Learning brain dynamics for decoding and predicting individual differences

Luiz Pessoa[1,2*], **Chirag Limbachia**[1+], **Joyneel Misra**[2+], **Srinivas Govinda Surampudi**[2+], **Manasij Venkatesh**[2+], **Joseph JaJa**[2]

***For correspondence:** pessoa@umd.edu (LP)

[1]Department of Psychology and Maryland Neuroimaging Center, University of Maryland, College Park, MD, USA; [2]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA ([+]authors with comparable contributions listed alphabetically)

Abstract Insights from functional Magnetic Resonance Imaging (fMRI), as well as recordings of large numbers of neurons, reveal that many cognitive, emotional, and motor functions depend on the multivariate interactions of brain signals. To *decode* brain dynamics, we propose an architecture based on recurrent neural networks to uncover distributed spatiotemporal signatures. We demonstrate the potential of the approach using human fMRI data during movie-watching data and a continuous experimental paradigm. The model was able to learn spatiotemporal patterns that supported 15-way movie-clip classification ($\sim 90\%$) at the level of brain regions, and binary classification of experimental conditions ($\sim 60\%$) at the level of voxels. The model was also able to learn individual differences in measures of fluid intelligence and verbal IQ at levels comparable or better than existing techniques. We propose a dimensionality reduction approach that uncovers low-dimensional trajectories and captures essential informational (i.e., classification related) properties of brain dynamics. Finally, saliency maps were employed to characterize brain-region/voxel importance, and uncovered how dynamic but consistent changes in fMRI activation influenced decoding performance. When applied at the level of voxels, our framework implements a dynamic version of multivariate pattern analysis. We believe our approach provides a powerful framework for visualizing, analyzing, and discovering dynamic spatially distributed brain representations during naturalistic conditions.[1]

Introduction

As brain data become increasingly *spatiotemporal*, there is a great need to develop methods that can effectively capture how information across space and time support behavior. In the context of functional Magnetic Resonance Imaging (fMRI), although data are acquired temporally, they are frequently treated in a relatively static manner (e.g., in event-related designs, responses to short trials are estimated with multiple regression assuming canonical hemodynamics). However, a fuller understanding of the mechanisms that support mental functions necessitates the characterization of *dynamic* properties, particularly during the investigation of more naturalistic paradigms, including movie watching (***Cutting et al.*** (***2012***)), and other continuous paradigms (***Meyer et al.*** (***2019***); ***Limbachia et al.*** (***2021***)).

A central goal of neuroscience is to *decode* brain activity: to infer mental processes and pro-

[1]Code will be made publicly available when the submission is peer reviewed.

cesses from brain signals. In fMRI, most approaches are spatially constrained and consider voxels in a local neighborhood ("searchlight") or across a few regions. Such models are useful to decode stimuli or mental state, particularly when the problem is well understood and localized anatomically (***Haxby et al.*** (***2001***); ***Haynes and Rees*** (***2005***); ***Kamitani and Tong*** (***2005***); ***Naselaris et al.*** (***2011***)). Additional techniques decode brain activity based on whole-brain functional connectivity matrices (***Richiardi et al.*** (***2011***)), as well as machine learning methods (***Rubin et al.*** (***2017***)). The vast majority of decoding methods are static, that is, the inputs to classification are patterns of activation that are averaged across time (“snapshots”) (***Allefeld and Haynes*** (***2015***)). Some studies have proposed using temporal information in addition to spatial data (***Mourão-Miranda et al.*** (***2007***); ***Nestor et al.*** (***2011***); ***Janoos et al.*** (***2011***); ***Loula et al.*** (***2018***); ***Hutchinson et al.*** (***2009***); ***Chu et al.*** (***2011***)). In such cases, features used for classification are extended by considering a temporal data segment instead of, for example, the average signal during the acquisition period of interest.

Despite some progress in utilizing temporal information in decoding methods, key issues deserve to be further investigated:

- How can we characterize brain signals generated by dynamic stimuli in terms of generalizable (across participants) spatiotemporal patterns? How are such patterns distributed across both space and time? For example, although a recent study used recurrent neural networks to decode brain states from fMRI data, they investigated different working memory conditions (remembering faces, bodies, tools, or places) which are associated with stable brain states at the temporal scale of technique (***Fan et al.*** (***2019***)).
- Understanding the dimensionality of brain representations has become an important research question in recent years (***Byron et al.*** (***2009***); ***Buonomano and Maass*** (***2009***); ***Gao et al.*** (***2017***)). For example, for simple movements, neuronal signals across populations of neurons can be well approximated by a projection onto a two-dimensional representation (***Russo et al.*** (***2018***)). This question is now starting to be addressed in the fMRI literature (***Shine et al.*** (***2019***); ***Gao et al.*** (***2020***)). Here, we tackle the problem of learning low-dimensional spatiotemporal signals from dynamic stimuli, including movie watching and continuous experimental paradigms.
- If spatiotemporal patterns capture important properties of brain dynamics, do they capture information about individual differences that are predictive of behavioral capabilities?
- Although multivariate techniques, including neural networks, can be successful at the level of prediction, *interpretability* is of key importance in the context of neuroscience. Therefore, here we quantified the relative importance of spatiotemporal features, specifically how brain regions contribute to input classification as a function of time.
- Can spatiotemporal decoding be applied at both the region-of-interest and voxel levels? Extending it to the voxel level allows investigators to probe finer spatial representations underlying task conditions or states of interest, and opens the way for dynamic representational similarity analysis (***Kriegeskorte et al.*** (***2008***)).

In the present paper, we sought to address the above challenges by developing a *dynamic* computational framework built upon recurrent artificial neural networks.

Results

Our goal was to develop a spatiotemporal decoding framework to predict a stimulus, mental state, behavior, and/or personality traits based on brain signals. A key element of our model was a recurrent neural network based on Gated Recurrent Units (GRUs; ***Cho et al.*** (***2014***)) sensitive to current and past inputs (Figure 1A). To determine the generalizability of our findings, all results reported below were obtained with not used in any fashion for training and/or tuning. For a complete description of the architecture, please refer to Methods and Materials.

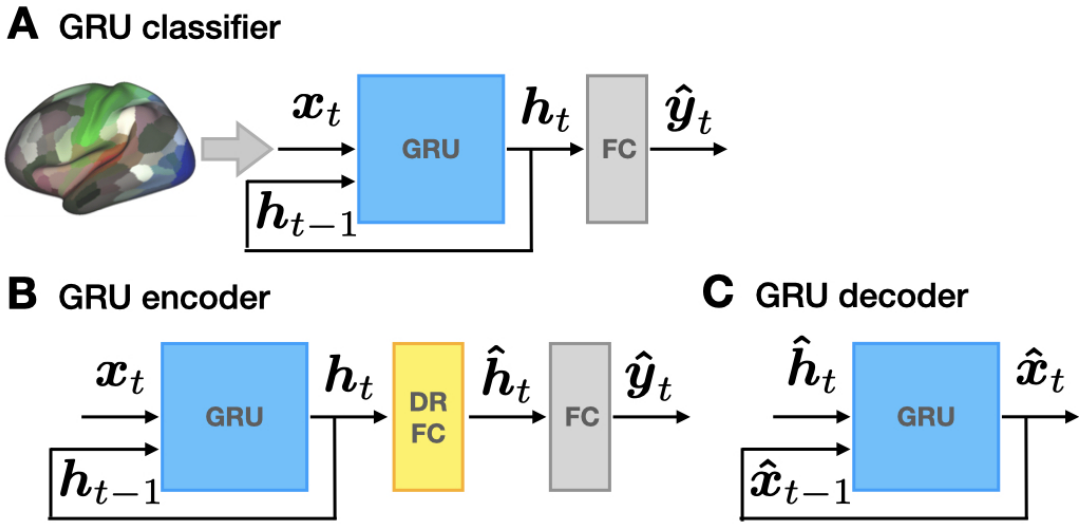


Figure 1. Model architectures based on neural networks with gated recurrent units (GRUs). (A) *Classifier*. At each time step, time series data, $\boldsymbol{x}_t$, provided inputs. The recurrent neural network transformed the inputs into a latent representation, $\boldsymbol{h}_t$, which then determined the output class scores, $\hat{\boldsymbol{y}}_t$. The unit with highest activation determined the model's prediction of the input stimulus at each time. (B) *Dimensionality reduction*. GRU outputs, $\boldsymbol{h}_t$, were first linearly projected to a lower-dimensional space using a fully-connected layer (DR-FC). Classification was then performed based on the low-dimensional representation, $\hat{\boldsymbol{h}}_t$. (C) *Input signal reconstruction*. A separate GRU was trained independently to reconstruct the original brain signals based on the low-dimensional signals, $\hat{\boldsymbol{h}}_t$.

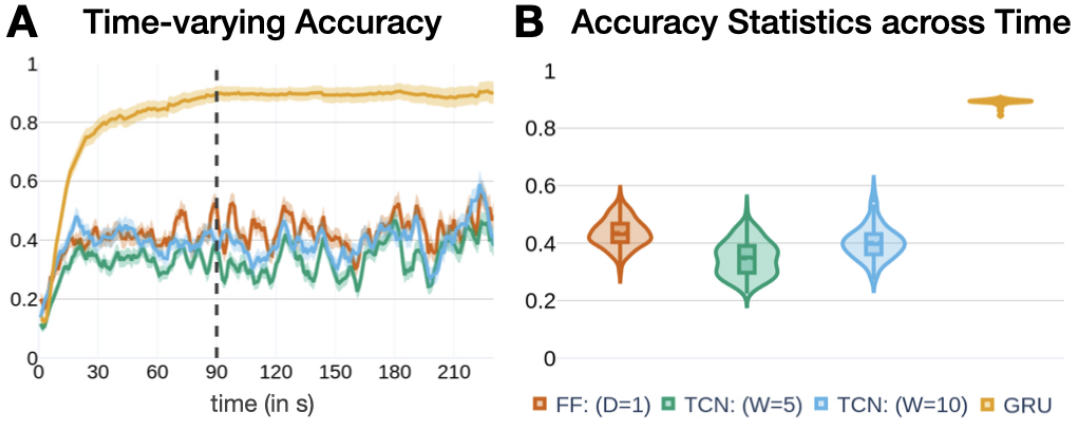


Figure 2. Prediction of movie clip (15-way classification). (A) Average clip prediction accuracy using the neural network architecture with gated recurrent units (GRUs) as a function of time. Accuracy increased sharply during the first 60 seconds, and stabilized around 90 seconds. Results using a feed-forward architecture (FF; 1 layer, 103 units), and temporal convolutional networks (TCN; kernel widths of 5 and 10) also shown. Error bars show the 95% confidence interval of the mean across test participants. (B) Summary of accuracy results after 90 seconds (transient period).

Generalizability of spatiotemporal patterns

To test the system's ability to generalize spatiotemporal brain patterns across participants, we employed the *GRU classifier* to predict movie clip labels from functional MRI data available from the Human Connectome Project. We performed 15-way classification as a function of time (Figure 2A). Accuracy increased sharply during the first 60 seconds, and stabilized around 90 seconds. Mean classification accuracy after this transient period was 89.46% (Fig. 2B). For completeness, a formal evaluation of chance performance based on the null distribution obtained through permutation testing resulted in a mean accuracy of 8.40% (1000 iterations; ***Ojala and Garriga*** (***2009***)).

Is temporal information necessary for clip prediction?

