Multi-view manifold learning of human brain state trajectories

Erica L. Busch[1], Jessie Huang[2], Andrew Benz[3], Tom Wallenstein[2], Guillaume Lajoie[7,8], Guy Wolf[7,8], Smita Krishnaswamy[2,4,5,6,†*], and Nicholas B. Turk-Browne[1,6,†]

[1] Department of Psychology, Yale University
[2] Department of Computer Science, Yale University
[3] Department of Mathematics, Yale University
[4] Department of Genetics, Yale University
[5] Program in Applied Mathematics, Yale University
[6] Wu Tsai Institute, Yale University
[7] Department of Mathematics and Statistics, Université de Montréal
[8]Mila – Quebec AI Institute
[†]Equal contribution
[*]Corresponding author: `smita.krishnaswamy@yale.edu`

1 Abstract

Brain activity as measured with functional magnetic resonance imaging (fMRI) gives the illusion of intractably high dimensionality, rife with collection and biological noise. Nonlinear dimensionality reductions like UMAP and tSNE have proven useful for high-throughput biomedical data, but have not been extensively used in fMRI, which reflects the redundancy and co-modulation of neural population activity. Here we extend the manifold-geometry preserving method PHATE for use with timeseries data, such as produced by fMRI, in a method we call temporal PHATE (T-PHATE). In addition to recovering the lower intrinsic dimensionality of fMRI data, T-PHATE exploits fMRI data's autocorrelative structure to faithfully denoise the signal and learn brain activation manifolds. We empirically validate T-PHATE on three fMRI tasks and show that T-PHATE improves visualization fidelity, stimulus feature classification, and neural event segmentation. These notable improvements over previous cutting-edge approaches to understanding the nature and dynamics of cognition from fMRI bode well for potential applications of T-PHATE to other high-dimensional datasets of temporally-diffuse processes.

2 Introduction

As we walk through the world, the human brain performs innumerable operations to process, represent, and integrate information. How does a system composed of limited computational units represent so much while handling a constant barrage of incoming information? How does the way the brain processes this information support our psychological experience? One possible explanation is that individual neural units operate in tandem to encode and update information in population codes, which exponentially expands the distinct ways information can be represented by a fixed number of units. Neural population codes better predict behavior than single neurons, particularly in how representations change over time in response to new information (Averbeck et al., 2006). Olfactory population codes represent odors dynamically, affording rapid classification and learning (Laurent, 2002). Motor population codes underlie the ability to plan, prepare, and execute movements (Churchland et al., 2010). Face population codes represent face identity in both static (Chang and Tsao, 2017) and invariant representations (Freiwald and Tsao, 2010). Recent studies have shown that although neural population codes exist in high ambient dimensions (Jazayeri and Ostojic, 2021), these dimensions are not orthogonal: they are redundant (Hennig et al., 2018) and from them emerge dominant latent signals that code for behavior and information (Cunningham and Yu, 2014; Nieh et al., 2021).

Principles of neural population codes were defined using direct neural recordings in nonhuman primates but can be extended to noninvasive, indirect measurements of brain activity using functional magnetic resonance imaging (fMRI). Multivariate pattern analysis (Cox and Savoy, 2003; Norman et al., 2006) of fMRI activation is one such extension that has yielded valuable insights into the structure and content of cognitive representations, including low-level (Kamitani and Tong, 2005) and high-level (Cox and Savoy, 2003; Haxby, 2001; Haynes and Rees, 2006) sensory stimuli, as well as higher-order cognitive processes like memory (Polyn et al., 2005), emotion (Peelen et al., 2010), narrative comprehension (Yeshurun et al., 2017), and theory of mind (Davatzikos et al., 2005). These insights have largely come from group-level analyses, requiring aggregation over multiple subjects and timepoints to draw conclusions because of the high levels of spatio-temporal noise inherent to fMRI. Like other types of high-throughput biomedical data, fMRI noise is pervasive at multiple levels, from subject movement and scanning conditions to measured blood-oxygenation level-dependent (BOLD) signal drift and physiological confounds (Birn et al., 2008). fMRI has low temporal resolution relative to the cognitive processes many studies attempt to measure, with typical brain volume acquisition time around 1-2 seconds (Turk-Browne, 2013). Further, the BOLD signal is a slow, indirect measurement of neuronal activity, peaking approximately 4-5 seconds after stimulation before returning to baseline. Finally, complex cognitive processes like memory (Hasson et al., 2015), narrative comprehension (Honey et al., 2012; Xu et al., 2005; Yeshurun et al., 2017), or learning are dynamic: they do not exist at a singular timepoint but unfurl over longer periods of time (Hasson et al., 2008). Such complex yet fundamental processes require integration across modalities and time. The brain likely performs computations on these representations at varying timescales to support conscious experience (Barack and

Krakauer, 2021; Golesorkhi et al., 2021; Himberger et al., 2018; Ito et al., 2020).

In sum, many brain representations of interest are coded in high-dimensional patterns of activation (Chang and Tsao, 2017; Churchland et al., 2010; Cox and Savoy, 2003; Laurent, 2002; Norman et al., 2006), which can be characterized by low-dimensional latent signals (Averbeck et al., 2006; Cunningham and Yu, 2014; Gao et al., 2020). fMRI affords unique insight into the healthy, behaving human brain, but the data are noisy, sampled slowly, and blurred in time, giving rise to high signal autocorrelation (Olszowy et al., 2019; Turk-Browne, 2013). Further, complex cognitive processes unfurl over time at differing intrinsic neural timescales along the brain's processing hierarchy (Golesorkhi et al., 2021; Ito et al., 2020). Addressing these issues requires consideration of the various ways in which time and dimensionality interact with the measured BOLD signal with the goal of optimally characterizing fMRI activity at the single-subject level.

We introduce a nonlinear, temporal manifold learning approach designed for the high noise and temporal properties of both fMRI data and complex cognitive processing. Most previous studies exploring fMRI activity with dimensionality reduction have relied upon linear heuristics to project brain activation patterns into a low-dimensional space (Allen et al., 2012; Shine et al., 2016; Shine et al., 2019). Nonlinear dimensionality reduction algorithms have been applied in past studies to characterize the geometry and underlying temporal dynamics in neural recordings (Gao et al., 2020; Jazayeri and Ostojic, 2021; Nieh et al., 2021) by integrating local similarities among data points into a global representation (Cunningham and Yu, 2014; Moon et al., 2019; Roweis and Saul, 2000; Salhov et al., 2015; Tenenbaum et al., 2000), but are underutilized in fMRI (Gao et al., 2020). We build upon the manifold learning algorithm PHATE (Moon et al., 2019) designed for high-throughput, high-noise biomedical data, with a second, explicit model of the temporal properties of BOLD activity. This second view is learned from the data to capture the temporal autocorrelation of the BOLD signal in different brain regions and the dynamics specific to a given stimulus (Fig. 1)A). Our algorithm, temporal PHATE or T-PHATE, better captures signal in simulated data under a range of noise conditions than other well-known manifold learning methods, including principal components analysis (PCA), uniform manifold approximation (UMAP) (Becht et al., 2019), and PHATE (Moon et al., 2019), using a benchmark of manifold preservation called DeMAP (Moon et al., 2019) (Fig. 1)). Unlike prior studies (Gao et al., 2020), we are interested in learning a manifold that captures temporal dynamics within a given naturalistic task that could later be extended to new subjects (Huang et al., 2021) or predict future brain states. Thus, we test T-PHATE on two movie-viewing fMRI datasets and find that it both denoises the data and affords enhanced access to brain-state trajectories than either voxel resolution data or a variety of other dimensionality reduction approaches. By mitigating noise and voxel covariance, this subspace yields clearer access to regional dynamics of brain signals, which we can then relate to time-dependent psychological phenomena. In all, we show that information about cortical dynamics during naturalistic stimuli lies in a low-dimensional latent space that can be best modeled with a nonlinear, temporal manifold.

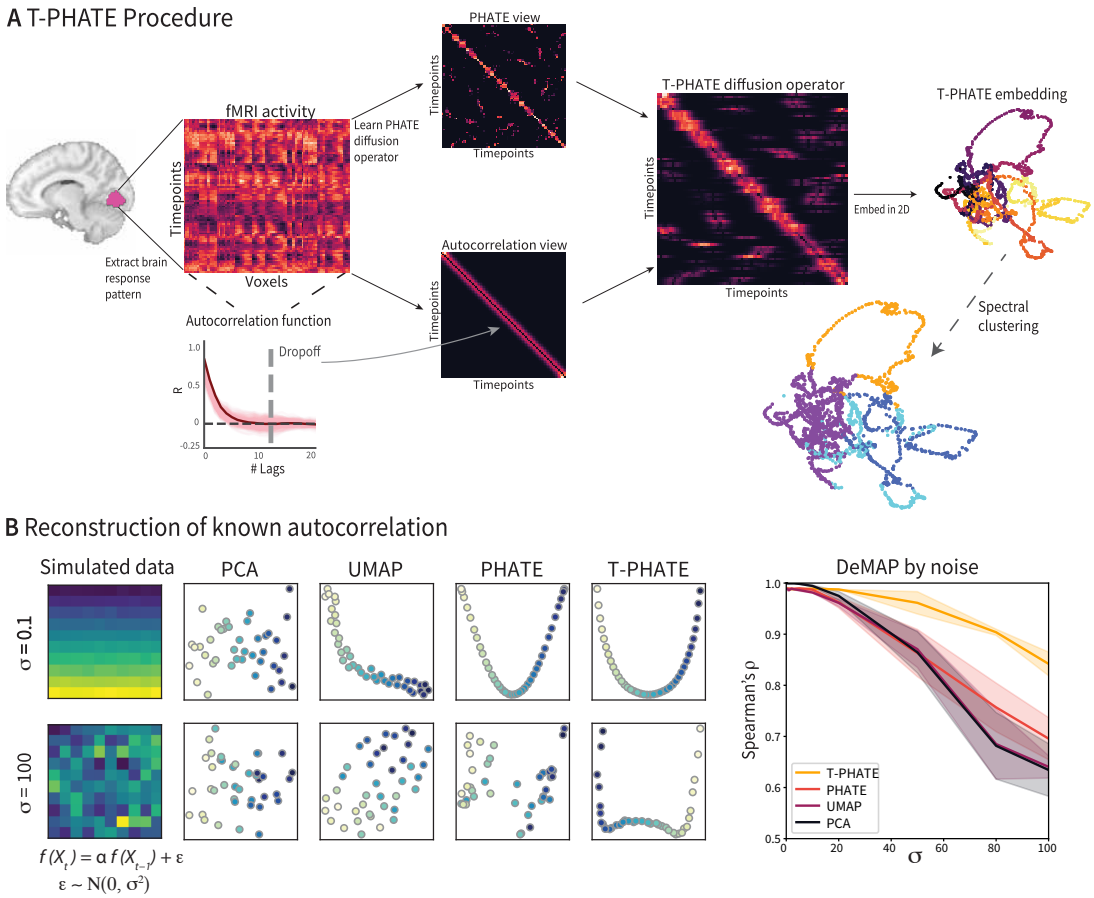

Figure 1: **A.** Temporal PHATE (T-PHATE) procedure. A multivariate pattern of fMRI activation over time (timepoints by voxels) is extracted from a region of interest (ROI). PHATE is then run to learn a PHATE-based affinity matrix between timepoints. The autocorrelation function is also estimated for the data separately to make an autocorrelation-based affinity matrix. These two views are combined into the T-PHATE diffusion operator then embedded into lower dimensions. **B.** We tested this approach on simulated data with a known autocorrelative function and varied the error added to the signal. T-PHATE both visually better reconstructs the autocorrelative signal from the noise and quantitatively better preserves manifold distances, compared with PCA, UMAP, and PHATE.

