eNeuro

Research Article: New Research

Cognition and Behavior

Neurophysiological Evidence for Cognitive Map Formation during Sequence Learning

Jennifer Stiso,[1] Christopher W. Lynn,[2,3] Ari E. Kahn,[4] Vinitha Rangarajan,[1] Karol P. Szymula,[1] Ryan Archer,[5] Andrew Revell,[5] Joel M. Stein,[6] Brian Litt,[5] Kathryn A. Davis,[5] Timothy H. Lucas,[5] and Dani S. Bassett[1,7,8,9,10,11,1,2]

https://doi.org/10.1523/ENEURO.0361-21.2022

[1]Department of Bioengineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104, [2]Initiative for the Theoretical Sciences, Graduate Center, City University of New York, New York, NY 10016, [3]Joseph Henry Laboratories of Physics, Princeton University, Princeton, NJ 08544, [4]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, [5]Department of Neurology, Hospital of the University of Pennsylvania, Philadelphia, PA 19104, [6]Department of Radiology, Hospital of the University of Pennsylvania, Philadelphia, PA 19104, [7]Department of Electrical and Systems Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104, [8]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, [9]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, [10]Department of Physics and Astronomy, College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, and [11]The Santa Fe Institute, Santa Fe, NM 87501

Abstract

Humans deftly parse statistics from sequences. Some theories posit that humans learn these statistics by forming cognitive maps, or underlying representations of the latent space which links items in the sequence. Here, an item in the sequence is a node, and the probability of transitioning between two items is an edge. Sequences can then be generated from walks through the latent space, with different spaces giving rise to different sequence statistics. Individual or group differences in sequence learning can be modeled by changing the time scale over which estimates of transition probabilities are built, or in other words, by changing the amount of temporal discounting. Latent space models with temporal discounting bear a resemblance to models of navigation through Euclidean spaces. However, few explicit links have been made between predictions from Euclidean spatial navigation and neural activity during human sequence learning. Here, we use a combination of behavioral modeling and intracranial encephalography (iEEG) recordings to investigate how neural activity might support the formation of space-like cognitive maps through temporal discounting during sequence learning. Specifically, we acquire human reaction times from a sequential reaction time task, to which we fit a model that formulates the amount of temporal discounting as a single free parameter. From the parameter, we calculate each individual's estimate of the latent space. We find that neural activity reflects these estimates mostly in the temporal lobe, including areas involved in spatial navigation. Similar to spatial navigation, we find that low-dimensional representations of neural activity allow for easy separation of important

Significance Statement

Humans are adept at learning the statistics of sequences. This ability is facilitated by learning a latent space of transition probabilities between items, or a cognitive map. However, work testing explicit theories of how these maps are built, vary across individuals, and are reflected in neural activity is sparse. We use a model that infers an individual's cognitive map from sequential reaction times and intracranial encephalography (iEEG) recordings to address these gaps. We find that neural activity in the temporal lobe most often reflects the structure of maps and easily identifies task-relevant features of the latent space. We also identify features of individual learning strategies and latent spaces that influence how quickly maps are learned. These discoveries advance our understanding of humans' highly generalizable ability to learn spaces.

features, such as modules, in the latent space. Lastly, we take advantage of the high temporal resolution of iEEG data to determine the time scale on which latent spaces are learned. We find that learning typically happens within the first 500 trials, and is modulated by the underlying latent space and the amount of temporal discounting characteristic of each participant. Ultimately, this work provides important links between behavioral models of sequence learning and neural activity during the same behavior, and contextualizes these results within a broader framework of domain general cognitive maps.

Key words: cognitive maps; intracranial electroencephalography; sequence learning

Introduction

A diverse range of behaviors requires humans to parse complex temporal sequences of stimuli. One can study this ability by exposing individuals to sequences evincing precise statistics, and by measuring how individuals react to or remember the stimuli. Sequence statistics can be fixed by (1) an underlying graph, or latent space, defining allowable transitions between stimuli; and by (2) a walk through the graph that determines which of the allowable transitions are taken and with what frequency (Fig. 1*A*). The graph representation of the latent space brings with it a rich toolbox of methods to quantify latent space topologies that are especially well-suited for abstract relational spaces connecting discrete objects (Butts, 2009). Recent studies have revealed that humans are sensitive to transition probabilities between neighboring elements (Saffran et al., 1996; Fogarty et al., 2019), higher-order statistical dependencies between non-neighboring elements like triplets or quadruplets (Newport and Aslin, 2004), and the global structure of the graph (Schapiro et al., 2013; Kahn et al., 2018). All of these relationships are important for naturalistic learning. For example, when learning a language, both human and artificial language processing algorithms require knowledge of which words tend to follow which others (transition probabilities), as well as of the grammar of sentences, structures of thought, and designs of paragraphs (higher-order structure; Bowerman, 1980; Pennington et al., 2014). Sensitivity to these relationships predicts language ability and problem solving skills (Kidd, 2012; Solway et al., 2014; Pudhiyidath et al., 2020).

Computational models of behavior that require learning an underlying latent space bear a striking resemblance to those used for learning and navigating Euclidean or abstract relational spaces (Gershman et al., 2012; Lynn et al., 2020a). Moreover, similar brain regions have been implicated in all three kinds of cognitive tasks (Buzsáki and Moser, 2013; Schapiro et al., 2016, 2017). However, this level of generalization across task domains has been difficult to replicate in artificial intelligent systems and remains an active area of research (Whittington et al., 2020; Wang, 2021). Work in sequence, relational, and spatial learning suggests that individuals may represent internal estimations of the latent spaces as cognitive maps that can be referenced during navigation and problem solving (Stachenfeld et al., 2014; Constantinescu et al., 2016; Epstein et al., 2017; Behrens et al., 2018). Recent progress in task generalizability in artificial systems has used similar techniques (Whittington et al., 2020; Wang, 2021). Uncovering the processes that guide latent space estimation, and investigating how these processes are implemented in the brain, will deepen our understanding of how humans map the world around them, and provide suggestions for artificial intelligence.

Some mathematical models of latent space estimation rely on individuals building internal estimates of which stimuli in the space are likely to follow which others (Dayan, 1993; Momennejad et al., 2017; Russek et al., 2017). Acquired through exploration, these estimates can be used to make predictions about which stimuli are likely to come next, and therefore allow individuals to navigate the space to reach desired goals (Momennejad et al., 2017). If we were designing a system to learn latent spaces, one strategy for building estimates would be to perfectly remember and log each observed transition, and then to make predictions from that stored estimate. Although such estimates are accurate, they require the learner to store each observed transition, a requirement that is not evidenced in or expected from human behavior (Bornstein et al., 2017; Momennejad et al., 2017; Lynn et al., 2020a). Instead, if estimates of future stimuli incorporate a broader, discounted temporal context, then some of the speed and flexibility of navigation can be restored, although at a cost to the fidelity of the estimate of the latent space (Lynn et al., 2020a; Fig. 1*B*).

Received September 7, 2021; accepted January 3, 2022; First published February 1, 2022.

The authors declare no competing financial interests.

Author contributions: J.S., V.R., and D.S.B. designed research; J.S., A.E.K., K.P.S., R.A., A.R., and J.S. performed research; J.S. and C.W.L. contributed unpublished reagents/analytic tools; J.S. analyzed data; J.M.S. and A.E.K. performed anatomical localization of depth electrodes; J.S., C.W.L., A.E.K., V. R., K.P.S., R.A., A.R., J.M.S., B.L., K.A.D., T.H.L., and D.S.B. wrote the paper.

D.S.B. and J.S. were supported by the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, the ISI Foundation, the Paul Allen Foundation, the Army Research Office (Grafton-W911NF-16-1-0474, DCIST-W911NF-17-2-0181), the National Institute of Mental Health (Grants 2-R01-DC-009209-11, R01-MH112847, R01-MH107235, and R21-M MH-106799), the National Institute of Neurological Disorders and Stroke (Grant R01 NS099348), and the National Science Foundation (Grant BCS-1631550). J.S. was also supported by the National Institute of Mental Health National Research Service Award F31MH120925.

Acknowledgements: We thank Dale Zhou for helpful discussions regarding dimensionality reduction in the brain. We also thank Everett Prince, Jacqueline Boccanfusco, Magda Wernovsky, Amanda Samuel, and all the employees at the CNT who make the collection and curation of iEEG data possible, and the patients and their families for their participation and support.

Correspondence should be addressed to Dani S. Bassett at dsb@seas.upenn.edu.

https://doi.org/10.1523/ENEURO.0361-21.2022

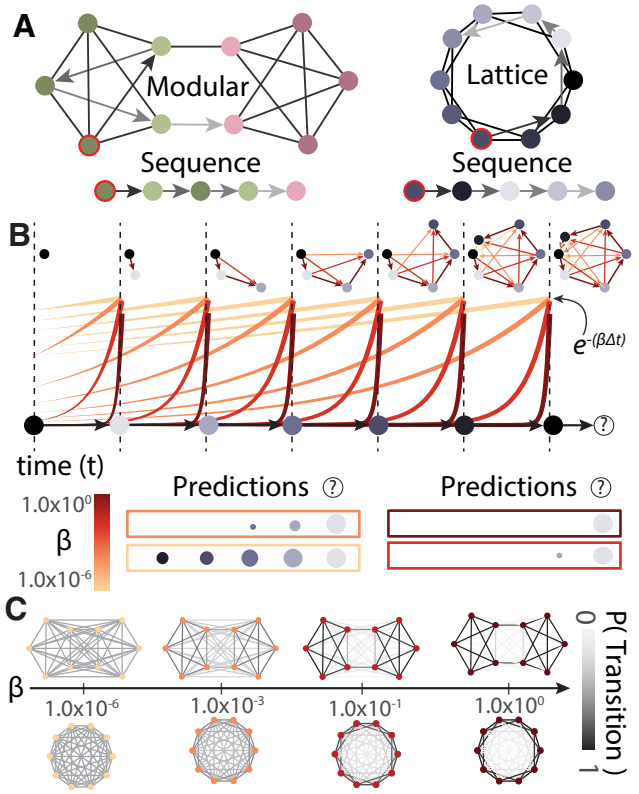

Figure 1. Schematic of latent space learning. ***A***, Visualization of the two graph types used to generate stimulus sequences in this study: modular (left) and lattice (right). An example sequence generated from a random walk on the graph, denoted by arrows, is shown below each visualization. ***B***, A schematic of how temporal discounting of previous stimuli leads to different predictions about which stimulus is likely to appear next in a sequence. As someone experiences each stimulus in a sequence, they will update their estimation of the latent space (shown in the top graph) with the temporal context preceding the currently viewed stimulus. The amount that each stimulus contributes to the temporal context for a given β is given by the height of the colored line. More yellow colors indicate less temporal discounting and smaller β values. Individuals with smaller β values will incorporate stimuli that happened further in the past into their estimate of the latent space. For example, individuals with high (maroon) β values will only estimate connections between stimuli seen one time point apart. However, individuals with small (yellow) β values will estimate connections between stimuli seen more than four time points apart, but with less confidence than connections between stimuli seen more recently. This confidence is depicted visually by the thickness of the connections in the estimated latent space. Individuals can use their estimate of the latent space to predict an upcoming stimulus, indicated by a "(?)." The size of each stimulus in the colored box indicates the confidence that the stimulus is going to be the next node in the sequence and is proportional to the weight of the connection between those stimuli in the estimated latent space. ***C***, A visualization of how different values of β result in different latent space estimations. As β approaches 0, all transitions are estimated to be equally likely. As β approaches ∞, estimations converge to the true structure.

Temporally extended models do not recreate the exact latent space of the true environment, but their modifications can have important behavioral benefits (Fig. 1*C*). For example, artificial intelligent agents using temporal discounting can quickly navigate to rewards in new environments and flexibly respond to changes in strategies or goal locations by using paths they have not explicitly traveled before (Dayan, 1993; Momennejad et al., 2017). Without the modifications from temporal discounting providing an extended context of future paths in space resulting from a given action, agents would only be able to traverse paths they had already encountered, which would limit their flexibility. Additionally, when applied to the free recall of word sequences, these models replicate the ability of humans to remember words presented in similar contexts (Howard et al., 2005). In these temporally extended models, when predicting which state is likely to follow the current state, the agent down-weights stimuli likely to occur far into the future relative to those in the near future; hence the term discounting. These temporally discounted estimates of the latent space can be constructed by applying the same discounting to the history of the previously visited stimuli (Dayan, 1993; Lynn et al., 2020a; Fig. 1*B*). Notably, temporal discounting is a biologically feasible process and can be implemented in brain regions thought to be important for building and manipulating cognitive maps: the hippocampus, entorhinal cortex, and prefrontal cortices (Stachenfeld et al., 2014; de Cothi and Barry, 2019). Activity in the hippocampus and entorhinal cortex has been shown to be more reflective of these discounted estimates of the latent space than the true latent space (Schapiro et al., 2016; Garvert et al., 2017). While most links to implementing temporal discounting have been uncovered in models of the medial temporal lobe, recent studies of latent space learning often find similar activity in diverse areas, suggesting that these algorithms might be implementable in any area of cortex (Bao et al., 2019; Viganò and Piazza, 2020). Taken together, these behavioral and neural insights support the conclusion that humans use temporally discounted estimates of latent spaces to solve a diverse set of problems.

When constructing representations of latent spaces, the brain must balance the need to accurately extract important features from the environment with the pressure to minimize resource consumption (Schapiro et al., 2017; Lai and Gershman, 2021). This balance between compressing information and retaining important features is evidenced behaviorally in the tendency to better remember events or items that occur within a given temporal context, rather than spanning multiple contexts (Brunec et al., 2018). The medial temporal lobe is thought to facilitate the separation and generalization of contexts by identifying key features of estimated latent spaces from low-dimensional projections (Stachenfeld et al., 2014). These lower dimensional projections can serve to identify important features of the space that might be relevant for decisionmaking, such as modules of similar items in relational spaces (Nassar et al., 2018) and borders in physical spaces (Stachenfeld et al., 2014). For cognitive maps specifically, these processes are thought to occur in the

entorhinal cortex, although evidence of similar low-dimensional bases in humans have been found in other regions (Stachenfeld et al., 2014; Constantinescu et al., 2016; Bao et al., 2019). Additionally, other medial temporal structures including the hippocampus have been modeled as variational autoencoders, which compress incoming sensory and structural information to predict future stimuli across domains (Whittington et al., 2020). Further verification that important task features can be identified from a low-dimensional basis of neural activity outside of Euclidean spatial navigation would help support the generalizability of these processes. Additionally, explicit mappings from trade-offs between accuracy and memory load to trade-offs between compressed and separable neural activity can help us better understand dependencies between important behavioral and cellular needs. Behavioral evidence suggests that all these features must be made available after relatively few exposures of different stimuli so that they can be used to make decisions (Lee et al., 2015). Neural recordings taken during latent space learning could help clarify the timescale over which these neural features arise.

