communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-026-10411-9

A view-engage-predict framework for enhancing brain-behavior mapping with naturalistic movie-watching fMRI

Yunrui Zhang [1], Emily S. Finn [2], Mert R. Sabuncu[1,3,4] & Amy Kuceyeski [3,5]

Most brain-behavior mapping studies rely on resting-state functional connectivity (FC), but this approach has known accuracy limits and can be outperformed by movie-watching FC. Here, we present a novel deep neural network framework to predict cognitive scores and sex from FC during naturalistic movie viewing, and examine how movie content and its ability to synchronize brain activity across individuals relate to prediction performance. We show that FC from movie-watching generally outperforms resting-state FC - even when compared to five times more temporal data - with sensory and higher-order brain networks emerging as the most important for prediction. Using both static and sliding-window dynamic FC approaches, we find that higher cognitive prediction accuracy is positively associated with greater inter-subject synchrony and the duration of human faces and voices in the movies; these effects were not found for sex prediction. This work underscores the promise of naturalistic movie viewing as a powerful tool for probing individual differences in the brain and revealing neural underpinnings of human behavior.

Understanding brain regions and circuits that underlie individual differences in cognition and behavior is a fundamental goal in neuroscience. Functional MRI, which allows non-invasive recording of brain activity, has revealed that functional connectivity (FC) patterns are uniquely identifiable across individuals, i.e., FC fingerprinting, and are associated with various behavioral traits[1–5]. Traditionally, FC fingerprinting and brain-behavior mapping have relied on resting-state fMRI, but naturalistic paradigms such as movie watching offer a more ecologically valid and engaging framework that at times results in more accurate brain-behavior mappings[6–10]. Watching a movie induces rich and dynamic neural responses, characterized by synchronized activity across individuals and structured temporal fluctuations in FC[11–14]. The use of movie stimuli provides a controlled yet immersive setting, capturing brain activity fluctuations that more closely resemble real-world cognitive processes compared to static resting-state paradigms, in which a person is told not to think about anything in particular.

Recent studies have used FC to predict behavior, i.e., cognitive scores, motor outcomes or diagnosis categories, or demographics like age or sex, but existing models and data remain limited in their predictive power. Most prior brain-behavior mapping studies have employed linear models like linear regression[1,15]; however, these methods may be limited, as the mapping between FC and behavior may involve nonlinear components. The application of deep learning to FC-based behavior prediction has been an area of recent exploration, and may have the potential to uncover complex mappings that traditional models overlook[16–21]. In this study, we introduce a novel deep neural network (DNN) framework for predicting cognition and sex from FC patterns derived from movie-watching fMRI data, leveraging machine learning techniques to enhance predictive performance and interpretability. One commonly cited drawback of DNNs is that they suffer from a lack of explainability, which is an important aspect of brain-behavior mapping. Being able to reliably identify the functional connections, regions or networks that are important for predicting a given task is of utmost importance for the field of network neuroscience[4,22–25]. Knowing the neural underpinnings of behavior or diagnoses would enable a better understanding of how we may intervene to enhance performance or reduce symptoms.

Another largely unexplored aspect of FC-based behavior prediction using movie-watching data is its relationship with neural synchrony, or the level of temporal alignment across individuals during shared experiences[26–30]. Increased neural synchrony has been associated with higher levels of attentional engagement and cognitive processing[31,32]. However, the connection between neural synchrony and brain-behavior prediction accuracy remains underexplored.

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. [2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [3]Department of Radiology, Weill Cornell Medicine, New York, NY, USA. [4]Cornell Tech, New York, NY, USA. [5]Department of Computational Biology, Cornell University, Ithaca, NY, USA. e-mail: amk2012@med.cornell.edu

Finally, prior research has largely focused on static FC, overlooking the temporal dynamics of FC during movie-watching. While static FC provides an aggregate measure of connectivity over an entire scan or movie clip, it cannot capture the moment-to-moment fluctuations that occur in response to evolving stimuli[33,34]. Dynamic functional connectivity (dFC) can reveal transient connectivity states that may be more informative for predicting behavior[35–40], and in the context of brain-behavior mapping, allow finer-scale analysis of the movie features that may be driving prediction accuracy.

In this study, we introduce a novel deep learning-based framework for predicting cognition and sex from FC during movie-watching and at rest. We compare the predictive performance of various movie clips with differing content, systematically analyzing the contributions of individual brain connections and regions to the predictions. Additionally, we explore the relationship between neural synchrony and brain-behavior prediction accuracy to test our hypothesis that higher neural synchrony generally correlates with better prediction performance. At a finer temporal scale, we segment the movie clips into sliding windows to capture the dynamic fluctuations in neural synchrony and prediction accuracy, further testing for correlations between synchrony and prediction performance. Finally, we investigate the visual, auditory, and semantic features of the movies, identifying key factors that correlate with predictive performance at both the level of the entire movie clip as well as the sliding windows within a clip. By integrating these approaches, our work provides novel insights into the relationship between functional brain connectivity, behavior, neural synchrony, and the features of naturalistic stimuli, offering a more comprehensive understanding of the brain and stimulus features that drive brain-behavior mapping using naturalistic paradigms.

Results

Brain-behavior mapping accuracy varies across movie clips

We applied a DNN to the static FC extracted from different movie clips (and resting-state) to predict individual traits, including cognitive scores and sex, in a cohort of 176 participants (106 female, 70 male, aged 22–36) from the 7T substudy in the Human Connectome Project's Young Adult dataset[41], see Fig. 1. Cross-validation was performed with 10-folds, ensuring family members did not span folds; this cross-validation was repeated 100 times to assess robustness. The prediction accuracy for cognitive scores was quantified using the Spearman correlation coefficient between true and predicted scores across participants in the test set, then averaged across the 100 iterations. To assess statistical significance, we used a cross-validated permutation test in which the training labels were randomly permuted within each CV fold and the DNN was fully refit for each permutation. For each of the 100 CV repetitions, 1000 such label-shuffled permutations were performed. The resulting permutation accuracies were pooled across all CV repetitions to construct a single empirical null distribution reflecting the entire model-fitting process. The p value was then computed as the proportion of pooled permutations with prediction accuracy greater than or equal to the observed accuracy.

For sex classification, prediction performance was evaluated using the area under the receiver operating characteristic curve (AUC) for the test set, also averaged over 100 iterations.

For the comparison between movie-watching and resting-state prediction performance, we constructed duration- and onset-matched resting-state segments ("rest blocks") corresponding to the 13 analyzed movie clips (see Methods). For visual clarity in Fig. 2, the results from these matched resting-state blocks are summarized within the cropped resting-state column rather than displayed as 13 separate bars. Detailed prediction results for the individual resting-state blocks are provided in Supplementary Fig. 15.

Our results show wide variability in the predictive performance across the different movie clips. Several clips numerically outperformed resting-state with the same length, whereas others were comparable to or worse than rest. Using the full 15 minutes of resting-state FC improved performance relative to cropped rest and exceeded some clips; however, a subset of highly predictive clips achieved comparable or higher accuracy despite being five times shorter (Fig. 2a, b).

For total cognition, the highest prediction accuracies were observed for models based on movie-watching FC, including *Social Network* (Spearman $r = 0.39 \pm 0.02$, permutation-based $p < 0.001$), *Ocean's 11* ($r = 0.38 \pm 0.02$, $p < 0.001$), and *Inception* ($r = 0.34 \pm 0.02$, $p < 0.001$). In contrast, the lowest predictive performance was associated with clips from the third movie session, including *Flower* ($r = 0.17 \pm 0.04$, $p = 0.031$), *Garden* ($r = 0.12 \pm 0.04$, $p = 0.088$), and *Dreary* ($r = 0.09 \pm 0.05$, $p = 0.147$). The cropped 164s resting-state blocks derived from *REST1* and *REST4* also demonstrate predictive power, with *REST1* ($r = 0.16 \pm 0.04$, $p = 0.022$)

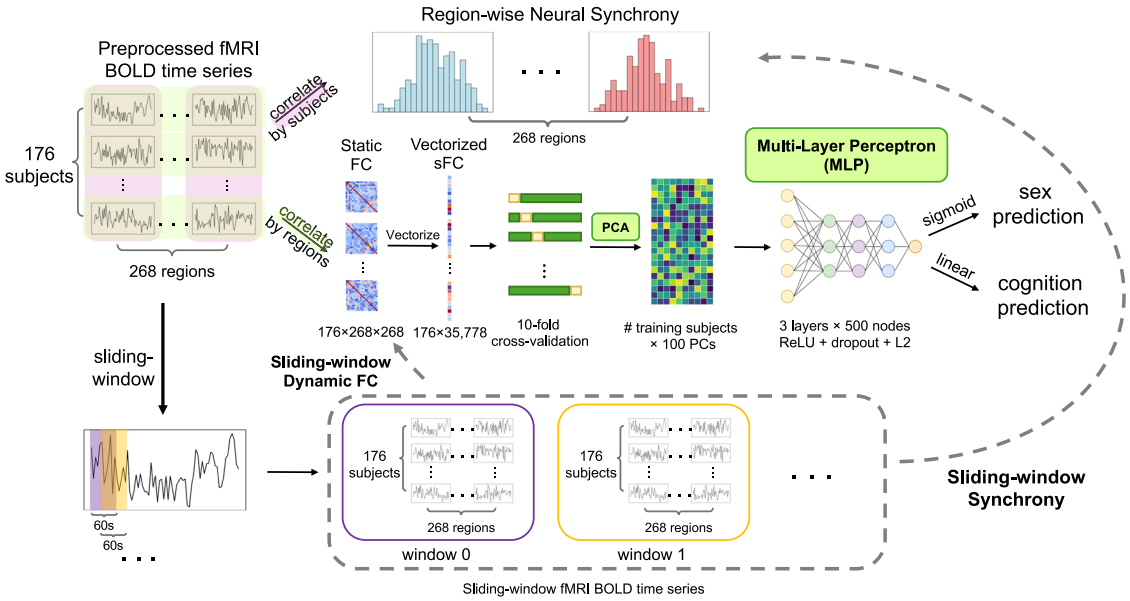

Fig. 1 | Workflow of the study. Static functional connectivity (sFC) is computed by correlating fMRI time series across pairs of brain regions, followed by vectorization and PCA to reduce dimensionality. Neural synchrony is defined by across-subject correlations of the same brain regions' fMRI time series and assessed by the mean/std of distribution across pairs of subjects. In parallel, a sliding-window approach is applied to compute dynamic functional connectivity (dFC) and sliding-window neural synchrony, capturing temporal fluctuations in connectivity patterns. The extracted features from both static and dynamic FC are fed into a 3-layer multi-layer perceptron (MLP) model with 500 nodes per layer for classification (sex prediction) and regression (cognition prediction).