If capturing temporal information and long-term dependencies are important for classification, performance should be affected by temporal order. To evaluate this, we shuffled the temporal order of fMRI time series for test data. To preserve the autocorrelation structure in fMRI data while shuffling, we used a wavestrapping approach ***Bullmore et al.*** (***2004***). Classification accuracy reduced considerably to 48.44%.

We then compared the GRU classifier to alternative schemes. Our objective was not to benchmark our proposal against possible competitors, but to further probe properties captured by our architecture. First, using a simple feed-forward network with one hidden layer, classification accuracy was 43.51%. We constrained our choice of units (103) such that the the total number of parameters of the feed-forward classifier ($32,563$ parameters) was approximately equal to that of the GRU classifier ($32,559$ parameters). As feed-forward classifiers do not capture temporal dependencies, we next employed temporal convolutional networks. We used 25 kernels with a width (W) of 5 time steps such that the number of parameters of the classifier ($37,915$ parameters) approximated that of the GRU classifier (see Methods for kernel details). Classification accuracy was only 35.65%. We then increased the kernel width to 10 time steps ($75,415$ parameters). This numerically improved accuracy to 41.85%.

Together, the results above are consistent with the notion that temporal information is essential for high-accuracy classification, and that our GRU-based architecture is capable of capturing long-term dependencies that are missed by temporal convolutional networks, for example. Note, however, that temporal shuffling of the time series still allowed the GRU classifier to attain classification at around 50% correct. As the temporal shuffling was unique for each training permutation, order information was broken down, indicating that non-temporal information supported considerably above-chance performance.

Low-dimensional trajectories as spatiotemporal signatures

We investigated the dimensionality of the inputs required for successful stimulus prediction by using a supervised non-linear dimensionality reduction approach (Figure 1B). The original dimensionality of the input, $\boldsymbol{x}_t$, was the number of regions ($N_x = 300$). After learning, the GRU outputs provide a low-dimensional latent spatiotemporal representation of the input, which is encoded in the vector $\boldsymbol{h}_t$ ($N_h = 32$). To further reduce dimensions, GRU signals, $\boldsymbol{h}_t$, were linearly projected onto a lower-dimensional space, $\hat{\boldsymbol{h}}_t$, using a fully connected layer. We refer to this weight matrix as the *Dimensionality Reduction Fully-Connected (DRFC)* layer, and to this model as the *GRU encoder*. Since $\hat{\boldsymbol{h}}_t$ is a low-dimensional representation of the history of $\boldsymbol{x}_t$, the inputs are not treated independently, effectively leading to non-linear temporal dimensionality reduction. As in the original classifier, a final FC layer was used to predict labels based on $\hat{\boldsymbol{h}}_t$.

Performance with low-dimensional encoding was surprisingly high; 3 dimensions yielded 63.30% accuracy. In Figure 3A, we show low-dimensional signals for each clip, which we refer to as *trajectories*: consecutive low-dimensional $\hat{\boldsymbol{h}}_t$ states, $(\hat{h}_1(t), \hat{h}_2(t), \hat{h}_3(t))$, describe the trajectory in *state space*. In other words, at each time t, the value of each unit is plotted along the (x, y, z) axes. To quantify a notion of proximity between trajectories, we computed the Euclidean distance between them as a function of time (Figure 3B). To compute the distance between clips A and a reference clip B,

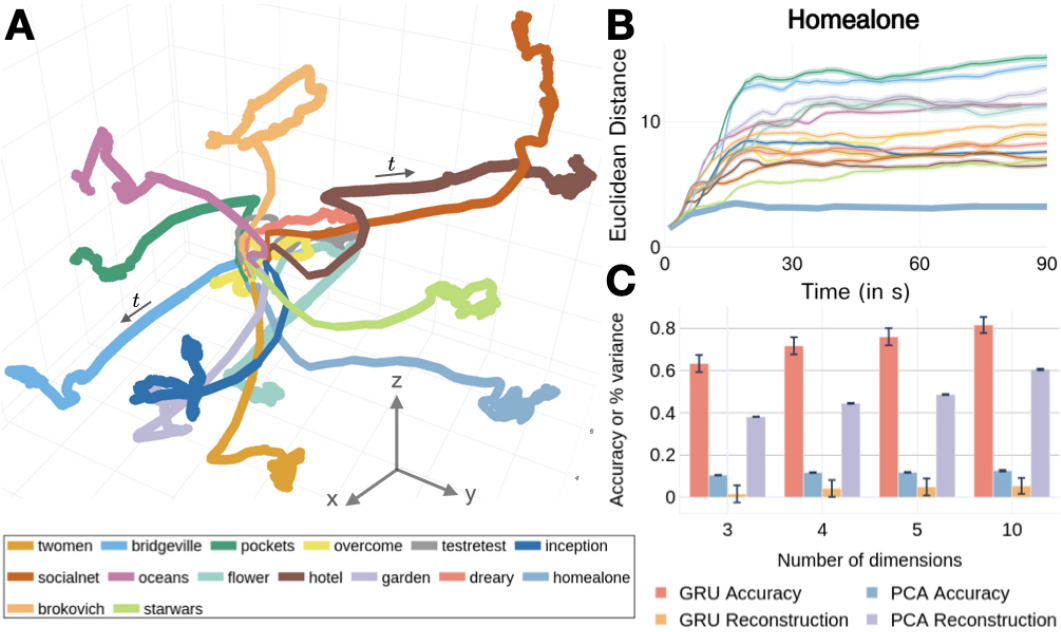


Figure 3. Low-dimensional trajectories. (A) Trajectories for all clips. Solid line: mean trajectory averaged across participants (line thickness is scaled by variance, which was highest at the end of the clip). All trajectories progressed away from the center (see sample arrows). The inset provides clip abbreviations. (B) Euclidean distance between trajectories. The Euclidean distance between the clip trajectory while watching *Home Alone* and the mean trajectory across participants for a second clip was computed. The thicker line corresponds to the distance of participants' *Home Alone* trajectories to the mean of this clip. (C) Clip prediction accuracy and fraction of variance captured after reconstruction using low-dimensional models. Error bars correspond to the standard error of the mean across participants.

Figure 3–Figure supplement 1. Euclidean distances between trajectories for all movie clips.

we first computed the mean trajectory of B averaged across participants. Then, for each participant's clip A trajectory, we computed the Euclidean distance between A and the mean trajectory of B at every time step. Accordingly, the proximity of a clip with itself was not zero (indicated by a thicker line), and reflected the variability of participant trajectories around the clip's average. As expected, the evolution of low-dimensional trajectories closely reflected the temporal accuracy obtained using the original GRU classifier. Clip trajectories were initially close to one another, but slowly separated during the first 60-90 seconds of the clip.

We further investigated low-dimensional projections with 4, 5, and 10 dimensions, at which point performance (81.56%) was not far (within 10%) from that of full dimensionality (Figure 3C). These results reveal that latent representations with as few as 10 dimensions captured essential discriminative information. The effectiveness of the GRU encoder in capturing low-dimensional information can be appreciated by applying principal components analysis on the input data, $\boldsymbol{x}_t$, which yielded very low prediction accuracy, although it was capable of recovering 60% of the input signal variance with 10 dimensions (Figure 3C).

To investigate the content of the latent space uncovered by the GRU encoder, we performed the following analysis. By construction, the low-dimensional vector captures information to classify the input signals. How much information about the input does this projection preserve? To evaluate this, we considered the low-dimensional signal, $\hat{\boldsymbol{h}}_t$, and used it to try to reconstruct the input signal, $\boldsymbol{x}_t$. Within our framework, it is natural to do so using a GRU decoder (Figure 1C), which reconstructs the input time series (the reconstructed signal is called $\hat{\boldsymbol{x}}_t$). Doing so captured a very modest amount of signal variance (less than 10%).

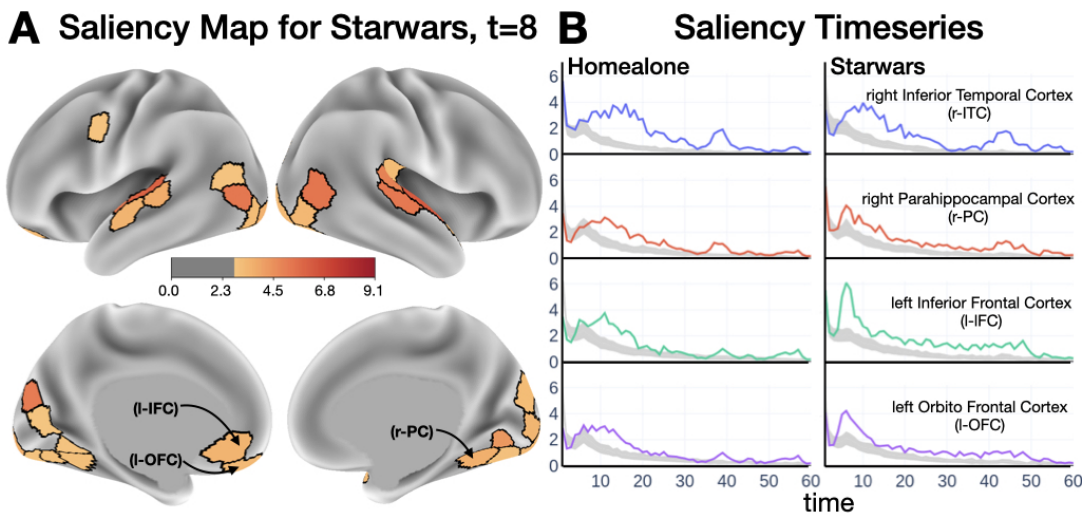


Figure 4. Determining brain contributions to classification. (A) Saliency map for the "Star Wars" clip. For illustration, the top 30 regions are shown. Color scale in arbitrary units. See also the video for dynamics. (B) Saliency values for the first 60 seconds of two clips. The gray bands corresponds to the 95-th percentile of null saliency values generated via permutation testing.

Figure 4–video 1. Saliency movie for Brokovich clip [link]
(https://drive.google.com/file/d/1Why34mgp4wedzLODZUl1Wyv3lbuPaOPN/view?usp=sharing)

Spatiotemporal importance maps

The GRU representation that supports classification lies in a latent space that is fairly disconnected from the original brain activation signals, $\boldsymbol{x}_t$. It is therefore important to relate GRU states to brain activation. We defined a *saliency* measure to capture the contribution of a brain region to classification as a function of time (see Methods). Figure 4A shows a saliency map at $t = 8$ for the *Star Wars* clip, also shown dynamically for the first 60 seconds (Video 1). The saliency time series in a few brain regions shows that values were relatively high initially and gradually decreased over the first minute of the clip (Figure 4B). The evolution of saliency for the *Home Alone* clip was somewhat similar (see also the saliency map video for the *Brokovich* clip in Figure 4).

Video 1. Saliency movie for the Star Wars clip [link]
(https://drive.google.com/file/d/1jaTgnHPhvouKM28-r3rGohZNPP3lvGfi/view?usp=sharing)

Predicting behavior

Recent studies have employed brain data to predict participants' behavioral capabilities or personality-based measures (***Beaty et al.*** (***2018***); ***Dubois et al.*** (***2018b***,a); ***Hsu et al.*** (***2018***); ***Jiang et al.*** (***2018***)). We tested the extent to which spatiotemporal information captures individual-based measures based on our recurrent neural network framework. Here, we targeted individual scores for fluid intelligence and verbal IQ. The same approach used for classification was employed, with the model predicting a scalar, namely, a participant's score. We used all clips for training (rather than training a separate model based on each clip) to promote learning representations that are not idiosyncratic to a particular clip.

