Extracting representations of cognition across neuroimaging studies improves brain decoding

Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux

Extracting representations of cognition across neuroimaging studies improves brain decoding

Arthur Mensch[1*], Julien Mairal[2], Bertrand Thirion[1], Gaël Varoquaux[1],

1 Inria, CEA, Univ. Paris Saclay, Paris, France
2 Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France

* arthur.mensch@m4x.org

Abstract

Cognitive brain imaging is accumulating datasets about the neural substrate of many different mental processes. Yet, most studies are based on few subjects and have low statistical power. Analyzing data across studies could bring more statistical power; yet the current brain-imaging analytic framework cannot be used at scale as it requires casting all cognitive tasks in a unified theoretical framework. We introduce a new methodology to analyze brain responses across tasks without a joint model of the psychological processes. The method boosts statistical power in small studies with specific cognitive focus by analyzing them jointly with large studies that probe less focal mental processes. Our approach improves decoding performance for 80% of 35 widely-different functional-imaging studies. It finds commonalities across tasks in a data-driven way, via common brain representations that predict mental processes. These are brain networks tuned to psychological manipulations. They outline interpretable and plausible brain structures. The extracted networks have been made available; they can be readily reused in new neuro-imaging studies. We provide a multi-study decoding tool to adapt to new data.

Author summary

Brain-imaging findings in cognitive neuroscience often have low statistical power, despite the availability of functional imaging data across hundreds of studies. Yet, with current analytic frameworks, combining data across studies that map responses to different tasks discards the nuances of the cognitive questions they ask. In this paper, we propose a new approach for fMRI analysis, where a predictive model is used to extract the shared information from many studies together, while respecting their original paradigms. Our method extracts cognitive representations that associate a wide variety of functions to specific brain structures. This provides quantitative improvements and cognitive insights when analyzing together 35 task-fMRI studies; the breadth of the functional data we consider is much higher than in previous work. Reusing the representations learned by our approach also improves statistical power in studies outside the training corpus.

Introduction

Cognitive neuroscience uses functional brain imaging to probe the brain structures underlying mental processes. The field is accumulating neural activity responses to

specific psychological manipulations. The diversity of studies that probe different mental processes gives a big picture on cognition [1]. However, as brain mapping has progressed in exploring finer aspects of mental processes, the statistical power of studies has stagnated or even decreased [2]—although sample size is increasing over years, it has not kept pace with the reduction of effect size. As a result, many, if not most individual studies often have low statistical power [3]. Large-scale efforts address this issue by collecting data from many subjects [4, 5]. For practical reasons, these efforts however focus on a small number of cognitive tasks. In contrast, establishing a complete view of the links between brain structures and the mental processes that they implement requires varied cognitive tasks [6], each crafted to recruit different mental processes. In this paper, we develop an analysis methodology that pools data across many task-fMRI studies to increase both statistical power and cognitive coverage. Standard meta analyses can only address *commonalities* across studies, as they require casting mental manipulations in a consistent overarching cognitive theory. They can bring statistical power at the cost of coverage and specificity in the cognitive processes. On the opposite, our approach uses the *specific* psychological manipulations of each study and extracts shared information from the brain responses across paradigms. As a result, it improves markedly the statistical power of mapping brain structures to mental processes. We demonstrate these benefits on 35 functional-imaging studies, all analyzed accordingly to their individual experimental paradigm.

Interpreting overlapping brain responses calls for multivariate analyses such as brain decoding [7]. Brain decoding uses machine learning to predict mental processes from the observed brain activity. It is a crucial tool to associate functions to given brain structures. Such inference endeavor calls for decoding across cognitive paradigms [8]. Indeed, a single study does not provide enough psychological manipulations to characterize well the functions of the brain structures that it activates [6], while covering a broader set of cognitive paradigms gives more precise functional descriptions. Moreover, the statistical power of functional data is limited by the sample size [3]. A single study seldom provides more than few hundreds of observations, which is well below machine-learning standards. Open repositories of brain functional images [9, 10] bring the hope of large-scale decoding with much larger sample sizes.

Yet, shoehorning such a diversity of studies into a decoding problem requires daunting manual annotation to build explicit correspondences across cognitive paradigms. We propose a different approach: we treat the decoding of each study as a single task in a multi-task linear decoding model [11, 12]. The parameters of this model are partially shared across studies to enable discovering potential commonalities. Model fitting—the training step of machine learning—is performed jointly, using non-convex training and regularization techniques [13, 14]. We thus learn to perform simultaneous decoding in many studies, to leverage the brain structures that they implicitly share. The extracted structures provide universal priors of functional mapping that improve decoding on new studies and can readily be reused in subsequent analyzes.

Models that generalize in measurable ways to new cognitive paradigms would ground broader pictures of cognition [15]. However, they face the fundamental roadblock that each cognitive study frames a particular question and resorts to specific task oppositions without clear counterpart in other studies [16]. In particular, a cognitive fMRI study results in *contrast* brain maps, each of which corresponds to an elementary psychological manipulation, often unique to a given protocol. Analyzing contrast maps across studies requires to model the relationships between protocols, which is a challenging problem. It has been tackled by labeling common aspects of psychological manipulations across studies, to build decoders that describe aspects of unseen paradigms [17, 18]. This annotation strategy is however difficult to scale up to a large set of studies as it requires expert knowledge on each study. The lack of complete cognitive ontologies to

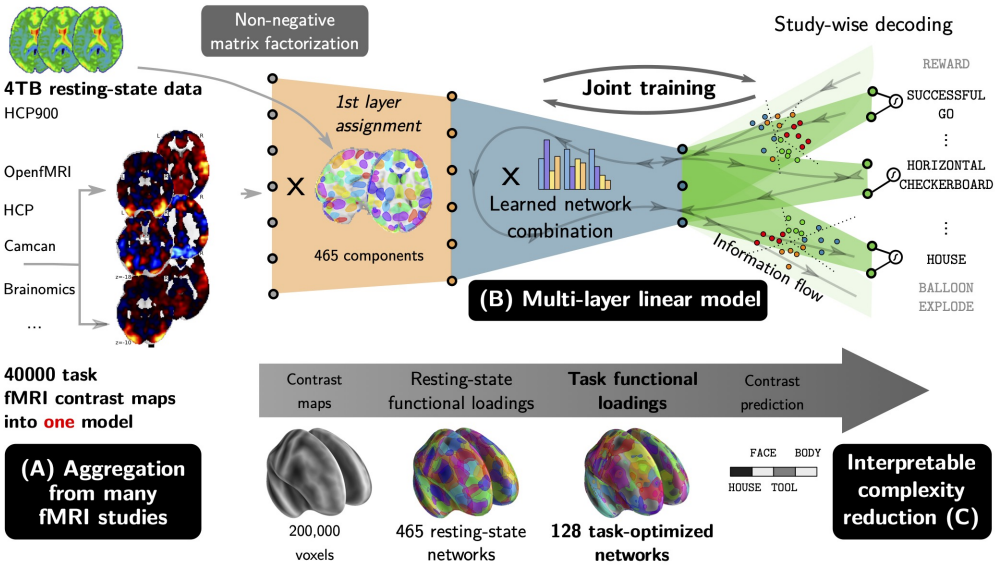


Fig. 1. General description of our multi-study decoding approach. We perform inter-subject decoding using a shared three-layer model trained on multiple studies. An initial layer projects the input images from all studies onto functional networks learned on resting-state data. Then, a second layer combines the functional networks loadings into common meaningful cognitive subspaces that are used to perform decoding for each study in a third layer. The second and third layers are trained jointly, fostering transfer learning across studies.

decompose psychological manipulations into mental processes [19] makes it even harder.

To overcome these obstacles, our multi-study decoding approach relies on the *original* labels of each study. Instead of relabeling data into a common ontology, the method extracts data-driven common cognitive dimensions. Our guiding hypothesis is that activation maps may be accurately decomposed into latent components that form the neural building blocks underlying cognitive processes [20]. This modelling overcomes the limitations of single-study cognitive subtraction models [19]. In particular, we show that it improves statistical power in individual studies: it gives better decoding performance for a vast majority of studies, and the improvement is particularly pronounced for studies with a small number of subjects. Our implicit modelling of functional information has the further advantage of providing explainable predictions. It decomposes the common aspects of psychological manipulations across studies onto latent factors, supported by spatial brain networks that are *interpretable* for neuroscience. These form by themselves a valuable resource for brain mapping: a functional atlas tuned to jointly decoding the cognitive information conveyed by various protocols. The trained model is a deep *linear* model. Building a linear model is important to bridge with classic decoding techniques in neuroimaging and ensures interpretability of intermediary representations.

Materials and methods

We first give an informal overview of the contributed methods for multi-study decoding. We review the mathematical foundations of the methods in a second part—a complete description is provided in S1A Appendix. Finally, we describe how we validate the performance and usability of the approach. A preliminary version of our method was

described in [21], with important differences and a less involved validation (S1B Appendix).

Table 1. Training and experiment set of fMRI studies. Note that even though some tasks are similar, they may feature different contrasts. Task correspondence is not encoded explicitly in our model. Supplementary Table C lists each contrast used in each study.

Study and task description	# contrasts	# subjects
[22] High level math & Localizer	31	30
[23] The ARCHI project	30	78
[24] Brainomics	19	94
[25] CamCAN	5	605
[26, 27] Music structure & Sentence structure	19	35
[28] Sentence/music complexity	25	20
[29] Balloon Analog Risk-taking	12	16
[30] Baseline trials & Classication learning	7	17
[31] Rhyme judgment	3	13
[32] Mixed-gambles	4	16
[33] Plain or mirror-reversed text	9	14
[34] Stop-signal	6	20
[35] Conditional stop-signal & Stop-signal	12	13
[36] Balloon analog risk task & Emotion regulation & Stop-signal & Temporal discounting task	23	24
[37] Classification probe without feedback & Dual-task weather classification & Single-task weather classification & Tone-counting	14	14
[38] Classification learning & Stop-signal	11	8
[38] Classification learning & Stop-signal	11	8
[39] Cross-language repetition priming	17	13
[40] Classification learning	3	13
[41] Simon task	8	7
[7] Visual object recognition	13	6
[42] Word & object processing	6	49
[43] Emotion regulation	26	34
[44] False belief	7	36
[45] Incidental encoding	26	18
[46] Covert verb generation & Line bisection & Motor & Overt verb generation & Overt word repetition	11	10
[47] Auditory oddball & Visual oddball	8	17
[48] Continuous house vs face & Discontinuous house (800ms) vs face & Discoutinuous house (400ms) vs face & House vs face	30	11
[48] Continuous house vs face & House vs face	23	13
[49] The Human Connectome Project	23	786
[50] Face recognition	5	16
[51] Arithmetic & Saccades	26	19
[52] UCLA LA5C consortium	24	189
[53] Foreign language & Localizer & Saccade	34	65
[54] Auditory compression & Visual compression	14	16
Total	545	2343

Method overview

The approach has three main components, summarized in Fig 1: aggregating many fMRI studies, training a deep linear model, and reducing this model to extract *interpretable* intermediate representations. These representations are readily reusable to apply the methodology to new data. Building upon the increasing availability of public task-fMRI data, we gathered statistical maps from many task studies, along with rest-fMRI data from large repositories, to serve as training data for our predictive model (Fig 1A). Statistical maps are obtained by standard analysis, computing z-statistics maps for either base conditions, or for contrasts of interest when available. We use

40,000 subject-level contrast maps from 35 different studies (detailed in Table 1), with 545 different contrasts; a few are acquired in cohorts of hundreds of subjects (e.g., HCP, CamCan, LA5C), but most of them feature more common sample sizes of 10 to 20 subjects. These studies use different experimental paradigms, though most recruit related aspects of cognition (e.g., motor, attention, judgement tasks, object recognition).

We use machine-learning classification techniques for inter-subject decoding. Namely, we associate each brain activity contrast map with a predicted contrast class, chosen among the contrasts of the map's study. For this, we propose a linear classification model featuring *three* layers of transformation (Fig 1B). This architecture reflects our working hypothesis: cognition can be represented on basic functions distributed spatially in the brain. The first layer projects contrast maps onto $k = 465$ functional units learned from resting-state data. This first dimension reduction should be interpreted as a projection of the brain signal onto small, smooth and connected brain regions, tuned to capture the resting-state brain signal with a fine grain. The second layer performs dimension reduction and outputs an embedding of the brain activity into $l = 128$ features that are *common* across studies. The embedded data from each study are then classified into their respective contrast class using a study-specific classification output from the third layer, in a setting akin to multi-task learning (see [55] for a review).

The second layer and the third layer are jointly extracted from the task-fMRI data using regularized stochastic optimization. Namely, the shared brain representation is optimized simultaneously with the third layer that performs decoding for every study. In particular, we use dropout regularization [56] in the layered model and stochastic optimization [13] to obtain good out-of-sample performance.

Study-specific decoding is thus performed on a shared low-dimensional brain representation. This representation is supported on 128 different combinations of the first 465 functional units identified with resting-state data. These combinations form diffuse networks of brain regions, that we call *multi-study task-optimized networks* (MSTONs). MSTONs differ from the notion of brain networks in the neuroscience literature—the later are typically obtained using a low-rank factorization of resting-state data, with a much lower number of components ($k \approx 20$) than what we use to extract the *functional units* of the first layer.

As we will show, projecting data onto MSTONs improves across-subject predictive accuracy, removing confounds while preserving the cognitive signal. Interpretability is guaranteed by the linearity of the model and a post-training identification of stable directions in the space of latent representations. These networks capture a general multi-study representation of the cognitive signal contained in statistical maps.

Mathematical modelling

Following this informal description, we now review the mathematical foundations of our decoding approach. The complete descriptions of the predictive models and of the training algorithms are provided in S1A Appendix.

We consider N task-fMRI studies, that we use for functional decoding. In this setting, each study j features n^j subjects, for which we compute c^j different contrasts maps, using the General Linear Model [57]. Masking them using a grey-mask filter in the MNI space, we obtain a set of z-maps $(\boldsymbol{x}_j^i)_{j\in[1,c^j n^j]}$, in $\mathbb{R}^p$, that summarizes the effect on brain activations of the psychological conditions $(y_i^j)_{i\in[1,c^j n^j]}$. The goal of functional decoding is to learn a predictor from z-maps to psychological conditions, namely a function $f^j : \mathbb{R}^p \to [1, c^j]$. This predictor will be evaluated on unseen subjects for validation.

Linear decoding with shared parameters. In our setting, we couple the predictors $(f^j)_{j\in[N]}$ by forcing them to share parameters. Each study corresponds to a classification task, and we cast the problem as multi-task learning (as first considered in [11]). For this, we consider a given z-map $\boldsymbol{x}_i^j$ in study j. We compute the predicted psychological condition using a factorized linear model:

$$\hat{y}_i^j = f^j(\boldsymbol{x}_i^j) = \underset{k\in[1,c^j]}{\operatorname{argmax}} \left(\boldsymbol{U}^j\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j + \boldsymbol{b}^j\right)_k.$$

The matrix $\boldsymbol{D} \in \mathbb{R}^{k\times p}$ and $\boldsymbol{L} \in \mathbb{R}^{l\times k}$ contain the basis for performing two successive projection of the z-map $\boldsymbol{x}_j^i$ onto low-dimension spaces. Those parameters are shared over all studies $j \in [N]$ and form the first and second layer of our model. The matrix $\boldsymbol{U}^j \in \mathbb{R}^{l\times c^j}$ and the bias vector $\boldsymbol{b}^j \in \mathbb{R}^{c^j}$ are the parameters of a multi-class linear classification model that labels the projected map $\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j$ with a psychological condition within the study j. Those parameters are specific to each study j, and form the third layer of our model.

First layer training from resting-state data. The first dimension reduction, contained in the matrix $\boldsymbol{D} \in \mathbb{R}^{k\times p}$, is learned using external resting-state data, from the HCP project [4]. Voxel time-series are stacked in a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ (with 4 millions brain-images), that is factorized so that $\boldsymbol{X} \approx \boldsymbol{D}\boldsymbol{A}$, with $\boldsymbol{D}$ non-negative and sparse (i.e. with mostly null coefficients). This forces the elements of $\boldsymbol{D}$ to delineate localized functional units. We use a sparse non-negative matrix factorization objective [58] and a recent scalable matrix factorization algorithm [59] to learn $\boldsymbol{D}$, as detailed in S1A Appendix. The non-negativity constraint allows to interpret functional units as a soft parcellation of the brain. We do not use additional spatial constraints, as non-negative sparse matrix factorization with $k = 465$ components readily finds smooth connected regions.

Joint training of the second and third layer. The matrix $\boldsymbol{L}$ and the multiple matrices $(\boldsymbol{U}^j)_{j\in[n]}$ and intercepts $(\boldsymbol{b}^j)_j$ are trained jointly to minimize the objective

$$\min_{\boldsymbol{L},\{\boldsymbol{U}^j,j\in[N]\}} \sum_{j=1}^{N} \frac{1}{n^j} \sum_{i=1}^{c^j n^j} \ell_j(\boldsymbol{U}^j\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j + \boldsymbol{b}^j, y_i^j),$$

where ℓ_j is the standard ℓ_2-regularized multinomial loss function for training a linear model with c^j classes (see S1A Appendix for details). This objective is trained using Adam [13]; at each step, we select a batch of examples from one study. To prevent specialization of the rows of matrix $\boldsymbol{L}$ to specific studies, we add a dropout noise [14] to the activations $\boldsymbol{D}\boldsymbol{x}_i^j$ and $\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j$ during training.

Model consensus. Although the atoms of $\boldsymbol{D}$ are naturally interpretable, the fact that the product $\boldsymbol{U}^j\boldsymbol{L}$ can always be rewritten as $\boldsymbol{U}^j\boldsymbol{M}^{-1}\boldsymbol{M}\boldsymbol{L}$ for an invertible matrix $\boldsymbol{M}$ prevents us from directly identifying meaningful directions in the low-dimensional space spanned by $\boldsymbol{L}\boldsymbol{D}$. On the other hand, we found this space to be remarkably stable across training runs. We therefore propose an ensemble technique to extract a non-negative matrix $\bar{\boldsymbol{L}} \in \mathbb{R}^{l\times k}$ such that $\bar{\boldsymbol{L}}\boldsymbol{D}$ captures meaningful directions (as above-mentioned non-negativity enables us to interpret MSTONs as soft brain parcellations).

For this, we train R decoding models with different sampling order and initialization, to obtain $(\boldsymbol{L}_r)_{r\in[R]}$. We stack these matrices into a tall matrix $\tilde{\boldsymbol{L}} \in \mathbb{R}^{lR\times k}$, that we factorize as $\tilde{\boldsymbol{L}} = \boldsymbol{K}\bar{\boldsymbol{L}}$, with $\bar{\boldsymbol{L}} \in \mathbb{R}^{l\times k}$ non-negative and sparse. This is in turn (see S1A Appendix) yields a consensus model $(\boldsymbol{D}, \bar{\boldsymbol{L}}, (\bar{\boldsymbol{U}}^j, \bar{\boldsymbol{b}}^j)_{j\in[N]})$, where $\bar{\boldsymbol{L}}\boldsymbol{D} \in \mathbb{R}^{l\times p}$ is sparse

and non-negative. It therefore holds interpretable brain networks, learned in a supervised manner from many studies—those form the MSTONs.

Layer widths. We chose $k = 465$ and $l = 128$ as those are a good compromise between model performance and interpretability—trade-offs in choosing the number of functional units k for fMRI analysis are discussed in e.g. [60], and we compare the model performance for different l in Fig E, S1B Appendix. Choosing l smaller than the number of classes enforces a low-rank structure over the set of 545 classification maps.

Validation

Quantitative measurements. The benefits of multi-study decoding may vary from study to study, and a single number cannot properly quantify the impact of our approach. We measure decoding *accuracy* on left-out subjects (half-split, repeated 20 times) for each study. For each split and each study, we compare the scores obtained by our model to results obtained by simpler baseline decoders, that classify contrast maps separately for each study, and directly from voxels. To analyse the impact of our method on the prediction accuracy specifically for each contrast, we also report the *balanced-accuracy* for each predicted class. For completeness, we report mean accuracy gain and the number of studies for which multi-study decoding improves accuracy—those hint at the benefit that one may expect when applying the method to a new fMRI study. Mathematical definitions of the metrics in use are reported in Section C.2.

Exploring MSTONs. Our model optimizes its second and third layers to project brain images on representations that help decoding. These representations boil down to MSTONs combinations: MSTONs form a valuable output of the model, as they can easily be reused to project data for new decoding tasks. We provide 2D and 3D views of the MSTONs, showing how they cover the brain. We evaluate the importance of each network for decoding a certain contrast by computing the cosine similarity between the MSTON and the classification map associated with this contrast. We represent these contrasts' names as specified in their original studies with word-clouds, with a size increasing with their similarity with a given MSTON.