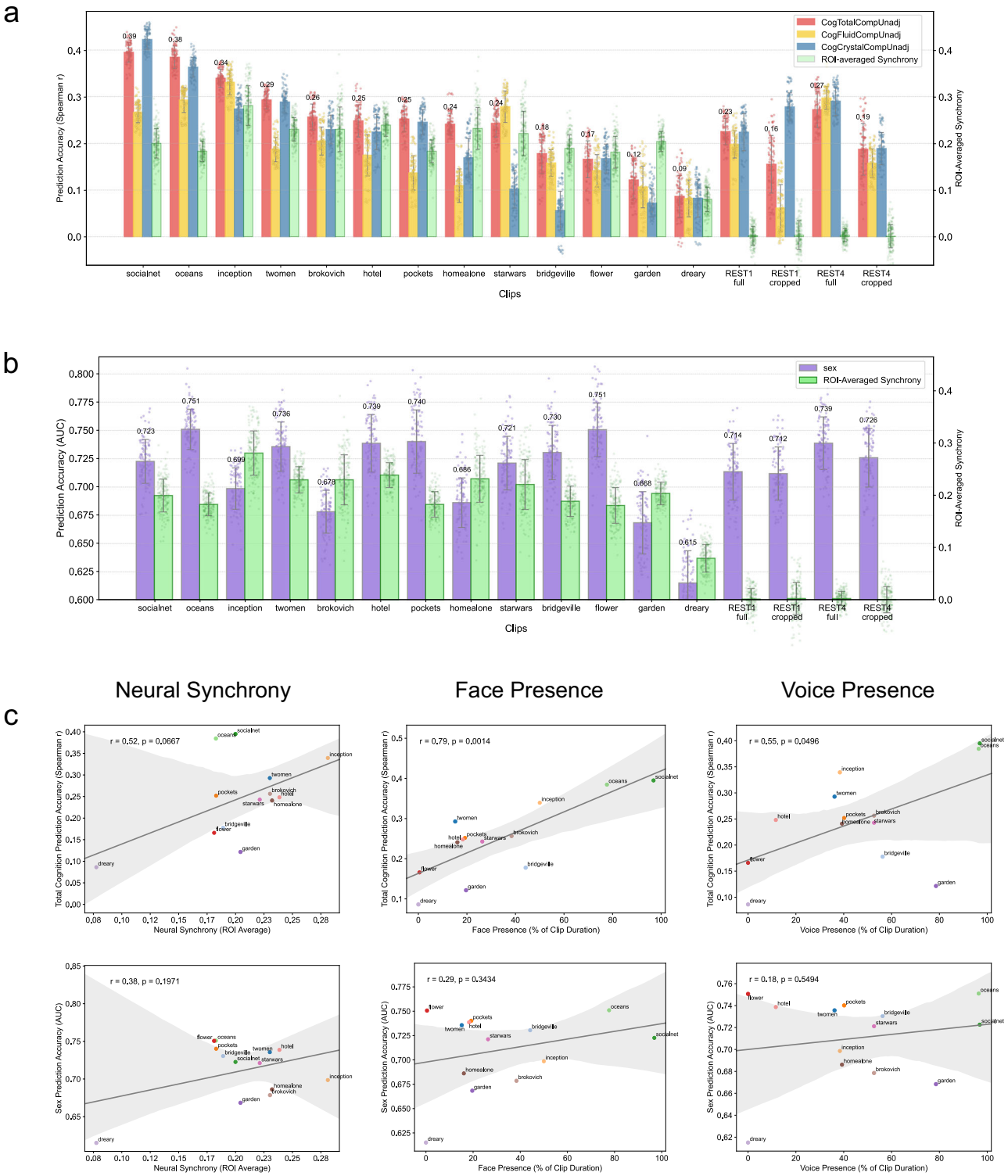

Fig. 2 | Brain-behavior mapping model performance. a Prediction accuracy for total, fluid, and crystallized cognition across all movie clips and resting-state sessions (full and cropped), measured by Spearman's rank correlation (*r*). Bars show mean accuracy across 100 cross-validation iterations, with gray error bars indicating standard deviation and points showing individual iterations. All sessions were cropped to 164s to ensure equal input length, except the full resting-state scan (15 min). Regional average neural synchrony (green) is computed across 55 regions showing significant synchrony in at least one clip, averaged over all subject pairs; points indicate per-subject mean synchrony (averaged across pairs with other subjects), with error bars showing standard deviation. **b** Prediction accuracy for sex classification across all movie clips and resting-state sessions, measured by AUC. Bars, points, and error bars are defined as in (**a**); regional average neural synchrony is shown in green. **c** Across movie clips, neural synchrony is positively associated with total cognition prediction accuracy ($r = 0.52$, $p = 0.0667$) and, to a lesser extent, related to sex classification accuracy ($r = 0.38$, $p = 0.1971$). Total cognition prediction accuracy is significantly correlated with the percentage of clip duration containing human faces ($r = 0.79$, $p = 0.0014$) and voices ($r = 0.55$, $p = 0.0496$). Sex prediction accuracy also shows positive, though weak and non-significant, correlations with face presence ($r = 0.29$, $p = 0.3434$) and voice ($r = 0.18$, $p = 0.5494$) presence.

corresponding to the same day as the first and second movie sessions, and *REST4* ($r = 0.19 \pm 0.03$, $p = 0.018$) aligning with the third and fourth sessions (Fig. 2a). Note that all movie clips and rest sessions were cropped to 164 seconds from their onset to ensure comparability across conditions. The 15-min full rest sessions show better prediction accuracy than the cropped rest sessions (full *REST1*: $r = 0.23 \pm 0.02$, $p = 0.004$; full *REST4*: $r = 0.27 \pm 0.03$, $p < 0.001$), but this result is numerically outperformed by the movie clips from *Social Network*, *Inception*, *Ocean's 11* and *Two Men*, which are five times shorter than the full rest sessions. The relative ranking of prediction accuracy for other cognitive measures, including crystallized cognition and fluid cognition, followed a similar trend to that of total cognition (Fig. 2a). However, there are slight variations in the prediction accuracy ranking for crystallized and fluid cognition; for example, while *Social Network* yielded the highest predictive accuracy for total cognition ($r = 0.39 \pm 0.02$, $p < 0.001$) and crystallized cognition ($r = 0.42 \pm 0.02$, $p < 0.001$), *Inception* exhibited the highest predictive accuracy for fluid cognition ($r = 0.33 \pm 0.03$, $p < 0.001$). Results for all 12 individual NIH toolbox cognitive measures and chronological age are presented in Supplementary Fig. 1; in the main text, we do not present the individual cognitive scores as they are noisier than the composites, and we do not present age, as the spread of age values is quite small in this population, and thus the accuracy is quite low.

For sex classification, the highest predictive accuracies were observed for *Ocean's 11* ($AUC = 0.75 \pm 0.02$), *Flower* (0.75 ± 0.02), and *Pockets* (0.74 ± 0.03), whereas *Erin Brockovich* (0.68 ± 0.02), *Garden* (0.67 ± 0.03), and *Dreary* (0.62 ± 0.03) demonstrated the lowest predictive power (Fig. 2b). The cropped 164s resting-state blocks derived from *REST1* (0.71 ± 0.03) and *REST4* (0.73 ± 0.02) demonstrated predictive accuracies that were either lower or higher than those of specific movie clips, as observed in cognition prediction. The 15-min full rest sessions show slightly better prediction accuracy (full *REST1*: $AUC = 0.71 \pm 0.02$, full *REST4*: 0.74 ± 0.02), which is numerically outperformed by shorter movie clips from *Ocean's 11*, *Flower*, *Pockets* and *Hotel*. Notably, while some movie clips consistently numerically outperformed the rest of the sessions in both cognition and sex prediction, the ranking of prediction accuracies for sex classification differed from that of total cognition.

Key brain connections and networks for multiple cognitive measures and sex prediction partially overlap in movie-watching sessions

We evaluated the contribution of specific brain connections and functional networks to cognition and sex prediction using the Integrated Gradients (IG) method. The importance of individual connections was quantified by calculating the absolute IG values derived from the DNN model, averaged across all 176 subjects. We identified the connections with the top 1% average importance scores and mapped to their corresponding Yeo 7 functional networks[42], and visualized using chord diagrams (Fig. 3 and, alternatively, glass brain visualizations in Supplementary Fig. 3). During movie-watching sessions, the most predictive connections involved interactions between sensory networks—including the visual (VIS) and auditory (within the somatomotor network, SMN)—and high-order networks such as the default mode network (DMN), frontoparietal network (FPN), dorsal attention network (DAN), and salience network (SAN). During the two rest sessions, DMN, FPN, and SAN are the most important for behavior prediction.

Although the most predictive connections across movie clips often originate from the same functional networks, the clip-specific top connections vary considerably. For example, in *Inception*, the most predictive connections were primarily within higher-order networks (DAN and DMN), whereas in *Ocean's 11*, they were more concentrated within the lower-order visual network. Notably, during both movie-watching and rest sessions, many connections important for cognition prediction also played a significant role in sex classification.

To directly quantify the relative contributions of clip identity and behavioral outcome to feature-map similarity (Figs. 3b and 4b and Supplementary Figs. SI4 and SI7), we modeled pairwise Fisher-z transformed correlations using a mixed-effects regression framework (Methods 4.3.2). Similarity was strongly explained by clip identity (edge-wise: $\beta = 0.53$, $p < 10^{-6}$; region-wise: $\beta = 0.74$, $p < 10^{-6}$), whereas outcome identity showed no detectable effect once clip and stimulus covariates were accounted for ($p > 0.6$ in both analyses). Thus, the importance map structure was primarily organized by stimulus context rather than behavioral target.

To test whether outcome-specific differences generalize across stimuli, we computed within-clip difference maps ($\Delta IG = IG_{\text{cognition}} - IG_{\text{sex}}$). Cross-clip correlations of ΔIG maps were uniformly near zero (Supplementary Fig. SI9), and the averaged consensus map had a small overall magnitude that did not exceed permutation-based null expectations ($p \approx 0.11$), indicating no stable clip-invariant outcome-specific pattern.

Finally, cross-clip generalization analyses (Supplementary Fig. SI10) showed that models trained on one clip predicted behavior above chance in other clips, but performance varied substantially across train-test pairs. Generalization was more robust for sex classification than for cognition prediction and tended to mirror within-clip performance levels.

Together, these analyses indicate that while predictive information transfers across stimuli, the specific FC features emphasized by the model are primarily structured by stimulus context. The similarity between feature importance maps was consistently stronger within the same clip for different outcomes than across different clips, suggesting that the relative importance of FC features is strongly shaped by stimulus context[43]. This pattern is also observed with the correlation of feature importance maps of all individual NIH Toolbox cognitive measures (Supplementary Fig. SI4).

Higher neural synchrony is associated with better FC-based cognition prediction accuracy under naturalistic viewing conditions

We observe that different movie clips elicit varying degrees of regional synchrony in the fMRI BOLD responses across subjects. Intuitively, movies that induce higher neural synchrony across the brain may be more engaging for most participants, as higher engagement likely means more time-locked, stimulus-driven brain activity. In addition, as neural activity increases in stimulus-driven regions during such movies, the signal-to-noise ratio (SNR) of the BOLD response and FC likely improves, potentially enhancing behavior prediction. However, very high levels of neural synchronization could mean reduced inter-individual variability, which may limit prediction capabilities across the population. Importantly, this intuition is specific to naturalistic viewing, where participants share a common time-locked stimulus; it is not expected to hold in resting-state scans, which contain minimal shared time-locked activity across individuals. This raises the question: how does neural synchrony impact brain-behavior prediction during naturalistic viewing?[44]

To investigate this, we first examined whether stimulus-evoked neural synchrony across all movie clips was correlated with prediction accuracy (Fig. 2c). For each brain region, neural synchrony was quantified using inter-subject correlation (ISC) of preprocessed fMRI BOLD signals across the duration of a movie clip or rest session, averaged over all subject pairs. The overall synchrony for a movie clip or rest session was then computed by averaging regional synchrony values across all 55 regions that had significantly synchronized activity in at least one movie clip. We found a positive, near-significant correlation between neural synchrony and total cognition prediction accuracy (Pearson $r = 0.52$, $p = 0.0667$), and a weaker, non-significant correlation between synchrony and sex prediction accuracy ($r = 0.38$, $p = 0.1971$). Although the relationships did not reach statistical significance ($p < 0.05$), the positive trend suggests that higher synchrony may contribute to improved total cognition prediction accuracy.