Here, we seek to better understand the neurophysiological basis of temporally discounted latent space estimation in humans. Additionally, we wish to test for similarities and divergences from processes of Euclidean spatial learning. To accomplish these goals, we will use an individual specific model of temporal discounting in patients undergoing intracranial encephalography (iEEG) monitoring while completing a probabilistic serial reaction time task. In this task, participants see cues generated from a random walk on either a modular or lattice graph (Fig. 1*A*). To each individual's reaction time data, we apply a maximum entropy model which determines the steepness of temporal discounting as parameterized by a single variable β (Lynn et al., 2020a; Fig. 1*B*). This parameter also determines the structure of the corresponding estimates of the latent space for that individual (Fig. 1*C*). We then use representational similarity analysis to identify the electrode contacts whose activity is most similar to the estimated latent space and identify common regions involved across participants. This analysis allows us to determine whether our model's estimation of latent spaces is reflected in neural activity, and also whether the regions involved are consistent across individuals and previously implicated in Euclidean space navigation. We find that for activity aligned to the stimulus (stimulus-locked), structures in the lateral and medial temporal lobe most often reflect the estimated latent space. In activity aligned to the response (response-locked), this similarity with the latent space shifts to frontal and premotor areas. We next tested whether low-dimensional neural activity could easily identify features of the latent space, as it does in Euclidean spatial learning. We find robust separability of modules in neural activity, consistent with the identification of borders and clusters in Euclidean and relational learning. Lastly, we wish to extend our understanding of the temporal dynamics of latent space estimation. In our sample of neural data, we find that neural activity reflects the latent space within 500 stimulus exposures, and that the steepness of temporal discounting and the structure of the underlying graph influence the learning rate.

Ultimately, our study provides a direct comparison between the distinct processes of latent space learning, coupled with an evaluation of their neurophysiological underpinnings. Additionally, it provides preliminary measurements of the timescales on which latent space estimations are formed, and an accounting of which factors influence their development. Lastly, we provide clear future directions for model development, and point out areas where neural data diverge from theoretical predictions.

Materials and Methods

Participants

All participants provided informed consent as specified by the Institutional Review Board of the University of Pennsylvania, and study methods and experimental protocols were approved by the Institutional Review Board of the University of Pennsylvania.

Amazon's Mechanical Turk (mTurk) cohort

We recruited 50 unique participants to complete our study on Amazon's mTurk, an online marketplace for crowdsourced work. Worker identifications were used to exclude any duplicate participants. Twenty-five of the participants completed a task with a sequence generated from a modular graph, and the other 25 participants performed the same task with a sequence generated from a ring lattice graph. All participants were paid $10 for their time (≈20 min). Three individuals started, but did not complete the task, leaving the sample size at 47 individuals. Interested candidates were excluded from participating if they had completed similar tasks for the lab previously (Kahn et al., 2018; Lynn et al., 2020a).

iEEG cohort

There was a total of 13 participants (10 female, mean age 33.9 years). See Table 1 for full demographics. This included three participants who completed a pilot version of the task that was largely similar. These participants were included to increase the number of participants when data collection paused during the COVID-19 pandemic. Two of these 13 participants did not have electrophysiological recordings that were synchronized with the task recordings; accordingly, these two participants were only included in behavioral analyses.

Behavior

For each participant in the iEEG and mTurk cohorts, we test their ability to learn the structure underlying a temporal sequence of stimuli by having them perform a probabilistic motor response task using a keyboard. We will first outline elements common to both tasks here, and then highlight differences.

Table 1: Participant demographics

Participant	Sex	Age (years)	Race	Ephys	Median RT	Graph	β	Coverage
1	F	36	White	Y	0.923	Lattice	0.024	Left: medial, ventral, and lateral temporal; central gyrus/sulcus; insula; middle and inferior frontal. Right: medial, ventral, and lateral temporal.
2	M	24	White	Y	0.546	Modular	0.17	Right: medial, ventral, and lateral temporal; insula; middle and inferior frontal.
3	F	25	White	Y	1.103	Lattice	1000	Left: medial, ventral, and lateral temporal; inferior parietal; middle and inferior occipital; insula; cingulate. Right: medial, ventral, and lateral temporal; inferior parietal; central gyrus/sulcus; insula.
4	F	32	White	Y	0.572	Modular	0.43	Left: medial, ventral, and lateral temporal; inferior occipital; insula; inferior frontal. Right: medial and lateral temporal; insula.
5	F	47	White	Y	0.657	Lattice	1000	Left: medial, ventral, and lateral temporal; inferior parietal; middle and inferior occipital.
6	F	58	White	Y	0.732	Modular	0.12	Right: superior and middle frontal; central gyrus/sulcus; supplemental motor; cingulate.
7	M	21	White	Y	0.724	Lattice	0.33	Left: medial, ventral, and lateral temporal; cingulate; central and middle occipital; superior parietal; superior, middle, and inferior frontal.
8	M	22	White	Y	0.482	Modular	0.024	Left: medial, ventral, and lateral temporal; insula; middle and inferior frontal.
9	F	22	White	Y	1.075	Modular	0.23	Left: medial and ventral temporal; insula; inferior and middle frontal; basal ganglia. Right: medial, ventral, and middle temporal; inferior and superior frontal.
10	F	37	White	Y	0.795	Modular	0.044	Left: medial, ventral, and lateral temporal; insula; inferior frontal; cingulate.
11	F	47	Black	Y	0.696	Modular	0	-
12	F	23	White	N	1	Modular	-	-
13	F	39	White	N	0.773	Modular	-	-

Demographic and task relevant information about each participant. The Ephys column indicates whether or not the participant had electrophysiological recordings (Y indicates 'yes', N indicates 'no'). The Median RT column provides the median reaction time. A '-' indicates that those data were not available.

Common experimental setup and procedure

First, participants were instructed that "In a few minutes, you will see 10 squares shown on the screen. Squares will light up in red as the experiment progresses. These squares correspond with keys on your keyboard, and your job is to watch the squares and press the corresponding key when the square lights up as quickly as possible to increase your score. The experiment will take around 20 min." For some participants, the sequence of stimuli was drawn from a random traversal through a modular graph (Fig. 1*A*, left); for other participants, the sequence of stimuli was drawn from a random traversal through a ring lattice graph (Fig. 1*A*, right). Both graphs have 10 distinct nodes, each of which is connected to four other nodes. Thus, the only difference between the two graphs lies in their higher-order structure. In the modular graph, the nodes are split into two modules of five nodes each, whereas in the lattice graph, the nodes are connected to their nearest and next-nearest neighbors around a ring. For each participant, the 10 stimuli are randomly assigned to the 10 different nodes in either the modular or lattice graph. The random assignment of stimuli to nodes ensures that modules are not distinguished by any stimulus features. Stimuli were each represented as a row of ten gray squares. Each square corresponds to and mimics the spatial arrangement of a key on the keyboard (Fig. 2*A*). To indicate a target key that the participant is meant to press, the corresponding square is outlined in red (Fig. 2*B*). If an incorrect key was pressed the message "Error!" displayed on the screen until the correct key was pressed. Participants had a brief training period (10 trials) to familiarize themselves with the key presses before engaging in the task for 1000 trials, which is a sufficient number of trials for participants to learn the structure of a similarly sized modular network (Kahn et al., 2018). To ensure that participants remain motivated and engaged for the full 1000 trials, participants receive points based on their average reaction time at the end of each of four stages (every 250 trials). The duration of the task is determined by how quickly participants respond, but on average it takes approximately 20 min. On average, participants in the mTurk cohort were 94.0 $\pm$ 3.76% accurate, and participants in the iEEG cohort were 97.7 $\pm$ 2.50% accurate.

mTurk experiment

Because no experimenter could be present for online mTurk data collection, a few additional measures were put in place to ensure that participants understood and were engaged with the task. First, participants were given a short quiz to verify that they had read and understood the instructions before the experiment began. If any questions were answered incorrectly, participants were shown

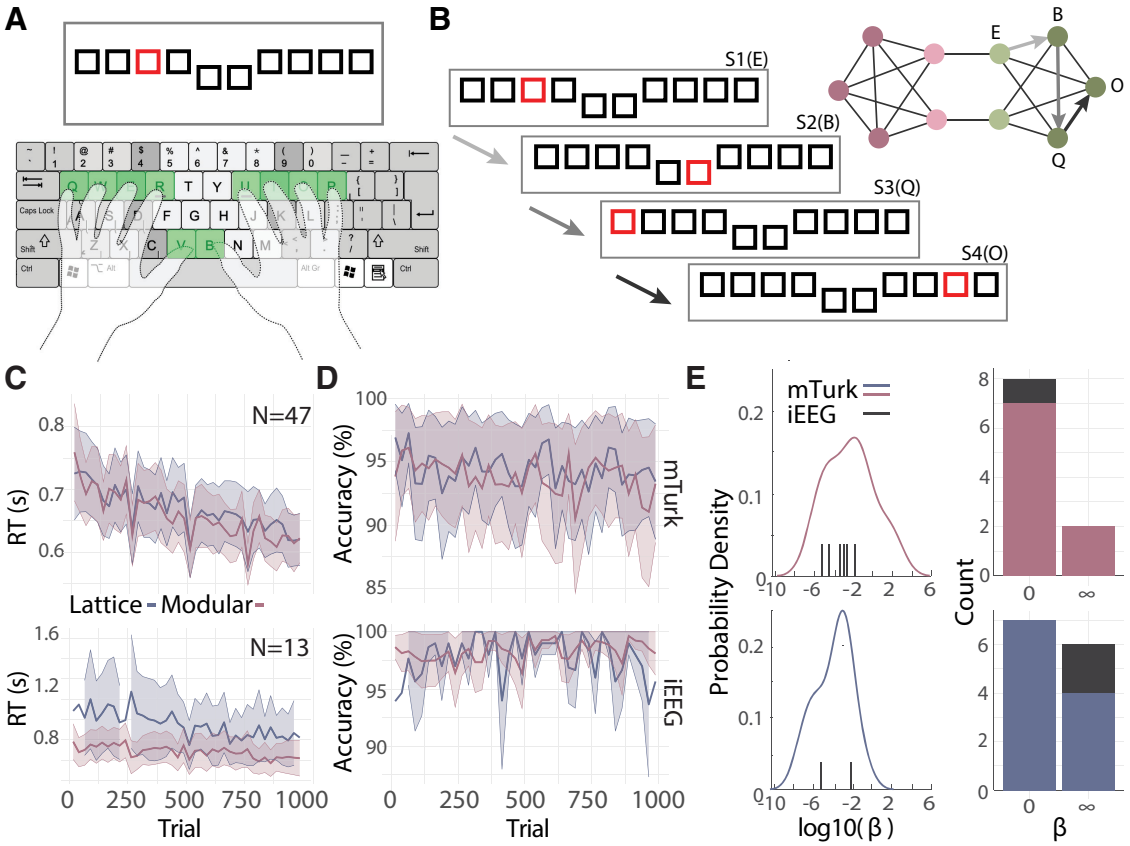

Figure 2. Task performance. ***A***, The hand position that the participants use to complete the task. ***B***, An example of four trials from the experiment. The first stimulus (S1) shows the third key highlighted in red, which corresponds to the letter *E*. This key maps to the light green node in the graph to the right. After the participant presses the correct key, they will advance to next trial, which in this case is the key *B*. ***C***, Average reaction time over trials for the mTurk (top) and iEEG (bottom) cohorts. Reaction times are shown for both modular (pink) and lattice (blue) graphs. For visualization purposes, reaction times were averaged across participants for each trial. Those average reaction times were then binned into 25-trial bins. Shaded regions indicate the standard deviation across participants for each bin. ***D***, The same plots as in panel ***C***, but for accuracy. ***E***, Distributions of β values for both cohorts (left) and the number of β values equal to 0 and ∞ (right). Participants who saw modular graphs are shown in pink on the top; participants who saw lattice graphs are shown in blue on the bottom. Plots were separated spatially to avoid overlap between individual data points. iEEG cohort β values are shown with black tick marks rather than as a population density because of the small sample size.

the instructions again and asked to repeat the quiz until they answered all questions correctly. Additionally, participants were instructed that if they took longer than 1 min to respond to any given trial, the experiment would end and they might not receive payment.

iEEG experiment

A member of the Hospital for the University of Pennsylvania (HUP) research staff was present during the experiment to ensure that participants understood the instructions. De-identified demographic information was collected and shared from all participants as part of the HUP research protocol. This information included age, race, and sex assigned at birth, as well as an estimate of how much of their day the participant typically spent typing at a computer. The iEEG experiment, unlike the mTurk experiment, also needed to be synchronized to ongoing neural recordings. To synchronize task events with neural recordings, the iEEG participants completed the task with a photodiode attached to the laptop where the test was being administered. A white square would appear in the lower corner of the screen when a stimulus appeared on the screen, which would be replaced by a black background when the correct response was made. The photodiode would record these luminance changes on the same system that was recording neural data, so that the two could be synchronized.

Participants in the cohort were also given the option to complete a second session of the same experiment with the same graph the following day. This option was taken by two participants. Because data collection was interrupted by the global pandemic, we also include three pilot iEEG participants who completed an earlier version of the task that did not contain breaks or points, but was otherwise identical.

Linear mixed-effects models

We used linear mixed-effects models to test whether each participant's reaction time decreased with increasing trial number. We took this decrease in reaction time as

evidence that participants were learning the probabilistic motor response task. Before fitting the mixed-effects models, we excluded trials that were shorter than 50 ms, or longer than 2 SDs above that participant's mean reaction time. Short trials were removed because 50 ms is not long enough to see and respond to a stimulus. We also excluded any incorrect trials. All participants in both cohorts had accuracy >80%.

Mixed-effects models were fit using the lme4 library in R (R version 3.5.0; lme4 version 1.1–17), using the lmer() function for continuous dependent variables and the glmer() function for categorical dependent variables. Predictors were centered to reduce multicollinearity. Some models of accuracy did not converge with the full set of variables, so variables were removed via backwards selection with reaction time model p-values until the accuracy model converged. Because of the slight task differences between iEEG and mTurk cohorts, different models were used to test for learning in each cohort. For the mTurk cohort, the reaction time model was *reaction_time*~*trial*graph*+*stage*graph*+*finger*+*hand*+*hand_transition*+*recency*+(1+*trial*+*recency*|*participant*). The accuracy model was correct = trial*graph+stage*graph+finger+hand_transition+recency+(1+trial|participant). Here, *hand_transition* indicated whether the current trial used a different hand than the previous trial, and *stage* indicated the set of 250 trials, ranging from 1 to 4. For the iEEG cohort, the reaction time model was reaction_time~trial*graph+stage*graph+finger+hand_transition+session+points+recency+(1+trial+recency|participant). The model for accuracy was correct~squared_trial*graph +**stage*graph+finger+hand_transition+session+recency+(1+squared_trial|participant). Here, *session* indicated whether the data were taken from the first or second recording session, *points* indicated whether these participants were given points according to their reaction time at breaks, and *typing_skill* was a self-reported value of how much time participants spent typing on a computer in a typical day, scaled to range from 1 to 4. The linear mixed-effects model used to assess cross-module differences in reaction time was the reaction time model, but with *is_cross_module* in place of the graph indicator. Similarly, the linear mixed-effects model used to assess differences in reaction time based on β values was the reaction time model, but with β in place of the graph indicator.

The *recency* term is meant to account for changes to reaction time based on the local properties of the current sequence. Participants will tend to react more quickly to items they have seen more recently (Karuza et al., 2017). To control for this effect, we included the log transform of the number of trials since the current stimulus was last seen, or the *recency*, as a covariate. The maximum number of trials was 10. This particular covariate was found to explain more variance in reaction time than other similar covariates in this dataset, as well as a similar dataset collected from Kahn et al. (2018; their Fig. S1). The specific covariates tested were the number of times the current stimulus was last seen (not log transformed, and not capped; χ^2 test $\chi^2 = 2448$, $p < 2.2 \times 10^{-16}$) and the number of times this stimulus appeared in the last 10 trials (χ^2 test $\chi^2 = 1295.8$, $p < 2.2 \times 10^{-16}$).

