Title: *Functional Annotation of Human Cognitive States using Deep Graph Convolution*

Running title: Brain decoding using graph convolutional networks

Yu Zhang [1,2] , Loïc Tetrel [1] , Bertrand Thirion [3] and Pierre Bellec [1,2,*]

[1] Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montreal, QC H3W 1W6, Canada

[2] Department of Psychology, Université de Montréal, Montreal QC H3C 3J7, Canada

[3] Parietal team, INRIA, Neurospin, CEA Saclay, Gif-sur-Yvette, France

*** Corresponding Author:**

Pierre Bellec

Département de Psychologie, Université de Montréal

4565, Chemin Queen-Mary, Montréal (Québec) H3W 1W5

pierre.bellec@gmail.com

Conflict of interest

The authors declare no competing financial interests.

Acknowledgment

This work was supported in part by the Courtois foundation through the Courtois NeuroMod Project and the IVADO Postdoctoral Scholarships Program. PB is supported by a salary award of “Fonds de recherche du Québec - Santé”, chercheur boursier junior 2.

Abstract

A key goal in neuroscience is to understand brain mechanisms of cognitive functions. An emerging approach is “brain decoding”, which consists of inferring a set of experimental conditions performed by a participant, using pattern classification of brain activity. Few works so far have attempted to train a brain decoding model that would generalize across many different cognitive tasks drawn from multiple cognitive domains. To tackle this problem, we proposed a domain-general brain decoder that automatically learns the spatiotemporal dynamics of brain response within a short time window using a deep learning approach. By leveraging our prior knowledge on network organization of human brain cognition, we constructed deep graph convolutional neural networks to annotate cognitive states by first mapping the task-evoked fMRI response onto a brain graph, propagating brain dynamics among interconnected brain regions and functional networks, and generating state-specific representations of recorded brain activity. We evaluated the decoding model on a large population of 1200 participants, under 21 different experimental conditions spanning 6 different cognitive domains, acquired from the Human Connectome Project task-fMRI database. Using a 10s window of fMRI response, the 21 cognitive states were identified with a test accuracy of 89% (chance level 4.8%). Performance remained good when using a 6s window (82%). It was even feasible to decode cognitive states from a single fMRI volume (720ms), with the performance following the shape of the hemodynamic response. Moreover, a saliency map analysis demonstrated that the high decoding performance was driven by the response of biologically meaningful brain regions. Together, we provide an automated tool to annotate human brain activity with fine temporal resolution and fine cognitive granularity. Our model shows potential applications as a reference model for domain adaptation, possibly making contributions in a variety of domains, including neurological and psychiatric disorders.

Keywords: fMRI, brain decoding, brain dynamics, graph convolutional network, deep learning.

Introduction

Identifying brain regions and networks involved in specific cognitive functions has been one of the main goals of neuroscience research. Modern imaging techniques, such as functional magnetic resonance imaging (fMRI), provide an opportunity to map cognitive function in-vivo, and even to decode the dynamics of cognitive processes. Brain decoding has been an active topic in neuroscience literature ever since Haxby and colleagues first proposed the idea of using fMRI brain responses to predict which visual stimuli were presented to a subject (Haxby, 2001). Since then, researchers have extended this line of work by greatly expanding on the type of stimuli used for brain decoding. For instance, researchers have successfully attempted to use brain activity to reconstruct the frames of movies (Nishimoto *et al.*, 2011), or to decode the semantic context from words (Mitchell *et al.*, 2008) and visual scenes (Huth *et al.*, 2012). Other works have moved away from well-controlled experimental conditions, to investigate fluid mental processes such as dreams (Horikawa *et al.*, 2013) and intentions (Haynes *et al.*, 2007). However, the vast majority of existing decoding studies, including the ones referenced in this paragraph, only probed a single cognitive domain at a time, and explored a population of less than ten subjects. The generalizability of these decoding models has not yet been thoroughly investigated in a large population, or across a variety of cognitive domains.

To train such a domain-general brain decoder requires a large collection of brain imaging data. One way to achieve such a large collection is to combine the results from a series of published studies, either using meta-analytic approaches (Rubin *et al.*, 2017; Bartley *et al.*, 2018), or by building linear classifiers based on the contrast maps (Poldrack, Halchenko and Hanson, 2009; Varoquaux *et al.*, 2018). However, these approaches neglect the temporal dynamics of cognitive processes, for which task-evoked brain responses are usually averaged across trials, functional scans or even subjects. Such brain dynamics may contain discriminative patterns of brain responses across different cognitive tasks that are shared among brain regions, or large-scale

functional networks (Gonzalez-Castillo *et al.*, 2012, 2015; Orban *et al.*, 2015). An alternative way is to train classifiers directly from a large set of fMRI data of a large population, for example the Human Connectome Project (HCP), that provides a detailed mapping of cognitive functions consisting of experimental conditions spanning seven cognitive domains (1 hour per subject) (Barch *et al.*, 2013; Van Essen *et al.*, 2013). Based on this powerful resource, several deep artificial neural networks (DNNs) have been recently proposed to map human cognition from recorded brain activity, for instance using the well-known convolutional (Wang *et al.*, 2019) and recurrent neural network architectures (Li and Fan, 2019). But these studies simplified the decoding task by either distinguishing the seven cognitive domains, or only focusing on experimental conditions from a single cognitive domain at a time.

Training a brain decoder that distinguishes task conditions across several cognitive domains may require the introduction of new machine learning tools, that can handle high-dimensional neural activities distributed across multiple brain systems, and that can at the same time accommodate inter-subject variations in brain organization. One promising approach is to model the variety of brain dynamics on a brain graph, which provides a network representation of brain organization by associating nodes to brain regions and defining edges via anatomical or functional connections (Bullmore and Sporns, 2009). Based on this architecture, graph signal processing provides a non-linear embedding tool to project brain activities onto Laplacian eigenspaces that integrate spatiotemporal neural dynamics among connected brain regions and networks (Ortega *et al.*, 2018). This approach has been previously used in the neuroscience literature to study the intrinsic organization of brain anatomy and functions. For instance, (Johansen-Berg *et al.*, 2004) separated the human supplementary motor area (SMA) and pre-SMA by mapping the second Laplacian eigenvector of the connectivity matrix derived from diffusion tractography. (Fan *et al.*, 2016) employed a set of Laplacian eigenvectors from the diffusion connectivity profiles and generated the "Brainnectome" whole-brain parcellation Atlas, which consist of 210 cortical and 36 subcortical subregions. Recently, (Margulies *et al.*, 2016) used the graph Laplacian to reveal the

gradients of functional organization in the human brain connectome, spanning from primary cortex to the default mode network. In terms of clinical applications, Raj and colleagues found a close correspondence between the Laplacian eigenvectors of whole-brain diffusion tractography profiles generated from healthy subjects and the atrophy patterns measured from Alzheimer's patients (Raj, Kuceyeski and Weiner, 2012). These Laplacian eigenvectors can also be used to build a predictive model of future progression to dementia (Raj *et al.*, 2015). Taken together, these studies suggest great potential of using graph Laplacian in neuroscience research.

In this study, we proposed a domain-general decoding model by embedding the graph Laplacian with the DNN architecture, called brain graph convolutional networks (GCN). The proposed approach leverages our prior knowledge on brain network organization using graphs, and automatically learns the spatiotemporal dynamics of cognitive processes during model training. Our decoding pipeline (as shown in Fig 1) takes a short series of fMRI volumes as input, maps the fMRI signals onto a predefined brain graph, propagates information of brain dynamics among inter-connected brain regions and networks, generates task-specific representations of recorded brain activities, and then predicts the corresponding task states. We tested the decoding pipeline on the Human Connectome Project (HCP) database by evaluating the performance across 1200 participants and 21 different cognitive tasks at the same time. The performance was compared with a classical brain decoding model, which applies multi-class linear support vector machines on trial-averaged brain activity. Moreover, a valid brain decoding model requires not only a high prediction accuracy but also good interpretability and generalizability. To evaluate whether the decoding inference was based on biologically meaningful features, we generated saliency maps for the input brain response and compared these saliency maps with prior results from the literature on brain anatomy and function. To investigate the temporal sensitivity of the proposed model, we evaluated the performance with time windows of variable length, ranging from a single fMRI volume to the entire block of task trials, and we explored to which extent the performance of

the decoding model was constrained by the shape of the hemodynamic response. The stability of the decoding model was finally evaluated by changing the number of subjects used for model training.

Results

State annotation using Brain Graph Convolutional Networks (GCN)

Cognitive states can be decoded with high accuracy from 10s of fMRI activity

The GCN state annotation model (Fig 1) was evaluated using the cognitive battery of HCP task-fMRI dataset acquired from 1200 healthy subjects. The entire dataset was split into training (70%), validation (10%) and test (20%) sets at the subject level. During model training and evaluation, fMRI response to different cognitive tasks acquired in HCP was collected and input to the decoding model at the same time. In our study we focused on 21 task conditions spanning six cognitive domains, namely: emotion, language, motor, relational, social, and working memory. The detailed description of these cognitive tasks can be found in (Barch *et al.*, 2013) and is also summarized in Table 2. Using a 10-second window of fMRI time series, the 21 conditions can be identified with an average test accuracy of 89.8%, significantly different from the chance level of 4.8%. The confusion matrix (see Fig 2), which indicates the proportion of true and false predictions given a cognitive task state, showed a nice block diagonal architecture which means the majority of the cognitive tasks were accurately identified. After summarizing the confusion matrix according to the six cognitive domains (see Fig 2-Supplement 1), each cognitive domain could be identified with an accuracy greater than 91%. Among the six cognitive domains, the language tasks (story vs math) and motor tasks (left/right hand, left/right foot and tongue) were the most recognizable conditions, and they showed the highest precision and recall scores (average f1-score = 95% and 94%, respectively for language and motor conditions).

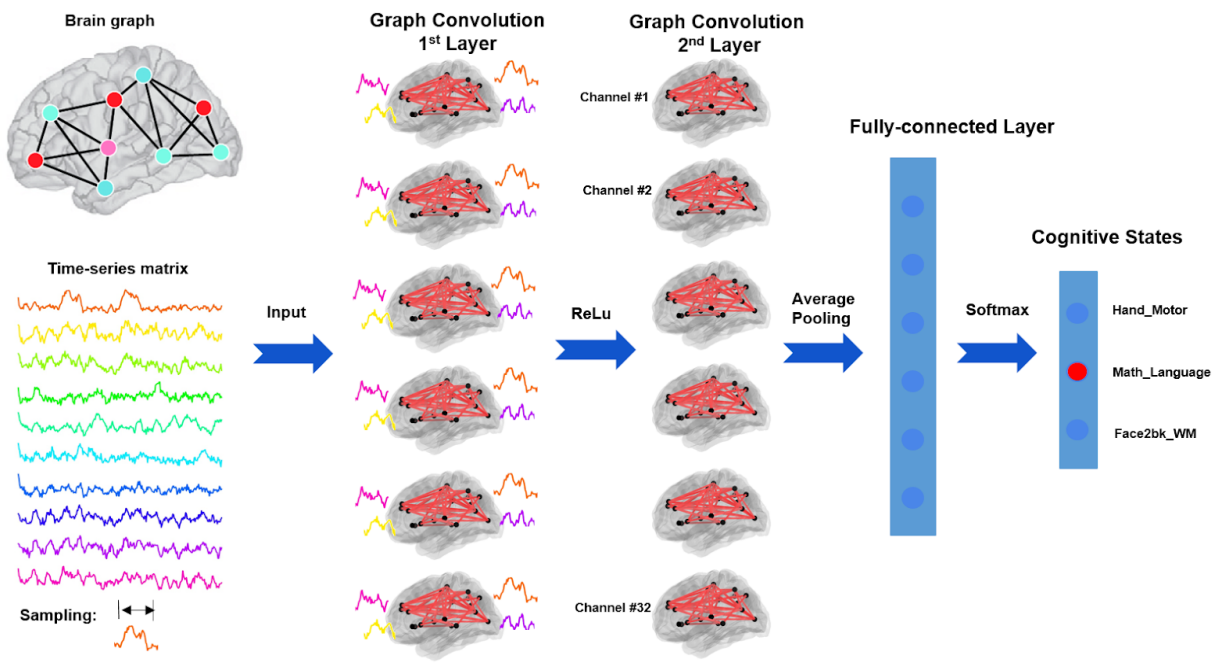

Fig 1. Pipeline of brain state annotation using deep graph convolution network. The proposed state annotation model consists of 6 graph convolutional layers with 32 graph filters at each layer, followed by a global average pooling layer and 2 fully connected layers. The brain graph is constructed by using multimodal cortical parcellation (Glasser *et al.*, 2016) to define the nodes and resting-state functional connectivity to indicate the weights on the edges, both of which were defined based on HCP subjects. A k-nearest-neighbour (k-NN) graph is then built by connecting each brain region only to its 8 neighbors with the highest connectivity. The annotation model takes a short time window of fMRI time series as input, maps the high-dimensional fMRI data onto the brain graph, propagates temporal dynamics of brain response among connected brain regions and networks, generates a high-order graph representation and finally predicts the corresponding cognitive task labels.

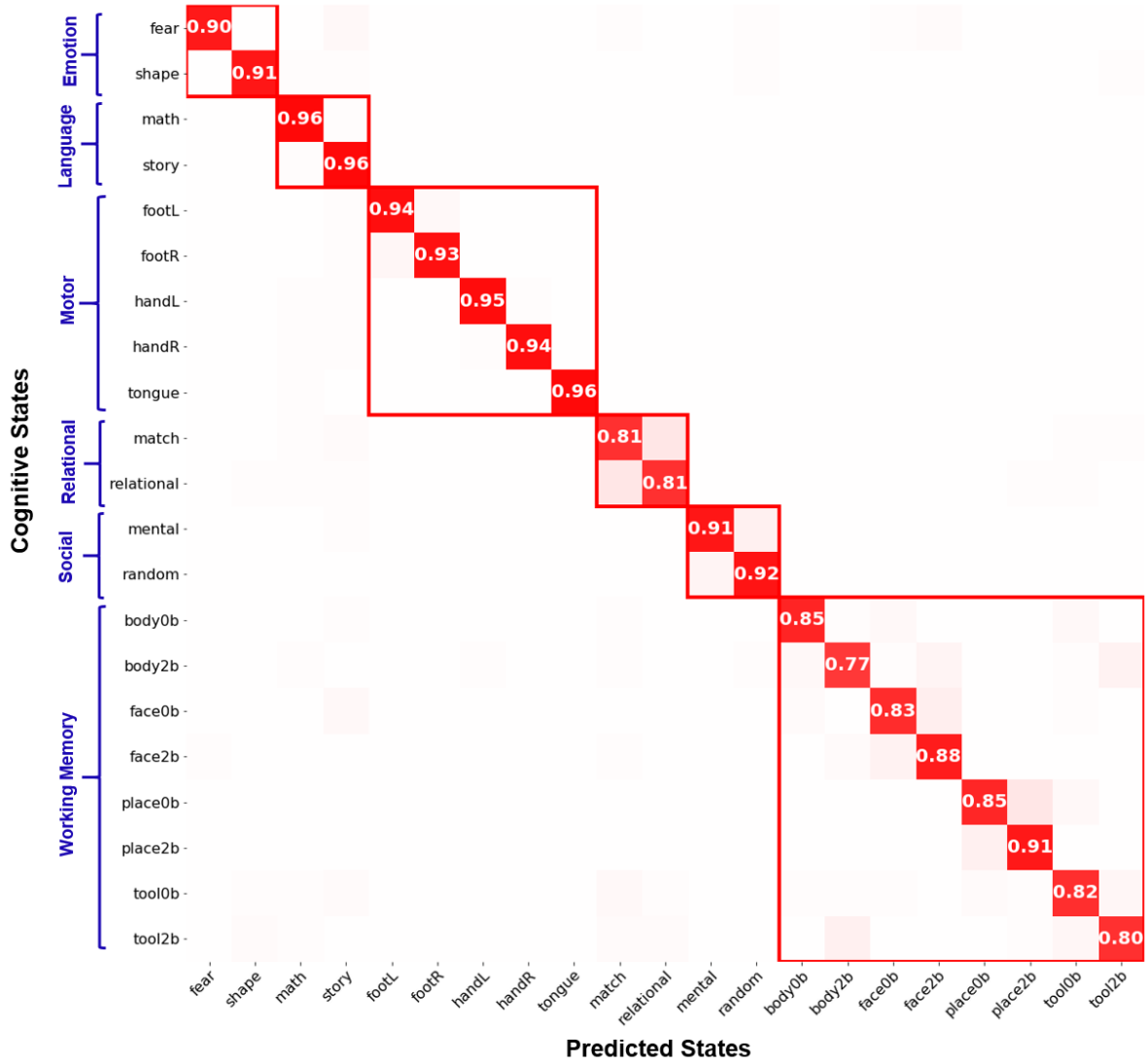

Fig 2. Confusion matrix of decoding 21 cognitive states.