Regional synchrony maps showing the average ISC across all pairs of individuals for each movie clip and rest session are shown in Fig. 4b, top row. During most movie-watching sessions, the highest synchrony was observed in auditory, visual, and language-related regions. For example, peak synchrony reached around 0.65 in the auditory cortex during *Inception* and 0.48 in *Ocean's 11*. Generally, auditory region synchrony was higher than visual

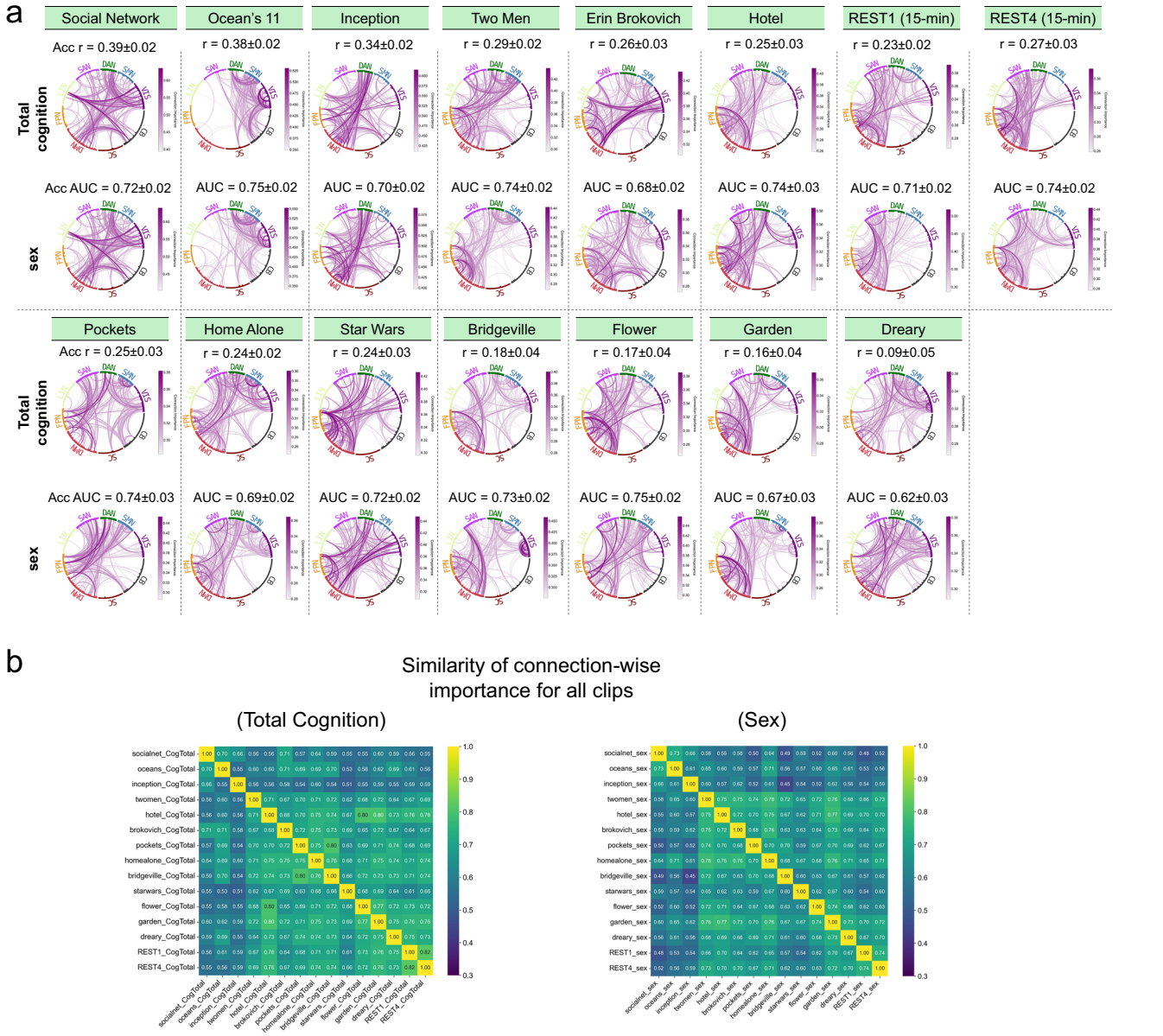

Fig. 3 | Brain connections important to total cognition and sex predictions. a For each clip and resting-state session, chord diagrams visualize the most important connections—specifically, those within the top 1% of average importance scores—for both total cognition and sex prediction models. The most important connections appear to vary across clips and remain consistent for total cognition and sex within a clip. These critical connections are predominantly located in sensory networks (VIS, SMN) and higher-order networks (DMN, FPN, SAN, DAN), though the specific top connections vary across different movie clips. **b** Correlation of edge-wise FC importance across all clips, for total cognition prediction and sex prediction. Edge-wise FC importance similarity is assessed by correlating the importance scores of all pairwise FCs across two fMRI sessions.

region synchrony, which has been shown in the past[45] and may be ascribed to more variability in gaze location during movie watching than in what the person is hearing, the latter of which should be identical across participants. Neural synchrony during rest sessions was minimal across all regions. This is expected given the absence of a shared time-locked stimulus at rest; nonetheless, resting-state FC may still predict behavior via stable, individual-specific network organization rather than shared stimulus-evoked responses.

Next, we show region-wise feature importance maps, computed by averaging the prediction importance scores of all 267 brain connections involving that region, for total cognition and sex prediction, see Fig. 4a (middle and bottom rows). We then assessed the correlation between regional synchrony and regional feature importance for each movie clip. The correlation between cognition and sex feature importance maps for models based on the FC from the same movie clip exceeded 0.90, highlighting a strong overlap in the predictive features for both outcomes. This similarity remained significant after controlling for spatial auto-correlation using a cortical spin-based permutation test[46] across all movie clips and resting-state sessions ($p_{spin} < 0.05$). Additionally, the synchrony maps showed consistent, moderate, positive correlations with both models' feature importance maps, indicating that regions exhibiting higher inter-subject synchrony also tended to have greater predictive relevance for cognition and sex. Spin-test results indicated that most of these correlations between synchrony maps and feature importance maps are significant ($p_{spin} < 0.05$), except for the *Dreary* clip (which elicited relatively low inter-subject synchrony) and the two resting-state sessions. See scatter plots showing these correlations in Supplementary Fig. 6. Notably, we also observed considerable similarity in regional synchrony and feature importance maps across different clips, with correlations ranging from 0.67 to 0.95, see in Fig. 4b.

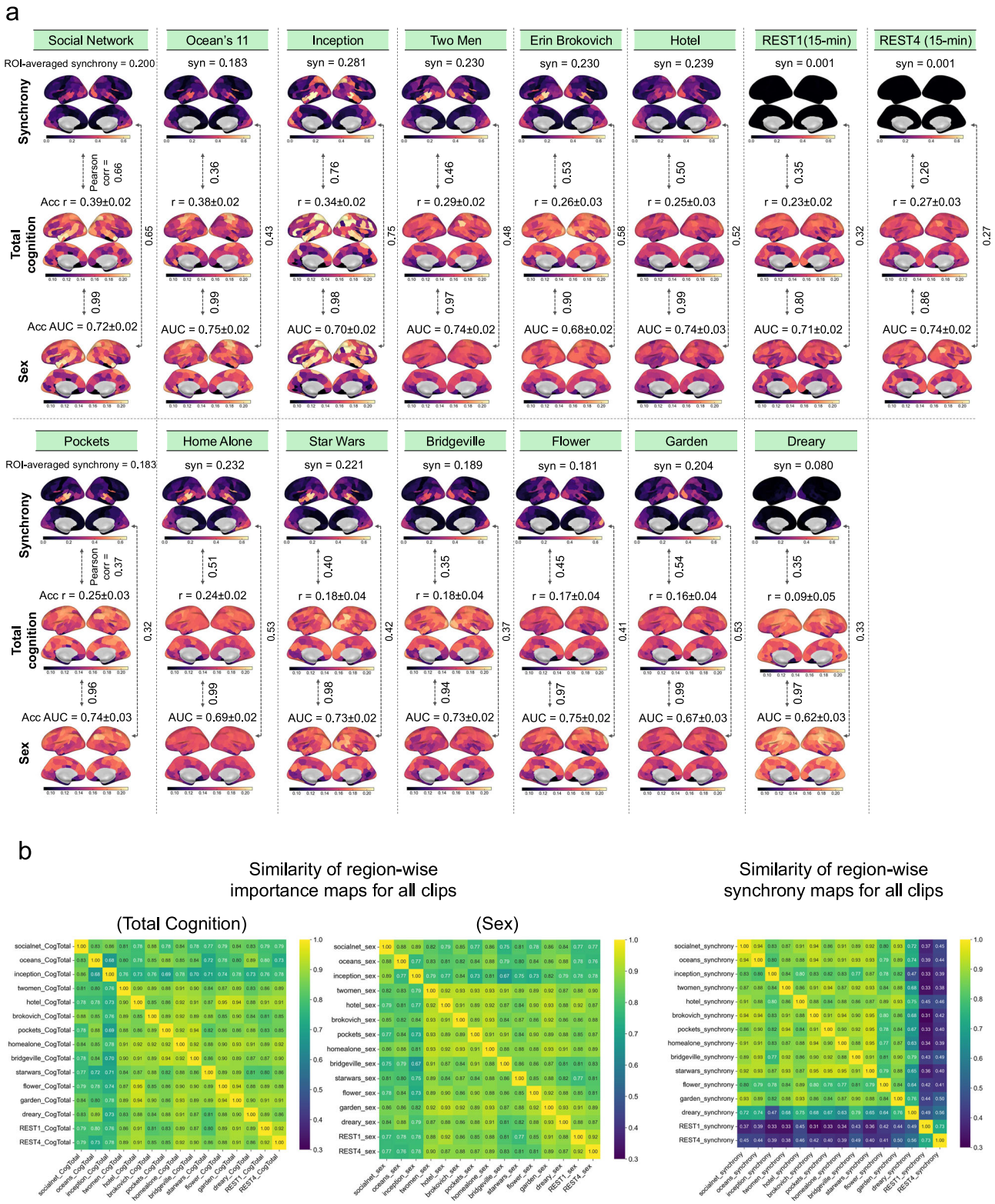

Fig. 4 | Neural synchrony and static FC feature importances. a For each clip, the first row shows regional neural synchrony maps (average across synchronized regions listed at the top of the panel). The second and third rows display region-wise prediction importance maps, i.e., the average of the importance scores of all edge-wise FCs to that region, for total cognition and sex, with the average prediction accuracy at the top of the panel. For all clips, neural synchrony and feature importance maps are positively correlated, indicating that regions with higher synchrony also tend to play a key role in behavior prediction. **b** Similarity of region-wise importance maps across all clips, for total cognition prediction, sex prediction and neural synchrony, respectively. Consistent with results from edge-level (see Fig. 3b), the region-level prediction importance is positively correlated across all clips. Neural synchrony is also consistent across clips, with the exception of clips showing particularly low synchrony, such as *Dreary*, and the two rest sessions.

To directly evaluate the role of high-synchrony regions in behavior prediction, we conducted targeted predictions using FC in subsets of brain regions with different levels of synchrony. As shown in Supplementary Fig. 8, models based on FC in regions exhibiting high synchrony consistently outperformed models based on FC of the same number of randomly selected regions in multiple movie clips, which in turn outperformed models based on FC of low-synchrony regions. These trends were consistent for both cognition and sex prediction. This finding highlights that during naturalistic movie viewing, regions with higher inter-subject synchrony tend to contribute more to FC-based prediction, particularly for cognitive outcomes.

Sliding-window FC prediction accuracy, neural synchrony and movie features are correlated

At a finer temporal resolution, movie clips elicit rich dynamic fluctuations in brain activity as visual and auditory stimuli unfold over time. To capture these moment-to-moment changes, we segmented each movie clip into overlapping sliding windows (window size = 60 s; step size = 1s) and tracked the temporal evolution of dFC prediction accuracy, neural synchrony, and high-level movie features within each window (Fig. 5a). We measure the coupling between sliding-window neural synchrony and behavior prediction accuracy as the Pearson correlation between their window-level time series. Because adjacent sliding windows are dependent due to overlap, we estimated statistical significance using a circular time-shift permutation test, which preserves each time series' temporal structure while disrupting their temporal alignment[47,48].

Several clips exhibited significant positive correlations between dynamic synchrony and total cognition prediction accuracy, including Ocean's 11 (r = 0.48, permutation-based p = 0.005), Inception (r = 0.37, p = 0.011), Erin Brockovich (r = 0.26, p = 0.032), Home Alone (r = 0.29, p = 0.025), Bridgeville (r = 0.31, p = 0.020), Flower (r = 0.33, p = 0.016), and Garden (r = 0.40, p = 0.008). Similarly, significant positive correlations between synchrony and sex prediction accuracy were observed in Two Men (r = 0.19, p = 0.049), Home Alone (r = 0.40, p = 0.009), Bridgeville (r = 0.48, p = 0.004), and Flower (r = 0.72, p < 0.001).

To gain a more comprehensive view across the different clips, we aggregated all sliding windows from all movie clips and recalculated the correlations. In this combined analysis, each point in Fig.5b represents a sliding window, colored by its corresponding movie clip. Across all windows, neural synchrony was significantly positively correlated with both total cognition dFC accuracy (Pearson r = 0.37, permutation-based p < 0.001) and, to a lesser extent, sex dFC accuracy (r = 0.22, p = 0.032). These findings are consistent with the clip-level correlations shown in Fig.2c, reinforcing the observation that periods of heightened neural synchrony are generally associated with improved cognitive and (lesser so) sex prediction.

