, since non-stimulus driven activity would not be expected to be time-locked to the stimulus in similar ways across people. Our use of across-participants temporal decoding as our measure of "informativeness" enables us to specifically characterize stimulus-driven informativeness, as opposed to (potentially ambiguous) timepoint "uniqueness" or other non stimulus-related aspects of the data.

We have also added some discussion about why the informativeness and compressibility of our brain patterns may change over time (pp. 15–17):

> **"Our explorations of informativeness and compressibility are related to a much broader literature on the correlational and causal structure of brain activity patterns and networks (Adachi et al., 2012; Bassett & Sporns, 2017; Brovelli et al., 2004; Bullmore & Sporns, 2009; Dhamala et al., 2008; Korzeniewska et al., 2008; Lynn & Bassett, 2021; Owen et al., 2021; Preti et al., 2017; Rogers et al., 2007; Rubinov & Sporns, 2010; Sizemore et al., 2018;**

Smith, Beckmann, et al., 2013; Smith, Vidaurre, et al., 2013; Sporns & Betzel, 2016; Sporns & Honey, 2006; Sporns & Zwi, 2004; Srinivasan et al., 2007; Tomasi & Volkow, 2011; Yeo et al., 2011). Correlations or causal associations between different brain regions simultaneously imply that full-brain activity patterns will be compressible and also that those activity patterns will contain redundancies. For example, the extent to which activity patterns at one brain area can be inferred or predicted from activity patterns at other areas (e.g., Owen et al., 2020; Scangos et al., 2021), reflects overlap in the information available in or represented by those brain areas. If brain patterns in one area are recoverable using brain patterns in another area, then a "signal" used to convey the activity patterns could be compressed by removing the recoverable activity. Predictable (and therefore redundant) brain activity patterns are also more robust to signal corruption. For example, even if the activity patterns at one region are unreadable or unreliable at a given moment, that unreliability could be compensated for by other regions' activity patterns that were predictive of the unreliable region. Whereas compressible brain patterns are robust to spatial signal corruption, high versus low informativeness reflects a similar (though dissociable; e.g., Fig. 1) tradeoff between expressiveness and robustness of temporal patterns. Highly informative brain patterns (by our measure; i.e., patterns that yield greater temporal decoding accuracy) are expressive about ongoing experiences or cognitive states, since each moment's pattern is reliably distinguishable from other moments' patterns. However, when each moment's pattern is unique, brain activity becomes less robust to temporal signal corruption. Our finding that brain activity patterns becomes more informative (i.e., less robust to temporal signal corruption) and compressible (i.e., more robust to spatial signal corruption) when cognitive engagement is higher suggests that our brain may optimize its activity patterns to prioritize either temporal or spatial robustness, according to task demands."

- The current work is based on analysis of a public dataset with several prior publications. This prior work has demonstrated more consistent intra-subject correlations (Simony et al., 2016) and higher temporal decoding with the HTFA method (Manning et al., 2018) for the story condition. As such, there's considerable overlap with the current study and what is novel is a little lost. From my read of it, the key contribution is characterizing how adding PCA into the mix as an index of compressibility relates to some of these prior findings and is linked to the reverse inference of Neurosynth topics. Given the concerns noted above, understanding the degree of independence between the measures of compressibility and informativeness (as operationally defined in the current work) is key to evaluating the level of contribution the present manuscript offers.

The key finding reported by Simony et al. (2016) is that consistency of neural responses across participants decreases across the intact → paragraph → word → rest conditions. In our paper, we use across participants decoding accuracy (our proxy for "informativeness") to

show that accuracy decreases across these conditions in a similar way. Essentially these are two ways of characterizing the same basic phenomenon.

In our paper, we *also* show that neural patterns are more *compressible* in the higher-level (intact and paragraph) conditions. For example, achieving a given decoding accuracy requires fewer components in the intact condition than in the paragraph condition; fewer components in the paragraph than the word condition; and so on. This is one novel contribution of our study. Whereas the Simony et al. (2016) paper nicely demonstrates that neural patterns are more consistent across participants in the higher-level conditions, we show that there is an additional shift in the timepoint-specific neural patterns across the different conditions. High-order conditions evoke more "redundant" (i.e., more compressible) neural activity patterns.

Another way of illustrating the compressibility shift is to imagine the brain as a medium for sending messages (e.g., along the lines of Shannon's information theory). If our goal is to understand which part of the story someone is listening to, compressibility tells us something about how many "messages" are needed to convey the answer (i.e., where in the story our participant is). We found that fewer messages are needed in the intact condition than, for example, the word condition. But we can also ask: how precisely could we nail down the participant's position in the story, even if we could send as many messages as we wanted? Our "informativeness" results show that, even with more messages, we can never achieve the same precision in the word condition as the intact condition. So in the intact condition each individual component of brain activity is more informative (i.e., the activity is highly compressible), *and* the overall information conveyed by the full-brain activity patterns is high (i.e., high informativeness), relative to the other conditions.