The confusion matrix was normalized by each cognitive state (row) such that each element in the matrix shows the recall score that among all predictions (column) how many of them are positive predictions. The confusion matrix showed a nice block diagonal architecture which means the majority of the cognitive tasks were accurately identified. Among the six cognitive domains, the language and motor tasks achieved the highest sensitivity, with the relational processing and working memory tasks as the lowest. Gambling task was excluded from this analysis due to the short events of the experimental design.

Classification errors are due to high similarity in task stimuli

Misclassifications of cognitive states not only existed within a cognitive domain but also across multiple cognitive domains. First of all, task trials within the same cognitive domain were relatively easy to be misclassified. For instance, most misclassifications of relational processing task trials were found between relational processing and pattern matching conditions. Similar misclassifications were noted between the 0-back and 2-back conditions for the working memory task (see Fig 2-Supplement 2A and B). Similar levels of false classification rates were observed when the decoding model was trained by exclusively using fMRI data from a single cognitive domain (misclassification rates as high as 13% for relational processing vs pattern-matching conditions, 10% for 0-back vs 2-back conditions). By contrast, for face and place working memory stimuli, brain decoding reached high accuracy, regardless of using a domain-general or single-domain classifier (misclassification rates less than 0.2%). This high accuracy is possibly driven by the known, strong spatial segregation of the neural representation for face vs place image, in the fusiform face area and parahippocampal place area respectively (Golarai *et al.*, 2007). Secondly, task trials can also be misclassified across different cognitive domains, probably due to similar cognitive demands of the underlying cognitive processes. For instance, we found some of the emotion and relational processing conditions were misclassified as working memory tasks. One of the reasons could be that the experimental design of the emotion task involves the matching of faces, overlapping with face encoding and retrieval in working memory tasks. Similarly, the relational processing task requires matching of drawn objects based on specific physical characteristics of target images, for instance, shape or texture, somewhat resembling the encoding and retrieval of bodies and tools in working memory tasks. These results suggest that the brain decoding model is mainly driven by the cognitive demands of the tasks and may not follow the original design of hierarchical organization among cognitive domains.

Decoding accuracy associated with in-scanner performance

We found a strong association between the prediction accuracy of GCN state annotation and participant's in-scanner performance, measured using the median reaction time (RT) and average accuracy (ACC) of repeated task trials (Fig 3). For instance, during relational processing task which consists of two conditions, i.e. relational processing and pattern matching, participants reacted faster to the matching condition than relational processing (mean RT=1.48s vs 2.02s, T-val=14.88, pval=3.9e-40) with higher accuracy (mean ACC=86% vs 65%, T-val=13.18, pval=3.4e-33). Similarly, GCN also achieved higher prediction for pattern matching than relational processing (mean F1-score=0.96 vs 0.91, T-val=4.24, pval=2.7e-5). Moreover, within each condition, GCN achieved higher accuracy on trials when participants were more engaged which was indicated as shorter reaction time (Spearman rank correlation rho= -0.21, pval= 0.002) and higher accuracy (rho=0.18, pval=0.012).

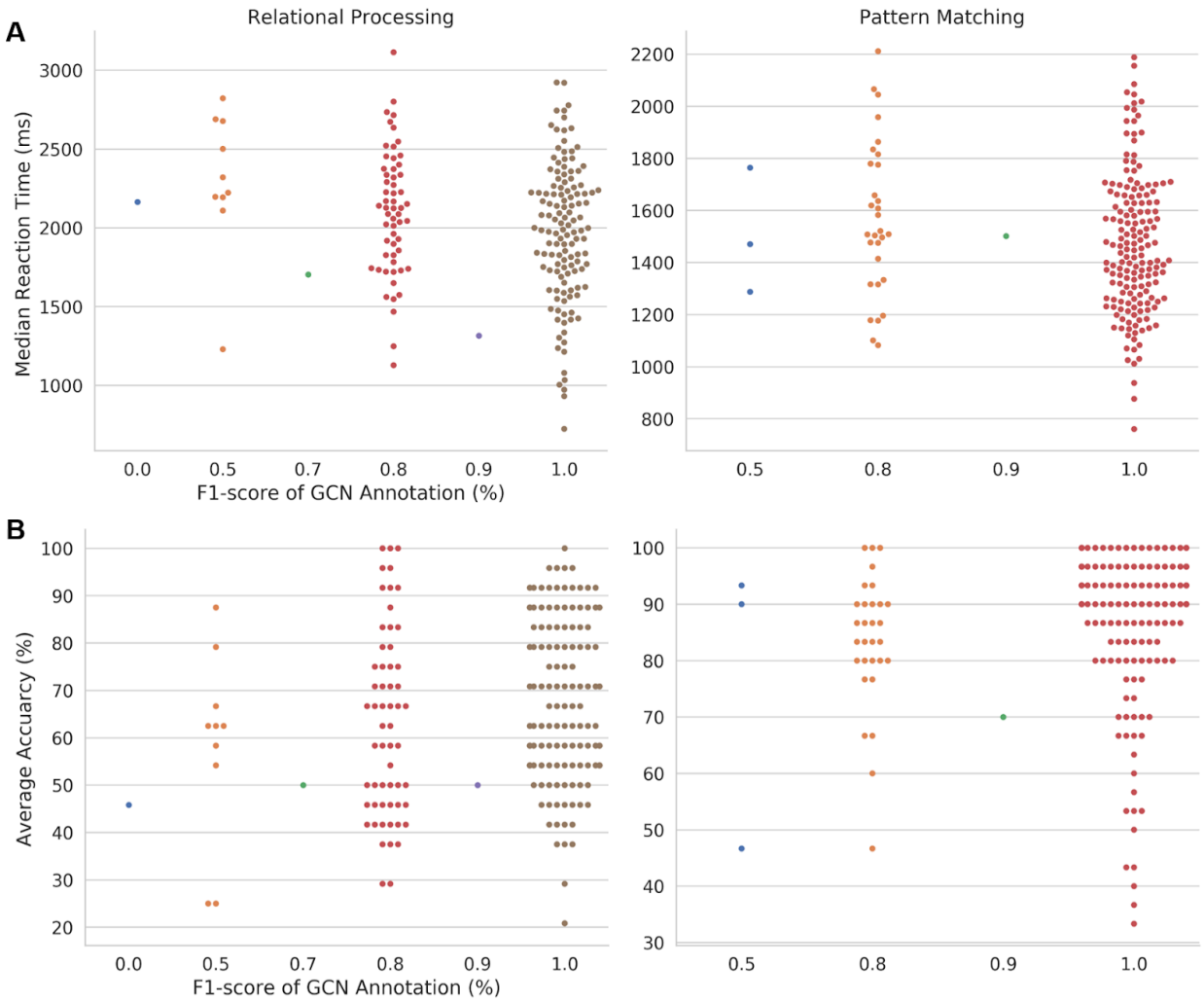

Fig 3. Association between prediction accuracy of GCN state annotation and participants' in-scanner performance for all trials of the relational processing task.

The relational processing task consists of two conditions, i.e. relational processing and pattern matching with each task block lasting for 16s. Two types of performance were measured for each task block, including median reaction time (A) and average accuracy (B) across repeated mini-trials. Comparing the two task conditions, participants reacted faster with higher accuracy for the pattern matching task than relational processing. Similarly, GCN also achieved higher prediction for matching (F1-score = 0.96) than relational processing (F1-score = 0.91). Within each task condition, GCN achieved higher accuracy on trials when participants responded faster (A) or achieved higher accuracy (B). The analysis was performed on 200 subjects from the test set.

Visualization of learned neural representations

To visualize the learned representations of cognitive functions, we projected the high-dimensional graph representations, i.e. the output of the last graph convolutional layer, onto a 2-dimensional space using t-SNE (Maaten and Hinton, 2008). We observed a high clustering effect in the learned representations (see Fig 4C). Specifically, the samples of different movement types were highly separated from each other, with the largest distance existing between tongue and foot movements. Meanwhile, the samples of the same type of movements were located closest to each other. Moderate distance was found between left and right for both hand and foot movements. A similar pattern was also observed by calculating the correlations of the learned representations across all trials (see Fig 4A and Fig 4B). But, this effect was not observed by directly projecting the input fMRI time-series or during the early stages of training process, for which the samples from all categories collapsed into a ball (Fig 4-Supplement 1).

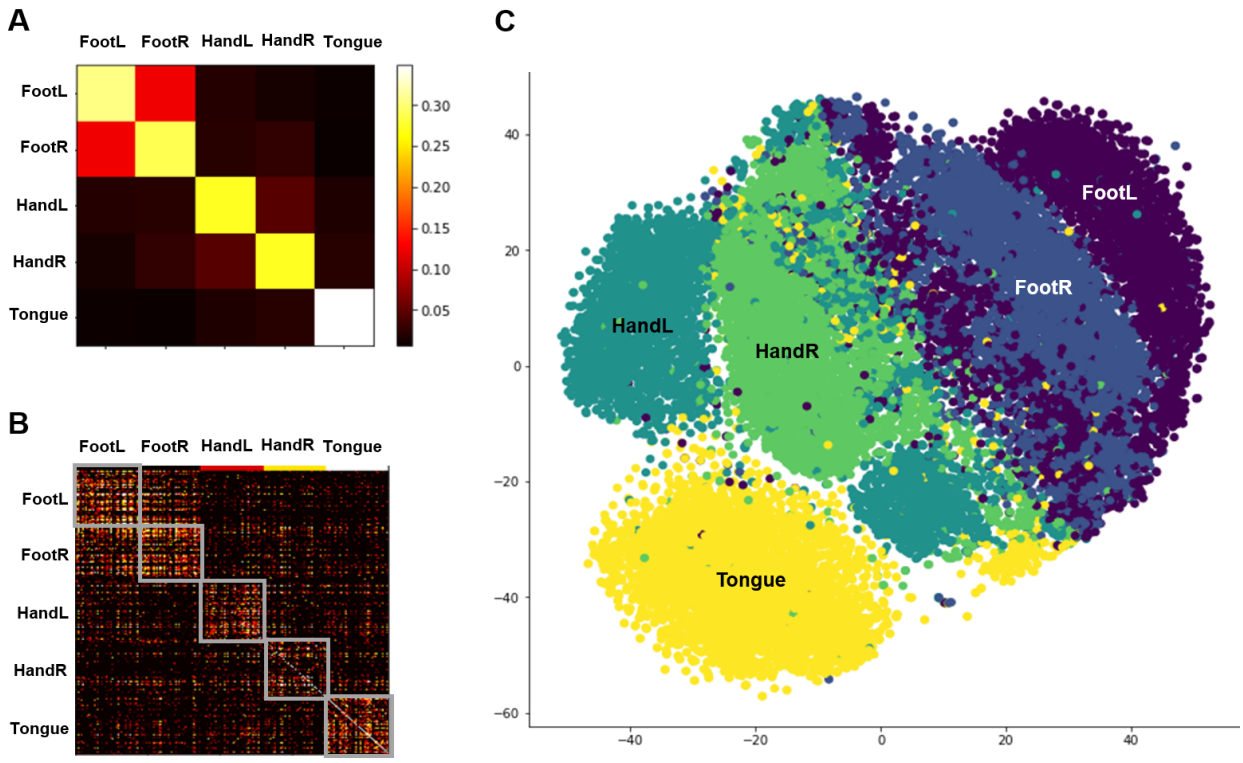

Fig 4. Similarity analysis of learned representations from the Motor task-fMRI data. A pre-trained single-domain GCN annotation model was used for this analysis, which meant the training set only included fMRI signals from the corresponding cognitive domain. Then, the fMRI time series from the test set was passed through the model as input and the layer activations of the last graph convolutional layer were extracted as the graph representations of brain dynamics. The similarity of graph representations was evaluated by calculating Pearson correlation between each pair of brain states (A) and experimental trials (B). Moreover, the learned representations were projected to 2-dimensional space by using t-SNE (C).

GCN outperformed linear and nonlinear decoding models

To establish whether the usage of deep GCN brings a substantial improvement over more traditional machine learning tools, we evaluated the same brain decoding tasks using a multi-class support vector machine classification (SVC) with a linear kernel, as our baseline model. The results showed that using 10s of fMRI data as the input features, SVC-linear achieved much lower prediction accuracy in classifying the 21 states (89% vs 63% respectively for GCN and SVC-linear) and took longer time for training (560s vs 9518s). Even when only focusing on a single cognitive domain, SVC-linear still showed much lower performance (96% vs 87% for the motor task; 86% vs 70% for working memory conditions). We also evaluated a simple multilayer perceptron (*MLP*) consisting of two hidden layers to decode brain activity over 21 states. MLP showed some improvements over the linear model, but not as high as GCN (89% vs 74% respectively for GCN and MLP).

Saliency maps demonstrate biologically meaningful features learned by GCN

We investigated whether GCN learns a set of biologically meaningful features during model training. For this purpose, we generated the saliency maps on the trained model by propagating the non-negative gradients backwards to the input layer (Springenberg *et al.*, 2014). An input

feature is *salient or important* only if its little variation causes big changes in the decoding output. Thus, high values in the saliency map indicate large contributions during the prediction of cognitive states. To note that the model used in this analysis was trained by exclusively using fMRI data from a single cognitive domain.

The two language conditions, story and mathematics, shared a number of salient features, likely related to shared cognitive processes. First, both conditions involve the processing of auditory statements, which may explain high salience in the primary auditory cortex and perisylvian language-related brain regions, consisting of inferior frontal gyrus (IFG), supramarginal gyrus/angular gyrus, and superior temporal gyrus (STG) (see Fig 5A). Second, the block design of both story and math conditions included a presentation and a response phase, and thus potentially imposed a high memory load on participants, and may explain the salience in the inferior parietal sulcus. There were also some salient features found only for either mathematics or story. For instance, the story condition involved salient features in more anterior part of left IFG, including pars triangularis and orbitalis. By contrast, mathematical statements involved more posterior parts, including pars opercularis of IFG and precentral sulcus. Additional inferior temporal regions were salient for mathematics only, which have been shown to be more involved in mathematical than non-mathematical judgment tasks (Amalric and Dehaene, 2016). Finally, left-lateralized salient features in IFG and STG were only revealed for the story condition, coinciding with the study showing strong lateralization for listening comprehension (Berl *et al.*, 2010).

As expected, no salient features in the perisylvian language-related brain regions were found for the motor task. Different types of movements were associated with high salience in the primary motor and somatosensory cortices (see Fig 5B), which have long been shown to be the main territories engaged during movements of the human body (Penfield and Boldrey, 1937). No clear somatotopic organization among different types of movements were identified here, which was somewhat expected because the primary motor and somatosensory cortex were parcellated into

single strips in the Glasser's atlas (Glasser *et al.*, 2016). Some category-specific salient features were still identified, for instance in medial primary motor cortex for foot movement and in lateral orbitofrontal cortex for tongue movement. Unexpectedly, salient features in the left temporal pole were found for all movements. This area has been shown to support language comprehension and production (Ardila, Bernal and Rosselli, 2014), which may be related to the word cues used to initiate different types of movements.