Classification maps. As our model is linear, we qualitatively compare the classification maps that it yields with maps obtained with a baseline single-study voxel-level decoding approach. For both approaches, we compute the correlation matrix between classification maps to uncover potential clusters of similar maps, using hierarchical clustering [61]. We compare this correlation matrix in term of how clustered it is, using the cophenetic correlation coefficient [62] and the mean absolute cosine similarity between maps.

Reusable tools and resources

Our approach can be used to improve statistical power of decoding in new fMRI studies. To facilitate its use, we have released resources and the *cogspaces* library (`http://cogspaces.github.io`). We include software to train the models. Pre-trained MSTONs networks (with associated word-clouds) can be downloaded and inspected on a dedicated page (`https://cogspaces.github.io/assets/MSTON/components.html`).The statistical maps used in the present study may be downloaded using our library, or on

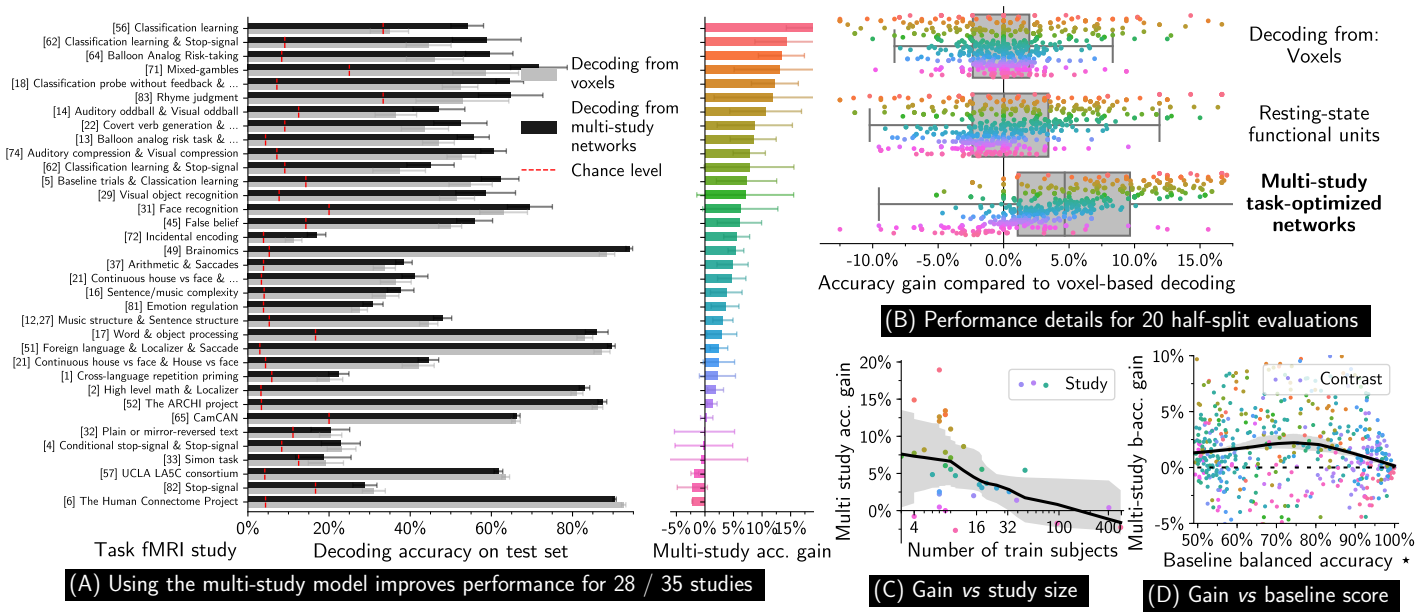


Fig. 2. Quantitative performance of multi-study decoding. (A) Multi-study decoding improves the performance of cognitive task prediction across subjects for most studies. **(B)** Overall, decoding from task-optimized networks leads to a mean improvement accuracy of 5.8% compared to voxel or networks based approaches. Each point corresponds to a study and a train/test split. **(C)** Studies of typical size strongly benefit from transfer learning, whereas little information is gained for very large studies. **(D)** Contrasts that are moderately difficult to decode benefit most from transfer. Error bars are calculated over 20 random data half-split. ⋆**(D)** shows *per-contrast* balanced accuracy (50% chance level), whereas *per-study* classification accuracy is used everywhere else. Numbers are reported in Table A

neurovault.org. The published MSTON networks hold the representations extracted from the 35 studies that we have considered; some of them are shown in Fig 4.

Results

We first detail the quantitative improvements brought by our approach, before exploring these results from a cognitive neuroscience point of view.

Improved statistical performance of multi-study decoding

Decoding from multi-study task optimized networks gives quantitative improvements in prediction of mental processes, as summarized in Fig 2. For 28 out of the 35 task-fMRI studies that we consider, the MSTON-based decoder outperforms single-study decoders (Fig 2A). It improves accuracy by 17% for the top studies, with a mean gain of 5.8% (80% experiments with net increase, 4.8% median gain) across studies and cross-validation splits (Fig 2B). *Jointly* minimizing errors on every study constructs second-layer representations that are efficient for many study-specific decoding tasks; the second layer parameters therefore incorporate information from all studies. This shared representation enables information transfer among the many decoding tasks performed by the third layer—predictive accuracy is thus improved thanks to *transfer learning*. Although we have not explicitly modeled how mental processes or

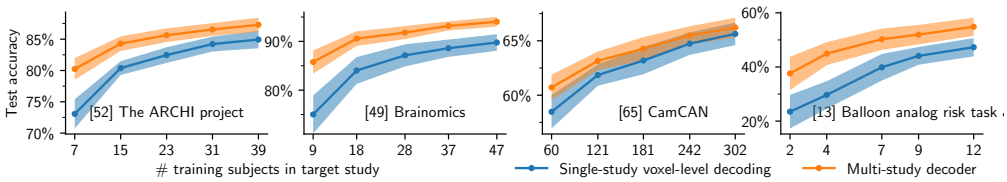


Fig. 3. Varying accuracy improvement with study size. Training an MSTON decoder increases decoding accuracy for many studies (see Fig 2A). Gains are higher as we reduce the number of training subjects in target studies—pooling multiple studies is especially useful to decode studies performed on small cohorts. Error bars are calculated over 20 random data half-splits.

psychological manipulations are related across experiments, our quantitative results show that these relations can be captured by the model—encoded into the second layer—to improve decoding performance.

Studies with diverse cognitive focus benefit from using multi-study modeling. The different decoding tasks have varying difficulties—we report performance sorted by chance level in Fig L, S1B Appendix. Among the highest accuracy gains, we find cognitive control (stop-signal), classification studies, and localizer-like protocols. Our corpus contains many of such studies; as a result, multi-study decoding has access to many more samples to gather information on the associated cognitive networks. The activation of these networks is better captured in the shared part of the model, thereby leading to the observed improvement. In contrast, for a few studies, among which HCP and LA5C, we observe a slight negative transfer effect. This is not surprising—as HCP holds 900 subjects, it may not benefit from the aggregation of much smaller studies; LA5C focuses on higher-level cognitive processes with limited counterparts in the other studies, which precludes effective transfer.

Fig 2B shows that simply projecting data onto resting state functional networks instead of using our three-layer model does not significantly improve decoding, although the net accuracy gain varies from study to study. Adding a second *task-optimized*—supervised—dimension reduction is thus necessary to improve overall decoding accuracy. Functional contrasts that are either easy or very hard to decode do not benefit much from multi-study modeling, whereas classes with a balanced-accuracy around 80% experience the largest decoding improvement (Fig 2). We attribute this to two causes: easy-to-decode studies do not benefit from the extra signal provided by other studies, while some studies in our corpus are simply too hard to decode due to a low signal-to-noise ratio. Fig 2D shows that the benefit of multi-study modeling is higher for smaller studies, confirming that the proposed method boosts their inter-subject decoding performance.

In Fig 3, we vary the number of training subjects in target studies, and compare the performance of the multi-study decoder with a more standard one. We observe that the smaller the study size, the larger the performance gain brought by multi-study modeling. Transfer learning in inter-subject decoding is thus particularly effective for small studies (e.g., 16 subjects), that still constitute the essential of task-fMRI studies. To confirm this effect, we trained a multi-study model on a subset of 15 subjects per study, considering studies that comprise more than 30 subjects. In this case, the transfer learning effect is positive for all studies (Fig K in S1B Appendix), including those for which negative transfer was observed when using full cohorts.

Finally, we show in Fig B (S1B Appendix) that training a three-layer model and reusing the first two layers as a fixed dimension reduction when decoding a new study improves decoding accuracy on average. The extracted functional networks (MSTONs) thus provide a study-independent prior that is likely to improve decoding for studies

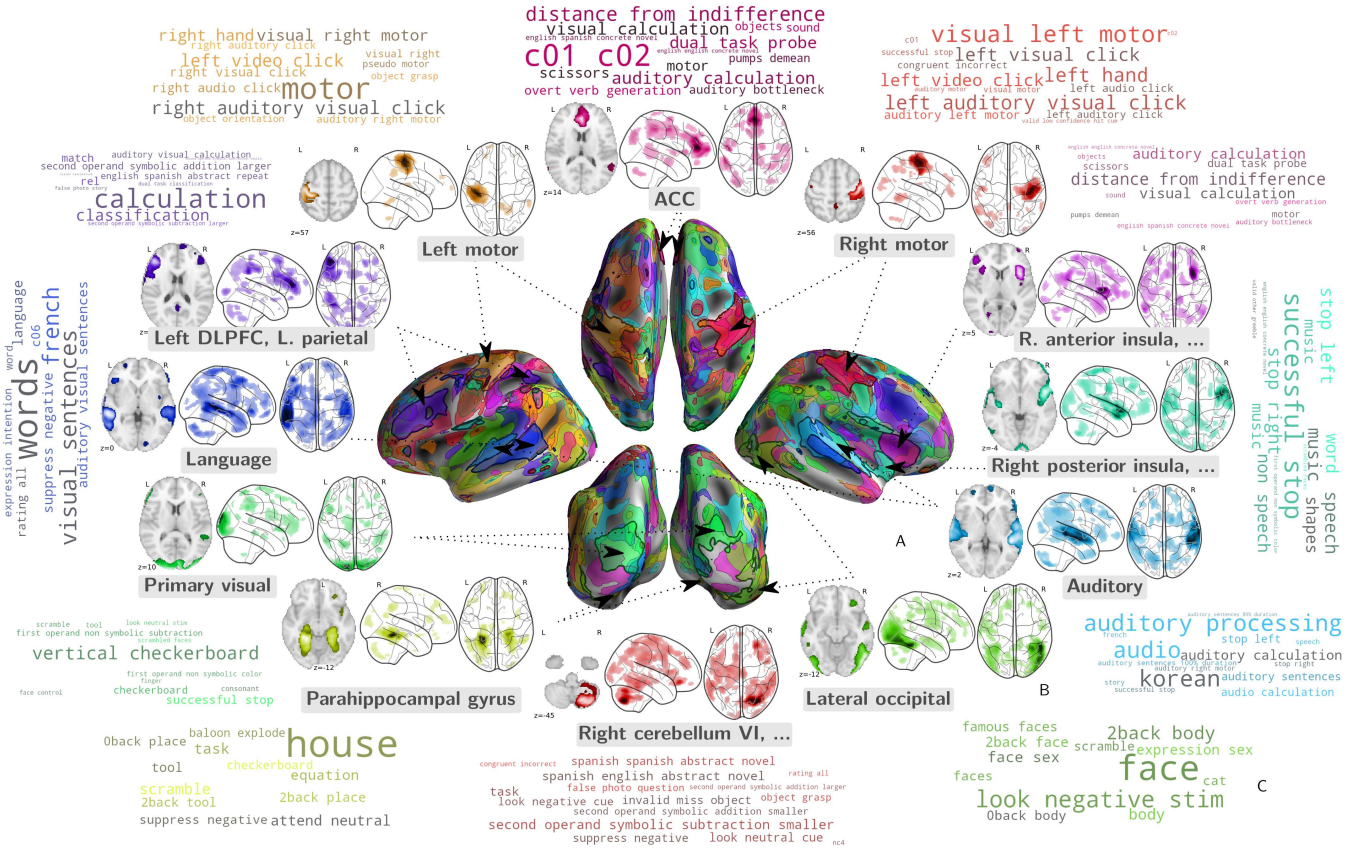


Fig. 4. Visualization of some of task-optimized networks. Our approach learns networks that are important for decoding across studies. These networks are individually focal and collectively well spread across the cortex. They are readily associated with the cognitive tasks that they contribute to predict. We display a selection of these networks on the cortical surface **(A)** and in 2D transparency **(B)**, named with the salient anatomical brain region they recruit, along with a word-cloud **(C)** representation of the stimuli whose likelihood increases with the network activation. The words in this word cloud are the terms used in the contrast names by the investigators; they are best interpreted in the context of the corresponding studies.

probing different cognitive questions than the ones considered in the training corpus.

Multi-study task-optimized networks capture broad cognitive domains

We outline the contours of the 128 extracted MSTONs in Fig 4A. The networks almost cover the entire cortex, a consequence of the broad coverage of cognition of the studies we gathered. Task-optimized networks must indeed capture information to predict 545 different cognitive classes from the resulting distributed brain activity. Brain regions that are systematically recruited in task-fMRI protocols, e.g., motor cortex, auditory cortex, and primary visual cortex, are finely segmented by MSTON: they appear in several different networks. Capturing information in these regions is crucial for decoding many contrasts in our corpus, hence the model dedicates a large part of its representation capability to it. As decoding requires capturing distributed activations, MSTON are formed of multiple regions (Fig 4B). For instance, both parahippocampal

gyri appear together in the yellow bottom-left network.

Most importantly, Fig 4B-C show that the model relates extracted MSTONs to specific cognitive information. The MSTONs each play a role in decoding a subset of contrasts. Components may capture low-level or high-level cognitive signal, though the low-level components are easier to interpret. Indeed, at a lower level, they outline the primary visual cortex, associated with contrasts such as checkerboard stimuli, and both hand motor cortices, associated with various tasks demanding motor functions. At a higher level, some interpretable components single out the left DLPFC and the IPS in separate networks, used to decode tasks related to calculation and comparison. Others delineate the language network and the right posterior insula, important in decoding tasks involving music [27]. Yet another MSTON delineates Broca's area, associated with language tasks (Fig 5).

Inspecting the tasks associated with the MSTONs reveals structure-function links. Once again, the results are more interpretable for low-level functions, although some well-known high-level functional associations are also well captured. For instance, several components on Fig 4 involve brain regions recruited across a wide variety of tasks, such as the anterior insula, engaged in auditory and visual tasks [63] and considered to tackle ambiguous perceptual information, or the ACC, associated with tasks with affective components [64] and reward-based decision making [65]. Some MSTONs are more distributed, but correspond to well-known patterns brain activity. For example, Fig 5 show components that reveal parts the default mode networks –associated with baseline conditions, theory-of-mind tasks and prospection [66, 67]–, parts of the fronto-parietal control network –associated with a variety of problem-solving tasks [68]– and the dorsal attentional network –associated with visuo-spatial attention tasks such as saccades [69].

Visualizing MSTONs along with word-clouds serves essentially an illustratory purpose. It yields more interpretable results with focal networks than with distributed networks. In both cases, the words in the contrasts related to the given MSTONs capture documented structure-function associations. Interpretability may be improved by reducing the number of extracted networks, at the cost of a quantitative loss in performance. In particular, with $k = 128$ components, the default mode network is split across several MSTONs (Fig 5). Such a splitting is common for high-dimensional decomposition of the fMRI signal, as noted in resting state [70], as a network such as the default-mode network has different sub-units with distinct functional contributions [71]. Conversely, some contrast maps are correlated with several distributed MSTONs, as illustrated in Fig A (S1B Appendix).

Impact of multi-study modeling on classification maps

To better understand how multi-study training and layered representations improve decoding performance, we compare classification maps obtained using our model to standard decoder maps in Fig 6. For contrasts with significant accuracy gains, the classification maps are less noisy and more focal. They single out determinant regions more clearly, e.g., the fusiform face area (FFA, row 1) in classification maps for the face-vs-house contrast, or the left motor cortex in maps (row 2) predicting pumping action in BART tasks [29]. The language network is typically better delineated by our model (row 3), and so is the posterior insula in music-related contrasts (row 4). These improvements are due to two aspects: First, projecting onto a lower dimension subspace has a denoising effect on contrast maps, that is already at play when projecting onto simple resting-state functional networks. Second, multi-study training finds more scattered classification maps, as these combine complex MSTONs, learned on a large set of brain images. Our method slightly decreases performance for a small fraction of contrasts, such as maps associated with vertical checkerboard (row 5), a condition well

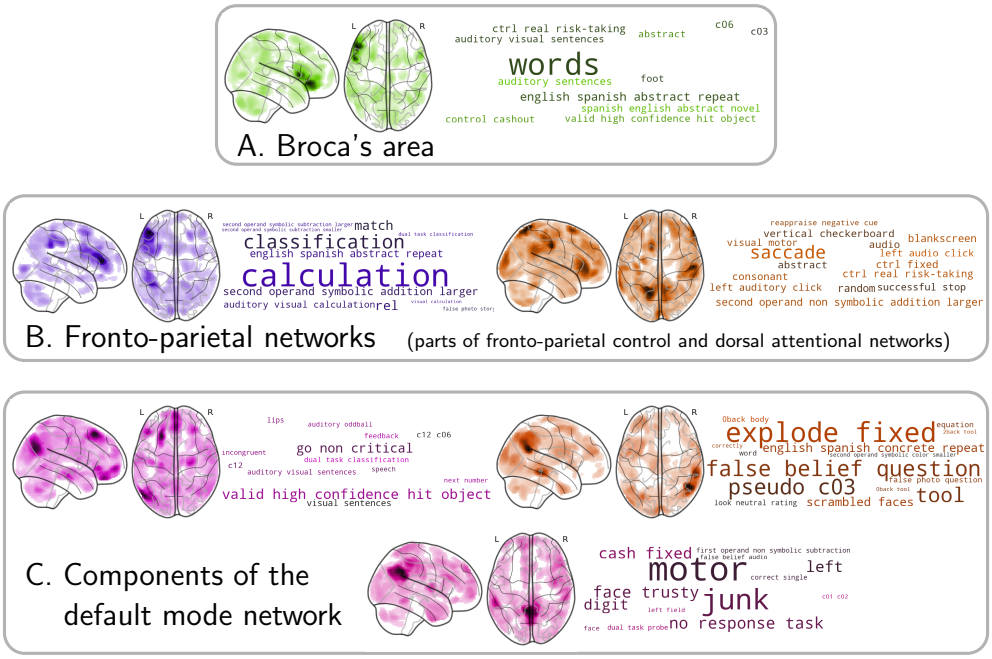


Fig. 5. Task-optimized networks associated with high-level functions. Some MSTONs outline brain-circuits that are associated with language, e.g. Broca's area **(A)**, or more abstract functions, e.g. fronto-parietal networks **(B)** or even part of the default mode network **(C)**. Those networks are more distributed than the ones displayed in Fig 4, but are associated with relatively interpretable word-clouds.

localized and easy to decode from the original data. Our model renders them too much distributed, an unfortunate consequence of multi-study modeling.

We also compare original input contrast maps to their transformation by the projection on task-optimized networks (Fig C in S1B Appendix). Projected data are more focal, i.e. spatial variations that are unlikely to be related to cognition are smoothed. This offers a new angle on the quantitative results (Fig 2): brain activity expressed as the activation of these networks captures better cognition and allows decoders to generalize better across subjects than when classifying raw input directly.