Figure 5 shows model performance temporally for the *Star Wars* clip, one of the best predicting clips. At every time point, predicted scores are correlated with actual participants' scores. For fluid intelligence, the observed correlations were often above chance levels (gray zone indicates 95% confidence level at specific time t), but more modestly for verbal IQ. As our model made predictions as a function of time, we developed a permutation-based test to evaluate the predictions while controlling for multiple comparisons (fluid intelligence, $p = 0.0013$; verbal IQ: $p = 0.0038$). The

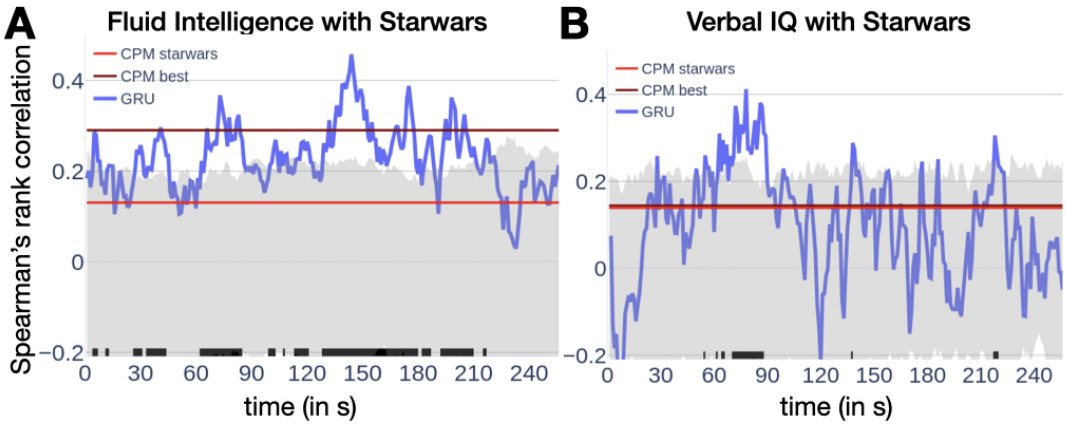


Figure 5. Prediction of behavior. (A) Prediction of fluid intelligence scores as a function of time. Prediction (blue) fluctuates considerably but consistently exceeds chance values (indicated by the tic marks at the bottom). Values obtained by connectome-predictive modeling (CPM) are indicated for comparison (red: CPM applied to *Star Wars* data; maroon: highest value applying CPM across all clips). The gray region indicates the 95th percentile region based on permutation testing (N.B.: applies to our method only, not CPM). (B) Prediction of verbal IQ as a function of time. Only short periods of time of the Star Wars clip exceeded chance levels.

Figure 5–Figure supplement 1. Fluid Intelligence predictions for all movie clips.

Figure 5–Figure supplement 2. Fluid Intelligence predictions: null distributions.

Figure 5–Figure supplement 3. Verbal IQ predictions for all movie clips.

Figure 5–Figure supplement 4. Verbal IQ predictions: null distributions.

supplementary figures shows results for all clips. For reference, we compared our approach to connectome-based predictive modeling based on patterns of functional connectivity (CPM; ***Shen et al.*** (***2017***)), possibly the state-of-the-art in this regard (see also ***Beaty et al.*** (***2018***); ***Dubois et al.*** (***2018a***)). We include values predicted with this technique to provide an informal comparison, as this method provides a single prediction per clip, unlike out approach which is time varying.

Dynamic multivariate pattern analysis

Thus far, we applied our model to region-based activation patterns. The approach can also be employed to perform voxel-based (or grayordinate-based) dynamic multivariate pattern analysis (MVPA). Here, we applied our model to the prediction of experimental conditions from a dataset collected in our laboratory (***Limbachia et al.*** (***2021***)). Briefly, in the "moving circles" paradigm, over periods of three-minute blocks, two circles moved on the screen, at times approaching and at times retreating from each other. If they touched, participants received a non-painful but highly unpleasant electrical shock. The movement of the circles was smooth but not predictable. In particular, the circle motions were set up to include multiple instances of "near misses": periods of approach followed by retreat; in such instances, the circles came close to each other and retreated just before colliding. Based on near-misses, we defined *approach* and *retreat* states: each lasted seven time steps (total of 11.25 seconds) during which the circles approached or retreated from each other.

We performed dynamic MVPA by using voxels from the anterior insula, a region strongly engaged by the experimental paradigm (***Meyer et al.*** (***2019***); ***Limbachia et al.*** (***2021***)). Classification accuracy ranged from 59-63% over the approach and retreat segments (the upper bound of the 95% confidence interval determined with permutation testing was 50.37%). We also computed saliency in a voxelwise manner; values were substantially higher at the outset of a approach/retreat segment and relatively lower for the subsequent time points. Figure 6 shows a snapshot at $t = 6$ seconds, illustrating the spatial organization of the map.

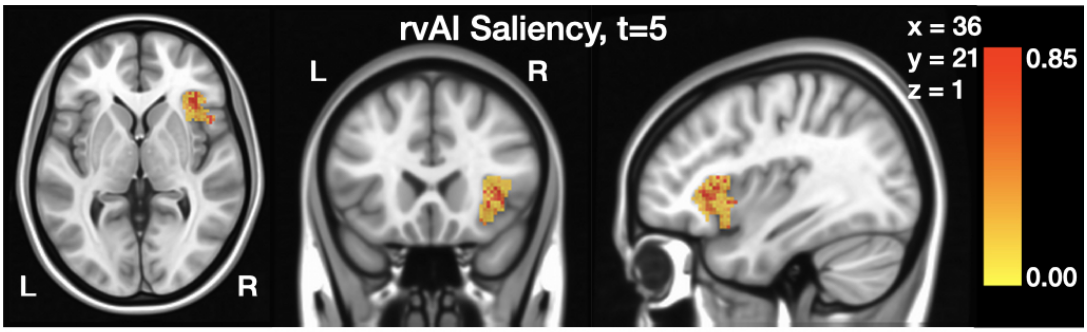


Figure 6. Voxelwise prediction of experimental conditions: threat vs. safe. Only voxels from the anterior insula were employed. Saliency values at $t = 6$ are shown; for illustration no thresholding was applied.

Discussion

We sought to characterize distributed spatiotemporal patterns of functional MRI data during movie watching and continuous task conditions. To do so, we employed recurrent neural networks sensitive to the long-term history of the input signals, and applied them to the problem of input classification. The model was tested on brain data both at the level of region of interest and voxels. All neural network training and model tuning were performed using a training set independent of the test set to establish the model's ability to generalize learned representations to unseen participants. The framework we developed is general and its subcomponents can be easily exchanged to utilize other algorithms (e.g., long short-term memory networks, or novel developments).

Spatiotemporal representations

Spatiotemporal information learned in a given set of participants generalized very well to new participants, where challenging 15-way classification was $\sim 90\%$ correct for movie clip data. These results are noteworthy because they suggest that activation patterns distributed across space and time are shared across participants when viewing naturalistic stimuli. To understand the type of information captured by our model, we compared it to a few simple schemes, including a simple feedforward network, and temporal convolutional networks. The low performance with these models ($\sim 40\%$) suggests that long-term dependencies are essential for input classification. Indeed, temporal shuffling of the input time series decreased performance to $\sim 50\%$. As temporal shuffling preserves the input signals but scrambles their sequence, these results reveal that the precise temporal order is key for the ability to generalize to unseen data.

Latent dimensionality

The question of the "inherent" dimensionality of brain signals has attracted considerable attention in recent years (***Byron et al.*** (***2009***); ***Buonomano and Maass*** (***2009***); ***Gao et al.*** (***2017***); ***Vyas et al.*** (***2020***)). Here, we investigated this issue for the movie clip data, where inputs were 300-dimensional based on the ROI time series. As the number of units in the hidden layer was 32, a considerable amount of dimensionality reduction was accomplished at the outset. Further investigation revealed that this signal could be further reduced to ~ 10 dimensions without compromising prediction accuracy appreciably. Strikingly, even with three dimensions, accuracy was relatively high (77.30%). Importantly, whereas the mapping onto a lower dimensionality space preserved classification information about the input category, it was very poor at reconstructing the input time series itself. Thus, the content of the latent low-dimensional representation was substantially distinct from the fMRI signal itself.

Saliency

Although prediction is valuable in applied settings (e.g., clinical diagnosis), our goal was to be able to characterize distributed spatiotemporal representations in the brain, too. A measure of saliency

uncovered brain regions and their time-varying importance to movie clip prediction. The contribution of brain areas was highest during the first minute in a manner that paralleled the time course of accuracy. Spatiotemporal saliency maps revealed that, whereas early visual and auditory regions contributed to prediction, multiple additional areas contributed too, supporting the idea that classification was not simply based on low-level movie features. For example, one of the regions with high salience (right inferior temporal cortex) included part of the fusiform gyrus, which is known to be highly responsive to faces. Other parts of inferior temporal cortex, including parahippocampal cortex which is involved in high-level object recognition, participated too. As further examples, the inferior frontal cortex and the orbitofrontal cortex supported clip classification, too. Together, these results support the idea that the overall model captured biologically relevant information, and shows how signal importance fluctuates spatiotemporally.

Individual differences

Recent work has sought to predict individual differences based on functional MRI data (***Shen et al.*** (***2017***); ***Beaty et al.*** (***2018***); ***Dubois et al.*** (***2018b***,a); ***Hsu et al.*** (***2018***); ***Jiang et al.*** (***2018***)). Our model was capable of predicting fluid intelligence and verbal IQ from unseen participants at levels comparable to, and possibly better than, established methods such as connectome-based predictive modeling, which is based on information from functional connectivity matrices (***Shen et al.*** (***2017***)). Some movies (e.g., *Star Wars*, *Social Net*, *Oceans*) performed much better at prediction than others, and at particular segments of the movie. This raises the intriguing possibility that those particular segments are better "tuned" to particular individual differences (in this case, fluid intelligence). Of course, larger datasets will be needed to adequately assess this possibility.

Dynamic multivariate pattern analysis

We applied the model developed here with inputs at the level of voxels, so as to implement dynamic MVPA. We tested the approach in a dataset involving continuous changes in threat level based on the proximity of two moving circles. Voxels from the anterior insula were used to predict stimulus condition (threat vs. safe). Although accuracy was relatively modest (∼60%), the results demonstrate the feasibility of the approach. Saliency maps also uncovered subsectors of the anterior insula with increased importance for prediction, illustrating how it is possible to identify voxels that contribute to decoding the experimental conditions, opening an avenue for the application of representational similarity analysis (***Kriegeskorte et al.*** (***2008***)) in a dynamic setting.

Other approaches

Decoding approaches using some temporal information have been used in the past (***Mourão-Miranda et al.*** (***2007***); ***Nestor et al.*** (***2011***); ***Janoos et al.*** (***2011***); ***Loula et al.*** (***2018***); ***Hutchinson et al.*** (***2009***); ***Chu et al.*** (***2011***)). Recurrent neural networks have also been employed for brain decoding more recently, albeit in some cases on relatively static task paradigms, such as working memory which generates stable brain states at the temporal scale of fMRI (***Fan et al.*** (***2019***); ***Thomas et al.*** (***2019***)); working memory conditions can be predicted even when temporal information is eliminated (***Venkatesh et al.*** (***2019***)). Because most fMRI datasets are typically small, they are often evaluated using cross-validation without assessing their generalization on held-out data (see discussion by ***Varoquaux et al.*** (***2017***)). The size of the datasets studied allowed us to test the model in a separate set of participants thereby evaluating the model's ability to generalize beyond trained data (see also ***Thomas et al.*** (***2019***)). Finally, we stress that the goal of the present study was not to describe a model that outcompetes others but to describe a general, modular modeling approach to investigating spatiotemporal dynamics of brain data, here applied to fMRI.