3 Results

3.1 Evaluating manifold quality

fMRI is a powerful tool for studying the neural bases of higher-order cognitive abilities. fMRI data are highly noisy in both space and time, with the measured BOLD signal canonically peaking 4-5 s after stimulus onset before slowly returning to baseline. Naturalistic, complex stimuli like movies are increasingly used to probe real-world cognition and reasoning in the brain. These stimuli are dynamic, with conversations and plot lines playing out over different timescales, which cause the autocorrelation of brain responses to extend beyond the curve of the BOLD signal and vary according to the temporal integrative

properties of individual brain regions. Thus, we designed T-PHATE as a variant of the PHATE algorithm that combines the robust manifold-geometry preserving properties of PHATE with an explicit model of the data's autocorrelation in a dual-diffusion operator. We compared manifolds learned with T-PHATE to those from PCA, PHATE, and UMAP.

We validated the quality of manifolds learned from data with a simulated high autocorrelative signal using Denoised Manifold Affinity Preservation (DeMAP)(Moon et al., 2019). DeMAP takes pristine, noiseless data (here, a simulated multivariate timeseries where $f(X_t) = \alpha f(X_{t-1})$) and computes the Spearman correlation between the geodesic distances of the noiseless data and the Euclidean distances on the embeddings learned from noisy data. Noisy data were generated by adding an error term ϵ to the pristine data, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We tested robustness to noise by varying $\sigma = (0, 100)$. Higher DeMAP scores suggest that an embedding method performs effective denoising and preserves the geometric relationships in the original data. DeMAP results for PCA, UMAP, PHATE, and T-PHATE are shown in Fig. 1B. At low levels of noise, all methods achieve high DeMAP scores, though PCA visualizations do not appear to reflect much structure. With increasing noise, T-PHATE outperforms the other methods at denoising the simulated data, providing visualizations that most closely resemble that of lower noise conditions.

Having validated that T-PHATE can learn meaningful manifolds from simulated data even under noisy conditions, we next applied T-PHATE to fMRI data. First, we embedded the movie-viewing data from the *Sherlock* and *StudyForrest* datasets with PCA, UMAP, PHATE, and T-PHATE in two dimensions to visually inspect the properties of the data highlighted by the manifold. Visually, T-PHATE embeddings denoise the timeseries data and capture stimulus-related trajectory structure better than the other methods, as shown in all brain regions for a sample subject (Fig. 2A) as colored by sequential timepoint. PCA visualizations show no apparent structure or clustering, whereas UMAP shows at least slight clustering of nearby timepoints in high visual (HV) and early auditory (EA) cortices but often creates small, shattered clusters with local structure placed through the latent space. PHATE visually yields slight improvements over UMAP, notably in HV and EA. Clusters in these plots appear larger and less disjointed, connected by tendrils that flow from a hub-like center space where the majority of clusters exist in the embedding. T-PHATE, however, reveals remarkable trajectories through the latent space that are clearly reflective of temporal structure but also shows a hub in the center of the space. T-PHATE manifolds also better reflect non-temporal attributes of the movie, as shown by classification of stimulus features at given timepoints (Fig. 2B). We ran a support vector classification to predict whether music was playing at a given timepoint from the embeddings of neural data (and as a baseline, the original voxel-resolution data). Significance of these predictions was tested by shifting the labels in time with respect to the brain data. Results are presented as the z-statistic of the true prediction accuracy between brain data and movie labels normalized to the mean and standard deviation of the null distribution from the shifted versions.

The clarity of the dynamic structure recovered by T-PHATE prompted two follow-up questions. First, does the performance of T-PHATE depend on the use of autocorrelation to represent temporal structure during manifold learning? To answer this question, we

tested two alternatives to T-PHATE, PHATE+Time and Smooth PHATE. PHATE+Time adds the TR label for each timepoint as an additional feature vector to the voxel data before embedding with PHATE. PHATE+Time manifolds show the clearest temporal trajectory through the embedding space (Fig. 3A), unraveling the brain data in time due to the disproportionate weighting of time in the input data, with less of the hub structure shown by T-PHATE. Smooth PHATE performs temporal smoothing along each voxel with a double gamma function (the canonical hemodynamic response function), which is a common preprocessing step for fMRI data generally, before embedding with PHATE. Smooth PHATE yields comparable structure to PHATE or UMAP, with shattered trajectories and mild clustering.

Second, are the benefits of T-PHATE for stimulus feature classification specific to fMRI tasks that are intrinsically linked to time, such as movie-watching? To answer this question, we tested T-PHATE on an object category localizer task from *StudyForrest* with no temporal structure, during which static images of faces, bodies, small objects, houses, scenes, and scrambled images were presented in a random order. We embedded these data with T-PHATE, PHATE+Time, and Smooth PHATE for early visual (EV) and high visual (HV) ROIs. Because of the lack of temporal structure, a meaningful manifold of fMRI activity in visual regions during this task would show a time-independent clustering of categories. T-PHATE manifolds show the best clustering of stimulus categories in the embedding space (Fig. 3B), compared with PHATE+Time and Smooth PHATE. Smooth PHATE showed some clustering but less pronounced, whereas PHATE+Time showed minimal clustering by category but instead still retained temporal structure. We quantified the ability to distinguish object categories in the latent space using a within-subject support vector classifier with leave-one-run-out cross-validation (6-way classification; chance = 1/6) (Fig. 3C). In EV, classification accuracy on T-PHATE embeddings surpassed all other dimensionality reduction methods (T-PHATE M=0.51, CI=[0.42, 0.6]; voxel M=0.46, CI=[0.38, 0.55]; PCA M=0.43, CI=[0.34, 0.52]; UMAP M=0.4, CI=[0.39, 0.42]; PHATE M=0.39, CI=[0.37, 0.42]; PHATE+Time M=0.38, CI=[0.37, 0.39]; Smooth PHATE M=0.42, CI=[0.38, 0.47]). In HV, T-PHATE again outperformed the other dimensionality reduction methods (T-PHATE M=0.55, CI=[0.47, 0.62]; voxel M=0.53, CI=[0.46, 0.61]; PCA M=0.48, CI=[0.46, 0.51]; UMAP M=0.42, CI=[0.41, 0.44]; PHATE M=0.43, CI=[0.42, 0.45]; PHATE+Time M=0.38, CI=[0.37, 0.39]; Smooth PHATE M=0.48, CI=[0.41, 0.56]). This analysis showed that T-PHATE captures structure in brain activity via autocorrelative denoising, whereas adding time explicitly into the manifold does little for capturing the data geometry aside from unrolling the signal in time. Temporal smoothing provides an advantage over time-agnostic dimensionality reduction (PCA, PHATE, UMAP), but less of an advantage than the autocorrelation kernel in both the dynamic movie task and the non-temporal localizer task.

3.2 Event segmentation

We designed T-PHATE to incorporate temporal dynamics (namely autocorrelation) into a manifold learning algorithm for two reasons. First, as shown above, this helps denoise fMRI data given its spatiotemporal noise and improve subsequent analyses such as fea-

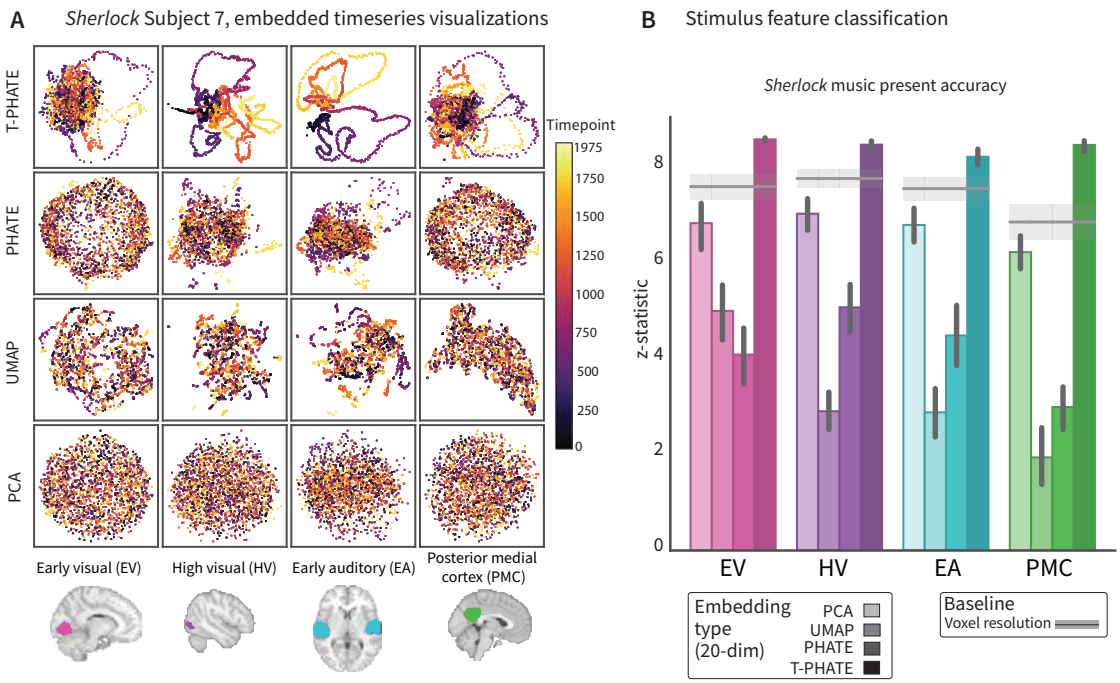

Figure 2: **A**. Manifold embeddings using PCA, UMAP, PHATE, and T-PHATE for the *Sherlock* dataset in an example subject (7) and 4 regions of interest. Points are colored by time index. **B**. Support vector classification accuracy predicting the presence of music in the show at the corresponding timepoint. Results are presented as a z-statistic of the true accuracy versus a null distribution of accuracies from time-shifting the timeseries (which preserves their autocorrelation). Extended SVC results included in Supplemental Fig. 2

ture classification. Second, this may help recover cognitive signals that are represented over time in fMRI data. Many cognitive processes operate on longer timescales, including our ability to segment experience into events (Radvansky and Zacks, 2017; Zacks et al., 2007; Zacks et al., 2010). That is, we do not perceive the world as transient from moment to moment, nor as amorphous and fully undifferentiated, but rather as a series of coherent and often hierarchically structured periods of time (Kurby and Zacks, 2008; Zacks, Tversky, et al., 2001). For example, consider the steps involved in taking a flight — traveling to the airport, passing through security, boarding, flying, deplaning, an transiting to the destination. These mental events are reflected in stable states of the brain, which can be captured with a Hidden Markov Model (HMM) to identify the boundaries between events (Baldassano et al., 2017; Baldassano et al., 2018; Lee et al., 2021). We hypothesized that an HMM would be better able to perform this event segmentation after embedding fMRI data into a manifold learned with T-PHATE. This would suggest that T-PHATE can increase sensitivity to the neural dynamics associated with cognitive processing of the stimulus.