Moreover, salient features in the ventral visual stream were identified for image recognition in the working memory task (see Fig 5C). Specifically, the place stimuli activated more medial areas in the ventral temporal cortex including parahippocampal gyrus; while the face stimuli activated more lateral ventral temporal regions including fusiform gyrus. This observation is consistent with the well-known segregation of the neural substrates for encoding faces vs places, in the fusiform face area and parahippocampal place area respectively (Golarai *et al.*, 2007).

We also found some overlap in brain regions between our saliency maps and meta-analysis activation maps from neuroquery (Fig 5-Supplement 1), as well as contrast maps from HCP dataset (Fig 5-Supplement 2). For instance, the inferior frontal gyrus and superior temporal gyrus were identified for the story condition in all three maps, while the inferior frontal sulcus and the adjacent middle frontal gyrus were identified for the mathematics condition, probably counting for the requirement of working memory for a sequence of mathematical operations. Although some consistent patterns of activations were observed for the motor tasks (Fig 5-Supplement 3), we found a large degree of divergence after mapping them onto the Glasser's atlas (Glasser *et al.*, 2016), probably due to the primary motor and somatosensory cortex being parcellated into single strips in the atlas and not differentiating the somatotopic areas in the feature space. For image recognition, the ventral visual stream was identified in all three maps, but with different specific spatial locations. Overall, although some overlap existed between saliency maps and meta-analysis maps, there was no systematic correspondence. This likely reflects the fact that features identified through traditional statistical tests and predictive models are to some extent

divergent (Bzdok and Ioannidis, 2019). Similar observations were made with contrast maps from the HCP dataset.

In summary, the regions highlighted by the saliency maps are consistent with prior knowledge from the neuroscience literature, and suggest that the GCN model has learned biologically meaningful features, rather than relying on confounding effects, for example motion artifacts.

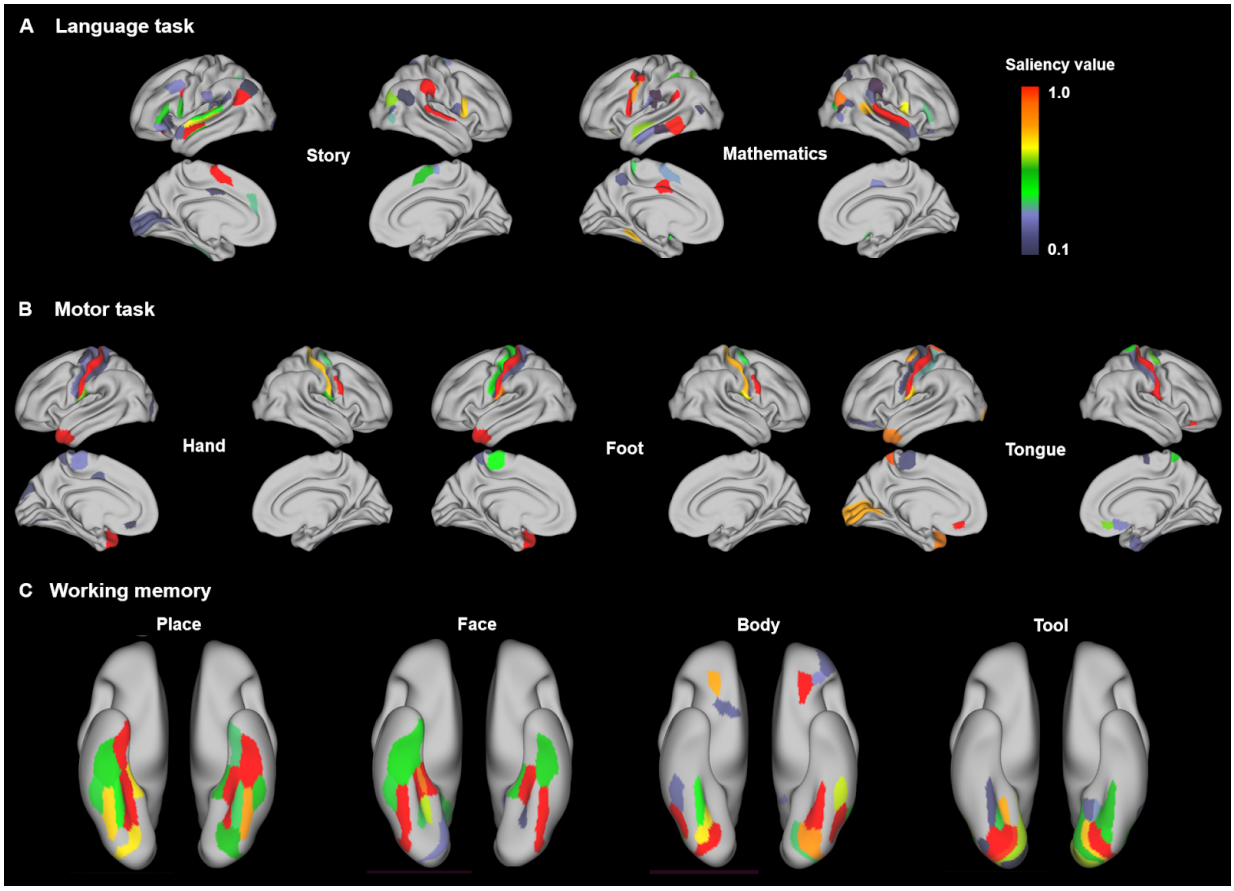

Fig 5. Saliency maps of language, motor and working memory tasks.

(A) The story and math conditions showed high salience in the primary auditory cortex and perisylvian language-related brain regions. (B) Different types of movements were associated with salient features in the motor and somatosensory cortex. (C) The 0-back working memory task mostly engaged the ventral visual stream for encoding different types of images. The saliency maps were estimated by using the guided backpropagation based on the pre-trained single-domain GCN annotation models that only used fMRI signals from the corresponding

cognitive domain during model training. A high saliency value indicates that little variation of the input features causes big changes in the decoding output. The saliency value was normalized to the range [0,1], with its highest value at 1 (a dominant effect for task prediction) and lowest at 0 (no contribution to task prediction). Only values above 0.1 were shown here to indicate a strong impact on the prediction.

Impact of the duration of fMRI time windows on cognitive annotations

Cognitive tasks showed different sensitivity levels to the duration of time windows

The temporal sensitivity of GCN was first evaluated by progressively increasing the duration of the fMRI time windows (Fig 6). At a temporal resolution of one fMRI volume (720ms), GCN could predict the 21 task conditions with an average accuracy of 56%, markedly lower than using 10 sec time windows, yet still significantly higher than chance level (4.8%). As the duration of fMRI time windows became larger, the prediction accuracy gradually increased and converged to a plateau of 89% at 10s of fMRI time series. Using 6s of fMRI data, GCN already showed good performance with an average prediction accuracy of 82%. The cognitive tasks showed different levels of sensitivity to the duration of time windows. Among the cognitive domains, the decoding accuracy of relational processing and working memory conditions were highly dependent on the duration of time windows and required more than 10s to reach stable performance (Fig 6-Supplement 1). These domains also showed the lowest prediction accuracy for all durations of time windows. By contrast, predictions on language and social tasks reached high accuracy for durations as small as one fMRI volume (70% and 66% for conditions of language and social tasks, respectively). This divergence on the temporal sensitivity might be driven by the form of stimuli that successive trials were used for the relational processing and working memory tasks while an auditory/video stream with continuous stimulation was presented for the language and social tasks.

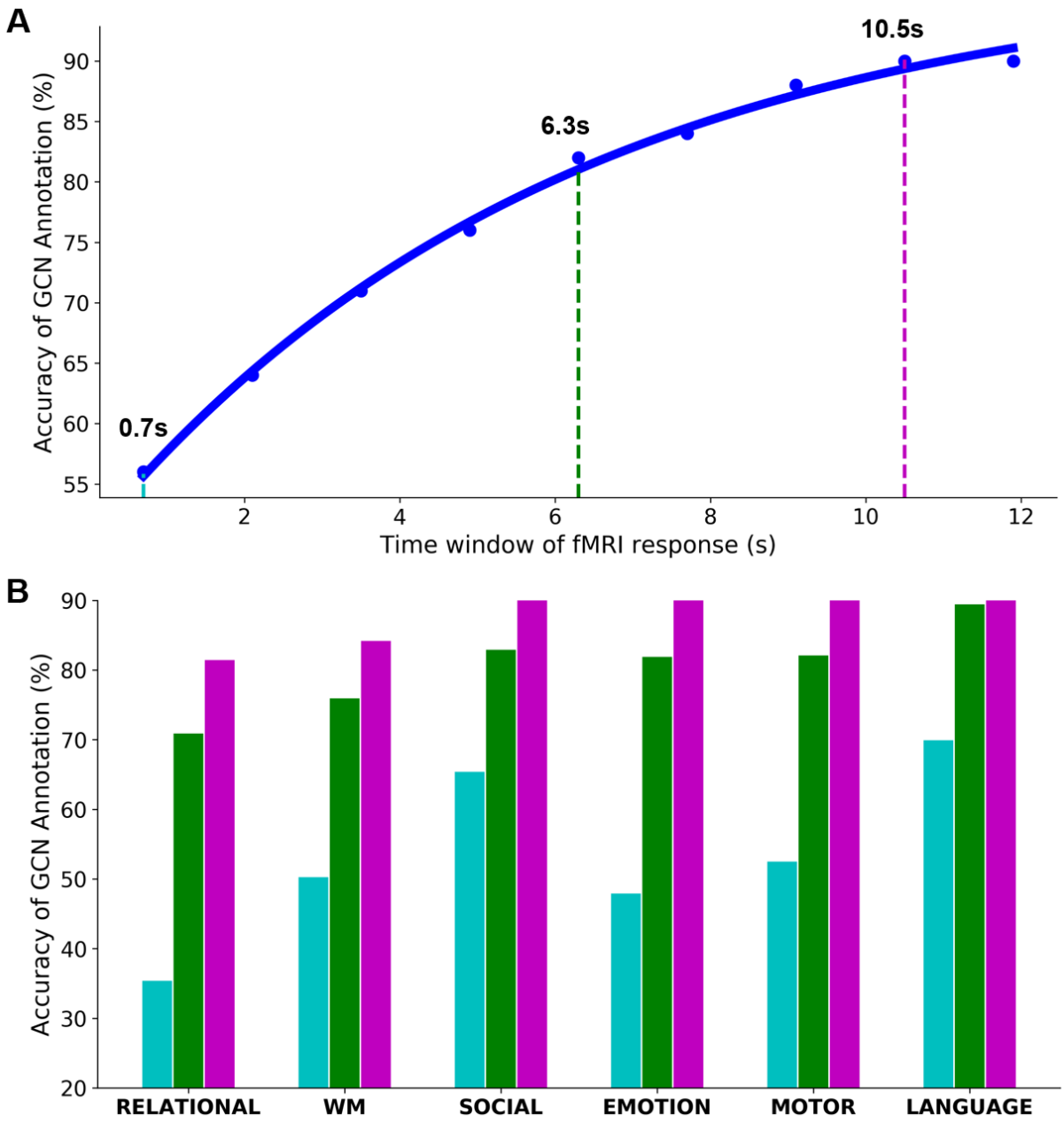

Fig 6. Temporal sensitivity of GCN on the fMRI time window.

The temporal sensitivity of GCN was investigated with variable lengths of time windows, ranging from a single fMRI volume (0.7s) to 10s with a step of 2 TRs (i.e. 1.4s). (A) The performance of GCN annotation gradually increased as prolonging the time window of fMRI time series. It started with 56% of test accuracy on a single fMRI volume (cyan), quickly increased to 82% with 6s of fMRI data (green), and reached a plateau of 89% at 10s (purple). (B) The cognitive tasks showed

high diversity in the sensitivity to the duration of time windows. Among the cognitive domains, the relational processing and working memory tasks were most sensitive to the time window and achieved the lowest decoding accuracy at all temporal scales.

The performance of GCN annotation is constrained by the hemodynamic response

The low performance at shorter fMRI time windows could be caused by two factors: 1) fewer parameters in decoding model especially in the first GCN layer (i.e. time window * graph filters); 2) a delay effect of the task-evoked hemodynamic response (HRF) of BOLD signals, that typically includes a dominant peak at 4-6s, and washes out around 8-12s after the end of the stimulus. To evaluate the impact of the hemodynamic response in GCN performance, we reformulated the prediction accuracy of GCN annotation on a single fMRI volume as a function of time-elapsed-from-onset. As shown in Fig 7, the GCN state annotation had an initial low performance at the cue phase, which gradually increased to reach a plateau at 6-8s after task onset. This effect was observed for all states of the motor and working memory tasks. For instance, the predictions on hand, foot and tongue movements reached an asymptotic performance of 95% for a single fMRI volume acquired 6s after task onset (Fig 7A). For the working memory task, the performance was more variable depending on the task conditions. Specifically, for the 0-back working memory task (Fig 7B), performance reached a plateau at around 8s and fluctuated around this asymptotic level. By contrast, for the 2-back working memory task (Fig 7C), the plateau was only reached at 10s after onset, and some conditions even showed a decreased performance after 20s, for example, the 2-back recognition of body and tool images.

These results suggest that for event-related designs (i.e. with short time duration of each trial), fMRI signals recorded at least 4s after the onset of the task will be required to achieve a stable GCN performance. This observation may also explain the low performance of GCN on the gambling task, where each trial only lasted 3.5s (1.5s for button press, 1s for feedback and 1s for

ITI). To verify whether this rule applies to longer event trials, each task trial was split into multiple bins of 6s-time window before and after the peak of HRF. The results in Table 1 and Fig 7-Supplement 1 indicated that, with the same length of time window, GCN achieved higher performance when the BOLD signals already reached the peak of HRF, but before reaching the post-stimulus undershoot.

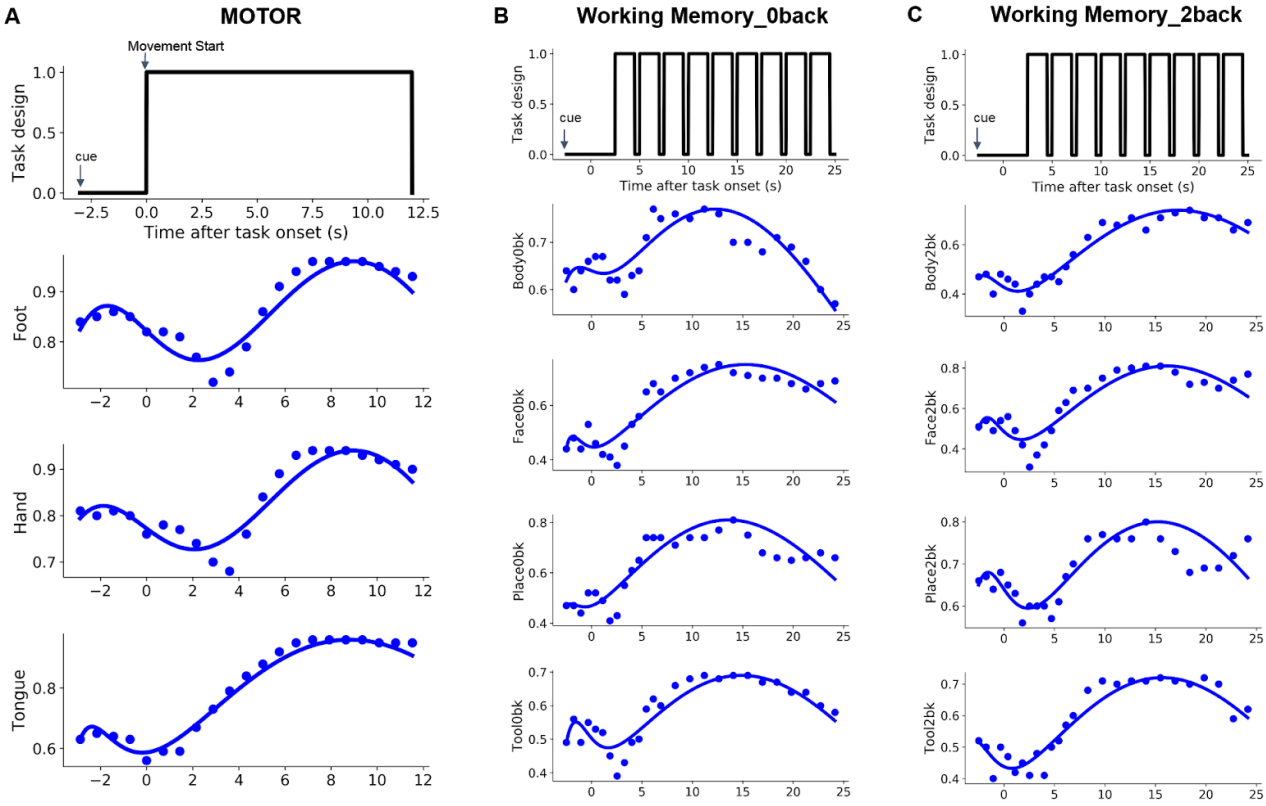

Fig 7. Performance of the GCN annotation as a function of time following onset. The pre-trained single-domain GCN annotation models were used for this analysis by exclusively using fMRI signals from a single cognitive domain during model training. The time window was set to 0.7s such that each single fMRI volume was treated as an independent sample. The trained model was then used to predict the cognitive state of all fMRI volumes from the test set as a function of time following onset. The state annotation of the motor (A) and working memory (B and C) tasks indicated an initial low performance at the cue phase, gradually converging to the plateau at 6-8s after the onset of a task, and then a variable post-stimulus undershoot. This

resembles the effect of the task-evoked hemodynamic response of fMRI signals. Notably, GCN annotation on the motor task even achieved over 90% of test accuracy by decoding on a single fMRI volume acquired 6s after the task onset. The performance was more variable for the working memory task, e.g. lower accuracy for the 2-back conditions compared to the 0-back, but with a reverse observation for the face recognition conditions (i.e. peak performance of 75% vs 81% respectively for the 0-back and 2-back face recognition conditions).