Conclusions

In the present paper, we developed a computational approach to study and characterize spatiotemporal brain signals in the context of fMRI. The framework can be applied to other types of brain

data for discovering and interpreting brain dynamics.

Methods and Materials

Data

Movie data

We employed Human Connectome Project (HCP; ***Van Essen et al.*** (***2013***)) data of participants scanned while watching movie excerpts (Hollywood and independent films) and other short videos, which we call "clips". The legend of Figure 3 provides further information. All 15 clips contained some degree of social content. Participants viewed each clip once, except for the *test-retest* clip that was viewed 4 times. Clip lengths varied from 65 to 260 seconds (average: 198 seconds).

We employed all available movie-watching data, except those of 8 participants with runs missing; thus we used $N = 176$ participants. Data were sampled every 1 second. The preprocessed HCP data included FIX-denoising, motion correction, and surface registration (***Glasser et al.*** (***2013***); ***T. Vu et al.*** (***2017***)). We analyzed data at the region of interest (ROI) level, with one time series per ROI (average time series across locations). We employed a standard 300-ROI cortical parcellation (***Schaefer et al.*** (***2018***)). Thus, the input time series consisted of a vector of $N_x = 300$ dimensions at each time, t. Individual ROI time courses were normalized by z-scoring (i.e. centered to zero mean and unit standard deviation) to help remove potential differences across runs/participants.

Moving-circles paradigm

We employed data from a separate dataset collected in our laboratory (***Limbachia et al.*** (***2021***)). Briefly, the experimental paradigm was as follows. Over a period of 3 minutes, two circles moved on the screen, at times approaching and at times retreating from each other. If they touched, participants received a non-painful, but highly unpleasant electrical shock. The movement of the circles was smooth but not predictable. In particular, the circle motions were set up with multiple instances of what we called "near misses": periods of approach followed by retreat; in such instances, the circles came close to each other and retreated just before colliding. Based on near-misses, we defined *approach* and *retreat* states: each lasted 6-9 seconds during which the circles approached or retreated from each other.

Data from $N = 61$ participants were employed, and were sampled every 1.25 seconds. Preprocessing steps included motion correction (ICA-AROMA via FSL). We analyzed data at the voxel level from the left anterior insula, as defined according to an anatomical template (***Faillenot et al.*** (***2017***)). The input time series consisted of a vector of $N_x = 745$ dimensions at each time, t.

Performance

Accuracy

At each point in time, t, we computed accuracy as the average true positive rate (TPR) across inputs of a given class (movie clip or experimental condition). Specifically, for each class label c:

$$\text{TPR}_c = \frac{\sum_{n=1}^{N_c} \text{I}\{\text{predicted_label}_n = c\}}{N_c} \tag{1}$$

where N_c is the total number of data samples belonging to class c, predicted_label_n is the label predicted by the model for the nth data sample, and I is an indicator function which equates to 1 if $\text{predicted_label}_n = c$ and 0 otherwise. We computed accuracy at each time step and visualized it as a function of time (see Figure 2).

Training/Testing split and Cross-Validation

Movie data from 100 participants were used for training, and the remaining 76 participants were used as a test set (thus completely invisible to model tuning/learning). To determine optimal hyperparameters for clip prediction, we employed a 5-fold cross-validation approach on the training

set. In each fold, participants in the training set were not included in the validation set. After determining optimal hyperparameters, we retrained the model on the entire training data. Final results were generated exclusively based on test data. A similar approach was used for the the moving-circles paradigm, where data from 42 and 19 participants were used as a training set and test set, respectively. For cross validation, the folds had size $\{9, 9, 8, 8, 8\}$.

Basic recurrent architecture for classification

Recurrent neural networks allow signals at the current time step to be influenced by long-term past information. Although it was originally difficult to develop effective learning procedures for these algorithms (e.g., vanishing/exploding gradients prevented them from learning relationships beyond 5-10 time steps; ***Bengio et al.*** (***1994***)), recent developments have largely overcome these challenges. Here, we employed a recurrent neural network architecture based on Gated Recurrent Units (GRUs; ***Cho et al.*** (***2014***)). Figure 1A shows the model. Briefly, inputs $\boldsymbol{x}_t$ were provided to a GRU network that produced a "hidden" representation, $\boldsymbol{h}_t$, of the incoming signal based on current and past signals. For classification purposes, the hidden-layer signals were transformed to create a vector of class scores, the maximum of which determined the model's prediction of the input. Next, we describe the model formally.

Time series data from units of interest (ROIs or voxels) comprised the input vector at each time, $\boldsymbol{x}_t$. The GRU network (with one or more layers) transformed its input, $\boldsymbol{x}_t$, non-linearly onto its output, $\boldsymbol{h}_t$, the hidden layer of the classifier system (see Appendix 1 for details). Because our goal was input classification, GRU outputs at each time, $\boldsymbol{h}_t$, passed through an additionally fully-connected (FC) layer of units generating a vector of class scores:

$$\hat{\boldsymbol{y}}_t = \text{Softmax}(W_{\text{FC}}\boldsymbol{h}_t + b_{\text{FC}}), \tag{2}$$

where Softmax is the generalized logistic activation function. As the outputs are in the range $[0, 1]$ and and add up to 1, they can be treated as probabilities. The output of the model was the unit label, i, such that it selected the class with largest probability: $\arg\max_i \hat{\boldsymbol{y}}_t$. As all operations were performed at every time, t, the final output was a time series of label predictions. For movie clips, we employed data from 300 ROIs, so the dimensionality of the input was $N_x = 300$. The GRU network employed a single layer with 32 units ($N_h = 32$). The dimensionality of $\hat{\boldsymbol{y}}_t$ was $N_y = 15$ to implement 15-way classification (based on the number of clips). For the moving-circles data, we employed data from 745 voxels ($N_x = 745$). The GRU network contained three layers of 32 nodes each; the third layer corresponded to the vector $\boldsymbol{h}_t$ ($N_y = 32$). The dimensionality of $\hat{\boldsymbol{y}}_t$ was $N_y = 2$ to implement threat vs. safe classification.

All weight matrices, W, required for GRU tuning (see Appendix 1) and the bias parameters, b, underwent supervised training. Training sought to minimize the cross-entropy (CE) loss between predicted labels, $\hat{\boldsymbol{y}}_t$, and true labels, $\boldsymbol{y}_t$ (as provided by the supervisor), and was evaluated at at each time step:

$$\mathcal{L}_{\text{CE}}\left(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t; \Theta\right) = -\sum_{i=1}^{N_y} y_t^i \log\left(\hat{y}_t^i\right), \tag{3}$$

where the superscript i indexes the dimension of the predicted and teaching signals. The set of trainable parameters was denoted as Θ. The cross-entropy attains its minimum value when the two distributions, $\hat{\boldsymbol{y}}_t and \boldsymbol{y}_t$, are identical. We defined the total loss function, $J(\Theta)$, as the average cross-entropy loss across time points (i.e., the average value across time of Equation 3), thus encouraging predicted labels to be as close to the true labels as possible across time. We minimized $J(\Theta)$ using the backpropagation through time algorithm ***Werbos*** (***1990***); gradient descent was optimised with the Adam optimizer (***Kingma and Ba*** (***2014***)). The model was implemented using TensorFlow (***Abadi et al.*** (***2015***)).

As our goal was not necessarily maximizing model performance, we explored a small set of potential architectures by varying the number of hidden layers and number of units per layer (Figure 7). For the region-based analysis, we explored the $\{1, 2, 3\} \times \{16, 32, 64\}$ space (units by layers).

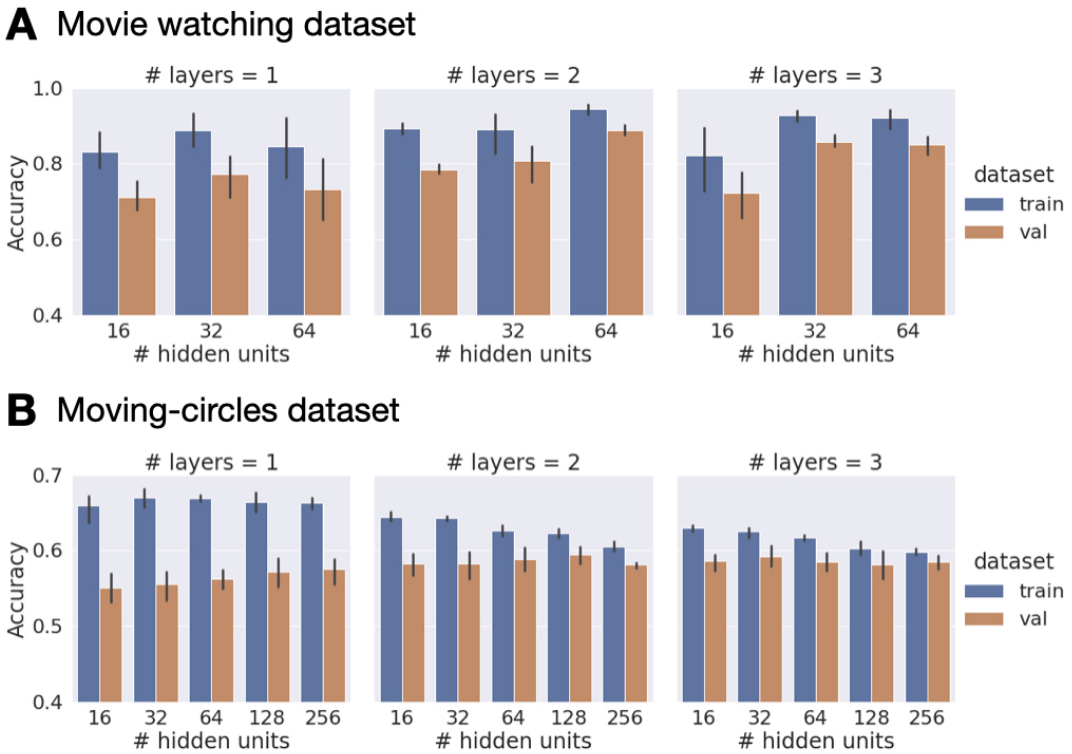


Figure 7. Selection of model architectures via cross validation on data from 100 participants. Val: validation portion of training data. Error bars show 95% confidence intervals across folds.

As accuracy was relatively high overall (>70%) and improvements were relatively modest as a function of these parameters, we chose the simplest case (one layer) but with 32 units which improved performance over 16 units (for evaluation performance, we considered all time points, even those during the first 60-90 seconds when accuracy increased sharply; see Figure 2). For the voxel-based analysis, we explored the $\{1, 2, 3\} \times \{16, 32, 64, 128, 256\}$ space. Classification of approach vs. retreat was a challenging problem, so we utilized 3 layers; we employed 32 units per layer for consistency with the clip classification architecture.

Dimensionality reduction

Several dimensionality reduction techniques exist that could be used to probe lower-dimensional representations in our system. To capture the temporal relationships in the time series data in the dimensionality reduction process, we combined the GRU network with a simple additional fully connected layer for dimensionality reduction (DR-FC). We refer to this model as the *GRU encoder* (Figure 1B). Formally, GRU outputs, $\boldsymbol{h}_t$, were linearly projected onto a lower-dimensional space containing units with linear activation:

$$\hat{\boldsymbol{h}}_t = (W_{\text{DRFC}}\boldsymbol{h}_t + b_{\text{DRFC}}), \tag{4}$$

where the weight matrix (W_{DRFC} had dimensions $N_h \times N_{\hat{h}}; N_{\hat{h}} < N_h$). We refer to the layer as the *Dimensionality Reduction Fully Connected (DR-FC)* layer. The output of this layer, $\hat{\boldsymbol{h}}_t$, is sensitive to past inputs, $\boldsymbol{x}_t$, thus effectively capturing temporal dependencies. An additional fully connected layer was used to predict labels, as in the previous subsection.