Maximum entropy model: β and $\hat{A}$

To estimate the amount of temporal discounting employed by each participant, we fit a maximum entropy model to the residuals of the linear mixed-effects models specified above. Residuals rather than raw reaction times are used to better isolate reaction variations because of the underlying expectations of the graph, rather than biomechanical or motor features like which hand or finger was used to respond. The model starts with the assumption that the fastest reaction times on this task would arise from accurate mental representations of the latent space. This would allow participants to accurately predict which stimuli could possibly follow any current stimulus, allowing them to react quickly to all transitions. However, these representations are costly to create and maintain because they require perfect memory of the sequence of stimuli. Allowing some inaccuracies in the memory of previous stimuli simplifies the learning process, but at the cost of erroneous predictions about future stimuli. In this model, an exponentially decaying memory distribution determines the time scale of errors in memory. The exponential form results in the fact that mistakes in memory will be temporally discounted, more likely to occur between stimuli that are temporally close than those that are temporally distant. The steepness of this discounting, and therefore the balance of cost and accuracy, is determined by a single parameter β that was fitted to the residuals of each participant's reaction times. Larger β values result in more temporal discounting in the memory distribution, indicating that participants were less likely to make memory errors, and the errors that were made tended to occur between stimuli in close temporal proximity. By contrast, smaller β values would result in less temporal discounting, indicating that participants made longer range errors in their estimates of the transition graphs. Mathematically, this is achieved by defining an individual's estimation of the latent space as $\hat{A} = (1 - e^{-\beta})A(I - e^{-\beta}A)^{-1}$, where A is the true latent space that defines transition probabilities between stimuli.

To drive more intuition for the model, we will explicitly walk through an example sequence from a lattice graph (Fig. 1*B*). Here, each colored circle represents a stimulus that is shown at a given time during a sequence. The bottom of the figure panel shows the exponentially decaying memory function for different values of β in different colors, ranging from maroon (highest β) to yellow (lowest β). The top of the figure shows the estimated latent space for individuals with different values of β at every time step, where transitions for different β values are shown in different colors ranging from maroon to yellow. At the first time point, just after the first stimulus is shown, all individuals (for any value of β) estimate the latent space to be a single node with no connections. At the second time step, all individuals update their latent space to contain one transition, between the first and second stimuli. At the third time step, latent spaces start to diverge for individuals with different β values. All individuals will once again update their estimate to reflect a transition between stimulus 2 and stimulus 3. However, individuals with lower β values (red, orange, and yellow) will additionally estimate

another weaker and erroneous connection between stimulus 1 and stimulus 3. The strength of the connection formed between two stimuli that are δt apart in time decreases exponentially as $e^{-\beta \delta t}$. Erroneous connections allow the estimated latent space to reflect the broader temporal context of stimuli preceding stimulus 3. This pattern continues at time step 4. All individuals add the transition between stimulus three and stimulus 4. Lower β values (red, orange, and yellow) add a weaker transition between stimulus 2 and stimulus 4, and even lower β values (orange and yellow) estimate an additional, even weaker connection between stimulus 1 and stimulus 4. At time point 7, stimulus 1 is repeated, and individuals have a chance to use their estimated latent spaces to predict which stimulus will come next in the sequence. The certainty of that prediction is calculated by summing the outgoing connections from stimulus 1. Individuals with the highest (maroon) β only have one connection in their latent space, so they will predict that stimulus 2 will come next. By contrast, individuals with the smallest β (yellow) think stimulus 2 is most likely to come next, think it slightly less likely that they will see stimulus 3, slightly less likely that they will see stimulus 4, and even less likely that they will see stimulus 5. We expect that individuals will react more quickly to cues that they more confidently predict; therefore, we would expect that individuals with different β values will react differently to the same sequence.

At time point 7, stimulus 1 is repeated, and individuals have a chance to use their estimated latent spaces to predict which stimulus will come next in the sequence. The certainty of that prediction is calculated by summing the outgoing connections from stimulus 1 in the estimated latent space. Individuals with the highest (maroon) β value only have one connection in their latent space, so they will predict that stimulus 2 will come next. By contrast, individuals with the smallest β value (yellow) think stimulus 2 is most likely to come next, and think it is slightly less likely that they will see stimulus 3, slightly less likely that they will see stimulus 4, and even less likely that they will see stimulus 5. We expect that individuals will react more quickly to cues that they more confidently predict. Therefore, we would expect that individuals with different β values will react differently to the same sequence. We will next explain how β is calculated from reaction times below.

Given an observed sequence of nodes $x_1, ..., x_{t-1}$, and given a parameter β, our model predicts each participant's internal estimates of transition probabilities $\hat{A}_{ij}(t-1)$, where i and j are different stimuli. Given a current stimulus x_{t-1}, we then model the participant's anticipation, or expectation, of the subsequent node x_t by $a(t) = \hat{A}_{x_{t-1}, x_t}(t-1)$. In order to quantitatively describe the reactions of a participant, we related the expectations $a(t)$ to predictions about a participant's reaction times $\hat{r}(t)$, and then learned the model parameters that best fit that participant's reaction times. The simplest possible prediction was given by the linear relation $\hat{r}(t) = r_0 + r_1 a(t)$, where the intercept r_0 represents a participant's reaction time with zero anticipation and where the slope r_1 quantifies the strength with which a participant's reaction times depend on their internal expectations. In total, our predictions $r(t)$ contain three parameters (β, r_0, and r_1), which must be estimated from the data for each participant. To estimate the model parameters that best describe a participant's reaction times $r(t)$ (more specifically, their reaction time residuals from the linear mixed-effects model described above), we minimized the root mean squared prediction error (RMSE) with respect to each participant's observed reaction times, $RMSE = \sqrt{\sum_t (r(t) - \hat{r}(t))^2}$. We note that, for a given β, the parameters r_0 and r_1 can be calculated using linear regression. Thus, the problem of estimating the model parameters can be restated with only one parameter; that is, by minimizing the RMSE with respect to β.

Because we wished to compare results from these models to neural data, we only run this analysis on each of the participants with neural data, and exclude trials that contained interictal epileptiform discharges (IEDs). To minimize the RMSE with respect to β, we began by calculating the RMSE along 100 logarithmically spaced values for β between 10^{-4} and 10. Then, starting at the minimum value of this search, we performed gradient descent until the gradient fell below an absolute value of 10^{-6}. The search also terminated if β reached 0, or was trending toward ∞ (>1000). The β values that were terminated at 0 or 1000 are referred to as extreme values throughout the manuscript.

Once β values were fitted for each participant, the estimated latent space $\hat{A}$ could be obtained with the equation: $\hat{A} = (1 - e^{-\beta}) A (I - e^{-\beta} A)^{-1}$, where A is the true latent space that defines transition probabilities between stimuli. This analytic prediction reflects the estimated latent space for a participant that viewed an infinite random walk, and does not take into account the statistics of the particular sequence observed by a given participant.

In addition to calculating each participant's estimated latent space, we also wished to understand how the estimate would evolve over time assuming a static β. A participant's expected likelihood of a transition between two elements i and j at time t is given by $\hat{A}(t) = \frac{\widetilde{n}_{ij}(t)}{\sum_k \widetilde{n}_{ik}(t)}$, where $\widetilde{n}_{ij}$ is a participant's recollection of the number of times they have observed stimulus i transition to stimulus j. We can then use β to solve for the expected number of transitions as $\widetilde{n}_{ij}(t+1) = \widetilde{n}_{ij}(t) + \sum_{\Delta t=0}^{t-1} \frac{1}{Z} e^{-\beta \Delta t} \delta(j = x_{t-\Delta t})$. Here, $\delta(.)$ is a δ function that gives a value of 1 when its argument is true and 0 otherwise, and Z is a normalization constant.

Intracranial recordings

All patients included in this study gave written informed consent in accord with the University of Pennsylvania Institutional Review Board for inclusion in this study. Deidentified data were retrieved from the online International

Epilepsy Electrophysiology Portal (Wagenaar et al., 2018). All data were collected at a 512-Hz sampling rate.

Preprocessing

Electric line noise and its harmonics at 60, 120, and 180 Hz were filtered out using a zero phase distortion fourth order stop-band Butterworth filter with a 1-Hz width. This filter was implemented using the butter() and filtfilt() functions in MATLAB. Impulse and step responses of these filters were visually inspected for major ringing artifacts.

We then sought to remove individual channels that were noisy, or had poor recording quality. We first rejected channels using both the notes provided and automated methods. After removing channels marked as low quality in the notes, we further marked electrodes that had (1) a line length greater than three times the mean (Ung et al., 2017); (2) a *z*-scored kurtosis >1.5 (Betzel et al., 2019); or (3) a *z*-scored power-spectral density dissimilarity measure >1.5 (Betzel et al., 2019). The dissimilarity measure was the average of one minus the Spearman's rank correlation with all channels. These automated methods should remove channels with excessive high frequency noise, electrode drift, and line noise, respectively. All contacts selected for removal were visually inspected by a researcher with six years of experience working with iEEG data (J.S.). The final set of contacts was also visually inspected to ensure that the remaining contacts had good quality recordings by the same researcher. Including removal of contacts outside of the brain, on average, 48.87 $\pm$ 22.50% of contacts were removed, leaving 89 ± 30 contacts.

Data were then demeaned and detrended. Detrending was used instead of a high-pass filter to avoid inducing filter artifacts (de Cheveigné and Nelken, 2019). Channels were then grouped by grid or depth electrode, and common average referenced within each group. Recordings from white matter regions have sometimes been used as reference channels (Li et al., 2018). However, work showing that channels in white matter contain unique information independent from nearby gray matter motivated us to include them in the common average reference (Mercier et al., 2017). Following the common average reference, plots of raw data and power spectral densities were visually inspected by the same expert researcher with six years of experience working with electrocorticography data (J.S.) to ensure that data were of acceptable quality.

Next, data were segmented into trials. A trial consisted of the time that a given stimulus was on the screen before a response occurred. iEEG recordings were matched to task events through the use of a photodiode during task completion (see above, iEEG experiment). Periods of high light content were automatically detected using a custom MATLAB script. Identified events were then visually inspected for quality. The times of photodiode change were then selected as the onset and offset of each trial. Two participants had poor quality photodiode data that could not be segmented, and these participants were accordingly not included in electrophysiological analyses, leaving 11 remaining participants.

Lastly, trials were rejected if they contained interictal epileptiform discharges (IEDs). IEDs have been shown to change task performance (Ung et al., 2017) and aspects of neural activity outside of the locus of IEDs (Dahal et al., 2019; Stiso et al., 2021). We chose to use an IED detector from Janca et al. (2015) because it is sensitive, fast, and requires relatively little data per participant. This Hilbert-based method dynamically models background activity and detects outliers from that background. Specifically, the algorithm first downsamples the data to 200 Hz and applies a 10- to 60-Hz bandpass filter. The envelope of the signal is then obtained by taking the square of this Hilbert-transformed signal. In 5-s windows with an overlap of 4 s, a threshold k is calculated as the mode plus the median and used to identify IEDs. The initial k value is set to 3.65, which was determined through cross-validation in Janca et al. (2015). In order to remove false positives potentially caused by artifacts, we apply a spatial filter to the identified IEDs. Specifically, we remove IEDs that are not present in a 50-ms window of IEDs in at least three other channels. The 50-ms window was consistent with that used in other papers investigating the biophysical properties of chains of IEDs, which tended to last <50 ms (Conrad et al., 2020).

Contact localization

Broadly, contact localization followed methodology similar to Revell et al. (2021). All contact localizations were verified by a board-certified neuroradiologist (J.M.S.). Electrode coordinates in participant T1w space were assigned to an atlas region of interest and also registered in participant T1w space. Brain region assignments were assigned first based on the AAL-116 (Tzourio-Mazoyer et al., 2002) atlas. This atlas extends slightly into the white matter directly below gray matter, but will exclude contacts in deeper white matter structures. For a list of the number of contacts in each region of this atlas, see Table 2. To provide locations for contacts outside the AAL atlas, we use the Talairach atlas (Brett et al., 2001). Assignment of contacts to a hemisphere was also done using the Talairach atlas label. If the contact was outside of the Talaraich atlas, then the AAL atlas hemisphere was used. If a contact was outside both atlases, then the contact name taken from https://www.ieeg.org/ was used (contact names include the hemisphere, electrode label, and contact label).

Similarity analysis

In this work, we sought to identify which electrode contacts have neural activity that reflected a participant's estimate of the latent space in a data driven manner. To identify these contacts, we used a similarity analysis that compared $\hat{A}$, the participant's estimation of latent space, to the similarity of neural activity evoked by each stimulus. This approach was used to abstract similarity patterns in high-dimensional neural activity into dissimilarity matrices, and allowed us to answer the

Table 2: Gray matter contacts

AAL region	Visual count	Latent count	Visual (%)	Latent (%)	Total count
Amygdala	1	3	9.09	27.2727273	11
Angular	1	0	11.11	0	9
Calcarine	2	0	28.57	0	7
Caudate	0	0	0	0	6
Cerebellum_4_5	1	0	100	0	1
Cingulum_Ant	0	0	0	0	3
Cingulum_Mid	0	0	0	0	1
Cingulum_Post	0	1	0	16.6666667	6
Cuneus	0	0	0	0	1
Frontal_Inf_Oper	0	1	0	14.2857143	7
Frontal_Inf_Orb	0	1	0	8.33333333	12
Frontal_Inf_Tri	10	2	19.23	3.84615385	52
Frontal_Med_Orb	0	0	0	0	2
Frontal_Mid	3	1	16.61	5.55555556	18
Frontal_Mid_Orb	0	0	0	0	3
Frontal_Sup	1	1	10	10	10
Frontal_Sup_Orb	0	0	0	0	2
Fusiform	4	5	8.69565217	10.8695652	46
Heschl	0	0	0	0	2
Hippocampus	4	2	7.84313725	3.92156863	51
Insula	1	2	4.76190476	9.52380952	21
Lingual	0	0	0	0	5
NotInAtl	25	11	9.65250965	4.24710425	259
Occipital_Inf	1	0	20	0	5
Occipital_Mid	0	0	0	0	9
Olfactory	0	0	0	0	1
Pallidum	0	0	0	0	1
ParaHippocampal	1	1	4.34782609	4.34782609	23
Parietal_Inf	1	0	16.6666667	0	6
Parietal_Sup	0	0	0	0	1
Postcentral	0	1	0	9.09090909	11
Precentral	4	2	23.5294118	11.7647059	17
Precuneus	1	0	33.3333333	0	3
Putamen	0	0	0	0	5
Rectus	1	0	100	0	1
Rolandic_Oper	0	0	0	0	12
Supp_Motor_Area	0	0	0	0	5
SupraMarginal	0	0	0	0	7
Temporal_Inf	5	5	4.62962963	4.62962963	108
Temporal_Mid	4	6	3.30578512	4.95867769	121
Temporal_Pole_Mid	2	0	18.1818182	0	11
Temporal_Pole_Sup	1	0	20	0	5
Temporal_Sup	2	1	6.06060606	3.03030303	33

The locations of all contacts in the AAL atlas.

question "Where does neural activity reflect the latent space?" (Kriegeskorte et al., 2008). These matrices can then be compared with similarity patterns obtained from our computational model, $\hat{A}$.

Here, we chose the cross-validated Euclidean distance as our neural similarity metric because it was shown to lead to more reliable classification accuracy when compared with other dissimilarity metrics (Walther et al., 2016). To compute similarity matrices for each contact, we first truncated all trials of preprocessed iEEG recordings to be the same length as the trial with the shortest reaction time. If the shortest reaction time was $<$200 ms, we instead used 200 ms as the minimum length and discarded trials shorter than that. This truncation was done in three ways: (1) stimulus aligned, where the end of trials was truncated; (2) middle aligned, where the middle of trials was truncated; and (3) response aligned, where the beginning of trials was truncated. We then calculated the leave-one-out cross-validated Euclidean distance between activity evoked from each of the 10 unique stimuli. This procedure resulted in one dissimilarity matrix for each contact. To compare these matrices to the estimated latent space, we then calculated the correlation between the lower diagonal of the neural dissimilarity matrix and $\hat{A}$. Because $\hat{A}$ reflects similarity rather than dissimilarity, we then multiplied the resulting correlation by -1.