We further explored how high-level, sliding window stimulus features modulate dynamic FC prediction performance. Total cognition dFC accuracy was positively correlated with the presence of human faces (r = 0.50, p < 0.001) and to a lesser extent voices (r = 0.30, p = 0.006). Here, face/voice presence was defined as the percentage of time points in which these features appeared across each 60-second window. Sex prediction accuracy was not correlated with face presence (r = 0.13, p = 0.134) or voice presence (r = 0.03, p = 0.756). These results align with prior findings from static FC prediction, where the presence of human faces was associated with improved cognitive prediction accuracy[6]. Together, these results suggest that socially rich moments in naturalistic stimuli may enhance both neural synchrony and the predictive power of FC patterns in mapping to cognitive scores - but not in sex classifications.

To further investigate the semantic content underlying accurate dFC predictions, we analyzed frame-level WordNet features of the movie clips. Within each sliding window, we computed the frequency of WordNet features and weighted them by dFC prediction accuracy (for either cognition or sex). We then visualized the weighted feature distributions as word clouds (Fig. 5c), where word size reflects the weighted average frequency (weighted by prediction accuracy so more predictive windows are given larger weights) and averaged across all windows and clips. For total cognition, human- and social-related features emerged as most predictive, including terms such as "male," "man," "person," "adult," "talk," and "look." This further supports our finding that naturalistic stimuli with high social content would elicit brain activity that has higher cognition predictive power. For sex prediction, the associated features were more abstract and general, including "entity," "physical entity," "object," "living thing," "artifact", "whole," and "causal agent", suggesting that sex-related neural patterns may be less tied to socially specific stimuli and more broadly distributed.

Discussion

In this study, we demonstrate the potential of DNNs to predict behavioral traits from the brain's FC during movie watching, using data from the HCP 7T substudy. We show that while certain movie clips induce FC that has higher predictive power in terms of cognitive scores and sex than resting-state, others do not. We identify that movie clips with higher across-subject synchrony of brain activity also have better prediction performance for cognition, but not for sex classification. On a dynamic, finer temporal scale, we show that neural synchrony and prediction performance are significantly correlated over time for both cognition and, to a lesser extent, sex. Importantly, these synchrony-performance relationships are evaluated only within naturalistic viewing, where subjects share a time-locked stimulus, and are not intended to generalize to resting-state scans, which lack shared stimulus-locked responses. Additionally, cognitive prediction accuracy is positively influenced by the presence of human faces and voices, while sex classification accuracy is not. Finally, we demonstrate that semantic content related to human and social stimuli enhances cognition prediction performance, and content related to non-human objects enhances sex classification performance. We find that connections within and between sensory networks (VIS, SMN) and higher-order networks (DMN, FPN, SAN, DAN) are key for predicting both cognition and sex, with the most salient connections for both cognition and sex being similar but varying across different movie stimuli and rest. In conclusion, our findings highlight the importance of visual, auditory, and higher-order network connections in behavior prediction under naturalistic stimuli, underscore the role of neural synchrony in prediction accuracy, and suggest that human- and social-related movie content is especially effective in stimulating cognition-related brain activity.

The proposed DNN for brain-behavior mapping introduces non-linearity and complex feature spaces, effectively capturing intricate relationships between FC and behavioral traits. We highlight that movie-watching FC yields a wide range of prediction accuracies for cognition and sex, depending on each clip's ability to elicit neural synchrony and the clip's content. Some clips numerically outperform resting-state FCs of the same duration, and some even outperform rest sessions that are five times longer. While the model's prediction results demonstrate its effectiveness, there are some limitations. Compared to previous studies on cognition prediction from movie-watching connectivity using linear models (connectome-based predictive modeling, CPM)[6], our results are similar overall, but the prediction accuracy is slightly higher for highly predictive clips (e.g., *Social Network*) and lower for less predictive clips (e.g., *Dreary*). We also note that some studies can achieve higher sex classification accuracy from a training set with more subjects and longer movie-watching[49]. This discrepancy may be due to overfitting, which can arise from the high complexity of the DNN, the high-dimensional input FC vectors, and the relatively small number of training subjects. To mitigate overfitting, we employed several strategies in the model design, applied strong regularization techniques, and reduced dimensionality via PCA. While the PCA+MLP framework performed better than other machine learning approaches (e.g., XGBoost, convolutional neural networks) we have tried, future research could explore more advanced non-linear approaches, including topography-based networks[16], convolutional-recurrent models[19] and graph neural networks[50,51].

Our findings also highlight the previously unexplored relationship between inter-subject neural synchrony during movie watching and FC

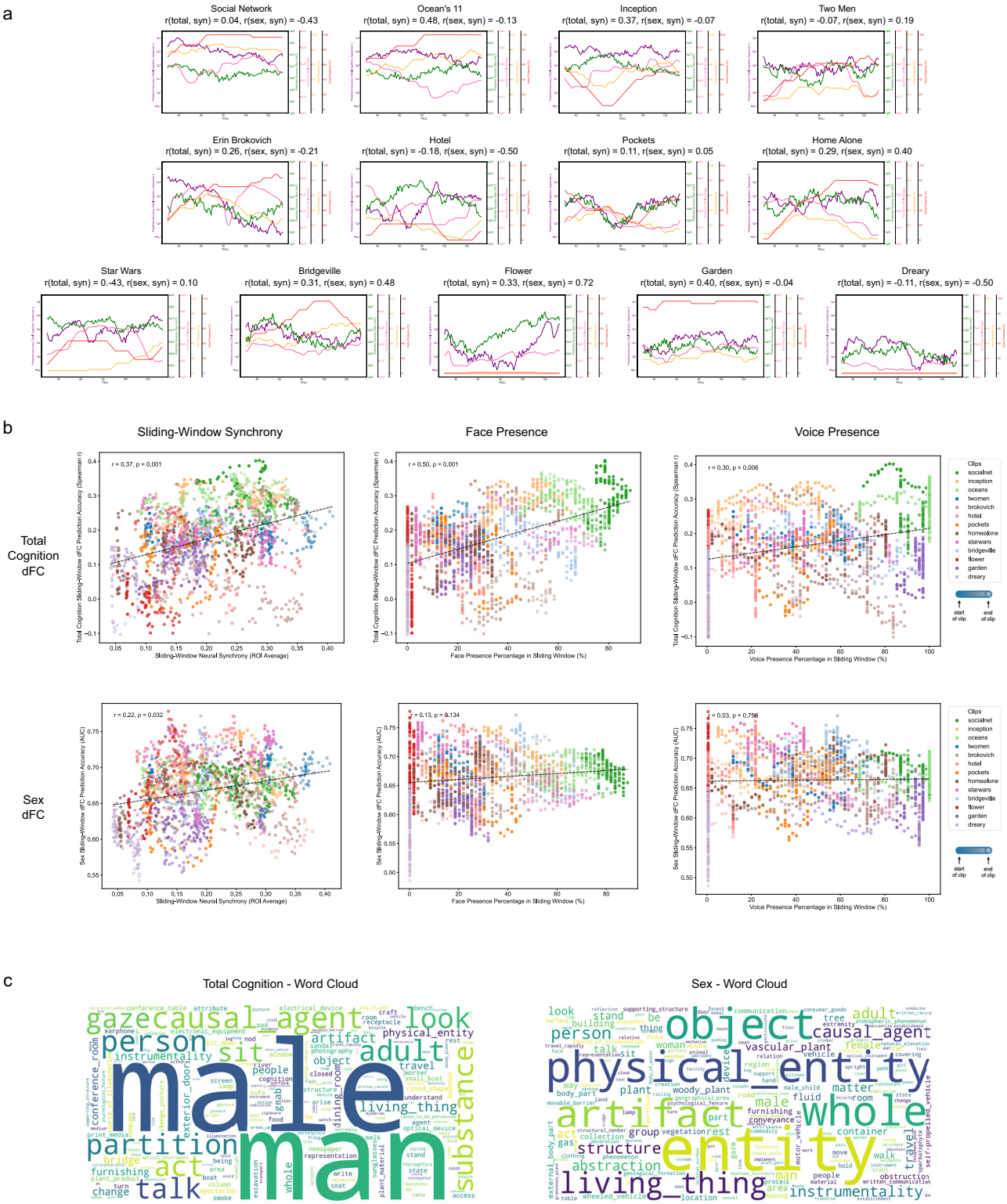

Fig. 5 | Sliding-window dynamic FC prediction, neural synchrony, and movie features co-fluctuate over time. a Sliding-window dynamic FC prediction accuracies for total cognition (purple) and sex (green), neural synchrony (pink), and percent of human face/voice presence (yellow/red, respectively) for all movie clips. **b** Correlations between sliding-window dynamic functional connectivity (dFC) prediction accuracies and sliding-window neural synchrony, face presence, and voice presence are computed across all sliding windows in all movie clips. Each point represents a sliding window, each color a corresponding movie clip. Point transparency indicates temporal position within the clip, ranging from 100% opacity at the beginning to 50% at the end. Sliding-window synchrony shows a positive correlation with both total cognition dFC prediction ($r = 0.37$, $p = 0.001$) and sex dFC prediction ($r = 0.22$, $p = 0.032$), with r values calculated by Pearson correlation and p values estimated using a 10,000-iteration permutation test. Total cognition dFC prediction is also positively correlated with face ($r = 0.50$, $p = 0.001$) and voice ($r = 0.30$, $p = 0.006$) presence percentage within sliding-windows. Sex dFC prediction is non-significantly positively correlated with face presence ($r = 0.13$, $p = 0.134$) but not correlated with voice presence ($r = 0.03$, $p = 0.756$). **c** The average frequency of WordNet features within each sliding window, weighted by that window's dFC prediction accuracy for total cognition (left panel) or sex (right panel). The size of the words represents the weighted average count; larger words correspond to a feature that is present during more predictive sliding windows.

prediction performance. Here, we focus exclusively on inter-subject synchrony during movie watching and use synchrony as a stimulus-evoked reliability marker within the naturalistic paradigm, rather than as a general predictor across paradigms. There is a potential tradeoff in inducing neural synchrony—on one hand, higher synchrony improves the SNR of BOLD activity and corresponding FC; on the other hand, higher synchrony means less inter-subject variability, which may negatively impact behavior prediction[52]. However, our overall finding of a positive correlation between neural synchrony and cognition prediction performance (and to a lesser extent sex classification) supports the conclusion that the primary benefit of synchrony lies in improving SNR of FCs—at least for this level of synchrony. While previous studies have pointed out that periods of high inter-subject FC similarity tend to align across viewers and are linked to the brain's modular architecture[33], we here directly examine how similarity in BOLD responses affects brain-behavior mapping. During movie watching, periods of higher synchrony tend to provide more accurate predictions, and brain regions of high synchrony tend to be more informative in terms of cognitive abilities and sex classification. This interpretation does not conflict with the moderate prediction performance we observe at rest, which is plausibly driven by stable, individual-specific connectivity organization rather than shared time-locked neural responses. These synchronizing moments are sometimes referred to as "events", though in the context of co-fluctuations in z-scored BOLD activity across brain connections[29]. This finding is especially relevant for future fMRI experiment design, suggesting that stimuli that evoke higher synchrony may be preferable. It would be of interest to try and maximize synchrony to test if accuracy in predicting individual metrics would hit a ceiling—it is our opinion that such a level of synchrony was not inducible due to inherent inter-individual variability in brain activity responses to the same stimuli.

At both static and dynamic scales, we found that the percentage of movie frames with human faces and voices was positively correlated with prediction accuracy for cognition, but not for sex. Additionally, we found that predictions of fluid and crystallized cognition are strongly associated with face presence (more so fluid), and to a lesser extent, voice presence, as shown in Supplementary Fig. 2. These findings suggest that movies rich in social content engage brain networks relevant to cognition, but not to sex. This result aligns with previous findings at both the clip and dynamic levels: at the clip level, studies have shown that the duration of face presence is positively correlated with prediction accuracy during movie watching[6]; at the dynamic level, movie watching has been shown to elicit co-fluctuations in both core and extended face-processing regions, which in turn correlate with face recognition scores[53]. The stronger correlation between cognitive prediction and face presence than voice presence is also consistent with previous findings that face-selective brain regions show more robust and distributed identity-related neural patterns than voice-selective regions[54]. On the other hand, objects not related to humans—both living and non-living— were related to better sex classification. These findings have practical implications for designing future naturalistic viewing experiments—the content of the movie may be varied to obtain more accurate brain-behavior mappings, but this accuracy appears to depend on the specific outcome one is trying to predict.