Table 1. Performance of GCN annotation using mini-blocks of a 6s-time window before and after the peak of HRF.

Task trials were split into mini-blocks with a temporal duration of 6s. Event blocks from the motor task last for 12s and thus were split into 2 mini-blocks of 6s time window. Event blocks from the working memory task last for 25s and thus were split into 4 mini-blocks of 6s time windows. These mini-blocks were treated as independent samples during model training. We also trained and evaluated separate decoding models for each of the time windows, by exclusively using the fMRI time series from the corresponding time bins. The last column indicates the average decoding performance on a mixture of 6s mini-blocks by including fMRI signals at all different phases.

Task domain	Decoding Performance on Time windows				
	0-6s	**6-12s**	12-18s	18-24s	Mixed (6s time window)
Motor	85.51 %	**94.58 %**	N/A	N/A	88.60 %
Working memory	75.54 %	**85.72 %**	82.59 %	81.89 %	80.37 %
ALL tasks	79.43 %	**88.38 %**	N/A	N/A	81.51 %

Impact of population sample size on cognitive annotations

GCN annotation reached a performance plateau with around 280 subjects

The sensitivity of GCN on sample size was investigated by changing the number of independent subjects selected from HCP task-fMRI dataset, ranging from 14 to 1060 subjects who have collected 2 sessions of all cognitive tasks. These subjects were again split into training (70%), validation (10%) and test (20%) sets. Generally, with more subjects, GCN achieved higher accuracy in decoding the 21 cognitive states (Fig 8). GCN annotation already achieved decent performance with a handful of subjects (average f1-score=46% using 14 subjects). Performance quickly increased to 77% by using 71 subjects and reached a plateau of 85% with around 280 subjects. After that, performance only showed slight improvement by using larger data samples.

Different cognitive tasks showed different highly variable sensitivity to sample size, and also varied in the asymptotic performance of the model. Specifically, the relational processing and working memory required the largest sample size: 284 subjects and 213 subjects, respectively, to reach 85% of the asymptotic performance. By contrast, the language and motor tasks only required 35 and 57 individuals, respectively, to reach 85% of the asymptotic performance. This variation on the sensitivity of sample size might be caused by different levels of inter-subject variability in the cognitive demands of the underlying cognitive processes. For instance, large individual variability has been reported in working memory tasks (Osaka *et al.*, 2003; Fougnie, Suchow and Alvarez, 2012), while the language network was consistently activated during the auditory language comprehension across different populations and languages (Friederici, 2011; Zhang *et al.*, 2017; Wu, Zaccarella and Friederici, 2019).

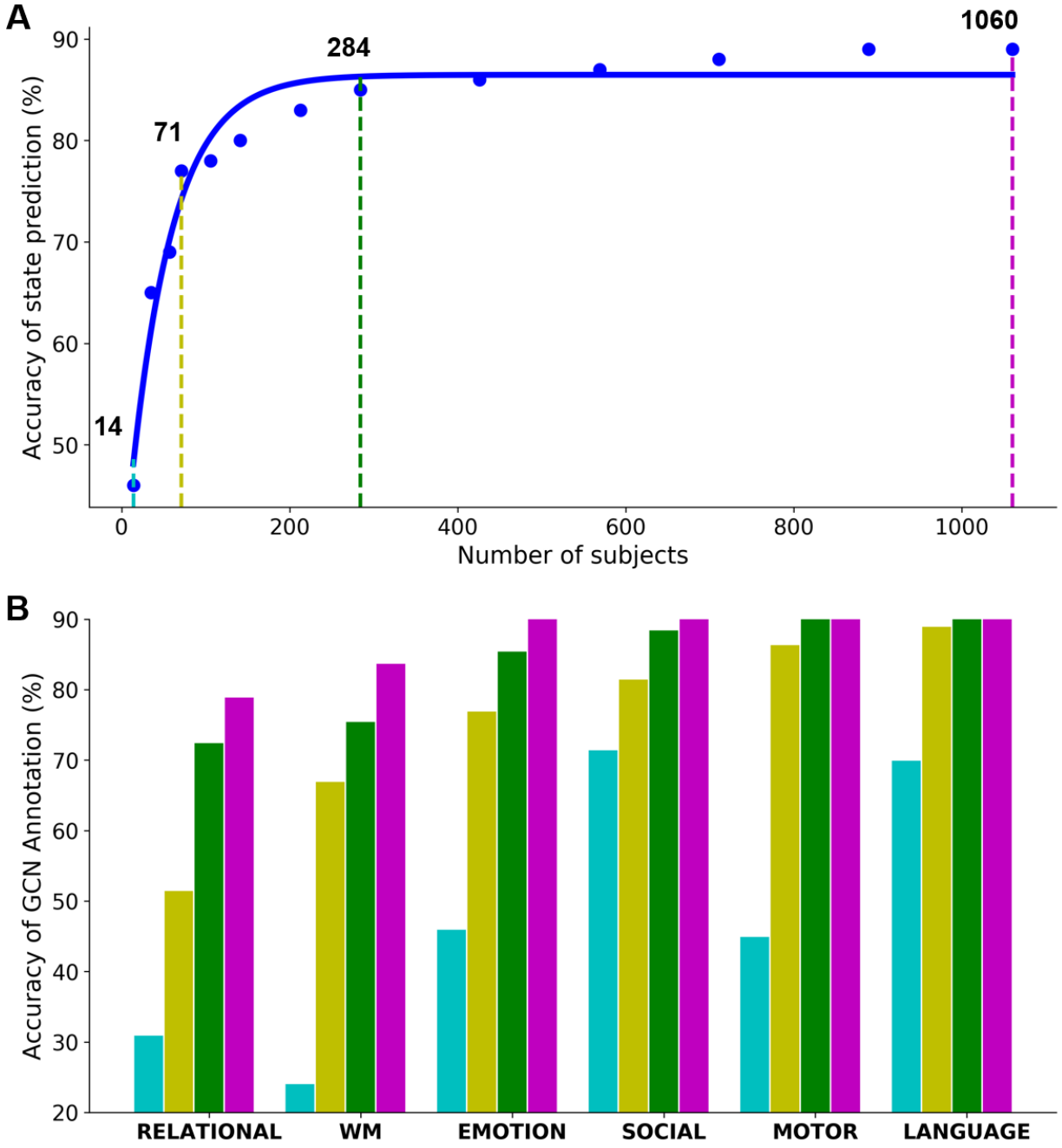

Fig 8. Sensitivity of GCN on sample size of independent subjects.

The sensitivity of GCN on sample size (A) was investigated by changing the number of independent subjects selected from HCP task-fMRI dataset, ranging from 14 to 1060 with a smaller step before the plateau and a larger step after. GCN annotation starts with 46% of test accuracy in decoding the 21 cognitive states by using only 14 subjects (cyan). Then, the performance quickly increased to 77% by using 71 subjects (yellow) and reached a plateau of

85% with around 280 subjects (green). Among the cognitive domains (B), the relational processing and working memory were the most demanding tasks on the sample size, while the language and motor tasks were more robust to the size of the dataset.

Discussion

The present study proposed a generalized brain decoding model which annotates brain dynamics of 21 cognitive functions using a short series of fMRI volumes. This approach relies on convolutional operations on a brain graph, which leverages our prior knowledge on network organization of human brain cognition. Graph convolution integrates information of brain dynamics among distributed brain networks and generates robust neural representations that could be generalizable across a large group of population and multiple cognitive domains. Specifically, our model identified 21 experimental conditions across 6 cognitive domains simultaneously with an accuracy of 89% on unseen subjects, by only using 10s window of fMRI signals. This high performance on brain annotation was mainly contributed by brain response of biologically meaningful brain areas, in line with the literature on functional localizers for each cognitive domain, as revealed by the saliency maps. By examining variable time windows, we found that our decoding model achieved above-chance annotation with a fine temporal resolution, as short as a single fMRI volume. Volume-to-volume performance followed the shape of a hemodynamic response, with a high accuracy achieved after at least 6 seconds following stimulus onset. Besides, the model converged to its stable performance by using a subset of 280 subjects. Together, our results provide an automated tool to annotate brain activity with fine temporal resolution and fine cognitive granularity, as well as high generalizability to new subjects.

Domain-general brain decoding

Brain decoding has been a popular topic in neuroscience literature for decades. The majority of studies still focused on the recognition of visual stimuli (Haxby, 2001; Huth *et al.*, 2012; Haxby, Connolly and Guntupalli, 2014). To build a decoding model that could generalize beyond visual stimuli and incorporate multiple cognitive domains is still a challenging topic. Researchers have attempted to tackle the issue of domain-general brain decoding by using meta-analytic

approaches based on thousands of reported brain coordinates (Bartley *et al.*, 2018) and a set of statistical contrast maps (Rubin *et al.*, no date) from a series of published studies. But meta-analyses bring other types of limitations, such as unbalanced samples across different cognitive domains (Alamolhoda, Ayatollahi and Bagheri, 2017), publication bias towards positive effects (Dubben and Beck-Bornholdt, 2005), as well as over-estimated effect sizes from small studies (Lin, 2018). These factors may bias the decoding analysis by falsely inferring the mental states given limited available studies (see discussion in (Lieberman and Eisenberger, 2015; Lieberman *et al.*, 2016; Wager *et al.*, 2016)).

To avoid these biases, an alternative approach has also been proposed by training linear classifiers on the activation maps collected from a group of individuals that have been scanned over a variety of cognitive tasks (Poldrack, Halchenko and Hanson, 2009; Bzdok *et al.*, 2016; Varoquaux *et al.*, 2018). It is worth noting that, by using parametric modelling and averaging brain response across multiple trials and even multiple runs, it is possible to achieve high accuracy on the task of distinguishing different experimental conditions, for instance classifying a subset of HCP tasks (Bzdok *et al.*, 2016). The challenge here is to achieve such high accuracy using a fully data-driven approach to infer cognitive states directly from a short time series. This requires the decoding model to take into account not only the overall discriminative patterns of brain response under different cognitive tasks, but also their temporal dynamics, i.e. changes of brain activations over time. Such brain dynamics are usually revealed by electro- or magnetoencephalography (Kietzmann *et al.*, 2019), but has also recently been investigated in fMRI studies. For instance, Gonzalez-Castillo and his collegues reported distinct shapes of task-evoked hemodynamic responses among distributed brain networks during a discrimination task of letters and numbers (Gonzalez-Castillo *et al.*, 2012). Similar findings were also revealed in our previous study that premotor and sensorimotor cortex showed different time courses during the preparation and execution stage of a motor sequential task (Orban *et al.*, 2015). These studies suggest that the differences in the shapes of hemodynamic response can help to distinguish different conditions

or stages of cognitive process. Accumulated evidence suggests that it is feasible to infer cognitive states directly from a short time window of hemodynamic response. Early attempts in the field include the reconstruction of visual scenes and prediction of semantic context from natural movies by fitting a linear regression model for fMRI signals for each individual voxel (Nishimoto *et al.*, 2011; Huth *et al.*, 2012). These studies neglected the modular, and hierarchical nature of brain organization by treating each brain voxel independently.

More complex and nonlinear decoding models are required in order to incorporate the high-dimensional spatiotemporal dynamics of brain response that are shared among distributed brain networks. Recently, promising results on brain decoding have been shown by using deep artificial neural networks (DNNs). For instance, multiple cognitive domains can be distinguished by applying convolutional neural networks on the whole-brain hemodynamic response (Wang *et al.*, 2019). But the temporal dependence of hemodynamic response was interrupted by choosing random time points from the entire fMRI scan. This effect can be corrected by applying a recurrent neural network to brain activity instead. Li and Fan proposed a long short-term memory (LSTM) architecture to predict the cognitive states from fMRI time-series of a set of functional networks (Li and Fan, 2019). But this decoding model only worked for a single cognitive domain with a fixed experimental design across all subjects. How to generalize it onto multiple cognitive domains consisting of variable duration of task events is unclear. In this study, we extend this line of work by combining the graph Laplacian with the DNN architecture and proposed a generalized brain decoding model that takes into account both the network architecture of the human brain (in space) and the fluctuations in the task-evoked BOLD signals (in time).

Brain decoding using graph convolution network

Graph Laplacian provides a powerful tool to map the intrinsic organization of the human brain, including parcellating brain areas (Johansen-Berg *et al.*, 2004; Fan *et al.*, 2016), identifying functional areas and networks (Craddock *et al.*, 2012; Atasoy, Donnelly and Pearson, 2016), and

generating connectivity gradients (Margulies *et al.*, 2016). This approach works not only on static brain connectome but also on dynamic brain signals that fluctuate over time. Recently, studies have shown that graph Laplacian captured different modes of brain dynamics by decomposing the task-evoked BOLD signals into different frequencies (Ortega *et al.*, 2018). Convergent evidence suggests that the low frequency modes, which have similar brain signals within a local community, corresponded to the low-level functions that localized within certain brain regions, such as motor learning (Huang *et al.*, 2016). On the other hand, the high frequency modes, which indicate high variational signals across brain networks, were associated with high-level cognitions that distributed among multiple brain systems, such as cognitive switch (Medaglia *et al.*, 2018). We generalized this approach by automatically learning a linear combination of the graph modes across multiple frequencies through graph convolutions, i.e. convolving the input fMRI signals with a graph filter. The resultant decoding model not only represented low-level functions like movements of body parts, but also embedded the high-level cognitions such as N-back working memory and language comprehension. The results showed that the decoding model achieved high classification accuracies on these cognitive tasks (Fig 2 and Fig 2-Supplement 1). Moreover, the saliency maps indicated that the task inference was drawn from brain response of biological meaningful brain regions, for instance, the motor and somatosensory cortex for the motor task, and the perisylvian language network for the story and math auditory statements (Fig 5), consistent with known brain anatomy and function (Penfield and Boldrey, 1937; Friederici, 2011).

Graph convolutions generated a new representation of brain activity by integrating neural dynamics from interconnected brain regions. A variety of neural representations were generated by training multiple graph filters at each GCN layer. Specifically, at the first GCN layer, various shapes of hemodynamic responses were captured by fitting different weights for each time point after task onset. By stacking several GCN layers, high-level graph representations were generated that integrated neural dynamics not only within specific brain networks but also across multiple networks, and even distributed across the whole brain. Our results demonstrated that the

generated graph representations already include task-specific information that discriminates different experimental conditions, for instance, showing the largest distances among different types of movements, moderate distance between left and right movements, and a small distance between the same type of movements (Fig 4). Moreover, a strong association was found between the model performance on classification of graph representations and human performance on recognition of visual patterns, e.g. reaction time of relational processing and pattern matching trials in scanner (Fig 3). A similar finding has been reported previously that the high-frequency graph mode was strongly associated with the response time of trials in a cognitive switch task (Medaglia *et al.*, 2018).