To identify electrode contacts with high similarity to the latent space, we compared the correlations between the neural dissimilarity and the estimated latent spaces to correlations between a distribution of 100 null neural dissimilarity matrices and estimated latent spaces. Null matrices were calculated from permuted data created by first

selecting a random trial number and then splitting and reversing the order of trials at that point. For example, if 128 were drawn as a random trial number, the corresponding permuted dataset would be neural data from trials 129–1000 followed by neural data from trials 1–128 matched with stimuli labels from the correct order of trials (1–1000). This model preserved natural features of autocorrelation in the neural data, unlike trial shuffling models (Aru et al., 2015). Contacts were determined to have activity similar to the latent space if they met two criteria: (1) the correlation between the neural dissimilarity matrix and the $\hat{A}$ was greater than at least 95 null models, and (2) the correlation between the neural dissimilarity matrix and the $\hat{A}$ was greater than the correlation between the neural dissimilarity matrix and the exact latent space A. Because we only require that contacts have similarity values $>$95% of null models and the correlation with the exact latent space, we expect a rate of false positives among contacts of close to (but less than) 5%. Therefore, we focus our discussion on regions where $>$5% of the total contacts were retained.

To test the specificity of our findings, we also examined the correlation between the dissimilarity matrices and a similarity space related to the lower-level features of the stimuli. We calculated a spatial similarity matrix that reflected the physical distance between stimuli on the screen. Since each stimulus consists of a single red square among nine black squares on the screen, we calculated the Euclidean distance between each square, and used this matrix as an estimate of spatial similarity. We then repeated the process detailed above for obtaining correlations relative to permuted neural data.

Low-dimensional projections and linear discriminability

For visualization purposes, we sought to obtain low-dimensional representations of the neural dissimilarity matrices. Classical multidimensional scaling (MDS) obtains low-dimensional (here, two dimensional) representations of Euclidean distance dissimilarity matrices that seek to preserve the distances of the original higher-dimensional data (Wang, 2013). Classical MDS was implemented using the cmds() function in MATLAB. For neural data, we first calculated a single neural dissimilarity matrix, rather than a single matrix per contact. This calculation was done by concatenating activity from every contact whose activity was similar to the latent space (see above, *Similarity analysis*), and then by repeating the process outlined above.

For some analyses, we wished to compare the low-dimensional representations of neural dissimilarity matrices with the low-dimensional representations of estimated latent spaces. Since estimated latent spaces are not Euclidean distance matrices, classical MDS is not an appropriate dimensionality reduction technique (Wang, 2013). Instead, we use principal components analysis (PCA). PCA yields the same low-dimensional embedding as classical MDS when the high-dimensional data are Euclidean distances, but not otherwise. We computed the principal components of the neural dissimilarity matrices and estimated latent spaces in MATLAB using the pca() function. The scaled and centered data were then projected onto the first two principal components to obtain two coordinates for each node.

From these low-dimensional data, we next sought to assess estimates of discriminability between modules. Module discriminability was calculated as the loss from a linear discriminant analysis. A linear classification model was fit to the low-dimensional coordinates using the fitdiscr() function in MATLAB. The proportion of nodes that were incorrectly classified using the best linear boundary, or the loss, was then reported as an estimate of the linear discriminability of modules.

Statistical analyses

Linear mixed-effects models were used to analyze reaction time data, and the results are displayed in Figure 2. Mixed-effects models were used to account for the fact that trials completed by the same participant constitute repeated measures and are not independent. The estimated β values were evaluated with t tests, and appear to be approximately normally distributed. Extreme values of β (0 or 1000) were removed from any statistical tests to ensure normality. Linear mixed-effects models were used to analyze changes in neural similarity over time, with participant included as a random effect. A paired t test was used to analyze changes in loss from a linear classifier.

Data and code

Code is available in the GitHub repository https://github.com/jastiso/statistical_learning. Electrophysiological data will be made available on request from the IEEG portal.

Citation diversity statement

Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are undercited relative to the number of such papers in the field (Maliniak et al., 2013; Mitchell et al., 2013; Caplar et al., 2017; Dion et al., 2018; Bertolero et al., 2020; Dworkin et al., 2020; Chatterjee and Werner, 2021; Fulvio et al., 2021; Wang et al., 2021). Here, we sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity, and other factors. First, we obtained the predicted gender of the first and last author of each reference by using databases that store the probability of a first name being carried by a woman (Dworkin et al., 2020; Zhou et al., 2022). By this measure (and excluding self-citations to the first and last authors of our current paper), our references contain 10.0% woman(first)/woman(last), 11.3% man/woman, 18.8% woman/man, and 55.0% man/man. This method is limited in that a) names, pronouns, and social media profiles used to construct the databases may not, in every case, be indicative of gender identity and b) it cannot account for intersex, nonbinary, or transgender people. Second, we obtained predicted racial/ethnic category of the first and last author of each reference by databases

that store the probability of a first and last name being carried by an author of color (Ambekar et al., 2009; Sood and Laohaprapanon, 2018). By this measure (and excluding self-citations), our references contain 8.86% author of color (first)/author of color(last), 10.82% white author/author of color, 20.19% author of color/white author, and 60.13% white author/white author. This method is limited in that (1) names and Florida Voter Data to make the predictions may not be indicative of racial/ethnic identity, and (2) it cannot account for Indigenous and mixed-race authors, or those who may face differential biases because of the ambiguous racialization or ethnicization of their names. We look forward to future work that could help us to better understand how to support equitable practices in science.

Results

Quantification of learning and temporal discounting

In this work, we are interested in the neural underpinnings of latent space estimation. Before investigating the neural dynamics directly, we tested whether participants learned the latent space and responded both faster and more accurately to stimuli over time. Our cohort of interest, the iEEG cohort, were all undergoing monitoring for medically refractory epilepsy. Because of the rarity of this population, it is often difficult to get large cohorts suitable for good estimates of behavioral effect sizes. Additionally, the epileptic population in the iEEG cohort has been shown to have cognitive impairments (Parvizi and Kastner, 2018), which requires tasks that have been designed to be comparatively easy and quick to complete. Because of these challenges, we also collected data from 50 participants from Amazon's mTurk.

In both cohorts, we were interested in the change over time of two estimates of learning: accuracy and reaction time. Across participants, we found that the average accuracy for the mTurk cohort was 94.0 ± 3.76%, with a median reaction time of 602.5 ± 134.0 ms. For the iEEG cohort, the mean accuracy was 97.7 ± 2.50% with a median reaction time of 721.2 ± 180.9 ms. We were also interested in determining whether the rate of learning differed between two graph types (Fig. 1*A*). We used linear mixed-effects models to assess learning based on increases in accuracy and decreases in reaction times on two time scales. The first, shorter timescale is that of individual trials; to examine learning on this timescale we tested for decreases in reaction time associated with increasing trial number. Since this task provided breaks after each 250-trial stage, we also assessed learning at the longer timescale of individual stages. To examine learning on this timescale, we tested for decreases in reaction time with increasing stages. In the mTurk cohort ($n = 47$), we found that reaction times tend to decrease only at the trial level (linear mixed-effects model $F_{trial} = 16.1, p_{trial} = 9.51 \times 10^{-5}$, F_{stage} = 0.005, p_{stage} = 0.946; Fig. 2*C*). In the iEEG cohort ($n = 13$), we found that reaction times decrease only at the stage level (linear mixed-effects model F_{trial} = 1.16, p_{trial} = 0.320, F_{stage} = 3.86, p_{stage} = 0.049; Fig. 2*C*). For accuracy, we found that the mTurk cohort shows a significant decrease in accuracy with trials (linear mixed-effects model z_{trial} = –2.48, p_{trial} = 0.013, z_{stage} = 1.93, p_{stage} = 0.054; Fig. 2*D*). For the iEEG cohort we observe no significant linear change with trial (linear mixed-effects model z_{trial} = –0.025, p_{trial} = 0.98, z_{stage} = 0.289, p_{stage} = 0.773; Fig. 2*D*). However, we qualitatively observed a quadratic relationship, where accuracy initially increased before decreasing with trial number. We tested the statistical significance of this observation with a mixed-effects model that relates accuracy to $trial^2$. We found that the quadratic trial estimate is a significant predictor of accuracy (linear mixed-effects model $z_{trial^2} = -2.6$, $p_{trial^2} = 0.009$; Fig. 2*D*).

We also sought to determine whether participants showed evidence of learning the underlying latent space in addition to the task. We tested for evidence of sensitivity to the underlying latent space in two ways: (1) a difference in reaction time based on graph type, and (2) a difference in reaction time based on module transition in the modular graph. In the mTurk cohort, we found that there was no difference in reaction time (linear mixed-effects model F_{graph} = 0.013, p_{graph} = 0.910) or in learning rate ($F_{trial*graph} = 0.043 p_{trial*graph} = 0.834$, $F_{stage*graph} = 0.002$, $p_{stage*graph} = 0.966$; Fig. 2*C*) between the graphs. There were also no significant changes in accuracy associated with graph type (linear mixed-effects model z_{graph} = –0.186, p_{graph} = 0.853, $z_{trial*graph} = -0.818$, $p_{trial*graph} = 0.414$, $z_{stage*graph} = 1.121$, $p_{graph*stage} = 0.225$; Fig. 2*D*). In the iEEG cohort, we found no differences in reaction time (F_{graph} = 1.63, p_{graph} = 0.300), but there was a significant interaction between learning rate and graph type at the stage level ($F_{graph*trial} = 4.70$, $p_{trial*graph} = 0.072$, $F_{stage*graph} = 14.3$, $p_{stage*graph} = 1.52 \times 10^{-4}$; Fig. 2*C*). There was also a significant interaction between accuracy and graph type (linear mixed-effects model, z_{graph} = –2.6, p_{graph} = 0.711, $z_{trial*graph} = 2.30$, $p_{trial*graph} = 0.022$, $z_{stage*graph} = 1.94$, $p_{stage*graph} = 0.052$; Fig. 2*D*).

We next tested whether reaction times reflected a sensitivity to the modules in the modular graph. We expected that reaction times would decrease for within-module transitions relative to between-module transitions (Kahn et al., 2018). In our mTurk cohort, we observe significantly slower reaction times on cross-cluster trials, compared to within-cluster trials (linear mixed-effects model, $F_{cross-module} = 11.0$, $p = 3.4 \times 10^{-3}$). In our iEEG cohort, we do not find a significant difference in reaction time across modules (linear mixed-effects model, $F_{cross-module}$ = 0.93, $p = 0.36$). Overall, we found that the iEEG cohort showed evidence of learning in both accuracy and reaction time. While the mTurk cohort showed quicker decreases in reaction time, these were coupled with decreases in accuracy. Additionally, we find that the mTurk cohort showed sensitivity to the latent space through cross-module transitions, while the iEEG cohort had steeper learning on the lattice graph ($n = 4$) compared with the modular graph ($n = 9$). Ultimately, both cohorts show evidence for learning the task and the latent space, but in different ways.

After we confirmed that participants learned the task, we quantified each participant's steepness of temporal discounting. For both cohorts, we calculated the parameter β by fitting a maximum entropy model to the residuals of reaction times from the linear mixed-effect model discussed above. This parameter indicates the prioritization of accurate latent space estimations against the cost of those accurate representations, as evidenced by each participant's behavior (Fig. 1*B*,*C*). The parameter β was fit with gradient descent, assuring that the fit for each participant was comparable, with the exception of the extremes of the distribution of possible β values (β = 0 and $\beta = \infty$). However, each extreme value has a different interpretation. A fitted value of β = 0 indicates that there is no evidence of temporal discounting or any sensitivity to sequence statistics in a participant's reaction times. The corresponding estimate $\hat{A}$ of the latent transition probabilities will show equally likely transitions between all nodes. A fitted value of $\beta = \infty$ indicates no influence of the cost of building accurate representations but a high sensitivity to the statistics of the sequence. At long time scales, their estimate $\hat{A}$ will converge to the actual latent space, but at short time scales, their estimates will reflect the local statistics of the specific sequence. To provide additional empirical evidence for our interpretation of extreme β values, we used a linear mixed-effects model to test for reaction time differences between individuals with β values of 0 and β values of ∞ in our mTurk cohort. We hypothesized that the high sensitivity to sequence statistics in individuals with $\beta = \infty$ would lead to initially slow reaction times that quickly sped up as the latent space was learned. Using a linear-mixed effect model containing an interaction between trial number and β, we find a significant effect for the β value ($F = 5.69$, $p = 0.029$), and for the interaction between β and reaction time ($F = 7.51$, $p = 0.01$). These results show that individuals with $\beta = \infty$ have larger reaction times and steeper learning than individuals with β = 0, which support our interpretation of the extreme values.

Because the gradient descent algorithm terminated if β approached 0 or ∞, we assessed the similarity of temporal discounting, operationalized as similar β values, between cohorts with two measures: (1) the percent of participants where β approached one of these extremes; and (2) the distribution of β values found between these two extremes. Additionally, all parametric statistical tests that used β values were applied after extreme values were removed, thus ensuring the normality of the β distribution. We first examined the percentage of participants who had β values at the extremes of the distribution. In the mTurk cohort, we found that 29.8% (or 14 participants: seven modular and seven lattice) had β values equal to 0, indicating no evidence for sensitivity to sequence statistics from their reaction times. For the iEEG cohort, we found that 7.7% (or one participant: modular) of participants had β values equal to 0. In the mTurk cohort, we found that 12.8% (or six participants: two modular and four lattice) had β values equal to ∞, indicating high sensitivity to sequence statistics. For the iEEG cohort, we found that 15.4% (or two participants: both lattice) had β values equal to 0.

We next assessed differences in the distribution of log-distributed β values in both cohorts. The mTurk cohort had a mean β value of 0.94 and showed no differences across graph types (permutation test: $p = 0.11$; Fig. 2*E*). The iEEG cohort had a mean β value of 0.17 and was not statistically different from the mTurk cohort (permutation test: $p = 0.53$; Fig. 2*E*). As these data indicate, we found similar temporal discounting levels among both groups, although the mTurk cohort had more participants with extreme values. We note that β values tend to be <1, indicating a high prioritization of the costs of building accurate representations. Since this amount of temporal discounting resulted in estimated latent spaces that are different from true latent spaces, we next investigated neural activity reflecting these estimated latent spaces.

Anatomical areas where activity reflects latent space estimation

We used a similarity analysis in a data driven manner to identify which contacts showed activity with a similar structure to the estimated latent space. First, we calculated the similarity structure of neural activity by calculating the cross-validated Euclidean distance between the activity evoked for each stimulus (Fig. 3*A*). To ensure that all stimuli had activity of the same length, the last time points of all trials were removed to create epochs the length of the shortest trial. We also report results based on removing the first and middle time points to reach the same length. We then selected the contacts where this neural similarity structure was closest to the estimated latent space, and closest to the visual Euclidean distance (see Materials and Methods; Fig. 3*A*).

The resulting contacts from all participants are visualized on a shared space (MNI; Fig. 3*B*). Between 2 and 10 contacts displayed activity whose dissimilarity matrices were similar to those of the latent space per participant. Qualitatively, we observed that contacts that reflect latent and Euclidean space appear in the frontal and temporal lobes, with some overlap between the two groups. Overall, 46 (5.0%) contacts spanning all participants were identified as reflecting the latent space, and 76 (8.3%) were identified as reflecting the Euclidean space. For the latent space, 32 (5.1%) contacts were from the right hemisphere and 14 (4.5%) contacts were from the left hemisphere. We note that we expect to select more visual than latent space sensitive contacts because visual space correlations were not required to be larger than the correlations to the true latent space. We also show separate visualizations for participants with modular and lattice graphs, respectively (Fig. 3*B*). Qualitatively, we observe a large overlap in the identified regions between the two graph types.