Information transfer among classification maps. In Fig 7, we compare the correlation between the 545 classification maps obtained using a multi-study decoder and using simple functional networks decoders. Classification maps learned using task-optimized networks are more correlated on average, and hierarchical clustering reveals a sharper correlation structure. This is because the whole classification matrix is low-rank (rank $l = 128 < c = 545$) and influenced by the many studies we consider—the classification maps of our model are supported by networks relevant for cognition. As a consequence, it is easier to cluster maps into meaningful groups using hierarchical clustering based on cosine distances. For instance, we outline inter-study groups of maps related to left-motor functions, or calculation tasks. Hierarchical clustering on baseline maps is less successful: the associated dendrogram is less structured, and the distortion introduced by clusters is higher (as suggested by the smaller cophenetic coefficient). Clusters are harder to identify, due to smaller contrast in the correlation matrix. Multi-study training thus acts as a regularizer, by forcing correlation across maps with discovered relations. This regularization partly explains the increase in

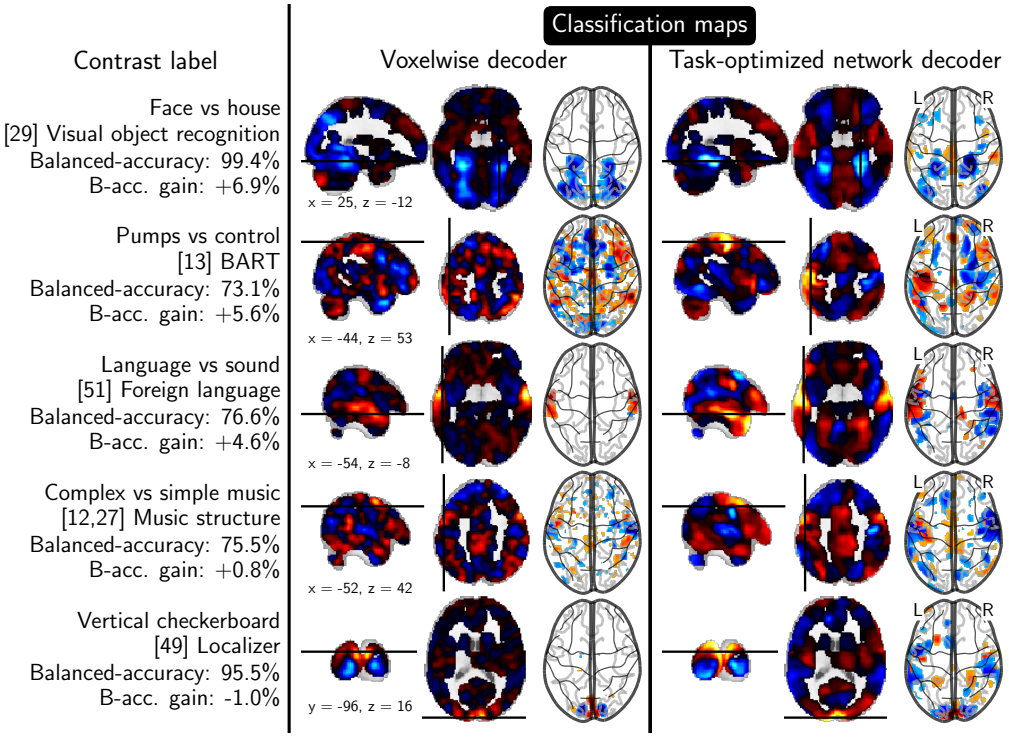


Fig. 6. Classification maps obtained from multi-study decoding (right). The maps are smoother and more focused on functional modules than when decoding from voxels (left). For contrasts for which there is a performance boost (top of the figure), relevant brain regions are better delineated, as clearly visible on the face vs house visual-recognition opposition, in which the fusiform gyrus stands out better. B-acc stands for balanced accuracy using multi-study decoding (see text).

decoding accuracy.

Discussion

The methodology presented in this work harnesses the power of deep representations to build multi-study decoding models for brain functional images. It brings an immediate benefit to functional brain imaging by providing a universal way to improve the accuracy of decoding in a newly acquired dataset. Decoding is a central tool to draw inferences on which brain structures implement the neural support of the observed behavior. It is most often applied to task-fMRI studies with 30 or less subjects, which tend to lack statistical power [72]. In this regime, aggregating existing studies to a new one using a multi-study model as the one we propose is likely to improve decoding performance. This is further evidenced in Fig B (S1B Appendix): using MSTONs as a decoding basis on a new decoding task outperforms using resting-state networks. Of course, such improvement can only occur if the cognitive functions probed by the new study are related to the ones probed in the multi-study corpus. We foresee limited benefits when analyzing strongly original task fMRI experiments, and experiments studying very specific and high-level cognitive functions, that MSTONs are only partially able to capture (Fig 5).

With increasing availability of shared and normalized data, multi-study modeling is an important improvement over simple decoders, provided that it can adapt to the

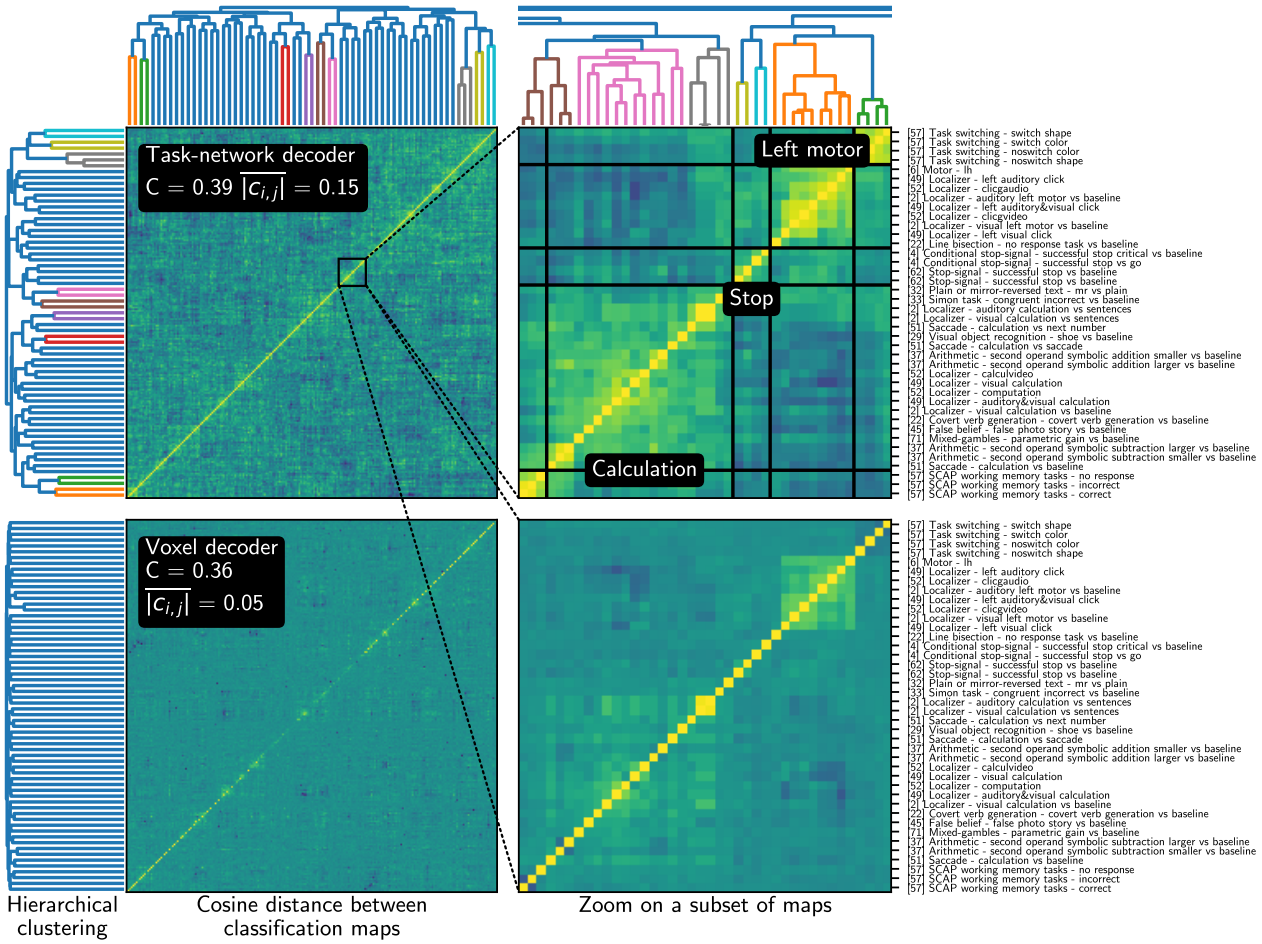


Fig. 7. Cosine similarities between classification maps, obtained with our multi-study decoder (top) and with decoders learned separately (bottom), clustered using average-linkage hierarchical clustering. The classification maps obtained when decoding from task-optimized networks are more easily clustered into cognitive-meaningful groups using hierarchical clustering—the cophenetic coefficient of the top clustering is thus higher. Maps may also be compared using the similarities of their loadings on MSTONs, with similar results.

diversity of cognitive paradigms. Our *transfer-learning* model has such flexibility, as it does not require explicit correspondence across experiments. Beyond quantitative benefits –the gain in prediction accuracy– the models also brings qualitative benefits, facilitating the interpretation of decoding maps (Fig 6). Pooling subjects across studies effectively increases the sample size, as advocated by [2]. The resulting increase in statistical power for cognitive modeling will help addressing the reproducibility challenge outlined by [3]. In our setting, each study (or *site*) provides a single decoding objective, which is predicting one contrast among all other contrasts from this study. This is a validated approach in decoding [73]. As some studies use different fMRI tasks, we may also use one decoding objective per *task*, with similar quantitative improvement in performance (see Fig F in S1B Appendix).

Our modeling choices were driven by the recent successes of deep non-linear models in computer vision and medical imaging. However, we were not able to increase performance by departing from linear models: introducing non linearities in our models provides no improvement on left-out accuracy. On the other hand, we have shown that pooling many fMRI data sources enables to learn deeper models, although these remain linear. Techniques developed in deep learning prove useful to fit models that generalize well across subjects: using dropout regularization [14] and advanced stochastic gradient techniques [13] is crucial for successful transfer and good generalization performance.

Sticking to linear models brings the benefit of easy interpretation of decoding models. The use of sparsity and non-negativity in the training and consensus phase allow to obtain interpretable networks. Using sparsity only in each phase (as originally advocated by [74]) yields "contrast" networks with both positive and negative regions, that are harder to interpret (see also [60]). In particular, this limits the occurence of non-zero weights that reflect noise suppression [75].

The models capture information relevant for many decoding tasks in their internal representations. From these internals, we extract interpretable cognitive networks, inspired by matrix factorization techniques used to interpret computer vision models [76]. The good predictive performance of MSTONs networks (Fig 2 and Fig B in S1B Appendix) provides quantitative support for their decomposition of brain function. Extracting a universal basis of cognition is beyond the scope of a single fMRI study, and should be done by analysis across many studies. We show that, across studies, a joint predictive model finds meaningful approximations of atomic cognitive functions spanning a wide variety of mental processes (Fig 4). This methodology provides a step forward towards defining mental processes in a quantitative manner, which remains a fundamental challenge in psychology [9, 77]. Yet, in the present work, the delineation of atomic cognitive functions remains coarse and incomplete. This is likely due to the limited scope of our corpus, and to the fact that we automatically align the cognitive functions probed by the various studies of the corpus. Expert annotation of mental process involved in the studies could greatly help establishing a clearer picture.

Our approach differs from commonly-used decomposition techniques in fMRI analysis (e.g. ICA [78], or dictionary learning [74]), that are used to extract *functional networks*. These techniques optimize an unsupervised reconstruction objective over resting-state data, in effect capturing co-occurrence of brain activity across distributed locations. They have traditionally been used with few components (e.g. $k \approx 20$). In contrast, after the first decomposition, performed without information from the tasks, we extract the MSTONs components to optimize the decoding performance on many tasks. Leaving a systematic comparison between MSTONs and classical functional networks for future work, we already make two observations. First, a fraction of functional networks extracted by unsupervised methods support non-Gaussian noise patterns in the BOLD time-series, and permits noise suppression [79, 80]. Typically, only a fraction of the networks extracted in an ICA analysis is interpreted. MSTONs, on the other hand,

optimize a supervised objective and focus on the fraction of the BOLD signal related to the tasks. Second, MSTONs (despite being more noisy) appears more skewed towards known coordinated brain networks (Figs 4 and 5), that differs from the networks recruited at rest (see e.g. [81] for a comparison of task and rest brain networks).

We use many different fMRI studies to distill MSTONs across various tasks. This data aggregation approach requires little supervison. The flip side is that it leads to coarse results by nature: our approach is obviously not sufficient to recover the detailed brain-to-mind mapping, collective knowledge of psychologists and neuroscientists, that has emerged from decades of research on multimodal datasets and careful behavioral experiments. Specific brain-to-mind associations are best resolved with dedicated experiments using experimental-pyschology paradigms tailored to the question at hand. Other data than fMRI, for instance more invasive, may also provide stronger evidence. For instance a double dissociation in brain-lesion patients give unambiguous evidence of distinct cognitive processes via distant neural supports, as with Broca and Wernicke's separation of language understanding and generation [82], or the more recent teasing out of emotional and cognitive empathy [83].

Finally, the current version of our framework does not model explicit inter-subject variability, and is rather focused on extracting commonalities across subjects. Future work may augment multi-study decoding with such information, as obtained by e.g., hyperalignment techniques [84].

Conclusion

The success of using distributed representations to bridge cognitive tasks supports a system-level view on how brain activity supports cognition.

Our multi-study model will become increasingly useful to brain imaging as the number of available studies grows. Such a growth is driven by the steady increase of publicly shared brain-imaging data, facilitated by online neuroimaging platforms and increased standardization [2, 85]. With a larger corpus of studies, the proposed methodology has the potential to build even better universal priors that overall improve statistical power for functional brain imaging. As such, multi-study decoding provides a path towards knowledge consolidation in functional neuroimaging and cognitive neuroscience.

References

1. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. Large-Scale Automated Synthesis of Human Functional Neuroimaging Data. Nature Methods. 2011;8(8):665–670.

2. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, et al. Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research. Nature Reviews Neuroscience. 2017;18(2):115–126.

3. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. Nature Reviews Neuroscience. 2013;14(5):365–376.

4. Van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, et al. The Human Connectome Project: A Data Acquisition Perspective. NeuroImage. 2012;62(4):2222–2231.

5. Miller KL, Alfaro-Almagro F, Bangerter NK, Thomas DL, Yacoub E, Xu J, et al. Multimodal Population Brain Imaging in the UK Biobank Prospective Epidemiological Study. Nature Neuroscience. 2016;19(11):1523–1536.

6. Poldrack RA. Can cognitive processes be inferred from neuroimaging data? Trends in cognitive sciences. 2006;10(2):59–63.

7. Haxby JV, Gobbini IM, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science. 2001;293(5539):2425–2430.

8. Poldrack RA, Halchenko YO, Hanson SJ. Decoding the Large-Scale Structure of Brain Function by Classifying Mental States Across Individuals. Psychological Science. 2009;20(11):1364–1372.

9. Poldrack RA, Barch DM, Mitchell J, Wager TD, Wagner AD, Devlin JT, et al. Toward Open Sharing of Task-Based fMRI Data: The OpenfMRI Project. Frontiers in Neuroinformatics. 2013;7:12.

10. Gorgolewski KJ, Varoquaux G, Rivera G, Schwarz Y, Ghosh SS, Maumet C, et al. NeuroVault.org: A Web-Based Repository for Collecting and Sharing Unthresholded Statistical Maps of the Human Brain. Frontiers in Neuroinformatics. 2015;9:8.

11. Ando RK, Zhang T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. Journal of Machine Learning Research. 2005;6:1817–1853.

12. Xue Y, Liao X, Carin L, Krishnapuram B. Multi-Task Learning for Classification with Dirichlet Process Priors. Journal of Machine Learning Research. 2007;8(Jan):35–63.

13. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: International Conference for Learning Representations; 2015.

14. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 2014;15(1):1929–1958.

15. Varoquaux G, Poldrack RA. Predictive models avoid excessive reductionism in cognitive neuroimaging. Current opinion in neurobiology. 2019;55:1–6.

16. Newell A. You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. Visual Information Processing. 1973; p. 1–26.

17. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-Based Neurologic Signature of Physical Pain. New England Journal of Medicine. 2013;368(15):1388–1397.

18. Varoquaux G, Schwartz Y, Poldrack RA, Gauthier B, Bzdok D, Poline JB, et al. Atlases of cognition with large-scale human brain mapping. PLoS computational biology. 2018;14(11):e1006565.

19. Poldrack RA, Yarkoni T. From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. Annual Review of Psychology. 2016;67(1):587–612.

20. Barrett LF. The Future of Psychology: Connecting Mind to Brain. Perspectives on Psychological Science. 2009;4(4):326–339.

21. Mensch A, Mairal J, Bzdok D, Thirion B, Varoquaux G. Learning Neural Representations of Human Cognition Across Many fMRI Studies. In: Advances in Neural Information Processing Systems; 2017. p. 5883–5893.

22. Amalric M, Dehaene S. Origins of the Brain Networks for Advanced Mathematics in Expert Mathematicians. Proceedings of the National Academy of Sciences. 2016;113(18):4909–4917.

23. Pinel P, Thirion B, Meriaux S, Jobert A, Serres J, Bihan DL, et al. Fast Reproducible Identification and Large-Scale Databasing of Individual Functional Cognitive Networks. BMC neuroscience. 2007;8:91.

24. Papadopoulos Orfanos D, Michel V, Schwartz Y, Pinel P, Moreno A, Le Bihan D, et al. The Brainomics/Localizer Database. NeuroImage. 2017;144:309–314.

25. Shafto MA, Tyler LK, Dixon M, Taylor JR, Rowe JB, Cusack R, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) Study Protocol: A Cross-Sectional, Lifespan, Multidisciplinary Examination of Healthy Cognitive Ageing. BMC Neurology. 2014;14:204.

26. Cauvet E. Traitement des structures syntaxiques dans le langage et dans la musique [PhD thesis]. Paris 6; 2012.

27. Hara N, Cauvet E, Devauchelle AD, Dehaene S, Pallier C, et al. Neural Correlates of Constituent Structure in Language and Music. NeuroImage. 2009;47:S143.

28. Devauchelle AD, Oppenheim C, Rizzi L, Dehaene S, Pallier C. Sentence Syntax and Content in the Human Temporal Lobe: An fMRI Adaptation Study in Auditory and Visual Modalities. Journal of Cognitive Neuroscience. 2009;21(5):1000–1012.

29. Schonberg T, Fox C, Mumford JA, Congdon C, Trepel C, Poldrack RA. Decreasing Ventromedial Prefrontal Cortex Activity During Sequential Risk-Taking: An fMRI Investigation of the Balloon Analog Risk Task. Frontiers in Neuroscience. 2012;6:80.

30. Aron AR, Gluck M, Poldrack RA. Long-Term Test–Retest Reliability of Functional MRI in a Classification Learning Task. NeuroImage. 2006;29:1000–1006.

31. Xue G, Poldrack RA. The Neural Substrates of Visual Perceptual Learning of Words: Implications for the Visual Word Form Area Hypothesis. Journal of Cognitive Neuroscience. 2007;19:1643–1655.

32. Tom SM, Fox CR, Trepel C, Poldrack RA. The Neural Basis of Loss Aversion in Decision-Making Under Risk. Science. 2007;315(5811):515–518.

33. Jimura K, Cazalis F, Stover ERS, Poldrack RA. The Neural Basis of Task Switching Changes with Skill Acquisition. Frontiers in Human Neuroscience. 2014;8.

34. Xue G, Aron AR, Poldrack RA. Common Neural Substrates for Inhibition of Spoken and Manual Responses. Cerebral Cortex. 2008;18:1923–1932.

35. Aron AR, Behrens TE, Smith S, Frank MJ, Poldrack RA. Triangulating a Cognitive Control Network Using Diffusion-Weighted Magnetic Resonance Imaging (MRI) and Functional MRI. The Journal of Neuroscience. 2007;27:3743–3752.

36. Cohen JR. The Development and Generality of Self-Control [PhD thesis]. University of the City of Los Angeles; 2009.

37. Foerde K, Knowlton B, Poldrack RA. Modulation of Competing Memory Systems by Distraction. Proceedings of the National Academy of Science. 2006;103:11778–11783.

38. Rizk-Jackson A, Aron AR, Poldrack RA. Classification Learning and Stop-Signal (one Year Test-Retest); 2011. https://openfmri.org/dataset/ds000017.

39. Alvarez RP, Jasdzewski G, Poldrack RA. Building Memories in Two Languages: An fMRI Study of Episodic Encoding in Bilinguals. In: Society for Neuroscience Abstracts; 2002. p. 179.12.

40. Poldrack RA, Clark J, Pare-Blagoev E, Shohamy D, Creso Moyano J, Myers C, et al. Interactive Memory Systems in the Human Brain. Nature. 2001;414(6863):546–550.

41. Kelly A, Milham M. Simon Task; 2011. https://openfmri.org/dataset/ds000101.

42. Duncan K, Pattamadilok C, Knierim I, Devlin J. Consistency and Variability in Functional Localisers. NeuroImage. 2009;46:1018–1026.

43. Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN. Prefrontal-Subcortical Pathways Mediating Successful Emotion Regulation. Neuron. 2008;59:1037–1050.

44. Moran JM, Jolly E, Mitchell JP. Social-Cognitive Deficits in Normal Aging. The Journal of Neuroscience. 2012;32:5553–5561.

45. Uncapher MR, Hutchinson JB, Wagner AD. Dissociable Effects of Top-Down and Bottom-Up Attention During Episodic Encoding. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2011;31(35):12613–12628.

46. Gorgolewski KJ, Storkey A, Bastin ME, Whittle IR, Wardlaw JM, Pernet CR. A Test-Retest fMRI Dataset for Motor, Language and Spatial Attention Functions. GigaScience. 2013;2(1):6.

47. Collier AK, Wolf DH, Valdez JN, Turetsky BI, Elliott MA, Gur RE, et al. Comparison of Auditory and Visual Oddball fMRI in Schizophrenia. Schizophrenia research. 2014;158:183–188.

48. Gauthier B, Eger E, Hesselmann G, Giraud AL, Kleinschmidt A. Temporal Tuning Properties Along the Human Ventral Visual Stream. The Journal of Neuroscience. 2012;32:14433–14441.

49. Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, et al. Function in the Human Connectome: Task-fMRI and Individual Differences in Behavior. NeuroImage. 2013;80:169–189.