Token together, using brain graph convolutional networks, our decoding model generates high-level neural representations from brain dynamics and provides a possible solution towards domain-general brain decoding by learning various shapes of hemodynamic response and integrating such neural dynamics among multiple brain systems.

Temporal resolution of brain decoding

The temporal resolution of brain decoding has been mostly ignored in previous studies, by either using meta-analytic approaches (Rubin *et al.*, no date; Bartley *et al.*, 2018), or training classifiers on activation maps (Poldrack, Halchenko and Hanson, 2009; Varoquaux *et al.*, 2018). The recent work of Loula and colleagues (2018) demonstrated the feasibility of decoding stimuli with short inter-stimuli intervals. Temporal resolution is thus becoming an important factor for brain annotation, especially when we tried to infer cognitive functions directly from brain response. A series of impressive work has been done in Gallant's group, in which the authors used brain response to reconstruct the visual frames of natural movies (Nishimoto *et al.*, 2011) or to map more abstract concepts of visual objects, e.g. semantic context (Huth *et al.*, 2012). But these studies did not directly attempt to characterize what amount of temporal data is required to perform meaningful brain decoding. The temporal resolution of fMRI decoding is intrinsically

constrained by two factors, including the acquisition time for a whole-brain fMRI scan (i.e. TR) and the delay effect of hemodynamic response. With a common setting as 2 second, the acquisition time was pushed down to a third (TR = 720ms) in HCP database by using simultaneous multislice acquisitions (Uğurbil *et al.*, 2013), which brings opportunities to investigate fine-grained temporal dynamics of brain activity.

Using this dataset, Li and Fan successfully predicted the entire experimental design of the working memory task by using a sliding window approach (Li and Fan, 2019). But each time window still took around 30s of fMRI signals as input for task inference. To which extent of a shorter time window the decoding model can work with is still unexplored. In this study, we applied graph convolutions on a short series of fMRI signals and investigated the temporal resolution of brain decoding by using variable time windows of fMRI scans, ranging from a single fMRI volume to the entire event trial. Leveraging the fast fMRI acquisition of HCP database, our model can annotate 21 cognitive conditions with a sub-second temporal resolution. In the meantime, the decoding performance was still impacted by the task-evoked hemodynamic response, for instance, higher decoding accuracy by using fMRI signals after the peak of HRF than before the peak. This phenomenon was observed not only for low-level functions like body movements, but also high-level cognitions such as working memory tasks, or even missing all experimental conditions together (Table 1 and Fig 6).

There are still a lot of challenges before achieving real-time brain decoding, for instance, to decode fast events with a short duration or even overlapping hemodynamic response.

Limitations and future applications

In the current project, we only explore the block design of task-fMRI dataset, i.e. consisting of long events with repeated trials that in total last for more than 10s. However, it is still unclear how to generalize the decoding pipeline to naturalistic stimuli, for instance visual scenes from movies, which consists of short and fast-switching events. The measured BOLD signals might be a

mixture of hemodynamic responses evoked by different task events. Early attempts have been made by adding independent regressors with delayed onsets (Nishimoto *et al.*, 2011). But the simple linear model only generates a blurred image from the average prediction of each category. One possible solution to this problem is to use a multi-label decoding model based on GCN. Specifically, given a short-series of fMRI signals, the model predicts a set of cognitive states instead of one single task condition. Due to the delay effect of hemodynamic response that reaches plateau around 6s past stimulus, we can modify the label matrix by prolonging each event duration until 8s after the task onset and allow multiple labels assigned to the same time point.

An interesting potential application of our work would be transfer learning. In natural image processing, it is common practice to take a model already trained on a large dataset, such as AlexNet trained on ImageNet (Krizhevsky, Sutskever and Hinton, 2012), and fine-tune the parameters of the trained model to accomplish a new task (Tajbakhsh *et al.*, 2016). This allows training complex models even in the absence of extensive training data. This problem of lacking a sufficiently large dataset for specific experimental questions is pervasive in medical imaging. Our model was made publicly available (https://github.com/zhangyu2ustc/GCN_fmri_decoding) and can be used as a reference model for domain adaptation, possibly making contributions in a variety of domains, including neurological and psychiatric disorders. It could also be applied in samples where extensive data is acquired on a few subjects, such as the individual brain charting (IBC) project (Pinho *et al.*, 2018) or the Courtois project on neuronal modelling (neuromod, https://docs.cneuromod.ca).

Materials and Methods

fMRI Datasets and Preprocessing

In this project, we are using the block-design task-fMRI dataset from the Human Connectome Project S1200 release. The minimal preprocessed fMRI data of the CIFTI format were used, which maps individual fMRI time-series onto the standard surface template with 32k vertices per hemisphere. The preprocessing pipelines includes two steps (Glasser *et al.*, 2013): 1) fMRIVolume pipeline generates “minimally preprocessed” 4D time-series that includes gradient unwarping, motion correction, fieldmap-based EPI distortion correction, brain-boundary-based registration of EPI to structural T1-weighted scan, non-linear (FNIRT) registration into MNI152 space, and grand-mean intensity normalization. 2) fMRISurface pipeline projects fMRI data from the cortical gray matter ribbon onto the individual brain surface and then onto template surface meshes, followed by surface-based smoothing using a geodesic Gaussian algorithm. Further details on fMRI data acquisition, task design and preprocessing can be found in (Barch *et al.*, 2013; Glasser *et al.*, 2013).

The task fMRI data includes seven cognitive tasks, which are emotion, gambling, language, motor, relational, social, and working memory. In total, there are 23 different experimental conditions. Considering the short event design nature of the gambling trials (1.5s for button press, 1s for feedback and 1s for ITI), we evaluated the decoding models (see the pipeline section below) with and without the two gambling conditions and found a much lower precision and recall scores for gambling task (average f1-score = 61%) than other cognitive domains (average f1-score $>$ 91%). In the following experiments, we excluded the two gambling conditions and only reported results on the remaining 21 cognitive states. The detailed description of the tasks can be found in (Barch *et al.*, 2013). A summary table is also shown in Table 2.

Table 2. Scanning parameters and experimental designs of HCP task-fMRI dataset.

Task Domains	#Subjects	#Runs	#Volumes per run	#Trials per run	#Conditions	Minimal duration per block (sec)
Working memory	1085	2	405	8	8	25
Motor	1083	2	284	10	5	12
Language	1051	2	316	8	2	10
Social Cognition	1051	2	274	5	2	23
Relational processing	1043	2	232	6	2	16
Emotion	1047	2	176	6	2	18

Motor task

Participants are presented with visual cues that ask them to either tap their fingers, or squeeze toes, or move the tongue. Each block of a movement type (hand, foot or tongue) is preceded by a 3s cue and lasts for 12s. In each of the two runs, there are 13 blocks in total, including 2 blocks of tongue movements, 4 of hand movements and 4 of foot movements, as well as 3 additional fixation blocks (15s) in the middle of each run.

Language task

The language task consists of two conditions, i.e. story or mathematics, with variable duration of auditory statements. During the story trials, participants listen to brief auditory stories (5-9 sentences) adapted from Aesop's fables, followed by a two-alternative-choice question and response on the topic of the story. In the math trials, participants are presented with a series of

arithmetic operations, e.g. addition and subtraction, followed by a two-alternative-choice question and response about the result of the operations. The math task is adaptive to maintain a similar level of difficulty across participants. Overall, the mathematical trials lasts around 12-15 seconds while the story trials lasts 25-30 seconds.

Working memory task

The working memory task involves two-levels of cognitive functions, with a combination of the category recognition task and N-Back memory task. Specifically, participants are presented with pictures of places, tools, faces and body parts. These 4 different stimulus types are presented in separate blocks, with half of the blocks using a 2-back working memory task (showing the same image after two image blocks) and the other half using a 0-back working memory task (showing the same image in the next block). Each of the two runs contains 8 task blocks and 4 fixation blocks (15s). Each task block consists of a 2.5s cue indicating the task type, followed by 10 task trials (2.5s each). For each task trial, the stimulus is presented for 2 seconds, followed by a 500 ms inter-task interval (ITI) when participants need to respond as target or not.

Social Cognition task

Participants are presented with short video clips of objects (squares, circles, triangles) that either interacted in some way, or moved randomly on the screen. After each video clip, participants need to judge whether the objects had a mental interaction , Not Sure, or No interaction. Each of the two runs contains 5 video blocks (20s) and 5 fixation blocks (15s). There are equal length of video blocks between the types of conditions among the 2 task runs (2 Mental and 3 Random in run 1, 3 Mental and 2 Random in run 2)

Relational Processing task

The task consists of two conditions, i.e. relational processing and matching. In the relational processing condition, participants are presented with 2 pairs of objects, which are shown in 6 different shapes and filled with 6 different textures. Participants need to first decide whether the

top pair of objects differ in shape or texture and then make the final decision whether the bottom pair differ along that same dimension. Each relational block consists of 4 task trials, where the stimuli are presented for 3500 ms followed by a 500 ms ITI. In the control matching condition, only one top pair of objects and one bottom object are presented. Additionally, the matching dimension is specified by a cue word presented in the middle of the screen (either “shape” or “texture”). Participants need to decide whether the bottom object matches either of the top objects on that dimension. Each matching block consists of 5 task trials, where the stimuli are presented for 2800 ms followed by a 400 ms ITI. In each of the two runs, there are 3 relational blocks, 3 matching blocks and 3 fixation blocks (16s). Each task block lasts 16 seconds.

Emotion Processing

The task consists of two conditions, i.e. face or shape images. Participants need to match the two images presented on the bottom of the screen to the target image whether the image shown at the top of the screen. The face images can have either an angry or fearful expression. In each of the two runs, there are 3 face blocks, 3 shape blocks and 1 fixation block (8s) at the end of each run. Each task block is preceded by a 3s task cue indicating the task type (“shape” or “face”), followed by 6 task trials (3s each). For each task trial, the stimulus is presented for 2 seconds, followed by a 1 second ITI when participants need to respond to which of the bottom images matches the target.

Convolutional Neural Networks on Brain Graphs

Graph Laplacian and graph signal processing (GSP) provides a generalized framework to analyze data defined on irregular domains, for instance social networks, biological interactions and brain graphs. A brain graph captures a network representation of brain organization by associating nodes with brain regions and defining edges via anatomical or functional connections (Bullmore and Sporns, 2009). Based on this representation, a non-linear embedding tool can be

used to project brain activity from large-scale noisy measures in the spatial domain to low-dimensional representations in the spectrum domain (Ortega *et al.*, 2018). This method has gained more and more attention in neuroscience studies, for instance parcellating brain areas (Johansen-Berg *et al.*, 2004; Fan *et al.*, 2016), identifying functional areas and networks (Craddock *et al.*, 2012; Atasoy, Donnelly and Pearson, 2016), and generating connectivity gradients (Margulies *et al.*, 2016). Recently, studies have found that, by decomposing the task-evoked fMRI signals using GSP, the resultant graph representations strongly associated with cognitive performance and learning (Huang *et al.*, 2016; Medaglia *et al.*, 2018). These findings brought new opportunities for the application of GSP on neuroimaging analysis.

Definition of Brain graph

Starting with assigning a brain signal $x \in \mathbb{R}^{N \times T}$, i.e. a short time-series with duration of T, to each of N brain regions, GSP maps the recorded brain activity onto a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$ that defines the network architecture among a set of brain regions. The set $\mathcal{V}$ is a parcellation of cerebral cortex into N regions, and $\mathcal{E}$ is a set of connections between each pair of brain regions, with its weights defined as $W_{i,j}$. Many alternative approaches can be used to build such brain graph $\mathcal{G}$, for instance using different brain parcellation schemes and constructing various types of brain connectomes (for a review, see (Bullmore and Sporns, 2009)). Here we used the multimodal cortical parcellation defined based on 210 subjects from Human Connectome Project (HCP) (Glasser *et al.*, 2016), which delineates 180 areas per hemisphere, bounded by sharp changes in cortical architecture, function, connectivity, and topography. The edges between brain areas were estimated by calculating the group averaged resting-state functional connectivity (RSFC) based on minimal preprocessed resting-state fMRI data from $N = 1080$ HCP subjects (Glasser *et al.*, 2013). Additional preprocessing steps were applied before the calculation of RSFC, including regressing out the signals from white matter and csf, and bandpass temporal filtering on frequencies between 0.01 to 0.1 HZ. Functional connectivity

was first calculated on individual brains using Pearson correlation and then normalized using Fisher z-transform before averaging among the entire group of subjects. After that, a k-nearest-neighbour (k-NN) graph was built by only connecting each brain region to its 8 neighbours with highest connectivity.

Graph Laplacian and Graph Fourier transform

The spectral analysis of the graph signals relies on the graph Laplacian, which maps the signal distributions from the spatial domain to the graph spectral domain and decomposes the signals into a series of graph modes with different frequencies. Specifically, the normalized graph Laplacian matrix is defined as:

$$L = I - D^{-1/2} W D^{-1/2} \qquad \text{(Eq. 1)}$$

where D is a diagonal matrix of node degrees and I is the identity matrix. As we assume the weights to be undirected and symmetric, the matrix L can be factored as $U \Delta U^T$, where $U = (u_0, \ldots, u_{N-1})$ is the matrix of Laplacian eigenvectors and is also called graph Fourier modes, and $\Delta = \mathrm{diag}(\lambda_0, \ldots, \lambda_{N-1})$ is a diagonal matrix of the corresponding eigenvalues, specifying the frequency of the graph modes. In other words, the eigenvalues quantify the smoothness of signal changes on the graph, while the eigenvectors indicate the patterns of signal distribution on the graph. This eigendecomposition can be interpreted as a generalization of the standard Fourier basis onto a non-Euclidean domain (Shuman *et al.*, 2013; Bronstein *et al.*, 2017). Based on the eigendecomposition, the graph Fourier transform is defined as $\hat{x} = \mathcal{L}\{x\} = U^T x$ and its inverse as $x = U\hat{x}$, where $x \in \mathbb{R}^{N \times T}$ is the graph signal and $\hat{x} \in \mathbb{R}^{K \times T}$ is the transformed signal with K selected eigenvectors or graph modes. The Laplacian matrix and these transformations are the fundamental basis of graph signal processing and graph convolutional networks.

Graph Convolutional Networks: spectral

Recently, graph convolutional neural networks (GCN) was proposed to merge graph signal processing with the deep neural network architecture (Bruna *et al.*, 2013). The key step is to generalize the convolution operations onto the graph domain. Instead of calculating a weighted sum among the spatial neighbours in the Euclidean space as in a classical convolutional neural network (Krizhevsky, Sutskever and Hinton, 2012), GCN generates a linear combination of graph Fourier modes across different frequencies by using graph filters. Specifically, the convolution between the graph signal $x \in \mathbb{R}^{N \times T}$ and a graph filter $g_\theta \in \mathbb{R}^{N \times T}$ (independent weight matrix for each temporal channel) based on graph $\mathcal{G}$, is defined as their element-wise Hadamard product in the spectral domain, i.e.:

$$x * g_\theta = U(U^T g_\theta) \odot (U^T x) = U G_\theta U^T x \qquad \text{(Eq. 2)}$$

where $G_\theta = \mathrm{diag}(U^T g_\theta)$ and θ indicates a parametric model for g_θ, and $U^T x$ is actually projecting the graph signal onto the full spectrum of graph modes. With different choices of θ GCN learns different types of graph filters and finds the optimal graph representations of the input signals for a given task.