We next sought to localize identified contacts in each participant's native space. The most common AAL atlas labels for latent space contacts are shown in Figure 3*C*. We found the most common regions identified were the middle temporal lobe (six contacts 5.0%), fusiform gyrus (five contacts, 11.0%), inferior temporal lobe (five contacts, 4.6%), and amygdala (three contacts, 27.3%). The middle temporal lobe and amygdala also showed the

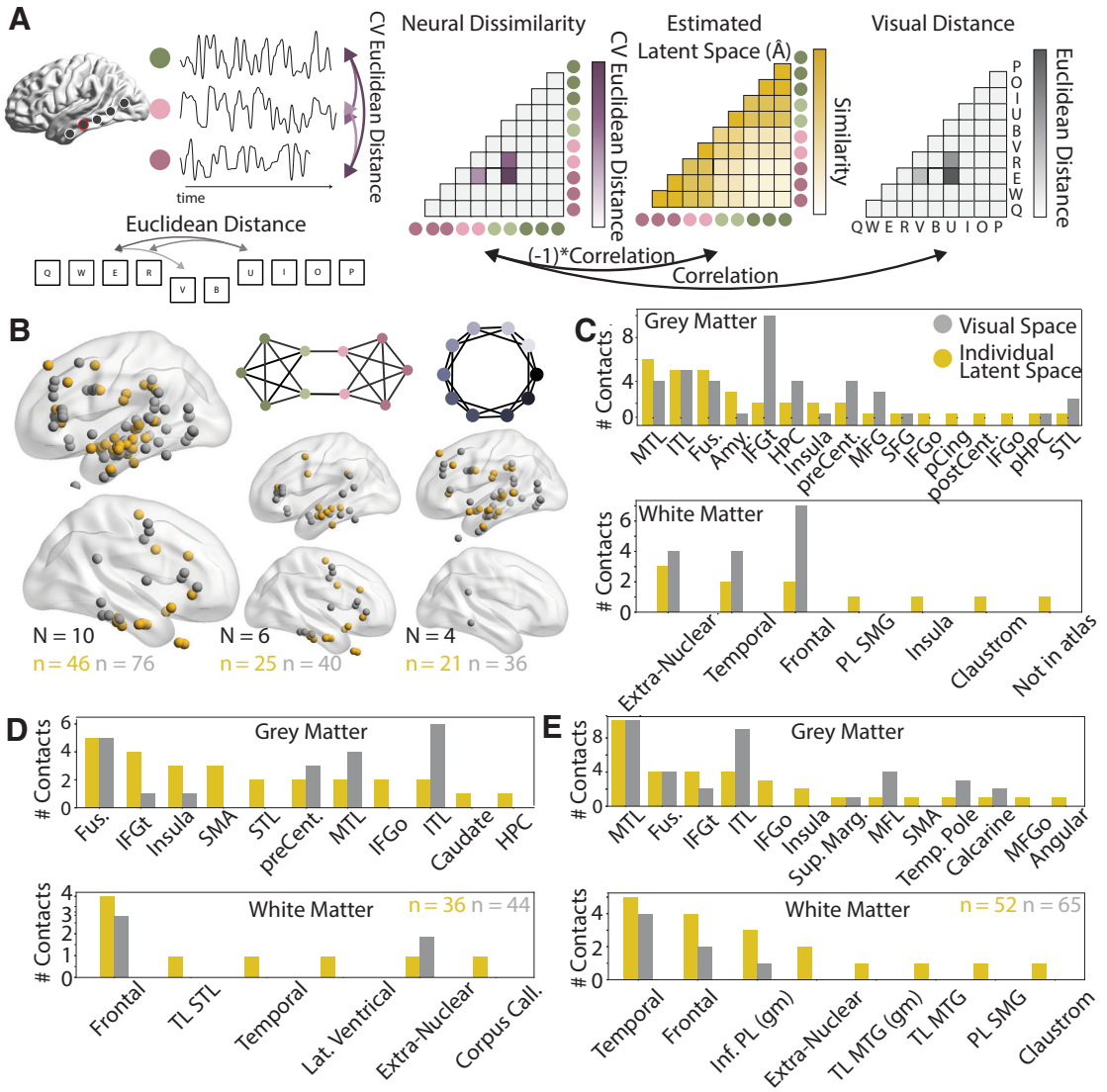

Figure 3. Location of contacts whose activity reflects the estimated latent space. ***A***, A schematic of the representational similarity analysis used here. Euclidean distances (top, left) are calculated from neural responses to each stimulus to create a neural dissimilarity matrix (leftmost matrix). The Euclidean distance between highlighted stimuli is also calculated (bottom, left) and used to make a visual distance matrix (rightmost matrix). The similarity between the neural dissimilarity matrix and both the estimated latent space $\hat{A}$ and the visual space are assessed with Pearson's correlation coefficients. Because the estimated latent space is a similarity matrix, rather than a dissimilarity matrix, the correlations are multiplied by −1. ***B***, Visualization of contacts with similar stimulus-aligned activity spaces on an MNI brain for all participants (left), only those who saw sequences from a modular graph (middle), and only those who saw sequences from a lattice graph (right). Contacts whose neural dissimilarity matrices are similar to those of the latent space are shown in gold; contacts whose neural dissimilarity matrices are similar to those of the visual space are shown in gray. The quantity *N* is the number of participants in each plot, and the quantity *n* is the number of contacts in each plot. ***C***, Anatomical localization of gray (top) and white (bottom) matter contacts for stimulus-aligned activity. ***D***, Anatomical localizations for gray and white matter for response-aligned activity. The number of contacts identified for estimated visual and latent spaces are written in gray and gold. ***E***, The same as panel ***D***, but with middle-aligned activity. ITL, inferior temporal lobe; MTL, middle temporal lobe; STL, superior temporal lobe; Fus, fusiform; pHPC, parahippocampal gyrus; Amy, amygdala; HPC, hippocampus; IFGt, inferior frontal gyrus (pars triangularis); IFGo, inferior frontal gyrus (pars orbitalis); MFG, middle frontal gyrus; SFG, superior frontal gyrus; preCent, precentral; postCent, postcentral; pCing, posterior cingulate; PL SMG, parietal lobe supramarginal gyrus; Inf PL (g), inferior parietal lobe (gray matter); TL MTG (g), temporal lobe middle temporal gyrus (gray matter).

most selectivity for the latent space compared to the visual space. Contacts located in white matter were localized with the Talaraich atlas. Most often, these contacts were in extranuclear, frontal or temporal sublobar white matter. We also looked explicitly at contacts in the hippocampus and entorhinal cortex, and identified 4 contacts in the hippocampus that reflect the estimated latent space and four that reflect that visual space (3.9%). We identified one contact in the parahippocamal gyrus (which contains both the parahippocampal cortex and the

entorhinal cortex in the AAL atlas) that reflects the estimated latent space (4.3%). Upon visual inspection, this contact is located in the anterior part of the parahippocampal gyrus, and therefore is likely part of the parahippocampal cortex rather than the entorhinal cortex. Information for all regions is given in Table 2. We note that the most common regions identified for latent space contacts were in lateral, ventral, and medial temporal lobe. While frontal contacts still showed neural dissimilarity consistent with that of the latent space, the specific anatomical location of these contacts were not consistent across participants.

The above similarity analysis was based on data that was aligned to stimulus presentation, and hence captured the initial evoked response to the stimulus. To examine whether other portions of the response data produced the same structure, we repeated this analysis with response-aligned trials (removing the first time points), and middle-aligned trials (removing the middle time points; Fig. 3*D*,*E*). When trials were aligned to the response, we found that fewer contacts were identified whose activity reflected the latent space, although such contacts were still identified in each participant. Overall, we found 36 (3.9%) contacts showed activity similar to the estimated latent space, and 44 (4.8%) contacts showed activity similar to the visual space. For latent space contacts, 24 (3.9%) were in the left and 12 (4.0%) in the right hemisphere. We found one region, the fusiform gyrus (5 contacts 10.9% for response-aligned), that included multiple contacts for both response-aligned and stimulus-aligned activity. Unlike the stimulus-aligned similarity, the response-aligned similarity also identified the inferior frontal gyrus (the pars orbitalis, pars triangularis, and pars opercularis; 4, 7.7%), insula (3, 14.3%), and supplemental motor area (3, 60%) as important regions. For the middle-aligned activity, we found 52 (5.7%) contacts showed activity reflecting the latent space and 65 (7.1%) showed activity reflecting the visual space (Fig. 3*D*). Among these latent space contacts, 39 (6.3%) were in the left hemisphere and 13 (4.4%) were in the right hemisphere. We found that most identified regions overlap with those identified for stimulus-aligned and response-aligned activity. Specifically, we found that the areas most commonly displaying activity that reflects the latent space are located in the middle temporal lobe (10, 8.3%), the fusiform gyrus (4, 8.7%), the inferior frontal gyrus (4, 7.7%), and the inferior temporal cortex (4, 3.7%; Fig. 3*E*). These results suggest that our findings from response-aligned and stimulus-aligned activity are not driven by the activity of the middle of the trial. Overall, we found that early stimulus-evoked activity shows greatest similarity to the estimated latent space in higher-order temporal regions, whereas later response-locked activity shows more similarity to the estimated latent space in frontal and especially premotor regions.

Module discriminability in low-dimensional space

Many of our hypotheses about low-dimensional projections of neural activity build on prior evidence in the hippocampus. To be more consistent with this literature, the remaining analyses considered only the stimulus-locked neural dissimilarity matrices, where the medial temporal lobe contacts most reflected the estimated latent space. We visualized low-dimensional projections of neural activity across all of the contacts that demonstrated similarity to the estimated latent space (Fig. 4). These low-dimensional projections were obtained for each participant by first creating a single dissimilarity matrix for all contacts whose activity was similar to the latent space, and then computing classical MDS on those matrices. From these low-dimensional projections, we can observe the diversity of estimated structures, and the ways in which they reflect and differ from the exact latent space that generated the sequences of images (Fig. 4). One notable property of participants who experienced sequences from modular graphs is that modules (green and pink) appear to be mostly separable. All participants appear to highly accurately separate the two modules, even when activity from diverse regions is included.

We next wished to test whether this modular separability in low-dimensional neural activity is also present in latent spaces estimated from behavior. If modules are separable in both spaces, then temporally discounted space estimations might be sufficient to explain separability. If separability is only present in neural spaces, then further computations are likely needed to explain separability. We test for module discriminability using a linear discriminant analysis on low-dimensional coordinates obtained from principal components analysis applied to neural distance matrices and on the estimated latent space (see Methods). We find that for most participants, discriminability varied little between the data from neural dissimilarity and from the estimated latent space (paired t test $t = -1.04$, $p = 0.344$; Fig. 5*A*). However, two participants showed much lower discriminability for the estimated latent space, compared with the neural dissimilarity space (Fig. 5*A*). We next sought to test whether discriminability for estimated latent spaces was specific to a particular range of β values. We find that participants with higher β values show perfect discriminability whereas participants with lower β values do not (Fig. 5*B*). Visualization of the low-dimensional projections from different β values shows that the poor discriminability was driven by the nodes with connections to other modules (Fig. 5*C*). More specifically, at a β value close to 0.1, we see an abrupt shift where nodes connecting two modules switch from being closer to their corresponding module, to being closer to the contrasting module. Taken together, these findings provide evidence for module discriminability in neural activity. However, whether that discriminability is predicted by the estimated latent space alone depends on the participant and diminishes for those participants characterized by low β values.

Temporal dynamics of latent space formation

In a final investigation, we sought to model how the estimated latent space might change during learning, and test whether neural activity showed similar temporal patterns. Assuming that β values are static during the course of learning, we can simulate how the estimated latent space $\hat{A}$ changes on each trial because of each new transition observed between stimuli. We then

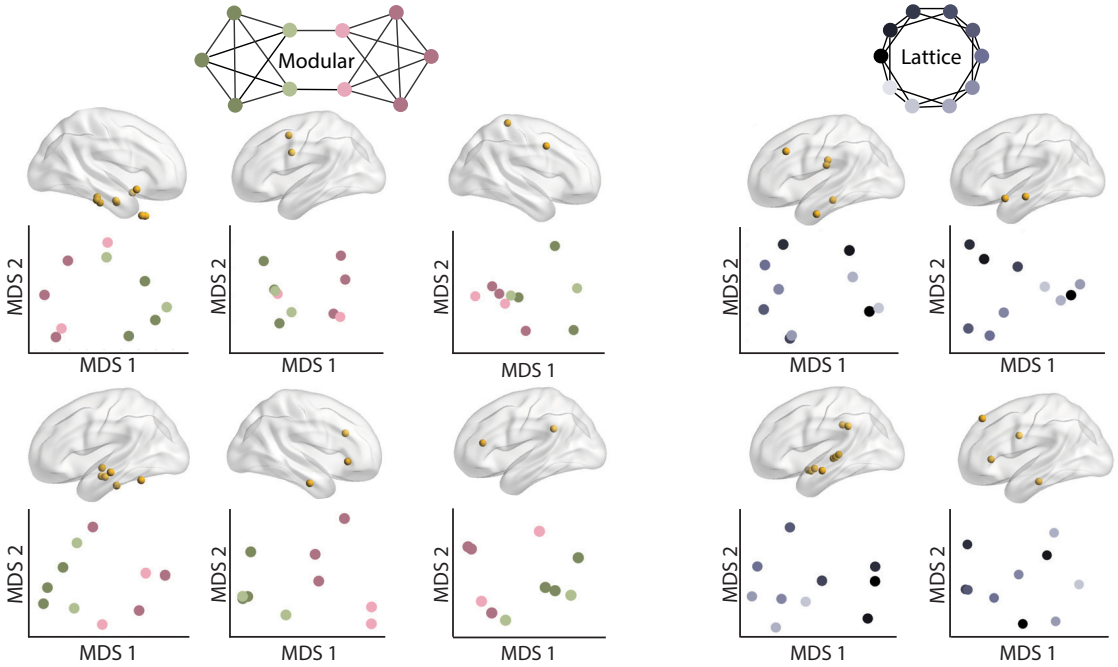

Figure 4. Diversity of low dimension projections of neural activity. The contacts used to create the underlying dissimilarity matrices (top). Low-dimensional projections of neural dissimilarity matrices for each participant (bottom). MDS 1 and 2 are the first and second dimensions obtained from multidimensional scaling. For visualization purposes, only the hemisphere with the most contacts is shown. Participants who viewed walks from modular graphs are shown in the three leftmost columns, and participants who viewed walks from lattice graphs are shown in the two rightmost columns.

calculated the correlation between the current estimated latent space at each trial $\hat{A}(t)$ and the estimated latent space obtained using the infinite trial limit $\hat{A}$ (Fig. 6*A*). Since participants only observed a finite walk, the quantity $\hat{A}(t)$ does not converge to exactly $\hat{A}$. However, most participants quickly show high agreement between the finite and infinite-time estimates as they learn. Qualitatively, we see that larger β values result in a faster convergence toward the final $\hat{A}$ regardless of the graph type (Fig. 6*A*,*B*).