50. Henson RN, Wakeman DG, Litvak V, Friston KJ. A Parametric Empirical Bayesian Framework for the EEG/MEG Inverse Problem: Generative Models for Multi-Subject and Multi-Modal Integration. Frontiers in Human Neuroscience. 2011;5.

51. Knops A, Thirion B, Hubbard EM, Michel V, Dehaene S. Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. Science. 2009;324:1583–1585.

52. Poldrack RA, Congdon E, Triplett W, Gorgolewski KJ, Karlsgodt K, Mumford JA, et al. A Phenome-Wide Examination of Neural and Cognitive Function. Scientific Data. 2016;3:160110.

53. Pinel P, Dehaene S. Genetic and Environmental Contributions to Brain Activation During Calculation. NeuroImage. 2013;81:306–316.

54. Vagharchakian L, Dehaene-Lambertz G, Pallier C, Dehaene S. A Temporal Bottleneck in the Language Comprehension Network. The Journal of Neuroscience. 2012;32:9089–9102.

55. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010;22(10):1345–1359.

56. Kingma DP, Salimans T, Welling M. Variational Dropout and the Local Reparameterization Trick. In: Advances in Neural Information Processing Systems; 2015. p. 2575–2583.

57. Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RS. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Human brain mapping. 1994;2(4):189–210.

58. Mairal J, Bach F, Ponce J, Sapiro G. Online Learning for Matrix Factorization and Sparse Coding. Journal of Machine Learning Research. 2010;11:19–60.

59. Mensch A, Mairal J, Thirion B, Varoquaux G. Stochastic Subsampling for Factorizing Huge Matrices. IEEE Transactions on Signal Processing. 2018;66(1):113–128.

60. Dadi K, Varoquaux G, Machlouzarides-Shalit A, Gorgolewski KJ, Wassermann D, Thirion B, et al. Fine-grain atlases of functional modes for fMRI analysis. To appear in NeuroImage. 2020;.

61. Gower JC, Ross GJ. Minimum Spanning Trees and Single Linkage Cluster Analysis. Journal of the Royal Statistical Society: Series C (Applied Statistics);18(1):54–64.

62. Sokal RR, Rohlf FJ. The Comparison of Dendrograms by Objective Methods. Taxon; p. 33–40.

63. Braver TS, Barch DM, Gray JR, Molfese DL, Snyder A. Anterior cingulate cortex and response conflict: effects of frequency, inhibition and errors. Cerebral cortex. 2001;11:825.

64. Stevens FL, Hurley RA, Taber KH. Anterior cingulate cortex: unique role in cognition and emotion. The Journal of neuropsychiatry and clinical neurosciences. 2011;23(2):121.

65. Bush G, Vogt BA, Holmes J, Dale AM, Greve D, Jenike MA, et al. Dorsal anterior cingulate cortex: a role in reward-based decision making. Proceedings of the National Academy of Sciences. 2002;99:523–528.

66. Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. A default mode of brain function. Proceedings of the National Academy of Sciences. 2001;98(2):676–682.

67. Spreng RN, Grady CL. Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network. Journal of cognitive neuroscience. 2010;22:1112.

68. Gratton C, Sun H, Petersen SE. Control networks and hubs. Psychophysiology. 2018;55:e13032.

69. Ptak R. The frontoparietal attention network of the human brain: action, saliency, and a priority map of the environment. The Neuroscientist. 2012;18(5):502–515.

70. Kiviniemi V, Starck T, Remes J, Long X, Nikkinen J, Haapea M, et al. Functional segmentation of the brain cortex using high model order group PICA. Human brain mapping. 2009;30(12):3865–3886.

71. Leech R, Kamourieh S, Beckmann CF, Sharp DJ. Fractionating the default mode network: distinct contributions of the ventral and dorsal posterior cingulate cortex to cognitive control. Journal of Neuroscience. 2011;31(9):3217–3224.

72. Varoquaux G. Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars. NeuroImage. 2018;180:68–77.

73. Bzdok D, Eickenberg M, Grisel O, Thirion B, Varoquaux G. Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. In: Advances in Neural Information Processing Systems; 2015. p. 3348–3356.

74. Varoquaux G, Gramfort A, Pedregosa F, Michel V, Thirion B. Multi-Subject Dictionary Learning to Segment an Atlas of Brain Spontaneous Activity. Proceedings of the International Conference on Information Processing in Medical Imaging. 2011;22:562.

75. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, et al. On the Interpretation of Weight Vectors of Linear Models in Multivariate Neuroimaging. NeuroImage;87:96–110.

76. Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, et al. The Building Blocks of Interpretability. Distill. 2018;3(3):e10.

77. Uttal WR. The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain. The MIT press; 2001.

78. McKeown MJ, Makeig S, Brown GG, Jung TP, Kindermann SS, Bell AJ, et al. Analysis of fMRI Data by Blind Separation Into Independent Spatial Components. Human Brain Mapping. 1998;6(3):160–188.

79. Perlbarg V, Bellec P, Anton JL, Pélégrini-Issac M, Doyon J, Benali H. CORSICA: correction of structured noise in fMRI by automatic identification of ICA components. Magnetic resonance imaging. 2007;25(1):35–46.

80. Salimi-Khorshidi G, Douaud G, Beckmann CF, Glasser MF, Griffanti L, Smith SM. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. Neuroimage. 2014;90:449–468.

81. Laird AR, Fox PM, Eickhoff SB, Turner JA, Ray KL, McKay DR, et al. Behavioral interpretations of intrinsic connectivity networks. Journal of cognitive neuroscience. 2011;23(12):4022–4037.

82. Friederici AD, Hahne A, Von Cramon DY. First-pass versus second-pass parsing processes in a Wernicke's and a Broca's aphasic: electrophysiological evidence for a double dissociation. Brain and language. 1998;62:311.

83. Shamay-Tsoory SG, Aharon-Peretz J, Perry D. Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. Brain. 2009;132:617.

84. Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, et al. A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. Neuron. 2011;72(2):404–416.

85. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments. Scientific Data. 2016;3:sdata201644.

86. Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, et al. Correspondence of the Brain's Functional Architecture During Activation and Rest. Proceedings of the National Academy of Sciences. 2009;106(31):13040–13045.

87. Yeo TBT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, et al. The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity. Journal of Neurophysiology. 2011;106(3):1125–1165.

88. Nocedal J. Updating Quasi-Newton Matrices with Limited Storage. Mathematics of Computation. 1980;35(151):773–782.

89. Neyshabur B. Implicit Regularization in Deep Learning [PhD thesis]. Toyota Technological Institute at Chicago; 2017.

90. Molchanov D, Ashukha A, Vetrov D. Variational Dropout Sparsifies Deep Neural Networks. In: Proceedings of the International Conference on Machine Learning; 2017. p. 2498–2507.

91. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: Proceedings of the International Conference on Machine Learning; 2015. p. 448–456.

92. Srebro N, Rennie J, Jaakkola TS. Maximum-Margin Matrix Factorization. In: Advances in Neural Information Processing Systems; 2004. p. 1329–1336.

93. Beck A, Teboulle M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. SIAM Journal on Imaging Sciences. 2009;2(1):183–202.

94. Rennie JDM, Srebro N. Fast Maximum Margin Matrix Factorization for Collaborative Prediction. In: Proceedings of the International Conference on Machine Learning; 2005. p. 713–719.

95. Bell RM, Koren Y. Lessons from the Netflix Prize Challenge. ACM SIGKDD Explorations Newsletter. 2007;9(2):75–79.

96. Wager S, Wang S, Liang PS. Dropout Training as Adaptive Regularization. In: Advances in Neural Information Processing Systems; 2013. p. 351–359.

97. Breiman L. Bagging Predictors. Machine Learning. 1996;24(2):123–140.

98. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine Learning for Neuroimaging with Scikit-Learn. Frontiers in Neuroinformatics. 2014;8:14.

99. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.

100. Paszke A, Gross S, Chintala S, Chanan G. PyTorch: Tensors and Dynamic Neural Networks in Python with Strong GPU Acceleration; 2017.

Supporting information

S1A Appendix. Detailed methods. This appendix discusses technical details of the multi-study decoding approach: the specific architecture, a 3-layer linear model, and the deep-learning technique used to regularize and train it.

S1B Appendix. Discussion on the model design. In this appendix, we perform supportive experiments to explain the observed results, An ablation study of the various model components is provided to further support modelling choices.

S1C Appendix. Reproduction details and tables. In this appendix, we provide implementation details for reproducibility, along with tables with quantitative results per contrast.

S1 Appendix

The appendix is structured as follow: in the first section, we formalize the learning setting and method, after describing decoding baselines. In the second section, we perform supportive experiments to explain the observed results, and discuss various alternatives for the model, to further support modelling choices. Finally, we provide reproduction details, along with data and software notes. A visualization of all MSTONs components (that reproduces https://cogspaces.github.io/assets/MSTON/components.html) is provided for completeness in S1 Components.

Notations. We denote scalars, vectors and matrices using lower-case, bold lower-case and bold upper-case letters, e.g., x, $\boldsymbol{x}$ and $\boldsymbol{X}$. We denote the elements of $\boldsymbol{X}$ by $x_{i,j}$ and its rows by $\boldsymbol{x}_i$. We write x^j a value that is specific to study number j. We denote $\bar{x}$ a value built from an ensemble of value $(x_s)_s$. Finally, we write $[l]$ the set of integers ranging from 1 to l.

A Methods

We describe in mathematical terms the multi-layer decoder at the center of our method and provide supporting experiments. We start by formalizing the joint objective loss and the model training process.

A.1 Inter-subject decoding setting

We consider N task functional MRI studies (detailed in Table 1), on which we perform inter-subject decoding. In study number j, n^j subjects are made to perform one (or sometimes several) tasks. Acquired BOLD time-series are registered to a common template using non-linear spatial registration, after motion and slice-timing corrections. BOLD time-series are then fed to a standard analysis pipeline, which fits a linear model relating the design matrix of each experiment to the signal in every voxel. We use the *nistats* library for this purpose. From the obtained beta maps, we compute z-statistics maps, either associated with each of the base conditions (stimulus or task) of the experiments, or with contrasts defined by the study's authors. In both cases, z-maps are labeled with a number $1 \leq y \leq c^j$ that corresponds to k-th contrast/base condition (called contrast in the following). Overall, this produces a set of z-maps $(\boldsymbol{x}_i^j)_{i \in [c^j n^j]}$ living in $\mathbb{R}^p$, where p is the number of voxels, associated with a sequence of contrast $(k_i^j)_{i \in [c^j n^j]}$. The transformation from 3D brain images to 1D vectors is done using a grey-matter mask after alignment with the MNI template. We compare using a grey-matter mask with using a full brain mask in Section B.4.1. Inter-subject decoding proposes a model $f_\theta^j : \mathbb{R}^p \to [1, c^j]$ that predicts contrast identity from z-maps, i.e., $\hat{y}_i^j \triangleq f_\theta^j(\boldsymbol{x}_i^j)$, where θ is learned from training data, and the performance of the model is assessed on left-out subjects.

A.2 Baseline voxel-space decoder

Baseline decoders are linear classifier models defined separately for each study j, which take full brain images as input. For every input map $\boldsymbol{x}_i$ in $\mathbb{R}^p$, we compute the logits $\boldsymbol{l}_i$ in $\mathbb{R}^c$ as

$$\boldsymbol{l}_i(\boldsymbol{W}, \boldsymbol{b}) \triangleq \boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{b},$$

where $\boldsymbol{W} \in \mathbb{R}^{c \times p}$ and $\boldsymbol{b} \in \mathbb{R}^c$ are the parameters of the linear model to be learned for study j—we drop the superscript j in this paragraph and the next for simplicity. Logits

are transformed into a classification probability vector using the softmax operator. At test time, we predict the label corresponding to the maximal logit, i.e., $\hat{y}_i = \text{argmax}_{1\leq y\leq c}\, l_{i,y}$. The model is trained on the data $(\boldsymbol{x}_i, y_i)_{i\in[n]}$ by minimizing the ℓ_2^2 regularized multinomial classification problem

$$\min_{\substack{\boldsymbol{W}\in\mathbb{R}^{c\times p}\\ \boldsymbol{b}\in\mathbb{R}^c}} -\frac{1}{n}\sum_{i=1}^{n}\Big(l_{i,y_i}(\boldsymbol{W},\boldsymbol{b}) + \log(\sum_{k=1}^{c}\exp l_{i,k}(\boldsymbol{W},\boldsymbol{b}))\Big) + \lambda\|\boldsymbol{W}\|_F^2, \tag{1}$$

where $\|\cdot\|_F^2$ is the Frobenius norm, that computes $\|\boldsymbol{W}\|_F^2 \triangleq \sum_{i,j=1}^{c,p} w_{i,j}^2$.

A.3 Baseline dimension reduced decoder

A variant of the voxel-based decoders is obtained by introducing a first-layer dimension reduction learned from resting-state data. This amounts to computing

$$\boldsymbol{l}_i(\boldsymbol{V},\boldsymbol{b},\boldsymbol{D}) \triangleq \boldsymbol{V}\boldsymbol{D}\boldsymbol{x}_i + \boldsymbol{b},$$

where $\boldsymbol{V}$ in $\mathbb{R}^{c\times k}$ forms the classifying weights of the model, and the matrix $\boldsymbol{D}$ in $\mathbb{R}^{k\times p}$ is *assigned* during training to functional networks learned on resting-state data, as detailed in A.5. Multiplying input data by $\boldsymbol{D}$ projects statistical images onto meaningful resting-state components, in an attempt to improve classification performance and reduce computation cost, akin to the methods proposed by [86, 87]. The model is trained by solving the convex objective (1) separately for each study, replacing $\boldsymbol{W}$ by $\boldsymbol{V}$ in $\mathbb{R}^{c\times k}$:

$$\min_{\substack{\boldsymbol{V}\in\mathbb{R}^{c\times k}\\ \boldsymbol{b}\in\mathbb{R}^c}} -\frac{1}{n}\sum_{i=1}^{n}\Big(l_{i,y_i}(\boldsymbol{V},\boldsymbol{b},\boldsymbol{D}) + \log(\sum_{k=1}^{c}\exp l_{i,k}(\boldsymbol{V},\boldsymbol{b},\boldsymbol{D}))\Big) + \lambda\|\boldsymbol{V}\|_F^2. \tag{2}$$

Our results (Fig 2C) show that decoding from functional networks is not significantly better than decoding from voxels directly. For both baselines, the parameter λ is found by half-split cross-validation. Training is performed using a L-BFGS solver [88]. We use non standardized maps $(\boldsymbol{x}_i)_i$ as input as we observed that standardization hinders performance.

A.4 Three-layer model description

Our three-layer model adds a second shared linear layer in between the projection on functional networks and the classification models. We still have

$$\boldsymbol{l}_i^j(\boldsymbol{W}^j,\boldsymbol{b}^j) \triangleq \boldsymbol{W}^j\boldsymbol{x}_i^j + \boldsymbol{b}^j,$$

for every z-map i and study j. However, we introduce a coupling between the various parameters $(\boldsymbol{W}^j)_{j\in[N]}$ of each study: they should decompose on common basis $\boldsymbol{L}\boldsymbol{D}$, where $\boldsymbol{L}$ is estimated from the whole corpus of data, and $\boldsymbol{D}$ is the resting-state dictionary presented above. Formally, we assume that there exist a matrix $\boldsymbol{L}$ in $\mathbb{R}^{l\times k}$ with $l < k < p$, and a set of matrices $(\boldsymbol{U}^j)_{j\in[N]}$ so that for all $j \in [N]$, the classification weights of (1) writes

$$\boldsymbol{W}^j \triangleq \boldsymbol{U}^j\boldsymbol{L}\boldsymbol{D}, \quad \text{where} \quad \boldsymbol{U}^j \in \mathbb{R}^{c^j\times l}. \tag{3}$$

The matrix $\boldsymbol{D}$ corresponds to the first-layer weights pictured in Fig 1, $\boldsymbol{L}$ to the second-layer weights, and $(\boldsymbol{U}^j, \boldsymbol{b}^j)_j$ to the various classification heads of the third layer. In this work, we choose $k = 465$ and $l = 128$. While $\boldsymbol{D}$ remains fixed, the second-layer matrix $\boldsymbol{L}$ and the N classification heads $(\boldsymbol{U}^j)_{j\in[N]}$ are jointly learned during training, a necessary step toward improving decoding accuracy. The "shared-layer" parameterization (3) is a common approach in multi-task learning [11, 12], and should allow *transfer learning* between decoding tasks, under certain conditions. In our setting, both the data distribution from the different studies and the classification task associated with each study differ—this is a particular case of *inductive transfer learning*[1], described by [55].

Modeling. Without refinement nor regularization, we seek a local minimizer of the following non-convex objective function, which combines the classification objectives (1) from all studies, with parameter sharing:

$$\min_{\substack{\boldsymbol{L}\in\mathbb{R}^{l\times k}\\(\boldsymbol{U}^j,\boldsymbol{b}^j)_j}} -\sum_{j=1}^{N}\frac{(n^j)^\beta}{n^j}\sum_{i=1}^{n^j}\Big(l^j_{i,y_i}(\boldsymbol{U}^j,\boldsymbol{b}^j,\boldsymbol{L}) - \log\Big(\sum_{k=1}^{c^j}\exp l^j_{i,k}(\boldsymbol{U}^j,\boldsymbol{b}^j,\boldsymbol{L})\Big)\Big), \tag{4}$$

where the dependence on $\boldsymbol{D}$ is left implicit. The scalar β in $[0, 1]$ is a parameter that regulates the importance of each study in the joint objective, that we further discuss in B.8. We note that the importance of the study j to find the latent parameter $\boldsymbol{L}$ depends on the amplitude of the gradient $\frac{\partial \ell_j}{\partial \boldsymbol{L}}$ that does not depend on the number of tasks c^j: in particular, for each study j, contrast $1 \le k \le c^j$ and subject $1 \le i \le n^j$, the susceptibility of the loss to the logits $l^j_{i,k}$ is such that $\frac{\partial \ell_j}{\partial l^j_{i,k}} \in [-1, 1]$, independent from c^j.

Regularization. We observe that minimizing (4) leads to strong overfitting and low performance on left-out data, with performance similar to fitting (1) without regularization, separately for each study. Adding ℓ_2 regularization to the second and third layer weights gives little benefit, as we discuss in Section B.2.3. On the other hand, introducing *dropout* [14] during training alleviates the overfitting issue and fosters transfer learning. Dropout is a stochastic regularization method that prevents the weights from each layer to co-adapt by perturbating them with multiplicative noise during training. It ensures that the information is well spread across coefficients rows and columns [89]. In our case, this favors transfer learning, as it ensures that no single row of $\boldsymbol{L}$, or in plain words no task-optimized network, becomes dedicated to a *single* study. We further compare the different methods that we can use to foster transfer of information between studies in Section B.2.

We use the variational flavor of dropout [56] to make the dropout rate for every study adaptive. This slightly improves performance compared to binary dropout: every decoding task requires a different level of regularization, depending on the size of the study and the hardness of the task, and it is beneficial to estimate it from data. In details, during training, at every iteration, for every input sample i of a mini-batch from study j, we randomly draw two multiplicative noise matrices

$$\boldsymbol{M}_D = \mathrm{Diag}([m_{D,t}]_{t\in[k]}), \quad \boldsymbol{M}^j_L = \mathrm{Diag}([m_{L,t}]_{t\in[l]}),$$

[1] This case is less studied than the classical multi-task setting where input data are single-source but learning tasks are multiple.

where $m_{D,t} \sim \mathcal{N}(1, \alpha)$ and $m_{L,t} \sim \mathcal{N}(1, \alpha^j)$, with α fixed and α^j estimated from data.[2] We then compute the noisy logits

$$\boldsymbol{l}_i^j \triangleq \boldsymbol{U}^j \boldsymbol{M}_L^j \boldsymbol{L} \boldsymbol{M}_D \boldsymbol{D} \boldsymbol{x}_i^j + \boldsymbol{b}^j,$$

and use these to compute the loss (5), to which we add a regularization term that regulates the learning of α^j, introduced by [90]. We compute the gradient with respect to $\boldsymbol{L}$, $\boldsymbol{U}^j$, $\boldsymbol{b}^j$ using the local reparametrization trick [56]. We refer to [90] for more details on variational dropout and a Bayesian grounding of this approach.

Optimization. We solve the problem (4) using stochastic optimization. Namely, at each iteration, we compute an unbiased estimate of the objective (4) and its gradient with respect to the model parameters, in order to perform a stochastic gradient step. For this, we randomly choose the study j with a probability proportional to $(n^j)^\beta$, and consider a mini-batch of z-maps $(\boldsymbol{x}_i^j)_{j \in B}$ that we use to compute the unbiased objective estimate

$$-\frac{1}{B} \sum_{i=1}^{n} -\Big(l_{i,k_i}^j \log(\sum_{k=1}^{c} \exp l_{i,k}^j) \Big), \tag{5}$$

from which we compute gradients with respect to $\boldsymbol{L}$, $\boldsymbol{U}^j$ and $\boldsymbol{b}^j$.