We started by learning the optimal *number* of events experienced by each ROI using a leave-one-subject-out cross-validation procedure on the voxel resolution data (Fig. 4A). We find a pattern of results consistent across datasets where the number of events

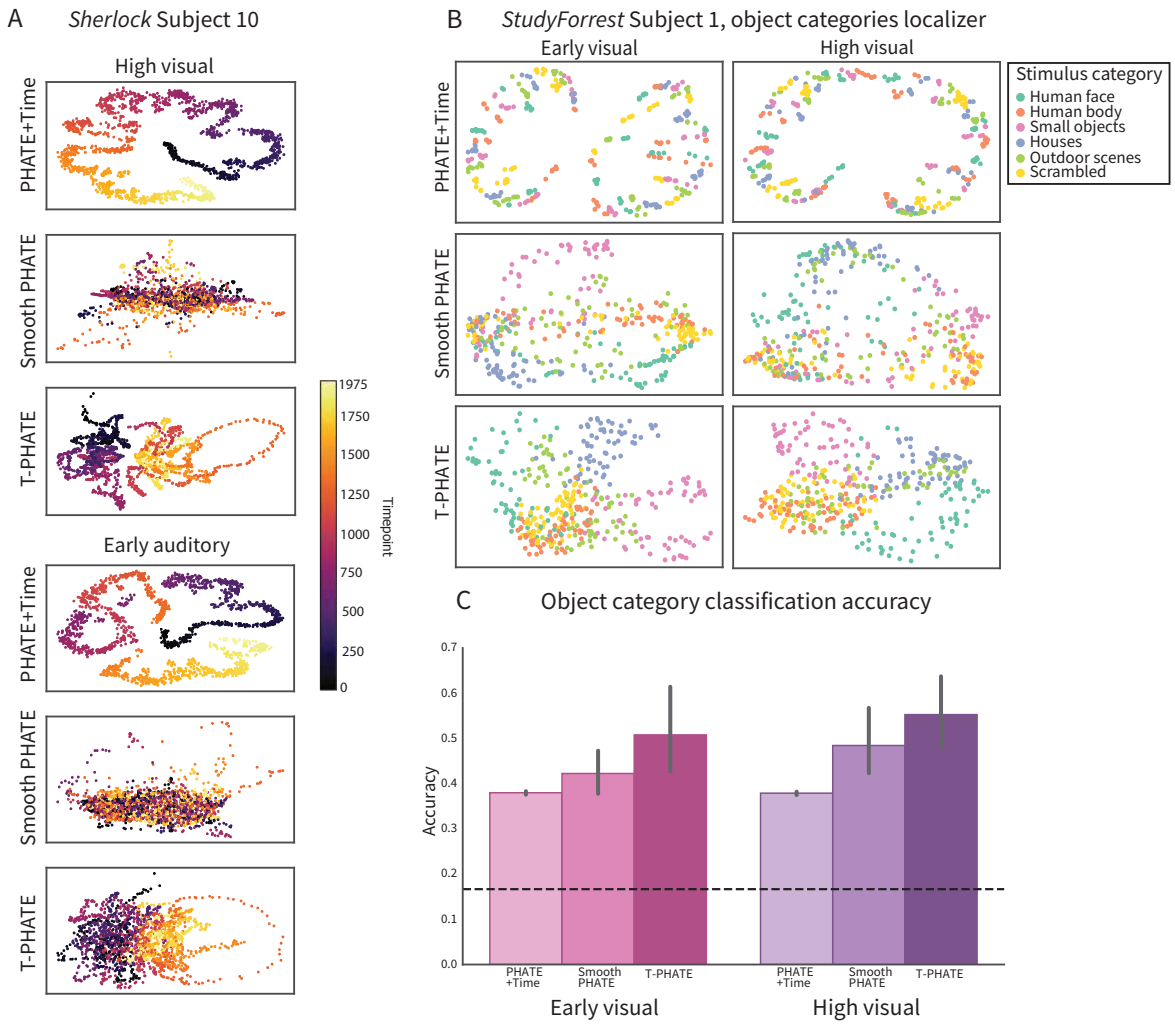

Figure 3: **A**. Manifold embedding of *Sherlock* data in high visual (HV) and early auditory (EA) ROIs with T-PHATE, PHATE+Time, and Smooth PHATE, colored by time. **B**. *StudyForrest* localizer data embeddings using T-PHATE, PHATE+Time, and Smooth PHATE in early visual (EV) and high visual (HV). T-PHATE reveals the best clustering by stimulus category, showing that the manifold successfully denoises data to learn a more accurate manifold. **C**. T-PHATE manifolds also afford better stimulus category classification accuracy. Bars represent average accuracy across subjects; error bars represent 95% CIs of the mean from bootstrap resampling; dashed line represents chance.

K represented by a brain region is higher for early sensory cortices (EV) than late sensory cortices (HV) and integrative cortices (PMC) (values presented as mean number of events $\pm$ s.d. across subjects): *Sherlock*: EV $= 53 \pm 12$, HV $= 21 \pm 7$, AE $= 18 \pm 5$, PMC $= 22 \pm 6$; *StudyForrest*: EV $= 111 \pm 43$, HV $= 69 \pm 17$, AE $= 74 \pm 18$, PMC $= 96 \pm 24$. After fixing the K parameter for each ROI based on the voxel data, we learned the number of manifold dimensions that would best reveal temporal structure for each region, subject, and method with the same cross-validation procedure. Number of dimensions is

low (between 3-6 dimensions) with small variance across subjects and method. This indicates substantial redundancy, covariance, and noise among voxels in the ROIs, obscuring meaningful, dominant signals (Supplemental Fig. 1).

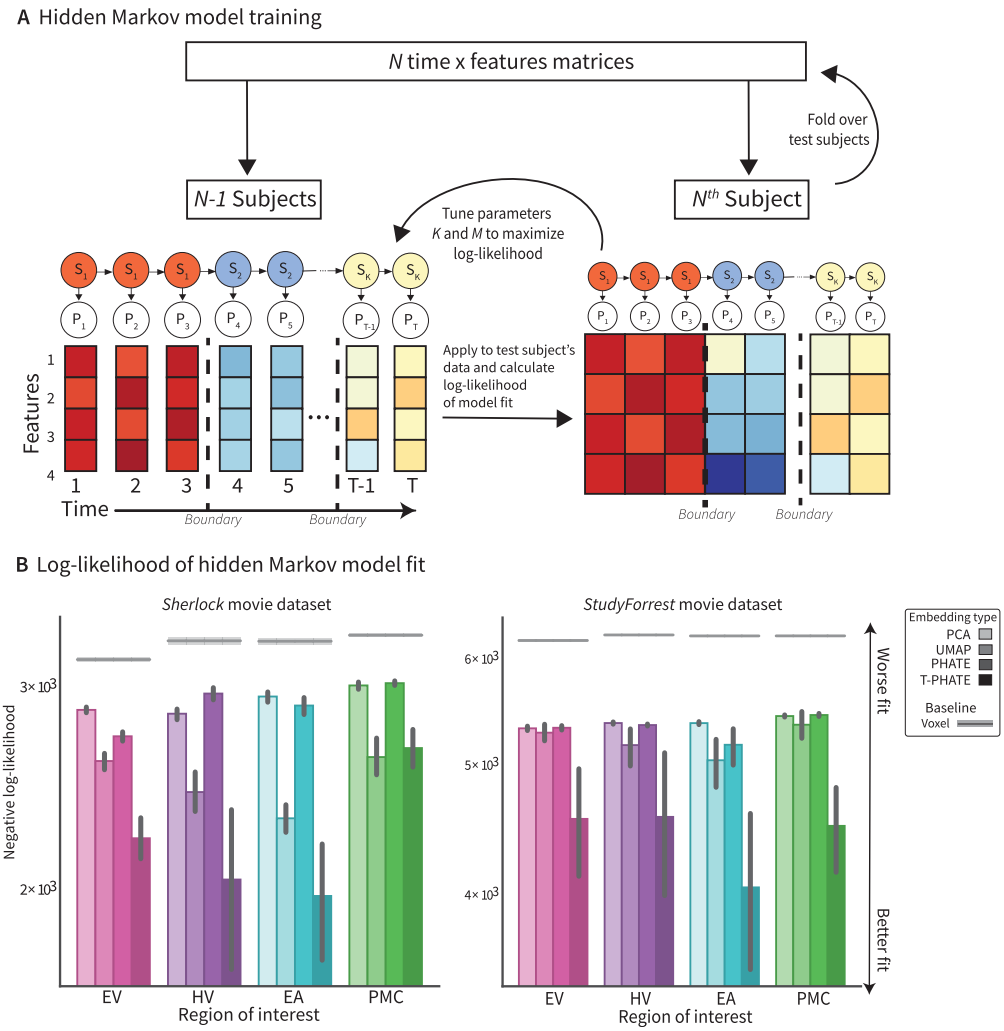

Figure 4: **A.** The HMM is trained on the average data for $N-1$ subjects to find K stable patterns of voxel activation, where the voxel patterns P for each timepoint belong to a given hidden state S. Model fit is evaluated on a test subject's data then repeated with leave-one-subject-out cross-validation. Event boundaries (dashed lines) are identified by the model as timepoints at a transition between hidden states. **B.** Log-likelihood of model fit is tested on a given subject using a HMM with hyper-parameters K and M tuned via leave-one-subject-out cross validation. Model fit here is presented as the negative log-likelihood, where a better model fit results in a *lower* negative log-likelihood. Different dimensionality-reduced approaches are shown as bars and the voxel resolution benchmark is shown as line with shading representing mean and 95% confidence interval.