Informed by these data, we hypothesized that neural activity structure would also reflect the estimated latent space $\hat{A}$ fairly early during learning. In order to ensure that we had enough trials to get stable estimates of activity structure, we tested this hypothesis by recalculating the correlation between the latent space and neural dissimilarity matrices in sliding blocks of 500 trials with a 100-trial offset. This process provided a total of 6 blocks. We recomputed these correlations only in individual contacts (*n* ranged from 2 to 10) whose activity was determined to be similar to the estimated latent space (Fig. 3). We also calculated the correlation to the visual space in these same contacts as a comparison (Fig. 6*C*). Since we wished to capture the dynamics of contacts converging to their final values rather than differences in those final values, we normalized all correlation coefficients to the

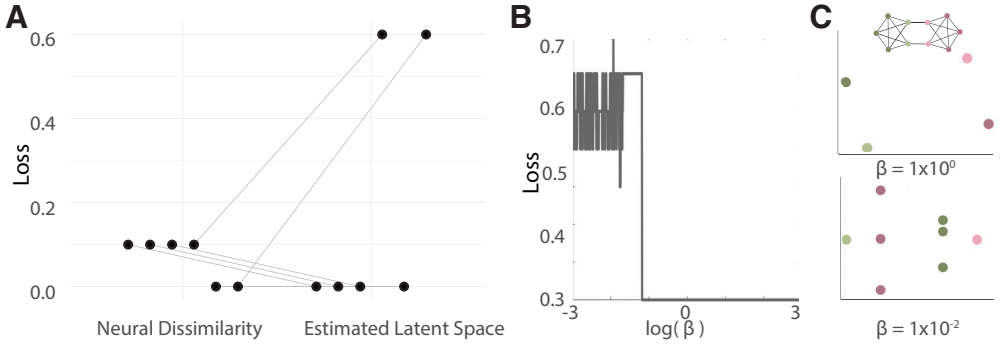

Figure 5. Module discriminability. ***A***, Loss, or incorrectly classified nodes, for each participant's neural distance matrix and their estimated latent space. Lines connect the loss of the neural dissimilarity and estimated latent spaces for a single participant. ***B***, The loss for estimated latent spaces at different β values. ***C***, Visualization of a low-dimensional projection of an estimated latent space from a large β value (top) and from a small β value (bottom) for one participant. Both plots show 10 points, although there is a high amount of overlap in the top plot, making only four points clearly distinguishable by eye.

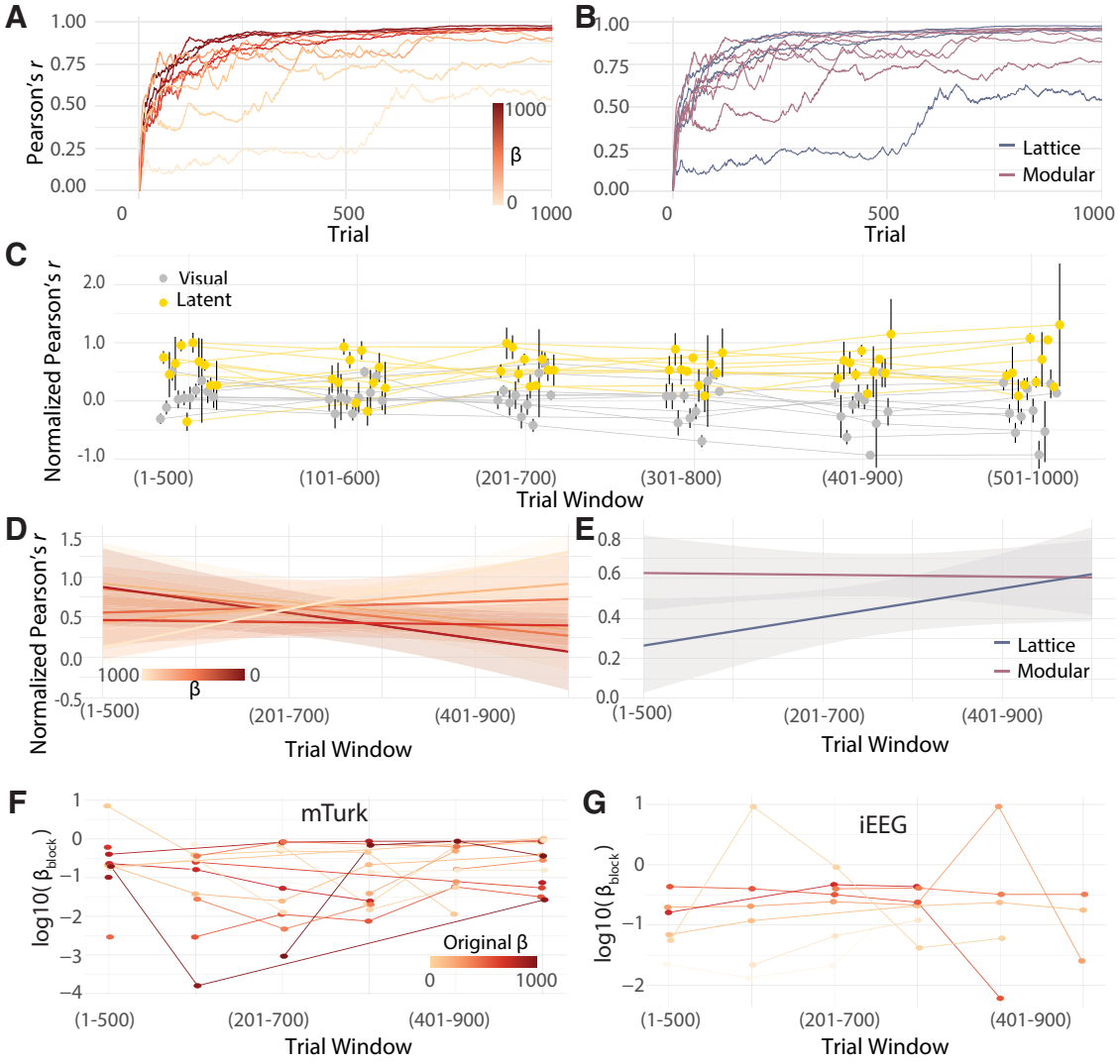

Figure 6. Time varying estimates of latent spaces. ***A***, The correlation between the infinite-time estimate of the latent space $\hat{A}$ and the finite-time estimate of the latent space $\hat{A}(t)$ at each trial t. Estimates with different β values are shown in different shades of orange. ***B***, The same information as that displayed in panel ***A***, but now colored by graph type. Walks from lattice graphs are shown in blue and walks from modular graphs are shown in pink. ***C***, The correlation between the neural activity dissimilarity matrices and two spatial templates in two blocks of 500 trials. The correlation to the estimated latent space template is shown in yellow, and the correlation to the visual template is shown in gray. Error bars indicate the standard deviation of correlations over contacts. ***D***, Change in correlation of neural dissimilarity matrices to the latent space as a function of trial window, colored by the β value of the participant. Shaded region indicates the 95% confidence interval of the estimated slope. Each line shows a linear fit of one participant's change in correlation over time. Shaded regions indicate the 95% confidence interval. ***E***, The same information as that displayed in panel ***D***, but now separated by graph type rather than by β value. ***F***, β estimates taken from 500 trial sliding windows of data from the mTurk cohort. The color of the line reflects the β value estimated from all of the data. Missing values indicate the β value was an extreme value of either 0 or 1000. ***G***, The same as in panel ***F***, but for the iEEG cohort.

values calculated using all trials. We then averaged similarity values over all contacts and used a linear mixed-effects model to assess whether participants' neural activity was more similar to the estimated latent space than the visual space, and if that similarity grew over time. In line with our hypothesis, we found significantly larger increases in correlation coefficients between the neural space and the latent space than in correlation coefficients between the neural space and visual space (linear mixed-effects model $F_{window*space} = 6.755$, $p_{window*space} = 0.011$), even in the first 500 trials (paired t test $t = 3.81$, $p = 0.004$).

We next asked whether these changes in similarity were modulated by β values or by graph type. Our simulations suggest that participants with larger β values should show greater similarity to the latent space early during learning. Accordingly, we tested whether β values predicted the magnitude and rate of change of the normalized correlation coefficients between neural activity and the estimated latent space. We found significant changes in the magnitude of the normalized correlations associated with β values, as evidenced by a significant interaction between β values and the change in similarity over blocks

(linear mixed-effects model $F_{\beta} = 2.65$, $p_{\beta} - 0.110$, $F_{window*\beta} = 6.12$, $p_{window*\beta} = 0.017$; Fig. 6*D*). Specifically, smaller β values have positive relationships between the neural similarity to the estimated latent space and the trial window (increasing with learning), whereas larger β values have negative relationships. While the observed variation in the convergence of neural data to $\hat{A}$ by β value was recapitulated in our simulations, we saw an additional trend toward less convergence over time that is not present in the model.

We also wished to determine whether changes in similarity (between neural activity and the latent space) over time were consistent across graph types, as predicted by our simulations. We found a significant interaction between the change in similarity over blocks and the graph type, which was not predicted by our model (linear mixed-effects model $F_{graph} = 5.06$, $p = 0.029$; Fig. 6*E*). Participants who saw sequences from lattice graphs tended to have more positive slopes, whereas participants who saw sequences from modular graphs tended to have more negative slopes. It is worth noting that this separation by graph type reflects the significant interaction observed between reaction time and graph type in this same cohort.

Lastly, we sought to examine whether the discrepancies between our simulations and observations could be narrowed by relaxing the assumption that β values remain static during learning, and instead hypothesizing that finite-time estimates of β per block would diverge more from the infinite-time estimates of β in the later windows. To test this hypothesis, we recalculated β values for each participant in the same blocks of 500 trials. We then tested whether β values changed consistently across the population over blocks of trials. While we observed substantial variability in β values for some participants, there were no consistent differences across the population (linear mixed-effects model $F_{window} = 0.14$, $p_{window} = 0.714$; Fig. 6*F*,*G*).

Discussion

In this work, we sought to better understand the neural correlates of latent space estimation from temporal sequences of stimuli that evince particular transition probability structures encoded as graphs. We used behavioral modeling to identify individual variations in temporal discounting and iEEG data recorded during learning to answer four main questions: (1) Do individuals in our iEEG cohort show behavioral evidence of learning an estimate of the latent space? (2) Which brain regions have neural activity that reflects these estimates? (3) Does the structure of neural activity facilitate the identification of task-relevant features? (4) Upon what time scale does neural structure appear, and is that timescale modulated by temporal discounting or graph structure? To answer question (1), we first had participants respond to cues generated from two different latent spaces: one with a modular structure, and one with a lattice structure. We found evidence that our iEEG cohort became faster and more accurate over time, consistent with participants learning the latent space and better anticipating upcoming stimuli. To answer question (2), we fit a model of learning that utilizes temporal discounting during latent space estimation, and found regions where neural activity has a similar structure to these estimates. For stimulus-evoked activity, most regions identified, regardless of the graph used to generate the sequences, were located in the temporal lobe, with some additional involvement of frontal structures.

Previous work investigating Euclidean spatial representations found that low-dimensional projections of the estimated space readily identified task-relevant features like boundaries and modules (Stachenfeld et al., 2014). This work motivated us to ask question (3), and accordingly to test whether there was evidence for the same identification of modules in our neural data. We found that for each participant who saw sequences drawn from a modular graph, low-dimensional projections of neural activity in the selected temporal and frontal regions accurately separated each module, misidentifying at most one stimulus. Interestingly, this separability was not achieved as consistently in the estimated latent space itself, suggesting the possibility that neural processing enhances the separability of task-relevant features such as modules. Lastly, we leveraged the neural recordings taken during latent space learning to ask question (4), and accordingly to test predictions about how quickly participants acquire their estimates of the latent space. Our model predicted that estimates of the latent space would be formed within the first 500 trials, and that participants with stronger temporal discounting would converge faster. We found evidence in support of these hypotheses, and also additional differences in latent space learning based on graph type that were not predicted by our model. Ultimately, we determined where and when neural activity during a sequential reaction time task reflects individual variation in behavior, and how that activity related to recent theories that extend concrete cognitive maps to abstract spaces.

Insights into probabilistic sequence learning

Previous work in probabilistic sequence learning has demonstrated that participants reacting to cues drawn from a random walk on a graph become sensitive to features of latent structure for a wide variety of graphs, with different numbers of stimuli, and across different sensory domains (Schapiro et al., 2013; Karuza et al., 2017, 2019; Kahn et al., 2018; Lange et al., 2019; Pudhiyidath et al., 2020). Here, we significantly extend this literature by adapting a version of these tasks for use in patient populations with iEEG recordings. Using our adapted task, we find that both a healthy cohort recruited via Amazon's mTurk and an iEEG cohort show evidence of learning, albeit with some differences in the nature of that learning.

We found that our mTurk cohort shows significant decreases in reaction time with increasing trial number, while our iEEG cohort shows decreases only across longer timescale blocks of 250 trials. While learning rates varied across the two cohorts, iEEG patients still performed the task with high accuracy (Fig. 2*A*,*B*), as expected given their cognitive capacities (Parvizi and Kastner, 2018). The slower learning is consistent with other work demonstrating poorer task performance in participants with epilepsy compared with controls (Bulteau et al., 2001; Parvizi and

Kastner, 2018). Patients with drug-resistant epilepsy were shown to have statistically significant decreases in task performance assessing motor function and cognitive attention (Motamedi and Meador, 2003), both of which are required for our experiment. However, our two cohorts are not matched on demographics or testing environment, making it difficult to determine whether differences are because of underlying epilepsy-related cognitive deficits or other factors.

We next tested for evidence of learning based on an increase in accuracy over time. While the mTurk cohort had relatively high accuracy throughout the experiment, their performance significantly decreased over time. While increasing speed and decreased accuracy are not necessarily indicative of disengagement from the task (Förster et al., 2003), this finding raises the possibility that some of the observed decreases in reaction time might be because of a decrease in correct responses. One possible explanation is a decrease in cognitive demand and arousal, leading to task disengagement and lower accuracy (Dehais et al., 2020; van der Wel and van Steenbergen, 2018). In contrast to the mTurk cohort, the iEEG cohort shows an initial increase in accuracy, followed by a decrease in accuracy during later trials. Individuals with temporal lobe epilepsy tend to perform worse on tasks that demand higher order cognition and attention (Hudson et al., 2014); and therefore, it may be easier to engage with simpler tasks. While this quadratic relationship with accuracy still suggests a lower engagement with the task as time goes on, it corroborates the conclusion that the task is better suited for the iEEG cohort.

The distributions of β values also provide further evidence that the task was not cognitively demanding enough for the mTurk cohort. We find that ~30% of individuals in the mTurk cohort had β values equal to 0. This proportion in larger than the iEEG cohort (approximately 8%), and previous studies using more complicated tasks (approximately 20%; Lynn et al., 2020a). Some of the disparity between previous research and our study likely stems from the fact that we use 500 fewer trials than previous work, which makes it more difficult for the gradient descent algorithm used to fit β values to converge. However, the disparity between cohorts in this study supports the conclusion that mTurk participants did not rely on the structure of the latent space as much as iEEG participants.

The decreases in reaction time and increases in accuracy with trial number suggest that participants are learning the task. It is true, however, that these improvements could be driven by learning features of the task that are independent of the underlying latent space. We do not think this explanation is likely, however, because most individuals have non-zero values of β despite the short task. To further demonstrate evidence for learning the underlying latent space, we tested for differences in reaction time based on graph type and across modules. Previous research has found that participants tend to react faster to cues drawn from modular graphs than to those drawn from lattice or random graphs, and from transitions between cues within the same module than between cues that span different modules (Kahn et al., 2018). We recapitulate the cross-module difference finding, but not the graph type difference finding in our mTurk population. The latter could be because of the fact that the task we used was significantly simpler than that previously employed (Kahn et al., 2018), and hence did not entail the same learning complexity. There were also design differences between previous work and the current study; for example, here we used simpler motor commands (using one rather than two fingers at a time), fewer trials, breaks with rewarding feedback, and fewer unique stimuli. In our iEEG cohort, we found no reaction time differences across modules, but different rates of learning based on underlying graph type. However, future work with either a more complex task or a larger number of subjects is necessary to further validate this result.