Together, our results suggest that naturalistic stimuli should not be treated as interchangeable: different clips can engage partially distinct perceptual and cognitive processes, modulating the reliability (e.g., synchrony/SNR) and apparent salience of FC patterns. This implies an important design consideration for naturalistic fMRI studies: stimulus choice should be matched to the scientific goal and target phenotype. At the same time, this stimulus dependence introduces a limitation for interpretability and generalization across arbitrary movies. However, the moderate-to-high similarity of feature-importance patterns across clips at the edge and network levels, together with recurring sensory-association network motifs, supports the existence of a shared substrate for prediction that is expressed to varying degrees depending on stimulus properties.

Our implementation of DNNs in the brain-behavior mapping allowed us to employ the IG method to identify the most predictive connections on an individual subject basis. The gradient-based approach measures feature importance at the connection level, while the interpolation method enhances robustness. Our way to assess general feature importance across the population was to average the individual-level feature importances from the IG framework, however, one could also look at the variability of the feature importances across individuals or subgroups within the overall group. We observe that the top FC features important for predicting cognition and sex are more consistent across movie clips than across the different outcomes that are being predicted, but that, generally speaking, the entire vector of feature importances was still moderately to highly correlated across movies (Figs. 3 and 4 and Supplementary Figs. 4 and 7). The most important connections across the movie-watching FC mapping included connections within and between sensory networks (VIS, SMN) and higher-order networks (DMN, FPN, SAN, DAN). This finding is consistent with prior work using support vector machine models[25], which identified the FPN, DMN, and DAN as key contributors to predicting cognitive states during movie-watching. It also partly agrees with work using convolutional models and deconvolution-based importance mapping[49], which showed that the DMN has the highest predictive power for sex classification, while the FPN is most predictive of fluid and crystallized intelligence.

Importantly, our results reveal a high level of similarity between cognition and sex prediction importance maps within a clip compared to across a clip for the same outcome. Looking at the individual NIH Toolbox cognition test scores, we see a similar pattern of moderate levels of agreement across all movies (and rest) but tighter correlations within a movie for different cognitive scores than across a movie for the same cognitive score. Rather than reflecting a clip-invariant, outcome-specific FC signature, this pattern suggests that different movies emphasize partially distinct subsets of predictive connections. Interestingly, we see that the highest performing movie FC (*Ocean's 11*, *Social Network* and *Inception*) feature importances are more similar to one another and tend to be more different from the other movie FC feature importance maps.

Finally, the movie FC feature importance maps at both the regional and edge-wise levels were correlated with those derived from resting-state FC to a degree comparable to correlations observed across different movies. We interpret this as evidence for a partially shared, trait-like predictive substrate that generalizes across brain states. At the same time, movies can introduce stimulus-dependent deviations in which connections are emphasized, and within the naturalistic viewing paradigm, higher inter-subject synchrony appears to enhance the reliability (SNR/stability) of FC estimates and thereby improve prediction accuracy.

The analysis of movie-watching FC-based brain-behavior mapping is crucial for understanding the unique value of this paradigm in understanding the neural circuits underlying behavior traits or demographics. We find that two key elements correlated with movie-watching FC brain-behavior mapping accuracy are across-subject neural synchrony and movie features; we find that the movie features driving accuracy may depend on the metric we are trying to predict. Specifically, we find that movies with human- and social-related content are most effective for cognition, and non-human object elements drive sex classification. In summary, these results highlight the importance of naturalistic viewing in fMRI neuroimaging and brain-behavior mapping, offering rich insights into brain mechanisms, and pointing us to further exploration of the mapping between movie features, neural synchrony, and accuracy in brain-behavior mapping using naturalistic paradigms.

Methods

Data

Dataset, participants, and behavioral measures. In this study, we use functional magnetic resonance imaging (fMRI) and behavioral data from the Human Connectome Project (HCP) 7T movie-watching dataset[41]. The dataset comprises n=184 healthy adult participants (ages 22–35), who underwent scanning using a Siemens Magnetom 7T scanner equipped with an optimized multiband echo-planar imaging sequence (TR = 1000ms, TE = 22.2ms, $voxelsize$ = 1.6mm^3). We use a subset of

$n = 176$ subjects with complete data for all resting state and movie fMRI runs, as well as behavioral data needed in this study.

In this study, we focus on predicting two items: biological sex (106 female, 70 male) and cognitive performance, both of which have been associated with FC patterns. We use the total cognition score, which is the average of crystallized cognition and fluid cognition, which are derived from the NIH toolbox cognition tests, including Picture Vocabulary, Reading Recognition, Dimensional Change Card Sort, Flanker Inhibitory Control and Attention, Picture Sequence Memory, List Sorting, and Pattern Comparison tests[55]. For completeness, we included the analysis for individual cognitive measures in Supplementary Information, but generally speaking, the individual cognitive scores are noisier than the composite scores.

fMRI acquisition and preprocessing. In the HCP 7T dataset, participants underwent four 15-min fMRI runs in which they viewed a total of 15 movie clips designed to engage cognitive and emotional processing. These clips, ranging from 1 to 4.3 minutes in length, consisted of both independent short films and excerpts from commercial Hollywood movies. The individual clips were concatenated into 12- to 14-minute-long .mp4 files, with a common Vimeo repeat validation clip included to assess test-retest reliability[56].

We excluded the test-retest clip, as well as another clip that was too short compared to the others. This resulted in a final set of 13 movie clips for our analysis. The movies were displayed using VisuaStim Digital MRI-compatible goggles to ensure consistent visual stimulation. We used brain activity data from all movie clips between the 10th and 174th seconds from the beginning of each clip, excluding the initial unstable period and aligning the end of the window with the shortest clip duration to ensure comparability across clips. The specific movie clips analyzed in this study were as follows:

- MOVIE1 (first session, first day): *Two Men, Welcome to Bridgeville, Pockets*
- MOVIE2 (second session, first day): *Inception, The Social Network, Ocean's 11*
- MOVIE3 (first session, second day): *Flower, Hotel, Garden, Dreary*
- MOVIE4 (second session, second day): *Home Alone, Erin Brockovich, Star Wars*

Clips in MOVIE1 and MOVIE3 are freely available independent films under Creative Commons licensing, and clips in MOVIE2 and MOVIE4 are from Hollywood movies as prepared and published by Cutting et al. 2012[57].

Resting-state fMRI (rfMRI) data were collected in four separate runs, each lasting approximately 16 minutes, at the beginning of each of the four 7T imaging sessions. During these runs, participants were instructed to keep their eyes open and maintain a relaxed fixation on a bright cross-hair projected against a dark background in a dimly lit scanning environment. For the primary movie-versus-rest comparisons in the main text, we used the first (REST1) and fourth (REST4) resting-state sessions, as these were acquired on the same days as the movie-watching sessions. To ensure a fair comparison between movie clips and resting-state data of matched duration and temporal position, we additionally segmented the resting-state runs into duration-matched "rest blocks" using the same temporal boundaries as the 13 analyzed movie clips. Specifically, for each movie clip, a corresponding resting-state block was extracted from the matched rest run using the identical start and end TRs. For example, because the clip *Two Men* spans TR 20 to TR 184 in MOVIE1, the corresponding resting-state block was defined as TR 20 to TR 184 in REST1. In this way, we obtained 13 resting-state blocks directly comparable to the 13 movie clips.

For visual directness in Fig. 2, the results from the matched resting-state blocks are summarized within the cropped resting-state column rather than displayed as 13 separate bars. Detailed prediction results for the individual resting-state blocks are provided in Supplementary Fig. 15. In addition, for a control analysis addressing the unequal number of movie and resting-state conditions, we also constructed matched resting-state blocks from REST2 and REST3 and report the corresponding results in Supplementary Fig. 15.

Head motion was quantified using the relative root-mean-square (RMS) framewise displacement provided by the HCP 7T preprocessing pipeline. For each subject and each fMRI run or movie clip, we computed the mean relative RMS motion across all TRs within the corresponding time interval. For full resting-state runs, we used the subject-level mean relative RMS motion averaged across the entire session. For individual movie clips, motion estimates were obtained by restricting the session-level motion time series to the TR range corresponding to each clip, yielding clip-specific mean relative RMS values. In analyses involving dFC, motion was quantified within each sliding window as the mean relative RMS across the TRs contained in that window, ensuring that motion estimates matched the temporal resolution of the connectivity analysis. Group-level summaries of head motion across sessions and clips are reported as mean ± SD across subjects in Supplementary Table SI1 and Supplementary Table SI2, and the distribution of motion values is illustrated in Supplementary Fig. SI11.

Standard preprocessing of fMRI data, including motion correction, distortion correction, high-pass filtering, and nonlinear alignment to MNI template space[58] was used in the FIX-denoised volume-space data from the HCP 7T release. We also applied global signal regression (GSR) by removing the global mean time series signal from each voxel's BOLD time series, thereby reducing potential confounds arising from widespread non-neuronal noise, including physiological artifacts and scanner-related fluctuations[59].

Static and dFC construction

Static functional connectivity. To quantify static functional connectivity (sFC), we employed the Shen 268-region brain atlas[60]. For each region in the atlas, we computed the mean BOLD time series by averaging voxel-wise preprocessed signals within that region for each subject, run, and time frame. We then computed pairwise Pearson correlation coefficients between all 268 regions, yielding a 268×268 FC matrix for each movie clip or rest run per subject. We then extracted the upper triangular portion of each matrix (excluding the diagonal) and vectorized it into a 35,778-dimensional feature vector[6].

Dynamic functional connectivity. To capture temporal fluctuations in FC during movie-watching, we adopted a sliding-window approach for dFC. Each movie clip was segmented into overlapping time windows of 60 seconds, with a step size of 1 second, resulting in 105 sliding windows for each movie clip. Within each window, the mean BOLD signals were computed, and Pearson correlation was used to generate a 268×268 FC matrix for each window, resulting in 105 sliding-window dFC matrices for each clip.

DNN prediction model

Model structure. We propose a novel DNN-based behavior prediction model designed to capture the complex relationship between dynamic/static FC and behavioral and demographic traits (Fig. 1). The model takes as input a set of $n = 176$ FC vectors, each of size (35,778), representing the upper triangular elements of the 268×268 FC matrix. The output corresponds to either cognition score (a continuous variable regression task) or biological sex (a binary classification task). Given the relatively small number of subjects, we employed a 10-fold cross-validation approach to ensure robust model evaluation. The 176 subjects were randomly divided into 10 groups, while making sure that individuals from the same family were always assigned to the same group. This step prevents data leakage and reduces the risk of inflated performance estimates due to genetic or familial similarities in brain connectivity. In each cross-validation iteration, the model was trained on nine groups and tested on the remaining one group, ensuring that every subject was included in the test set exactly once. Predicted values for all subjects were collected across the 10 folds and then correlated with the true values to evaluate model performance. The division into 10 folds was performed 100 times to assess

uncertainty due to varied groupings of individuals into training and test sets.

The prediction model consists of two primary components:

(1) Principal component analysis (PCA) block:

To reduce dimensionality and minimize noise, we apply PCA to the input FC vectors across all training subjects, retaining the first 100 principal components. This results in a (#training subjects, 100) feature matrix, capturing generally around 84% of the variance.

(2) Multi-layer perceptron (MLP) block:

- The PCA-reduced data is fed into a three-layer MLP, each layer containing 500 nodes with ReLU activation.
- To reduce overfitting, we employ dropout (rate = 0.5) in each hidden layer and apply L2 weight decay (0.01) as regularization.
- The model is optimized using the Adam optimizer with a learning rate of 0.001.
- The output layer is:
  - – Linear for cognition prediction (regression task).
  - – Sigmoid for sex classification (binary classification task).
- The loss function is:
  - – Mean Squared Error (MSE) loss for cognition prediction.
  - – Binary Cross-Entropy Loss for sex prediction.