- The temporal decoding approach has been used previously and is considered a decent measure of deeper levels of cognitive processing, at least when listening to a story. However, informativeness is a highly variable factor that depends on the task, paradigm, goals, stimuli, etc. Even within the story comprehension paradigm, each participant could be generating highly complex and personal semantic retrievals throughout the story. The temporal decoding approach would specifically ignore these more complex patterns related to an individual's unique brain activity. Thus, the authors' conclusion that deeper levels of processing lead to more compressible yet informative neural codes seem to be linked to both the specifics of the story listening task and the decoding approach. If instead, a study asked participants to perform size judgements or semantic elaborations for different object words and measured informativeness as similarity to representations of semantic networks, one might find a different relationship between compressibility and informativeness (e.g., higher dimensionality linked to higher informativeness). I think the generalization of the current findings to "rules" governing neural codes should be tempered to better represent the empirical and analytic context of the work.

These are important points. To address them, we have added some additional discussion about the limitations of our work (pp. 18–19), and we have also re-worded the "Concluding

remarks" section where we had attempted to generalize our findings to the "rules governing neural codes" to both temper our claims and to stick closer to the scope of the current work.

To unpack our thinking, we agree that participants may have idiosyncratic thoughts while listening to the story. As an approximation, and following the logic outlined by Simony et al., 2016 (Nature Communications), the across-participant similarity in neural responses while experiencing a common stimulus can serve as a means of homing in on stimulus-driven neural activity. Intuitively, only activity patterns that are driven by the stimulus would be expected to synchronize (e.g., be time-locked to the stimulus) across participants. This approach implicitly removes idiosyncratic responses (e.g., neural patterns that are *not* similar across people). This is the basic intuition we leveraged in our current study.

There are also some published examples, including from our own group, that indicate that some types of stimulus-evoked activity will be *missed* by these sorts of across-participant comparisons. For example, in Chang et al., 2021 (Science Advances), we showed how brain regions like the ventromedial prefrontal cortex show stimulus-driven responses that are (for the most part) *not* similar across people. In that paper (and drawing on other work), we suggest that the vmPFC seems to represent or support highly idiosyncratic internal states, like affective responses. Although we would consider the vmPFC to be a "high-level" region (e.g., we consider affect to be a relatively high-level aspect of cognition), our measure of informativeness that we used in our current study would identify regions like the vmPFC as having low informativeness. This is because across-participant decoding accuracy (our proxy measure for informativeness) will only be high for representations or responses that are common across people.

Another related point is that, even in the conditions we describe as "less cognitively engaging," we do not think high-level thought or cognitive processing is *absent*. Rather, we suggest that these high-level representations tend to be more idiosyncratic when the stimulus is less engaging, and therefore less constraining on people's thoughts. The same would be true of the sorts of processes the reviewer is suggesting, like retrieval of personal information as people listen to the story. Those retrievals could happen at different times for different people according to each individual's prior experiences, and even when those sorts of retrievals happen to be temporally synchronized across people, the specific memories or information being retrieved might still be idiosyncratic. As the reviewer notes, our analyses are insensitive to these processes.

The reviewer also suggests that the specific task, even in response to an identical stimulus, could change the relationship between compressibility and informativeness. This idea is consistent with studies like Mack et al., 2020 (Nature Communications), that show that the "dimensionality" of neural representations can change systematically with task complexity, even in response to an identical stimulus.

- I think that prior work with this dataset has shown an association between neural measures and recall performance. The link to behaviour was a key missing aspect in the current work. One way to demonstrate that temporal decoding is capturing some cognitively relevant aspect of informativeness would be to link the degree of temporal decoding (or compressibility) to some behavioural measure of story comprehension/recall.

We agree that this would be very interesting. However unfortunately (to our knowledge) the authors of the original study have only published the neural data, and not the behavioral data.

Although we did not directly relate our neural findings to behavior, we do think that our temporal decoding procedure *implies* cognitive relevance, to the extent that stimulus-driven activity is cognitively relevant. In particular, since all of our decoding is done across participants, only stimulus-driven activity patterns (i.e., time-locked to the stimulus) would be expected to provide useable signal to the decoders. Activity unrelated to the stimulus should not be similar across participants, since such activity would not show any consistent temporal alignment with the stimuli.

- Although the introduction provides a nice description of how network dynamics can be characterized with information theory, I think this manuscript space would be better devoted to providing a clear description of the research questions and the analytic methods for evaluating informativeness and levels of processing. To be honest, the analysis linking components to Neurosynth topics came out of nowhere. Retrospectively, I think this provides a nice approach to characterize the kinds of cognitive operations that may be at play, but foreshadowing this approach would strengthen the presentation.

To help foreshadow both the network parcellation and neurosynth results, we have added a note to the introduction (pp. 5–6) suggesting that tradeoffs between informativeness and compressibility might vary by brain region or network.

- I was surprised to see Chat-GPT used as a omniscient ranker of cognitive domains. I did appreciate that the authors included the full prompt, but I think a rationale for using Chat-GPT is warranted (e.g., provides a window into the zeitgeist). I do worry about the contribution of non-scientific information to the rankings, but the results seem reasonable.

We appreciate that ChatGPT is not actually an omniscient ranker, so we approached its use with a healthy dose of skepticism. In practice, we reviewed ChatGPT's responses by hand to check that they made some sense to us (as cognitive neuroscientists). As the reviewer notes (and we agree), the rankings seemed reasonable. An alternative approach might have been to construct the rankings of cognitive functions by hand, as has been done in many prior studies. We saw our use of ChatGPT in this case as a small additional "sanity check" on our rankings that helped us to be slightly more objective than if we had simply created the rankings ourselves manually. We have added a note to this effect on page 26.