Insights into neural involvement in latent space estimation

To complement our study of behavior, we next probed the neural correlates of latent space estimation. We identified regions whose activity has a structure most similar to each individual's estimated latent space (rather than to the true latent space). In performing this identification, we used a short window of activity locked to the stimulus, to the response, or to the middle of all trials. In contrast to the slower temporal resolution of metabolic neuroimaging techniques, iEEG allows for the use of short temporal windows to investigate neural activity structure in a time-resolved manner thereby providing insight not only into where, but also into when, structural representations emerge. We also compared two different similarity matrices to rule out possible alternative explanations of the observed structure that were unrelated to the estimated latent spaces. The first is a null model that takes the empirical trial data, and reorders it around a single point. Unlike shuffling trial order, this model preserves autocorrelative features of the data and ensures that the observed similarity is specific to the observed walk sequence (Aru et al., 2015); the second comparison is to a lower-level feature of stimulus appearance: the visual distance between highlighted stimuli on the screen. We expected this structure to be reflected in neural activity, and indeed many regions included contacts that were similar to both latent and visual spaces. Including these comparisons allows us to assess the selectivity of regional activity for structural, rather than visual, information.

Stimulus- and response-locked activity implicate different brain regions

We found that for stimulus-locked activity, the most common regions identified were in the lateral, medial, and inferior temporal lobes. It is important to note that the temporal lobes also have more electrode coverage, and the identified regions made up between 4.6% and 27.3% of contacts in those areas; although not all highly sampled areas (e.g., superior temporal lobe) showed any contacts that were similar to either space. The presence of structure in this early evoked response is consistent with work

demonstrating that changes in tuning curves of neurons in the hippocampal part of the medial temporal lobe (in both human and nonhuman primates) reflect statistical similarities between stimuli (Miyashita, 1988; Reddy et al., 2015). For response-locked contacts, common areas still include the fusiform gyrus, but also include the inferior frontal gyrus, somatomotor area, and insula. This anatomic distribution is consistent with work showing that later stages of processing involve frontal regions receiving structural information from the hippocampus. Further, the involvement of motor regions is certainly intuitive during response planning (Siapas et al., 2005; Dehaene et al., 2015).

Amygdala involvement in cognitive map formation

The region with both the highest percentage of contacts identified and the highest selectivity for the latent space was the amygdala, followed by the middle temporal lobe. The amygdala is a region often associated with processing of emotional and rewarding stimuli, and is highly connected to the hippocampus, with which it interacts during emotional memory (Phelps, 2004). Notably, some previous work using single unit human iEEG recordings has also shown activity reflective of cognitive map building in the amygdala (Fried et al., 1997; Ekstrom et al., 2003). For example, in a study of single cell place selectivity in patients undergoing iEEG recording, the hippocampus demonstrated the most place-selective activity, yet cells in other parts of the medial temporal lobe, including the amygdala, showed selectivity as well (Ekstrom et al., 2003). Additionally, nonhuman primate studies have shown representations of abstract contexts for nonemotional stimuli in the amygdala (Saez et al., 2015). Ultimately, our results corroborate these findings that amygdala activity can reflect abstract spaces.

Middle temporal lobe involvement in cognitive map formation

The second region identified, the middle temporal lobe, has also been identified in other iEEG studies of statistical learning. Previous work studying lower-level statistical learning using iEEG also identified primarily lateral temporal cortex, and little involvement of the hippocampus and entorhinal cortex (Henin et al., 2021). Much work in human iEEG and fMRI implicates a broader range of temporal regions than comparable work in rodents (Buzsáki and Moser, 2013; Schapiro et al., 2013; Henin et al., 2021). This trend is likely partially because of the different cognitive and behavioral demands between species, but also raises the possibility of compensatory mechanisms in cohorts undergoing iEEG monitoring because of pathology in medial temporal lobes. This possibility cannot be ruled out completely, and therefore findings should ideally be corroborated in recordings from a healthy population. Some evidence of the role of lateral temporal lobe activity in learning a latent space from sequences exists in nonepileptic populations. Specifically, fMRI studies using similar tasks have also identified the interior temporal cortex to be reflective of some features of higher-order structure but not reflective of the estimates of latent spaces as a whole (Schapiro et al., 2013).

Medial temporal lobe involvement in cognitive map formation

Much of the work in rodent and human latent space learning has focused on the hippocampal, rather than lateral temporal lobes and amygdala (Buzsáki and Moser, 2013). Here, the hippocampus and sublobar temporal white matter are both implicated in our similarity analysis at higher levels than many neighboring medial temporal regions. For example, the entorhinal cortex activity is most commonly associated with low-dimensional projections of temporally discounted maps rather than the maps themselves (Stachenfeld et al., 2014; although there are some exceptions; Garvert et al., 2017). Similarly, the parahippocampal gyrus is most often associated with features of spatial exploration not tied to the underlying latent space (Epstein, 2008; although there are some exceptions; Aguirre et al., 1996). In line with these theories, we find that the hippocampus has more contacts than either the entorhinal or parahippocampal cortices, supporting its important role in latent space learning. However, these areas are less common and less selective than the nearby lateral temporal structures and the amygdala in our data.

Other brain regions' involvement in cognitive map formation

While the most common regions identified in our study and in previous work were in the temporal or frontal lobes, we observed multiple contacts in a wider distributed set of regions, including the insula, supplementary motor area, and precentral gyrus. Frontal areas, especially those in the medial prefrontal and orbitofrontal cortices, have been implicated in latent space learning, and are thought to be required at later stages than are medial temporal regions (Wilson et al., 2014; Desrochers et al., 2015; Brunec and Momennejad, 2019). Consistent with these observations, we found that response-locked activity shows more involvement of these frontal areas. Additionally, some work in humans has shown that activity that reflects the estimated latent structure (e.g., in place and grid cells) is much more spatially distributed than in rodents, leading to theories that most of the cortex is actually capable of forming these representations (Bao et al., 2019; Viganò and Piazza, 2020). Our results are in line with these theories, and support the conclusion that diverse brain regions could support temporally discounted estimates of the latent space. Taken together, neural activity most represented latent space estimates in the amygdala and middle temporal lobe when locked to the stimulus, whereas they most represented latent space estimates in the supplementary motor area and inferior frontal gyrus when locked to the response. These observations indicate that brain representations of learning are spatially distributed.

Importance of low-dimensional separation of task features

Studies investigating representations of spatial environments have pointed out the usefulness of low-dimensional representations for learning to navigate (Stachenfeld et al., 2014; Whittington et al., 2020).

Evidence for dimensionality reduction of neural signals has been observed in neural structures at three distinct scales: single neurons, anatomic regions, and the whole brain (Pang et al., 2016; Stringer et al., 2019; Mack et al., 2020; Zhou et al., 2020a). Broadly, dimensionality reduction of neural signals is thought to enable the brain to easily extract important, often changing information and facilitating the development of a sparse, efficient neural code for items in the environment (Pang et al., 2016; Beyeler et al., 2019). For dimensionality reduction of cognitive maps specifically, much work has focused on the medial temporal lobe. For example, the hippocampus has been functionally modeled as a variational autoencoder that continuously compresses incoming structural and sensory information to identify similar contexts (Whittington et al., 2020). Additionally, properties of grid cells (Stachenfeld et al., 2014), commonly but not exclusively found in the entorhinal cortex (Long and Zhang, 2021), can be explained by the eigenvectors of a temporally discounted estimate of the latent space. Importantly, these low-dimensional bases identify borders and modules in simulated spaces, the same features thought to be useful for successful navigation (Stachenfeld et al., 2014).

We asked whether these modeling observations were recreated in an abstract relational space. Using linear discriminant analysis, we found that modules are highly discriminable in individuals who saw sequences drawn from a modular graph. Interestingly, many of the estimated latent spaces show the same level of discriminability, although some show levels far lower. Upon further investigation, we find that the discriminability of estimated latent spaces was determined by the associated β value. We chose linear discrimination as a conservative estimate of separability, which is biologically implementable in theory by few neurons whose firing mimics the low-dimensional bases (Pagan et al., 2013); however, other methods of identifying modules are theoretically possible (Fusi et al., 2016).

The discussion of these findings raises the possibility that neural systems are transforming or building estimates of latent spaces in a way that enhances the separability of modules. One hypothesis is that the increased separability in low-dimensional space arises from neurons with high-dimensional, combinatorial responses to individual stimuli (Fusi et al., 2016). These types of neurons are thought to be present in associative areas such as the frontal cortex and medial temporal lobes (Fusi et al., 2016). It is hence intuitively plausible that regions in lower-order areas are less able to separate modules, but potentially more able to distinguish individual stimuli (Fusi et al., 2016). While these theories are based on the function of individual neurons, similar ideas can be extended to neuronal populations. Accordingly, future work could test whether divergences of neural activity from the estimated latent space increase the separability of modules at all places on the neural processing hierarchy, or only at more transmodal areas. The observation that neural dissimilarity better separates modules than the corresponding latent space estimation presents interesting directions for further investigation independent of validations of those theories.

Limitations

Here, we have put forth new evidence for neural correlates of latent space learning, although these results should be interpreted in light of the various limitations of our study. Many of our analyses focused on individual participants, an approach that is especially well-suited for iEEG analysis given the small and heterogeneous samples. However, some results, including evidence for learning and temporal changes in neural similarity structure, were assessed at the group level. To supplement these findings, we also present larger behavioral cohorts and numerical simulations. Despite these techniques, our group level results would be further strengthened by replication in larger samples.

Additionally, we sought to identify the regions whose activity was structured most similarly to the estimated latent space. Our approach involved selecting contacts with stronger correlations than 95% of null models. This selection process means that there is a chance that some contacts would be retained because of basic features of neural activity, and not because of task structure. Because of this fact, we highlight the regions where multiple contacts were identified, reducing the likelihood that our conclusions depend on false positives. We approached identifying contacts with activity structure similar to the latent space in a data driven manner and, therefore, expected the same pattern of activity in all regions. We also grouped all identified regions together when investigating properties of low-dimensional projections of neural activity structure. However, there is good evidence that specific regions, or even locations within the same region might be active at different times (Brunec and Momennejad, 2019) or use slightly altered transformations of the estimated space (Garvert et al., 2017). Identifying these differences is an important pathway for future analysis, but would require a larger cohort, where more individuals reliably show activity in the regions of interest, or a hypothesis-driven rather than data-driven assessment of regional contributions.

Lastly, we show evidence that some predictions of how latent spaces are learned over time are borne out in neural data. However, our model only uses one relatively simple learning rule. Other work has tested a variety of learning rules that all give rise to temporally discounted latent space estimations, and has shown that some are more consistent with neural activity than others (Chien and Honey, 2019). Here, we do not intend to claim that the implemented rule was more accurately reflecting changes in neural activity than others, but simply to identify the ways that estimated latent spaces appear in neural activity. Future work investigating and comparing different learning rules would be a welcome contribution to the field.

Future directions

Studying latent space learning presents an exciting opportunity in neuroscience to connect theoretical models to both behavior and hypothesized neural mechanisms for the implementation of these models. Work in rodents has suggest that temporally discounted estimates of relational spaces are built through synchronization of cell

populations to θ rhythms (4–10 Hz; O'Keefe and Recce, 1993). Distinct populations of cells in the CA1 subfield of the hippocampus synchronize their firing to the peaks of θ rhythms; the firing of different cells then becomes bound together via plasticity to represent unique temporal contexts (Battaglia et al., 2011). Within the hippocampus, map-like firing patterns of these linked assemblies of neurons reflect physical relationships after exploring new environments (Skaggs et al., 1996). The phase of θ rhythms also synchronizes with activity in cortical areas such as the prefrontal cortex where information about temporal context is used for other processes (Siapas et al., 2005). In humans, θ rhythms have been implicated in tasks requiring estimates of an underlying latent space, including episodic memory, spatial navigation, and semantic memory (Johnson and Knight, 2015). However, there is also evidence that these rhythms are less important for human learning than for rodent learning, and some investigators even hypothesize that other mechanisms, such as saccades, are responsible for the synchronization of cell populations (Buzsáki and Moser, 2013). Similar studies to clarify the role of θ rhythms during latent space learning would extend the field appreciably.

While models that incorporate some form of temporal discounting are common in studies of cognitive maps (Dayan, 1993), the exact rules that govern updates to those maps are not well agreed on and can lead to different predicted behaviors across tasks (Chien and Honey, 2019). In our specific task, most of these update rules would lead to highly similar estimated latent spaces, although tasks with other features such as erroneous transitions and explicit forgetting can separate the models (Chien and Honey, 2019). Other individual-level designs similar to this one would be well suited to study individual variability in the types of rules used to update cognitive maps and would be an important addition to the literature. In addition to testing other existing models, the current maximum entropy framework could be updated to incorporate more features that we know to be important for statistical learning, such as rewarding reinforcement of edges, the length of the sequence, or explicit forgetting of old transitions seen earlier in the sequence. Identifying how these additions change estimates of the latent space and individual differences in the weighting of different components would also represent a significant improvement in our understanding of latent space learning.

Beyond connections to mechanistic neural implementations of these models, further extensions to more ecological contexts would also benefit our understanding of latent space learning, and how they influence diverse cognitive processes. Extensions of this theory to ecological network structures, and to different exploration strategies and walk types have already been discussed and implemented (Karuza et al., 2017; Lynn et al., 2020b; Zhou et al., 2020b). Nevertheless, our work suggests that further advancements could expand the theory to incorporate temporal variability in learning strategies. Here, we show preliminary evidence for a change in learning rates based on the extent of temporal discounting, and also a shift in the extent of temporal discounting used over time. One would expect that different amounts of temporal discounting might be better suited to different tasks, tasks occurring at different timescales, or even different stages of the same task. This intuition is consistent with work demonstrating that different brain regions, or even different parts of the hippocampus, are sensitive to different timescales of information (Brunec and Momennejad, 2019), which could potentially be related to different amounts of temporal discounting. Extension of this work to incorporate more dynamic models of learning would help us to better understand domain-general latent space learning, and further align the models of these behaviors with the evidence of their implementation in the brain.

References

Aguirre GK, Detre JA, Alsop DC, D'Esposito M (1996) The parahippocampus subserves topographical learning in man. Cereb Cortex 6:823–829.

Ambekar A, Ward C, Mohammed J, Male S, Skiena S (2009) Name-ethnicity classification from open sources. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 49–58. 28 June 2009- 1 July 2009, Paris, France.

Aru J, Aru J, Priesemann V, Wibral M, Lana L, Pipa G, Singer W, Vicente R (2015) Untangling cross-frequency coupling in neuroscience. Curr Opin Neurobiol 31:51–61.

Bao X, Gjorgieva E, Shanahan LK, Howard JD, Kahnt T, Gottfried JA (2019) Grid-like neural representations support olfactory navigation of a two-dimensional odor space. Neuron 102:1066–1075.e5.

Battaglia FP, Benchenane K, Sirota A, Pennartz CM, Wiener SI (2011) The hippocampus: hub of brain network communication for memory. Trends Cogn Sci 15:310–318.

Behrens TE, Muller TH, Whittington JC, Mark S, Baram AB, Stachenfeld KL, Kurth-Nelson Z (2018) What is a cognitive map? Organizing knowledge for flexible behavior. Neuron 100:490–509.

