Published in final edited form as:

Neuroimage. 2020 August 15; 217: 116865. doi:10.1016/j.neuroimage.2020.116865.

Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space

Samuel A. Nastase[a,*], **Yun-Fei Liu**[c], **Hanna Hillman**[d], **Kenneth A. Norman**[a,b], **Uri Hasson**[a,b]

[a]Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

[b]Department of Psychology, Princeton University, Princeton, NJ, USA

[c]Department of Psychological & Brain Sciences, Johns Hopkins University, Baltimore, MD, USA

[d]Department of Psychology, Harvard University, Cambridge, MA, USA

Abstract

Connectivity hyperalignment can be used to estimate a single shared response space across disjoint datasets. We develop a connectivity-based shared response model that factorizes aggregated fMRI datasets into a single reduced-dimension shared connectivity space and subject-specific topographic transformations. These transformations resolve idiosyncratic functional topographies and can be used to project response time series into shared space. We evaluate this algorithm on a large collection of heterogeneous, naturalistic fMRI datasets acquired while subjects listened to spoken stories. Projecting subject data into shared space dramatically improves between-subject story time-segment classification and increases the dimensionality of shared information across subjects. This improvement generalizes to subjects and stories excluded when estimating the shared space. We demonstrate that estimating a simple semantic encoding model in shared space improves between-subject forward encoding and inverted encoding model performance. The shared space estimated across all datasets is distinct from the shared space derived from any particular constituent dataset; the algorithm leverages shared connectivity to yield a consensus shared space conjoining diverse story stimuli.

Keywords

Data harmonization; fMRI; Functional connectivity; Hyperalignment; Naturalistic stimuli

[*]Corresponding author. sam.nastase@gmail.com (S.A. Nastase).

CRediT authorship contribution statement

Samuel A. Nastase: Conceptualization, Methodology, Software, Writing - original draft. **Yun-Fei Liu:** Conceptualization, Methodology, Data curation. **Hanna Hillman:** Data curation, Resources. **Kenneth A. Norman:** Conceptualization, Supervision, Writing - review & editing. **Uri Hasson:** Conceptualization, Funding acquisition, Supervision, Writing - review & editing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2020.116865.

1. Introduction

The developing infrastructure for data sharing (alongside evolving incentives) has led to a proliferation of publicly available "open" neuroimaging data. Although still overshadowed by traditional task and resting-state acquisitions, we are beginning to see more public data collected during rich, naturalistic paradigms (e.g., Hanke et al., 2014, 2016; Taylor et al., 2017; DuPre et al., 2019). Although the neuroimaging community unequivocally benefits from the increasing availability of public data (Poldrack and Gorgolewski, 2014; Milham et al., 2018), this trend introduces a challenge. Namely, datasets are often markedly heterogeneous—i.e., collected on different scanners, using different acquisition parameters, with different samples of subjects—and require sophisticated harmonization (e.g., Yamashita et al., 2019). Furthermore, in the context of naturalistic stimuli (e.g., movie-watching, story--listening), stimuli vary considerably from experiment to experiment. Here we focus on a particular aspect of harmonization: finding a shared functional response space across heterogeneous naturalistic datasets.

In order to fully realize the potential of "big" neuroimaging data for prediction and translational purposes, we need to obtain some level of correspondence across individuals (Gabrieli et al., 2015; Dubois and Adolphs, 2016; Woo et al., 2017). Typically, each individual brain is spatially normalized to a standard space based on macroanatomical features such as sulcal curvature (Fischl et al., 1999; Coalson et al., 2018). However, fine-grained functional response topographies (e.g., Brett et al., 2002; Duncan et al., 2009; Frost and Goebel, 2012; Haxby et al., 2014; Zhen et al., 2015, 2017) and connectivity patterns (e.g., Langs et al., 2016; Braga and Buckner, 2017; Gordon et al., 2017; Bijsterbosch et al., 2019) are not tightly coupled to macroanatomical features and are markedly idiosyncratic across individuals. Information encoded at this finer scale may be inaccessible based on anatomical alignment alone (Feilong et al., 2018; Kong et al., 2019). In order to leverage large volumes of data for prediction in individuals, we need to resolve these idiosyncrasies in functional–anatomical correspondence. Hyperalignment is a family of algorithms for normalizing functional data into a common space by resolving topographic idiosyncrasies (Haxby et al., 2011; Guntupalli et al., 2016). These methods hinge on functional commonalities to drive normalization—typically a rich stimulus is used to evoke stereotyped, time-locked response trajectories across subjects. However, recent work by Guntupalli et al. (2018) demonstrates that functional connectivity can be used to effectively drive functional normalization absent any shared stimulus. Each voxel's participation in functional networks elsewhere in the brain can provide a shared functional signature sufficient for resolving topographic idiosyncrasies. One important product of this development, which the authors examined in detail, is the extension of hyperalignment to resting-state functional connectivity. In this case, the "rest" task yields rich and consistent enough connectivity patterns to support functional normalization. A separate, largely unexplored avenue is that connectivity hyperalignment may allow us to define a single common space across distinct naturalistic stimuli or tasks.

Here, we use a variant of connectivity-based hyperalignment (Guntupalli et al., 2018) to showcase the utility of aggregating disjoint naturalistic story-listening datasets into a single, shared response space. For a given region of interest (ROI), we first compute intersubject

functional correlations (ISFC; Simony et al., 2016) between each voxel and a set of parcels tiling the cortex (i.e., connectivity targets; Glasser et al., 2016). We then apply the shared response model (SRM; Chen et al., 2015) to these connectivity patterns to find a reduced-dimension connectivity space shared across both subjects and stimuli. Critically, in the context of a task such as listening to spoken stories, we expect these coarse connectivity patterns to be well-preserved across subjects and stimuli. The SRM effectively decomposes the connectivity data across all datasets into a shared connectivity space, and a set of subject-specific transformation matrices that resolve topographic idiosyncrasies. Although the shared model is derived from functional connectivity, the subject-specific topographic transformations can be used to project response time series into shared space. We benchmark this algorithm on a large, heterogeneous collection of story-listening functional MRI datasets assembled over the course of approximately seven years. This data collection comprises 10 unique auditory story stimuli across 300 scans with 149 unique subjects. We evaluate the shared space using between-subject time-segment classification (e.g., Haxby et al., 2011), temporal and spatial intersubject correlations (e.g., Nastase et al., 2019a), and between-subject semantic model-based encoding and decoding (e.g., Huth et al., 2016).

2. Materials and methods

2.1. Participants

We aggregated fMRI datasets collected between 2011 and 2018 comprising 10 story stimuli and 149 subjects totalling 300 scans (mean age = 22.6 years, *SD* = 6.25, range: 18–53; 84 reported female). Subjects with behavioral comprehension scores (where applicable) lower than 25% accuracy were excluded. Furthermore, we computed leave-one-subject-out ISCs in a left early auditory cortex ROI (Glasser et al., 2016) for all subjects in each dataset at temporal lags ranging from −100 to 100 TRs and excluded any subjects with a peak ISC at lags exceeding ±1 TR. Here we briefly summarize the resulting sample size and demographics for each dataset, and point to previously published work using these data (see Table 1). The datasets are named according to the names of the corresponding story stimuli, with abbreviated aliases used in analysis and figures. The “Pie Man” data (alias: *pieman*) comprised 46 subjects (mean age = 22.4 years, *SD* = 3.8, 23 reported female; (Simony et al., 2016). The “Pretty Mouth and Green My Eyes” data (alias: *prettymouth*) comprised 19 subjects (mean age = 20.2 years, SD = 2.1, 9 reported female) from the “cheating” condition of the context manipulation described by Yeshurun and colleagues (2017b). The “Milky Way” data comprised 16 subjects (mean age = 19.9 years, *SD* = 1.5, 7 reported female) from one condition (Story1) of the word-substitution manipulation described by Yeshurun and colleagues (2017a). The previously unpublished “Slumlord” and “Reach for the Stars One Small Step at a Time” stories were presented in a single scanning run (alias: *slumlordreach*) and comprised 16 subjects (mean age = 21.1 years, *SD* = 2.4, 8 reported female). The “It’s Not the Fall That Gets You” data (alias: *notthefall*) comprised 18 subjects (mean age = 21.4 years, *SD* = 2.5, 8 reported female) from the “intact” condition described by Chien and Honey (2020). The “The 21st Year” data (alias: *21styear*) comprised 24 subjects (mean age = 23.3 years, SD = 6.7, 14 reported female) and described by Chang et al. (2020). Finally, two stories recorded at the Princeton Neuroscience Institute (PNI) served as stimuli: “Pie Man (PNI)” (alias: *pieman (PNI)*) and “Running from the Bronx (PNI)” (alias: *bronx (PNI)*).

The “Pie Man (PNI)” data comprised 39 subjects (mean age = 23.3 years, *SD* = 7.7, 28 reported female). The “Running from the Bronx (PNI)”, “I Knew You Were Black” (alias: *black*), and “The Man Who Forgot Ray Bradbury” (alias: *forgot*) data were collected at the same time and comprised roughly the same sample of 40 subjects (mean age = 23.3 years, *SD* = 7.6, 29 reported female; Lin et al., 2019). Overall, 83 subjects (56% of the total sample) contributed a single scan, 22 (15%) contributed to two scans, 5 (3%) contributed three scans, 37 (25%) contributed four scans (~40 subjects were acquired for the four “pieman (PNI),” “bronx (PNI),” “forgot,” and “black” stories by design), and 2 (1%) contributed five scans (Fig. S1). All data used herein are publicly available as part of the “Narratives” collection (Nastase et al., 2019b) on the OpenNeuro repository: https://openneuro.org/datasets/ds002345.

2.2. Stimuli and design

Story stimuli were presented auditorily and ranged from ~7 to 56 min in duration (summarized in Table 1). The stimuli included professional storytellers performing for a live audience, actors performing written narratives, and authors reading their written works. For each dataset, TRs corresponding to the story stimulus were isolated by discarding any TRs corresponding to silence or music (padding the beginning or end of a run). In general, participants were instructed to maintain fixation on a centrally-presented crosshair or dot and listen to the story. Behavioral questionnaires assessing narrative comprehension were acquired for the “Pretty Mouth and Green My Eyes”, “Milky Way”, “Slumlord” and “Reach for the Stars One Small Step at a Time”, “The 21st Year”, “Pie Man (PNI)”, “Running from the Bronx (PNI)”, “I Knew You Were Black”, and “The Man Who Forgot Ray Bradbury”. These comprehension scores were used to initially exclude poor-performing or noncompliant participants (participants with accuracies lower than 25%), but were not used in subsequent analyses.

2.3. Image acquisition

MRI data for the “Pie Man”, “Pretty Mouth and Green My Eyes”, “Milky Way”, “Slumlord”, “Reach for the Stars One Small Step at a Time”, “It’s Not the Fall that Gets You”, and “The 21st Year” were collected using a 3T Siemens Skyra with a 20-channel phased-array head coil. Functional blood-oxygenation-level-dependent (BOLD) images were acquired in an interleaved fashion using gradient-echo echo-planar imaging with an in-plane acceleration factor of 2 using mSENSE: TR/TE = 1500/28 ms, flip angle = 64°, bandwidth = 1445 Hz/Px, in-plane resolution = 3 × 3 mm, slice thickness = 4 mm, matrix size = 64 × 64, FoV = 192 ×192 mm, 27 axial slices with roughly full brain coverage and no gap, anterior-posterior phase encoding. At the beginning of each run, three dummy scans were acquired and discarded by the scanner to allow for signal stabilization. T1-weighted structural images were acquired using a high-resolution single-shot MPRAGE sequence with an in-plane acceleration factor of 2 using GRAPPA: TR/TE/TI = 2300/3.08/900 ms, flip angle = 9°, bandwidth = 240 Hz/Px, in-plane resolution 0.859 × 0.859 mm, slice thickness 0.9 mm, matrix size = 256 × 256, FoV = 172.8 × 219.9 × 219.9 mm, 192 sagittal slices, ascending acquisition, no fat suppression.

MRI data for the “Pie Man (PNI)”, “Running from the Bronx (PNI)”, “I Knew You Were Black”, and “The Man Who Forgot Ray Bradbury” stores were collected using a 3T Siemens Prisma with a 64-channel head coil. Functional images were acquired in an interleaved fashion using gradient-echo echo-planar imaging with a multiband acceleration factor of 3 and no in-plane acceleration: TR/TE 1500/31 ms, flip angle = 67°, bandwidth = 2480 Hz/Px, in-plane resolution = 2.5 × 2.5 mm, slice thickness 2.5 mm, matrix size = 96 × 96, FoV = 240 × 240 mm, 48 axial slice with full brain coverage and no gap, anterior-posterior phase encoding, three dummy scans. T1-weighted structural images were acquired using a high-resolution single-shot MPRAGE sequence with an in-plane acceleration factor of 2 using GRAPPA: TR/TE/TI = 2530/3.3/1100 ms, flip angle = 7°, bandwidth = 200 Hz/Px, in-plane resolution = 1.0 × 1.0 mm, slice thickness 1.0 mm, matrix size = 256 × 256, FoV = 176 × 256 × 256 mm, 176 sagittal slices, ascending acquisition, no fat suppression. T2-weighted structural images were acquired using a high-resolution single-shot MPRAGE sequence with an in-plane acceleration factor of 2 using GRAPPA: TR/TE = 3200/428 ms, flip angle = 120°, bandwidth = 200 Hz/Px, in-plane resolution 1.0 × 1.0 mm, slice thickness 1.0 mm, matrix size = 256 × 256 mm, FoV = 176 sagittal slice, ascending acquisition, fat suppression.

2.4. Preprocessing

All MRI data were preprocessed using fMRIPrep (Esteban et al., 2019), which uses Nipype (Gorgolewski et al., 2011) to adaptively construct workflows based on metadata. Anatomical T1-weighted images were corrected for intensity non-uniformity (Tustison et al., 2010) and skull-stripped based on the OASIS template using ANTs (Avants et al., 2008). Cortical surfaces were reconstructed using FreeSurfer (Dale et al., 1999) and tissue segmentation was performed using FSL (Zhang et al., 2001). T2-weighted images were also supplied to surface reconstruction where applicable.