While overfitting remains a challenge due to the limited sample size and the high-dimensional, noisy nature of fMRI-based FC measures[61–64], the model's hyperparameters were carefully tuned to strike a balance between learning the complex FC behavior mapping and avoiding excessive overfitting. The model is developed using Pytorch on a CUDA GPU environment. To evaluate the model's predictive performance, we computed Spearman's correlation coefficient (r) between the predicted and true behavior scores for cognition. For sex classification, we assessed the model using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.

Hyperparameter tuning and early stopping. To prevent optimistic bias from tuning on the held-out test folds, all model selection was performed strictly within the training data of each outer cross-validation fold (nested cross-validation). Specifically, for each outer 10-fold split (constructed family-wise so that related individuals were always assigned to the same fold), we performed an inner fivefold family-wise cross-validation on the outer training set only. Subjects from the same family were always assigned to the same fold in both outer and inner splits.

The grid search space included: PCA dimensionality {10, 20, 50, 100, 200, 500, 1000, 5000}; number of hidden layers {3, 4, 5}; number of nodes per layer {100, 200, 500, 1000}; learning rate $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$; weight decay $\{10^{-2}, 10^{-4}, 10^{-6}\}$; and dropout rates {0.3, 0.5, 0.7, 0.9}. Within each inner split, PCA (and feature scaling) was fit using only the inner training subjects and then applied to the corresponding inner validation subjects. Hyperparameters were selected by maximizing inner-loop validation performance (Spearman's r for cognition prediction; ROC-AUC for sex classification).

We examined the distribution of optimal hyperparameters selected across outer folds and across movie/rest sessions. The selection frequencies were highly consistent: PCA dimensionality of 100 was selected in 95% of folds; a 3-layer MLP architecture in 90% of folds; 500 nodes per layer in 96%; learning rate 10^{-3} in 100%; weight decay 10^{-6} in 98%; and dropout configuration (0.5, 0.5, 0.3) in 90% of folds. Given this high consistency in optimal hyperparameter selection, we adopted a single consensus hyperparameter configuration for all clips and sessions. This design choice was made to minimize the influence of clip-specific model architecture differences and ensure that differences in prediction accuracy across movie clips reflected stimulus-related effects rather than variability in model configuration. Based on this consensus selection, the final hyperparameter configuration was fixed as follows: PCA dimensionality = 100; MLP with three hidden layers, each containing 500 nodes; learning rate = 10^{-3}; weight decay = 10^{-6}; and dropout rates of 0.5, 0.5, and 0.3 for the three hidden layers, respectively.

We employed early stopping to reduce overfitting: during training, validation loss (MSE for cognition; binary cross-entropy for sex) was monitored on the inner validation fold, and training stopped when validation loss failed to improve for a fixed number of epochs (patience); the model parameters from the epoch with the best validation loss were retained. The model was refit on the full outer training set (again fitting PCA/scaling on training subjects only) and evaluated once on the untouched outer test fold. The outer 10-fold process was repeated 100 times with different family-wise fold assignments.

Confound control in behavior prediction. Potential confounds related to head motion and acquisition time were controlled in the behavior prediction model. These two factors have been shown in previous work on the HCP dataset to correlate with cognitive measures and can therefore influence predictive modeling results[6,65]. Because the movie stimulus is identical for all subjects within a given clip, confound correction was performed separately for each clip.

For cognition prediction (continuous outcome), head motion was treated as a nuisance covariate and regressed out from the behavioral target within each training fold using linear regression, following methods in prior works[6]. The fitted confound model was then applied to the held-out test fold without refitting, ensuring that no information from the test subjects influenced the confound correction procedure. This approach prevents information leakage while removing variance in the behavioral target that can be attributed to motion.

Consistent with prior reports on the HCP 7T dataset[6], both head motion and acquisition time show negative correlations with cognition scores, but these relationships do not differ across runs. Therefore, these factors are unlikely to explain the observed differences in prediction performance between movie-watching and resting-state conditions. Additional clip-level analyses confirming this pattern are reported in Supplementary Fig. SI12.

For binary outcomes (sex classification), residualizing the 0/1 label is not appropriate. Instead, we performed two complementary analyses to evaluate the influence of potential confounds. First, we trained a confounds-only baseline classifier using logistic regression with head motion (mean relative RMS) and acquisition time as predictors. This model achieved near-chance performance (ROC-AUC ≈ 0.5), indicating that these variables alone do not carry predictive information about biological sex (Supplementary Fig. 13).

Second, following the reviewer's suggestion to include motion-related covariates directly in the prediction model, we re-ran the FC-based classifier after concatenating head motion and session-related covariates with the FC-derived features (FC+confounds). Sex classification performance using FC +confounds was unchanged relative to the FC-only model, indicating that prediction accuracy was not driven by motion or session-related variables. The results of these analyses are summarized in Supplementary Fig. 13.

Session-level factors such as imaging day, run identity, and movie type do not vary across subjects within a given clip and therefore cannot act as subject-level confounds in the within-clip prediction models. These variables were instead considered in analyses that compare neural measures across clips or sessions.

Statistical analysis of predictive connections and networks

Calculation of edge-wise and region-wise feature importance. To quantify the predictive contribution of individual brain connections to the predictions, we employed IG, a technique designed to estimate feature importance in DNNs[66,67]. IG uses backward propagation to compute gradients of the model's output with respect to the input, and improves robustness by integrating these gradients across a path between a baseline and the actual input.

In our DNN behavior prediction model, the IG procedure is as follows:

- PCA transformation: First, we perform Principal Component Analysis (PCA) on the training FC vectors, resulting in a (100,1) principal component(PC) vector for each training subject.
- Baseline Definition: The baseline reference input is defined as the average of each of the PC vectors across all training subjects (i.e., one vector per PC).
- Interpolation: A series of intermediate vectors is generated by linearly interpolating between the baseline and the actual input principal component (PC) vector of a subject. Specifically, for an input PC vector $\mathbf{x}$ and baseline $\bar{\mathbf{x}}$, we compute the kth step:

$$\mathbf{x}^{(k)} = \bar{\mathbf{x}} + \frac{k}{m}(\mathbf{x} - \bar{\mathbf{x}})$$

  where $m = 50$ is the number of interpolation steps.
- Gradient Computation: For each interpolated vector $\mathbf{x}^{(k)}$, we compute the gradient of the DNN's output (predicted cognition score or sex classification probability) with respect to its ith feature $\mathbf{x}_i^{(k)}, i = 0, \ldots, 99$.
- Integration Along the Path: The importance of each feature i of the PC vector is estimated by integrating these gradients over all interpolation steps:

$$\begin{aligned} IG(x_i) &= \frac{1}{m}(x_i - \bar{x}_i)\sum_{k=1}^{m}\frac{\partial F(\mathbf{x}^{(k)})}{\partial x_i} \\ &= \frac{1}{m}(x_i - \bar{x}_i)\sum_{k=1}^{m}\frac{\partial F(\bar{\mathbf{x}} + \frac{k}{m}(\mathbf{x} - \bar{\mathbf{x}}))}{\partial x_i}, i = 0, \ldots, 99 \end{aligned}$$

  where $F(x)$ is the DNN's prediction function, and $\frac{\partial F}{\partial x_i}$ represents the gradient of the model output with respect to the ith entry of the input PC vector, computed by backward propagation.
- PCA inverse: The gradient is multiplied by the inverse of PCA loadings to project back to the gradients of the whole set of 35,778 brain connections.
- Feature Importance Assignment: For each subject, we obtain a set of IG values for each functional connection (i.e., each of the 35,778 connections). The absolute values of the IG for each connection are averaged across all subjects to compute its importance score, reflecting its contribution to behavior prediction.

At the regional level, we computed the average of the absolute values of the importance scores for the 267 connections associated with each region, which was used as its relative importance. At the functional network level, we mapped the most important connections to their corresponding Yeo canonical functional networks[42] and analyzed which functional networks contained the most predictive connections.

Analysis of feature map similarity and decomposition of clip vs. outcome effects

Decomposing similarity into clip vs. outcome effects using a mixed-effects regression model. To quantify the similarity structure of predictive feature maps across movie clips and behavioral outcomes, we computed pairwise Pearson correlations between all importance maps (both edge-wise and region-wise), see Supplementary Fig. SI4 (edge-wise) and Supplementary Fig. SI7 (region-wise). Correlations were Fisher-z transformed prior to statistical modeling.

To explicitly disentangle stimulus-driven (clip-related) effects from outcome-specific effects, we modeled feature-map similarity using a mixed-effects regression framework. For each pair of feature maps (i, j), the Fisher-z transformed similarity $z(r_{ij})$ was modeled as:

$$\begin{aligned} z(r_{ij}) &= \beta_0 + \beta_1\,\text{same\_clip}_{ij} + \beta_2\,\text{same\_outcome}_{ij} \\ &\quad + \beta_3\,\Delta\text{face}_{ij} + \beta_4\,\Delta\text{voice}_{ij} + \beta_5\,\Delta\text{sync}_{ij} \\ &\quad + u_{\text{clip}_i} + u_{\text{clip}_j} + \varepsilon_{ij}. \end{aligned} \tag{1}$$

Here, `same_clip` indicates whether the two maps were derived from the same movie clip, and `same_outcome` indicates whether they correspond to the same behavioral outcome (cognition or sex). The Δ terms represent absolute differences in clip-level stimulus covariates (face presence, voice presence, and neural synchrony) between the two maps. Random intercepts were included for both clips in each pair to account for non-independence.

This framework allowed us to quantify the relative contributions of clip identity, behavioral outcome, and stimulus covariates to the observed similarity structure of predictive FC maps.

Meta-analysis of outcome-specific ΔIG patterns. To test whether outcome-specific FC patterns generalize across movie stimuli, we computed within-clip difference maps defined as:

$$\Delta IG^{(c)} = IG^{(c)}_{\text{cognition}} - IG^{(c)}_{\text{sex}}, \tag{2}$$

for each clip c. These ΔIG maps quantify outcome-specific differences in feature importance within a given stimulus context.

We first assessed cross-clip generalizability by computing pairwise Pearson correlations between $\Delta IG^{(c)}$ maps across clips (edge-wise). Correlations were Fisher-z transformed prior to analysis.

To derive a potential clip-invariant consensus pattern, we averaged ΔIG maps across clips:

$$\overline{\Delta IG} = \frac{1}{C}\sum_{c=1}^{C}\Delta IG^{(c)}. \tag{3}$$

We quantified the global magnitude of this consensus map as the mean absolute value across edges:

$$\| \overline{\Delta IG} \| = \frac{1}{E}\sum_{e}|\overline{\Delta IG}_e|. \tag{4}$$

Statistical significance was assessed using a sign-flip permutation test across clips, generating a null distribution under the assumption that outcome-specific effects are not directionally consistent across stimuli.

Cross-clip generalization analysis. To evaluate whether outcome-relevant FC information generalizes across stimuli, we performed a cross-clip generalization analysis. For each movie clip c_{train}, we trained the PCA+MLP model using FC derived from that clip and evaluated prediction performance on FC from a different clip c_{test}, including both movie clips and resting-state sessions.

Model training followed the same nested cross-validation framework described above, with PCA fitted exclusively on the training data for each iteration. For each train-test clip pairing, performance was averaged across 10 cross-validation iterations.

Prediction performance was quantified using Spearman correlation for cognition and ROC-AUC for sex classification. The resulting train-test performance matrices summarize the degree to which predictive information learned from one stimulus transfers to another.

Neural synchrony and its relationship to prediction performance

Neural synchrony. During naturalistic viewing, people tend to exhibit higher neural synchrony when watching engaging and cognitively stimulating movies, whereas less engaging content often results in more variable neural responses across individuals[26–30]. To quantify neural synchrony, we calculated the inter-subject Pearson correlation of

regional fMRI BOLD signals over time, reflecting shared neural responses across participants. This resulted in 268 region-specific synchrony values per subject pair. We then averaged these values across all subject pairs for each region to obtain regional population-level synchrony.

