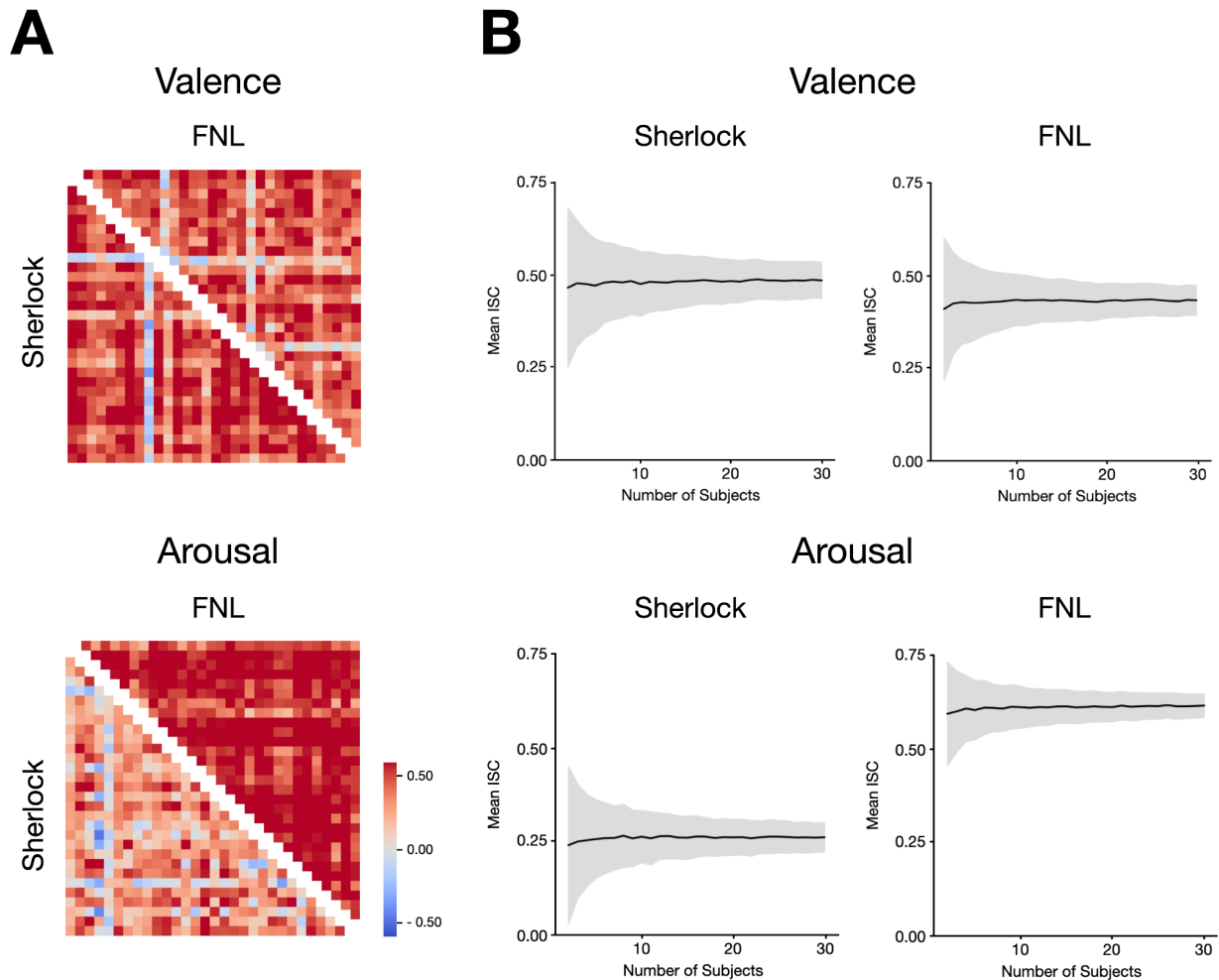
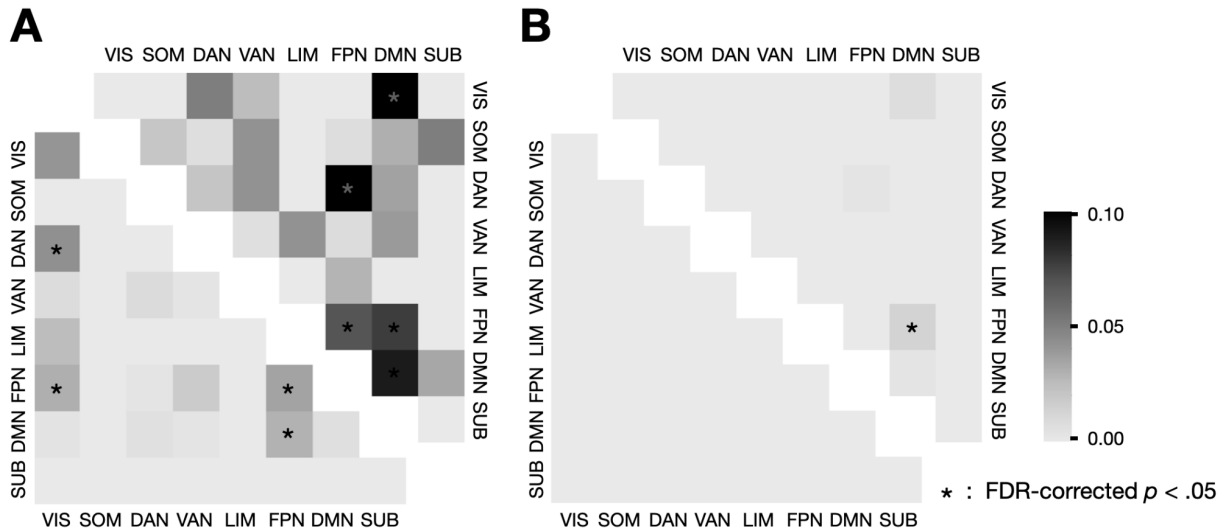