Functional data were slice-time corrected using AFNI (Cox, 1996, 2012) and motion corrected using FSL (Jenkinson et al., 2002, 2012). “Fieldmap-less” susceptibility distortion correction was performed by co-registering the functional image to the T1-weighted image for that subject with intensity inverted (Wang et al., 2017) constrained with an average field map template (Treiber et al., 2016; Wang et al., 2017) using ANTs. Functional images were next co-registered to the corresponding T1-weighted image using FreeSurfer’s boundary-based registration (Greve and Fischl, 2009). Transformations for performing motion correction, susceptibility distortion correction, and functional to anatomical registration were concatenated and applied in a single step with Lanczos interpolation using ANTs. Functional data were then resampled to the subject-specific cortical surface models by averaging samples at six intervals along the normal between the white matter and pial surfaces. Functional data were then spatially normalized to the fsaverage surface template based on sulcal curvature and downsampled to the fsaverage6 template (Fischl et al., 1999). All subsequent analyses (including functional normalization) were performed on surface data (Van Essen and Glasser, 2018), and functional normalization algorithms are compared to relatively high-performing nonlinear surface-based anatomical normalization (Klein et al., 2010). Note that the terms “voxel” and “vertex” are effectively interchangeable for the

analyses of interest; although we sometimes refer to voxels in keeping with conventions in the literature, all analyses of interest were performed explicitly on surface vertices.

The following confound variables were regressed out of the signal in a single step (Lindquist et al., 2019) using AFNI's 3dTproject: linear and quadratic trends, sine/cosine bases for high-pass filtering (cutoff: 0.00714 Hz; ~140 s), six head motion parameters and their derivatives, framewise displacement (Power et al., 2014), and six principal component time series from anatomically-defined cerebrospinal fluid and white matter segmentations (Behzadi et al., 2007).

2.5. Regions of interest

We evaluated the shared model in several regions of interest (ROIs) defined according to a multimodal parcellation (MMP) based on anatomical and functional data from the Human Connectome Project (Glasser et al., 2016). The surface-based parcellation was projected to the fsaverage surface template and downsampled to the fsaverage6 template (Mills, 2016). We focused on four large cortical regions (Fig. 1), each of which comprises several smaller cortical areas. Following the MMP, early auditory cortex (EAC) comprised five areas (A1, MBelt, LBelt, PBelt, RI) and contained 808 and 638 vertices in the left and right hemispheres, respectively. Auditory association cortex (AAC) comprised eight areas (A4, A5, STSdp, STSda, STSvp, STSva, STGa, TA2) and contained 1,420 (left hemisphere) and 1,493 (right hemisphere) vertices. The temporo-parieto-occipital junction (TPOJ) comprised five areas (TPOJ1, TPOJ2, TPOJ3, STV, PSL) and contained 847 (left hemisphere) and 1,188 (right hemisphere) vertices. To reduce the posterior cingulate cortex region to a more comparable size (originally 14 areas containing over 2, 500 vertices per hemisphere), we selected seven core areas (POS1, POS2, v23ab, d23ab, 31pv, 31pd, 7m) containing 1,198 (left hemisphere) and 1,204 (right hemisphere) vertices; we refer to this region as posterior medial cortex (PMC). These ROIs (EAC, AAC, TPOJ, and PMC) span a cortical hierarchy supporting language and narrative comprehension (Lerner et al., 2011; Huth et al., 2016; Baldassano et al., 2017). The selection of ROIs was not intended to be exhaustive; rather, we aimed to benchmark the shared model in a sample of relevant ROIs ranging from low-level sensory cortex to high-level association cortex. We analyze ROIs in both hemispheres (left and right) separately in all subsequent analyses, but generally collapse across hemispheres for statistical summarization.

2.6. Connectivity-based shared response model

Here we apply a variant of connectivity hyperalignment to multiple datasets with largely non-overlapping subjects, where all datasets share a common task (i.e., story-listening) and each "dataset" corresponds to a unique, naturalistic stimulus (auditory recordings of spoken stories). We use the term "hyperalignment" (via Haxby et al., 2011) to refer to the superordinate class of functional normalization algorithms that leverage commonality of function to transform subject-specific, topographic responses into a common response space. The current work develops a specific algorithm within this overarching class, which we refer to as connectivity-based shared response model (connectivity SRM or cSRM; Fig. 2). Note, however, that this algorithm differs in several ways from the core implementations of hyperalignment and connectivity hyperalignment, which, for example, use iterative

Procrustes transformations (Haxby et al., 2011; Guntupalli et al., 2016, 2018). We aim to describe the method in enough detail to provide a recipe for others. Our method was implemented using the Brain Imaging Analysis Kit (BrainIAK; https://brainiak.org), and code to perform these analyses is publicly available at https://github.com/snastase/connectivity-srm. To validate our approach, we first split each story in half; we designate the first half as the training set and the second half as the test set. All functional normalization algorithms are estimated from the training set and validated on the test set.

The following describes the cSRM algorithm for a given ROI. For each subject within a given dataset, we first estimate the functional connectivity between each vertex in the ROI and a set of connectivity targets. In the current context, because the subjects in a given dataset are all exposed to the same time-locked naturalistic story stimulus, we use ISFC to estimate stimulus-related functional connectivity and filter out idiosyncratic noise and intrinsic fluctuations (Simony et al., 2016; Nastase et al., 2019a). Leave-one-subject-out ISFCs are computed by correlating the response time series in one subject with the average response time series across the remaining subjects in the same dataset (i.e., exposed to the same stimulus). To construct intersubject connectivity targets, we first extract for each subject the regional-average response time series for 360 areas spanning the cortex derived from a multimodal surface-based cortical parcellation (Glasser et al., 2016). For a given subject, we then average the 360 response time series across all other subjects (excluding the current subject). Finally, we compute the Pearson correlation between the response time series at each vertex in the ROI for the left-out subject and the regional-average response time series for 360 connectivity targets derived from the remaining subjects. This yields an asymmetric ISFC matrix for each subject with a number of rows corresponding to the number of vertices in the ROI and 360 columns representing the connectivity targets. The logic so far may seem counterintuitive; i.e., how can poorly-aligned connectivity targets be used to “bootstrap” fine-grained alignment? This approach hinges on the assumption that the voxels within an ROI are characterized by topographic “signatures” of long-range functional connectivity (e.g., Heinzle et al., 2011; Jbabdi et al., 2013; Arcaro et al., 2015). Connectivity targets are deliberately coarse (to mitigate deficiencies of anatomical alignment), but widespread enough to afford distinct connectivity signatures for driving fine-grained alignment. Although the overall aim is similar, our method of estimating connectivity differs from the core implementation of connectivity hyperalignment by Guntupalli et al. (2018) in several ways. First, we estimate functional connectivity using ISFC rather than within-subject functional connectivity. ISFC analysis relies on a shared stimulus to effectively isolate stimulus-related connectivity and is not applicable to resting-state data. Second, Guntupalli and colleagues (2018, pp. 20–21) used finer-grained, higher-dimensional connectivity vectors. To construct connectivity targets, they applied an initial, coarse connectivity hyperalignment within each of 1,284 regularly-spaced searchlights, then extracted the top three principal component time per searchlight, yielding 3,852 target times series. These 3,852 connectivity targets were then used to drive connectivity hyperalignment for the vertices of interest. Instead we simply use the average time-series per parcel, yielding an order of magnitude fewer targets, limiting ourselves to coarser-grained, lower-dimensional connectivity vectors. This was an arbitrary decision for the sake of simplicity and computational efficiency; however, we expect that similar alternatives (e.g., using the

first principal component) would yield similar performance. Third, we use a predefined multimodal cortical parcellation to delineate connectivity targets rather than regularly-spaced searchlights agnostic to areal borders.

In contrast to time-series hyperalignment, which relies on a shared stimulus to evoke time-locked response trajectories across subjects, moving to connectivity abstracts away from the time series. Following the notation in Fig. 2, time-series hyperalignment operates on response matrices where the number of rows corresponds to the number of voxels or vertices in the ROI and the number of columns corresponds to the number of time points in the experiment. (Notation conventions vary, and sometimes the data matrix is transposed in prior publications—e.g., Haxby et al., 2011—but this is not a substantive difference). Each row corresponds to the response time series for a single voxel and each column corresponds to the distributed response pattern for a given time point. Different stimuli result in different response trajectories that cannot effectively be aligned and may yield different shared spaces (although these spaces may converge for sufficiently rich stimuli). On the other hand, connectivity hyperalignment operates on connectivity matrices where the number of columns instead corresponds to the number of connectivity targets elsewhere in the brain (Fig. 2C). In this framework, each row corresponds to the connectivity vector (a coarse whole-cortex connectivity profile) for a given “seed” voxel in the ROI; each column in the matrix corresponds to a fine-grained, spatially distributed pattern of connectivities across voxels in the ROI relative to a given connectivity target. Critically, the shape of the connectivity matrices is dictated by the number of connectivity targets, not the number of time points in a stimulus. These connectivity matrices are isomorphic across stimuli and can be more readily aggregated than disparate response trajectories. The “second-order isomorphism” of connectivity matrices allows us to aggregate (i.e., stack) ISFC matrices across both subjects and story stimuli.

The goal of hyperalignment is then to leverage commonality of function to find a set of transformations (e.g., rotations) that map each subject’s idiosyncratic voxel response space into an abstract, shared response space. The SRM implementation used here frames this in terms of matrix factorization (Chen et al., 2015; Anderson et al., 2016), where the aggregate data matrix across subjects is decomposed into a reduced-dimension shared space and a set of subject-specific topographic transformation matrices. Applying this algorithm to connectivity matrices yields a shared connectivity space. Intuitively, it may seem that the resulting subject-specific transformations are only suitable for aligning connectivity data—but this is not the case (Guntupalli et al., 2018). Importantly, these subject-specific matrices are topographic transformations and can be used to project response time series into the shared space—assuming the connectivity matrices capture sufficient functional commonality to effectively align response trajectories. Concretely, we applied the SRM to the ISFC matrices derived from the training data (first half of each story), resulting in a shared connectivity space and subject-specific transformations. We then used these transformations to project response time series from the test set (second half of each story) into the shared space for evaluation. Note that the SRM algorithm can differ in performance relative to, e.g., the iterative Procrustes transformations used by Guntupalli and colleagues (2018; see, e.g., Chen et al., 2015, for comparisons among algorithms).

For subjects participating in multiple datasets, we computed their connectivity matrices separately per dataset, then averaged these prior to estimating the SRM. This ensures that each subject only submits one connectivity matrix to the SRM and only receives one transformation matrix into shared space. Note that this is not strictly necessary; multiple connectivity matrices could be submitted for a single subject participating in multiple datasets, yielding multiple dataset-specific transformations. However, this would also bias the shared space toward subjects contributing multiple connectivity matrices, so we did not explore this alternative here. Although the majority of subjects (56%) only participated in one scan (i.e., story), more densely sampled subjects may contribute more robust connectivity estimates due to this averaging procedure.

We can also exclude a subset of stories and subjects when estimating the shared space. First, we estimate a shared connectivity space based on the training data from a subset of stories. In the schematic depicted in Fig. 2, this corresponds to excluding, e.g., “dataset 1” (green) entirely when estimating the shared space; i.e., the shared space is estimated from datasets (e.g., “dataset 2”, orange) comprising stimuli and subjects not included in dataset 1. Subsequently, we compute connectivity matrices on the training data for a set of left-out subjects and stories not used to estimate the shared space. Given the preexisting shared connectivity space and a connectivity matrix for a given left-out subject, we can solve for that subject’s topographic transformation into the predefined shared space (as described in Chen et al., 2015). We can then use this transformation to project the left-out subject’s test data into shared space. That is, in Fig. 2, we use the training half of dataset 1 (light green) to define a transformation into the preexisting shared space, and project the test half of dataset 1 (dark green) into this independent shared space. Projecting left-out subjects into a predefined shared space in this manner does not alter the shared space (or the preexisting transformations for subjects used to estimate the shared space). Note that in the context of connectivity SRM, the training data used to project left-out subjects into shared space comprise connectivity matrices that are not strictly tied to a given stimulus; this allows novel subjects listening to novel stories to be transformed into an independent, predefined shared space.

Although we focus on defining a single shared connectivity space across datasets, we can also apply cSRM separately to each story dataset in isolation to create story-specific shared spaces. We directly compare the single connectivity-based shared space defined across all stories to story-specific connectivity-based shared spaces defined separately for each dataset. We also compare cSRM to conventional within-story time-series hyperalignment using the analogous SRM implementation (tSRM; Chen et al., 2015). SRM yields a reduced-dimension shared space at a specified dimensionality of k shared features; in several cases we compare shared spaces at varying dimensionality. To control for uninteresting effects of dimensionality reduction with SRM, we also performed principal components analysis (PCA) at matched dimensionality (as in, e.g., Chen et al., 2015). PCA imposes an orthogonality constraint analogous to SRM, but when applied to the aggregated subject data yields the same projection across subjects. Intuitively, PCA can be thought of as a control condition implementing similar dimensionality reduction, but without accounting for topographic idiosyncrasies across subjects. When interpreting results, cSRM performance should be compared to PCA at the matching dimensionality.

2.7. Time-segment classification

We evaluated cSRM against alternative normalization schemes using between-subject story time-segment classification (Haxby et al., 2011). This analysis measures how accurately brief spatiotemporal response trajectories corresponding to unique segments of a story can be matched across subjects. We divided the test data (second half of each story) into 10-TR (15-second) segments and concatenated the response patterns across TRs into a single spatiotemporal response vector (or response trajectory) per segment. To perform between-subject classification for a given test subject, we first averaged the response vectors for each time segment over N–1 subjects excluding the test subject. We then computed the Pearson correlation between each response vector in the test subject and the average response vectors from the remaining subjects. A given response vector in the test subject was correctly classified if it is most highly correlated with the correct average vector from the remaining subjects. This is effectively a correlation-based 1-nearest neighbor classifier with leave-one-subject-out cross-validation (Haxby, 2012). Chance accuracy is 1 over the number of time segments in the test data for a given story and varies across stories. Note that, by design, this analysis can capitalize on any information shared across subjects and is agnostic to the type of information (e.g., sensory, semantic) encoded in response trajectories.