For each movie clip, we identified significantly synchronized regions based on the Pearson correlation of BOLD signals across all subject pairs. To correct for the large number of pairwise comparisons (176 × 175/2 total pairs), we applied False Discovery Rate (FDR) correction using the Benjamini-Hochberg procedure[68]. Regions with adjusted p values below $\alpha = 0.05$ were considered significantly synchronized. We took the union of significant regions across clips (collected into the set called ROI), and then averaged these to obtain a whole-brain synchrony measure; this procedure excludes regions with non-significant synchrony across all movie clips.

- Pairwise neural synchrony for a pair of subjects (i, j), for a region v:

$$r_{i,j;v} = Pearsonr(BOLD_{i;v}, BOLD_{j;v})$$

- Region-wise synchrony for a region v for all subject pairs:

$$r_v = \frac{1}{n(n-1)/2}\sum_{i=1}^{n}\sum_{j=i+1}^{n} r_{i,j;v}, n = 176$$

- Global, whole brain synchrony for a pair of subjects (i, j):

$$r_{i,j} = \frac{1}{\#ROI}\sum_{k=1}^{\#ROI} r_{i,j;v_k}, v_k \in \{ROI\}$$

Relationship between neural synchrony and behavior prediction performance. To investigate the relationship between neural synchrony and behavior prediction, we analyzed both whole-brain and regional levels. At the whole-brain level, we computed an overall synchrony score for each movie clip by averaging synchrony across all regions of interest (ROIs). We then correlated this score with the prediction accuracy for cognition and sex across all movie clips to assess whether a significant relationship exists.

At the regional level, we examined the correlation between the region-wise synchrony map and the region-wise prediction importance map. The prediction importance map was calculated by averaging the importance scores of all 267 functional connections linked to each brain region. A positive correlation would indicate that regions with higher synchrony are also more predictive. To verify the robustness of the correlation, we used both Pearson and Spearman correlation metrics. This dual approach helps ensure the correlation is not driven solely by low-synchrony regions, where predictive power might still vary despite minimal synchrony.

To further evaluate the impact of high- and low-synchrony brain regions on behavior prediction, we conducted targeted predictions using subsets of brain regions exhibiting distinct levels of synchrony. For each movie clip, we identified high-synchrony regions as those with ISC values satisfying FDR-corrected $p < 0.05$. We then selected an equal number of randomly chosen regions and an equal number of low-synchrony regions (ranked in descending order of p values). Using these subsets, we trained separate DNN prediction models while maintaining the same model architecture and compared their prediction accuracies. This approach allowed us to determine whether high-synchrony regions contribute more strongly to behavior prediction compared to randomly selected or low-synchrony regions.

Control for clip-level confounds in cross-clip analyses. When examining relationships between clip-level neural synchrony, prediction accuracy, and stimulus features, we additionally evaluated whether these associations could be influenced by clip-level confounds. Because these analyses operate at the level of movie clips rather than individual subjects (n = 13 clips), potential confounds arise from properties of the experimental design and stimulus characteristics rather than subject-level variability.

For each clip c, we defined the clip-level prediction accuracy y_c (Spearman correlation for cognition prediction; ROC-AUC for sex classification), the clip-level predictor x_c (e.g., mean ROI-averaged neural synchrony, or stimulus features such as face or voice presence), and a set of clip-level covariates C_c. The covariates included: mean head motion averaged across subjects within the clip (*motion_mean*), imaging day (Day 1 vs. Day 2), movie category (*movie_type*, Hollywood vs. independent), and the clip start time within the run (*start_tr*), which served as a proxy for order or fatigue effects.

We first report the raw Pearson correlation between synchrony (or stimulus features) and prediction accuracy across clips. To assess robustness to potential confounds, we then performed two complementary analyses. First, we computed partial correlations by residualizing both variables with respect to the covariates and correlating the resulting residuals:

$$y_c = C_c\gamma + \varepsilon_{y,c},\ x_c = C_c\delta + \varepsilon_{x,c},\ r_{\text{partial}} = \text{corr}(\varepsilon_{y,c}, \varepsilon_{x,c}).$$

Second, we fit an equivalent multiple linear regression model,

$$y_c = \beta_0 + \beta_1 x_c + C_c\beta + \varepsilon_c,$$

and evaluated the coefficient β_1 for the predictor of interest.

Across clips, neural synchrony showed a positive association with cognition prediction accuracy in the raw analysis ($r = 0.52$, $p = 0.0667$). After adjusting for motion, imaging day, movie type, and order, the effect size remained similar (partial $r = 0.50$, $p = 0.0792$; regression $\beta_1 = 0.79$, $t = 1.54$, $p = 0.17$). For stimulus content features, face presence was strongly correlated with cognition prediction accuracy in the raw analysis (r = 0.79, p = 0.0014) but was reduced after covariate adjustment (partial r = 0.52, p = 0.0711), while voice presence showed a moderate raw association (r = 0.55, p = 0.0496) that was attenuated after adjustment (partial $r = 0.31$, $p = 0.30$).

When the movie-type covariate was excluded (due to its overlap with stimulus content), the synchrony-accuracy relationship became stronger after controlling for the remaining confounds (partial $r = 0.62$, $p = 0.0244$), and face presence remained strongly associated with cognition accuracy (partial r = 0.71, p = 0.0067). In contrast, none of the tested predictors showed robust associations with sex prediction accuracy either before or after covariate adjustment.

Full scatter plots and statistics for all confound-controlled analyses are shown in Supplementary Fig. 14.

Sliding-window dynamic FC, neural synchrony and movie features

Sliding-window dynamic FC. To capture temporal dynamics, we employed a sliding-window approach that segments each 164-second movie clip into overlapping 60-second windows with a 1-second step size. This resulted in 105 windows per movie clip, with each window defined by its midpoint (e.g., the 30th second corresponds to the window from 0 to 60 seconds).

We trained separate DNN models to predict cognitive score and sex for each 60-second sliding window dFC, ensuring that the model performance reflected the predictive power of the FC for that specific time window under the given naturalistic stimulus. While the structure of the DNN model and cross-validation method were identical to those used for static FC, the key difference was that the input FC vector was based on shorter temporal segments. This approach allowed us to generate an accuracy curve over time during the movie clip, which we later compared with fluctuations in neural synchrony and movie features.

Sliding-window neural synchrony. We use the same set of overlapping 60-second sliding windows and compute the ISC of fMRI BOLD time series within the window to generate a dynamic sequence of whole-brain neural synchrony levels. Similar to static neural synchrony, we took the average of ISC across ROIs to obtain the whole-brain synchrony level for each window. We can thus obtain the time fluctuation of the synchrony level among the subjects. At the movie clip level, we computed the Pearson correlation between neural synchrony and dynamic FC prediction accuracies across the 105 windows for each clip. A strong positive correlation would indicate a high degree of co-fluctuation and similar temporal patterns between synchrony and FC prediction performance over time. At a more aggregate level, we combined the sliding windows from all 13 movie clips and computed the same Pearson correlation between neural synchrony and dynamic FC prediction accuracies across the 13 × 105 windows. This approach provides many more data points compared to static FC, allowing for a more robust analysis and helping to further validate the relationship between synchrony and behavioral prediction performance.

For comparison, we also applied the same sliding-window analysis to resting-state data. Resting-state runs (REST1,2,3,4) were segmented into duration- and onset-matched blocks corresponding to the 13 analyzed movie clips, using the same start and end TRs (164 s per block; see Methods above). Sliding-window analyses (window size = 60 s; step size = 1 s) were then performed within each rest block to compute dynamic FC prediction accuracy and neural synchrony using the identical procedures applied to the movie data. The resulting resting-state sliding-window trajectories and synchrony-prediction relationships are presented in Supplementary Fig. 16. As expected, resting-state windows exhibited near-zero inter-subject synchrony and showed no significant association between neural synchrony and dFC prediction accuracy for either cognition or sex.

Sliding-window movie features. We also investigated whether specific high-level features of the movie, including the presence of human faces and voices, influenced the accuracy of behavior prediction. To quantify these features, we calculated the percentage of time each feature was present within each 60-second sliding window, defined as the proportion of TRs (time points) where at least one human face or one person talking was detected. The presence of human faces in each frame was identified using the open-source Python package *pliers*[69]. Human voice presence in each frame is identified using the open-source Python package *webrtcvad* (https://github.com/wiseman/py-webrtcvad). Note that the TR for fMRI scanning in the HCP-7T dataset is 1 second, meaning the "frame" here is also 1 second to align with the dataset's temporal resolution. We then examined the correlations between these features and the sliding-window dFC prediction accuracy to explore how visual and auditory elements shape dFC's predictive power.

Additionally, we used WordNet features to analyze the semantic content driving high prediction accuracy for cognition and sex. The HCP-7T dataset includes frame-wise word descriptions that label the content of each frame, as specified by the HCP movie labeling pipeline[70,71]. For each sliding window, we counted the occurrences of each word, then weighted these occurrences by the behavior prediction accuracy for that window. The average of these weighted occurrences was computed across all 13 × 105 sliding windows. This produced a weighted average for each word in the movie labeling dictionary, reflecting the words most influential for behavior prediction, which we visualized using word clouds.

Statistical inference for sliding-window correlations. Sliding-window estimates obtained with overlapping windows are temporally autocorrelated because adjacent windows share most time points (60s window; 1s step). Consequently, treating the 105 window-wise observations as independent and using i.i.d. permutations of windows would violate exchangeability by destroying the temporal dependence structure. To obtain valid p values for correlations between sliding-window measures (e.g., neural synchrony, dFC prediction accuracy, face/voice presence), we therefore used a within-clip circular time-shift permutation test that preserves each series' temporal autocorrelation while disrupting their temporal alignment[47,48].

For each clip c, let $x_c(t)$ and $y_c(t)$ denote two window-level time series of length T_c (here $T_c = 105$), indexed by window t. We computed the observed Pearson correlation r_{obs} either (i) within each clip using (x_c, y_c), or (ii) at the aggregated level by concatenating windows across clips, $x = [x_1, \ldots, x_C]$ and $y = [y_1, \ldots, y_C]$ where $C = 13$. To generate a null distribution that respects temporal dependence, for each permutation b, we circularly shifted $y_c(t)$ within each clip by a random offset $s_c^{(b)}$ (sampled independently for each clip and permutation), yielding $y_c^{(b)}(t) = y_c((t + s_c^{(b)}) \bmod T_c)$. We excluded small offsets to ensure meaningful misalignment (minimum shift = 60 windows), and then concatenated the shifted series across clips, $y^{(b)} = [y_1^{(b)}, \ldots, y_C^{(b)}]$, and computed the permuted correlation $r_b = \text{corr}(x, y^{(b)})$. The two-sided empirical p value was estimated as

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{I}(|r_b| \geq |r_{\text{obs}}|)}{B + 1}$$

with B permutations.

Statistics and reproducibility

All prediction models were evaluated using 10-fold cross-validation repeated 100 times with different family-wise fold assignments, where individuals from the same family were always assigned to the same fold to prevent data leakage. Performance was averaged across the 100 repetitions; reported values reflect the mean ± standard deviation across iterations. For cognition prediction, performance was quantified using Spearman's rank correlation coefficient between predicted and true scores. For sex classification, performance was quantified using the area under the receiver operating characteristic curve (AUC).

Statistical significance of prediction accuracy was assessed using a cross-validated permutation test in which training labels were randomly permuted within each cross-validation fold and the model was fully refit for each permutation (100 permutations per repetition, 100 repetitions). p values were computed as the proportion of permutations achieving accuracy greater than or equal to the observed accuracy, and the median p value across the 100 repetitions is reported.

For sliding-window analyses, correlations between time series (e.g., dynamic FC prediction accuracy, neural synchrony, face/voice presence) were assessed using a circular time-shift permutation test with 10,000 iterations to account for temporal autocorrelation introduced by overlapping windows. A minimum shift of 60 windows was enforced to ensure meaningful temporal misalignment. Two-sided empirical p values were estimated as the proportion of permuted correlations with absolute value greater than or equal to the observed correlation.

Regional correlations between neural synchrony maps and feature importance maps were assessed using Pearson and Spearman correlation, with spatial autocorrelation accounted for using a cortical spin-based permutation test. Feature map similarity analysis was modeled using a mixed-effects regression framework with Fisher-z transformed pairwise correlations as the outcome, with random intercepts for clip identity. ISC significance was assessed using FDR correction (Benjamini–Hochberg procedure) across all subject pairs, with regions considered significantly synchronized at adjusted $p < 0.05$.