Graph Convolutional Networks: ChebNet

To avoid calculating the spectral decomposition of the graph Laplacian, especially for large-scale graphs, ChebNet convolution (Defferrard, Bresson and Vandergheynst, 2016) uses a truncated expansion of the Chebychev polynomials, which are defined recursively by:

$$T_k(x) = 2T_{k-1}(x) - T_{k-2}(x), \qquad T_0(x) = 1, T_1(x) = x \qquad \text{(Eq. 3)}$$

Consequently, the graph convolution is defined as:

$$x * g_\theta = \sum_{k=0}^{K} \theta_k T_k(\widetilde{L}) x \qquad \text{(Eq. 4)}$$

where $\widetilde{L}$ is a normalized version of graph Laplacian, equals to $2L/\lambda_{\max} - I$, with $\lambda_{\max}$ being the largest eigenvalue, θ_k is the parameter to be learned for each order of the Chebychev polynomials,

Graph Convolutional Networks: 1st-order

Kipf and colleagues (Kipf and Welling, 2016) introduced a simplified version of GCN by taking a first-order approximation of the above Chebychev polynomial expansion and $\lambda_{\max} \approx 2$:

$$x * g_\theta = \theta(I + D^{-1/2} W D^{-1/2})x \qquad \text{(Eq. 5)}$$

where θ is a single parameter to be learned and W is the weight matrix for brain connectome.

Graph Convolutional Networks: multi-layer

Complex signal representations can be learned by stacking multiple layers of graph convolutions. The output of a graph convolution layer is defined as:

$$X^{l+1} = \sigma(\widetilde{W} X^l \Theta^l), \qquad \widetilde{W} = I + D^{-1/2} W D^{-1/2} \qquad \text{(Eq. 6)}$$

where $X^l \in \mathbb{R}^{N \times F}$ denotes the matrix of input graph signals on layer l, with N brain regions and F graph filters. To be noted that in the first graph convolution layer, F is equal to the number of temporal channels of the input graph signal T. $\Theta^l \in \mathbb{R}^{F_{in} \times F_{out}}$ is the parameters to be learned on layer l with F_{in} income filters (equals to the input temporal channels for the first graph convolution layer) and F_{out} outcome filters. These parameters are shared among all nodes on layer l. $\sigma(.)$ denotes an activation function, such as the $\mathrm{ReLU}(x) = \max(0, x)$. It's worth noting that the first-order GCN only takes into account the direct neighbours for each brain region which are indicated by the adjacency matrix of the graph. By stacking multiple GCN layers, we could propagate brain activity among the k_{th}-order neighbourhood, i.e. connecting two nodes by passing $k - 1$ other neighbours in between, with k is the number of convolution layers.

In the following analysis, we are using the multi-layer architecture of 1st-order GCN for brain decoding.

Brain State Annotation pipeline

We propose a brain state annotation model consisting of 6 graph convolutional layers with 32 graph filters at each layer, followed by a global average pooling layer and 2 fully connected layers. Specifically, in the first GCN layer, we treat the short series of fMRI volumes as multiple input channels, with $X^1 \in \mathbb{R}^{N \times T}$ being a 2D matrix consisting of N brain regions and T time steps. During model training, the first GCN layer learns various versions of the spatiotemporal convolution kernel (integrating information from graph neighbors in space, and training separate kernels for each time step) for fMRI time-series, as a replacement of the canonical hemodynamic response function (HRF). The model takes a short series of fMRI data as input, propagates information among inter-connected brain regions and networks, generates a high-order graph representation and finally predicts the corresponding cognitive labels as a multi-class classification problem. An overview of the fMRI decoding model was illustrated in Fig 1.

The entire dataset was split into training (70%), validation (10%), test (20%) sets using a subject-specific split scheme, which ensures that all fMRI data from the same subject was assigned to one of the three sets. Specifically, for each subject and each cognitive domain, individual fMRI time-series on the 64k surface template (including both hemispheres) was first mapped onto the 360 areas of Glasser atlas (Glasser *et al.*, 2016), by averaging the BOLD signals within each parcel. The time-series of each task trial was extracted and saved into a 2D matrix, by first realigning fMRI signals with experimental designs of event tasks using task onsets and durations and then cutting the time-series into bins of selected time window (see *Time window* section below). Next, the time-series matrices from all training subjects were collected into a pool of data samples. At each step of model training, a set of data samples (e.g. 128 time-series matrices) was input to the decoding model and the parameter matrix Θ^l of each layer

were optimized through gradient descent. After all data samples have been trained (i.e. finishing one epoch), the model was then evaluated on the samples from the validation set before the next epoch started. The best model with the highest prediction score on the validation set was saved and then evaluated separately on the test set. There are mainly two types of decoding models used in this study, either training by exclusively using fMRI data from a single cognitive domain or combining fMRI data from multiple cognitive domains. The rectified linear unit (ReLU) function (Maas, Hannun and Ng, 2013) was used as the activation function for all layers except the last layer where the softmax function was used to predict the cognitive labels. The network was trained for 100 epochs with the batch size set to 128. We used Adam as the optimizer with the initial learning rate as 0.001. Additional l2 regularization of 0.0005 on weights was used to control model overfitting and the noise effect of fMRI signals. Dropout of 0.5 was additionally applied to the neurons in the last two fully connected layers. The implementation of the GCN model was based on Tensorflow 1.12.0, and was made publicly available in the following repository: https://github.com/zhangyu2ustc/GCN_fmri_decoding.git .

Time window of fMRI data

As mentioned above, we treated the fMRI time windows as multiple input channels in the first layer of GCN model. There are several benefits of using multiple input channels. First, the network is enriched with more low-level graph filters, which provides more diverse features for the high-level graph convolutions. Second, with long enough fMRI time series, the network trains its own versions of the convolution kernel based on the fluctuation of task-evoked BOLD signals, as a replacement of the canonical HRF, that typically includes a small initial dip, followed by a dominant peak at 4-6s after the onset of neural activity, and then a variable post-stimulus undershoot around 8-12s after onset (Buxton, Wong and Frank, 1998). In the meantime, the different shapes of fluctuations are also informative regarding the cognitive states and could help the GCN model in state annotation.

To test this effect, we first trained a GCN model with only one input channel, i.e. using a single fMRI volume as input and predicting the cognitive label associated with that fMRI volume. It's worth noting that, according to this design, each fMRI volume during the task event (from task onset to the end of each task trial) was treated as an independent data sample. As a result, brain response at different stages of task-evoked hemodynamic response was embedded by learning multiple graph filters during model training. Thus, we could evaluate the performance of GCN annotation as a function of time-elapsed-from-onset, ranging from 0 to the length of the entire task trial. F1-score (Powers, 2011) was used as a measure of the prediction accuracy, which is the harmonic average of the precision and recall, with its best value at 1 (perfect precision and recall) and worst at 0.

Considering the low temporal signal-to-noise ratio of fMRI acquisition, especially for a single fMRI volume, we tested the same procedure with 6s of fMRI time series which includes 8 input channels at the first convolution layer. Specifically, the fMRI time-series of all task trials were first cut into non-overlapping mini-blocks of 6s time window. For instance, as for the 12s movement trials from the motor task, we compared the GCN performance in predicting different types of movements at time bins of 0-6s vs 6-12s after task onset. These short bins of time-series were treated as independent data samples during model training. For those task trials shorter than 12s, we applied a neighborhood wrapping method by using numpy.take. For instance, some of the mathematical task trials only last for 10s. In order to match the time window of the input fMRI data, we repeated the fMRI scan at the end of the task trial several times matching for 12s.

Other time windows were also evaluated, ranging from a single fMRI volume (0.72s) to the minimal duration of all task trials (10s) at a step of two TRs (1.4s). The decoding accuracies on the test set were fitted with an exponential function and summarized by averaging the performance within each cognitive domain.

Size of the dataset

The Human Connectome Project recruits 1200 healthy participants. It also provides us an opportunity to evaluate the sample size effect, i.e. how many independent subjects were sufficient to reach the stable performance of GCN. To test that, we scanned over the entire task-fMRI dataset and selected the first N complete subjects, who had completed the 7 cognitive tasks with 2 runs. The tested sample size ranges from 14 to 1060 subjects. The time window was fixed as 10s for this test.

Saliency map of graph convolutions

In addition to high classification accuracy, good interpretability is also very important for brain decoding. In our case, we need to map which discriminative features in the brain help to differentiate different cognitive task conditions. There are several ways to visualize a deep neural network, including visualizing layer activation (Springenberg *et al.*, 2014) and filters (Olah, Mordvintsev and Schubert, 2017), and heatmaps of class activation (Selvaraju *et al.*, 2017). Here, we chose the first method due to its easy implementation and generalization to graph convolutions. The basic idea is that if an input is relevant, a little variation on it will cause high change in the layer activation. This can be characterized by the gradient of the output given the input, with the positive gradients indicating that a small change to the input signals increases the output value. To visualize the gradients, we could simply use a backward pass of the activation of a single unit through the network. However, this type of map is usually very noisy, and uninversible pooling operations and nonlinear activation functions can bias the gradient. To alleviate these problems, Springenberg and his colleagues proposed to suppress the flow of gradients through neurons wherein either of input or incoming gradients were negative (Springenberg *et al.*, 2014). Specifically, for the graph signal X^l of layer l and its gradient R^l, the overwritten gradient $\nabla_{X^l} R^l$ can be calculated as follows:

$$\nabla_{X^l} R^l = (X^l > 0) \odot (\nabla_{X^{l+1}} R^{l+1} > 0) \odot \nabla_{X^{l+1}} R^{l+1} \qquad \text{(Eq. 7)}$$

In order to generate the saliency map, we started from the output layer of a pre-trained model and used the above chain rule to propagate the gradients at each layer until reaching the input layer. This guided-backpropagation approach can provide a high-resolution saliency map which has the same dimension as the input data. Since we have used multiple time channels in the first layer of the GCN model, the approach also provides one saliency map per time step. We further calculated a heatmap of saliency maps by taking the variance across the time steps for each parcel. Since each task condition can evoke different shapes of hemodynamic response, the variance of the saliency curve provides a simplified way to evaluate the contribution of task-evoked hemodynamic response. This saliency value was additionally normalized to the range [0,1], with its highest value at 1 (a dominant effect for task prediction) and lowest at 0 (no contribution to task prediction). Note that the saliency maps were generated by using the decoding model trained from a single cognitive domain with a time window as long as the event trials.

References

Alamolhoda, M., Ayatollahi, S. M. T. and Bagheri, Z. (2017) ‘A comparative study of the impacts of unbalanced sample sizes on the four synthesized methods of meta-analytic structural equation modeling’, *BMC research notes*, 10(1), p. 446.

Amalric, M. and Dehaene, S. (2016) ‘Origins of the brain networks for advanced mathematics in expert mathematicians’, *Proceedings of the National Academy of Sciences of the United States of America*, pp. 4909–4917.

Ardila, A., Bernal, B. and Rosselli, M. (2014) ‘The Elusive Role of the Left Temporal Pole (BA38) in Language: A Preliminary Meta-Analytic Connectivity Study’, *International Journal of Brain Science*. Hindawi, 2014. doi: 10.1155/2014/946039.

Atasoy, S., Donnelly, I. and Pearson, J. (2016) ‘Human brain networks function in connectome-specific harmonic waves’, *Nature communications*, 7, p. 10340.

Barch, D. M. *et al.* (2013) ‘Function in the human connectome: task-fMRI and individual differences in behavior’, *NeuroImage*, 80, pp. 169–189.

Bartley, J. E. *et al.* (2018) ‘Meta-analytic evidence for a core problem solving network across multiple representational domains’, *Neuroscience and biobehavioral reviews*, 92, pp. 318–337.

Berl, M. M. *et al.* (2010) ‘Functional anatomy of listening and reading comprehension during development’, *Brain and language*, 114(2), pp. 115–125.

Bronstein, M. M. *et al.* (2017) ‘Geometric Deep Learning: Going beyond Euclidean data’, *IEEE Signal Processing Magazine*, 34(4), pp. 18–42.

Bruna, J. *et al.* (2013) ‘Spectral Networks and Locally Connected Networks on Graphs’, *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1312.6203.

Bullmore, E. and Sporns, O. (2009) ‘Complex brain networks: graph theoretical analysis of structural and functional systems’, *Nature reviews. Neuroscience*, 10(3), pp. 186–198.

Buxton, R. B., Wong, E. C. and Frank, L. R. (1998) ‘Dynamics of blood flow and oxygenation changes during brain activation: the balloon model’, *Magnetic resonance in medicine: official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, 39(6), pp. 855–864.

Bzdok, D. *et al.* (2016) ‘Formal Models of the Network Co-occurrence Underlying Mental Operations’, *PLoS computational biology*, 12(6), p. e1004994.

Bzdok, D. and Ioannidis, J. P. A. (2019) ‘Exploration, Inference, and Prediction in Neuroscience and Biomedicine’, *Trends in neurosciences*, 42(4), pp. 251–262.

Craddock, R. C. *et al.* (2012) ‘A whole brain fMRI atlas generated via spatially constrained spectral clustering’, *Human brain mapping*, 33(8), pp. 1914–1928.

Defferrard, M., Bresson, X. and Vandergheynst, P. (2016) ‘Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering’, *arXiv [cs.LG]*. Available at:

http://arxiv.org/abs/1606.09375.

Dockès, J. *et al.* (2020) ‘NeuroQuery, comprehensive meta-analysis of human brain mapping’, *eLife*, 9. doi: 10.7554/eLife.53385.

Dubben, H.-H. and Beck-Bornholdt, H.-P. (2005) ‘Systematic review of publication bias in studies on publication bias’, *BMJ* , 331(7514), pp. 433–434.

Fan, L. *et al.* (2016) ‘The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture’, *Cerebral cortex* , 26(8), pp. 3508–3526.

Fougnie, D., Suchow, J. W. and Alvarez, G. A. (2012) ‘Variability in the quality of visual working memory’, *Nature communications*, 3, p. 1229.

Friederici, A. D. (2011) ‘The brain basis of language processing: from structure to function’, *Physiological reviews*, 91(4), pp. 1357–1392.

Glasser, M. F. *et al.* (2013) ‘The minimal preprocessing pipelines for the Human Connectome Project’, *NeuroImage*, 80, pp. 105–124.

Glasser, M. F. *et al.* (2016) ‘A multi-modal parcellation of human cerebral cortex’, *Nature*, 536(7615), pp. 171–178.

Golarai, G. *et al.* (2007) ‘Differential development of high-level visual cortex correlates with category-specific recognition memory’, *Nature neuroscience*, 10(4), pp. 512–522.

Gonzalez-Castillo, J. *et al.* (2012) ‘Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(14), pp. 5487–5492.

Gonzalez-Castillo, J. *et al.* (2015) ‘Tracking ongoing cognition in individuals using brief, whole-brain functional connectivity patterns’, *Proceedings of the National Academy of Sciences of the United States of America*, 112(28), pp. 8762–8767.

Haxby, J. V. (2001) ‘Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex’, *Science*, pp. 2425–2430. doi: 10.1126/science.1063736.

Haxby, J. V., Connolly, A. C. and Guntupalli, J. S. (2014) ‘Decoding neural representational spaces using multivariate pattern analysis’, *Annual review of neuroscience*, 37, pp. 435–456.

Haynes, J.-D. *et al.* (2007) ‘Reading Hidden Intentions in the Human Brain’, *Current Biology*, pp. 323–328. doi: 10.1016/j.cub.2006.11.072.

Horikawa, T. *et al.* (2013) ‘Neural decoding of visual imagery during sleep’, *Science*, 340(6132), pp. 639–642.

Huang, W. *et al.* (2016) ‘Graph Frequency Analysis of Brain Signals’, *IEEE journal of selected topics in signal processing*, 10(7), pp. 1189–1203.

Huth, A. G. *et al.* (2012) ‘A continuous semantic space describes the representation of thousands of object and action categories across the human brain’, *Neuron*, 76(6), pp. 1210–1224.

Johansen-Berg, H. *et al.* (2004) ‘Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex’, *Proceedings of the National Academy of Sciences of the*

United States of America, 101(36), pp. 13335–13340.

Kietzmann, T. C. *et al.* (2019) ‘Recurrence required to capture the dynamic computations of the human ventral visual stream’, *arXiv preprint arXiv:1903. 05946*. Available at: https://arxiv.org/abs/1903.05946.