2.8. Intersubject correlation

We used two varieties of intersubject correlation analysis to further dissect how cSRM affects spatially distributed response time series. We first examined how cSRM impacts vertex- or feature-wise intersubject time-series correlations (Hasson et al., 2004, 2010; Nastase et al., 2019a). For each subject, we computed the Pearson correlation between the response time series at each vertex or feature and the average response time series at that vertex or feature across N – 1 remaining subjects (i.e., leave-one-out temporal ISC). For each vertex or feature, we then averaged ISCs across subjects to summarize the shared signal for that vertex/feature. To evaluate how cSRM affects spatially distributed response patterns, we computed intersubject pattern correlations at each time point (Chen et al., 2017; Zadbood et al., 2017; Nastase et al., 2019a). More specifically, for each subject, we computed the Pearson correlation between the response pattern at each TR and the average response pattern for N – 1 remaining subjects, then averaged these correlations across all time points per subject (i.e., leave-one-out spatial ISCs). We summarized these values by averaging the resulting ISCs across time points. Note that cSRM could conceivably produce the same degenerate response pattern across all time points and still yield high spatial ISCs, despite effectively discarding the stimulus-specific information of interest. However, this would be inconsistent with high time-segment classification accuracies; therefore, we consider spatial ISCs as a view into the observed time-segment classification performance rather than a benchmark in isolation. Evaluating the efficacy of cSRM using ISCs may seem “circular” at first, but there are two reasons why this is not the case: (*a*) cSRM optimizes intersubject similarity of connectivity matrices, not response trajectories; (*b*) the shared space is estimated from a subset of training data (the first half of each story) and the ISCs are evaluated on a separate subset of test data (the second half of each story).

2.9. Semantic encoding model

We also evaluated how cSRM impacts model-based encoding and decoding for two stories (Fig. 3; Güçlü and van Gerven, 2017; Vodrahalli et al., 2017; Van Uden et al., 2018; Wen et al., 2018). To quantify the semantic content of the stories, we first used a semi-supervised forced-alignment algorithm (Yuan and Liberman, 2008) to extract time-stamped transcripts from each story stimulus (see Fig. 2 for an example). We then assigned semantic vectors to each word from the 300-dimensional word2vec embedding space trained on ~100 billion words from the Google News corpus (Mikolov et al., 2013). More semantically similar words are located nearer to each other in this vector space; that is, they have more similar word embeddings (Turney et al., 2010). For each TR, all words with onsets occurring within that TR were assigned to the TR, and any words spanning two TRs were assigned to both. For TRs containing multiple words, we simply averaged the corresponding word embeddings to produce a single semantic vector per TR (cf. Vodrahalli et al., 2017). TRs in which no words occurred were assigned zero vectors. To account for varying hemodynamic lag, the 300-dimensional model was concatenated at delays of 2, 3, 4, and 5 TRs (3.0, 4.5, 6.0, 7.5 s), yielding a 1200-dimensional vector per TR (similarly to Huth et al., 2016). Delays were applied separately for the training and test data so as to avoid leakage across the train-test boundary.

To assess whether semantic information encoded in the embedding space captures variability in brain activity, we used a forward encoding model (Mitchell et al., 2008; Wehbe et al., 2014; Huth et al., 2016; Pereira et al., 2018). Following work by Huth et al. (2016), we used ridge regression to estimate coefficients (weights) for these 1200 semantic model features. *L2*-regularized linear regression effectively imposes a prior on the feature weights: the identity matrix scaled by a ridge coefficient (Diedrichsen and Kriegeskorte, 2017). Ideally, the optimal ridge coefficient is selected using nested cross-validation. However, this is computationally intensive, and will tend to yield different ridge coefficients for each subject, voxel or vertex, and cross-validation fold, complicating model comparison. For example, Huth et al. (2016) used a resampling approach to find optimal ridge coefficients, then averaged these across voxels and subjects to arrive at a single, consensus ridge coefficient (183.3 in that case). Here, to simplify numerous model comparisons, we use an arbitrary ridge coefficient of 100 throughout. This of course handicaps the absolute performance of our model. Our goal is not to engineer a novel or high-performing encoding model (see, e.g., Huth et al., 2016; Pereira et al., 2018), but to explore how functional normalization algorithms such as cSRM impact model performance under simplifying assumptions. We are interested in the encoding model insofar as it can provide insights into the performance of normalization algorithms.

Ridge regression was used to estimate weights so as to best predict the response time series at each vertex (or feature) in the training data (implemented using scikit-learn; Pedregosa et al., 2011). In the case of functional normalization (e.g., cSRM), note that we first estimated transformations into shared space from the training data. In order to fit the semantic encoding model, we used these transformations to project the training data (from which the transformations were derived) into shared space. That is, both the SRM transformations and the semantic model weights were estimated from the training data (first half of each story),

affording unbiased validation on the test data (second half of each story). We focus on leave-one-subject-out cross-validation to evaluate the semantic encoding models: for each cross-validation fold, regression weights were estimated on the training data for $N-1$ subjects, and evaluated on the test data for the left-out subject. In the current work, we average the training data across the $N-1$ subjects in shared space before training; however, these data could be concatenated instead.

The regression weights estimated from the training data can then be used to predict response time series from the semantic vectors in a left-out subject. This approach—predicting vertex-wise response time series from the semantic model—is referred to as "forward encoding". To evaluate the quality of these predictions, we computed the Pearson correlation between the predicted response time series and the actual response time series. We also perform a model-based decoding analysis, referred to as an "inverted encoding model," to predict semantic vectors from response patterns in the test set (Thirion et al., 2006; Brouwer and Heeger, 2009; Sprague et al., 2018; Gardner and Liu, 2019). Note that this approach differs from generic classification analyses (e.g., Norman et al., 2006; Haxby, 2012; see Naselaris and Kay, 2015), which do not specify an explicit feature space for decoding. We first averaged coefficients across the four delays to obtain a single 300-dimensional weight matrix (as in Huth et al., 2012). We then computed the pseudo-inverse of the weight matrix comprising all vertices or features in an ROI. We multiplied the response pattern at each time point by this inverted weight matrix, resulting in a predicted semantic vector for each time point. We evaluated the quality of these predictions by computing the Pearson correlation between the predicted semantic vector for each time point and the actual semantic vectors for all time points in the test set. We then assess the rank of the correct semantic vector in the test set and normalize this by the number of semantic vectors in the test set to obtain a normalized rank accuracy score (Pereira et al., 2018). We average these rank accuracies across all time points in the test set. The rank accuracy score ranges from 0 to 1 where a score of 1 indicates that the correct semantic vector was the most similar to the predicted semantic vector and thus the highest-ranked vector. If there were no systematic relationship between predicted and actual semantic vectors, this would yield a chance rank accuracy score of approximately 0.5.

3. Results

3.1. Time-segment classification

We evaluated functional normalization algorithms in terms of between-subject story time-segment classification (Haxby et al., 2011). We divided the test data into 10-TR (15-second) response trajectories and supplied these to a between-subject correlation-based classifier (chance is 1 over the number of time segments for a given story). We first assessed between-subject time-segment classification in AAC across all 10 story stimuli (see Fig. 4). We compared classification performance for several implementations of functional normalization against the anatomically normalized data ("no SRM"), including time-series SRM defined within each story, connectivity SRM defined within each story, and connectivity SRM defined across all stories. For this representative example, we fit the SRMs with $k = 100$ shared features. We estimated 95% bootstrap confidence intervals

surrounding the mean classification accuracy across left-out subjects and hemispheres by resampling subjects with replacement. We avoid performing gratuitous null-hypothesis statistical tests, but note that, considered in isolation, cases in which the 95% confidence interval for the mean of one condition does not cross the mean of another imply statistically significant differences (at $p < .05$; Nakagawa and Cuthill, 2007). The visual depiction of confidence intervals does not account for the within-subjects design within a story, and is therefore more conservative than a paired test (Loftus and Masson, 1994).

In general, all functional normalization algorithms provided considerable gains in time-segment classification over surface-based anatomical normalization. In most cases, the connectivity SRM was comparable to time-series SRM. Defining a connectivity-based shared space across all stories yielded comparable results to cSRMs defined separately for each story. In some cases—for example the “Pie Man” and “Running from the Bronx” stimuli recorded at PNI—the cSRM defined across stories provided marked improvement over other functional normalization algorithms. In general, classification performance on average improved from 40.3% with anatomical alignment to 63.7% with cSRM defined across all stories; tSRM and within-story cSRM yielded summary accuracies of 59.8% and 58.1% respectively.

Finally, we estimated a separate connectivity space across all stories excluding the four most recently collected stories (and the constituent subjects). We then computed connectivity matrices from the training data for each left-out subject in the left-out stories “I Knew You Were Black” and “The Man Who Forgot Ray Bradbury.” We excluded the “Pie Man (PNI)” and “Running from the Bronx (PNI)” stories from model estimation to rule out the effect of shared subjects. We also excluded the “Pie Man (PNI)” and “Running from the Bronx” stories from model evaluation due to the relatively low audio quality of these stories. Given the predefined shared space and a connectivity matrix for each subject, we can derive a transformation for each left-out subject into the preexisting shared space (Chen et al., 2015). Projecting the test data for left-out subjects into this independent shared space yielded comparable improvements in accuracy over anatomical alignment (from 39.0% with anatomical alignment to 65.3% with the independent cSRM). This suggests that (*a*) the shared space generalizes to novel subjects and stimuli, and (*b*) the connectivity estimates for the training half of these left-out stories are sufficient for aligning these data to the shared space. In addition to comprising a completely non-overlapping sample of subjects viewing a different stimulus, these data were collected on a different scanner model using a different acquisition sequence.

We next assessed time-segment classification for two example stories (“I Knew You Were Black” and “The Man Who Forgot Ray Bradbury”) using cSRM at varying dimensionality across all four ROIs (Fig. 5). We compared classification performance for cSRM at dimensionalities $k = 100$, 50, and 10 shared features to anatomical normalization and PCA at matching dimensionality. PCA provides a control for the dimensionality reduction of SRM without resolving functional topographies across subjects. In EAC, cSRM afforded minimal improvements over anatomical alignment and only at low dimensionalities (from 35.3% to 42.5% at the best-performing $k = 10$ across both stories; chance ≈ 3.7%). However, in AAC and TPOJ, cSRM markedly improved classification performance over anatomical alignment:

from 39.0% to 70.2% at the best-performing $k = 50$ in AAC; and from 22.9% to 48.5% at the best-performing $k = 100$ in TPOJ. Classification performance was only modestly improved in PMC (from 16.1% to 21.9% at $k = 50$). The PCA control analysis indicates that decreasing dimensionality biases time-segment classification performance upward, but that reduced dimensionality alone cannot account for the improvement due to cSRM. Furthermore, the consistent pattern of increasing performance with decreasing dimensionality for PCA suggests—particularly for AAC and TPOJ, which do not perform best at lowest dimensionality—that a higher-dimensional shared space (e.g., $k = 100, 50$) may encode information lost at lower dimensionality (e.g., $k = 10$) for some ROIs.

3.2. Intersubject correlation

The connectivity SRM maps each subject's idiosyncratic responses into a shared space maximizing intersubject alignment of connectivity matrices—it does not explicitly optimize intersubject response time-series or pattern correlations. However, improvements in between-subject time-segment classification suggest that connectivity-based normalization in fact does implicitly align temporally-specific responses. To better understand this effect, we first examined how cSRM impacts response time series. For each subject, we computed vertex- or feature-wise intersubject time-series correlations (Hasson et al., 2004; Nastase et al., 2019a). We compared temporal ISCs at each vertex with anatomical alignment alone, with PCA to control for dimensionality reduction in SRM, and with cSRM at dimensionalities $k = 100, 50$, and 10. This comparison is not straightforward: the anatomically-aligned ROIs contain many hundreds of correlated features; on the other hand, cSRM (or PCA) reduces this feature space to fewer dimensions accounting for orthogonal (uncorrelated) components of response variance (each feature reflects a distributed response topography across the entire ROI).

We first visualized the average ISC across subjects for every vertex or feature in the ROI for each hemisphere (Fig. 6A). This reveals, for example, that cSRM and PCA isolate very few features in EAC (~1 in each hemisphere) that capture the vast majority of the shared signal across subjects. We did not observe large differences in the distribution temporal ISCs across hemispheres (Fig. S2). In downstream ROIs, however, cSRM yields a considerably larger number of dimensions capturing shared signal than PCA. This implies that functional alignment reveals higher-dimensional shared information across subjects. To more intuitively visualize this, we computed the number of features in each ROI exceeding an arbitrary ISC threshold of $r > 0.1$ (Fig. 6B). Note that in the reduced-dimension spaces, the absolute number of features exceeding this threshold is limited by the specified number of features k. We considered visualizing instead the proportion of features exceeding threshold relative to the maximum possible number of features k, but this obscures the fact that in most cases cSRM at, e.g., $k = 100$ yields several times the absolute number features exceeding threshold as cSRM at $k = 10$; although the proportion of features exceeding threshold increases at lower values of k, the absolute number of features exceeding threshold is considerably higher at higher values of k. These features represent largely orthogonal, non-redundant components of the response and a greater absolute number of features reflects higher-dimensional information shared across subjects. In AAC, for example, cSRM at $k = 100$ yields on average 79 features with ISC exceeding 0.1, while PCA at $k = 100$

yields only 21. Similarly, in TPOJ, cSRM at $k = 100$ increases the number of features exceeding this threshold from 15 to 66. In PMC, cSRM at $k = 100$ increases the number of features exceeding threshold from 12 to 50. Even in EAC, cSRM at $k = 100$ yields over twice the number of features with ISCs exceeding 0.1 as PCA at matched dimensionality (39 and 16 features, respectively). We can infer that these numerous, largely orthogonal features with moderate ISCs encode high-dimensional information about the stimulus, because the stimulus is what drives shared responses across subjects (Nastase et al., 2019a). Again, unlike vertices in anatomical space, these features represent distributed topographies across the ROI and capture orthogonal components of response variance. Interestingly, although we find that time-segment classification in EAC is maximal at $k = 10$, we find considerably more than 10 features with moderate ISC at $k = 50$ and 100. These features, despite having moderate ISCs, may not be informative for time-segment classification.

Can the transformations derived from connectivity SRM align spatially distributed response patterns across subjects? To provide another window into the improvement in time-segment classification, we computed intersubject pattern correlations at each time point (Chen et al., 2017; Zadbood et al., 2017; Nastase et al., 2019a). We assessed these spatial ISCs with anatomical alignment, PCA, and cSRM (Fig. 7). We found that spatial ISCs generally increased with reduced dimensionality, but that cSRM provides a boost over both anatomical alignment and PCA. In EAC, the benefit of cSRM over anatomical alignment was negligible, but exceeded PCA at matching dimensionality. In AAC and TPOJ, however, cSRM significantly improved the alignment of spatial response topographies across subjects. For example, in AAC, spatial ISCs almost doubled, from $r = .110$ with anatomical alignment ($r = .058$ with PCA) to $r = .211$ with cSRM at $k = 100$. An effect of similar relative magnitude was observed in TPOJ: from $r = .082$ with anatomical alignment ($r = .029$ with PCA) to $r = .153$ at cSRM at $k = 100$. In PMC, spatial ISCs increased from $r = .074$ ($r = .021$ with PCA) to $r = .100$ with cSRM at $k = 100$. This matches with the expectation that the SRM algorithm will enforce spatial patterns that are increasingly correlated across subjects at lower dimensionality k.