Statistical significance was considered at $p < 0.05$. The study used $n = 176$ participants from the HCP 7T dataset with complete fMRI and behavioral data across all sessions.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The preprocessed data from HCP 7T Dataset is publicly available at https://www.humanconnectome.org/. Preprocessed shen268 atlas movie-watching BOLD time series data can be found at: https://github.com/esfinn/movie_cpm.

Code availability

The code used in this study, along with usage instructions, is publicly available at: https://github.com/Kyrrego/MovieFC_Behav_Mapping.git.

Received: 6 August 2025; Accepted: 26 May 2026;
Published online: 09 June 2026

References

1. Finn, E. S. et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* **18**, 1664–1671 (2015).
2. Shen, X. et al. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *Nat. Protoc.* **12**, 506–518 (2017).
3. Miranda-Dominguez, O. et al. Connectotyping: model based fingerprinting of the functional connectome. *PLoS ONE* **9**, e111048 (2014).
4. Rosenberg, M. D. et al. A neuromarker of sustained attention from whole-brain functional connectivity. *Nat. Neurosci.* **19**, 165–171 (2016).
5. Kong, R. et al. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cereb. Cortex* **29**, 2533–2551 (2019).
6. Finn, E. S. & Bandettini, P. A. Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage* **235**, 117963 (2021).
7. Vanderwal, T. et al. Individual differences in functional connectivity during naturalistic viewing conditions. *NeuroImage* **157**, 521–530 (2017).
8. Gal, S., Coldham, Y., Tik, N., Bernstein-Eliav, M. & Tavor, I. Act natural: functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity. *NeuroImage* **258**, 119359 (2022).
9. Vanderwal, T., Eilbott, J. & Castellanos, F. X. Movies in the magnet: naturalistic paradigms in developmental functional neuroimaging. *Dev. Cogn. Neurosci.* **36**, 100600 (2019).
10. Kröll, J.-P. et al. Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks. *NeuroImage* **273**, 120083 (2023).
11. Zamani Esfahlani, F. et al. High-amplitude cofluctuations in cortical activity drive functional connectivity. *Proc. Natl. Acad. Sci.* **117**, 28393–28401 (2020).
12. Bolton, T. A., Jochaut, D., Giraud, A.-L. & Van De Ville, D. Brain dynamics in ASD during movie-watching show idiosyncratic functional integration and segregation. *Hum. Brain Mapp.* **39**, 2391–2404 (2018).
13. Raz, G. et al. Functional connectivity dynamics during film viewing reveal common networks for different emotional experiences. *Cogn. Affect. Behav. Neurosci.* **16**, 709–723 (2016).
14. Simony, E. et al. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* **7**, 12141 (2016).
15. Tian, L., Ye, M., Chen, C., Cao, X. & Shen, T. Consistency of functional connectivity across different movies. *NeuroImage* **233**, 117926 (2021).
16. Li, X., Friedrich, P., Patil, K. R., Eickhoff, S. B. & Weis, S. A topography-based predictive framework for naturalistic viewing fMRI. *NeuroImage* **277**, 120245 (2023).
17. Hüpen, P. et al. Functional brain network of trait impulsivity: whole-brain functional connectivity predicts self-reported impulsivity. *Hum. Brain Mapp.* **45**, e70059 (2024).
18. Peng, S. et al. Heterogenous brain activations across individuals localize to a common network. *Commun. Biol.* **7**, 1–11 (2024).
19. Almeida de Souza, E., Vieira, B. H. & Salmon, C. E. G. Individual cognitive traits can be predicted from task-based dynamic functional connectivity with a deep convolutional-recurrent model. *Cereb. Cortex* **34**, bhae412 (2024).
20. Gao, J. et al. Prediction of cognitive scores by movie-watching fMRI connectivity and eye movement via spectral graph convolutions. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)* 1–5 (IEEE, 2022).
21. Kawahara, J. et al. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* **146**, 1038–1049 (2017).
22. Greene, A. S., Gao, S., Scheinost, D. & Constable, R. T. Task-induced brain state manipulation improves prediction of individual traits. *Nat. Commun.* **9**, 2807 (2018).
23. Yoo, K. et al. Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage* **197**, 212–223 (2019).
24. Smith, S. M. et al. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat. Neurosci.* **18**, 1565–1567 (2015).
25. Sanchez-Alonso, S., Rosenberg, M. D. & Aslin, R. N. Functional connectivity patterns predict naturalistic viewing versus rest across development. *NeuroImage* **229**, 117630 (2021).
26. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G. & Malach, R. Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640 (2004).
27. Dziura, S. L. et al. Effects of social and emotional context on neural activation and synchrony during movie viewing. *Hum. Brain Mapp.* **42**, 6053–6069 (2021).
28. Nguyen, M., Vanderwal, T. & Hasson, U. Shared understanding of narratives is correlated with shared neural responses. *NeuroImage* **184**, 161–170 (2019).
29. Tanner, J. C. et al. Synchronous high-amplitude co-fluctuations of functional brain networks during movie-watching. *Imaging Neurosci.* **1**, 1–21 (2023).
30. Lyons, K., Stevenson, R., Owen, A. & Stojanoski, B. Examining the relationship between measures of autistic traits and neural synchrony during movies in children with and without autism. *NeuroImage: Clin.* **28**, 102477 (2020).
31. Ohad, T. & Yeshurun, Y. Neural synchronization as a function of engagement with the narrative. *NeuroImage* **276**, 120215 (2023).
32. Song, H., Finn, E. S. & Rosenberg, M. D. Neural signatures of attentional engagement during narratives and its consequences for event memory. *Proc. Natl. Acad. Sci.* **118**, e2021905118 (2021).
33. Betzel, R. F., Byrge, L., Esfahlani, F. Z. & Kennedy, D. P. Temporal fluctuations in the brain's modular architecture during movie-watching. *NeuroImage* **213**, 116687 (2020).
34. Bolton, T. A. W., Jochaut, D., Giraud, A.-L. & Van De Ville, D. Dynamic inter-subject functional connectivity reveals moment-to-moment brain network configurations driven by continuous or communication paradigms. *J. Vis. Exp.* **147**, e59083 (2019).
35. Calhoun, V. D., Miller, R., Pearlson, G. & Adalı, T. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron* **84**, 262–274 (2014).
36. Shine, J. M., Koyejo, O. & Poldrack, R. A. Temporal metastates are associated with differential patterns of time-resolved connectivity, network topology, and attention. *Proc. Natl. Acad. Sci.* **113**, 9888–9891 (2016).

37. Vidaurre, D., Smith, S. M. & Woolrich, M. W. Brain network dynamics are hierarchically organized in time. *Proc. Natl. Acad. Sci.* **114**, 12827–12832 (2017).
38. Liégeois, R. et al. Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nat. Commun.* **10**, 2317 (2019).
39. Lurie, D. J. et al. Questions and controversies in the study of time-varying functional connectivity in resting fMRI. *Netw. Neurosci.* **4**, 30–69 (2020).
40. Preti, M. G., Bolton, T. A. & Van De Ville, D. The dynamic functional connectome: state-of-the-art and perspectives. *NeuroImage* **160**, 41–54 (2017).
41. Van Essen, D. C. et al. The WU-Minn human connectome project: an overview. *NeuroImage* **80**, 62–79 (2013).
42. Thomas Yeo, B. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
43. Grall, C. & Finn, E. S. Leveraging the power of media to drive cognition: a media-informed approach to naturalistic neuroscience. *Soc. Cogn. Affect. Neurosci.* **17**, 598–608 (2022).
44. Finn, E. S. et al. Idiosynchrony: from shared responses to individual differences during naturalistic neuroimaging. *NeuroImage* **215**, 116828 (2020).
45. Khosla, M., Ngo, G. H., Jamison, K., Kuceyeski, A. & Sabuncu, M. R. Cortical response to naturalistic stimuli is largely predictable with deep neural networks. *Sci. Adv.* **7**, eabe7547 (2021).
46. Alexander-Bloch, A. F. et al. On testing for spatial correspondence between maps of human brain structure and function. *NeuroImage* **178**, 540–551 (2018).
47. Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V. & Stefanovska, A. Surrogate data for hypothesis testing of physical systems. *Phys. Rep.* **748**, 1–60 (2018).
48. Kauppi, J.-P., Jääskeläinen, I. P., Sams, M. & Tohka, J. Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. *Front. Neuroinform.* **4**, 5 (2010).
49. Fan, L., Su, J., Qin, J., Hu, D. & Shen, H. A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction. *Front. Neurosci.* **14**, 881 (2020).
50. Ding, J.-E., Luo, D., Zilverstand, A., Kulkarni, K. & Liu, F. NeuroTree: Hierarchical functional brain pathway decoding for mental health disorders. Preprint at https://arxiv.org/abs/2502.18786 (2025).
51. Kim, B.-H. & Ye, J. C. Understanding graph isomorphism network for rs-fMRI functional connectivity analysis. *Front. Neurosci.* **14**, 630 (2020).
52. Finn, E. S. et al. Can brain state be manipulated to emphasize individual differences in functional connectivity? *NeuroImage* **160**, 140–151 (2017).
53. Levakov, G., Sporns, O. & Avidan, G. Fine-scale dynamics of functional connectivity in the face-processing network during movie watching. *Cell Rep.* **42**, 112585 (2023).
54. Lally, C., Lavan, N., Garrido, L., Tsantani, M. & McGettigan, C. Neural representations of naturalistic person identities while watching a feature film. *Imaging Neurosci.* **1**, 1–19 (2023).
55. Gershon, R. C. et al. NIH toolbox for assessment of neurological and behavioral function. *Neurology* **80**, S2–S6 (2013).
56. Vu, A. et al. High resolution whole brain diffusion imaging at 7 T for the human connectome project. *NeuroImage* **122**, 318–331 (2015).
57. Cutting, J. E., Brunick, K. L. & Candan, A. Perceiving event dynamics and parsing hollywood films. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 1476–1490 (2012).
58. Glasser, M. F. et al. The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80**, 105–124 (2013).
59. Li, J. et al. Global signal regression strengthens association between resting-state functional connectivity and behavior. *NeuroImage* **196**, 126–141 (2019).
60. Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. *NeuroImage* **82**, 403–415 (2013).
61. Grady, C. L., Rieck, J. R., Nichol, D., Rodrigue, K. M. & Kennedy, K. M. Influence of sample size and analytic approach on stability and interpretation of brain-behavior correlations in task-related fMRI data. *Hum. Brain Mapp.* **42**, 204–219 (2021).
62. Marek, S. et al. Reproducible brain-wide association studies require thousands of individuals. *Nature* **603**, 654–660 (2022).
63. Helmer, M. et al. On the stability of canonical correlation analysis and partial least squares with application to brain-behavior associations. *Commun. Biol.* **7**, 1–15 (2024).
64. Scheinost, D. et al. Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage* **193**, 35–45 (2019).
65. Siegel, J. S. et al. Data quality influences observed links between functional connectivity and behavior. *Cereb. Cortex* **27**, 4492–4502 (2017).
66. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. 34th International Conference on Machine Learning* 3319–3328 (PMLR, 2017).
67. Wang, D. et al. Deep neural network heatmaps capture alzheimer's disease patterns reported in a large meta-analysis of neuroimaging studies. *NeuroImage* **269**, 119929 (2023).
68. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
69. McNamara, Q., De La Vega, A. & Yarkoni, T. Developing a comprehensive framework for multimodal feature extraction. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1567–1574 (Association for Computing Machinery, 2017).
70. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
71. Nishimoto, S. et al. Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* **21**, 1641–1646 (2011).

Acknowledgements

This work was supported in part by the Ann S. Bowers Foundation through the Ann S. Bowers Women's Brain Health Initiative (A.K.), and by NIH grants 1R01MH137166-01A1 (A.K.). Data were provided by the Human Connectome Project (HCP), WU-Minn Consortium (Principal Investigators: D. Van Essen and K. 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.

Author contributions

A.K. and M.S. conceived the experiments and interpreted the results. Y.Z. conducted the experiments, analyzed, and interpreted the results. E.F. interpreted the results. Y.Z. and A.K. wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-026-10411-9.

Correspondence and requests for materials should be addressed to Amy Kuceyeski.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Jasmine Pan. A peer review file is available.