A goal of T-PHATE is to recover dynamic cognitive signals that may have been previously obscured by the noise of redundancy and noise of the data. If dynamical structure during movie watching is unveiled in the T-PHATE latent space, we would see better success of the HMM on data embedded with T-PHATE than to the other methods or the

voxel resolution data. This can be framed as: (1) how well does the model fit the data and (2) how well do boundaries learned by the model capture structure of the neural data. For each subject, ROI, and embedding method, we fit a HMM with K hidden states on a M-dimensional embedding (where hyper-parameters K and M were cross-validated at the subject level for ROIs and methods, respectively). Model fit is quantified after hyper-parameter tuning using log-likelihood. T-PHATE shows significantly better model fit across both the *Sherlock* and *StudyForrest* datasets in almost all regions (except PMC for *Sherlock*), visualized as negative log-likelihood, where closer to zero means a better model fit (Fig. 4B). This indicates a higher probability of the HMM's learned segmentation fitting the test data. To validate whether the boundaries identified actual differences in the data, we then applied the boundaries learned during the HMM fitting procedure to the data and asked whether the neural data from pairs of timepoints that fall within the same event are more similar than pairs that fall across different events (equating temporal distance) (Fig. 5A). T-PHATE embeddings result in significantly better within-versus-between event boundary distinctions in almost all regions (except EV for *StudyForrest*, where T-PHATE is marginally greater than PCA) (Fig. 5B). In the *Sherlock* dataset (and 2/4 ROIs of the *StudyForrest* data), T-PHATE at least doubles the performance of the second-best embedding type. Notably, all dimensionality reduction outperform the original voxel resolution data. This indicates that, though there is an inherent advantage in revealing dominant signal structure in data in lower dimensions, the temporal diffusion step of T-PHATE captures dynamic signals beyond a temporally-agnostic manifold.

In the previous analysis, event boundaries are learned and applied from a HMM fitted within-subject. To ensure the high performance in this analysis was not driven by overfitting from training/testing the boundaries in the same subject, we performed cross-validation between subjects by applying the event boundaries identified for $N-1$ subjects to subject N's data individually. In other words, for each of N subjects, we ran the analysis $N-1$ times using each of the other subject's boundaries. Again, we found the strongest effect for T-PHATE relative to the other embedding types (Fig. 6), albeit weaker than when the event boundaries were learned within-subject, as expected. Interestingly, the voxel resolution data outperformed each of the methods other than T-PHATE in a few cases (voxel $>$ others in: *Sherlock*: HV UMAP and PHATE, EA UMAP, PMC PHATE; *StudyForrest*: EV PHATE, HV PCA and PHATE, PMC PCA, PHATE and UMAP). This was not seen in the within-subject version, suggesting that some overfitting to subject-specific latent spaces occurred for the other dimensionality reduction methods.

Identifying neural event boundaries with HMM is a data-driven approach to understanding how the brain makes sense of continuous stimulation. However, the identified boundaries have no ground truth. Another way to assess whether learned neural boundaries are meaningful is to use a separate behavioral task where individuals watch the same movie and indicate where they believe something changed and there was a shift to a new event. Ample literature supports the use of this type of parsing task for studying human event segmentation (Radvansky and Zacks, 2017; Speer et al., 2007; Zacks et al., 2007; Zacks et al., 2010), allowing us to treat human-rated boundaries as a ground truth. A separate cohort of human subjects watched the *Sherlock* movie and were asked to indicate where they thought an event boundary occurred (Fig. 7A). We applied these

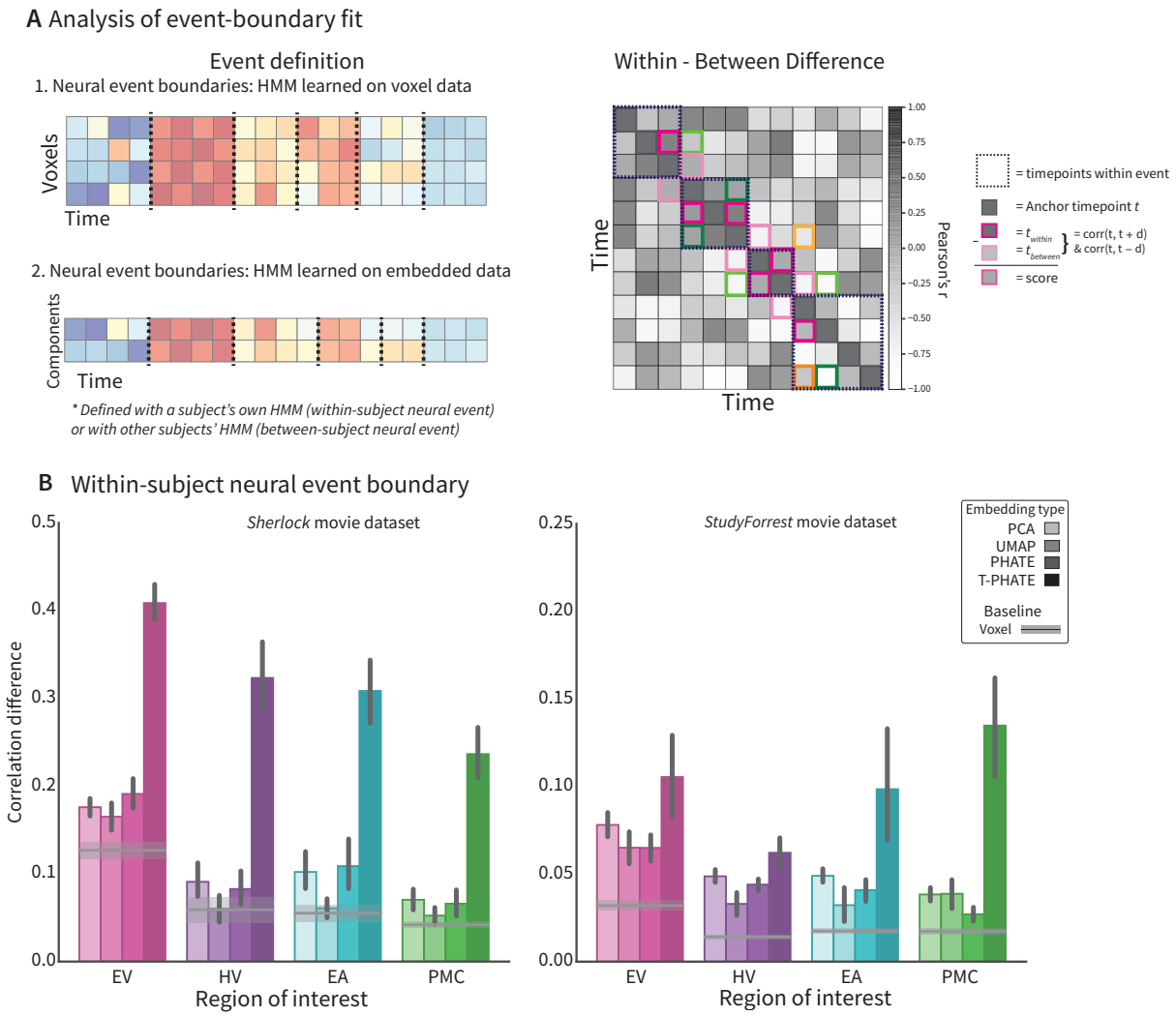

Figure 5: **A.** Event boundaries are learned both from the data (by applying the HMM on voxel resolution or embedded timeseries data and applying those boundaries either within or across subjects) and from independent sources including human behavioral ratings. (Right) Representation of dynamic information in a latent space is quantified by applying event boundaries to embedded data. Data emphasizes brain state transitions if the boundaries maximize the difference between brain activity at timepoints across events versus within events. For each TR t_0 in the timeseries, we correlate the activation pattern across features at t_0 with that of t_1 and t_2 such that t_1 is in the same event as t_0 (within a dashed box) and t_2 is in a different event. The within-between score is the subtraction of $corr(t_2, t_0)$ from $corr(t_0, t_1)$. To account for intrinsically higher correlations of proximal timepoints, we restrict comparisons to where $|t_0 - t_1| = |t_0 - t_2|$. Here, $corr(t_0, t_1)$ is denoted by a darker colored box and its corresponding $corr(t_2, t_0)$ a lighter shade. **B.** Neural event boundaries for subject N are identified with the HMM and then applied to calculate the subject's within-between score. Different dimensionality-reduced approaches are shown as bars and the voxel resolution benchmark is shown as line with shading representing mean and 95% confidence interval.

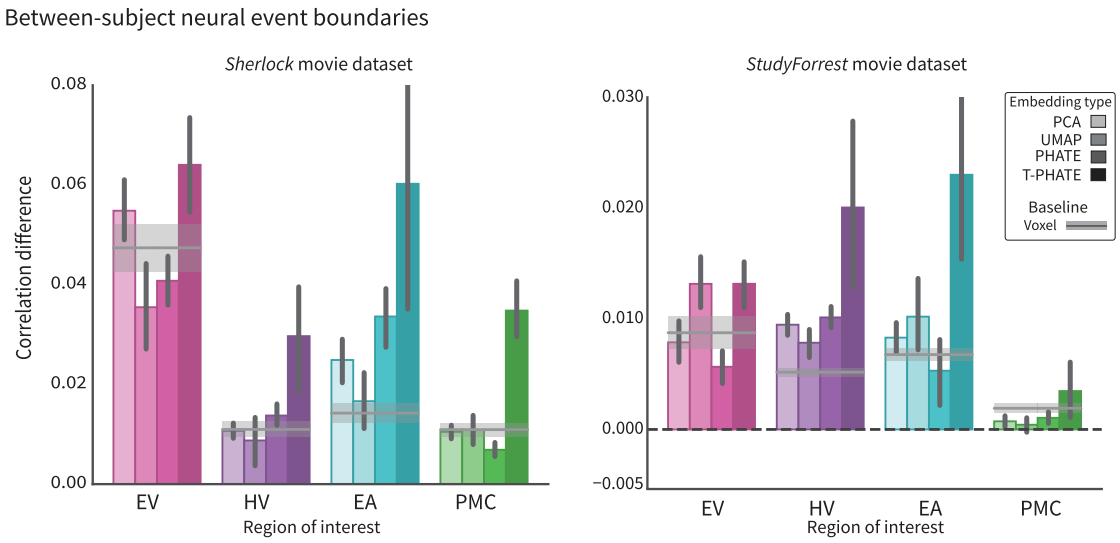

Figure 6: Event boundaries were rotated between subjects by testing subject *N* iteratively with the event boundaries learned from HMMs applied to each of the remaining *N-1* subjects (within ROI and embedding type). T-PHATE embeddings reveal more robust event-related structure even when event boundaries are learned from different subjects.

behavioral boundaries to the neural data in the same fashion as we did the neural events and measured the neural distance within and across boundaries (again equating temporal distance). The ground-truth behavioral event boundaries resulted in significantly stronger within-versus-between neural differences when applied to the T-PHATE embedding than to any of the other methods (Fig. 7B). The PMC is a higher-order integrative region that has previously been shown to predict behavioral boundaries (Baldassano et al., 2017; Chen et al., 2017). We replicated these findings, showing a significant effect in PMC on voxel data (and with other methods), but critically T-PHATE improved on this performance by an order of magnitude. This gain is particularly striking because the behavioral data being predicted were generated by other subjects, in a different task, and did not rely on an HMM of neural data to find the boundaries.

4 Discussion

In this study, we investigated the latent space of regional brain activity during naturalistic viewing from fMRI data. We found that this latent space best captured dynamic information about the experience of a stimulus using a low-dimensional subspace identified with T-PHATE, our novel method that incorporates the temporal dynamics of a signal to learn a more robust manifold. By reducing into nonlinear dimensions that capture not just the geometry of multivoxel activity patterns but also their temporal dynamics, we reveal trajectories through cognitive states that were previously obscured by considering the data in its full resolution. T-PHATE can capture meaningful cognitive trajectories in as few as three dimensions from single-subject fMRI data, which is inherently noisy in both space and time.