3.3. Semantic encoding model

Encoding models have been increasingly adopted as a means of testing explicit feature spaces capturing representational content ranging from low-level visual features to high-level semantic content (Naselaris et al., 2011; Serences and Saproo, 2012). However, these models are often estimated independently per subject using large volumes of data (e.g., Huth et al., 2016), which poses problems of both scalability (in terms of data collection) and generalizability across subjects (cf. Güçlü and van Gerven, 2017; Vodrahalli et al., 2017; Van Uden et al., 2018). Here we used a simplistic semantic encoding model to explore how cSRM impacts model performance. We assigned semantic word embeddings to each time point, then used ridge regression to estimate a weight matrix mapping between word embeddings and brain responses (see Fig. 3).

In the forward encoding analysis, we used the regression weights estimated from the training data to predict vertex- or feature-wise response time series in the test set for a left-out subject. To evaluate the forward model, we computed the Pearson correlation between the

predicted time series and actual time series for each vertex or feature. We compared forward encoding performance with surface-based anatomical alignment, PCA to control for dimensionality reduction in SRM, and cSRM at dimensionalities $k = 100$, 50, and 10. Although we emphasize a predictive approach using leave-one-subject-out cross-validation, we also present the typical within-subject performance (as in, e.g., Huth et al., 2016). Similarly to the temporal ISC analysis, this makes for a difficult comparison because vertices in the anatomical ROI are highly redundant (correlated), while PCA and cSRM reduce this feature space to fewer, orthogonal dimensions. We first visualized the vertex- and feature-wise performance averaged across subjects (Fig. 8A), then plotted the number of features in each ROI exceeding an arbitrary performance threshold of $r > 0.1$ (Fig. 8B). Although the maximum number of features exceeding threshold is limited by the specified k, greater absolute numbers of largely orthogonal features surpassing this threshold reflect higher-dimensional semantic information shared across subjects. In all cases, cSRM dramatically increased the number of well-predicted features relative to the dimensionality-matched PCA control. In EAC, cSRM at $k = 100$ doubled the number of features with performance exceeding $r > 0.1$ from 11 to 22. In AAC and TPOJ, cSRM at $k = 100$ roughly tripled the number of features exceeding the forward encoding performance threshold—from 16 to 49 and from 12 to 36, respectively. In PMC, the number of features exceeding this threshold increased from 9 with PCA to 23 with cSRM at $k = 100$.

We next inverted the encoding model to predict semantic vectors from spatially distributed response patterns (Thirion et al., 2006; Brouwer and Heeger, 2009; Sprague et al., 2018; Gardner and Liu, 2019). While between-subject time-segment classification can capitalize on any diagnostic information shared across subjects, between-subject model-based decoding is strictly limited to the representational content encoded in the model. For each time point in the test set, we multiplied the distributed response pattern by the inverted weight matrix to recover a predicted semantic vector for that time point. To evaluate decoding performance, we computed the Pearson correlation between the predicted semantic vector and the actual semantic vectors in the test set and summarized this using a normalized rank accuracy score (Pereira et al., 2018). We compared between-subject model-based decoding using anatomical alignment, PCA, and cSRM (Fig. 9). Although we focus on between-subject decoding using leave-one-subject-out cross-validation, we also computed within-subject decoding performance for comparison. In general, model-based decoding accuracies were low; likely due to our simplistic model estimation procedure. However, we observed some interesting trends. First, within-subject performance exceeded between-subject performance using surface-based alignment in AAC and TPOJ, suggesting that anatomical normalization alone fails to translate some semantic information across brains. Second, dimensionality reduction did not consistently increase performance. Crucially, in almost all cases, between-subject decoding performance using cSRM exceeded both between- and within-subject performance using anatomical alignment. For example, in AAC, the average rank accuracy across test subjects increased from 52.1% with anatomical alignment (54.1% for within-subject decoding) to 56.5% with cSRM at $k = 100$ (with a rank accuracy of 63.5% for the best-performing test subject).

3.4. Consensus space across stimuli

Finally, we examined the consequences of deriving a single shared space across numerous distinct stories. Is the single, connectivity-based shared space defined across all datasets notably different from shared connectivity spaces derived separately for each story? For example, we may expect that each story considered in isolation will nonetheless yield similar shared spaces due to the nature of functional connectivity. To address this, we computed ISFC matrices from the test data for each subject and projected these ISFC matrices into either (*a*) the connectivity-based shared space defined across all stories (across-story cSRM), or (*b*) the unique connectivity-based shared spaces defined separately for each story (within-story cSRM). The subjects contributing to each within-story cSRM are a strict subject of the subjects contributing to the across-story cSRM. For illustrative purposes, we used AAC and cSRM at $k = 100$. We then flattened and averaged the ISFC matrices across subjects per story in their respective shared spaces and computed the pairwise correlations between the mean ISFC matrices across stories (Fig. 10). In fact, ISFCs projected into the single, shared connectivity space are more similar across all pairs of stories than those projected into story-specific shared spaces: the average correlation of ISFC matrices increased from $r = .623$ to $r = .764$ across all pairs of stories (averaged across hemispheres). We did not observe a substantive difference between hemispheres (Fig. S3). These results are expected but serve as a useful sanity check, and suggest that the transformations defined across all stories point to a consensus space. Interestingly the magnitude of this effect may increase in higher-level ROIs, where the within-story shared connectivity spaces are more distinct (Figs. S4–8). Datasets with longer story stimuli (more TRs) and larger samples of subjects likely have a larger “pull” on the consensus space (Fig. S9).

We next considered whether constructing a shared space across distinct stimuli impacts responses within a story. Similarly to before, we projected the response time series for the test data into either a shared connectivity space defined across all stories or story-specific shared connectivity spaces (in AAC, using cSRM at $k = 100$). We then concatenated these response patterns over time into a single spatiotemporal response trajectory for each subject, and computed leave-one-subject-out ISC on these spatiotemporal vectors within each story (Fig. 11; Nastase et al., 2019a). Interestingly, we found that the response trajectories were more similar across subjects within a given story when projected into the shared space defined across stories. This increase in similarity was small, from $r = .261$ to $r = .286$, but statistically significant across all stories ($p \leq .001$ for all stories, nonparametric Wilcoxon signed-rank test, Bon-ferroni corrected).

4. Discussion

We have demonstrated that connectivity hyperalignment can be used to estimate a shared response space across disjoint datasets with unique stimuli and non-overlapping samples of subjects. Although the shared space is defined in terms of connectivity, the subject-specific topographic transformation matrices are suitable for projecting response time series into this space (Guntupalli et al., 2018). We have shown that transformations derived from intersubject functional connectivity are sufficient to precisely align response trajectories in a way that is both spatially and temporally specific; in the case of time-segment classification,

this effect was consistent across all 10 stories. The features in the shared space derived using cSRM reflect orthogonal components of response variability; projecting data into shared space dramatically increases the dimensionality of information shared across subjects. The consensus space introduced here conjoins diverse story stimuli, and effectively regularizes subject- and story-specific transformations.

There are three key concepts underlying the success of this algorithm. First, relatively coarse connectivity targets spanning cortex can provide sufficiently rich signatures of functional connectivity to drive fine-grained topographic alignment (Heinzle et al., 2011; Jbabdi et al., 2013; Arcaro et al., 2015). Here we define connectivity targets according to a multimodal cortical parcellation (Glasser et al., 2016), which yields a relatively low-dimensional, more computationally efficient connectivity space. However, there are a variety of ways to construct connectivity targets (cf. Guntupalli et al., 2018); for example, in the context of naturalistic stimuli, vertices with temporal ISCs exceeding some threshold could serve as potentially finer-grained connectivity targets. The second, related concept is that constructing a shared space based on functional connectivity yields subject-specific topographic transformations suitable for aligning response time series. This may seem counterintuitive, but, again, suggests that long-range connectivity profiles capture information about local response topographies. Third, the datasets used here, despite showcasing story stimuli with diverse topics and speakers, are all examples of the superordinate story-listening “task.” This common task may yield commensurate connectivity patterns across stimuli, allowing the algorithm to find a consensus shared space. Future work is required to explore the boundary conditions of this algorithm and determine whether a shared space can be defined across qualitatively different tasks or paradigms.

Our analyses demonstrate that projecting data into a shared space derived from functional connectivity improves semantic model-based encoding and decoding. Interestingly, between-subject semantic model-based decoding with cSRM exceeded within-subject decoding in all ROIs. How can this be possible? It may be that our simplistic semantic encoding model only captures relatively coarse-grained information across subjects. However, between-subject encoding models allow us to leverage much larger volumes of data than could be acquired in a single subject. Aggregating data across subjects can yield much cleaner training samples, thus improving performance over limited and often noisy within-subject data. Here, we fit the semantic encoding model on the averaged response time series across training subjects; although averaging yields clean response time series, subjects in the shared space could instead be concatenated, dramatically increasing the number of training samples available for modeling. Furthermore, our rudimentary fitting procedure (e.g., using an arbitrary, fixed ridge coefficient) will necessarily yield suboptimal model performance. Although we adopt this approach to reduce the computational burden and simplify model comparison, it could be argued that models are not fairly compared unless they can arrive at their own optimal ridge coefficients. For any use-case where evaluating or comparing encoding models is the primary scientific goal, we recommend grid search using nested cross-validation or resampling procedures to identify optimal hyperparameters (as in, e.g., Huth et al., 2016).

Interestingly, our functional normalization algorithm differentially benefited certain ROIs and certain stories. For example, EAC, an early sensory ROI, was only marginally improved by cSRM, while downstream association cortices such as AAC and TPOJ were more dramatically improved. On the other hand, the putatively high-level PMC was only modestly improved by cSRM. There are several possible reasons for these discrepancies. First, the quality of the cSRM derives from the richness of functional connectivity for a given ROI; early sensory areas may have limited or less-informative connectivity with the rest of the brain, whereas association cortices are in effect defined by their broad, integrative connectivity. Second, some cortical areas may have relatively stereotyped functional architecture across individuals or inherently coarse response topographies; both scenarios would reduce the benefits of functional normalization over anatomical normalization. For example, PMC may host coarser-scale response topographies that better match across subjects (Chen et al., 2017; Zadbood et al., 2017) than lower-level perceptual areas (Cox and Savoy, 2003; Haxby et al., 2011). Third, cortical areas vary in the extent to which their processing is strictly stimulus-locked; early sensory areas may be better aligned by temporal hyperalignment, while association cortices may be better aligned by connectivity hyperalignment (Guntupalli et al., 2018). Certain stories, such as “Pie Man (PNI)” and “Running from the Bronx (PNI)” seem to have received an outsized benefit from cSRM derived across stories. We suspect that there are two related reasons for this. First, these story stimuli were recorded while the speaker was undergoing an fMRI scan, resulting in relatively low audio quality. Responses to stimuli with low audio quality may see an increased benefit with cSRM derived across stimuli with higher audio quality. Second, the subjects in these studies also received the “I Knew You Were Black” and “The Man Who Forgot Ray Bradbury” stories (both of which have high audio quality), likely yielding more robust connectivity estimates.

All of the analyses reported here have focused on generalization across subjects and we have not explicitly examined individual differences among subjects (Dubois and Adolphs, 2016). Although there is a common assumption that hyperalignment methods will necessarily quash individual differences, recent work by Feilong et al. (2018) has shown that, despite increasing intersubject similarity, hyperalignment in fact preserves—and increases—the reliability of individual differences. We suspect this occurs for two related reasons: (*a*) hyperalignment effectively resolves idiosyncratic functional-anatomical mapping, thus factoring out topographic idiosyncrasies and isolating individual differences in representational geometry; and (*b*) hyperalignment reveals individual differences in fine-grained response topographies that are otherwise obscured or inaccessible with anatomical alignment. Procrustes-based hyperalignment methods are constrained to orthogonal transformations—effectively rotating and reflecting the feature space-—and therefore do not distort subject-specific response trajectories (or representational geometries). Although the dimensionality reduction of SRM will begin to distort subject-specific representational geometries at low dimensionalities, previous work has shown that SRM can be used to factor out shared signals, thus accentuating individual differences or differences due to experimental manipulations (Chen et al., 2015). We expect to see continuing developments in “neural fingerprinting” at the intersection of hyperalignment methods and naturalistic paradigms (Vanderwal et al., 2017; Feilong et al., 2018; Finn et al., 2019).

The approach described here has several limitations. We constrain our analysis to a handful of ROIs and do not provide a whole-cortex searchlight-based solution as in the core implementation of connectivity hyperalignment (Guntupalli et al., 2018). However, it would be straightforward to extend the implementation used here to parcels tiling the entire cortex. Unlike Guntupalli and colleagues, we take advantage of the shared story stimulus within each dataset and use ISFC to filter out idiosyncratic, intrinsic fluctuations and isolate stimulus-related connectivity (Simony et al., 2016). This approach, however, is not applicable to resting-state paradigms where there is no shared stimulus. Furthermore, we do not currently account for the fact that certain stories have considerably more subjects than others. In our implementation of cSRM, this will tend to bias the shared space toward the stories with the most subjects. The contribution of each story to the shared space could conceivably be normalized by the proportion of subjects for that story relative to the total number of subjects. However, in practice we may want to bias the shared space toward stories with the largest number of subjects, as these may provide the most robust shared model for limited data.

This raises the important question of whether it is worthwhile to "sacrifice" functional data for the purposes of normalization. We generally advocate for estimating a shared space on independent data to avoid circularity (Kriegeskorte et al., 2009). In the current work we estimate the shared space from the first half of all stories and use the second half for evaluation, potentially undermining generalizability across stories. However, we demonstrate that left-out stories still benefit from connectivity SRM defined on independent data. On the other hand, many analyses (e.g., model-based encoding and decoding) already require independent training data for model estimation. Here we adopt an approach where both functional normalization and encoding model weights are estimated from the same training set, effectively negating the price paid for normalization.

Precision neuroscience is fundamentally limited by the feasibility of collecting large volumes of data in experimental subjects (let alone patients). The fact that connectivity hyperalignment benefits a completely independent set of subjects and stories not used to estimate the shared space has important implications. This capacity for generalization suggests that many existing datasets, from resting-state to naturalistic movies, could benefit from connectivity hyperalignment. This provides a means for using existing data to "bootstrap" improvements in functional registration and build better, more generalizable predictive models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank James V. Haxby, Feilong Ma, Eshin Jolly, Asieh Zadbood, Qihong Lu, Kristina M. Rapuano, Christopher Baldassano, Francisco Pereira, and Erez Simony for helpful comments, as well as Janice Chen, Christopher Honey, Yaara Yeshurun, and Claire Chang for sharing previously-collected data.