Optimization is performed using *Adam* [13], a flavor of stochastic gradient descent that depends less on the step-size. We use batch normalization [91] between the second and third layer, as it slightly improves performance—it reduces potential negative transfer learning—and training speed.

A.5 Resting-state data

As mentioned above, we use resting-state data to compute the first-layer weights $\boldsymbol{D}$ in $\mathbb{R}^{k \times p}$, where $k = 512$. Such high-order dictionaries are known to perform well for decoding [60]. We consider data from the HCP900 release, and stack all records to obtain a data matrix $\boldsymbol{X}$ in $\mathbb{R}^{n \times p}$. We then use an online solver [59] to solve the sparse non-negative matrix factorization problem

$$\boldsymbol{A}, \boldsymbol{D} \triangleq \underset{\boldsymbol{D} \in \mathcal{C}, \boldsymbol{A} \in \mathbb{R}^{k \times n}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{A}\boldsymbol{D}\|_F^2 + \lambda \|\boldsymbol{A}\|_F^2, \tag{6}$$

where the constraint $\mathcal{C} = \left\{ \boldsymbol{D} \in \mathbb{R}^{k \times p}, \boldsymbol{D} \geq 0, \|\boldsymbol{d}_j\|_1 \leq 1 \, \forall \, j \in [k] \right\}$ enforces every dictionary component to live in the simplex of $\mathbb{R}^p$, ensuring sparsity and non-negativity of the functional networks. The sparsity level is chosen so that the functional networks $\boldsymbol{D}$ cover the whole brain with as little overlap as possible. Larger overlap leads to more correlated activations input to the second layer, yielding a harder learning problem. With lower coverage, we would miss important information to decode some of the predicted psychological conditions. We refer to [60] for further discussion on selecting sparsity level when using dictionary learning in fMRI analysis.

Second-layer initialization. To initialize the weights of the second layer, we learn a smaller dictionary $\boldsymbol{D}_l$ in $\mathbb{R}^{l \times p}$ as in (6), where $l = 128$. We then compute the initial weights $\boldsymbol{L}_l$ so that $\boldsymbol{D}_l \approx \boldsymbol{L}_l \boldsymbol{D}$ using least-square regression. This way, applying the first two layers initially amount to projecting data onto $l = 128$ larger functional networks $\boldsymbol{D}_l$, which is a reasonable prior for reducing the dimension of brain statistical maps. Using this resting-state based initialization slightly improves performance, as we discuss in Section B.4.

[2] This *Gaussian* dropout has a similar behavior to the more commonly used binary dropout with parameter $p = \frac{\alpha}{\alpha+1}$.

Grey matter restriction. To help interpreting the obtained model, we found it helpful to remove from $\boldsymbol{D}$ the fraction (9%) of the functional networks components located in the white matter and the cerebrospinal fluid areas, turning $k = 512$ into $k = 465$. We discuss the effect of this restriction in Section B.4.1.

A.6 Model introspection with ensembling

Given any invertible matrix $\boldsymbol{M}$ in $\mathbb{R}^{l\times l}$, the non regularized version of the objective (4) is left invariant when transforming $\boldsymbol{L}$ into $\boldsymbol{M}\boldsymbol{L}$ and each $\boldsymbol{U}^j$ into $\boldsymbol{U}^j\boldsymbol{M}^{-1}$. This prevents us from interpreting the coefficients of $\boldsymbol{L}$ at the end of the training procedure, and to retrieve relevant networks by reading the weights of the second weight. The only aspect of $\boldsymbol{L}$ that remains unchanged after a linear parameter transformation is its span. Dropout regularization, which favors the canonical directions in matrix space [14], should break this symmetry, but does not help to uncover meaningful directions in the span of $\boldsymbol{L}$ in practice.

On the other hand, we found that this span was remarkably stable across runs on the same data, whether when varying initialization or simply the order in which data are streamed during stochastic gradient descent. More precisely, we trained our model 100 times with different seeds, and concatenated the weights $(\boldsymbol{L}_r)_r$ of the second-layer into a big matrix $\bar{\boldsymbol{L}}$. We performed a SVD on this matrix, and observed that the first $l = 128$ components captured 98% of the variance of $\bar{\boldsymbol{L}}$ when using the same initialization but different streaming order, and 96% when also using a different random initialization. Despite the many local minima that objective (4) admits, the span of $\boldsymbol{L}$ thus remains close to some reference span that we can extract with a matrix factorization method.

The above remark suggested the following ensemble method. We run the learning algorithm $R = 100$ times, and store the weights $(\boldsymbol{L}_r)_r$ of the second layer for each run, along with the average matrices and biases

$$\bar{\boldsymbol{W}}^j = \frac{1}{R}\sum_{r=1}^{R} \boldsymbol{U}_r^j \boldsymbol{L}_r \qquad \bar{\boldsymbol{b}}^j = \frac{1}{R}\sum_{r=1}^{R} \boldsymbol{b}_r^j, \quad \forall\, j \in [N],$$

that combine the second and third-layer weights and biases for each study j and run N, and average them across runs. We then stack the second-layer weights $(\boldsymbol{L}_r)_r$ into a tall matrix $\tilde{\boldsymbol{L}} \in \mathbb{R}^{lR\times k}$ on which we perform sparse non-negative matrix factorization. Namely, we compute $\bar{\boldsymbol{L}} \in \mathbb{R}^{l\times k}$, the new weight matrix for the second layer, solving

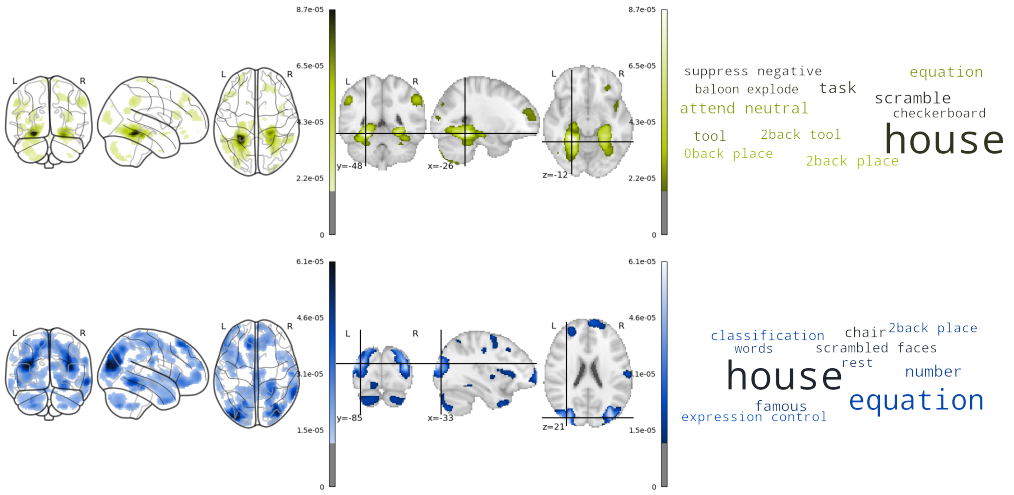
$$\bar{\boldsymbol{L}} \triangleq \operatorname*{argmin}_{\boldsymbol{L}\in\mathcal{C}} \min_{\boldsymbol{K}\in\mathbb{R}^{lR\times l}} \frac{1}{2}\|\tilde{\boldsymbol{L}} - \boldsymbol{K}\boldsymbol{L}\|_F^2 + \lambda\|\boldsymbol{K}\|_F^2,$$

where $\mathcal{C} = \left\{\boldsymbol{L} \in \mathbb{R}^{l\times k}, \boldsymbol{L} \geq 0, \|\boldsymbol{l}_j\|_1 \leq 1\, \forall\, j \in [l]\right\}$ and λ regulates the sparsity of $\bar{\boldsymbol{L}}$—performance little depends on λ provided it leads to finding $\bar{\boldsymbol{L}}\boldsymbol{D}$ with more than 50% non-zero voxels (see Section C.1). Higher λ leads to sparser maps with lower performance as brain coverage is reduced, while lower λ gives good performances but lower interpretability of the extracted networks. Finally, we compute new weights $\bar{\boldsymbol{U}}^j$ for all the classification heads of the third layer, so that $\bar{\boldsymbol{W}}^j \approx \bar{\boldsymbol{U}}^j\bar{\boldsymbol{L}}$, from a least-square point of view, for each study j. The new model is then formed of parameters $\boldsymbol{D}, \bar{\boldsymbol{L}}, (\bar{\boldsymbol{U}}^j, \bar{\boldsymbol{b}}^j)_{j\in[N]}$. In plain words, we obtain sparse non-negative second-layer weights $\bar{\boldsymbol{L}}$, and define from these weights a new model that is as close as possible to the ensemble of all learned models $\left\{\boldsymbol{D}, \boldsymbol{L}_r, (\boldsymbol{U}_r^j, \boldsymbol{b}_r^j)_j\right\}_{r\in[R]}$.

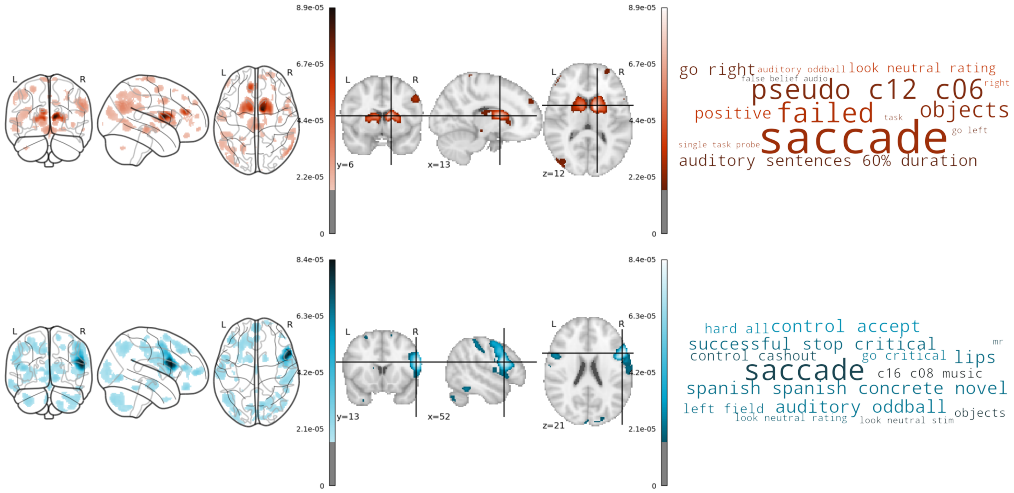
The rows of $\bar{\boldsymbol{L}}$ are now interpretable separately, as the non-negative and sparse constraints have broken the inherent parameter invariance of the original model. The rows of $\bar{\boldsymbol{L}}$ hold the coefficients for combining resting-state networks held in $\boldsymbol{D}$ into l multi-study task-optimized networks $\bar{\boldsymbol{L}}\boldsymbol{D}$ in $\mathbb{R}^{l\times p}$. We initialize the sparse NMF

algorithm with the weights $\boldsymbol{L}_l$ computed in Section A.5, to inject a small prior regarding final MSTON distribution: before running NMF, those are set to $\boldsymbol{L}_l\boldsymbol{D} \approx \boldsymbol{D}_l$, i.e., are close to large resting-state functional networks.

We observed that directly enforcing negativity/sparsity over $\boldsymbol{L}$ during the training of the model led to a strong loss in accuracy. Finding a consensus model through a post-hoc ensembling transformation thus proves to be the right solution for obtaining both performance improvement *and* interpretability.



A. MSTONs recruited by the "house" base condition



A. MSTONs recruited by the "saccade" base condition

Fig. A. Examples of MSTONs that are activated in many different tasks.

B Discussion on the model design

In this section, we discuss various choices made for designing our model and training procedures. To this end, we perform diverse quantitative and qualitative comparisons of model variants.

B.1 Understanding the role of task-optimized networks

We first provide new examples of MSTONs to enlight their properties. Then, we propose several measurements and experiments that allow to better understand how the dimension reduction performed by projecting on multi-study task-optimized networks brings quantitative improvements in decoding.

B.1.1 Other examples of MSTONs

Fig 4 shows a selection of MSTONs that are well associated with relevant clusters of base psychological conditions. Other MSTONs are of interest to discuss the multi-study decoding approach, as we now discuss.

Some base conditions recruit many MSTONs. We observe that some base psychological conditions are strongly correlated with many different MSTONs, as exemplified in Fig A. The "saccade" condition [24] triggers a very distributed response of the brain, which is the reason why it appears often in the word-clouds. The base condition "house" is in particular part of the HCP Working Memory [4] task. Decoding it versus the other HCP conditions gives a classification map for which much of the lateral visual cortex is positively activated, hence the appearance of the word "house" in the MSTONs that includes a fraction of these regions.

B.1.2 Performance of MSTONs on new studies

We argue that using the joint objective (4) improves decoding performance because the data from every study influence the model weights in both the second layer *and* all components of the third layer. This can be measured as follows. We compare the performance of learning task-optimized networks on all studies but a target one, before using the second layer as a fixed dimension reduction for fitting a decoder from the target (unobserved) study. Using this technique, information transfer from the corpus to the new study can only be imputed to the fact that the second layer has captured a dimension reduction for brain images that is efficient for decoding in general. In other words, the task optimized networks learned on N - 1 studies form a universal prior of cognition that generalizes to new paradigms.

We observe in Fig B that decoding cognitive processes from externally learned MSTON indeed performs better than decoding from voxels (3.7% mean accuracy gain, 67% experiments with net increase[3]). On the other hand, leveraging a low-dimensional representation of brain images using all studies, including the target one, during training (1.9% mean accuracy gain, 75% experiments with net increase) performs even better. This can only be explained by the fact that joint objective also fosters transfer between the classification heads of the third layer during training.

B.1.3 Effect of brain-map dimension reduction

In a dual perspective, we study the effect of reducing the dimension of the input data with the first two linear layers. We set $\boldsymbol{M} = \bar{\boldsymbol{L}}\boldsymbol{D}$ in $\mathbb{R}^{l \times p}$ to hold the task-optimized networks on each row, and compute, for all input statistical map $\boldsymbol{x}$ in $\mathbb{R}^p$, the projection of $\boldsymbol{x}$ onto span($\boldsymbol{M}$), namely

$$\boldsymbol{x}_{\text{proj}} = \boldsymbol{M}^{\boldsymbol{T}}(\boldsymbol{M}\boldsymbol{M}^{\boldsymbol{T}})^{-1}\boldsymbol{M}\boldsymbol{x} \in \mathbb{R}^p.$$

$\boldsymbol{x}_{\text{proj}}$ is thus a denoised, low-dimensional representation of the brain map $\boldsymbol{x}$, held in the span of the l multi-study task-optimized networks contained in matrix $\boldsymbol{M}$. We compare different maps $\boldsymbol{x}$ to their projection $\boldsymbol{x}_{\text{proj}}$ in Fig C.

B.2 Fostering transfer learning

We now discuss the various way in which we can foster information sharing across studies in training our multi-layer model.

[3] Due the fact that half-split folds are overlapping and performance betweens studies are interacting, model comparison experiments are not independent. This suggests to report the amount of advantageous model comparisons instead of classical null hypothesis testing, that assumes independence of trials.

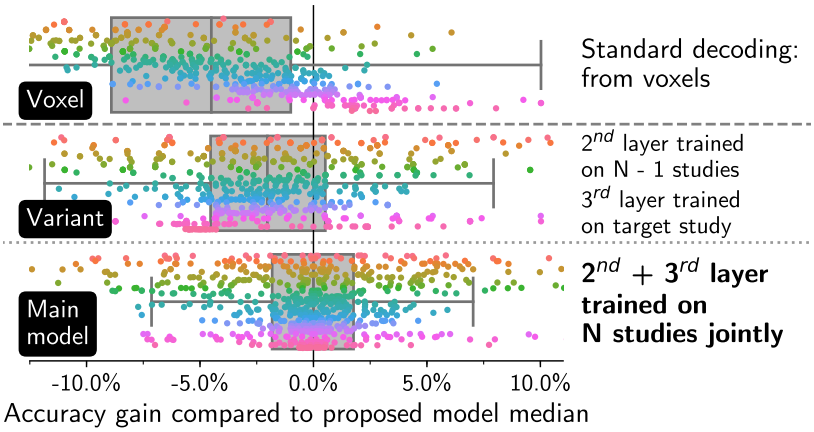


Fig. B. Quantitative improvement linked to training the model on the joint objective (4), versus improvement linked to transfer in the second-layer only. Box plots calculated over 20 random data half-split and all studies.

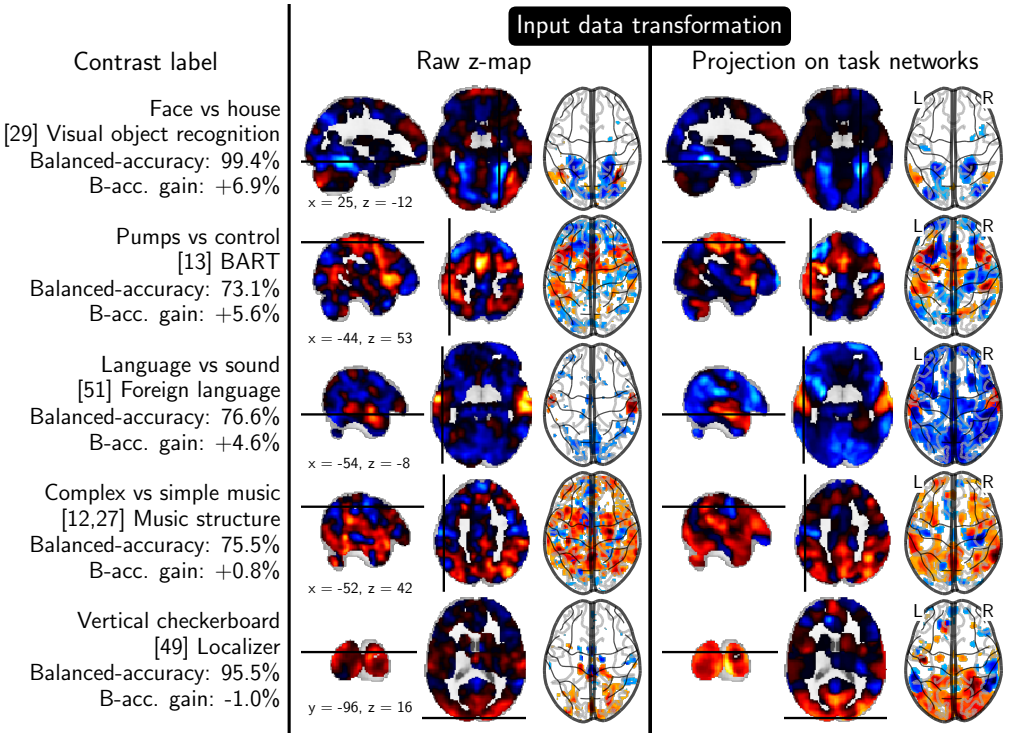


Fig. C. Effect of projecting z-maps onto MSTONs. In a dual perspective to Fig 6, input data are simplified by the projection onto task-optimized networks, and become easier to classify.

B.2.1 The need for objective coupling

Without modification nor constraint on the second layer output size l, we cannot expect to observe any transfer learning by solving the joint objective (4). Indeed, in the general case where we allow $l \geq c \triangleq \sum_{j=1}^{N} c^j$, we let $(\tilde{\boldsymbol{V}}^j, \boldsymbol{b}^j)_j$ be the unique solutions of the N non-regularized convex problems (2). We let $\tilde{\boldsymbol{V}} \in \mathbb{R}^{c \times k}$ be the vertical concatenation of $(\boldsymbol{V}^j)_j$. We then form the matrices

$$\boldsymbol{L} = \begin{bmatrix} \tilde{\boldsymbol{V}} \\ \mathbf{o} \in \mathbb{R}^{l-c\times k} \end{bmatrix} \in \mathbb{R}^{l\times k} \qquad \text{and} \tag{8}$$
$$\begin{bmatrix} \boldsymbol{U}^1 \\ \vdots \\ \boldsymbol{U}^N \end{bmatrix} \triangleq \left[\boldsymbol{I}_c \in \mathbb{R}^{c\times c}, \mathbf{o} \in \mathbb{R}^{l-c\times l}\right],$$

where $\boldsymbol{I}_c$ is the identity matrix of $\mathbb{R}^{c\times c}$. $\boldsymbol{L}$ is thus split into row-blocks $(\tilde{\boldsymbol{V}}^j)_j$, dedicated to and learned on *single studies*. It follows from elementary considerations that the matrices $(\boldsymbol{L}, (\boldsymbol{U}^j, \boldsymbol{b}^j)_j)$ form a global minimizer of (4), that is formed from the solutions of the *separated* problems (2). It is therefore possible to find solutions of (4) for which no transfer occurs. Two possible modifications of the objective (4) allow to enforce transfer: Dropout regularization and low-rank constraints, that we present and compare.