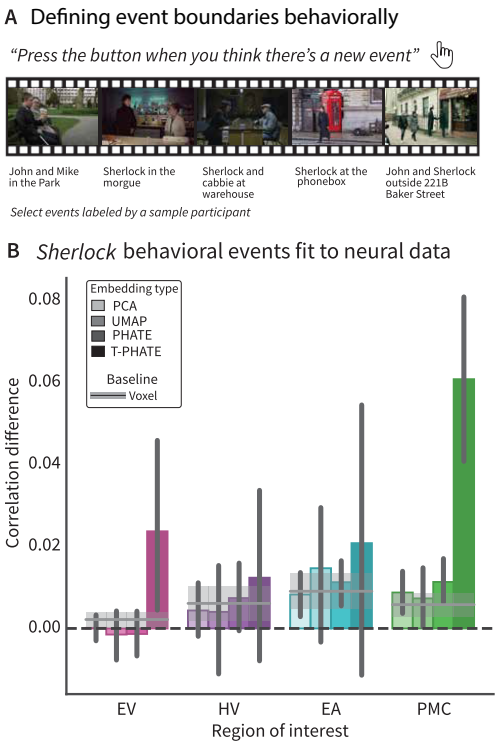

Figure 7: **A.** Event boundaries were also identified by behavioral ratings in a separate experiment on the *Sherlock* dataset with other subjects. Behaviorally-defined event boundaries were then applied to the neural data and within-versus-between event correlation was evaluated as in Fig. 5A) **B.** These behavioral boundaries best mirror event structure in PMC and are best revealed with T-PHATE. Different dimensionality-reduced approaches are shown as bars and the voxel resolution benchmark is shown as line with shading representing mean and 95% confidence interval.

We quantified the degree to which manifold representations of brain activity reflect the dynamics of the stimulus and cognitive processing using a neural event segmentation framework. We applied Hidden Markov Models to learn transitions between stable neural states that reflect boundaries between event representations in different brain regions. The event boundaries learned on data embedded with T-PHATE provided a better fit to the neural data of same subject, predicted the neural similarity of timepoints from that subject, and generalized to other subjects. Moreover, the data embedded with T-PHATE were able to predict where in the movie other human raters would indicate behaviorally that they sensed an event boundary subjectively. To our knowledge, previous investigations of neural event segmentation have not used dimensionality reduction to either denoise neural data or model the structure of neural dynamics. All dimensionality reduction methods we presented did a better job at modeling event dynamics than the voxel-dimensional data (Fig. 4B), likely because they denoised the data in low dimensions. Nevertheless, T-PHATE vastly outperforms PHATE, UMAP, and PCA. T-PHATE manifolds reveal neural dynamics previously obscured by the high-dimensionality, noise, and redundancy of both fMRI data and the neural population activity underlying it by modeling the latent spatiotemporal geometry of brain activity.

Importantly, the advantage T-PHATE gains in dynamic modeling is not at the expense of static decoding. Multivariate pattern classification of stimulus features on low-dimensional T-PHATE embeddings outperform or match decoding from the voxel resolution data (Fig. 2B; Supplemental Fig. 2) and outperform the other dimensionality reduction methods we consider. This shows that while both time-considerate and time-agnostic manifolds can encode static information, only the former can additionally unveil dynamic information.

Cognitive processes like learning, language comprehension (Xu et al., 2005), theory of mind (Gallagher and Frith, 2003; Saxe et al., 2004), narrative understanding (Honey et al., 2012; Xu et al., 2005; Yeshurun et al., 2017), and memory (Hasson et al., 2015) do not occur at a single point in time — they require the accumulation of information over longer temporal windows (Hasson et al., 2008). These automatic processes define our conscious experience of the world in real-time, structuring continuous input into meaningful events that shape perception and memory (Baldassano et al., 2017; Radvansky and Zacks, 2017; Yates et al., 2021; Zacks, Braver, et al., 2001; Zacks et al., 2007; Zacks et al., 2010). To understand how such complex, integrative processing can happen so seamlessly during conscious experience, it is helpful to consider the geometry of event representations in the brain. The fMRI literature suggests that flexible population codes underlie complex cognitive processing (Baldassano et al., 2017; Chen et al., 2017; Davatzikos et al., 2005; Haxby, 2001; Haxby et al., 2014). Further, the animal literature shows that population representations underlying cognitive processes can be modeled in low-dimensional manifolds (Churchland et al., 2010; Cunningham and Yu, 2014; Nieh et al., 2021), and the temporal unfoldings of behavior are best predicted with nonlinear, temporally-dependent manifolds (Low et al., 2018; Nieh et al., 2021).

To date, linear manifolds have been used to track task-related brain states across block-design cognitive tasks (Shine et al., 2016; Shine et al., 2019) and nonlinear manifolds have been used to model the dynamics of cognition across these tasks (Gao et al., 2020). These approaches have required aggregation of whole-brain data across dozens of subjects to define a manifold, and the dynamics captured are coarse, population-average representations of traditional, task-defined cognitive states (such as working memory or social tasks (Gao et al., 2020), limiting applicability in understanding fine-scale, individual brain dynamics. To our knowledge, no one has yet investigated individual subject manifolds learned from single-subject fMRI data collected with naturalistic stimuli to investigate temporal dynamics of individual brain regions. This is likely due to the noise of single-subject, naturalistic fMRI data. With the proper choice of dimensionality reduction algorithm (Cunningham and Yu, 2014; Jazayeri and Ostojic, 2021), we accounted for these challenges to unveil latent dynamics of the cognitive processes related to movie watching.

In conclusion, we found that nonlinear, temporal manifold learning can identify a low-dimensional subspace of brain activity from continuous, naturalistic stimuli that captures higher-order cognitive representations in the brain. Critically, latent dimensions learned with this combination of nonlinear dimensionality reduction and temporal diffusion outperforms methods lacking either of those components. This suggests that the relation between the dimensions of individual voxels and the dominant signals that emerge from them are tied by nonlinear relations. By modeling a subject's BOLD timeseries in these

dimensions, we can gauge the trajectory of their brain states throughout the course of the experiment, and we can relate this trajectory to theoretical frameworks like event segmentation theory that facilitate higher-order cognitive processes in the brain. Because this is performed at the level of individual subjects, future investigations could use this approach to probe individual differences, developmental trajectories, and/or clinical disorders in the native latent space. The characteristics of these manifolds could inform theories about the nature and dynamics of human cognition.

5 Methods

5.1 Manifold embeddings

Different dimensionality reduction algorithms focus on different aspects of a dataset, thus amplify distinct signals depending upon the structure targeted by the algorithm. We hypothesized that a nonlinear manifold learning algorithm designed to handle noisy, high-dimensional biological data, would be best suited for the signal and intrinsic noise of fMRI data. Potential of heat diffusion for affinity-based transition embedding (PHATE) (Moon et al., 2018) is a diffusion-based manifold learning method that models local and global structures simultaneously in nonlinear dimensions. Brain activity measured with fMRI is highly noisy in both space and time, with the BOLD signal canonically peaking 4-5 s after stimulus onset before slowly returning to baseline. With a temporally-dependent stimulus such as a movie, where conversations and plot lines play out over different timescales, the autocorrelation of nearby timepoints will likely extend beyond the curve of the BOLD signal and vary by ROI along the hierarchy of temporal integration in the brain (Himberger et al., 2018). We estimate an autocorrelation kernel for each ROI by correlating each voxel timeseries with lagged versions of itself until the correlation drops and stays below zero, then averaging the functions across voxels to get a regional autocorrelation function. This function is then expanded to calculate the transitional probability between all pairs of timepoints based solely on their estimated autocorrelation and then combined with the PHATE-based transitional probability matrix and embedded in lower dimensions (Fig. 1A).

5.1.1 PHATE

Given a dataset of voxel time-series data, $\mathcal{X} = x_1, x_2, \ldots, x_T$, where $x_t \in \mathbb{R}^n$ at time t is a n-dimensional vector and n is the number of voxels. Construction of the PHATE diffusion geometry starts by computing the Euclidean distance matrix D between data pairs, where $D(i,j) = ||x_i - x_j||^2$. D is then converted into a local affinity matrix K using an adaptive bandwidth Gaussian kernel. The affinity matrix K is then row-normalized to get the initial probabilities P_D, which are used for the Markovian random-walk diffusion process. The diffusion operator P_D is then raised to the t_Dth power, where $t_D$is the PHATE optimal diffusion time scale, to perform a t_D-step random walk over the data graph. This specific diffusion process infers more global relations beyond the local affinities. To further preserve both local and manifold-intrinsic global structure of data, PHATE computes

the diffusion potential U by taking the logarithm of $P_D^{t_D}$ and the diffusion potential distance D_U using distance between row pairs of U. This diffusion potential distance can be reduced to 2-3 dimensions for visualization or to any other dimensionality (such as that obtained by cross-validation) by performing metric MDS (MMDS).

5.1.2 T-PHATE

We designed T-PHATE as a variant of PHATE that uses a dual-view diffusion operator to embed timeseries data in a low-dimensional space. The first view, P_D, of the diffusion operator uses the same process as PHATE to build an affinity matrix based on the Euclidean distance between data samples (here, timepoints) and then row-normalize it to obtain the transition probability matrix, where P_D is a TxT matrix. The second view P_T is based on an affinity matrix that summarizes the autocorrelation function of the data (Brockwell and Davis, 2002). First we calculate a direct estimator of the autocovariance of the timeseries vector V from each voxel using $|V| - 1$ lags. Then the functions are averaged across voxels to obtain a single autocorrelation function (acf) and smoothed with a rolling average over w timepoints, to account for possible jittering around where $acf = 0$. Next, we find the first lag lag_{max} where $acf = 0$, which defines the maximum width of smoothing for the temporal affinity matrix A. The temporal affinity matrix A is calculated as $A(i,j) = acf(i,j), 0 < |i-j| <= lag_{max}$ and zero elsewhere. The autocorrelative probability P_t is obtained by row-normalizing the affinity matrix A. These views are combined with alternating diffusion into a single operator $P = P_D P_T$, which is used as the initial probabilities for diffusion. The rest of constructing diffusion geometry and data visualization is performed the same as in PHATE. This dual-view diffusion step allows T-PHATE to learn data geometry and latent signals that represent cognitive processes that play out over longer temporal windows.

We compare T-PHATE's performance at learning dynamic neural manifolds with common dimensionality reduction algorithms that are agnostic to time including principal components analysis (PCA), uniform manifold approximation and projection (UMAP) (Becht et al., 2019), and vanilla PHATE (Moon et al., 2018). To test whether our autocorrelation kernel was the best approach to modeling temporal dynamics in this manifold, we tested two additional versions of PHATE: incorporating time as a feature in the input data (PHATE+Time) and smoothing the data temporally (Smooth PHATE). fMRI data were extracted from regions of interest (ROIs) and normalized before embedding.