Funding

This work was supported by the National Institutes of Health (R01 MH112566) and the Defense Advanced Research Projects Agency (DARPA; Brain-to-Brain Seedling contract number FA8750-18-C-0213).

References

Anderson MJ, Capota M, Turek JS, Zhu X, Willke TL, Wang Y, Chen P, Manning JR, Ramadge PJ, Norman KA, 2016. Enabling factor analysis on thousand-subject neuroimaging datasets. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 1151–1160. 10.1109/bigdata.2016.7840719.

Arcaro MJ, Honey CJ, Mruczek REB, Kastner S, Hasson U, 2015. Widespread correlation patterns of fMRI signal across visual cortex reflect eccentricity organization. eLife 4. 10.7554/elife.03952.

Avants BB, Epstein CL, Grossman M, Gee JC, 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal 12, 26–41. 10.1016/j.media.2007.06.004. [PubMed: 17659998]

Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA, 2017. Discovering event structure in continuous narrative perception and memory. Neuron 95, 709–721. 10.1016/j.neuron.2017.06.041 e5. [PubMed: 28772125]

Behzadi Y, Restom K, Liau J, Liu TT, 2007. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage 37, 90–101. 10.1016/j.neuroimage.2007.04.042. [PubMed: 17560126]

Bijsterbosch JD, Beckmann CF, Woolrich MW, Smith SM, Harrison SJ, 2019. The relationship between spatial configuration and functional connectivity of brain regions revisited. eLife 8. 10.7554/eLife.32992.

Braga RM, Buckner RL, 2017. Parallel interdigitated distributed networks within the individual estimated by intrinsic functional connectivity. Neuron 95, 457–471. 10.1016/j.neuron.2017.06.038 e5. [PubMed: 28728026]

Brett M, Johnsrude IS, Owen AM, 2002. The problem of functional localization in the human brain. Nat. Rev. Neurosci 3, 243–249. 10.1038/nrn756. [PubMed: 11994756]

Brouwer GJ, Heeger DJ, 2009. Decoding and reconstructing color from responses in human visual cortex. J. Neurosci 29, 13992–14003. 10.1523/jneurosci.3577-09.2009. [PubMed: 19890009]

Chen J, Leong YC, Honey CJ, Yong CH, Norman KA, Hasson U, 2017. Shared memories reveal shared structure in neural activity across individuals. Nat. Neurosci 20, 115–125. 10.1038/nn.4450. [PubMed: 27918531]

Chang CHC, Lazaridi C, Yeshurun Y, Norman KA, Hasson U, 2020. Relating the past with the present: information integration and segregation during ongoing narrative processing. bioRxiv. 10.1101/2020.01.16.908731.

Chen P-H, Chen J, Yeshurun Y, Hasson U, Haxby J, Ramadge PJ, 2015. A reduced-dimension fMRI shared response model. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (Eds.), Advances in Neural Information Processing Systems, vol. 28. Curran Associates, Inc., pp. 460–468. https://papers.nips.cc/paper/5855-a-reduced-dimension-fmri-shared-response-model

Chien H-YS, Honey CJ, 2020. Constructing and forgetting temporal context in the human cerebral cortex. Neuron. 10.1016/j.neuron.2020.02.013.

Coalson TS, Van Essen DC, Glasser MF, 2018. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. Proc. Natl. Acad. Sci. U.S.A 115, E6356–E6365. 10.1073/pnas.1801582115. [PubMed: 29925602]

Cox RW, 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res 29, 162–173. 10.1006/cbmr.1996.0014. [PubMed: 8812068]

Cox RW, 2012. AFNI: what a long strange trip it's been. NeuroImage 62, 743–747. 10.1016/j.neuroimage.2011.08.056. [PubMed: 21889996]

Cox DD, Savoy RL, 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270. 10.1016/S1053-8119(03)00049-1. [PubMed: 12814577]

Dale AM, Fischl B, Sereno MI, 1999. Cortical surface-based analysis: I. segmentation and surface reconstruction. NeuroImage 9, 179–194. 10.1006/nimg.1998.0395. [PubMed: 9931268]

Diedrichsen J, Kriegeskorte N, 2017. Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput. Biol 13, e1005508 10.1371/journal.pcbi.1005508. [PubMed: 28437426]

Dubois J, Adolphs R, 2016. Building a science of individual differences from fMRI. Trends Cognit. Sci 20, 425–443. 10.1016/j.tics.2016.03.014. [PubMed: 27138646]

Duncan KJ, Pattamadilok C, Knierim I, Devlin JT, 2009. Consistency and variability in functional localisers. NeuroImage 46, 1018–1026. 10.1016/j.neuroimage.2009.03.014. [PubMed: 19289173]

DuPre E, Hanke M, Poline J-B, 2019. Nature abhors a paywall: how open science can realize the potential of naturalistic stimuli. NeuroImage 116330. 10.1016/j.neuroimage.2019.116330. [PubMed: 31704292]

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, Kent JD, Goncalves M, DuPre E, Snyder M, Oya H, Ghosh SS, Wright J, Durnez J, Poldrack RA, Gorgolewski KJ, 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16, 111–116. 10.1038/s41592-018-0235-4. [PubMed: 30532080]

Feilong M, Nastase SA, Guntupalli JS, Haxby JV, 2018. Reliable individual differences in fine-grained cortical functional architecture. NeuroImage 183, 375–386. 10.1016/j.neuroimage.2018.08.029. [PubMed: 30118870]

Finn ES, Glerean E, Khojandi AY, Nielson D, Molfese PJ, Handwerker DA, Bandettini PA, 2020. Idiosynchrony: from shared responses to individual differences during naturalistic neuroimaging. NeuroImage 215, 116828. 10.1016/j.neuroimage.2020.116828. [PubMed: 32276065]

Fischl B, Sereno MI, Tootell RB, Dale AM, 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp 8, 272–284. 10.1002/(sici)1097-0193(1999)8:4<272::aid-hbm10>3.0.co;2-4. [PubMed: 10619420]

Frost MA, Goebel R, 2012. Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59, 1369–1381. 10.1016/j.neuroimage.2011.08.035. [PubMed: 21875671]

Gabrieli JDE, Ghosh SS, Whitfield-Gabrieli S, 2015. Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron 85, 11–26. 10.1016/j.neuron.2014.10.047. [PubMed: 25569345]

Gardner JL, Liu T, 2019. Inverted encoding models reconstruct an arbitrary model response, not the stimulus. eNeuro 6. 10.1523/eneuro.0363-18.2019.

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC, 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. 10.1038/nature18933. [PubMed: 27437579]

Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, Ortega M, Hoyt-Drazen C, Gratton C, Sun H, Hampton JM, Coalson RS, Nguyen AL, McDermott KB, Shimony JS, Snyder AZ, Schlaggar BL, Petersen SE, Nelson SM, Dosenbach NUF, 2017. Precision functional mapping of individual human brains. Neuron 95, 791–807. 10.1016/j.neuron.2017.07.011 e7. [PubMed: 28757305]

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS, 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Front. Neuroinf 5, 13. 10.3389/fninf.2011.00013.

Greve DN, Fischl B, 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48, 63–72. 10.1016/j.neuroimage.2009.06.060. [PubMed: 19573611]

Güçlü U, van Gerven MAJ, 2017. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage 145, 329–336. 10.1016/j.neuroimage.2015.12.036. [PubMed: 26724778]

Guntupalli JS, Feilong M, Haxby JV, 2018. A computational model of shared fine-scale structure in the human connectome. PLoS Comput. Biol 14, e1006120 10.1371/journal.pcbi.1006120. [PubMed: 29664910]

Guntupalli JS, Hanke M, Halchenko YO, Connolly AC, Ramadge PJ, Haxby JV, 2016. A model of representational spaces in human cortex. Cerebr. Cortex 26, 2919–2934. 10.1093/cercor/bhw068.

Hanke M, Adelhöfer N, Kottke D, Iacovella V, Sengupta A, Kaule FR, Nigbur R, Waite AQ, Baumgartner F, Stadler J, 2016. A studyforrest extension, simultaneous fMRI and eye gaze

recordings during prolonged natural stimulation. Sci. Data 3, 160092. 10.1038/sdata.2016.92. [PubMed: 27779621]

Hanke M, Baumgartner FJ, Ibe P, Kaule FR, 2014. A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie. Sci. Data 10.1038/sdata.2014.3.

Hasson U, Malach R, Heeger DJ, 2010. Reliability of cortical activity during natural stimulation. Trends Cognit. Sci 14, 40–48. 10.1016/j.tics.2009.10.011. [PubMed: 20004608]

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R, 2004. Intersubject synchronization of cortical activity during natural vision. Science 303, 1634–1640. 10.1126/science.1089506. [PubMed: 15016991]

Haxby JV, 2012. Multivariate pattern analysis of fMRI: the early beginnings. NeuroImage 62, 852–855. 10.1016/j.neuroimage.2012.03.016. [PubMed: 22425670]

Haxby JV, Connolly AC, Guntupalli JS, 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci 37, 435–456. 10.1146/annurev-neuro-062012-170325. [PubMed: 25002277]

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ, 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416. 10.1016/j.neuron.2011.08.026. [PubMed: 22017997]

Heinzle J, Kahnt T, Haynes J-D, 2011. Topographically specific functional connectivity between visual field maps in the human brain. NeuroImage 56, 1426–1436. 10.1016/j.neuroimage.2011.02.077. [PubMed: 21376818]

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL, 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. 10.1038/nature17637. [PubMed: 27121839]

Huth AG, Nishimoto S, Vu AT, Gallant JL, 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224. 10.1016/j.neuron.2012.10.014. [PubMed: 23259955]

Jbabdi S, Sotiropoulos SN, Behrens TE, 2013. The topographic connectome. Curr. Opin. Neurobiol 23, 207–215. 10.1016/j.conb.2012.12.004. [PubMed: 23298689]

Jenkinson M, Bannister P, Brady M, Smith S, 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841. 10.1006/nimg.2002.1132. [PubMed: 12377157]

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM, 2012. FSL. NeuroImage 62, 782–790. 10.1016/j.neuroimage.2011.09.015. [PubMed: 21979382]

Klein A, Ghosh SS, Avants B, Yeo BTT, Fischl B, Ardekani B, Gee JC, Mann JJ, Parsey RV, 2010. Evaluation of volume-based and surface-based brain image registration methods. NeuroImage 51, 214–220. 10.1016/j.neuroimage.2010.01.091. [PubMed: 20123029]

Kong R, Li J, Orban C, Sabuncu MR, Liu H, Schaefer A, Sun N, Zuo X-N, Holmes AJ, Eickhoff SB, Yeo BTT, 2019. Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. Cerebr. Cortex 29, 2533–2551. 10.1093/cercor/bhy123.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI, 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci 12, 535–540. 10.1038/nn.2303. [PubMed: 19396166]

Langs G, Wang D, Golland P, Mueller S, Pan R, Sabuncu MR, Sun W, Li K, Liu H, 2016. Identifying shared brain networks in individuals by decoupling functional and anatomical variability. Cerebr. Cortex 26, 4004–4014. 10.1093/cercor/bhv189.

Lerner Y, Honey CJ, Silbert LJ, Hasson U, 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci 31, 2906–2915. 10.1523/jneurosci.3684-10.2011. [PubMed: 21414912]

Lindquist MA, Geuter S, Wager TD, Caffo BS, 2019. Modular preprocessing pipelines can reintroduce artifacts into fMRI data. Hum. Brain Mapp 40, 2358–2376. 10.1002/hbm.24528. [PubMed: 30666750]

Lin X, Sur I, Nastase SA, Divakaran A, Hasson U, Amer MR, 2019. Data-efficient mutual information neural estimator. arXiv. https://arxiv.org/abs/1905.03319.

Loftus GR, Masson ME, 1994. Using confidence intervals in within-subject designs. Psychon. Bull. Rev 1, 476–490. 10.3758/bf03210951. [PubMed: 24203555]

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J, 2013. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (Eds.), Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Inc., pp. 3111–3119. https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality

Milham MP, Craddock RC, Son JJ, Fleischmann M, Clucas J, Xu H, Koo B, Krishnakumar A, Biswal BB, Castellanos FX, Colcombe S, Di Martino A, Zuo X-N, Klein A, 2018. Assessment of the impact of shared brain imaging data on the scientific literature. Nat. Commun 9, 2818. 10.1038/s41467-018-04976-1. [PubMed: 30026557]

Mills K, 2016. HCP-MMP1.0 projected on fsaverage. 10.6084/m9.figshare.3498446.v2.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA, 2008. Predicting human brain activity associated with the meanings of nouns. Science 320, 1191–1195. 10.1126/science.1152876. [PubMed: 18511683]

Nakagawa S, Cuthill IC, 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol. Rev. Camb. Phil. Soc 82, 591–605. 10.1111/j.1469-185x.2007.00027.x.

Naselaris T, Kay KN, 2015. Resolving ambiguities of MVPA using explicit models of representation. Trends Cognit. Sci 19, 551–554. 10.1016/j.tics.2015.07.005. [PubMed: 26412094]

Naselaris T, Kay KN, Nishimoto S, Gallant JL, 2011. Encoding and decoding in fMRI. NeuroImage 56, 400–410. 10.1016/j.neuroimage.2010.07.073. [PubMed: 20691790]

Nastase SA, Gazzola V, Hasson U, Keysers C, 2019. Measuring shared responses across subjects using intersubject correlation. Soc. Cognit. Affect Neurosci 14, 667–685. 10.1093/scan/nsz037. [PubMed: 31099394]

Nastase SA, Liu Y-F, Hillman H, Zadbood A, Hasenfratz L, Keshavarzian N, Chen J, Honey CJ, Yeshurun Y, Regev M, Nguyen M, Chang CHC, Baldassano C, Lositsky O, Simony E, Chow MA, Leong YC, Brooks PP, Micciche E, Choe G, Goldstein A, Halchenko YO, Norman KA, Hasson U, 2019b. Narratives: fMRI data for evaluating models of naturalistic language comprehension. 10.18112/openneuro.ds002345.v1.0.0.