Kipf, T. N. and Welling, M. (2016) ‘Semi-Supervised Classification with Graph Convolutional Networks’, *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1609.02907.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) ‘ImageNet Classification with Deep Convolutional Neural Networks’, in Pereira, F. et al. (eds) *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc., pp. 1097–1105.

Lieberman, M. D. *et al.* (2016) ‘Reply to Wager et al.: Pain and the dACC: The importance of hit rate-adjusted effects and posterior probabilities with fair priors’, *Proceedings of the National Academy of Sciences of the United States of America*, pp. E2476–9.

Lieberman, M. D. and Eisenberger, N. I. (2015) ‘The dorsal anterior cingulate cortex is selective for pain: Results from large-scale reverse inference’, *Proceedings of the National Academy of Sciences of the United States of America*, 112(49), pp. 15250–15255.

Li, H. and Fan, Y. (2019) ‘Interpretable, highly accurate brain decoding of subtly distinct brain states from functional MRI using intrinsic functional networks and long short-term memory recurrent neural networks’, *NeuroImage*, p. 116059. doi: 10.1016/j.neuroimage.2019.116059.

Lin, L. (2018) ‘Bias caused by sampling error in meta-analysis with small sample sizes’, *PloS one*, 13(9), p. e0204056.

Maas, A. L., Hannun, A. Y. and Ng, A. Y. (2013) ‘Rectifier nonlinearities improve neural network acoustic models’, in *Proc. icml*, p. 3.

Maaten, L. van der and Hinton, G. (2008) ‘Visualizing Data using t-SNE’, *Journal of machine learning research: JMLR*, 9(Nov), pp. 2579–2605.

Margulies, D. S. *et al.* (2016) ‘Situating the default-mode network along a principal gradient of macroscale cortical organization’, *Proceedings of the National Academy of Sciences of the United States of America*, 113(44), pp. 12574–12579.

Medaglia, J. D. *et al.* (2018) ‘Functional Alignment with Anatomical Networks is Associated with Cognitive Flexibility’, *Nature human behaviour*, 2(2), pp. 156–164.

Mitchell, T. M. *et al.* (2008) ‘Predicting human brain activity associated with the meanings of nouns’, *Science*, 320(5880), pp. 1191–1195.

Nishimoto, S. *et al.* (2011) ‘Reconstructing visual experiences from brain activity evoked by natural movies’, *Current biology: CB*, 21(19), pp. 1641–1646.

Olah, C., Mordvintsev, A. and Schubert, L. (2017) ‘Feature visualization’, *Distill*, 2(11), p. e7.

Orban, P. *et al.* (2015) ‘The Richness of Task-Evoked Hemodynamic Responses Defines a Pseudohierarchy of Functionally Meaningful Brain Networks’, *Cerebral cortex* , 25(9), pp. 2658–2669.

Ortega, A. *et al.* (2018) ‘Graph Signal Processing: Overview, Challenges, and Applications’,

Proceedings of the IEEE, 106(5), pp. 808–828.

Osaka, M. *et al.* (2003) ‘The neural basis of individual differences in working memory capacity: an fMRI study’, *NeuroImage*, 18(3), pp. 789–797.

Penfield, W. and Boldrey, E. (1937) ‘Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation’, *Brain: a journal of neurology*. Citeseer, 60(4), pp. 389–443.

Pinho, A. L. *et al.* (2018) ‘Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping’, *Scientific data*, 5, p. 180105.

Poldrack, R. A., Halchenko, Y. O. and Hanson, S. J. (2009) ‘Decoding the large-scale structure of brain function by classifying mental States across individuals’, *Psychological science*, 20(11), pp. 1364–1372.

Powers, D. M. (2011) ‘Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation’. Bioinfo Publications. Available at: https://dspace2.flinders.edu.au/xmlui/handle/2328/27165.

Raj, A. *et al.* (2015) ‘Network Diffusion Model of Progression Predicts Longitudinal Patterns of Atrophy and Metabolism in Alzheimer’s Disease’, *Cell Reports*, pp. 359–369. doi: 10.1016/j.celrep.2014.12.034.

Raj, A., Kuceyeski, A. and Weiner, M. (2012) ‘A network diffusion model of disease progression in dementia’, *Neuron*, 73(6), pp. 1204–1215.

Rubin, T. N. *et al.* (2017) ‘Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition’, *PLoS computational biology*, 13(10), p. e1005649.

Rubin, T. N. *et al.* (no date) ‘Decoding brain activity using a large-scale probabilistic functional-anatomical atlas of human cognition’. doi: 10.1101/059618.

Selvaraju, R. R. *et al.* (2017) ‘Grad-cam: Visual explanations from deep networks via gradient-based localization’, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626.

Shuman, D. I. *et al.* (2013) ‘The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains’, *IEEE Signal Processing Magazine*, 30(3), pp. 83–98.

Springenberg, J. T. *et al.* (2014) ‘Striving for Simplicity: The All Convolutional Net’, *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1412.6806.

Tajbakhsh, N. *et al.* (2016) ‘Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?’, *IEEE transactions on medical imaging*, 35(5), pp. 1299–1312.

Uğurbil, K. *et al.* (2013) ‘Pushing spatial and temporal resolution for functional and diffusion MRI in the Human Connectome Project’, *NeuroImage*, 80, pp. 80–104.

Van Essen, D. C. *et al.* (2013) ‘The WU-Minn Human Connectome Project: an overview’, *NeuroImage*, 80, pp. 62–79.

Varoquaux, G. *et al.* (2018) ‘Atlases of cognition with large-scale human brain mapping’, *PLOS*

Computational Biology, p. e1006565. doi: 10.1371/journal.pcbi.1006565.

Wager, T. D. *et al.* (2016) ‘Pain in the ACC?’, *Proceedings of the National Academy of Sciences*. National Acad Sciences, 113(18), pp. E2474–E2475.

Wang, X. *et al.* (2019) ‘Decoding and mapping task states of the human brain via deep learning’, *Human brain mapping*. doi: 10.1002/hbm.24891.

Wu, C.-Y., Zaccarella, E. and Friederici, A. D. (2019) ‘Universal neural basis of structure building evidenced by network modulations emerging from Broca’s area: The case of Chinese’, *Human brain mapping*, 40(6), pp. 1705–1717.

Zhang, Y. *et al.* (2017) ‘Cross-cultural consistency and diversity in intrinsic functional organization of Broca’s Region’, *NeuroImage*, 150, pp. 177–190.

Supplementary Materials - Functional Annotation of Human Cognitive States using Deep Graph Convolution

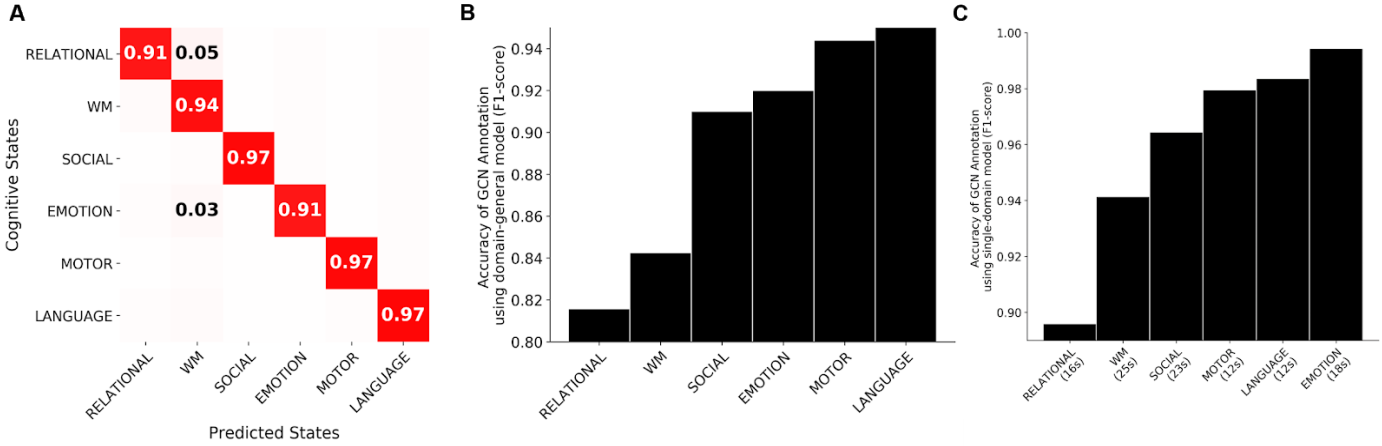

Fig 2-Supplement 1. Confusion matrix and F1-scores of the six cognitive domains. The normalized confusion matrix (A) indicates the sensitivity of each cognitive domain by averaging the recall score within each of the six domains. Relational processing and working memory showed the lowest sensitivity, with some misclassifications between emotion/relational processing and working memory tasks. A similar trend was shown in the F1-scores of GCN annotation using the decoding model either trained on multiple domains simultaneously (B) or exclusively using a single domain (C). Both of them showed the highest decoding accuracy for language and motor tasks and the lowest for relational processing and working memory tasks. Comparing the two models, a significant improvement of prediction accuracy was also shown for all cognitive domains.

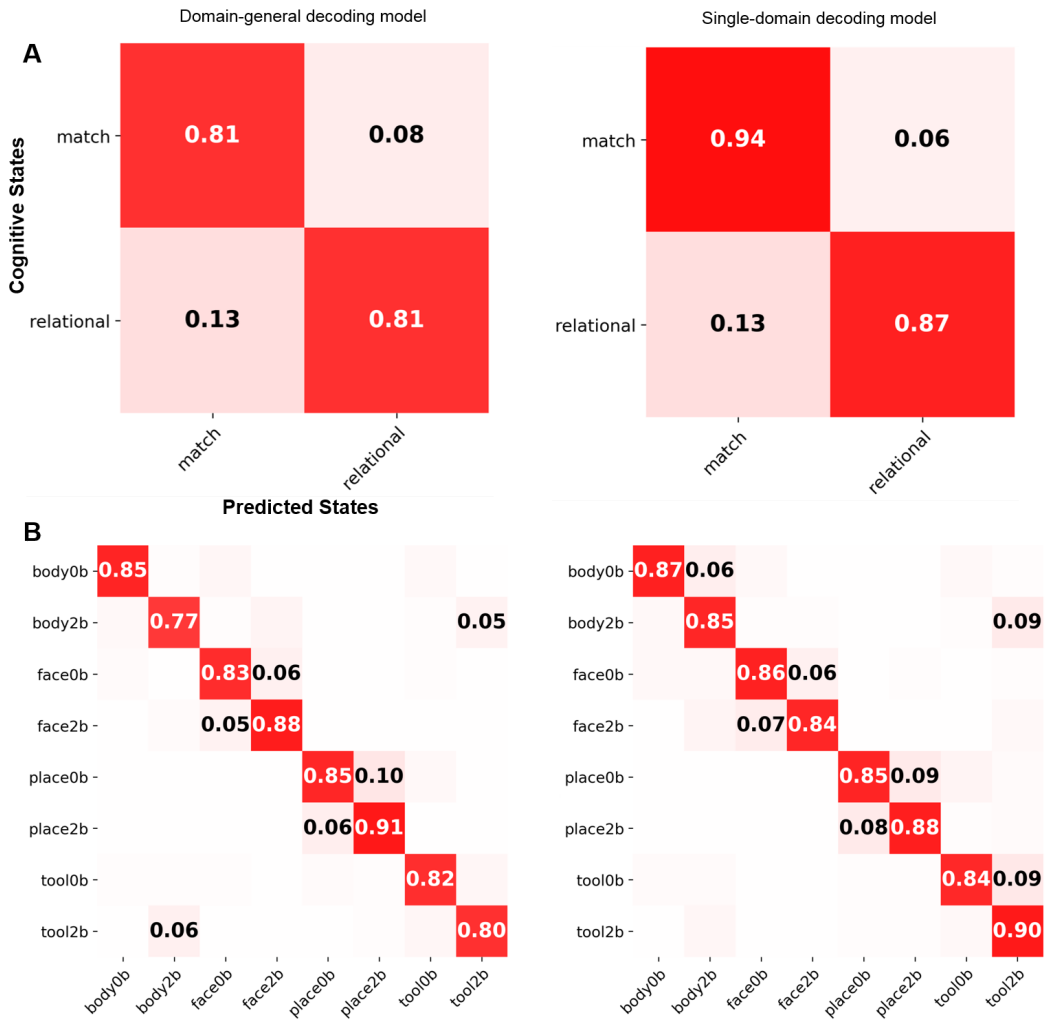

Fig 2-Supplement 2. Misclassification table for the relational processing (A) and working memory tasks (B).

The confusion matrix was either extracted from the domain-general decoding model which encodes 21 cognitive conditions simultaneously (left panel) or calculated using a separate decoding model for every single cognitive domain (right panel). ALL decoding models were trained using 10s of fMRI time series. A similar level of misclassification rates was found for the two types of decoding models, with a slight improvement of prediction accuracy for the model trained exclusively from a single domain.

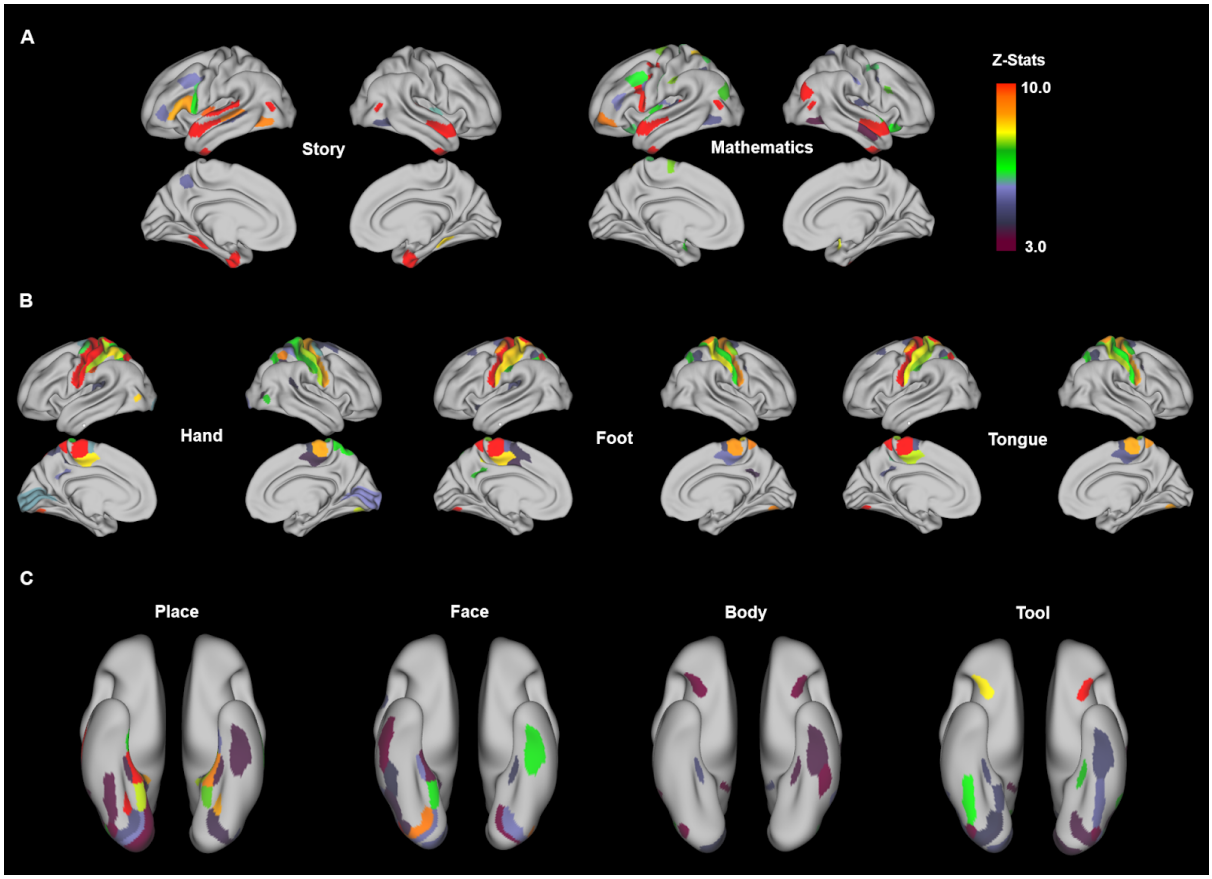

Fig 5-Supplement 1. Meta-analysis of language, motor, and working memory tasks. Meta-analysis was conducted by searching for keywords in neuroquery (Dockès *et al.*, 2020). For language task (A), we used the keyword “story” for language condition and “addition+subtraction” for the mathematical condition. For motor task (B), we used the keyword “hand movement” for hand condition, “foot+motor” for foot condition, “tongue+motor” for tongue condition. For the 0-back working memory task (C), we used the keyword “face recognition” for face condition, “body image” for body condition, “place+image” for place condition, “tool+image” for tool condition. The downloaded brain maps were first projected to the template surface “HCP_S1200_GroupAvg_v1 " using the ciftify tool (https://github.com/edickie/ciftify) and then mapped onto Glasser’s atlas (Glasser *et al.*, 2016) for visualization. Only brain parcels with z-score above 3.0 were shown here to represent significant involvement of brain regions under the corresponding condition. Note that the activation maps of the three conditions of the motor

task were not easily differentiated here mainly due to the primary motor and somatosensory cortex being parcellated into single strips in the Glasser's atlas (Glasser *et al.*, 2016).