5.2 Event segmentation modeling

Human experience of real life is continuous, yet perception and conception of this experience is typically divided into discrete moments or events. Event segmentation theory explains that humans automatically divide continuous streams of information into discrete events (Radvansky and Zacks, 2017; Zacks et al., 2007) in order to form, organize, and recollect memories, make decisions, and predict the future (Shin and DuBrow, 2021). Participants show high consistency in explicitly segmenting continuous stimuli (Zacks et al., 2010) and also in how their brains represent these event boundaries (Baldassano et al.,

2017; Yates et al., 2021). Event boundaries can be represented, both behaviorally and in brain activity, at different timescales depending on the information being used to draw event boundaries (Zacks, Braver, et al., 2001). In the brain, event boundaries are reflected by shifts in the stability of regional activity patterns, and we can learn the structure of events with a variant of a Hidden Markov Model (Baldassano et al., 2017). During a continuous stimulus such as a movie, different brain regions represent events along different timescales, which reflect the dynamics of the information being represented by a given region. For example, early sensory regions represent low-level information about a stimulus. In most movie stimuli, low-level sensory features shift quickly as cameras change angles or characters speak, and so early sensory regions show more frequent event shifts (Baldassano et al., 2017; Baldassano et al., 2018). Later sensory regions represent longer timescale events, such as conversations and physical environments that change less frequently than features like camera angles. Regions associated with memory or narrative processing represent events on longer timescales, and these boundaries best correspond with scene changes marked by human raters (Lee et al., 2021).

We used the Hidden Markov Model (HMM) variant presented in (Baldassano et al., 2017) and implemented in BrainIAK (Kumar et al., 2020) to learn from different representations of BOLD timeseries data where a brain region experiences event boundaries, or transitions between stable states of activation. Given an activation timeseries of a brain region during a continuous stimulus, the HMM identifies stable activity patterns or "events" that are divided by boundaries, where the timeseries transitions between two stable patterns of activation. This is done iteratively and the log-likelihood of the model fit is evaluated at each iteration, and model fitting stops when the log-likelihood begins to decrease. The first step of this process is to use the HMM to learn from the data the number of events a brain region represents for a given stimulus. Past studies have used HMMs for event segmentation on multivoxel activity patterns and have validated this approach against behavioral segmentation (Baldassano et al., 2017; Lee et al., 2021; Yates et al., 2021). This shows that voxel resolution data reflect meaningful event segmentation, so we chose to estimate the optimal number of events (K) for each brain region using the voxel resolution data (Fig. 4), which also prevents overfitting to the embedding data.

To run an HMM for event segmentation on our manifold embedding data, we needed to tune two parameters for each subject: K, or the number of hidden states through which a brain region passes during the stimulus, and M, the dimensionality of the latent space that captures this signal.

5.2.1 Optimizing number of neural events.

We estimated K and M for each brain region and embedding method with a leave-one-subject-out cross-validation scheme. Using BOLD data from $N-1$ training subjects, we created an average timeseries matrix by averaging response profiles for each voxel across the training subjects. We define K as the number of hidden states in an HMM and optimize K by fitting a model with K states on the training data, applying it on the BOLD data of the held-out test subject, and calculating the log probability of the test data according to the learned event segmentation of the model (hence referred to as the model

log-likelihood). We tested a range of K from 2 to 200 for *Sherlock* and 2 to 400 for *Study-Forrest*. For each training fold, we calculated the best K value that maximized the model log-likelihood of the voxel resolution data, setting subject N's K value to the average of the $N-1$ training subjects' Ks. After definition on the voxel resolution data, K for each subject and region was held constant for all embedding types.

5.2.2 Optimizing manifold dimensionality.

We used the same leave-one-subject-out cross-validation scheme to select the dimensionality of manifold embeddings for each subject. We embedded each subject's timeseries data into an array of possible M values then fit an HMM on each M-dimensional embedding with the subject's cross-validated K value, which had been learned in the previous step from all the other subjects' voxel resolution data. Event boundary fit evaluation is described in more detail below. The M parameter that maximizes the event boundary fit was recorded for each test subject. We then set the M value to be used for subject N as the average of all $N-1$ subjects' best M values.

5.2.3 Evaluating the fit of event boundaries.

To quantify how well a learned manifold embedding amplifies dynamic structure in the data, we calculated difference in correlation across latent dimensions for pairs of timepoints within and across event boundaries. For this calculation, we restricted the timepoints that went into the calculation as follows. We calculated the length of the longest event, and only considered timepoints that were less than that temporal distance apart. We anchored a comparison on each timepoint t in the timeseries, and for each temporal distance n ranging from 1 timepoint to the length of the longest event, we considered a pair of timepoints $t-n$ and $t+n$ if one timepoint fell in the same event as the anchor and the other a different event. We then took the correlation between the anchor and each timepoint, binning them according to whether they were a within-event comparison or a between-event comparison. This process assures that there are equal numbers of timepoints going into each comparison type (Fig. 5A).

One issue of defining and testing event boundaries within-subject is overfitting. To circumvent this, we cross-validated the event boundaries across subjects, testing $N-1$ sets of event boundaries on each Nth subject's neural data using the same procedure outlined above.

To assess the behavioral relevance of the neural event segmentation, we used a set of event boundaries identified by a separate cohort of *Sherlock* study participants. These participants watched the *Sherlock* episode outside of the scanner and were asked to indicate where they believed a new event began (Vodrahalli et al., 2018) (Fig. 7). We applied these human-labeled boundaries to the neural data and measured the fit of the boundaries (as outlined above) to gauge how the embeddings not only highlight neural dynamics but how those neural dynamics relate to the conscious, real-world experience of events.

5.3 Data

5.3.1 Sherlock dataset

For full details on the *Sherlock* data, see the original publication of this dataset (Chen et al., 2017). Here, we used data from the 16 participants who viewed a 48-minute clip of the BBC television series "Sherlock." Data were collected in two fMRI runs of 946 and 1030 timepoints (repetition time; TRs) and was downloaded from the DataSpace public repository (http://arks.princeton.edu/ark:/88435/dsp01nz8062179). The data were collected on a Siemens Skyra 3-T scanner with a 20-channel head coil. Functional images were acquired with a T2*-weighted echo-planar imaging sequence (TE 28ms, TR 1.5s, 64-degree flip angle, whole brain coverage with 27 slices of 4mm thickness and 3x3 mm^2 in-plane resolution, 192x192mm^2 FOV). Anatomical images were acquired with a T1-weighted MPRAGE pulse sequence with 0.89mm^3 resolution. Slice-time correction, motion correction, linear detrending, high-pass filtering (140s cutoff), and co-registration and affine transformation of functional volumes to the Montreal Neurological Institute (MNI) template were all performed with FSL. Functional images were then resampled from native resolution to 3mm isotropic voxels for all analyses, z-scored across time at every voxel, and smoothed with a 6mm kernel.

5.3.2 StudyForrest dataset

For full details on the *StudyForrest* data, see the original publication of the movie dataset (Hanke et al., 2014) and the localizer extension dataset (Sengupta et al., 2016). Here, we included data from 15 participants who completed both tasks. All participants were native German speakers. In the movie-viewing task, participants watched a 2-hour version of the movie "Forrest Gump." These data were collected in eight fMRI runs resulting in a full timeseries of 3599 TRs (451, 441, 438, 488, 462, 439, 542, and 338 per run). Movie data were collected on a 7-T Siemens scanner with a 32-channel headcoil. Functional images were acquired with a T2*-weighted echo-planar imaging sequence (TE 22ms, TR 2s, whole brain coverage with 36 slices of 1.4 mm thickness and 1.4 x 1.4 mm^2 in-plane resolution, 224mm FOV). In the localizer task, the same fifteen participants viewed 24 unique grayscale images from 6 categories (human faces, human bodies without heads, houses, small objects, outdoor scenes, and scrambled images) in four (156 TR) block-design runs with two 16s blocks per stimulus category per run. Localizer data were collected on a 3-T Philips Achieva scanner with a 32-channel headcoil. Functional images for the localizer task were acquired with a T2*-weighted echo-planar imaging sequence (TE 30ms, TR 2s, 90-degree flip angle, 35 slices of 3mm thickness and 3x3 mm^2 in-plane resolution, 240mm FOV). Structural images were acquired with a 3-T Philips Achieva using a 32-channel headcoil. T1-weighted anatomical images were acquired with a 3D turbo field echo sequence with 0.67mm isotropic resolution. Slice-time correction, co-registration and affine transformation to MNI template were performed with FSL where functional images were resampled to 3mm isotropic voxels. Additional preprocessing included linear detrending, high-pass filtering (100s cutoff), spatial smoothing with a 6mm kernel, and nuisance regression (including 6 motion parameters, global signal, white matter, and cerebrospinal

fluid), and z-scoring within voxel to match the *Sherlock* preprocessing.

5.4 Region of interest (ROI) selection

We selected four regions of interest (ROIs) based on the original publication of the *Sherlock* dataset (Chen et al., 2017), subsequent publications (Baldassano et al., 2017), and regions known to have reliably strong signals in response to audiovisual movie stimuli (Haxby et al., 2021). The early visual (EV), early auditory (EA), and high visual (HV) region masks were based on a functional atlas defined with resting-state connectivity (Shirer et al., 2012). As in the original *Sherlock* publication, we defined a posterior medial cortex (PMC) ROI as the posterior medial cluster of voxels within the dorsal default mode network (Chen et al., 2017). In the *Sherlock* data, the dimensionality of voxel-resolution ROIs are as follows: EV = 307, HV = 571, EA = 1018, PMC = 481. In the *StudyForrest* data, the dimensionality of voxel-resolution ROIs are as follows: EV = 166, HV = 456, EA = 657, PMC = 309.

5.5 Code availability

T-PHATE is available as a python package at: https://github.com/KrishnaswamyLab/TPHATE. All analysis scripts will also be made available at time of publication.

5.6 Author contributions

E.L.B., J.H., G.W., G.L., S.K., N.B.T-B. conceived the idea. E.L.B., J.H., A.B., and S.K. designed the T-PHATE algorithm. E.L.B. and J.H. developed software. E.L.B. curated data and designed analyses, and E.L.B, J.H., and T.W. performed the analyses. E.L.B., S.K., and N.B.T-B. wrote the initial manuscript draft. All authors revised the paper. S.K. and N.B.T-B. jointly supervised the work.

References

Allen, E. A., Damaraju, E., Plis, S. M., Erhardt, E. B., Eichele, T., & Calhoun, V. D. (2012). Tracking Whole-Brain Connectivity Dynamics in the Resting State. *Cerebral Cortex*, *24*(3), 663–676. https://doi.org/10.1093/cercor/bhs352

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience*, *7*(5), 358–366.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.e5. https://doi.org/https://doi.org/10.1016/j.neuron.2017.06.041

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699.

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. *Nature Reviews Neuroscience*, *22*(6), 359–371. https://doi.org/10.1038/s41583-021-00448-6

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, *37*(1), 38–44.