For dimensionality reduction, only the DRFC and final FC layers were trained (via standard back-progapation). In other words, the training weights of the GRU network for classification were frozen in place after that initial learning phase. For visualization, when the dimensionality of $\hat{\boldsymbol{h}}_t$ was reduced to three, we plotted "trajectories" in a state-space representation (Figure 3A).

GRU decoder

Can low-dimensional representations obtained for classification be used to reconstruct the original data? Another GRU-based module was used to reconstruct brain activation from the learned low-dimensional representations, $\hat{\boldsymbol{h}}_t$, which we refer to as the *GRU decoder* (Figure 1C). The GRU decoder was trained independently (that is, separately from the GRU encoder), by minimizing the mean squared error between the GRU decoder output, $\hat{\boldsymbol{x}}_t$, and the original time series input, $\boldsymbol{x}_t$, at each time step.

Saliency maps for spatiotemporal importance

To help evaluate the contribution of individual inputs (ROIs or voxels) to classification, we employed a *saliency* measure (***Simonyan et al.*** (***2014***)). Saliency assigns importance scores to each input feature based on their contribution to the model's predictions. Once the model is fully trained, saliency can be computed at every at time, t. Given input activation, $\boldsymbol{x}_t$, the GRU classifier generates an output vector, $\hat{\boldsymbol{y}}_t$. As the process is supervised, the true class label, $c \in \{1, \cdots, C\}$, is known. Thus, the cth element of the output vector, $\hat{y}_t^c$, is the class score for the input's true class label.

Although $\hat{y}_t^c$ is determined by a complex, non-linear transformation, it can be approximated using a first-order Taylor approximation:

$$\hat{y}_t^c \approx \boldsymbol{w}_t^{\top} \boldsymbol{x}_t + b_t, \tag{5}$$

where the operator $\top$ is the transpose, and the the weight vector, $\boldsymbol{w}_t$, is defined as the gradient of $\hat{y}_t^c$ with respect to the activation input:

$$\boldsymbol{w}_t \equiv \frac{\partial \hat{y}_t^c}{\partial \boldsymbol{x}_t}. \tag{6}$$

The value of the ith element, w_t^i, was the saliency value for the ith input (ROI or voxel) at time t. The saliency captures the contribution of the input activation at time t to the true-class score, $\hat{y}_t^c$. Thus, the higher the value, the stronger the effect on predicting the class in question. To consider saliency values across participants, we z-scored the gradient values, w_t^i, across inputs and time steps for each participant. Because positive or negative contributions were potentially equally important (i.e., increases or decreases in brain activation altered the class prediction), we considered the absolute values of the z-scores above. Finally, we averaged these values across participants to obtain group-level saliency estimates.

To evaluate the robustness of the saliency results, we compared them to those obtained from a null model. To generate a null distribution, using test data, class labels were randomly shuffled, and saliency values computed as defined above. The procedure was repeated 1000 times, yielding a distribution of null-saliency values at each time step (see Figure 4).

Baseline models

We compared the GRU classifier to two baseline models. Training was based on computing cross-entropy loss between predicted and true labels at each time step (similar to Equation 3). The standard backpropagation algorithm was employed (Adam optimizer), as implemented in TensorFlow.

Feed-forward network

The feed-forward (*FF classifier*) network consisted of three fully-connected (FC) layers. Brain inputs at time, $\boldsymbol{x}_t$, were fed into a fully-connected hidden layer, and then transformed into a vector of class scores, as with the GRU classifier. Formally, the hidden layer, $\boldsymbol{h}_t^{\text{FF}}$, and output layer, $\hat{\boldsymbol{y}}_t^{\text{FF}}$, activations were obtained as follows:

$$\begin{aligned} \boldsymbol{h}_t^{\text{FF}} &= \text{ReLU}(W_{\text{FC1}} \boldsymbol{x}_t + b_{\text{FC1}}), \\ \hat{\boldsymbol{y}}_t^{\text{FF}} &= \text{Softmax}(W_{\text{FC2}} \boldsymbol{h}_t^{\text{FF}} + b_{\text{FC2}}). \end{aligned} \tag{7}$$

where the ReLU and Softmax operations are standard rectified linear and generalized logistic functions, respectively. The vectors $\boldsymbol{h}_t^{\text{FF}}$ and $\hat{\boldsymbol{y}}_t^{\text{FF}}$ had dimensions $N_h^{\text{FF}} = 103$ and $N_{\hat{y}}^{\text{FF}} = 15$ (for 15-way classification) respectively. The number of hidden units was set to 103 in order to maintain the total number of parameters of the classifier (32563) comparable to that of the GRU classifier (32559).

Temporal Convolutional Network

We investigated whether static fixed-sized windows would be sufficient for modeling dynamics. A temporal convolutional network (*TCN classifier*) uses fixed-sized convolutional filters, where the output at a time depends upon a fixed temporal window, which is determined by the filter length l ((***Lea et al.*** (***2016***); ***Bai et al.*** (***2018***)). Activations after temporal convolution were then transformed to obtain class scores. Formally, brain inputs at each time, $\boldsymbol{x}_t$, were convolved temporally with k filters, each with length of l time steps. A single filter of size $N_x \times l$ convolved along the time dimension (input time series) and generated a scalar time series. Therefore, k such filters produced a k-dimensional convolutional layer activation at each time, $\boldsymbol{h}_t^{\text{TCN}}$, which was used to generate class scores:

$$\hat{\boldsymbol{y}}_t^{\text{TCN}} = \text{Softmax}(W_{\text{FC}}\boldsymbol{h}_t^{\text{TCN}} + b_{\text{FC}}). \tag{8}$$

The vectors $\boldsymbol{h}_t^{\text{TCN}}$ and $\hat{\boldsymbol{y}}_t^{\text{TCN}}$ were of dimensions $k = 25$ and $N_{\hat{y}}^{\text{TCN}} = 15$, respectively. The choice of $k = 25$ filters, each of width of $l = 5$ time steps, was chosen to keep the the total number of parameters (37915) similar to that of the GRU classifier (32559).

Predicting behavior

To predict participants' scores, we employed a GRU-based model, essentially a *GRU regressor*. Separate models were trained to predict fluid intelligence and verbal IQ (***Barch et al.*** (***2013***)). In each case, all movie clips were used for training (rather than training a separate model for each clip) to promote learning representations that were not idiosyncratic to a particular clip, and thus generalizable across clips.

We applied the GRU classifier approach (Figure 1). The GRU outputs, $\boldsymbol{h}_t$ (dimensionality: $N_h = 32$), were input to a fully-connected layer with a single unit having a linear activation function. The models learned to predict a single score, $\hat{y}_t$, at every time step t:

$$\hat{y}_t = W_{\text{FC}}\boldsymbol{h}_t + b_{\text{FC}}. \tag{9}$$

Thus, we were able to generate predicted scores as a function of time, which could be compared with the true score, allowing evaluation of how model predictions evolved. We computed the mean squared error (MSE) between true, y_t, and predicted, $\hat{y}_t$, scores at each time point:

$$\mathcal{L}_{\text{MSE}}\left(y_t, \hat{y}_t; \Theta\right) = \frac{1}{2}\left(y_t - \hat{y}_t\right)^2, \tag{10}$$

where Θ was the set of learnable parameters. The total loss was defined as the average MSE loss across all time points. The models were trained using backprogagation, using the Adam optmizer for gradient descent.

We compared our approach to connectome-based predictive modeling (CPM; ***Shen et al.*** (***2017***)), possibly the state-of-the-art in this regard. In this approach, a functional connectivity matrix is initially formed for each clip based on the Pearson correlation between every pair of brain region time series. Subsequently, the entries in the matrix (or edges) that correlate with the individual differences measure beyond a set threshold (here, 0.2) are retained and a linear model is fit to predict scores based on the functional connectivity values. Given that functional connectivity is employed, CPM predictions rely on activations during the entire clip.

We tested our model statistically at two levels. At each time t, once the model was trained (training set only), we evaluated chance performance based upon a null distribution obtained through permutation testing. For each iteration, we randomly shuffled the true scores of test data across participants and then calculated Spearman's rank correlation between predicted and shuffled scores (1000 iterations). The 95% confidence interval is shown by the gray zone in Figure 5. Because of multiple comparisons across time, we performed an additional test considering all time points simultaneously. The permutation-based test draws from ideas used in the MEG/EEG field to evaluate if two time series differ in time (***Maris and Oostenveld*** (***2007***)). Consider a sliding window of length l over the entire clip time series, and compute the value of the average correlation over

Table 1. Summary of models implemented in the paper. Hyperparameters and other parameters: LR: controlled the rate of update for gradient descent; Dr: dropout for the GRU layer (***Srivastava et al.*** (***2014***)); L2: L2 regularization coefficient for the GRU layer; Batch size: number of training samples passed through the model before updating model parameters; Epochs: number of passes through the training dataset during training.

Model (Dataset)	Architecture	LR	Dr	L2	Batch size	Epochs
GRU classifier (movie-watching)	GRU (32 units) FC (15 units)	6×10^{-3}	10^{-6}	10^{-3}	32	45
FF classifier (movie-watching)	FC (103 units) FC (15 units)	10^{-3}	NA	NA	32	45
TCN classifier (movie-watching)	TCN (25 filters) FC (15 units)	10^{-3}	NA	NA	32	50
GRU encoder (movie-watching)	GRU (32 units) FC ($N_{\hat{h}}$ units) FC (15 units)	6×10^{-3}	10^{-6}	10^{-3}	32	45
GRU regressor (movie-watching)	GRU (1 unit) FC (1 unit)	10^{-3}	NA	NA	30	20
GRU classifier (moving-circles)	GRU (3 layers, 32 units) FC (1 unit)	6×10^{-3}	10^{-6}	10^{-3}	32	28

the sliding window. This allows us to determine the window that has the highest average correlation. For example, in Figure 5, the window of length $l = 10$ extended from $t = 139, \ldots, 148$, for fluid intelligence. To determine the probability that such value would be observed by chance, we computed the same highest average correlation based on values for shuffled participants (1000 iterations), providing a null distribution (Figure 5, supplemental figures).

Summary of models used

Table 1 summarizes all model architectures and documents all hyperparameters used for training.

Acknowledgments

M.V. was supported by a fellowship by the Brain and Behavior Initiative, University of Maryland, College Park. L.P. is supported by the National Institute of Mental Health (R01 MH071589 and R01 MH112517). Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado G, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. . 2015; http://download.tensorflow.org/paper/whitepaper2015.pdf.

Allefeld C, Haynes JD. Multi-voxel Pattern Analysis. In: Toga AW, editor. *Brain Mapping* Waltham: Academic Press; 2015.p. 641–646. https://www.sciencedirect.com/science/article/pii/B9780123970251003456, doi: 10.1016/B978-0-12-397025-1.00345-6.

Bai S, Kolter JZ, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:180301271. 2018; .

Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, Glasser MF, Curtiss S, Dixit S, Feldt C, et al. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage. 2013; 80:169–189.

Beaty RE, Kenett YN, Christensen AP, Rosenberg MD, Benedek M, Chen Q, Fink A, Qiu J, Kwapil TR, Kane MJ, et al. Robust prediction of individual creative ability from brain functional connectivity. Proceedings of the National Academy of Sciences. 2018; 115(5):1087–1092.

Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks. 1994; 5(2):157–166.

Bullmore E, Fadili J, Maxim V, Şendur L, Whitcher B, Suckling J, Brammer M, Breakspear M. Wavelets and functional magnetic resonance imaging of the human brain. Neuroimage. 2004; 23:S234–S249.

Buonomano DV, Maass W. State-dependent computations: spatiotemporal processing in cortical networks. Nature Reviews Neuroscience. 2009; 10(2):113–125.

Byron MY, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In: *Advances in neural information processing systems*; 2009. p. 1881–1888.

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. arXiv:14061078 [cs, stat]. 2014 Sep; http://arxiv.org/abs/1406.1078, arXiv: 1406.1078.

Chu C, Mourão-Miranda J, Chiu YC, Kriegeskorte N, Tan G, Ashburner J. Utilizing temporal information in fMRI decoding: Classifier using kernel regression methods. NeuroImage. 2011 Sep; 58(2):560–571. https://www.sciencedirect.com/science/article/pii/S1053811911006999, doi: 10.1016/j.neuroimage.2011.06.053.

Cutting J, Brunick K, Candan Simsek A. Perceiving Event Dynamics and Parsing Hollywood Films. Journal of experimental psychology Human perception and performance. 2012 Mar; 38. doi: 10.1037/a0027737.

Dubois J, Galdi P, Han Y, Paul LK, Adolphs R. Resting-State Functional Brain Connectivity Best Predicts the Personality Dimension of Openness to Experience. Personality Neuroscience. 2018 Jul; 1:e6. https://www.cambridge.org/core/product/identifier/S2513988618000081/type/journal_article, doi: 10.1017/pen.2018.8.

Dubois J, Galdi P, Paul LK, Adolphs R. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philosophical Transactions of the Royal Society B: Biological Sciences. 2018; 373(1756):20170284.

Faillenot I, Heckemann RA, Frot M, Hammers A. Macroanatomy and 3D probabilistic atlas of the human insula. NeuroImage. 2017 Apr; 150:88–98. https://www.sciencedirect.com/science/article/pii/S1053811917300964, doi: 10.1016/j.neuroimage.2017.01.073.

Fan C, Wang J, Gang W, Li S. Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. Applied Energy. 2019 Feb; 236:700–710. https://www.sciencedirect.com/science/article/pii/S0306261918318221, doi: 10.1016/j.apenergy.2018.12.004.

Gao P, Trautmann E, Byron MY, Santhanam G, Ryu S, Shenoy K, Ganguli S. A theory of multineuronal dimensionality, dynamics and measurement. BioRxiv. 2017; p. 214262.

Gao S, Mishne G, Scheinost D. Non-linear manifold learning in fMRI uncovers a low-dimensional space of brain dynamics. bioRxiv. 2020 Nov; p. 2020.11.25.398693. https://www.biorxiv.org/content/10.1101/2020.11.25.398693v1, doi: 10.1101/2020.11.25.398693, publisher: Cold Spring Harbor Laboratory Section: New Results.

Gers FA, Schmidhuber JA, Cummins FA. Learning to Forget: Continual Prediction with LSTM. Neural Computation. 2000 Oct; 12(10):2451–2471.

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR, et al. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 2013; 80:105–124.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001; 293(5539):2425–2430.

Haynes JD, Rees G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nature Neuroscience. 2005 May; 8(5):686–691. https://www.nature.com/articles/nn1445, doi: 10.1038/nn1445, number: 5 Publisher: Nature Publishing Group.

Hsu WT, Rosenberg MD, Scheinost D, Constable RT, Chun MM. Resting-state functional connectivity predicts neuroticism and extraversion in novel individuals. Social Cognitive and Affective Neuroscience. 2018 Feb; 13(2):224–232. https://doi.org/10.1093/scan/nsy002, doi: 10.1093/scan/nsy002.

Hutchinson RA, Niculescu RS, Keller TA, Rustandi I, Mitchell TM. Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models. NeuroImage. 2009 May; 46(1):87–104. https://www.sciencedirect.com/science/article/pii/S1053811909000688, doi: 10.1016/j.neuroimage.2009.01.025.

Janoos F, Machiraju R, Singh S, Morocz I Spatio-temporal models of mental processes from fMRI. NeuroImage. 2011 Jul; 57(2):362–377. https://www.sciencedirect.com/science/article/pii/S1053811911003259, doi: 10.1016/j.neuroimage.2011.03.047.

Jiang R, Calhoun VD, Zuo N, Lin D, Li J, Fan L, Qi S, Sun H, Fu Z, Song M, Jiang T, Sui J. Connectome-based individualized prediction of temperament trait scores. NeuroImage. 2018 Dec; 183:366–374. https://www.sciencedirect.com/science/article/pii/S1053811918307365, doi: 10.1016/j.neuroimage.2018.08.038.

Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nature Neuroscience. 2005 May; 8(5):679–685. https://www.nature.com/articles/nn1444, doi: 10.1038/nn1444, number: 5 Publisher: Nature Publishing Group.

Kingma DP, Ba J, Adam: A Method for Stochastic Optimization; 2014. http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Kriegeskorte N, Mur M, Bandettini PA. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience. 2008; 2. https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008/full?utm_source=FWEB&utm_medium=NBLOG&utm_campaign=ECO_10YA_top-research, doi: 10.3389/neuro.06.004.2008, publisher: Frontiers.

Lea C, Vidal R, Reiter A, Hager GD. Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In: Hua G, Jégou H, editors. *Computer Vision – ECCV 2016 Workshops* Lecture Notes in Computer Science, Cham: Springer International Publishing; 2016. p. 47–54. doi: 10.1007/978-3-319-49409-8_7.

Limbachia C, Morrow K, Khibovska A, Meyer C, Padmala S, Pessoa L. Controllability over stressor decreases responses in key threat-related brain areas. Communications Biology. 2021 Jan; 4(1):1–11. https://www.nature.com/articles/s42003-020-01537-5, doi: 10.1038/s42003-020-01537-5, number: 1 Publisher: Nature Publishing Group.

Loula J, Baroni M, Lake BM. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. arXiv:180707545 [cs]. 2018 Jul; http://arxiv.org/abs/1807.07545, arXiv: 1807.07545.

Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods. 2007 Aug; 164(1):177–190. https://www.sciencedirect.com/science/article/pii/S0165027007001707, doi: 10.1016/j.jneumeth.2007.03.024.

Meyer C, Padmala S, Pessoa L. Dynamic Threat Processing. Journal of Cognitive Neuroscience. 2019 Apr; 31(4):522–542. https://doi.org/10.1162/jocn_a_01363, doi: 10.1162/jocn_a_01363.

Mourão-Miranda J, Friston KJ, Brammer M. Dynamic discrimination analysis: A spatial–temporal SVM. NeuroImage. 2007 May; 36(1):88–99. https://www.sciencedirect.com/science/article/pii/S1053811907001073, doi: 10.1016/j.neuroimage.2007.02.020.

Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. NeuroImage. 2011 May; 56(2):400–410. https://www.sciencedirect.com/science/article/pii/S1053811910010657, doi: 10.1016/j.neuroimage.2010.07.073.

Nestor A, Plaut DC, Behrmann M. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proceedings of the National Academy of Sciences. 2011 Jun; 108(24):9998–10003. http://www.pnas.org/cgi/doi/10.1073/pnas.1102433108, doi: 10.1073/pnas.1102433108.

Ojala M, Garriga GC. Permutation Tests for Studying Classifier Performance. In: *2009 Ninth IEEE International Conference on Data Mining* Miami Beach, FL, USA: IEEE; 2009. p. 908–913. http://ieeexplore.ieee.org/document/5360332/, doi: 10.1109/ICDM.2009.108.

Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, Van De Ville D. Decoding brain states from fMRI connectivity graphs. NeuroImage. 2011 May; 56(2):616–626. https://www.sciencedirect.com/science/article/pii/S1053811910008335, doi: 10.1016/j.neuroimage.2010.05.081.

Rubin TN, Koyejo O, Gorgolewski KJ, Jones MN, Poldrack RA, Yarkoni T. Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition. PLOS Computational Biology. 2017 Oct; 13(10):e1005649. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005649, doi: 10.1371/journal.pcbi.1005649, publisher: Public Library of Science.

Russo AA, Bittner SR, Perkins SM, Seely JS, London BM, Lara AH, Miri A, Marshall NJ, Kohn A, Jessell TM, Abbott LF, Cunningham JP, Churchland MM. Motor Cortex Embeds Muscle-like Commands in an Untangled Population Response. Neuron. 2018 Feb; 97(4):953–966.e8. https://www.sciencedirect.com/science/article/pii/S0896627318300072, doi: 10.1016/j.neuron.2018.01.004.

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex. 2018 Sep; 28(9):3095–3114. https://doi.org/10.1093/cercor/bhx179, doi: 10.1093/cercor/bhx179.

Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, Constable RT. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nature Protocols. 2017 Mar; 12(3):506–518. https://www.nature.com/articles/nprot.2016.178, doi: 10.1038/nprot.2016.178, number: 3 Publisher: Nature Publishing Group.

Shine JM, Hearne LJ, Breakspear M, Hwang K, Müller EJ, Sporns O, Poldrack RA, Mattingley JB, Cocchi L. The Low-Dimensional Neural Architecture of Cognitive Complexity Is Related to Activity in Medial Thalamic Nuclei. Neuron. 2019 Dec; 104(5):849–855.e3. https://www.sciencedirect.com/science/article/pii/S0896627319307755, doi: 10.1016/j.neuron.2019.09.002.

Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:13126034 [cs]. 2014 Apr; http://arxiv.org/abs/1312.6034, arXiv: 1312.6034.

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. The journal of machine learning research. 2014; 15(1):30.

T Vu A, Jamison K, Glasser MF, Smith SM, Coalson T, Moeller S, Auerbach EJ, Uğurbil K, Yacoub E. Tradeoffs in pushing the spatial resolution of fMRI for the 7T Human Connectome Project. NeuroImage. 2017 Jul; 154:23–32. https://www.sciencedirect.com/science/article/pii/S1053811916306681, doi: 10.1016/j.neuroimage.2016.11.049.

Thomas AW, Heekeren HR, Müller KR, Samek W. Analyzing Neuroimaging Data Through Recurrent Deep Learning Models. Frontiers in Neuroscience. 2019 Dec; 13:1321. https://www.frontiersin.org/article/10.3389/fnins.2019.01321/full, doi: 10.3389/fnins.2019.01321.

Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: An overview. NeuroImage. 2013 Oct; 80:62–79. https://www.sciencedirect.com/science/article/pii/S1053811913005351, doi: 10.1016/j.neuroimage.2013.05.041.

Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage. 2017 Jan; 145:166–179. https://www.sciencedirect.com/science/article/pii/S105381191630595X, doi: 10.1016/j.neuroimage.2016.10.038.

Venkatesh M, Jaja J, Pessoa L. Brain dynamics and temporal trajectories during task and naturalistic processing. NeuroImage. 2019 Feb; 186:410–423. https://www.sciencedirect.com/science/article/pii/S105381191832086X, doi: 10.1016/j.neuroimage.2018.11.016.