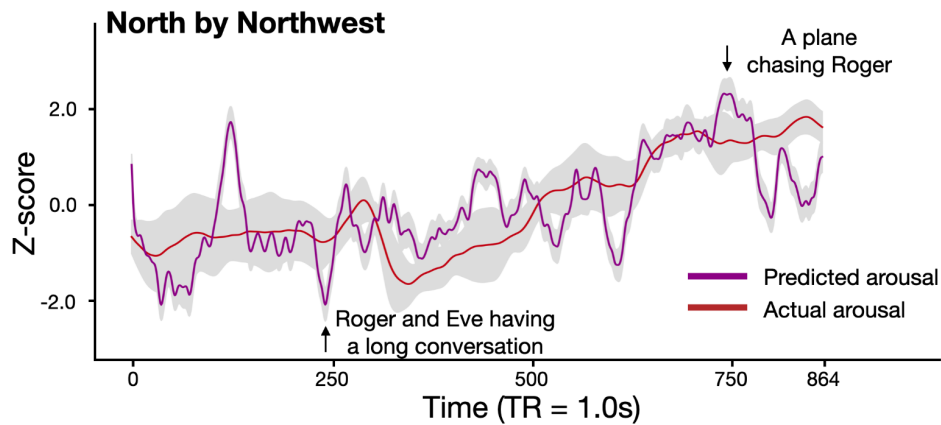
Supplementary Figures



Supplementary Figure 1. A. Heatmaps of pairwise intersubject correlations (ISC) of behavioral ratings in the four conditions. Each cell represents the Pearson's r value between the behavioral rating time courses from a pair of subjects. Red color indicates positive correlation, and blue color indicates negative correlation. Higher r values indicates stronger pairwise synchrony. . The top plot and the bottom plot represent the heatmap of valence and arousal ISC by subject respectively. Within each heatmap, the upper right displays correlations for *Friday Night Lights* and the lower left displays correlations for *Sherlock*. **B. Pairwise behavioral intersubject correlation as a function of the number of subjects.** The x-axis represents the number of subjects. In each condition, the number of subjects, k , increased from 2 to 30. Corresponding to the point k on the x-axis, k subjects were randomly selected from all subjects for 1000 times with replacement, where each time the Fisher's z-transformed group-average ISC was computed. The gray area represents the standard deviation of the distribution of permutations. The group-average rating similarity stabilizes as the number of subjects increases, suggesting it is unlikely for the group-average behavioral rating to increase in precision with more subjects.