Birn, R. M., Smith, M. A., Jones, T. B., & Bandettini, P. A. (2008). The respiration response function: The temporal dynamics of fmri signal fluctuations related to changes in respiration. *NeuroImage*, *40*(2), 644–654. https://doi.org/https://doi.org/10.1016/j.neuroimage.2007.11.059

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, *169*(6), 1013–1028.e14. https://doi.org/https://doi.org/10.1016/j.cell.2017.05.011

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., & Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature neuroscience*, *20*(1), 115–125. https://doi.org/10.1038/nn.4450

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Ryu, S. I., & Shenoy, K. V. (2010). Cortical preparatory activity: Representation of movement or first cog in a dynamical machine? *Neuron*, *68*(3), 387–400. https://doi.org/10.1016/j.neuron.2010.09.015

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fmri) "brain reading": Detecting and classifying distributed patterns of fmri activity in human visual cortex. *NeuroImage*, *19*(2), 261–270. https://doi.org/https://doi.org/10.1016/S1053-8119(03)00049-1

Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, *17*(11), 1500–1509. https://doi.org/10.1038/nn.3776

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., Gur, R. C., & Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, *28*(3), 663–668. https://doi.org/https://doi.org/10.1016/j.neuroimage.2005.08.009

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845. https://doi.org/10.1126/science.1194908

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in cognitive sciences*, *7*(2), 77–83.

Gao, S., Mishne, G., & Scheinost, D. (2020). Non-linear manifold learning in fmri uncovers a low-dimensional space of brain dynamics. *bioRxiv*, 2020.11.25.398693. https://doi.org/10.1101/2020.11.25.398693

Golesorkhi, M., Gomez-Pilar, J., Zilio, F., Berberian, N., Wolff, A., Yagoub, M. C. E., & Northoff, G. (2021). The brain and its time: Intrinsic neural timescales are key for input processing [Number: 1 Publisher: Nature Publishing Group]. *Communications Biology*, *4*(1), 1–16. https://doi.org/10.1038/s42003-021-02483-6

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*(1), 140003. https://doi.org/10.1038/sdata.2014.3

Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: Memory as an integral component of information processing. *Trends in cognitive sciences*, *19*(6), 304–313. https://doi.org/10.1016/j.tics.2015.04.006

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, *28*(10), 2539–2550. https://doi.org/10.1523/JNEUROSCI.5487-07.2008

Haxby, J. V. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis [PMID: 25002277]. *Annual Review of Neuroscience*, *37*(1), 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Haxby, J. V., Guntupalli, J. S., Nastase, S. A., & Feilong, M. (2021). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife*, *9*. https://doi.org/10.7554/eLife.56601

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Hennig, J. A., Golub, M. D., Lund, P. J., Sadtler, P. T., Oby, E. R., Quick, K. M., Ryu, S. I., Tyler-Kabara, E. C., Batista, A. P., Yu, B. M., & Chase, S. M. (2018). Constraints on neural redundancy. *eLife*, *7*, e36774. https://doi.org/10.7554/eLife.36774

Himberger, K. D., Chien, H.-Y., & Honey, C. J. (2018). Principles of temporal processing across the cortical hierarchy [Sensory Sequence Processing in the Brain]. *Neuroscience*, *389*, 161–174. https://doi.org/https://doi.org/10.1016/j.neuroscience.2018.04.030

Honey, C. J., Thesen, T., Donner, T. H., Silbert, L. J., Carlson, C. E., Devinsky, O., Doyle, W. K., Rubin, N., Heeger, D. J., & Hasson, U. (2012). Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, *76*(2), 423–434. https://doi.org/https://doi.org/10.1016/j.neuron.2012.08.011

Huang, J., Busch, E. L., Wallenstein, T., Gerasimiuk, M., Benz, A., Lajoie, G., Wolf, G., Turk-Browne, N. B., & Krishnaswamy, S. (2021). Learning shared neural manifolds from multi-subject fmri data. *arXiv preprint arXiv:2201.00622*.

Husler, C., & Hanke, M. (2016). An annotation of cuts, depicted locations, and temporal progression in the motion picture "forrest gump" [version 1; peer review: 3 approved]. *F1000Research*, *5*(2273). https://doi.org/10.12688/f1000research.9536.1

Ito, T., Hearne, L. J., & Cole, M. W. (2020). A cortical hierarchy of localized and distributed processes revealed via dissociation of task activations, connectivity changes, and intrinsic timescales. *NeuroImage*, *221*, 117141. https://doi.org/10.1016/j.neuroimage.2020.117141

Jazayeri, M., & Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *arXiv preprint arXiv:2107.04084*.

Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685. https://doi.org/10.1038/nn1444

Kumar, M., Anderson, M. J., Antony, J. W., Baldassano, C., Brooks, P. P., Cai, M. B., Chen, P.-H. C., Ellis, C. T., Henselman-Petrusek, G., & Huberdeau, D. (2020). Brainiak: The brain imaging analysis kit.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72–79.

Laurent, G. (2002). Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience*, *3*(11), 884–895. https://doi.org/10.1038/nrn964

Lee, C. S., Aly, M., & Baldassano, C. (2021). Anticipation of temporally structured events in the brain (M. V. Peelen, T. E. Behrens, & L. Geerligs, Eds.). *eLife*, *10*, e64972. https://doi.org/10.7554/eLife.64972

Low, R. J., Lewallen, S., Aronov, D., Nevers, R., & Tank, D. W. (2018). Probing variability in a cognitive map using manifold inference from neural dynamics. *bioRxiv*. https://doi.org/10.1101/418939

Moon, K. R., Stanley, J. S., Burkhardt, D., van Dijk, D., Wolf, G., & Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, *7*, 36–46. https://doi.org/https://doi.org/10.1016/j.coisb.2017.12.008

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., & Krishnaswamy, S. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, *37*(12), 1482–1492. https://doi.org/10.1038/s41587-019-0336-3

Nieh, E. H., Schottdorf, M., Freeman, N. W., Low, R. J., Lewallen, S., Koay, S. A., Pinto, L., Gauthier, J. L., Brody, C. D., & Tank, D. W. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature*, *595*(7865), 80–84. https://doi.org/10.1038/s41586-021-03652-7

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fmri data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Olszowy, W., Aston, J., Rua, C., & Williams, G. B. (2019). Accurate autocorrelation modeling substantially improves fMRI reliability [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *10*(1), 1220. https://doi.org/10.1038/s41467-019-09230-w

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience*, *30*(30), 10127–10134. https://doi.org/10.1523/JNEUROSCI.2161-10.2010

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*(5756), 1963–1966. https://doi.org/10.1126/science.1117645

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition [Memory in time and space]. *Current Opinion in Behavioral Sciences*, *17*, 133–140. https://doi.org/https://doi.org/10.1016/j.cobeha.2017.08.006

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Salhov, M., Bermanis, A., Wolf, G., & Averbuch, A. (2015). Approximately-isometric diffusion maps. *Applied and Computational Harmonic Analysis*, *38*(3), 399–419. https://doi.org/https://doi.org/10.1016/j.acha.2014.05.002

Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.*, *55*, 87–124.

Sengupta, A., Kaule, F. R., Guntupalli, J. S., Hoffmann, M. B., Häusler, C., Stadler, J., & Hanke, M. (2016). A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Scientific Data*, *3*(1), 160093. https://doi.org/10.1038/sdata.2016.93

Shin, Y. S., & DuBrow, S. (2021). Structuring memory through inference-based event segmentation. *Topics in Cognitive Science*, *13*(1), 106–127.

Shine, J. M., Bissett, P. G., Bell, P. T., Koyejo, O., Balsters, J. H., Gorgolewski, K. J., Moodie, C. A., & Poldrack, R. A. (2016). The dynamics of functional brain networks: Integrated network states during cognitive task performance. *Neuron*, *92*(2), 544–554. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5073034/pdf/nihms-820230.pdf

Shine, J. M., Breakspear, M., Bell, P. T., Ehgoetz Martens, K. A., Shine, R., Koyejo, O., Sporns, O., & Poldrack, R. A. (2019). Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nature Neuroscience*, *22*(2), 289–296. https://doi.org/10.1038/s41593-018-0312-0

Shirer, W. R., Ryali, S., Rykhlevskaia, E., Menon, V., & Greicius, M. D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb Cortex*, *22*(1), 158–65. https://doi.org/10.1093/cercor/bhr099

Speer, N. K., Zacks, J. M., & Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological science*, *18*(5), 449–455.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *science*, *290*(5500), 2319–2323.

Turk-Browne, N. B. (2013). Functional interactions as big data in the human brain. *Science*, *342*(6158), 580–584. https://doi.org/10.1126/science.1238409

Vodrahalli, K., Chen, P.-H., Liang, Y., Baldassano, C., Chen, J., Yong, E., Honey, C., Hasson, U., Ramadge, P., Norman, K. A., & Arora, S. (2018). Mapping between fmri responses to movies and their natural language annotations. *NeuroImage*, *180*(Pt A), 223–231. https://doi.org/10.1016/j.neuroimage.2017.06.042

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *Neuroimage*, *25*(3), 1002–1015.

Yates, T. S., Skalaban, L. J., Ellis, C. T., Bracher, A. J., Baldassano, C., & Turk-Browne, N. B. (2021). Neural event segmentation of continuous experience in human infants. *bioRxiv*, 2021.06.16.448755. https://doi.org/10.1101/2021.06.16.448755

Yeshurun, Y., Nguyen, M., & Hasson, U. (2017). Amplification of local changes along the timescale processing hierarchy. *Proceedings of the National Academy of Sciences*, *114*(35), 9475–9480.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651–655. https://doi.org/10.1038/88486

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: Segmentation of narrative cinema. *Frontiers in human neuroscience*, *4*, 168.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of experimental psychology: General*, *130*(1), 29.

1 Supplemental methods

1.1 Movie annotations

***Sherlock* annotations** We used annotations for the *Sherlock* movie stimulus released with Vodrahalli et al., 2018 and available freely on Github. These annotations were created by human annotators who viewed the film and labeled each timepoint on a variety of metrics. Specifically here we used two sets of binary labels that correspond with visual and auditory aspects of the stimulus: whether a given TR featured an indoor or outdoor scene and whether or not the TR featured music.

***StudyForrest* annotations** Annotations of scene cuts and locations used in later analysis of the Forrest Gump movie viewing data were downloaded from a subsequent *StudyForrest* publication Husler and Hanke, 2016. These provide additional information about a number of shots from the film, including the setting of the shot, whether a setting is interior or exterior, the temporal progression of a shot with reference to the previous shot, and the time of day. These are used in the movie feature decoding analysis for the *StudyForrest* dataset.