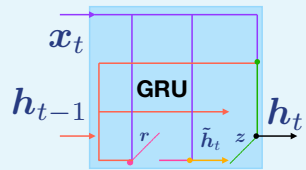
Vyas S, Golub MD, Sussillo D, Shenoy KV. Computation Through Neural Population Dynamics. Annual Review of Neuroscience. 2020; 43(1):249–275. https://doi.org/10.1146/annurev-neuro-092619-094115, doi: 10.1146/annurev-neuro-092619-094115, _eprint: https://doi.org/10.1146/annurev-neuro-092619-094115.

Werbos PJ. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE. 1990 Oct; 78(10):1550–1560. doi: 10.1109/5.58337, conference Name: Proceedings of the IEEE.

Appendix 1

Gated Recurrent Units

We employed recurrent neural networks with Gated Recurrent Units (GRUs) which overcome challenges in learning temporal information 1) by adaptively updating temporal history, and 2) by resetting/dropping temporal history that is irrelevant for future predictions (***Cho et al.*** (***2014***); ***Gers et al.*** (***2000***)).



Appendix 1 Figure 1. Formal specification of a GRU unit. The input, $\boldsymbol{x}_t$, and previous GRU output, $\boldsymbol{h}_{t-1}$, together generate a reset gate vector, $\boldsymbol{r}$, and an update gate vector, $\boldsymbol{z}$.

Given the input, $\boldsymbol{x}_t$, and previous output vector, $\boldsymbol{h}_{t-1}$, GRU units compute a reset gate vector, $\boldsymbol{r}$, and an update gate vector, $\boldsymbol{z}$:

$$\begin{aligned} \boldsymbol{r} &= \sigma\left(W_r \boldsymbol{x}_t + U_r \boldsymbol{h}_{t-1} + b_r\right), \\ \boldsymbol{z} &= \sigma\left(W_z \boldsymbol{x}_t + U_z \boldsymbol{h}_{t-1} + b_z\right), \end{aligned} \tag{11}$$

where W_r, U_r, W_z, U_z are weight matrices and b_r, b_z are bias vectors that are defined during the learning phase. The symbols $\sigma(\cdot)$, $\tanh(\cdot)$, and $\odot$ represent the sigmoid activation function, the hyperbolic tangent function, and the Hadamard product, respectively.

The two gate vectors are then used to compute a candidate state vector:

$$\tilde{h}_t = \tanh\left(W_h \boldsymbol{x}_t + U_h\left(\boldsymbol{r} \odot \boldsymbol{h}_{t-1}\right) + b_h\right), \tag{12}$$

using weight matrices W_h, U_h and bias vector b_h, which are defined during the learning phase ($N_{\tilde{h}} = 32$ dimensions). The reset vector, $\boldsymbol{r}$, controls the amount of information from the past output, $\boldsymbol{h}_{t-1}$, used to compute the candidate activation. Thus, when $\boldsymbol{r}$ was close to zero, the prior memory, $\boldsymbol{h}_{t-1}$, was de-emphasized and $\tilde{\boldsymbol{h}}_t$ was "reset" with information from the current (brain signal) input, $\boldsymbol{x}_t$. Finally, the update gate vector, $\boldsymbol{z}$, governs the fraction of the previous state activation to be carried forward:

$$\boldsymbol{h}_t = (\boldsymbol{1} - \boldsymbol{z}) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z} \odot \tilde{\boldsymbol{h}}_t. \tag{13}$$

Each GRU unit has its own reset and update gate units, thus allowing each unit to learn dependencies over multiple time scales. Units that learn long-term dependencies have their update gates frequently active, while those that learn shorter-term dependencies have their reset gates frequently active.

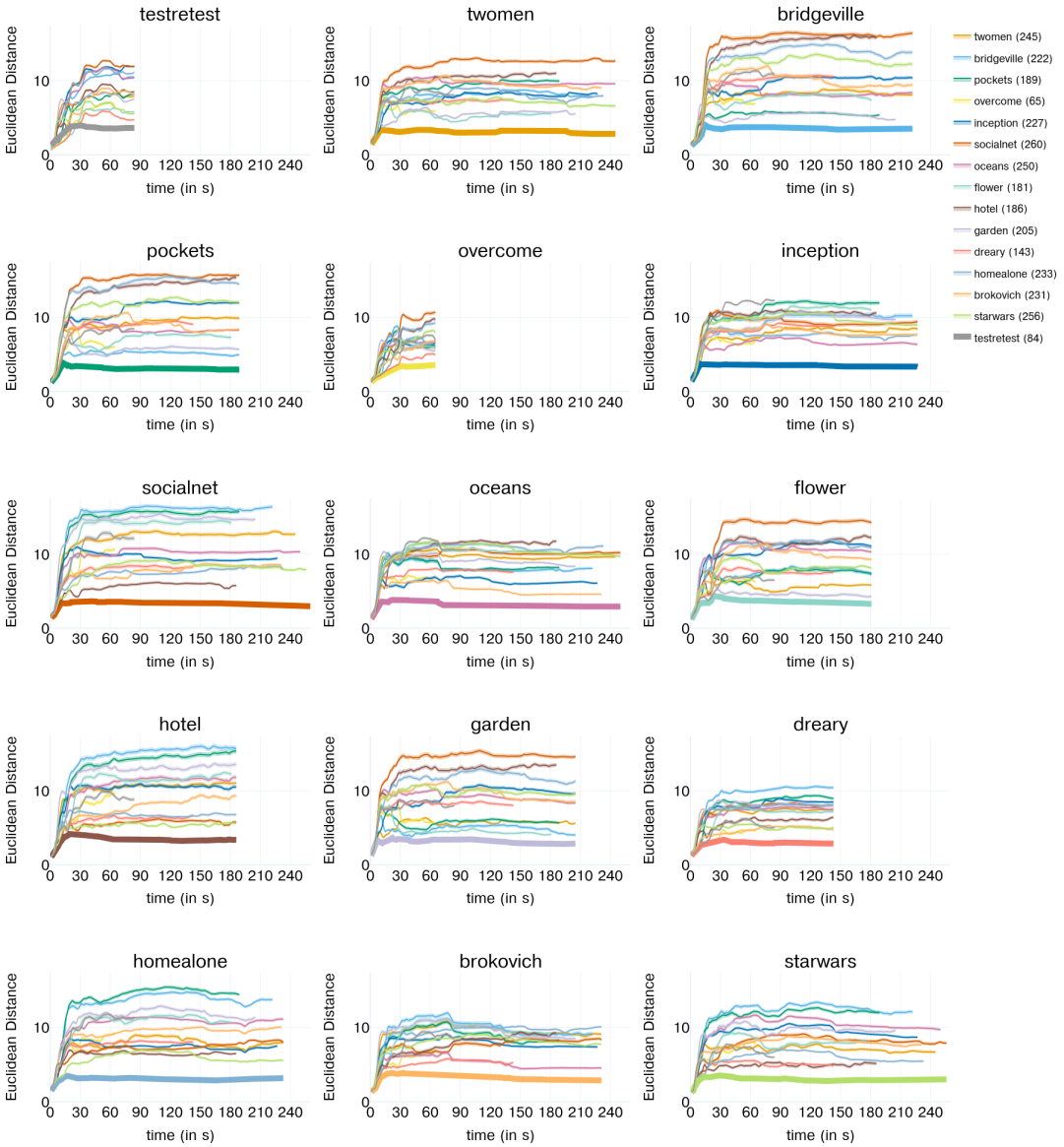


Figure 3–Figure supplement 1. Euclidean distances between trajectories. In the inset, the duration of every clip is indicated in parenthesis.

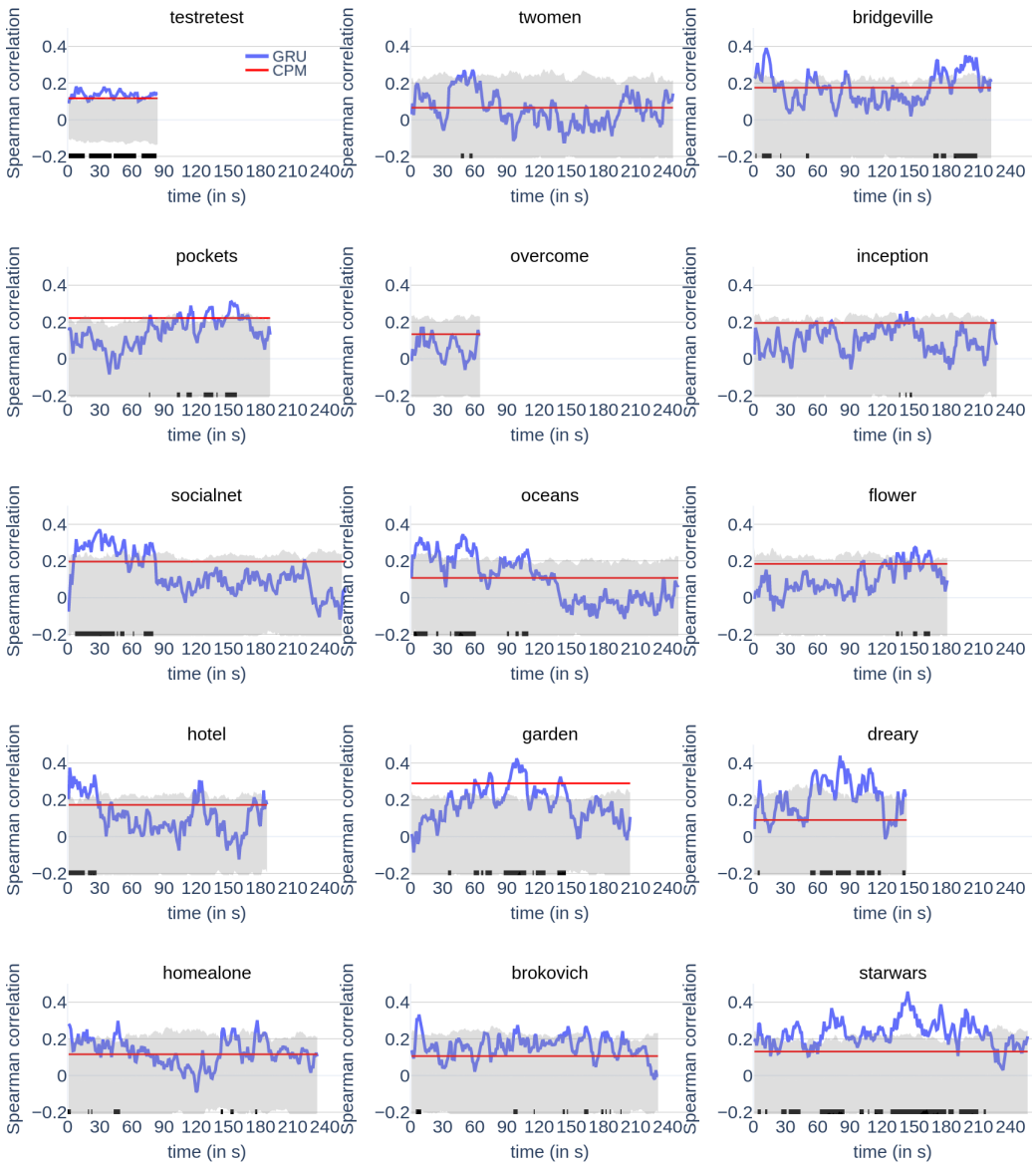


Figure 5–Figure supplement 1. Fluid Intelligence predictions for all movie clips. Conventions as in the figure in the main text.