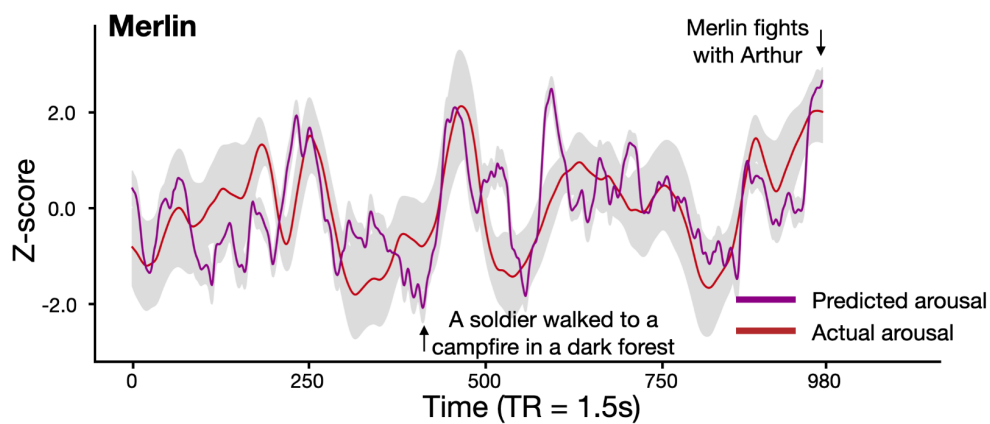
Supplementary Figure 2. Functional anatomy of the overlap network between arousal and engagement. Engagement network acquired from Song et al., 2021. In each figure, the upper triangle represents the network which contains functional connections that positively correlate with both arousal and engagement. The lower triangle represents the network which contains functional connections that negatively correlate with both arousal and engagement. Each cell represents the proportion of selected FCs among all possible FCs between each pair of functional networks. Networks pairs with above-chance selected FCs are indicated with an asterisk (one-tailed t -test, $\text{fdr-corrected } p < .05$). **A. The overlap between the arousal network and engagement network in the *Sherlock* dataset.** The connections between DMN and FPN, between DMN and VIS, between DAN and FPN as well as connections within the DMN and FPN positively predicted both arousal and engagement. The connections between VIS and DAN, between VIS and FPN, between FPN and DMN, as well as connections within the FPN negatively predicted both arousal and engagement (FDR-corrected $P < .05$). **B. The overlap between the *Sherlock-Friday Night Lights* arousal network and *Sherlock-Paranoia* engagement network.** Only the connection between DMN and FPN positively predicted both arousal and engagement in this cross-dataset manner.