B.2.2 Dropout as a transfer incentive

First, as presented in Section A, we can use dropout between the second layer weight L and the third layer head weights $\boldsymbol{U}^j$. Dropout prevents constructions of block-separated solution of objective (4) similar to the one proposed in (8). Indeed, every reduced sample $\boldsymbol{L}\boldsymbol{D}\boldsymbol{x}_i^j$ fed to the third layer classification head j can see any of his features corrupted by multiplicative noise $\boldsymbol{M}_L$ during training. This pushes the model to capture information relevant for all studies in every activation of the second layer. In other word, the projection performed on any task-optimized network $\boldsymbol{l}_h\boldsymbol{D}$, for $h \in [l]$ should be relevant for decoding every study. This fosters transfer learning as $\boldsymbol{L}$ carries multi-study aggregated information at the end of training, unlike in (8).

B.2.3 Transfer through low-rank constraints/penalty

A second approach to transfer is to force the matrices

$$\boldsymbol{V} \triangleq \begin{bmatrix} \boldsymbol{V}^1 \\ \vdots \\ \boldsymbol{V}^N \end{bmatrix} \triangleq \begin{bmatrix} \boldsymbol{U}^1 \\ \vdots \\ \boldsymbol{U}^N \end{bmatrix} \boldsymbol{L},$$

formed of the parameters of the joint objective (4) to be *low-rank*. In this case, the subspace of $\mathbb{R}^{c\times k}$ in which $\boldsymbol{V}$ evolves is strictly smaller than $\mathbb{R}^{c\times k}$, and we cannot always find a global minimum of the joint objective (4) formed with the solutions $\tilde{\boldsymbol{V}}$ of the separate objectives (2), as we did in the construction (8). As a consequence, the data from studies truly influence the solutions $(\boldsymbol{L}, (\boldsymbol{U}^j, \boldsymbol{b}_j)_j)$ of (4), and transfer is theoretically possible.

The low-rank property may be enforced in two ways. First, we may set it as a hard constraint, setting $l < c$ in the joint objective (4). This is in practice what we do when selecting $l = 128$, as $c = 545$ in our experiments.

Alternatively, following [92], we may resort to a convex objective function parameterized by $\boldsymbol{V}$ in $\mathbb{R}^{c\times k}$, that penalizes the rank of $\boldsymbol{V}$. We learn $\boldsymbol{V}^j$ in $\mathbb{R}^{c^j\times k}$ for

all study j in $[N]$ solving the joint objective

$$\min_{(\boldsymbol{V}^j,\boldsymbol{b}^j)_j} -\sum_{j=1}^{N}\frac{(n^j)^\beta}{n^j}\sum_{i=1}^{n^j}\Big(l^j_{i,y_i}(\boldsymbol{V}^j,\boldsymbol{b}^j) - \log(\sum_{k=1}^{c^j}\exp l^j_{i,k}(\boldsymbol{V}^j,\boldsymbol{b}^j))\Big) + \lambda\left\|\left[\boldsymbol{V}^{1\top}\dots\boldsymbol{V}^{N\top}\right]\right\|_\star, \tag{9}$$

where $\|\boldsymbol{V}\|_\star$ is the nuclear norm of $\boldsymbol{V}$, defined as $\sum_{i=1}^{\min(c,k)}\sigma_i(\boldsymbol{V})$, where $(\sigma_i(\boldsymbol{V}))_i$ are the singular values of $\boldsymbol{V}$. The nuclear norm is a convex proxy for the rank of matrix $\boldsymbol{V}$. As a consequence, the rank of the solution decreases from $\min(c,k)$ to 0 as λ increases. The objective (9) is solvable using proximal methods, e.g., FISTA [93]. However, these methods become unpractical when c becomes large—it requires to perform a $c\times c$ singular value decomposition at each iteration. Fortunately, there exists a non-convex objective [94], amenable to stochastic gradient descent [95], that includes the solution of (9) as a minimizer. It is obtained by setting $l=\max(x,k)$ and adding ℓ_2^2 penalties to the objective (4):

$$\min_{\substack{\boldsymbol{L}\in\mathbb{R}^{l\times k}\\(\boldsymbol{U}^j,\boldsymbol{b}^j)_j}} -\sum_{j=1}^{N}\frac{(n^j)^\beta}{n^j}\sum_{i=1}^{n^j}\Big(l^j_{i,y_i}(\boldsymbol{U}^j,\boldsymbol{b}^j,\boldsymbol{L}) - \log(\sum_{k=1}^{c^j}\exp l^j_{i,k}(\boldsymbol{U}^j,\boldsymbol{b}^j,\boldsymbol{L}))\Big) + \frac{\lambda}{2}\Big(\|\boldsymbol{L}\|_F^2+\sum_{j=1}^{N}\|\boldsymbol{U}^j\|_F^2\Big).$$

We solve this objective using *Adam*, similarly to the main method. It is possible to continue using dropout in between the first and second layer while enforcing $\boldsymbol{V}$ to be low-rank—this can then be understood as a regularization technique through feature noising [96].

B.2.4 Empirical comparison of transfer penalties

Dropout versus ℓ_2. Both the dropout and low-rank approaches are a priori competitive to foster transfer learning. Our final method uses a combination of both, as it enforces a hard low-rank constraint and uses dropout. This choice was motivated by a first experiment, summarized in Fig D. We compare three regularization variants by measuring the improvement due to hard low-rank constraints and the difference between dropout and ℓ_2. The three estimators use input dropout ($p=0.25$). The first two estimators use ℓ_2 regularization. Dropout between layer 2 and 3 is initialized to $p=0.75$ in the third estimators. The first estimator does not use a hard-rank constraint ($l=c=545$), while others use $l=128$.[4] We observe that forcing $\boldsymbol{V}$ to be low-rank is beneficial (0.7% mean accuracy gain, 72% experiments with net increase) in the absence of dropout, and that dropout regularization performs significantly better than low-rank inducing ℓ_2 penalties (2.7% mean accuracy gain, 79% experiments with net increase). This justifies using dropout regularization.

[4]The reported ℓ_2 accuracy gain is larger than its actual performance when λ is set with cross-validation, as we take the highest performing λ on the *test* sets. Symmetrically, we may slightly improve results by setting dropout rates using cross-validation—we choose not to, to avoid the fragility of cross-validation in neuro-imaging [72].

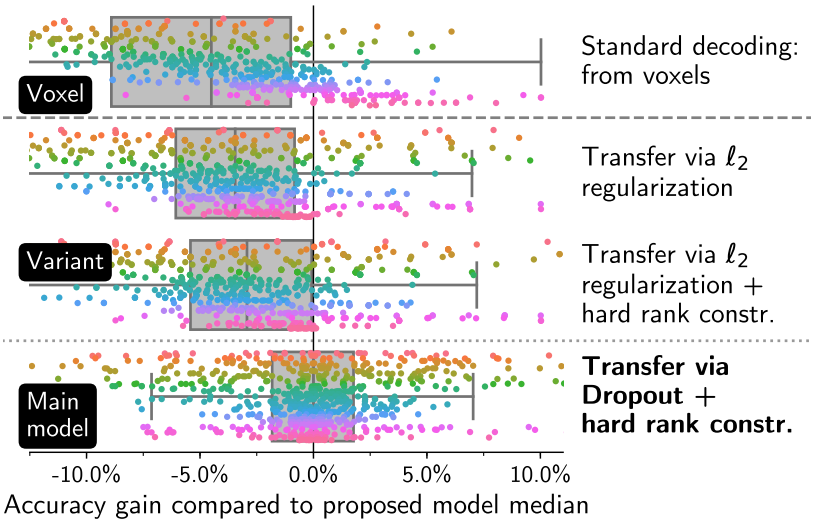


Fig. D. Quantitative comparison of transfer inducing regularizations. Dropout with hard-rank constraints outperforms ℓ_2 regularization with and without hard-rank constraints. Box plots calculated over 20 random data half-split and all studies.

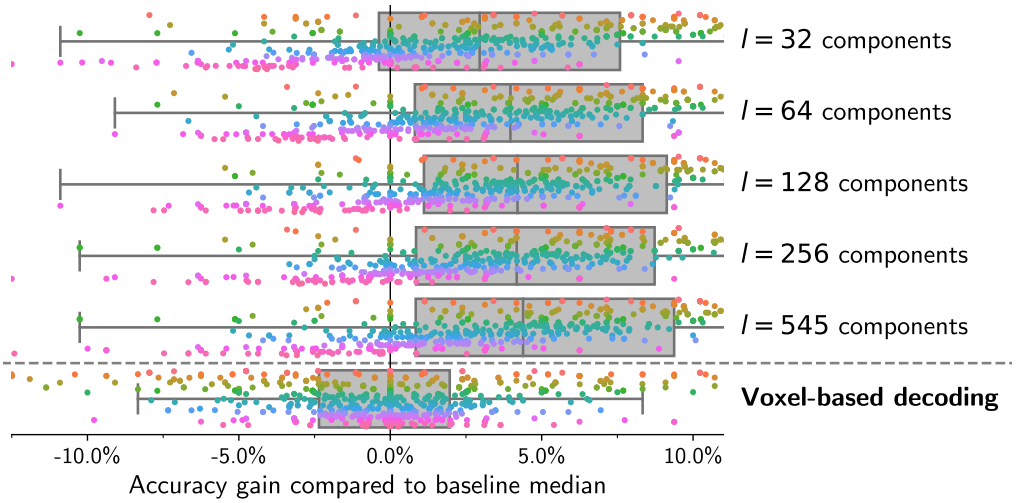


Fig. E. Performance of multi-study decoding for varying second layer width l.

Low-rank constraints and second-layer width. With dropout, the performance of multi-study decoding varies with the size of the latent space l, as displayed in Fig E. The performance reaches a plateau at $l \approx 128$. Setting a high l results in more scattered networks, so that different but similar MSTONs may be recruited to decode the same psychological condition (see examples in Fig A). Choosing a low l leads to slightly worse performances but more interpretable components. We therefore use $l = 128$, as if offers the best performance/interpretability trade-off.

First-layer width. Some previous work [60] studies the impact of projecting brain signal onto k functional units, for varying k and different fMRI analysis tasks. The conclusion of this work applies here: setting a high k ensures the best performances. We use $k = 465$ grey-matter components extracted from 512 full-brain components due to constraints in training—higher k may be used in future work.

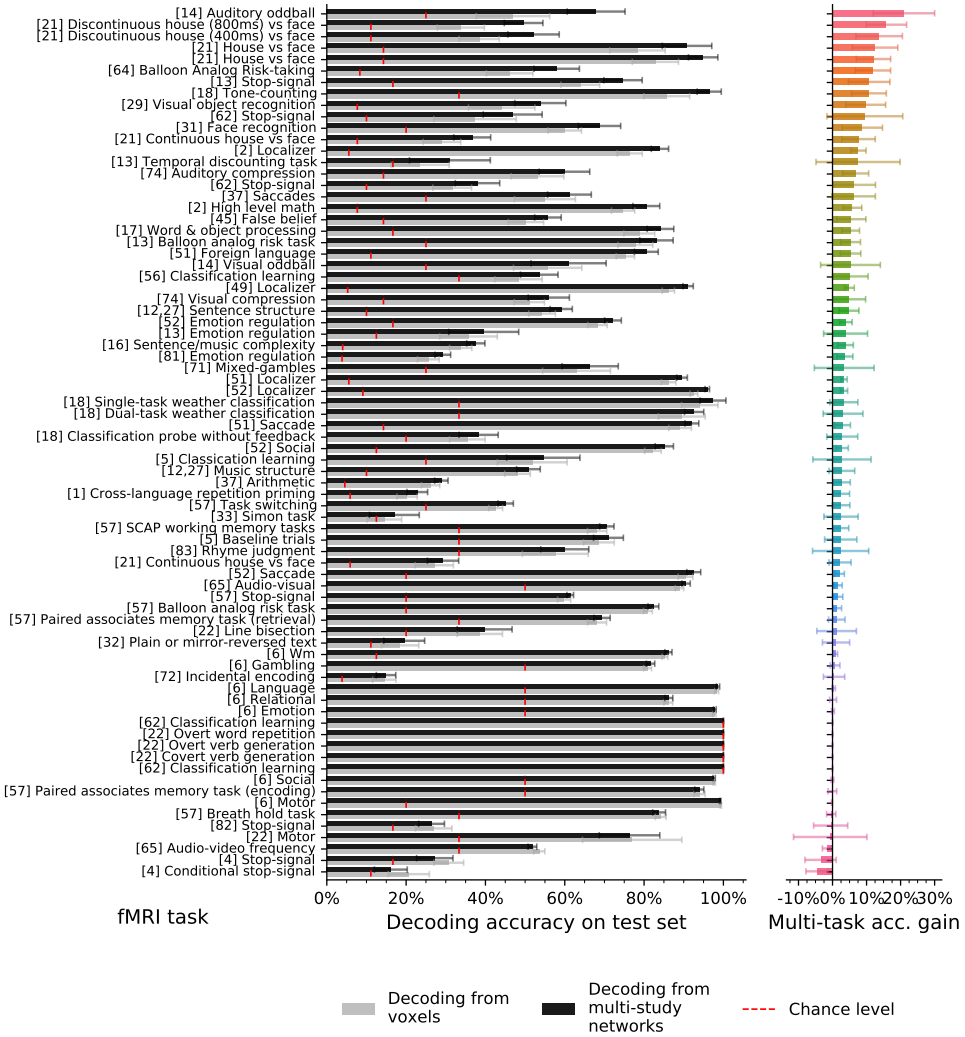


Fig. F. Performance of multi-study multi-task decoding, versus single-study single-task decoding from resting-state functional units. Numbers are reported in Table B.

B.3 Multi-study multi-task decoding

We have validated the multi-study decoding approach in a *per-site* setting, in which each study defines a single decoding task. Some studies include different fMRI tasks: we can also use each of these tasks to define a single decoding problem, and perform *multi-study multi-task decoding*. To evaluate this approach, we use the task annotations from the 35 studies of our corpus and obtain 76 classification tasks to be solved simultaneously. We compare the performance of the three-layer model, versus single-task decoding from the resting-state functional units. We use the exact same architecture as for multi-study training.

Results are displayed in Fig F. Multi-task training brings an improvement for 62/76 tasks. Quantitatively, the mean improvement is lower than the one obtained for within-study decoding (+3.9% vs +5.8%). This was expected, as the average chance-level in within-task decoding is higher than in within-study decoding. Using multi-task or multi-site modelling should depend on the purpose of the study.

B.4 Interpretability incentives

A core feature of our approach is model interpretability. Three aspects allow to find cognitive meaningful task-optimized networks. First, the initial first layer, learned on resting-state data, coarsens the resolution of networks in a way adapted to typical brain signals. Second, we compute a consensus model, so that the task-optimized network loadings held in $\boldsymbol{L}$ are non-negative and interpretable. Third, we initialize the second-layer weights so that $\boldsymbol{L}_{\text{init}}\boldsymbol{D}$ corresponds to resting-state functional networks $\boldsymbol{D}_l$, coarser than $\boldsymbol{D}$. This initialization is used both during the training phase and the consensus phase.

Consensus model and resting-state initialization. In Fig H, we measure the quantitative effects of the two later factors on decoder accuracy. Learning a consensus model using sparse NMF is crucial for finding interpretable direction in the span of $\boldsymbol{L}$. Without this refinement, the directions we obtain are similar to the one displayed in Fig GA, and are less interpretable. Both the consensus phase and the resting-state initialization contributes positively to the model decoding performance (0.6% mean accuracy gain, 66% experiments with net increase). We attribute this improvement to an ensembling effect similar to the benefits of bagging [97], as the final model summarizes 100 training runs on the same data, with different random seeds, and to the fact that resting-state networks form a good prior for task-optimized network.

Qualitatively, we show examples of three components found without resting-state initialization in Fig GB. Two of those are scattered networks, that capture various connected components whose co-occurrence is not interpretable: those components are likely artifacts due to random initialization. Using resting-state initialization finds such networks much less frequently. It remains interesting to note that most of the components found without resting-state based prior bear cognitive meaning, similar to the third components displayed in Fig GB.

B.4.1 Effect of selecting grey-matter components

We project data onto a subset of 465 out of 512 functional networks learned on HCP resting-state data, selecting the networks that intersect with an anatomical grey-matter mask. This avoids finding MSTONs that are distributed or formed with non grey-matter regions. In Fig GC, we show that without those precautions, our model finds networks located in the white matter and the cerebro-spinal fluid zones. Quantitatively (Fig I),

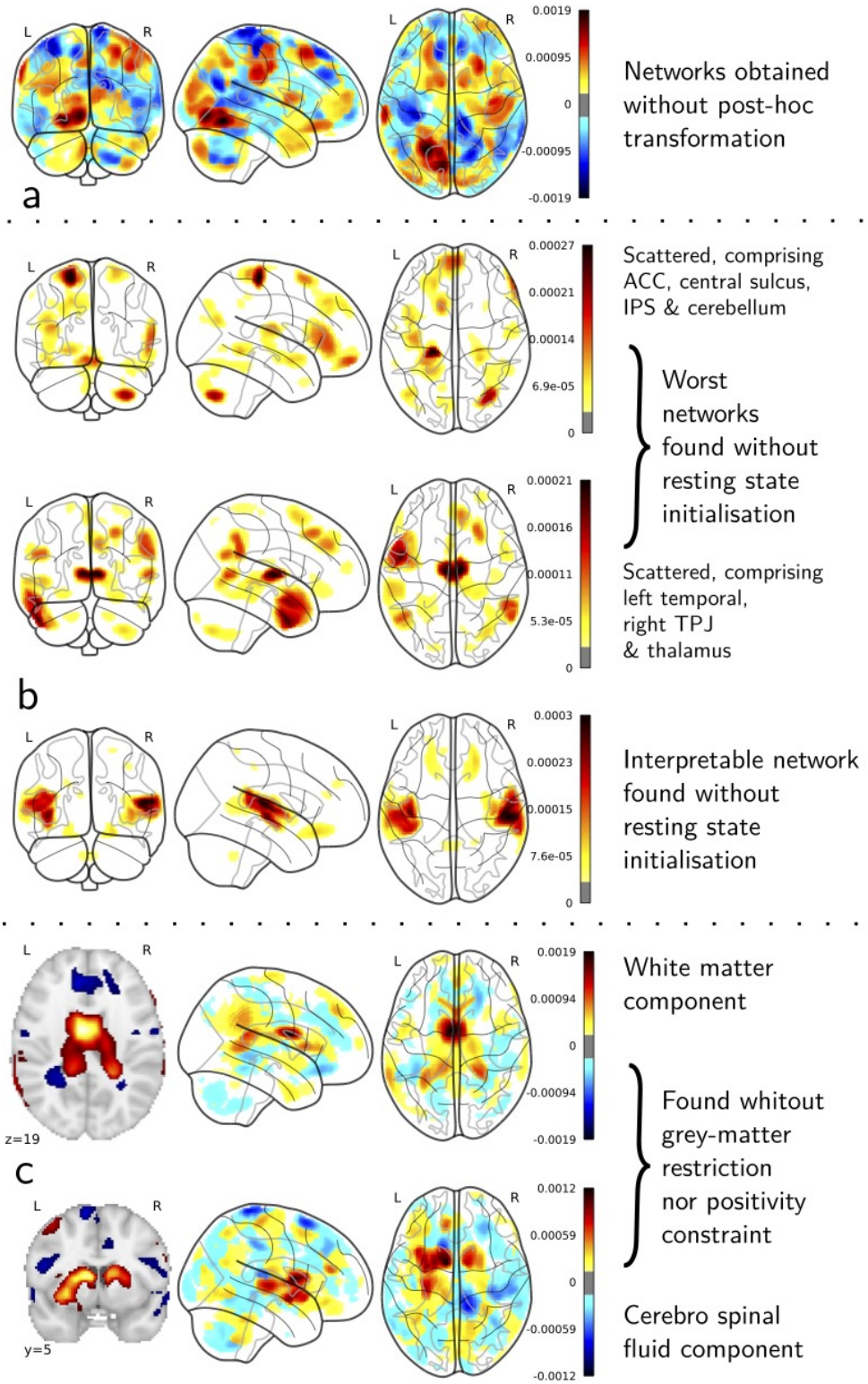


Fig. G. Effects of components selection. Without post-hoc transformation **(A)** resting-state based initialization **(B)** and grey matter components selection **(C)**, some task-optimized networks may be hard to interpret or not relevant from a cognitive perspective.