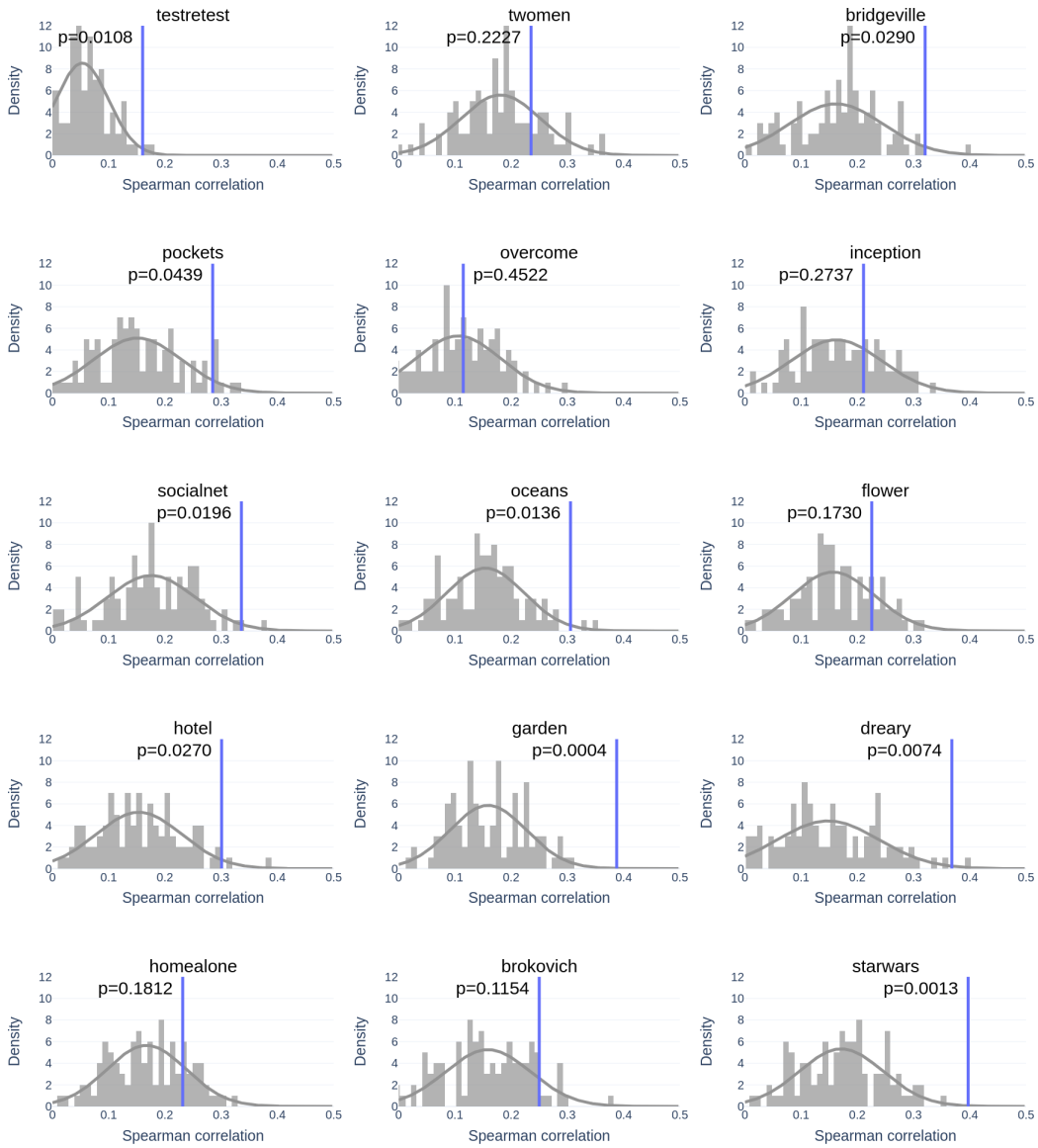


Figure 5–Figure supplement 2. Null distributions of fluid intelligence predictions. Vertical blue lines indicate the prediction based on actual data.

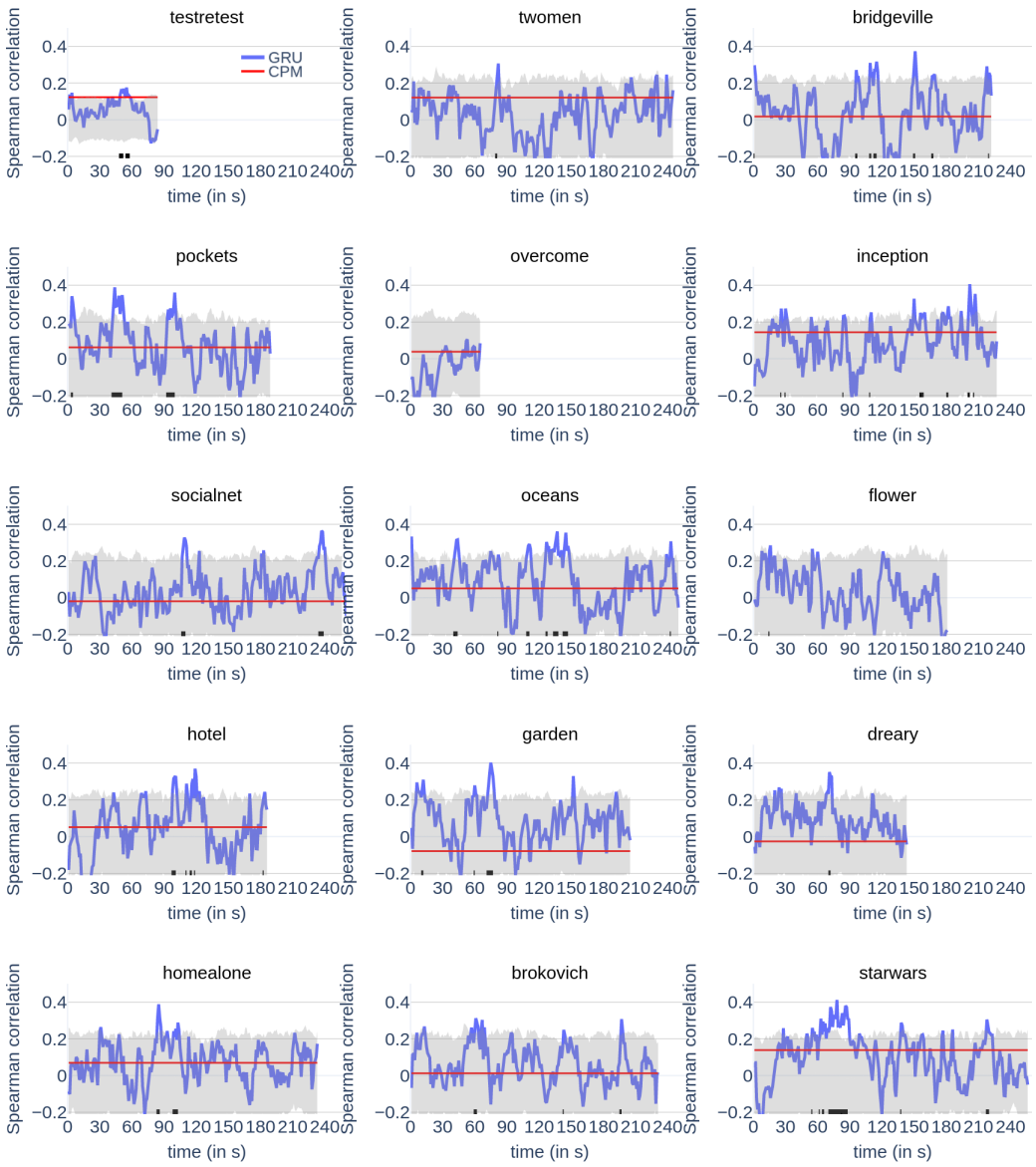


Figure 5–Figure supplement 3. Verbal IQ predictions for all movie clips. Conventions as in the figure in the main text.

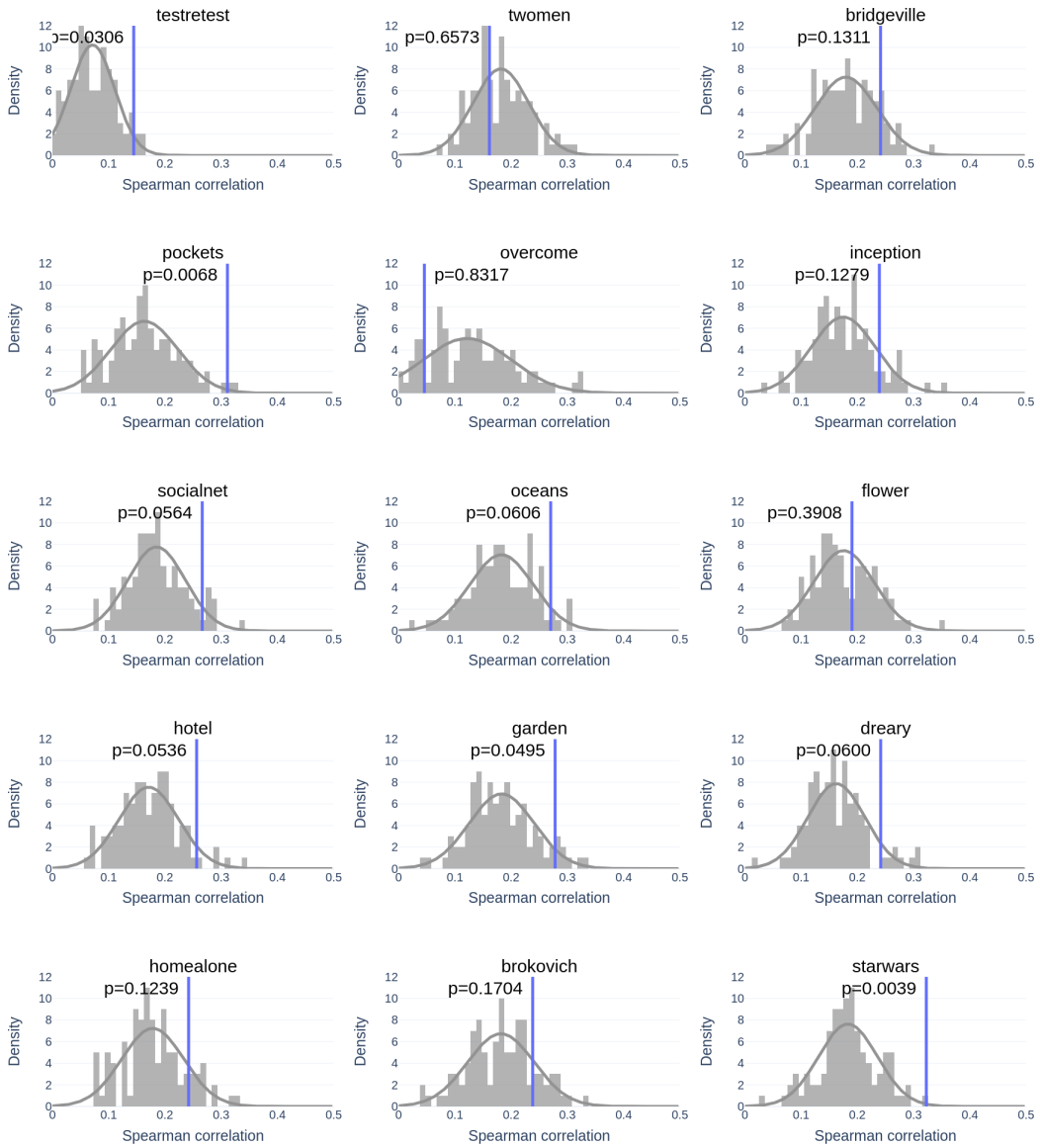


Figure 5–Figure supplement 4. Verbal IQ predictions: null distributions.Vertical blue lines indicate the prediction based on actual data.