Norman KA, Polyn SM, Detre GJ, Haxby JV, 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cognit. Sci 10, 424–430. 10.1016/j.tics.2006.07.005. [PubMed: 16899397]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Others, 2011. scikit-learn: machine learning in Python. J. Mach. Learn. Res 12, 2825–2830. http://www.jmlr.org/papers/v12/pedregosa11a.

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E, 2018. Toward a universal decoder of linguistic meaning from brain activation. Nat. Commun 9, 963. 10.1038/s41467-018-03068-4. [PubMed: 29511192]

Poldrack RA, Gorgolewski KJ, 2014. Making big data open: data sharing in neuroimaging. Nat. Neurosci 17, 1510–1517. 10.1038/nn.3818. [PubMed: 25349916]

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE, 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage 84, 320–341. 10.1016/j.neuroimage.2013.08.048. [PubMed: 23994314]

Serences JT, Saproo S, 2012. Computational advances towards linking BOLD and behavior. Neuropsychologia 50, 435–446. 10.1016/j.neuropsychologia.2011.07.013. [PubMed: 21840553]

Simony E, Honey CJ, Chen J, Lositsky O, Yeshurun Y, Wiesel A, Hasson U, 2016. Dynamic reconfiguration of the default mode network during narrative comprehension. Nat. Commun 7, 12141. 10.1038/ncomms12141. [PubMed: 27424918]

Sprague TC, Adam KCS, Foster JJ, Rahmati M, Sutterer DW, Vo VA, 2018. Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. eNeuro 5. 10.1523/eneuro.0098-18.2018.

Taylor JR, Williams N, Cusack R, Auer T, Shafto MA, Dixon M, Tyler LK, Cam-Can, Henson RN, 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural

and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage 144, 262–269. 10.1016/j.neuroimage.2015.09.018. [PubMed: 26375206]

Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline J-B, Lebihan D, Dehaene S, 2006. Inverse retinotopy: inferring the visual content of images from brain activation patterns. NeuroImage 33, 1104–1116. 10.1016/j.neuroimage.2006.06.062. [PubMed: 17029988]

Treiber JM, White NS, Steed TC, Bartsch H, Holland D, Farid N, McDonald CR, Carter BS, Dale AM, Chen CC, 2016. Characterization and correction of geometric distortions in 814 diffusion weighted images. PLoS One 11, e0152472. 10.1371/journal.pone.0152472. [PubMed: 27027775]

Turney PD, Pantel P, Others, 2010. From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res 37, 141–188. 10.1613/jair.2934.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, Gee JC, 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imag 29, 1310–1320. 10.1109/TMI.2010.2046908.

Vanderwal T, Eilbott J, Finn ES, Craddock RC, Turnbull A, Castellanos FX, 2017. Individual differences in functional connectivity during naturalistic viewing conditions. NeuroImage 157, 521–530. 10.1016/j.neuroimage.2017.06.027. [PubMed: 28625875]

Van Uden CE, Nastase SA, Connolly AC, Feilong M, Hansen I, Gobbini MI, Haxby JV, 2018. Modeling semantic encoding in a common neural representational space. Front. Neurosci 12, 437. 10.3389/fnins.2018.00437. [PubMed: 30042652]

Vodrahalli K, Chen P-H, Liang Y, Baldassano C, Chen J, Yong E, Honey C, Hasson U, Ramadge P, Norman KA, Arora S, 2017. Mapping between fMRI responses to movies and their natural language annotations. NeuroImage 180, 223–231. 10.1016/j.neuroimage.2017.06.042. [PubMed: 28648889]

Wang S, Peterson DJ, Gatenby JC, Li W, Grabowski TJ, Madhyastha TM, 2017. Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. Front. Neuroinf 11, 17. 10.3389/fninf.2017.00017.

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A, Mitchell T, 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PLoS One 9, e112575. 10.1371/journal.pone.0112575. [PubMed: 25426840]

Wen H, Shi J, Chen W, Liu Z, 2018. Transferring and generalizing deep-learning-based neural encoding models across subjects. NeuroImage 176, 152–163. 10.1016/j.neuroimage.2018.04.053. [PubMed: 29705690]

Woo C-W, Chang LJ, Lindquist MA, Wager TD, 2017. Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci 20, 365–377. 10.1038/nn.4478. [PubMed: 28230847]

Yamashita A, Yahata N, Itahashi T, Lisi G, Yamada T, Ichikawa N, Takamura M, Yoshihara Y, Kunimatsu A, Okada N, Yamagata H, Matsuo K, Hashimoto R, Okada G, Sakai Y, Morimoto J, Narumoto J, Shimada Y, Kasai K, Kato N, Takahashi H, Okamoto Y, Tanaka SC, Kawato M, Yamashita O, Imamizu H, 2019. Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. PLoS Biol. 17, e3000042 10.1371/journal.pbio.3000042. [PubMed: 30998673]

Yeshurun Y, Nguyen M, Hasson U, 2017a. Amplification of local changes along the timescale processing hierarchy. Proc. Natl. Acad. Sci. U.S.A 114, 9475–9480. 10.1073/pnas.1701652114. [PubMed: 28811367]

Yeshurun Y, Swanson S, Simony E, Chen J, Lazaridi C, Honey CJ, Hasson U, 2017b. Same story, different story: the neural representation of interpretive frameworks. Psychol. Sci 28, 307–319. 10.1177/0956797616682029. [PubMed: 28099068]

Yuan J, Liberman M, 2008. Speaker identification on the SCOTUS corpus. J. Acoust. Soc. Am 123, 3878. http://www.ling.upenn.edu/~jiahong/publications/c09.pdf.

Zadbood A, Chen J, Leong YC, Norman KA, Hasson U, 2017. How we transmit memories to other brains: constructing shared neural representations via communication. Cerebr. Cortex 27, 4988–5000. 10.1093/cercor/bhx202.

Zhang Y, Brady M, Smith S, 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imag 20, 45–57. 10.1109/42.906424.

Zhen Z, Kong X-Z, Huang L, Yang Z, Wang X, Hao X, Huang T, Song Y, Liu J, 2017. Quantifying the variability of scene-selective regions: interindividual, interhemispheric, and sex differences. Hum. Brain Mapp 38, 2260–2275. 10.1002/hbm.23519. [PubMed: 28117508]

Zhen Z, Yang Z, Huang L, Kong X-Z, Wang X, Dang X, Huang Y, Song Y, Liu J, 2015. Quantifying interindividual variability and asymmetry of face-selective regions: a probabilistic functional atlas. NeuroImage 113, 13–25. 10.1016/j.neuroimage.2015.03.010. [PubMed: 25772668]

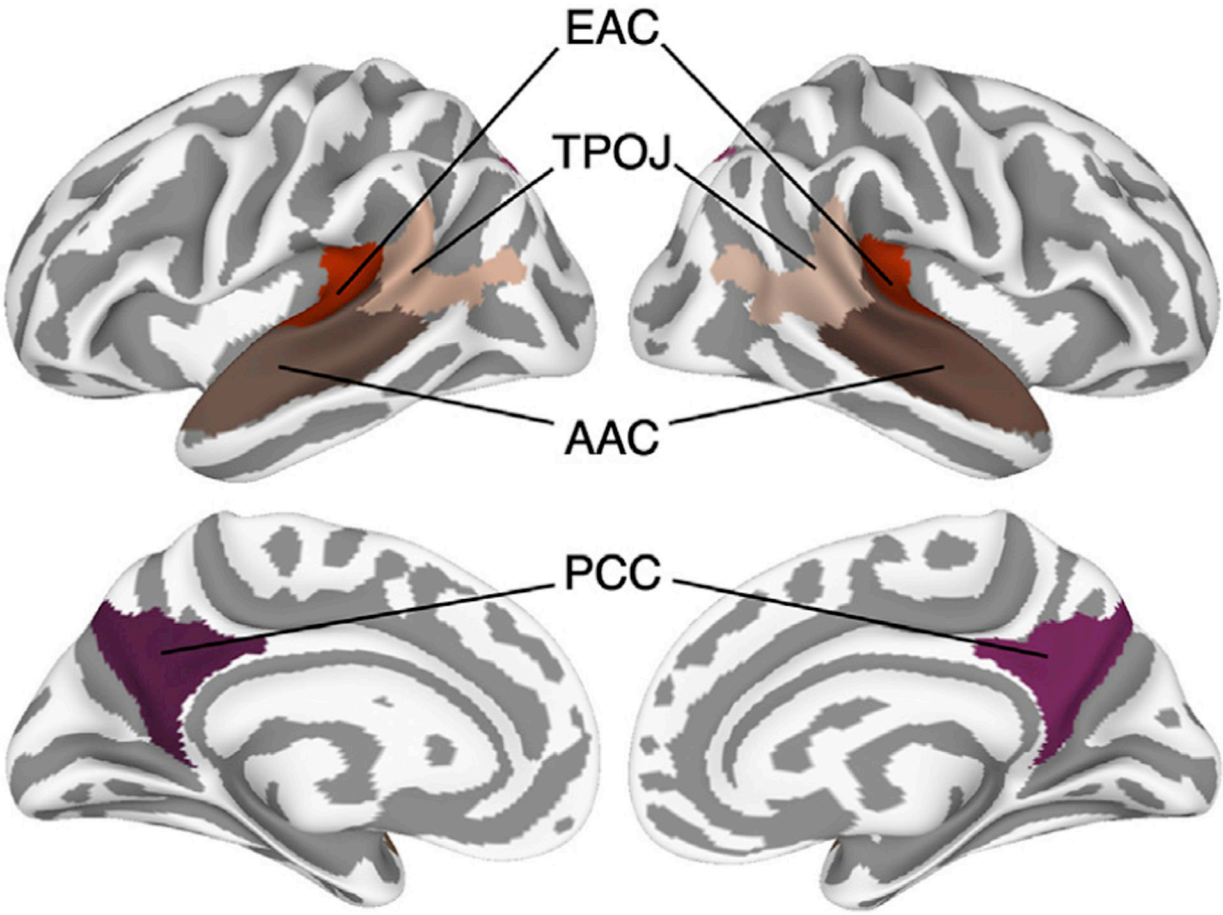

Fig. 1.

Regions of interest. Four large cortical regions roughly capturing the processing hierarchy for language and narrative comprehension were defined according to a multimodal parcellation (Glasser et al., 2016): early auditory cortex (EAC), auditory association cortex (AAC), temporo-parieto-occipital junction (TPOJ), and posterior medial cortex (PMC).

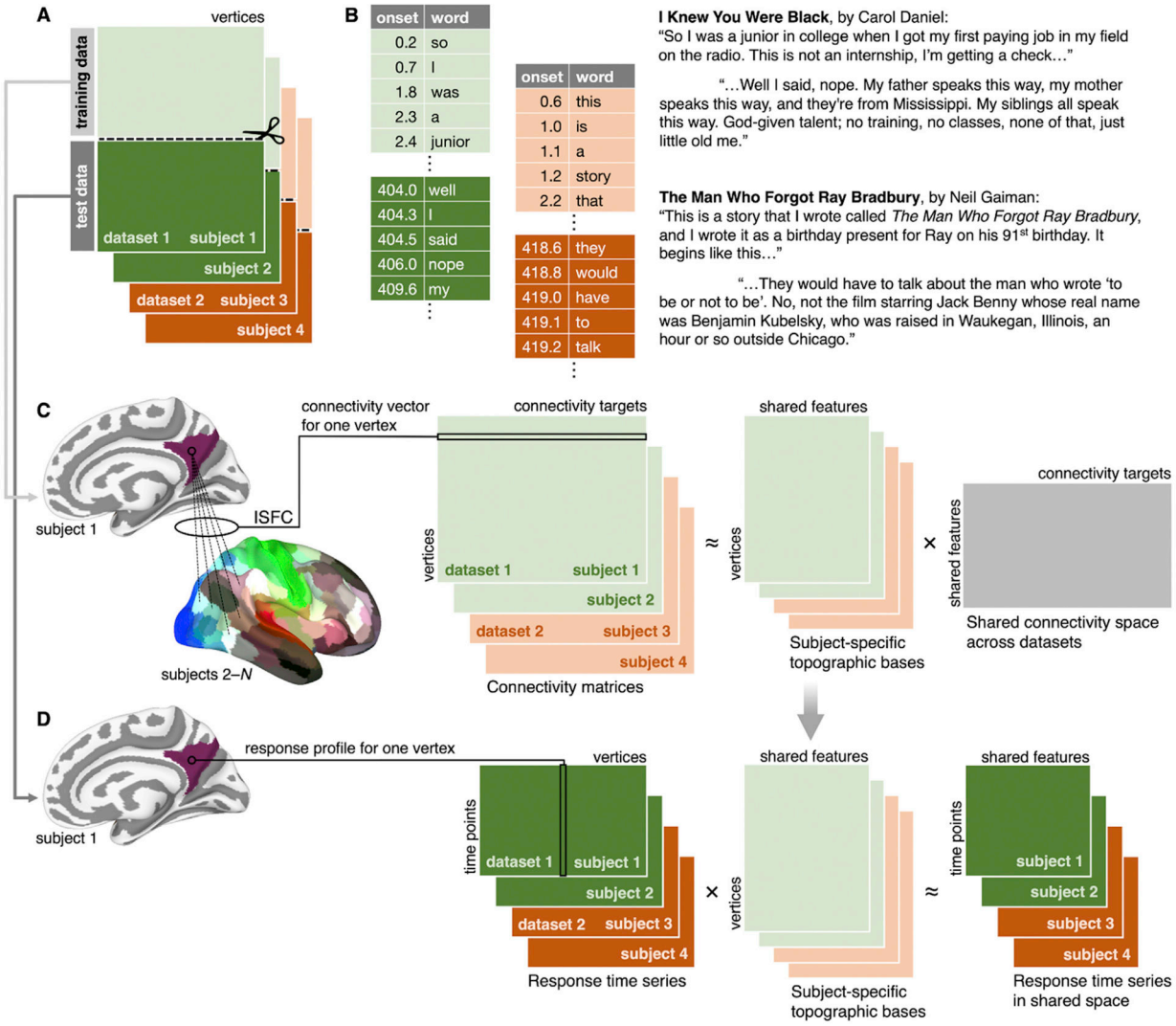

Fig. 2.