The most arousing



The least arousing



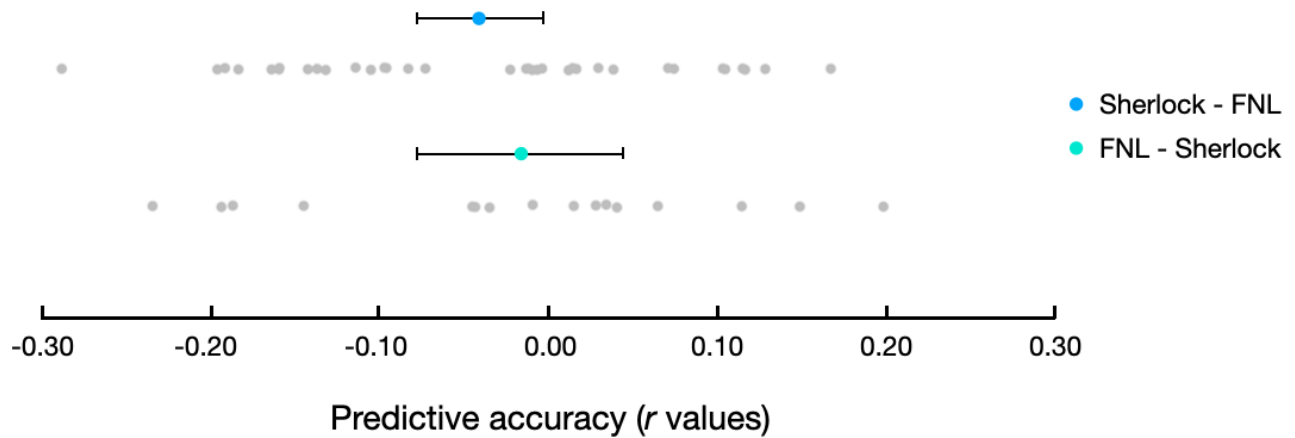
The most arousing



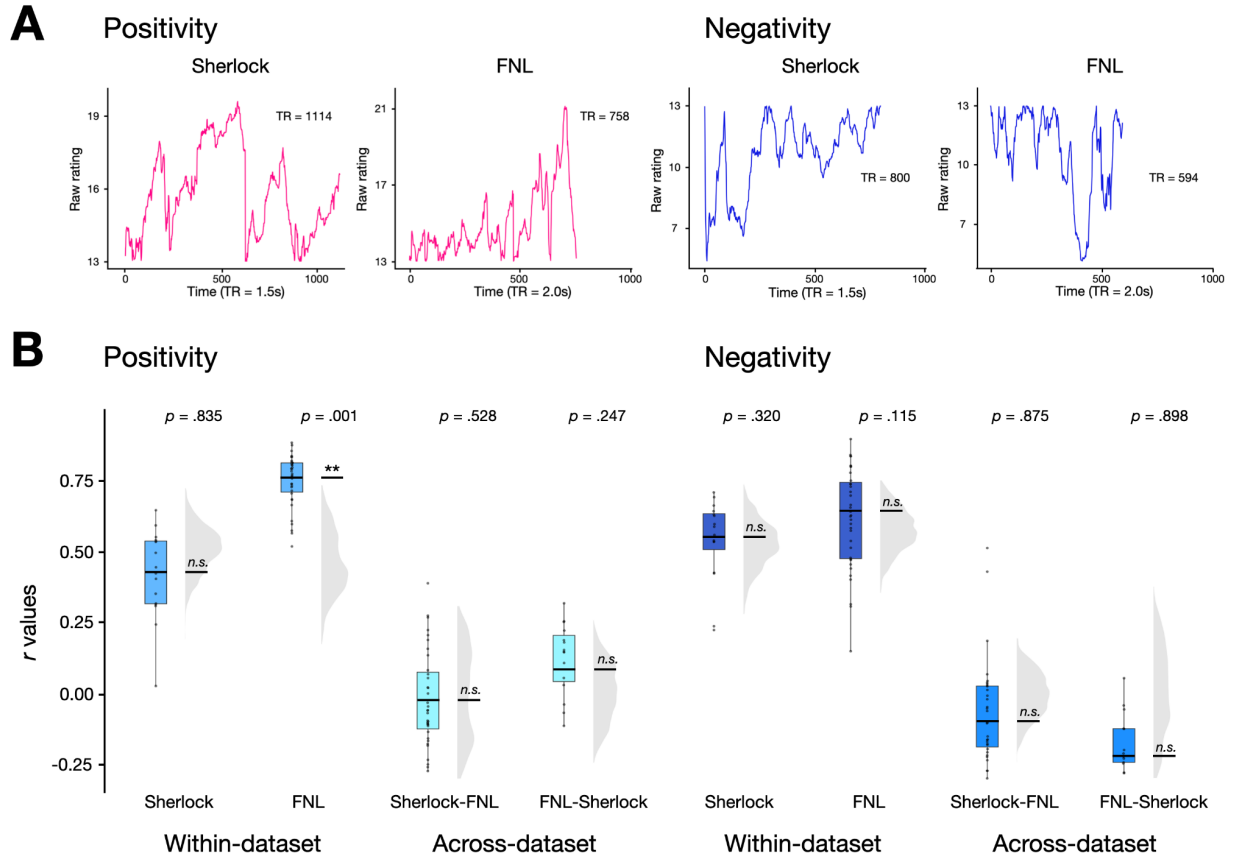
The least arousing



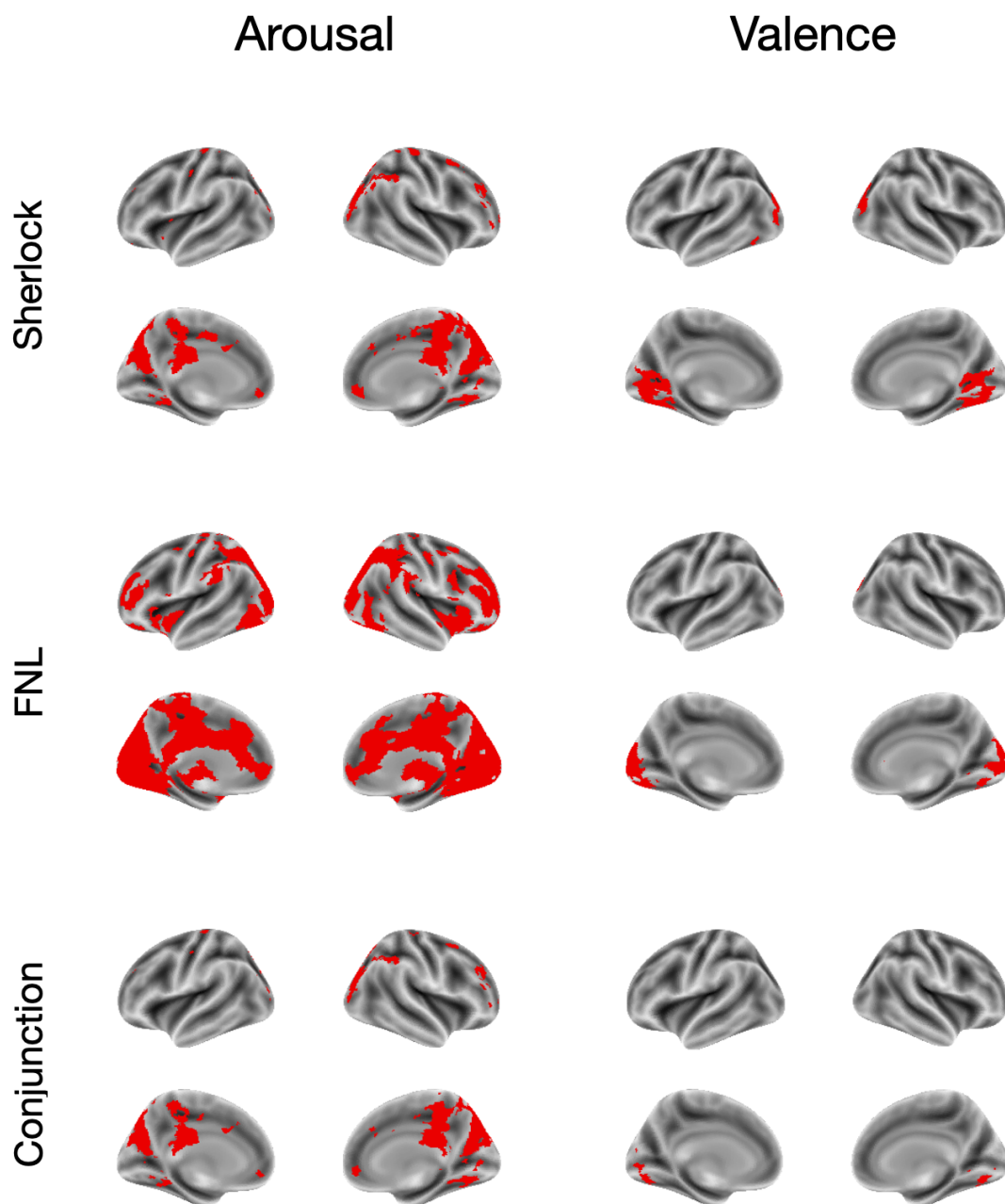
Supplementary Figure 3. Arousal CPM predicts group-average arousal rating in two additional datasets. Model predicted time courses across participants watching the same movie was averaged. This averaged predicted arousal time course significantly correlated with the actual group-average arousal rating. Model-predicted arousal time courses also corresponded with the plot of each movie. In *North by Northwest*, the most arousing moment predicted by the model occurred in the scene when the protagonist was being chased by a plane, while the least arousing moment occurred during a long conversation between characters; In *Merlin*, the most arousing moment predicted by the model occurred when the two main protagonists had a brawl in a tavern, while the least arousing moment occurred when a nameless soldier walked towards a campfire.



Supplementary Figure 4. Equivalence tests were performed to assess whether the across-dataset predictive accuracies in predicting valence fell within a predefined range around zero, which would suggest that they were not only statistically non-significant but also practically insignificant. We defined an equivalence interval of $[-.100, .100]$, with the bounds determined based on a small effect size of $r = .100$. Each datapoint in the box plot represents the predictive accuracy of each round of cross-validation. The black horizontal lines show the 95% percent CI of the mean r value. The equivalence test was significant for both across dataset accuracies (*Sherlock* - *Friday Night Lights*: $p = .368$, *Friday Night Lights* - *Sherlock*: $p = .306$), indicating that the observed predictive accuracies were statistically indistinguishable from zero within the defined bound.