2 Supplemental results

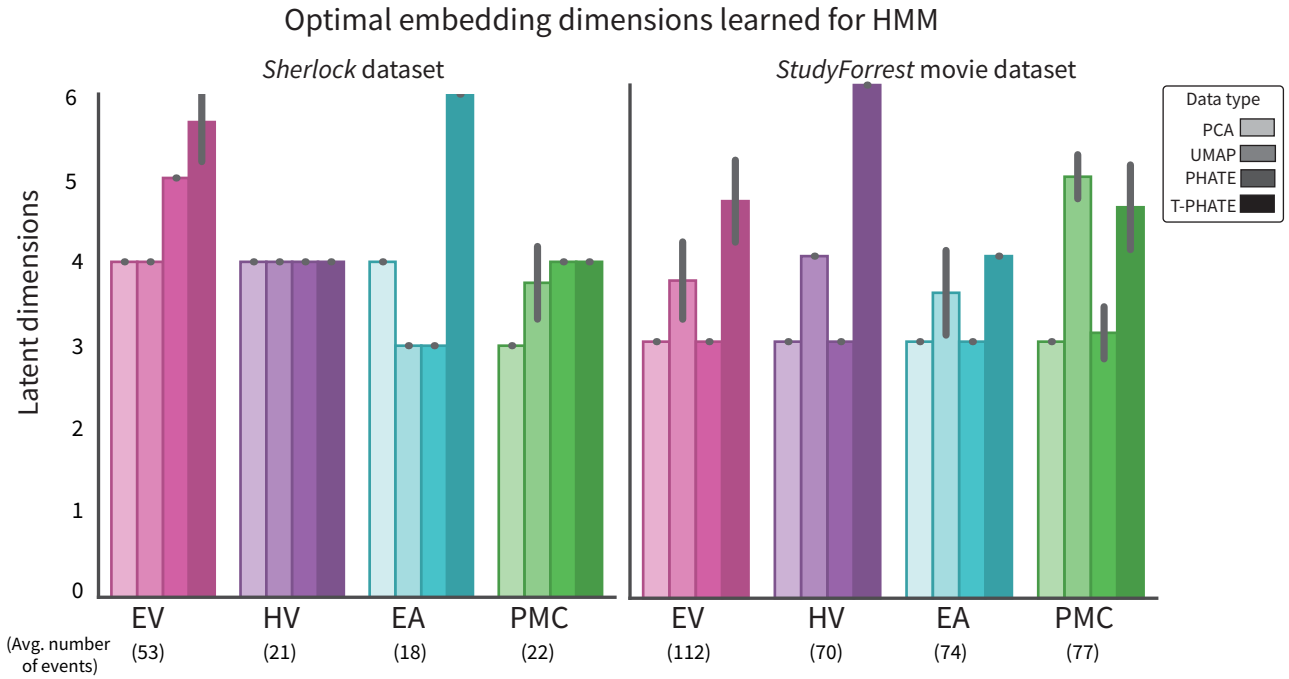

Figure S1: HMM parameters K and M learned with leave-one-subject-out cross validation. Dimensionality reduction methods are able to optimize event boundaries in relatively few dimensions. K values follow the predicted event hierarchy in the brain, with lower-level sensory regions experiencing more events than higher-order cognitive regions.

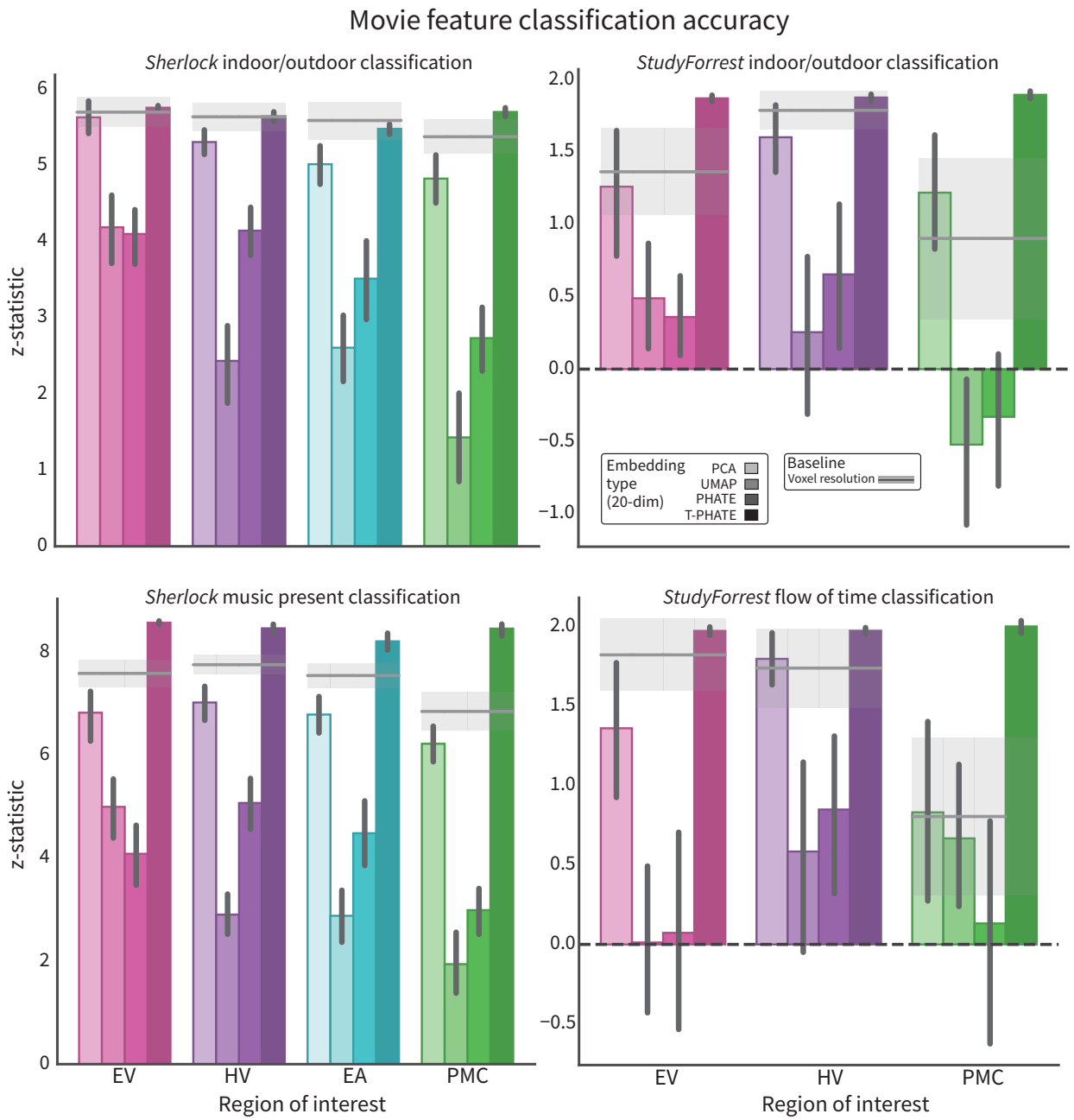

Figure S2: Support vector classification of movie features in the *Sherlock* and *StudyForrest* movie datasets. All labels were binary aside from the Flow of Time labels, which had four levels. Results presented are the z-statistic of true classification accuracy versus 1000 repetitions of time-shifting labels and re-training the model to generate a null distribution.

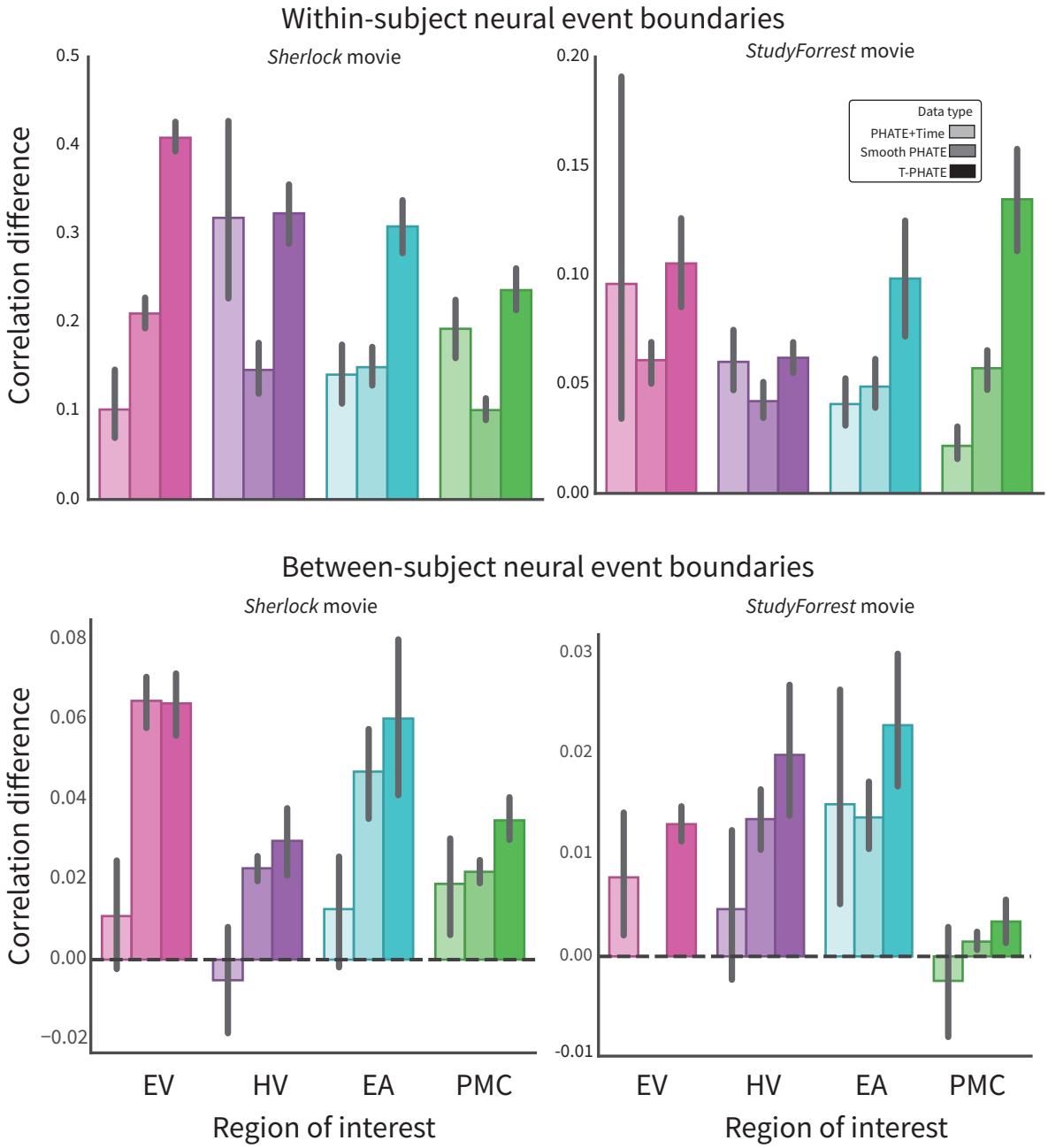

Figure S3: Within-vs-between neural event boundaries for the time alternative methods. **A.** Event boundaries defined as neural events within-subject. **B.** Event boundaries cross-validated across subjects with leave-one-subject-out boundaries.