Schematic of connectivity-based shared response model. (*A*) Data comprise multiple stories (e.g., dataset 1–2). and largely non-overlapping samples of subjects (e.g., subject 1–4). The green and orange colors indicate distinct datasets corresponding to different story stimuli. Data were partitioned into training and test sets for cross-validation: the first half of each story was assigned to the training set (light colors; i.e., light green and light orange), and the second half was assigned to the test set (dark colors; i.e., dark green and dark orange). (*B*) Time-stamped transcripts are shown for example stories "I Knew You Were Black" by Carol Daniel and "The Man Who Forgot Ray Bradbury" by Neil Gaiman. (*C*) For a given ROI, we computed ISFC between the response time series at each vertex and the average response time series for each of 360 cortical areas (connectivity targets) in the training set. We then used SRM to decompose these ISFC matrices into a set of subject-specific transformation matrices (topographic bases) and a single shared connectivity space across all datasets. (*D*) We then apply the subject-specific transformations to response time series from the test set.

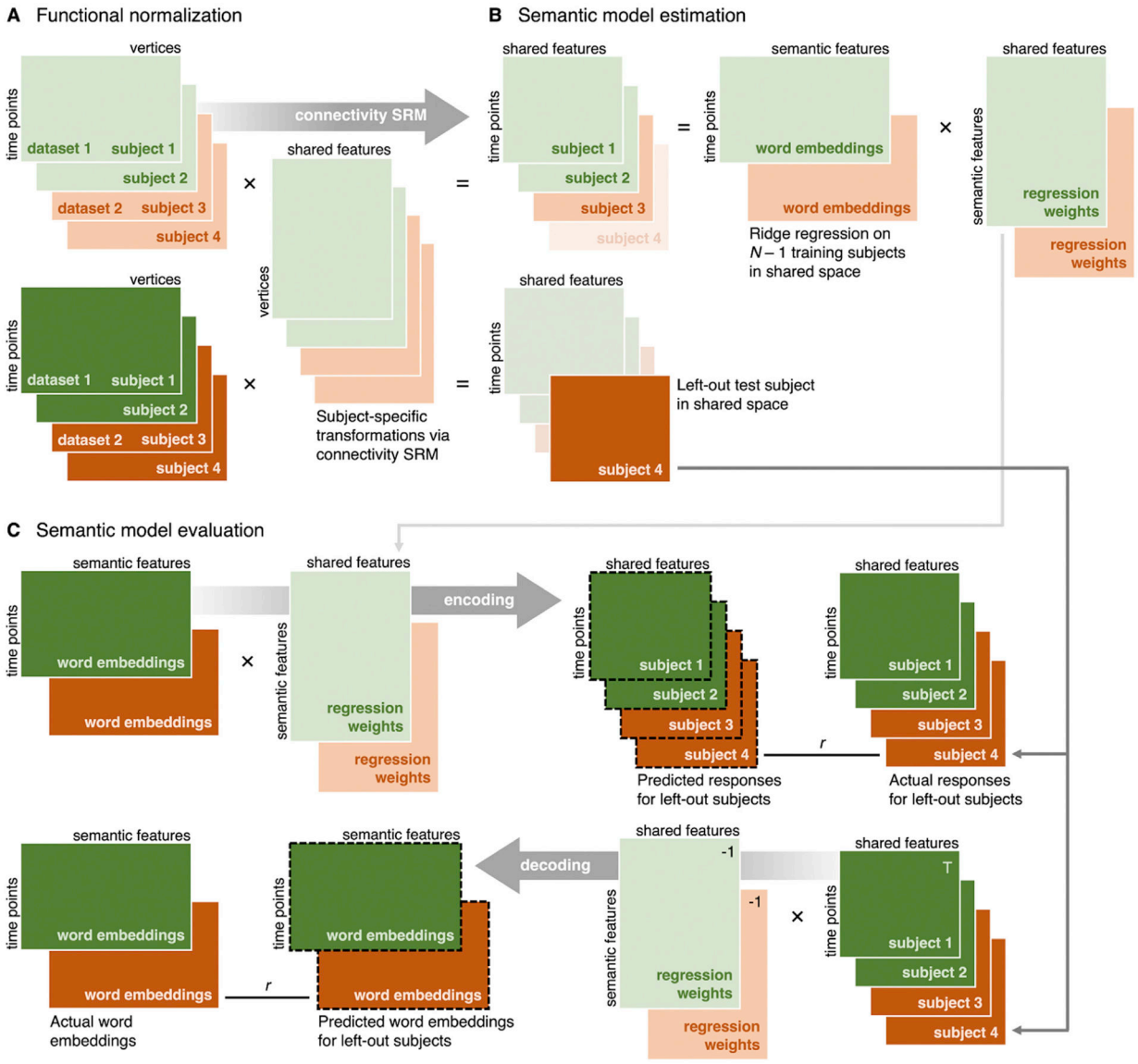

Fig. 3.

Schematic of semantic model-based encoding and decoding. (*A*) Connectivity SRM is used to project both the training and test data into the shared space. The green and orange colors indicate distinct datasets corresponding to different story stimuli. Light colors (i.e., light green and light orange) indicate training data (first half of each story) and dark colors (i.e., dark green and dark orange) indicate test data (second half of each story; as in Fig. 2). Here we average response time series for the $N-1$ training subjects. (*B*) Ridge regression is then used to find a set of coefficients (weights) mapping from the semantic feature space (word embeddings) to the response time series at each vertex or feature. (*C*) In the forward encoding analysis, we use the weight matrix estimated from the training data to predict vertex- or feature-wise response time series from the semantic vectors in the test set. In the decoding (or inverse encoding) analysis, we use the inverse of this weight matrix to predict semantic vectors from the response patterns at each time point. In both cases, we use

correlation to evaluate that match between the predicted and actual response time series or semantic vectors.

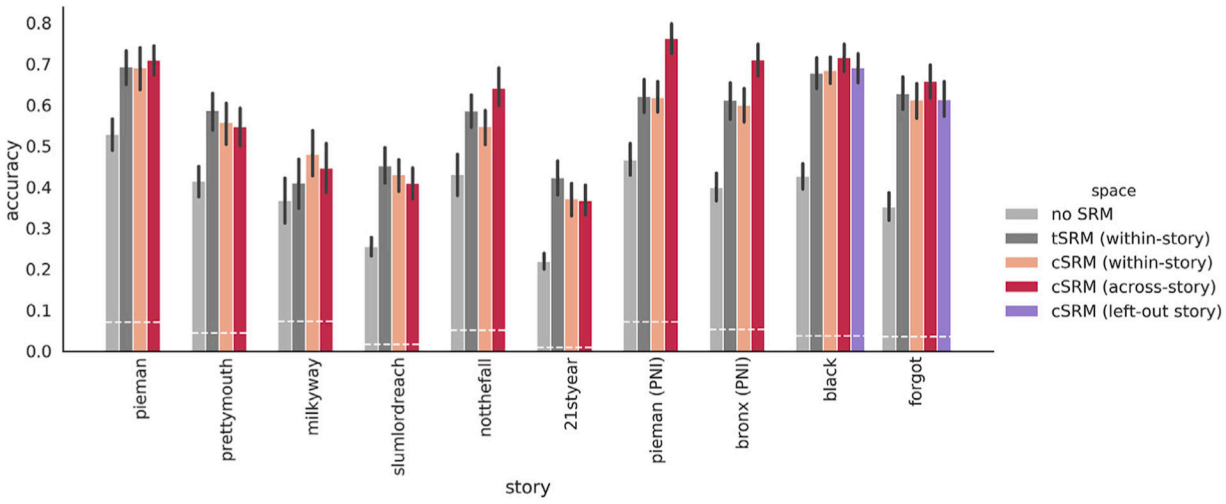

Fig. 4.

Time-segment classification across all stories in auditory association cortex (AAC). For each story, we compared surface-based anatomical alignment with no SRM (light gray), time-series SRM (necessarily defined within each story; dark gray), connectivity SRM defined separately within each story (pink), and a single connectivity SRM defined across all stories (red). We also recomputed a single connectivity SRM across all stories excluding subjects in the rightmost four datasets, and projected the “black” and “forgot” stories into this independent shared space (purple). The *y*-axis indicates between-subject time-segment classification accuracy averaged across left-out subjects and hemispheres. Dotted horizontal white lines indicate chance accuracy for each story (1 over the number of time segments in the test data). Error bars indicate 95% bootstrap confidence intervals estimated by resampling left-out subjects with replacement.

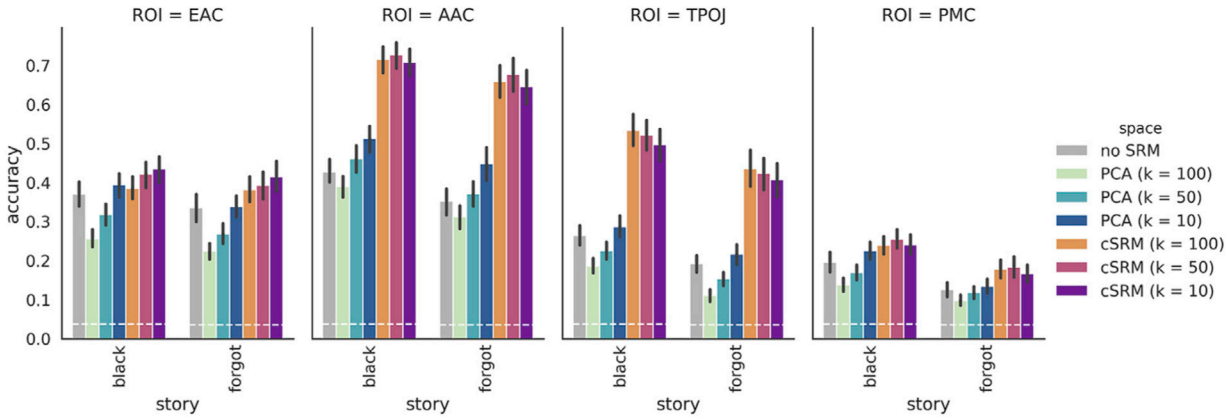

Fig. 5.

Time-segment classification at varying dimensionality for each ROI. For two example stories, we compared surface-based anatomical alignment with no SRM (gray), PCA controlling for the dimensionality reduction of SRM without resolving topographic idiosyncrasies (green–blue), and cSRM at dimensionalities $k = 100$, 50, and 10 (orange–purple). When interpreting reduced-dimension model performance, cSRM performance should be compared to PCA at the matching dimensionality. The *y*-axis indicates between-subject time-segment classification accuracy averaged across left-out subjects and hemispheres. Dotted horizontal white lines indicate chance accuracy for each story (1 over the number of time segments in the test data). Error bars indicate 95% bootstrap confidence intervals estimated by resampling left-out subjects with replacement.

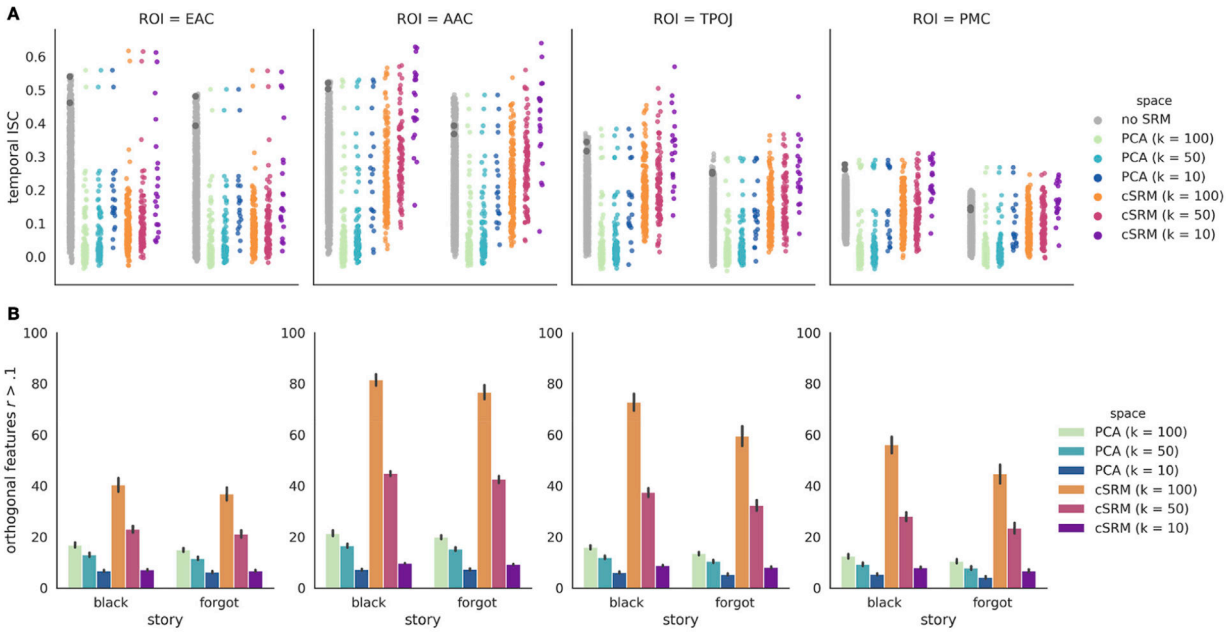

Fig. 6.

Intersubject time-series correlations per vertex/feature. (*A*) We computed the average ISC across subjects for each vertex with anatomical alignment and no SRM (light gray). Each marker corresponds to a vertex or feature for each hemisphere; for example, cSRM with $k = 10$ yields 20 ISC values corresponding to 10 features from the left hemisphere and 10 features from the right hemisphere. See Fig. S2 for the distribution of temporal ISC values split by hemisphere. We also computed ISCs on the regional-average response time series per hemisphere for the purpose of comparison (dark gray markers overlaid on the "no SRM" strip). We visualized the average ISC across subjects for each feature after dimensionality reduction using PCA (green–blue) and cSRM at dimensionalities $k = 100$, 50, and 10 (orange–purple). When interpreting reduced-dimension model performance, cSRM performance should be compared to PCA at the matching dimensionality. The *y*-axis indicates the average temporal ISC across subjects per vertex or feature in each hemisphere. (*B*) We computed the number of features with leave-one-out ISCs exceeding a threshold of $r > 0.1$ per subject (the maximum of which is limited by the specified k). The *y*-axis indicates the average number of features with ISCs exceeding this threshold across subjects. We visualized the absolute number of features exceeding threshold (rather than, e.g., the proportion) to demonstrate that considerably more orthogonal (non-redundant) features with high ISC are observed at higher values of k. Note, however, that the absolute number of features exceeding threshold is bounded by the specified total number of features k. Error bars indicate 95% bootstrap confidence intervals estimated by resampling subjects with replacement. See Fig. S2 for the distribution of temporal ISC values split across hemispheres.