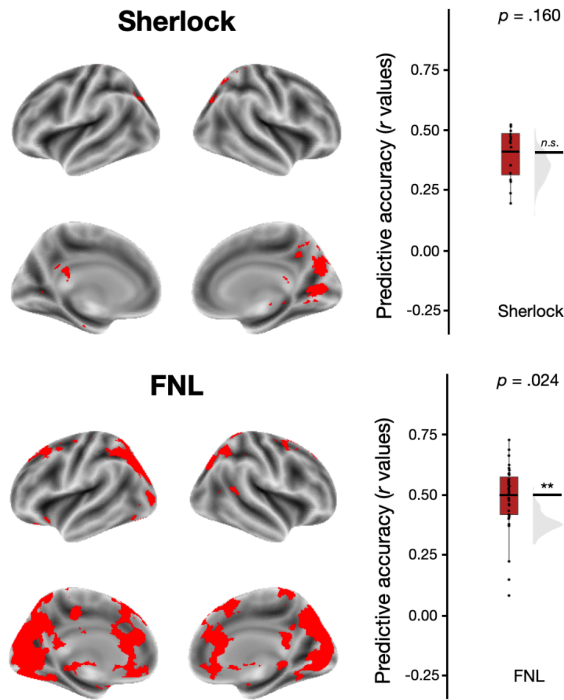


Supplementary Figure 5. A. Participants' subjective affective experience of positivity (left) and negativity (right) fluctuates over time during naturalistic movie watching. **B.** Dynamic functional connectivity does not predict subjective feelings of positivity and negativity. CPM performance in predicting positivity and valence for within-dataset (the left panel) and between-dataset (the right panel). The y-axis represents the predictive accuracy, as measured by Pearson's correlation between the model predicted time course and the observed group-average time course. Each datapoint in the box plot represents the predictive accuracy of each round of cross-validation. The black horizontal lines show the fisher-z transformed mean r value. The gray half-violin plots show the null distribution of 1000 permutations, generated by phase-randomizing the observed group-average before training and testing the models. *: $p < 0.05$, **: $p < 0.01$, n.s.: $p > 0.05$, as assessed by comparing the empirical mean predictive accuracy against the null-distribution.

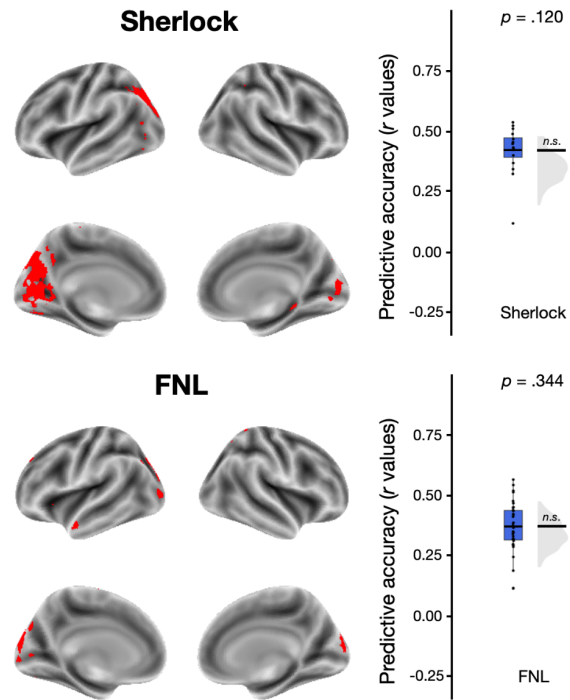


Supplementary Figure 6. Univariate parametric maps of arousal (left) and valence (right) in the *Friday Night Lights* and *Sherlock* datasets, as well as the conjunction between the two datasets. We built generalized linear models (GLM) for *Friday Night Lights* and *Sherlock* separately to identify the voxels whose BOLD activity were associated with valence or arousal in each movie. A MS/CN test was performed as a conjunction analysis to identify the voxels whose activations were consistently significantly related to valence or arousal across datasets (TFCE-corrected $p < .01$). We observed clusters of voxels that were more positively associated with arousal in the cingulate gyrus, precuneus cortex, and insular. However, few voxels exhibit a significant positive association with valence.

A. arousal



B. valence



Supplementary Figure 7. Multivariate patterns of brain activities do not predict the subjective feelings of valence and arousal within-dataset. Multivariate-pattern based predictive model performance in predicting arousal (**A**) and valence (**B**) within-dataset. The brain map in each sub-plot indicates the voxels selected in the predictive modeling. The y-axis represents the predictive accuracy, as measured by Pearson's correlation between the model predicted time course and the observed group-average time course. Each datapoint in the box plot represents the predictive accuracy of each round of cross-validation. The black horizontal lines show the Fisher-z transformed mean r value. The gray half-violin plots show the null distribution of 1000 permutations, generated by phase-randomizing the observed group-average before training and testing the models. * $p < 0.05$, ** $p < 0.01$, n.s.: $p > 0.05$, as assessed by comparing the empirical mean predictive accuracy against the null distribution.