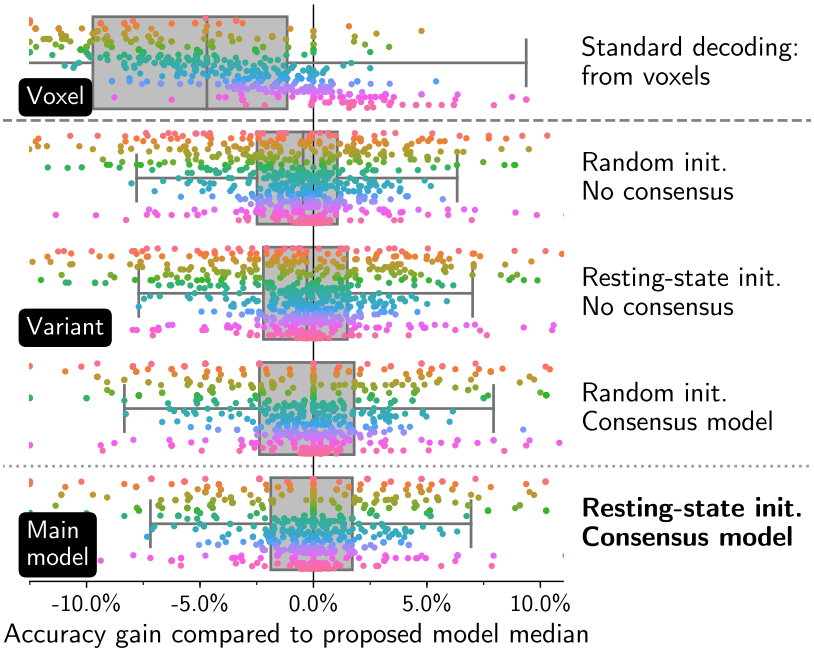


Fig. H. Quantitative improvement linked to ensembling and resting-state initialization. Box plots calculated over 20 random data half-split and all studies.

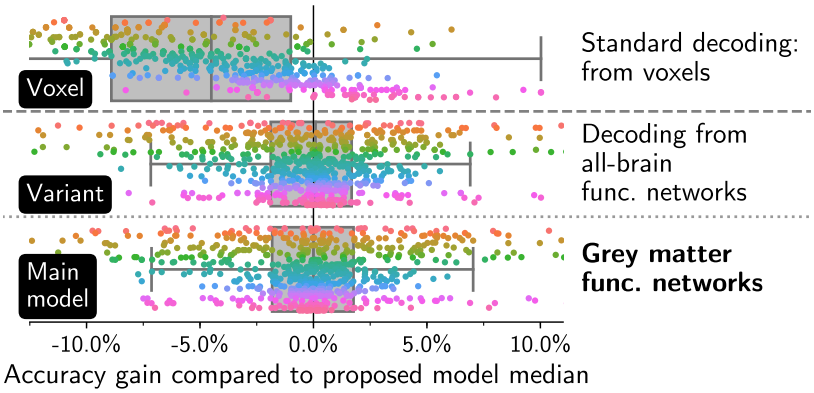


Fig. I. Quantitative improvement linked to working with a grey-matter mask. Working with functional networks located in the grey matter only do not have a significant impact on performance. Box plots calculated over 20 random data half-split and all studies.

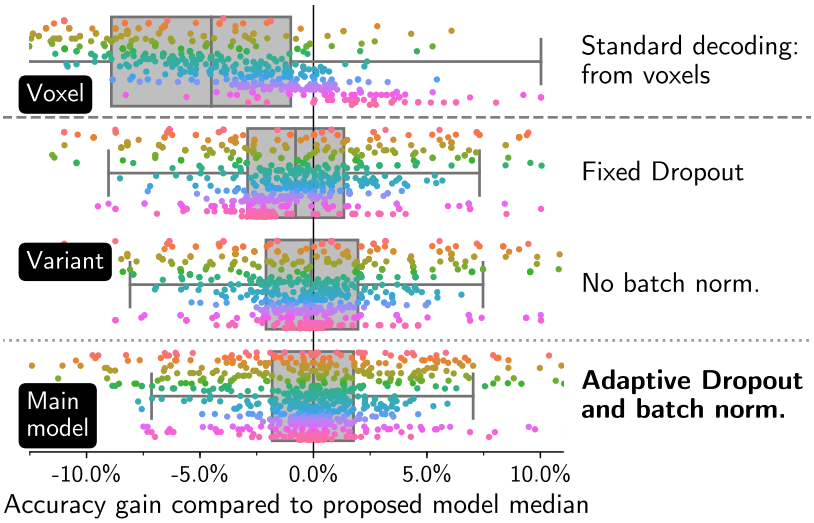


Fig. J. Batch normalization and adaptive variational dropout both have a beneficial impact on classification accuracy of the final learned decoder. Box plots calculated over 20 random data half-split and all studies.

as expected, performing classification from grey-matter components only brings a non-significant performance loss (0.03% median accuracy gain).

B.5 Effect of variational dropout and batch normalization

We introduced variational dropout and batch normalization in the training procedure of our algorithm. Fig J shows that it is indeed beneficial. Variational dropout brings a mean accuracy gain of 0.7% (64% experiments with net increase) compared to binary dropout; batch normalization benefit is smaller but positive (0.1% mean accuracy gain, 55% experiments with net increase), and allows faster training—in line with its original purpose [91].

B.6 Stronger improvement for smaller studies

To verify the finding of Fig 3 and evaluate the impact of training-size on multi-study decoding, we perform the following experiment. We restrict the study corpus to studies with more than 30 subjects, train the three-layer model on 15 subjects from each study, and evaluate its performance on the remaining population. We repeat this experiment 20 times.

We report results in Fig K. Transfer learning is positive for all studies (mean accuracy gain +4.8%.). This includes studies with a large complete cohort, for which transfer learning is uneffective when considering all available subjects (e.g. HCP, Fig 2 and data from the UCLA consortium). The multi-study approach is therefore particularly efficient for studies with less than 30 subjects, that are still the most common in the literature.

B.7 Effect of decoding difficulty

We investigate how the difficulty of a given decoding task (provided by a single study) influences the performance improvement due to multi-study decoding. For this, we report in Fig L the same numbers as in Fig 2, sorting studies by their chance level: lower chance level means "harder" decoding tasks, as contrasts must be selected in

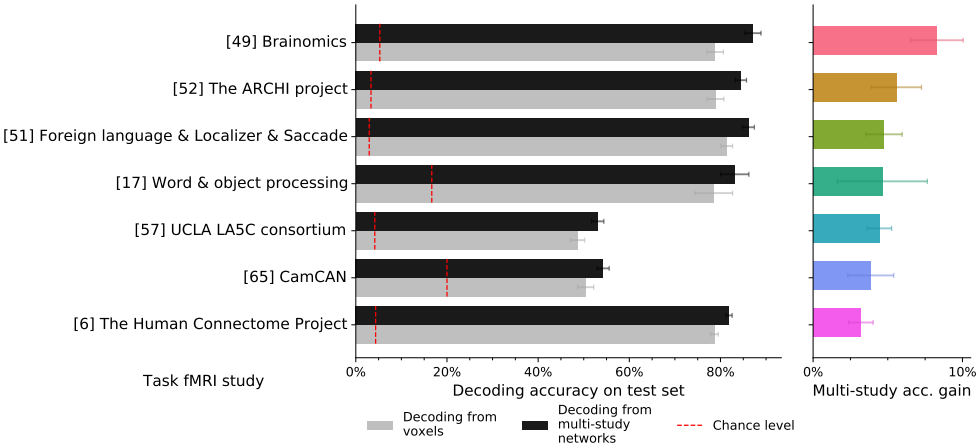


Fig. K. Performance of multi-study decoding with 15 training subjects per study.

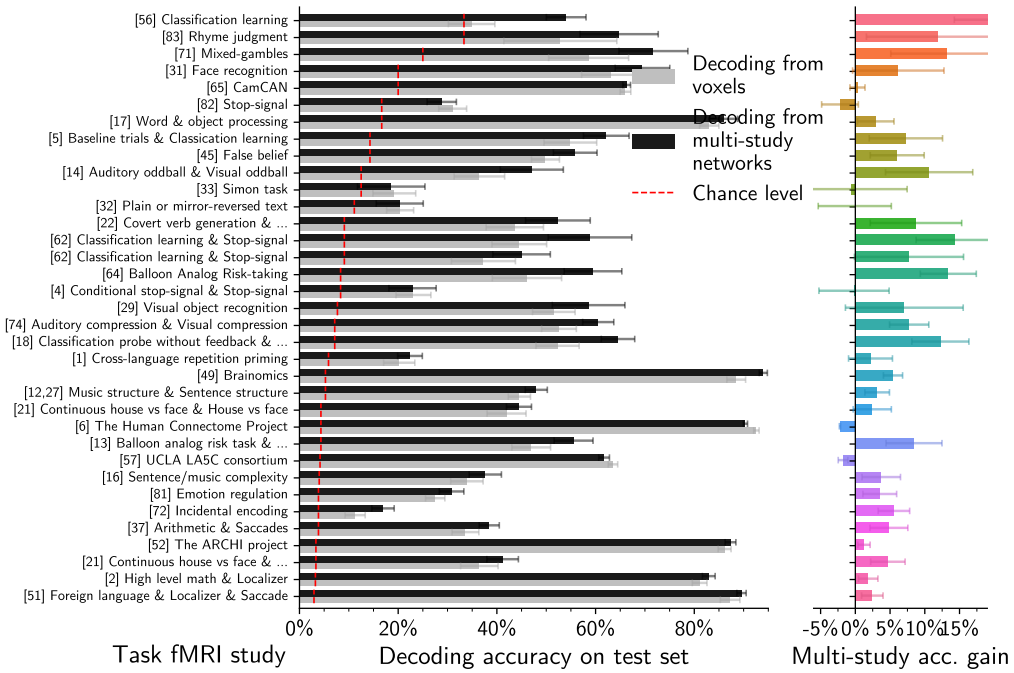


Fig. L. Performance improvement of multi-study decoding vs voxel-level decoding, **sorted by the chance level** of the decoding task of each study.

larger sets. We observe a slight tendency of higher improvement for easier tasks, although no strong conclusion may be drawn.

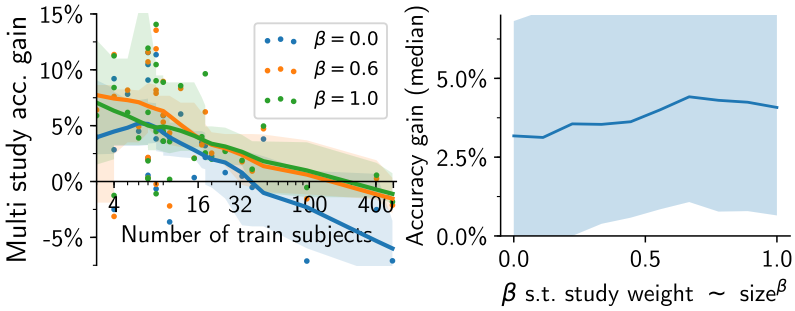


Fig. M. Impact of changing the study weight in the joint objective. Giving more weight ($\beta \to 1$) to large studies prevents negative transfer learning but may reduce overall performance. Small studies should not be given too much weight ($\beta \to 0$), as this voids the benefits of jointly training over bigger studies. An intermediary $\beta = 0.6$ gives the best performances. Error bars calculated over 20 random data half-split and all studies.

B.8 Effect of study weights

Our model learns the second and third layer weights by solving

$$\min_{\substack{\boldsymbol{L}\in\mathbb{R}^{l\times k}\\(\boldsymbol{U}^j,\boldsymbol{b}^j)_j}} -\sum_{j=1}^{N}\frac{(n^j)^\beta}{n^j}\sum_{i=1}^{n^j}\Big(l^j_{i,y_i}(\boldsymbol{U}^j,\boldsymbol{b}^j,\boldsymbol{L}) - \log(\sum_{k=1}^{c^j}\exp l^j_{i,k}(\boldsymbol{U}^j,\boldsymbol{b}^j,\boldsymbol{L}))\Big),$$

in which the many studies can be given various weights. At one extreme, we may consider that all studies of the corpus should be weighted the same, which amounts to setting $\beta = 0$ in (4). At the opposite, we can consider that each brain map from each study should have the same importance, which amounts to setting $\beta = 1$. As Fig MB shows, it is beneficial to set an intermediary β, typically $\beta = 0.6$. On the one hand, we want to give the smallest study of our corpus a non negligible importance; on the other hand, we want the large studies to remain more weighted than the smaller ones, as they should provide more accurate information. Our reweighting amounts to giving every study j an "effective sample size"

$$n^j_{\text{eff}} = \sum_{i=1}^{N} n^i \frac{{n^j}^\beta}{\sum_{i=1}^{N} {n^i}^\beta},$$

that is larger than the true sample size for smaller studies and smaller for larger studies. We observe on Fig MA that the negative transfer learning endured by large-study decoders such as HCP and LA5C reduces as these studies are given more weight ($\beta \to 1$). On the other hand, the performance on small datasets slightly reduces for $\beta > 0.6$. It also reduces for low β, hinting at the importance of using large studies for improving small studies decoding.

We thus have provided justifications for all the technical design choices made in training our decoding model: regularization, joint training, training refinements, choice of study weights.

B.9 Comparison with earlier work

We proposed a proof-of-concept, smaller-scale and harder to interpret multi-study decoding approach in [21]. This earlier work already relies on a three-layer linear model, with joint training of the second and third layer. Beyond its extended cognitive neuroscience point-of-view, the present work strongly improves the multi-study decoding methods and results.

Model interpretability. From a methodological point of view, [21] fail short of providing a principled way for interpreting results and extracting meaningful task-optimized networks, as those outlined in Fig 4. Their approach yields networks akin to Fig GA, which are not relevant from a cognitive perspective. A template-extracting approach that clusters the low-dimensional brain map representations is proposed; yet it remains exogenous to the model and does not perform convincingly. The consensus post-hoc transformation method we propose in this work addresses the issue of interpretability and finds cognitive directions that efficiently capture mental state information. As Fig B shows, these meaningful networks can be used as a cognitive atlas for improving decoding on newly acquired datasets, without joint training. Consensus through matrix factorization of the model weights also increases model performance (Fig B).

Architecture, constraints, training. The functional atlases used as a first-layer by [21] are smaller (up to 256 components) and not constrained to be non-negative. As we discovered, enforcing non-negativity of the first layer $\boldsymbol{D}$ and the second layer $\boldsymbol{L}$ (after ensembling) is crucial to interpret the prediction of the model. Using a larger functional atlas extracted from resting-state data ensures that no information is lost when reducing the dimension of brain maps. Initialization of the second-layer with resting-state information increases the model performance (Fig H), as well as the use of variational dropout [56] and batch normalization [91] (Fig J).

Data and validation. [21] pool only the results of 5 studies, which prevents the observation heavy transfer effects, and the extraction of broadly-valid cognitive directions. The present work validates the approach on 7 times more studies, proving that our multi-study approach is valid beyond proof-of-concept, and truly promising for the neuroscience community. To better explain the transfer of information across studies, we compare several transfer approaches (convex models, low-rank constraints, stochastic regularization: see Section B.2), and assess how classification maps are affected by the use of task-optimized network (Figs 6, 7 and C); this endeavor is missing in earlier work.

C Reproduction details and tables

In this last section, we detail our experiment pipeline, the numerical parameters needed for reproducing this study, and the sources from which we obtained our corpus of studies.

C.1 Software and parameters

We used *nilearn* [98] and *scikit-learn* [99] in our experiment pipelines, the stochastic solver from [59] to learn resting state dictionaries and *pytorch* [100] for model design and training. The *cogspaces* package that we have published provides the multi-scale resting-state dictionaries extracted from HCP, as those are costly to learn. It also provides the reduced representations of the data from the 35 studies we consider.

General cross-validation scheme. For every validation experiment and comparison, we perform 20 half-split of all data. Namely, we consider half of the subjects of every study for training, and test the decoder on the other half. As two studies [38] share subjects, we also ensure that no single subject appears in both the training and the test sets across studies.

Baseline parameter selection. We cross validate the λ parameter for the baseline multinomial regression classifiers, on a grid

$$\{10^i, i = \{-3, -2, -1, 0, 1, 2, 3\}\}.$$

Dropout rate. We use a dropout rate of $p = 0.25$ in between the first and second layer and initialize study-specific dropout rates with $p = 0.75$ in between the second-layer and third-layer classification heads (i.e., we set $\alpha = \frac{p}{1-p}$ in variational dropout).

Resting-state dictionaries. We obtain the 512-components and 128 components resting-state dictionaries by choosing λ on a grid

$$\{10^i, i = \{-5, -4, -3, -2, -1, 0, 1\}\},$$

so to obtain components that cover the whole brain with minimal overlap.

Consensus phase. We run the training procedure 100 times with different random seeds. We set $\lambda = 10^{-4}$, so as to obtain 80% sparsity. We tried $\lambda \in \{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$. Higher sparsity leads to a slight decrease in performance, lower sparsity is softer on symmetry breaking, which may reduce interpretability. This parameter has little influence as long as the sparsity remains higher than 50%.

Word-clouds. In Fig 4, we form word-clouds associated with the k-th MSTON network $\boldsymbol{Dl}_k$ as follows. We compute the correlations between each classification map $\boldsymbol{w}_c$, associated with a condition c, and the network $\boldsymbol{Dl}_k$ as

$$d_{k,c} = \frac{\langle \boldsymbol{Dl}_k, \boldsymbol{w}_c \rangle}{\|\boldsymbol{Dl}_k\|_2 \|\boldsymbol{w}_c\|_2}.$$

We then show the 20 contrast names with highest correlation values—this corresponds to the contrasts whose likelihood increases the most when the input data is pushed in the direction of $\boldsymbol{Dl}_k$. The height of the contrast name c in the word-cloud reflects the rank of the contrast in the sorted values $(d_{k,c})_c$ and the value $d_{k,c}$, using heuristics from the Python *word_cloud* package (`https://github.com/amueller/word_cloud`).

C.2 Validation metrics

We used two metrics to measure the performance of our models. To compare per-study decoding accuracy, we use the multi-class accuracy, defined as

$$a^j = \frac{\#\{i \in [c^j n^j], \hat{y}_i^j = y_i^j\}}{c^j n^j},$$

for study j, where $(\hat{y}_i^j)_{i\in[c^j n^j]}$ and $(y_i^j)_{i\in[c^j n^j]}$ encodes the predicted and ground-truth contrasts, respectively. Box plots presented in Fig 2 and Figs D–I reports the median and 25%, 75% quantiles of

$$\{a_r^j - \bar{a}_0^j, j \in [1, \dots, N], r \in [1, 2, \cdots, 20]\},$$

where r is the half-split run index and $\bar{a}_0^j$ is the median accuracy obtained for study j over 20 half-split.

We use balanced accuracy to measure the performance relative to a single contrast $y \in [1, \dots, c^j]$. It corresponds to the average of 1) the proportion of z-maps being correctly classified into y and 2) the proportion of z-maps being correctly classified into other classes. This metric has the advantage of being comparable across studies, as its chance level is always 50% no matter the number of contrasts in the study. We recall that the balanced accuracy b_y^i for study j and contrast y in $[1, \dots, c^j]$ is defined as

$$b_y^j \triangleq \frac{1}{2}\Big(\frac{n^j}{\#\{i \in [1, 2, \dots, c^j n^j], \hat{y}_i^j = y\}} + \frac{n^j(c^j - 1)}{\#\{i \in [1, 2, \dots, c^j n^j], \hat{y}_i^j \neq y\}}\Big).$$

C.3 Quantitative results per study, task and contrast

We report the accuracies displayed in Fig 2 in Table A (multi-study decoding), and the ones displayed in Fig F in Table B (multi-task decoding). We report the list of all contrasts used in this paper in Table C, as provided by the authors of each study. We report the associated balanced-accuracy when performing multi-study decoding, i.e. when we predict each contrasts among the set of all contrasts of a given study.

Table B. Accuracies per task in multi-site multi task decoding.