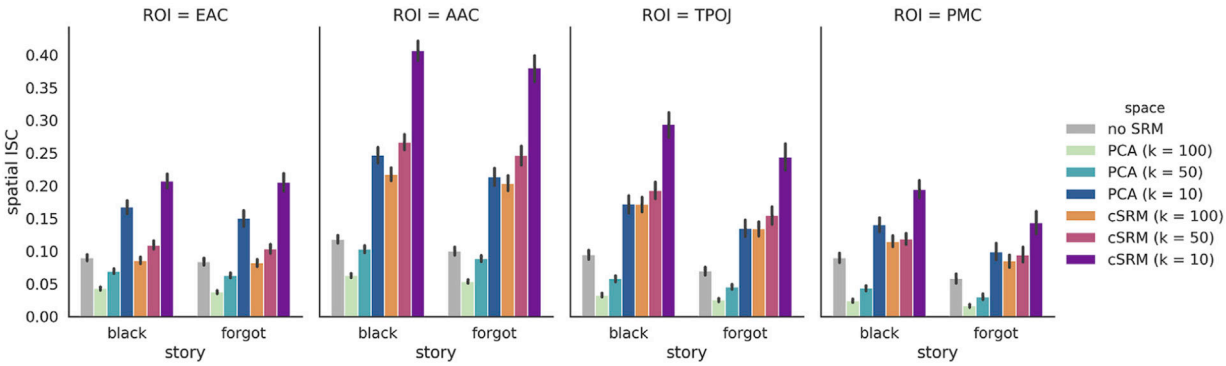

Fig. 7.

Intersubject pattern correlations. We compared spatial ISCs with anatomical alignment (gray), PCA to control for the reduced dimensionality of SRM (green–blue), and cSRM with dimensionalities $k = 100, 50, 10$ (orange–purple). When interpreting reduced-dimension model performance, cSRM performance should be compared to PCA at the matching dimensionality. The y-axis represents spatial ISC for each time point averaged across time points and subjects. Error bars indicate 95% bootstrap confidence intervals estimated by resampling subjects with replacement.

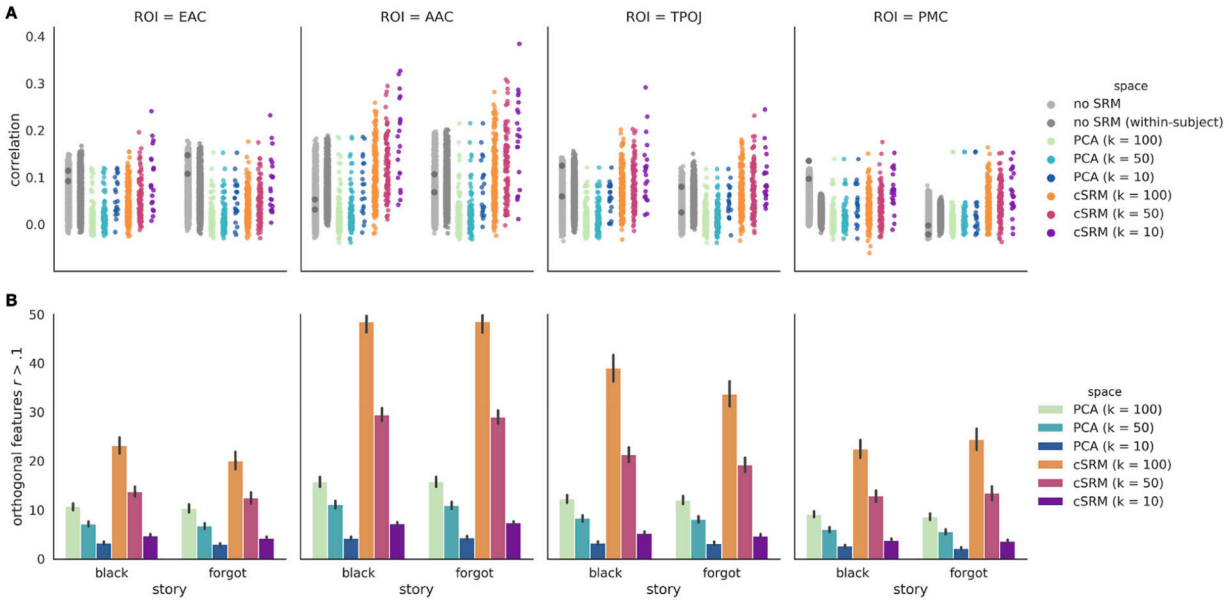

Fig. 8.

Forward encoding model performance. (*A*) We evaluated the vertex-wise between-subject (light gray) and within-subject (dark gray) forward encoding models with anatomical alignment (no SRM). Forward encoding performance for the regional-average response time series per hemisphere is visualized for comparison (dark gray markers overlaid on the "no SRM" strip). We also compared between-subject forward encoding performance for each feature after dimensionality reduction using PCA (green–blue) and cSRM at dimensionalities $k = 100$, 50, and 10 (orange–purple). When interpreting reduced-dimension model performance, cSRM performance should be compared to PCA at the matching dimensionality. The *y*-axis represents the average correlation between predicted and actual response time series across test subjects for each vertex or feature in each hemisphere. (*B*) We computed the number of features with forward encoding performance exceeding a threshold of $r > 0.1$ per subject. The *y*-axis represents the average number of features with performance exceeding this threshold across test subjects (and hemispheres). Error bars indicate 95% bootstrap confidence intervals estimated by resampling test subjects with replacement.

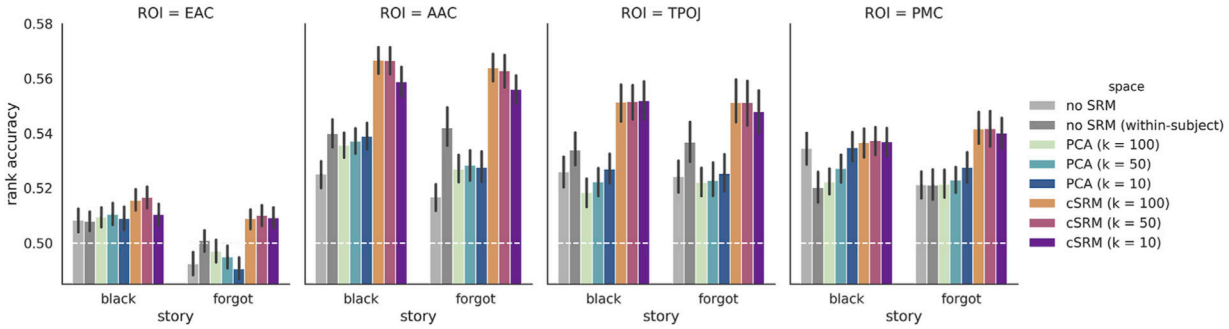

Fig. 9.

Model-based decoding performance. We compared between-subject decoding performance using anatomical alignment (light gray), PCA matching the dimensionality reduction of SRM (blue–green), and cSRM at dimensionalities $k = 100$, 50, and 10 (orange–purple). When interpreting reduced-dimension model performance, cSRM performance should be compared to PCA at the matching dimensionality. We also provide within-subject decoding performance for comparison (dark gray). The *y*-axis indicates rank accuracy averaged across time points, subjects, and hemispheres. Dotted horizontal lines indicate chance accuracy of approximately 50% for the rank accuracy score. Error bars indicate 95% bootstrap confidence intervals estimated by resampling test subjects with replacement.

Fig. 10.

Story similarity in consensus and story-specific shared spaces. We projected AAC ISFCs estimated from test data into either a shared connectivity space ($k = 100$) defined across all stories (across-story cSRM) or story-specific shared connectivity spaces (within-story cSRM). We then computed the pairwise correlations of ISFC matrices across stories. Correlation matrices were computed separately per hemisphere then averaged (see Fig. S3 for correlation matrices computed separately for each hemisphere). The difference between these cSRMs (right) indicates that the cSRM defined across all stories projects into a consensus space. The average (off-diagonal) correlation value for the across-story cSRM is 0.764, while the average correlation for the within-story cSRM is 0.623; the average difference between across- and within-story cSRM correlations is 0.145, and the maximum difference is 0.190. See Figs. S4–8 for all four ROIs and varying dimensionality *k*, as well as Fig. S9 for effects of stimulus duration and sample size.

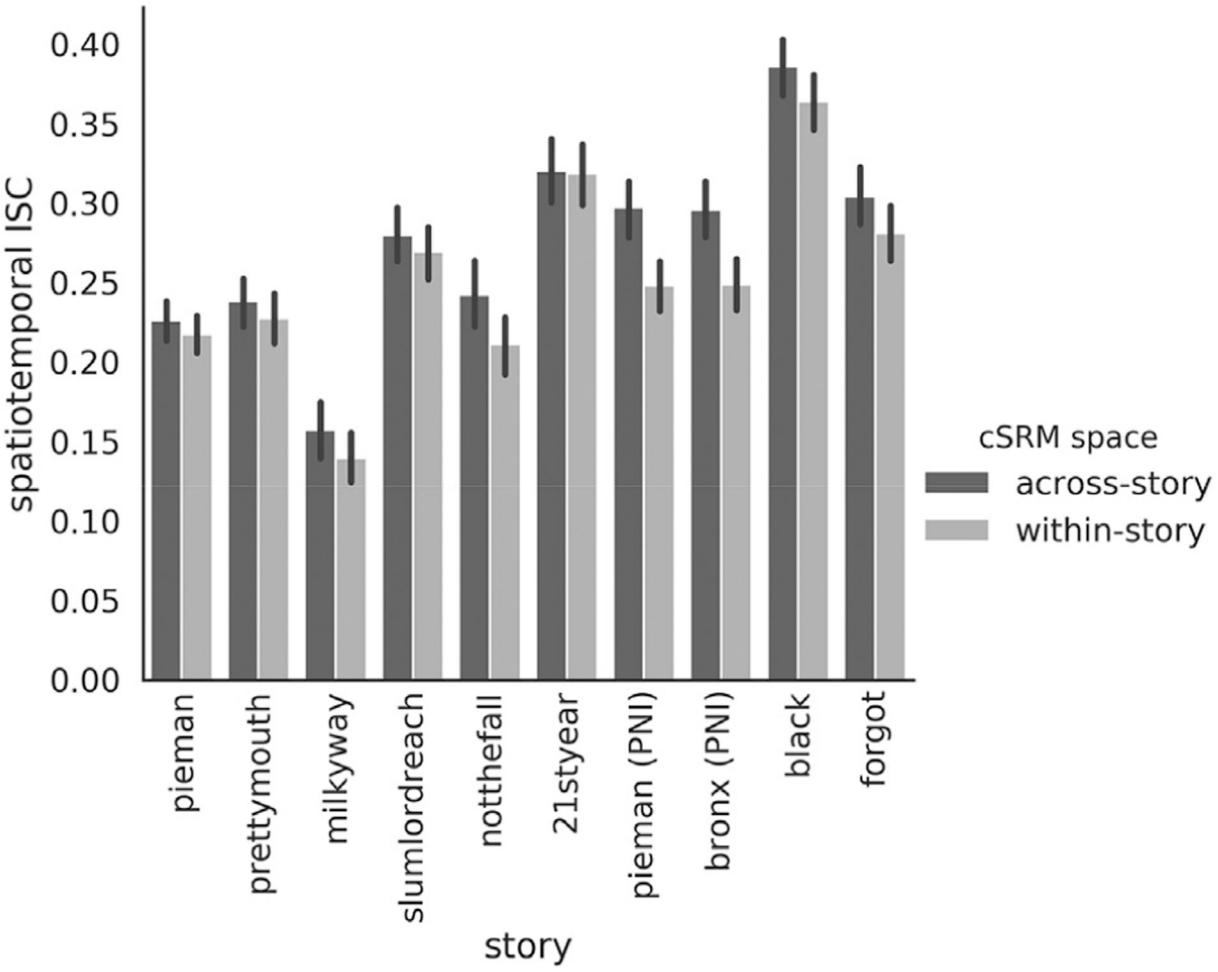

Fig. 11.

Intersubject spatiotemporal correlations in consensus and story-specific shared spaces. We projected spatiotemporal response trajectories into either a shared space defined across all stories (across-story) or story-specific shared spaces (within-story). We then computed the intersubject correlations of these response trajectories within each story. The *y*-axis indicates the average spatiotemporal ISC across subjects and hemispheres. Error bars indicate 95% bootstrap confidence intervals estimated by resampling subjects with replacement.

Table 1

Summarization of 10 benchmark story-listening fMRI datasets. Stimuli comprised 10 spoken stories. In addition to the names of the stories, we use abbreviated aliases in analysis and figures. Story durations are listed in “minutes:seconds” format, and exclude any silence or music bookending the story itself. The number of TRs for each story also excludes any TRs corresponding to silence or music; a 1.5-second TR was used for all acquisitions. The sample size listed for each story corresponds to the number of subjects used for subsequent analyses after applying exclusion criteria (see Participants section). The “total duration” is simply the sum of story durations; i.e., the duration of unique stimuli across datasets (not accounting for the number of subjects in each dataset). The “total duration across subjects” takes into account the number of subjects acquired for each story and reflects the grand total duration if all data were concatenated across both subjects and stories. Note that “Slumlord” and “Reach for the Stars One Small Step at a Time” are distinct stories but were presented one after the other in a single scanning run. The first six stories—from “Pie Man” to “The 21st Year”—were collected on a Siemens Skyra, whereas the remaining four—from “Pie Man (PNI)” to “The Man Who Forgot Ray Bradbury ”—were collected on a Siemens Prisma. The “Pie Man (PNI)” and “Running from the Bronx (PNI)” stimuli were recorded while the speaker underwent an fMRI scan, and those have relatively low audio quality. The “Pie Man (PNI)” stimulus recorded at PNI differs from the original “Pie Man” stimulus recorded at a live storytelling event. See Fig. S1 for more details on which subjects received which story stimuli.

Story	Alias	Duration	TRs	Subjects
“Pie Man”	pieman	07:03	282	46
“Pretty Mouth and Green My Eyes”	prettymouth	11:17	451	19
“Milky Way”	milkyway	06:50	273	16
“Slumlord”, “Reach for the Stars One Small Step at a Time”	slumlordreach	29:26	1,177	16
“It’s Not the Fall That Gets You”	notthefall	09:45	390	18
“The 21st Year”	21styear	55:38	2,225	24
“Pie Man (PNI)”	pieman (PNI)	06:57	278	39
“Running from the Bronx (PNI)”	bronx (PNI)	09:21	374	40
“I Knew You Were Black”	black	13:21	534	40
“The Man Who Forgot Ray Bradbury”	forgot	13:57	558	40
Total duration:		2.7 h	6,542	
Total duration across subjects:		3.1 days	179,093	