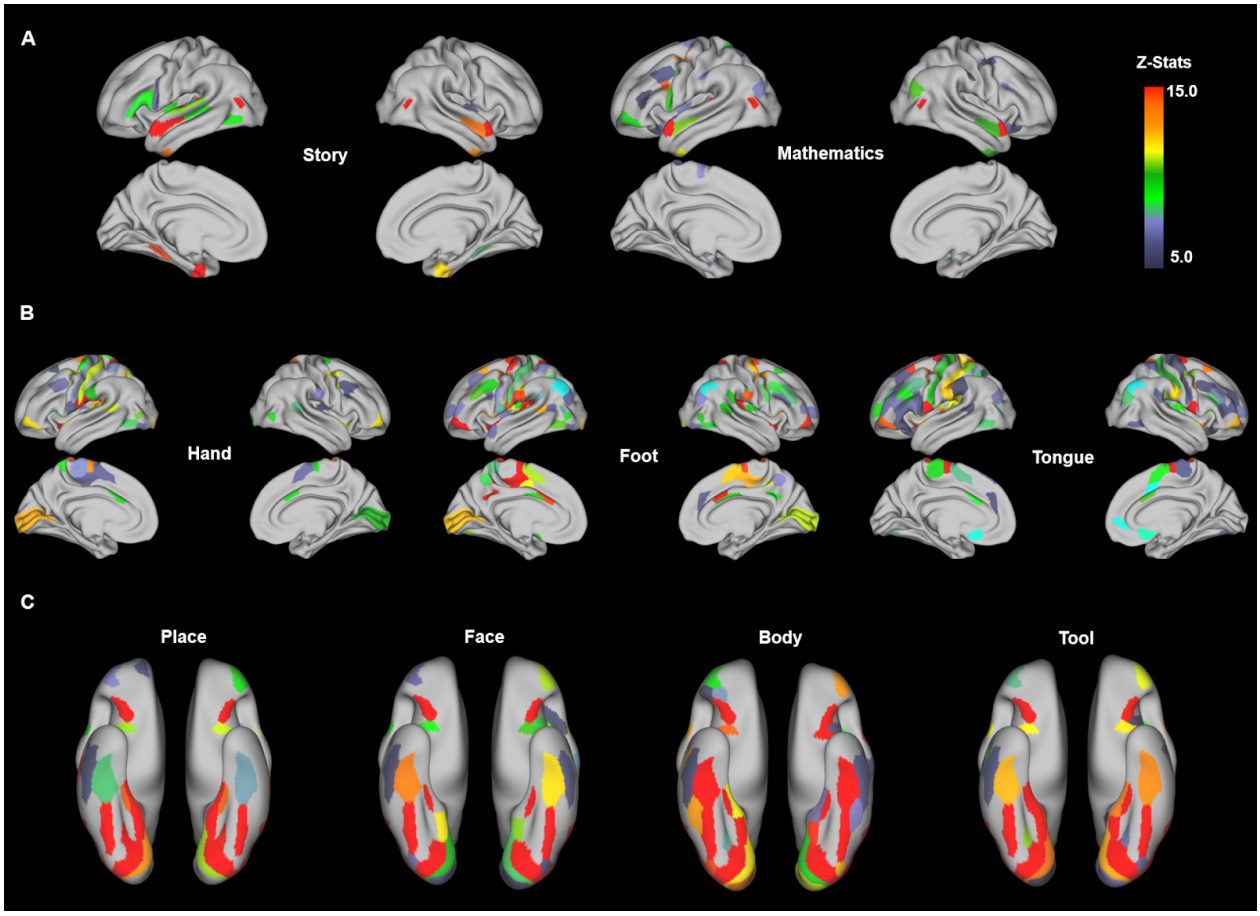

Fig 5-Supplement 2. Activation maps of language, motor and working memory tasks from HCP.

The contrast maps of HCP tasks (Barch *et al.*, 2013) were downloaded from neurovault (https://neurovault.org/collections/457/), which contained a list of group-level z-stat maps for the task conditions. For language task (A), we showed the contrast of "Story vs Baseline" and "Math vs Baseline". For motor task (B), we showed the contrast of "Right Hand vs Baseline", "Right foot vs Baseline" and "Tongue vs Baseline". For the 0-back working memory task (C), we showed the

contrast of “0back Place vs Baseline”, “0back Face vs Baseline”, “0back Body vs Baseline” and “0back Tool vs Baseline”. The downloaded contrast maps were first projected to the template surface “HCP_S1200_GroupAvg_v1 " using the ciftify tool (https://github.com/edickie/ciftify) and then mapped onto Glasser’s atlas (Glasser *et al.*, 2016) for visualization. Only brain parcels with z-score above 5.0 were shown here to represent strong brain activations under the corresponding condition.

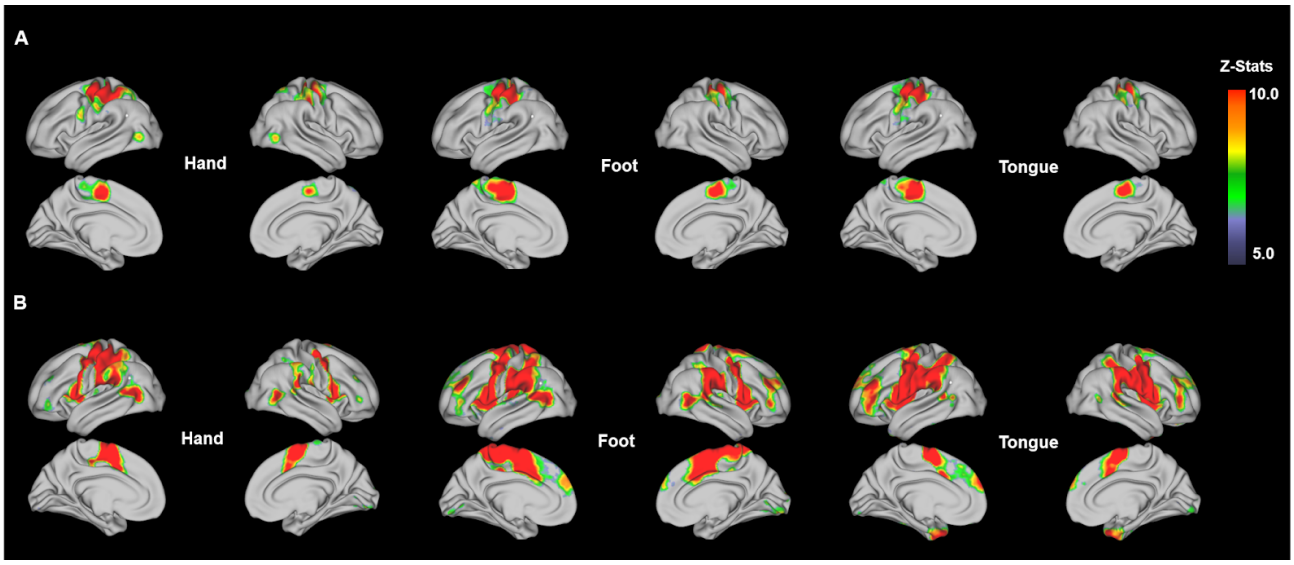

Fig 5-Supplement 3. Meta-analysis and contrast maps for the motor task

Meta-analysis (A) was conducted by searching for the keywords in neuroquery (Dockès *et al.*, 2020). We used the keyword “hand movement” for hand condition, “foot+motor” for foot condition, “tongue+motor” for tongue condition. The contrast maps of HCP tasks (B) were downloaded from neurovault (https://neurovault.org/collections/457/). We only showed the contrast of “Right Hand vs Baseline”, “Right foot vs Baseline” and “Tongue vs Baseline” here. Both activation maps from meta-analysis and contrast maps from the HCP database were projected to the template surface “HCP_S1200_GroupAvg_v1 " for visualization.

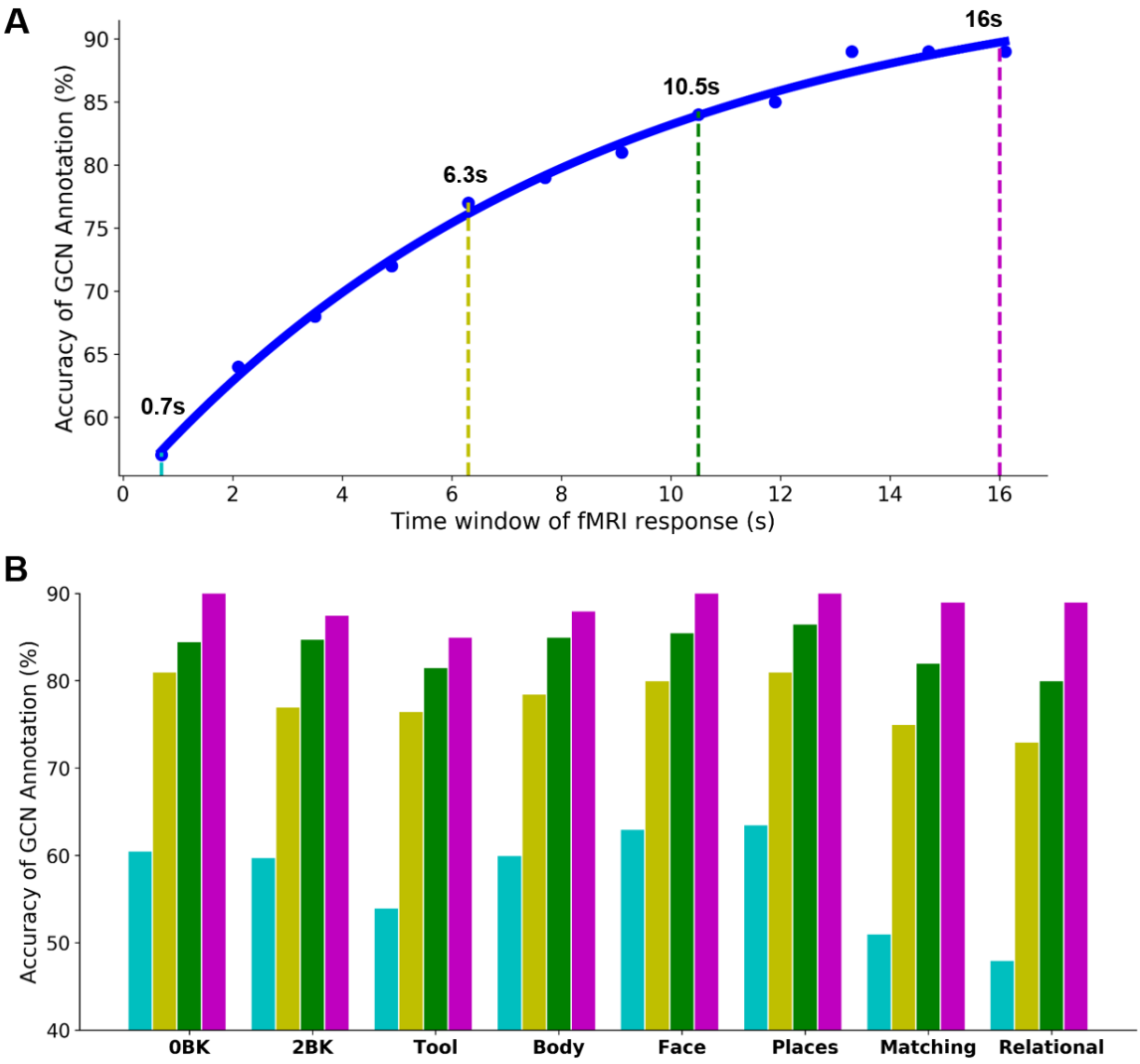

Fig 6-Supplement 1. State annotation of relational processing and working memory conditions requires more than 10s to reach a plateau.

The GCN model was trained based on the combination of all conditions from the relational processing and working memory tasks. With the minimal duration of working memory task trials lasting for 25s and relational reprocessing trials lasting for 16s, we evaluated the model with variable time windows, including a single fMRI volume (cyan: 0.7s), 9 TRs (yellow: 6.3s), 15 TRs (green: 10.5s) and 22 TRs (purple: 16s). Among all the experimental conditions, relational processing and recognition of tool images showed the lowest prediction scores at all levels of time windows.

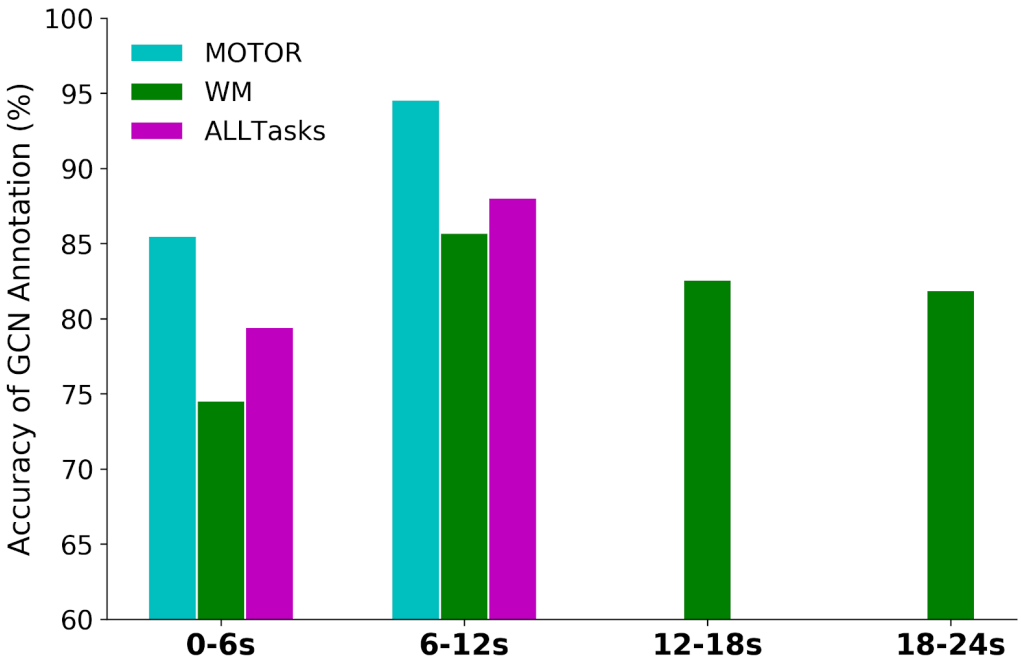

Fig 7-Supplement 1. Performance of GCN annotation using a 6s window of fMRI signals. Task trials were split into mini-blocks with a temporal duration of 6s. Event blocks from the motor task last for 12s and thus were split into 2 mini-blocks of 6s time window. Event blocks from the working memory task last for 25s and thus were split into 4 mini-blocks of 6s time windows. These mini-blocks were treated as independent samples during model training. We also trained and evaluated separate decoding models for each of the time windows, by exclusively using the fMRI time series from the corresponding time bins.