Bertolero MA, Dworkin JD, David SU, Lloreda CL, Srivastava P, Stiso J, Zhou D, Dzirasa K, Fair DA, Kaczkurkin AN, Marlin BJ, Shohamy D, Uddin LQ, Zurn P, Bassett DS (2020) Racial and ethnic imbalance in neuroscience reference lists and intersections with gender. bioRxiv. doi: 10.1101/2020.10.12.336230.

Betzel RF, Medaglia JD, Kahn AE, Soffer J, Schonhaut DR, Bassett DS (2019) Structural, geometric and genetic factors predict interregional brain connectivity patterns probed by electrocorticography. Nat Biomed Eng 3:902–916.

Beyeler M, Rounds EL, Carlson KD, Dutt N, Krichmar JL (2019) Neural correlates of sparse coding and dimensionality reduction. PLoS Comput Biol 15:e1006908.

Bornstein AM, Khaw MW, Shohamy D, Daw ND (2017) Reminders of past choices bias decisions for reward in humans. Nat Commun 8:15958.

Bowerman M (1980) The structure and origin of semantic categories in language-learning child. In: Symbol as sense. New York: Academic Press.

Brett M, Christoff K, Cusack R, Lancaster J (2001) Using the Talairach atlas with the MNI template. Neuroimage 13:85.

Brunec IK, Momennejad I (2019) Predictive representations in hippocampal and prefrontal chierarchies. bioRxiv. doi: 10.1101/786434.

Brunec IK, Moscovitch M, Barense MD (2018) Boundaries shape cognitive representations of spaces and events. Trends Cogn Sci 22:637–650.

Bulteau C, Jambaqué I, Dellatolas G (2001) Epilepsy, cognitive abilities and education. J Child Epilepsy 269:269–274.

Butts CT (2009) Revisiting the foundations of network analysis. Science 325:414–416.

Buzsáki G, Moser EI (2013) Memory, navigation and theta rhythm in the hippocampal-entorhinal system. Nat Neurosci 16:130–138.

Caplar N, Tacchella S, Birrer S (2017) Quantitative evaluation of gender bias in astronomical publications from citation counts. Nat Astron 1:e0141.
Chatterjee P, Werner RM (2021) Gender disparity in citations in high-impact journal articles. JAMA Netw Open 4:e2114509.
Chien HYS, Honey CJ (2019) Constructing and forgetting temporal context in the human cerebral cortex. bioRxiv 1.
Conrad EC, Tomlinson SB, Wong JN, Oechsel KF, Shinohara RT, Litt B, Davis KA, Marsh ED (2020) Spatial distribution of interictal spikes fluctuates over time and localizes seizure onset. Brain 143:554–569.
Constantinescu AO, O'Reilly JX, Behrens TE (2016) Organizing conceptual knowledge in humans with a gridlike code. Science 352:1464–1468.
Dahal P, Ghani N, Flinker A, Dugan P, Friedman D, Doyle W, Devinsky O, Khodagholy D, Gelinas JN (2019) Interictal epileptiform discharges shape large-scale intercortical communication. Brain 142:3502–3513.
Dayan P (1993) Improving generalisation for temporal difference learning: the successor representation. Neural Comput 5:613–624.
de Cheveigné A, Nelken I (2019) Filters: when, why, and how (not) to use them. Neuron 102:280–293.
de Cothi W, Barry C (2019) Neurobiological successor features for spatial navigation. bioRxiv 9:789412.
Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C (2015) The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. Neuron 88:2–19.
Dehais F, Lafont A, Roy R, Fairclough S (2020) A neuroergonomics approach to mental workload, engagement and human performance. Front Neurosci 14:1.
Desrochers TM, Chatham CH, Badre D (2015) The necessity of rostrolateral prefrontal cortex for higher-level sequential behavior. Neuron 23:918–926.
Dion ML, Sumner JL, Mitchell SM (2018) Gendered citation patterns across political science and social science methodology fields. Polit Anal 26:312–327.
Dworkin JD, Linn KA, Teich EG, Zurn P, Shinohara RT, Bassett DS (2020) The extent and drivers of gender imbalance in neuroscience reference lists. Nat Neurosci 23:918–926.
Ekstrom AD, Kahana MJ, Caplan JB, Fields TA, Isham EA, Newman EL, Fried I (2003) Cellular networks underlying human spatial navigation. Nature 425:184–188.
Epstein RA (2008) Parahippocampal and retrosplenial contributions to human spatial navigation. Trends Cogn Sci 12:388–396.
Epstein RA, Patai EZ, Julian JB, Spiers HJ (2017) The cognitive map in humans: spatial navigation and beyond. Nat Neurosci 20:1504–1513.
Fogarty JS, Barry RJ, Steiner GZ (2019) Sequential processing in the classic oddball task: ERP components, probability, and behavior. Psychophysiology 56:e13300.
Förster J, Higgins ET, Bianco AT (2003) Speed/accuracy decisions in task performance: built-in trade-off or separate strategic concerns? Organ Behav Hum Decis Process 90:148–164.
Fried I, MacDonald KA, Wilson CL (1997) Single neuron activity in human hippocampusand amygdala during recognition of faces and objects. Neuon 18:753–765.
Fulvio JM, Akinnola I, Postle BR (2021) Gender (im)balance in citation practices in cognitive neuroscience. J Cogn Neurosci 33:3–7.
Fusi S, Miller EK, Rigotti M (2016) Why neurons mix: high dimensionality for higher cognition. Curr Opin Neurobiol 37:66–74.
Garvert MM, Dolan RJ, Behrens TE (2017) A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. Elife 6:e17086.
Gershman SJ, Moore CD, Todd MT, Norman KA, Sederberg PB (2012) The successor representation and temporal context. Neural Comput 24:1553–1568.
Henin S, Turk-Browne NB, Friedman D, Liu A, Dugan P, Flinker A, Doyle W, Devinsky O, Melloni L (2021) Learning hierarchical sequence representations across human cortex and hippocampus. Sci Adv 7:1.
Howard MW, Fotedar MS, Datey AV, Hasselmo ME (2005) The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. Psychol Rev 112:75–116.
Hudson JM, Flowers KA, Walster KL (2014) Attentional control in patients with temporal lobe epilepsy. J Neuropsychol 8:140–146.
Janca R, Jezdik P, Cmejla R, Tomasek M, Worrell GA, Stead M, Wagenaar J, Jefferys JG, Krsek P, Komarek V, Jiruska P, Marusic P (2015) Detection of interictal epileptiform discharges using signal envelope distribution modelling: application to epileptic and non-epileptic intracranial recordings. Brain Topogr 28:172–183.
Johnson EL, Knight RT (2015) Intracranial recordings and human memory. Curr Opin Neurobiol 31:18–25.
Kahn AE, Karuza EA, Vettel JM, Bassett DS (2018) Network constraints on learnability of probabilistic motor sequences. Nat Hum Behav 2:936947.
Karuza EA, Kahn AE, Thompson-Schill SL, Bassett DS (2017) Process reveals structure: how a network is traversed mediates expectations about its architecture. Sci Rep 7:12733.
Karuza EA, Kahn AE, Bassett DS (2019) Human sensitivity to community structure is robust to topological variation. Complexity 2019:1–8.
Kidd E (2012) Implicit statistical learning is directly associated with the acquisition of syntax. Dev Psychol 48:171–184.
Kriegeskorte N, Mur M, Bandettini PA (2008) Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.
Lai L, Gershman SJ (2021) The psychology of learning and motivation, Ed 1, pp 1–38. New York: Elsevier Inc.
Lange KV, Miller CA, Weiss DJ, Karuza EA (2019) Sensitivity to temporal community structure in the language domain. Proceedings of The 41st Annual Meeting of the Cognitive Science Society 2071. July 24-27 2019, Montreal, Canada.
Lee SW, O'Doherty JP, Shimojo S (2015) Neural computations mediating one-shot learning in the human brain. PLoS Biol 13:e1002137.
Li G, Jiang S, Paraskevopoulou SE, Wang M, Xu Y, Wu Z, Chen L, Zhang D, Schalk G (2018) Optimal referencing for stereo-electroencephalographic (SEEG) recordings. Neuroimage 183:327–335.
Long X, Zhang SJ (2021) A novel somatosensory spatial navigation system outside the hippocampal formation. Cell Res 31:649–663.
Lynn CW, Kahn AE, Nyema N, Bassett DS (2020a) Abstract representations of events arise from mental errors in learning and memory. Nat Commun 11:2313.
Lynn CW, Papadopoulos L, Kahn AE, Bassett DS (2020b) Human information processing in complex networks. Nat Phys 16:965–973.
Mack ML, Preston AR, Love BC (2020) Ventromedial prefrontal cortex compression during concept learning. Nat Commun 11:46.
Maliniak D, Powers R, Walter BF (2013) The gender citation gap in international relations. Int Org 67:889–922.
Mercier MR, Bickel S, Megevand P, Groppe DM, Schroeder CE, Mehta AD, Lado FA (2017) Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: a glimpse on white matter signal. Neuroimage 147:219–232.
Mitchell SM, Lange S, Brus H (2013) Gendered citation patterns in international relations journals. Int Stud Perspect 14:485–492.
Miyashita Y (1988) Neuronal correlate of visual associative long-term memory in the primate temporal cortex. Nature 335:817–820.
Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ (2017) The successor representation in human reinforcement learning. Nat Hum Behav 1:680–692.
Motamedi G, Meador K (2003) Epilepsy and cognition. Epilepsy Behav 4 Suppl. 2:S25–38.
Nassar MR, Helmers JC, Frank MJ (2018) Chunking as a rational strategy for lossy data compression in visual working memory. Psychol Rev 125:486–511.

Newport EL, Aslin RN (2004) Learning at a distance I. Statistical learning of non-adjacent dependencies. Cogn Psychol 48:127–162.
O'Keefe J, Recce ML (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus 3:317–330.
Pagan M, Urban LS, Wohl MP, Rust NC (2013) Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. Nat Neurosci 16:1132–1139.
Pang R, Lansdell BJ, Fairhall AL (2016) Dimensionality reduction in neuroscience. Curr Biol 26:R656–R660.
Parvizi J, Kastner S (2018) Promises and limitations of human intracranial electroencephalography. Nat Neurosci 21:474–483.
Pennington J, Socher R, Manning CD (2014) GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing ({EMNLP}), pp 1532–1543. Association for Computational Linguistics.
Phelps EA (2004) Human emotion and memory: interactions of the amygdala and hippocampal complex. Curr Opin Neurobiol 14:198–202.
Pudhiyidath A, Roome HE, Coughlin C, Nguyen KV, Preston AR (2020) Developmental differences in temporal schema acquisition impact reasoning decisions. Cogn Neuropsychol 37:25–45.
Reddy L, Poncet M, Self MW, Peters JC, Douw L, Van Dellen E, Claus S, Reijneveld JC, Baayen JC, Roelfsema PR (2015) Learning of anticipatory responses in single neurons of the human medial temporal lobe. Nat Commun 6:8556.
Revell AY, Silva AB, Campbell Arnold T, Stein JM, Das SR, Shinohara RT, Bassett DS, Litt B, Davis KA (2021) A framework for brain atlases: lessons from seizure dynamics. bioRxiv. doi: 10.1101/2021.06.11.448063.
Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND (2017) Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput Biol 13: e1005768.
Saez A, Rigotti M, Ostojic S, Fusi S, Salzman CD (2015) Abstract context representations in primate amygdala and prefrontal cortex. Neuron 87:869–881.
Saffran JR, Aslin RN, Newport EL (1996) Statistical learning by 8-month-old infants. Science 274:1926–19LP.
Schapiro AC, Rogers TT, Cordova NI, Turk-Browne NB, Botvinick MM (2013) Neural representations of events arise from temporal community structure. Nat Neurosci 16:486–492.
Schapiro AC, Turk-Browne NB, Norman KA, Botvinick MM (2016) Statistical learning of temporal community structure in the hippocampus. Hippocampus 26:3–8.
Schapiro AC, Turk-Browne NB, Botvinick MM, Norman KA (2017) Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. Philos Trans R Soc Lond B Biol Sci 372:20160049.
Siapas AG, Lubenov EV, Wilson MA (2005) Prefrontal phase locking to hippocampal theta oscillations. Neuron 46:141–151.
Skaggs WE, McNaughton BL, Wilson MA, Barnes CA (1996) Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus 6:149–172.
Solway A, Diuk C, Córdova N, Yee D, Barto AG, Niv Y, Botvinick MM (2014) Optimal behavioral hierarchy. PLoS Comput Biol 10: e1003779.
Sood G, Laohaprapanon S (2018) Predicting race and ethnicity from the sequence of characters in a name. arXiv:1805.02109.
Stachenfeld KL, Botvinick MM, Gershman SJ (2014) Design principles of the hippocampal cognitive map. Adv Neural Inf Process Syst 27:1.
Stiso J, Caciagli L, Hadar P, Kathryn A, Lucas TH, Bassett DS (2021) Fluctuations in functional connectivity associated with interictal epileptiform discharges (IEDs) in intracranial EEG. bioRxiv. doi: 10.1101/2021.05.14.444176.
Stringer C, Pachitariu M, Steinmetz N, Carandini M, Harris KD (2019) High-dimensional geometry of population responses in visual cortex. Nature 571:361–365.
Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–289.
Ung H, Cazares C, Nanivadekar A, Kini L, Wagenaar J, Becker D, Krieger A, Lucas T, Litt B, Davis KA (2017) Interictal epileptiform activity outside the seizure onset zone impacts cognition. Brain 140:2157–2168.
van der Wel P, van Steenbergen H (2018) Pupil dilation as an index of effort in cognitive control tasks: a review. Psychon Bull Rev 25:2005–2015.
Viganò S, Piazza M (2020) Distance and direction codes underlie navigation of a novel semantic space in the human brain. J Neurosci 40:2727–2736.
Wagenaar J, Worrell GA, Ives Z, Dumpelmann M, Litt B, Schulze-Bonhage A (2018) Collaborating and sharing data in epilepsy research. J Clin Neurophysiol 176:139.
Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J (2016) Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137:188–200.
Wang J (2013) Classical multidimensional scaling. In: Geometric structure of high-dimensional data and dimensionality reduction. Berlin: Springer.
Wang JX (2021) Meta-learning in natural and artificial intelligence. Curr Opin Behav Sci 38:90–95.
Wang X, Dworkin JD, Zhou D, Stiso J, Falk EB, Bassett DS, Zurn P, Lydon-Staley DM (2021) Gendered citation practices in the field of communication. Ann Int Commun Assoc 45:134–153.
Whittington JC, Muller TH, Mark S, Chen G, Barry C, Burgess N, Behrens TE (2020) The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. Cell 183:1249–1263.e23.
Wilson RC, Takahashi YK, Schoenbaum G, Niv Y (2014) Orbitofrontal cortex as a cognitive map of task space. Neuron 81:267–279.
Zhou D, Lynn CW, Cui Z, Ciric R, Baum GL, Moore TM, Roalf DR, Detre JA, Gur RC, Gur RE, Satterthwaite TD, Bassett DS (2020a) Efficient coding in the economics of human brain connectomics. bioRxiv. doi: 10.1101/2020.01.14.906842.
Zhou D, Lydon-Staley DM, Zurn P, Bassett DS (2020b) The growth and form of knowledge networks by kinesthetic curiosity. Curr Opin Behav Sci 35:125–134.
Zhou D, Bertolero M, Stiso J, Cornblath E, Teich E, Blevins AS, Oudyk K, Michael C, Urai A, Matelsky J, Virtualmario, Camp C, Castillo RA, Saxe R, Dworkin J, Bassett D (2022) dalejn/cleanBib: v1.1.1 (1.1.2). Zenodo. Avaibale at. https://doi.org/10.5281/zenodo.4104748.