Study	Task	Chance level	Multi-task accuracy	Single-task accuracy	Accuracy gain
[22]	High level math	8%	$81 \pm 3\%$	$75 \pm 3\%$	$6 \pm 3\%$
	Localizer	6%	$84 \pm 2\%$	$76 \pm 3\%$	$8 \pm 2\%$
[23]	Emotion regulation	17%	$72 \pm 2\%$	$68 \pm 2\%$	$4 \pm 2\%$
	Localizer	9%	$96 \pm 1\%$	$93 \pm 1\%$	$3 \pm 1\%$
	Saccade	20%	$93 \pm 2\%$	$90 \pm 2\%$	$2 \pm 1\%$
	Social	12%	$85 \pm 2\%$	$82 \pm 2\%$	$3 \pm 2\%$
[24]	Localizer	5%	$91 \pm 1\%$	$86 \pm 2\%$	$5 \pm 1\%$
[25]	Audio-video frequency	33%	$52 \pm 1\%$	$54 \pm 1\%$	$-2 \pm 1\%$
	Audio-visual	50%	$91 \pm 1\%$	$89 \pm 1\%$	$2 \pm 1\%$
[26, 27]	Music structure	10%	$51 \pm 3\%$	$48 \pm 3\%$	$3 \pm 4\%$
	Sentence structure	10%	$59 \pm 3\%$	$54 \pm 3\%$	$5 \pm 3\%$
[28]	Sentence/music complexity	4%	$38 \pm 2\%$	$34 \pm 3\%$	$4 \pm 2\%$
[29]	Balloon Analog Risk-taking	8%	$58 \pm 6\%$	$46 \pm 6\%$	$12 \pm 5\%$
[30]	Baseline trials	33%	$71 \pm 4\%$	$69 \pm 4\%$	$2 \pm 5\%$

Continued on next page

Study	Task	Chance level	Multi-task accuracy	Single-task accuracy	Accuracy gain
	Classication learning	25%	$55 \pm 9\%$	$52 \pm 9\%$	$3 \pm 9\%$
[31]	Rhyme judgment	33%	$60 \pm 6\%$	$58 \pm 8\%$	$2 \pm 8\%$
[32]	Mixed-gambles	25%	$66 \pm 7\%$	$63 \pm 9\%$	$3 \pm 9\%$
[33]	Plain or mirror-reversed text	11%	$20 \pm 5\%$	$18 \pm 5\%$	$1 \pm 4\%$
[34]	Stop-signal	17%	$26 \pm 3\%$	$27 \pm 5\%$	$-1 \pm 5\%$
[35]	Conditional stop-signal	11%	$16 \pm 4\%$	$21 \pm 5\%$	$-5 \pm 3\%$
	Stop-signal	17%	$27 \pm 5\%$	$31 \pm 4\%$	$-4 \pm 5\%$
[36]	Balloon analog risk task	25%	$83 \pm 4\%$	$78 \pm 4\%$	$5 \pm 3\%$
	Emotion regulation	12%	$40 \pm 9\%$	$36 \pm 7\%$	$4 \pm 6\%$
	Stop-signal	17%	$75 \pm 5\%$	$64 \pm 5\%$	$11 \pm 6\%$
	Temporal discounting task	17%	$31 \pm 10\%$	$24 \pm 7\%$	$8 \pm 12\%$
[37]	Classification probe without feedback	20%	$38 \pm 5\%$	$35 \pm 4\%$	$3 \pm 5\%$
	Dual-task weather classification	33%	$93 \pm 2\%$	$90 \pm 6\%$	$3 \pm 6\%$
	Single-task weather classification	33%	$97 \pm 3\%$	$94 \pm 5\%$	$3 \pm 4\%$
	Tone-counting	33%	$96 \pm 3\%$	$86 \pm 6\%$	$11 \pm 5\%$
[38]	Classification learning	100%	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
	Stop-signal	10%	$38 \pm 6\%$	$32 \pm 5\%$	$6 \pm 6\%$
	Classification learning	100%	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
	Stop-signal	10%	$47 \pm 7\%$	$37 \pm 10\%$	$9 \pm 11\%$
[39]	Cross-language repetition priming	6%	$23 \pm 3\%$	$20 \pm 3\%$	$3 \pm 2\%$
[40]	Classification learning	33%	$54 \pm 5\%$	$48 \pm 6\%$	$5 \pm 5\%$
[41]	Simon task	12%	$17 \pm 6\%$	$15 \pm 4\%$	$2 \pm 5\%$
[7]	Visual object recognition	8%	$54 \pm 6\%$	$44 \pm 8\%$	$10 \pm 6\%$
[42]	Word & object processing	17%	$84 \pm 3\%$	$79 \pm 4\%$	$5 \pm 3\%$
[43]	Emotion regulation	4%	$29 \pm 2\%$	$26 \pm 3\%$	$4 \pm 2\%$
[44]	False belief	14%	$56 \pm 3\%$	$50 \pm 4\%$	$6 \pm 4\%$
[45]	Incidental encoding	4%	$15 \pm 2\%$	$14 \pm 3\%$	$0 \pm 3\%$
[46]	Covert verb generation	100%	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
	Line bisection	20%	$40 \pm 7\%$	$39 \pm 6\%$	$1 \pm 6\%$
	Motor	33%	$76 \pm 8\%$	$77 \pm 13\%$	$-1 \pm 11\%$
	Overt verb generation	100%	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
	Overt word repetition	100%	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
[47]	Auditory oddball	25%	$68 \pm 7\%$	$47 \pm 9\%$	$21 \pm 9\%$
	Visual oddball	25%	$61 \pm 9\%$	$56 \pm 9\%$	$5 \pm 9\%$
[48]	Continuous house vs face	8%	$37 \pm 5\%$	$29 \pm 5\%$	$8 \pm 5\%$
	Discontinuous house (800ms) vs face	11%	$50 \pm 5\%$	$34 \pm 6\%$	$16 \pm 6\%$
	Discoutinuous house (400ms) vs face	11%	$52 \pm 6\%$	$38 \pm 5\%$	$14 \pm 7\%$
	House vs face	14%	$91 \pm 6\%$	$78 \pm 7\%$	$13 \pm 7\%$
	Continuous house vs face	6%	$29 \pm 4\%$	$27 \pm 5\%$	$2 \pm 3\%$
	House vs face	14%	$95 \pm 4\%$	$83 \pm 6\%$	$12 \pm 5\%$
[49]	Emotion	50%	$98 \pm 0\%$	$98 \pm 0\%$	$0 \pm 0\%$
	Gambling	50%	$82 \pm 1\%$	$81 \pm 1\%$	$1 \pm 1\%$
	Language	50%	$99 \pm 0\%$	$98 \pm 0\%$	$0 \pm 1\%$
	Motor	20%	$99 \pm 0\%$	$99 \pm 0\%$	$-0 \pm 0\%$
	Relational	50%	$86 \pm 1\%$	$86 \pm 1\%$	$0 \pm 1\%$
	Social	50%	$98 \pm 0\%$	$98 \pm 0\%$	$-0 \pm 0\%$
	Wm	12%	$86 \pm 1\%$	$85 \pm 1\%$	$1 \pm 1\%$
[50]	Face recognition	20%	$69 \pm 5\%$	$60 \pm 4\%$	$9 \pm 6\%$
[51]	Arithmetic	5%	$29 \pm 2\%$	$26 \pm 2\%$	$3 \pm 3\%$
	Saccades	25%	$61 \pm 6\%$	$55 \pm 8\%$	$6 \pm 6\%$
[52]	Balloon analog risk task	20%	$82 \pm 1\%$	$81 \pm 1\%$	$1 \pm 1\%$
	Breath hold task	33%	$84 \pm 2\%$	$84 \pm 1\%$	$-0 \pm 1\%$
	Paired associates memory task (encoding)	50%	$94 \pm 1\%$	$94 \pm 1\%$	$-0 \pm 1\%$

Continued on next page

Study	Task	Chance level	Multi-task accuracy	Single-task accuracy	Accuracy gain
	Paired associates memory task (retrieval)	33%	$69 \pm 2\%$	$68 \pm 2\%$	$1 \pm 2\%$
	SCAP working memory tasks	33%	$71 \pm 2\%$	$68 \pm 2\%$	$2 \pm 2\%$
	Stop-signal	20%	$61 \pm 1\%$	$60 \pm 2\%$	$2 \pm 1\%$
	Task switching	25%	$45 \pm 2\%$	$43 \pm 2\%$	$3 \pm 3\%$
[53]	Foreign language	11%	$81 \pm 3\%$	$75 \pm 2\%$	$5 \pm 3\%$
	Localizer	6%	$90 \pm 1\%$	$86 \pm 2\%$	$3 \pm 1\%$
	Saccade	14%	$92 \pm 2\%$	$89 \pm 3\%$	$3 \pm 2\%$
[54]	Auditory compression	14%	$60 \pm 6\%$	$53 \pm 7\%$	$7 \pm 4\%$
	Visual compression	14%	$56 \pm 5\%$	$51 \pm 4\%$	$5 \pm 5\%$

Table C. List of all contrasts used in this paper, per study of origin and task.

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
[22]	High level math	body vs baseline	$90 \pm 5\%$	$90 \pm 5\%$	$1 \pm 5\%$
		body vs checkerboard	$94 \pm 5\%$	$97 \pm 3\%$	$-3 \pm 4\%$
		checkerboard vs baseline	$88 \pm 4\%$	$89 \pm 5\%$	$-1 \pm 5\%$
		equation vs baseline	$83 \pm 4\%$	$84 \pm 5\%$	$-1 \pm 4\%$
		equation vs number	$97 \pm 2\%$	$96 \pm 3\%$	$2 \pm 3\%$
		face vs baseline	$90 \pm 5\%$	$92 \pm 3\%$	$-2 \pm 4\%$
		face vs house	$100 \pm 0\%$	$98 \pm 3\%$	$2 \pm 3\%$
		house vs baseline	$88 \pm 6\%$	$88 \pm 8\%$	$0 \pm 7\%$
		number vs baseline	$86 \pm 5\%$	$89 \pm 6\%$	$-3 \pm 6\%$
		number vs word	$95 \pm 4\%$	$88 \pm 6\%$	$7 \pm 5\%$
		tool vs baseline	$83 \pm 6\%$	$84 \pm 8\%$	$-1 \pm 8\%$
		tool vs checkerboard	$90 \pm 5\%$	$86 \pm 5\%$	$3 \pm 5\%$
		word vs baseline	$85 \pm 8\%$	$81 \pm 6\%$	$3 \pm 6\%$
	Localizer	auditory calculation vs baseline	$95 \pm 3\%$	$93 \pm 2\%$	$2 \pm 2\%$
		auditory calculation vs sentences	$72 \pm 8\%$	$71 \pm 12\%$	$2 \pm 16\%$
		auditory left motor vs baseline	$98 \pm 2\%$	$96 \pm 2\%$	$2 \pm 2\%$
		auditory motor vs sentences	$94 \pm 3\%$	$91 \pm 5\%$	$3 \pm 4\%$
		auditory right motor vs baseline	$98 \pm 3\%$	$93 \pm 3\%$	$4 \pm 5\%$
		auditory right vs left motor	$80 \pm 5\%$	$80 \pm 6\%$	$-1 \pm 6\%$
		auditory sentences vs baseline	$100 \pm 0\%$	$99 \pm 2\%$	$1 \pm 2\%$
		horizontal checkerboard vs baseline	$95 \pm 3\%$	$98 \pm 2\%$	$-2 \pm 3\%$
		horizontal vs vertical checkerboard	$92 \pm 3\%$	$95 \pm 3\%$	$-3 \pm 5\%$
		vertical checkerboard vs baseline	$92 \pm 4\%$	$98 \pm 2\%$	$-6 \pm 5\%$
		visual calculation vs baseline	$96 \pm 3\%$	$93 \pm 3\%$	$3 \pm 3\%$
		visual calculation vs sentences	$70 \pm 9\%$	$67 \pm 10\%$	$4 \pm 14\%$
		visual left motor vs baseline	$99 \pm 2\%$	$99 \pm 2\%$	$0 \pm 1\%$
		visual motor vs sentences	$99 \pm 2\%$	$96 \pm 3\%$	$3 \pm 3\%$
		visual right motor vs baseline	$97 \pm 3\%$	$95 \pm 3\%$	$1 \pm 3\%$
		visual right vs left motor	$82 \pm 7\%$	$77 \pm 7\%$	$5 \pm 6\%$
		visual sentences vs baseline	$98 \pm 2\%$	$96 \pm 3\%$	$2 \pm 3\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		visual sentences vs checkerboard	$100 \pm 0\%$	$97 \pm 2\%$	$2 \pm 2\%$
[23]	Emotion regulation	expression control	$81 \pm 4\%$	$83 \pm 4\%$	$-1 \pm 4\%$
		expression intention	$96 \pm 1\%$	$94 \pm 2\%$	$2 \pm 1\%$
		expression sex	$84 \pm 4\%$	$85 \pm 4\%$	$-1 \pm 3\%$
		face control	$79 \pm 4\%$	$78 \pm 4\%$	$1 \pm 5\%$
		face sex	$81 \pm 2\%$	$81 \pm 4\%$	$0 \pm 4\%$
		face trusty	$90 \pm 2\%$	$89 \pm 3\%$	$2 \pm 2\%$
	Localizer	audio	$97 \pm 2\%$	$93 \pm 4\%$	$5 \pm 3\%$
		calculaudio	$97 \pm 2\%$	$96 \pm 2\%$	$2 \pm 1\%$
		calculvideo	$97 \pm 2\%$	$94 \pm 2\%$	$2 \pm 3\%$
		clicdaudio	$98 \pm 1\%$	$98 \pm 1\%$	$0 \pm 1\%$
		clicdvideo	$98 \pm 1\%$	$98 \pm 1\%$	$1 \pm 1\%$
		clicgaudio	$98 \pm 1\%$	$98 \pm 2\%$	$1 \pm 2\%$
		clicgvideo	$98 \pm 1\%$	$96 \pm 2\%$	$2 \pm 1\%$
		computation	$97 \pm 2\%$	$96 \pm 2\%$	$1 \pm 2\%$
		damier h	$97 \pm 2\%$	$98 \pm 1\%$	$-1 \pm 2\%$
		damier v	$98 \pm 1\%$	$99 \pm 1\%$	$-1 \pm 1\%$
		motor-cognitive	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
	Saccade	object grasp	$97 \pm 2\%$	$95 \pm 2\%$	$1 \pm 2\%$
		object orientation	$94 \pm 2\%$	$92 \pm 3\%$	$2 \pm 3\%$
		rotation hand	$97 \pm 1\%$	$96 \pm 2\%$	$0 \pm 1\%$
		rotation side	$95 \pm 1\%$	$94 \pm 2\%$	$1 \pm 2\%$
		saccade	$98 \pm 1\%$	$96 \pm 2\%$	$2 \pm 2\%$
	Social	false belief audio	$98 \pm 1\%$	$97 \pm 1\%$	$1 \pm 1\%$
		false belief video	$92 \pm 3\%$	$91 \pm 3\%$	$1 \pm 3\%$
		mecanistic audio	$96 \pm 3\%$	$94 \pm 3\%$	$2 \pm 2\%$
		mecanistic video	$88 \pm 2\%$	$90 \pm 3\%$	$-2 \pm 2\%$
		non speech	$92 \pm 3\%$	$93 \pm 3\%$	$-1 \pm 3\%$
		speech	$90 \pm 4\%$	$92 \pm 3\%$	$-2 \pm 2\%$
		triangle intention	$90 \pm 3\%$	$89 \pm 3\%$	$0 \pm 2\%$
		triangle random	$91 \pm 3\%$	$92 \pm 2\%$	$-1 \pm 2\%$
[24]	Localizer	auditory calculation	$99 \pm 1\%$	$99 \pm 0\%$	$0 \pm 1\%$
		auditory processing	$100 \pm 0\%$	$93 \pm 5\%$	$7 \pm 5\%$
		auditory sentences	$98 \pm 1\%$	$97 \pm 2\%$	$1 \pm 1\%$
		auditory&visual calculation	$97 \pm 2\%$	$93 \pm 3\%$	$5 \pm 3\%$
		auditory&visual sentences	$95 \pm 2\%$	$90 \pm 3\%$	$5 \pm 3\%$
		checkerboard	$90 \pm 3\%$	$86 \pm 3\%$	$4 \pm 2\%$
		effects of interest	$100 \pm 0\%$	$85 \pm 4\%$	$14 \pm 4\%$
		horizontal checkerboard	$89 \pm 3\%$	$93 \pm 2\%$	$-4 \pm 2\%$
		left auditory click	$99 \pm 1\%$	$98 \pm 1\%$	$0 \pm 1\%$
		left auditory&visual click	$97 \pm 2\%$	$93 \pm 3\%$	$4 \pm 3\%$
		left visual click	$99 \pm 1\%$	$98 \pm 2\%$	$2 \pm 2\%$
		motor	$97 \pm 1\%$	$91 \pm 3\%$	$5 \pm 3\%$
		right auditory click	$98 \pm 1\%$	$96 \pm 2\%$	$2 \pm 2\%$
		right auditory&visual click	$95 \pm 3\%$	$93 \pm 4\%$	$2 \pm 3\%$
		right visual click	$98 \pm 1\%$	$96 \pm 2\%$	$2 \pm 1\%$
		vertical checkerboard	$95 \pm 2\%$	$98 \pm 1\%$	$-3 \pm 1\%$
		visual calculation	$97 \pm 1\%$	$97 \pm 2\%$	$0 \pm 1\%$
		visual processing	$99 \pm 1\%$	$93 \pm 3\%$	$6 \pm 3\%$
		visual sentences	$97 \pm 2\%$	$97 \pm 2\%$	$1 \pm 3\%$
[25]	Audio-video frequency	audvid1200	$75 \pm 1\%$	$75 \pm 1\%$	$0 \pm 1\%$
		audvid300	$70 \pm 1\%$	$71 \pm 1\%$	$-0 \pm 2\%$
		audvid600	$66 \pm 1\%$	$67 \pm 1\%$	$-1 \pm 1\%$
	Audio-visual	audonly	$91 \pm 1\%$	$89 \pm 1\%$	$1 \pm 1\%$
		vidonly	$93 \pm 1\%$	$92 \pm 1\%$	$1 \pm 1\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
[26, 27]	Music structure	c01 c02 vs c16 c08 music	$75 \pm 5\%$	$72 \pm 6\%$	$3 \pm 7\%$
		c01 music vs baseline	$57 \pm 4\%$	$56 \pm 4\%$	$1 \pm 3\%$
		c02 music vs baseline	$58 \pm 3\%$	$57 \pm 4\%$	$1 \pm 5\%$
		c04 music vs baseline	$54 \pm 3\%$	$54 \pm 4\%$	$0 \pm 4\%$
		c08 music vs baseline	$55 \pm 4\%$	$57 \pm 4\%$	$-2 \pm 4\%$
		c16 c08 music vs motor	$89 \pm 6\%$	$85 \pm 6\%$	$4 \pm 4\%$
		c16 c08 vs c01 c02 music	$75 \pm 5\%$	$71 \pm 5\%$	$4 \pm 6\%$
		c16 music vs baseline	$55 \pm 4\%$	$57 \pm 6\%$	$-3 \pm 3\%$
		motor vs baseline	$91 \pm 2\%$	$86 \pm 3\%$	$5 \pm 2\%$
		motor vs c16 c08 music	$86 \pm 6\%$	$86 \pm 6\%$	$-0 \pm 3\%$
	Sentence structure	c01 c02 vs c16 c08 language	$87 \pm 5\%$	$85 \pm 5\%$	$1 \pm 3\%$
		c01 language vs baseline	$61 \pm 6\%$	$62 \pm 5\%$	$-1 \pm 3\%$
		c02 language vs baseline	$63 \pm 6\%$	$60 \pm 5\%$	$3 \pm 6\%$
		c04 language vs baseline	$59 \pm 5\%$	$60 \pm 6\%$	$-1 \pm 6\%$
		c08 language vs baseline	$60 \pm 4\%$	$57 \pm 5\%$	$3 \pm 6\%$
		c16 c08 language vs motor	$86 \pm 5\%$	$84 \pm 4\%$	$2 \pm 4\%$
		c16 c08 vs c01 c02 language	$88 \pm 4\%$	$85 \pm 5\%$	$3 \pm 4\%$
		c16 language vs baseline	$74 \pm 4\%$	$72 \pm 5\%$	$3 \pm 4\%$
		motor vs baseline	$91 \pm 2\%$	$86 \pm 3\%$	$5 \pm 2\%$
		motor vs c16 c08 language	$85 \pm 6\%$	$84 \pm 6\%$	$1 \pm 5\%$
[28]	Sentence/music complexity	c01 c02 vs c12 c06	$90 \pm 9\%$	$81 \pm 11\%$	$8 \pm 7\%$
		c01 vs baseline	$57 \pm 6\%$	$55 \pm 5\%$	$2 \pm 6\%$
		c02 vs baseline	$56 \pm 6\%$	$54 \pm 7\%$	$2 \pm 6\%$
		c03 vs baseline	$53 \pm 4\%$	$50 \pm 3\%$	$3 \pm 4\%$
		c04 vs baseline	$54 \pm 5\%$	$54 \pm 4\%$	$0 \pm 6\%$
		c06 vs baseline	$62 \pm 6\%$	$58 \pm 6\%$	$4 \pm 6\%$
		c12 c06 vs c01 c02	$90 \pm 8\%$	$80 \pm 10\%$	$10 \pm 7\%$
		c12 c06 vs motor	$91 \pm 7\%$	$88 \pm 9\%$	$3 \pm 6\%$
		c12 vs baseline	$74 \pm 9\%$	$67 \pm 8\%$	$7 \pm 8\%$
		motor vs baseline	$98 \pm 2\%$	$99 \pm 1\%$	$-0 \pm 2\%$
		motor vs c12 c06	$89 \pm 9\%$	$88 \pm 8\%$	$1 \pm 6\%$
		nc3 vs baseline	$53 \pm 4\%$	$51 \pm 4\%$	$1 \pm 5\%$
		nc4 vs baseline	$56 \pm 6\%$	$57 \pm 9\%$	$-1 \pm 6\%$
		pseudo c01 c02 vs c12 c06	$71 \pm 10\%$	$73 \pm 10\%$	$-2 \pm 9\%$
		pseudo c01 vs baseline	$55 \pm 5\%$	$55 \pm 4\%$	$0 \pm 5\%$
		pseudo c02 vs baseline	$52 \pm 3\%$	$52 \pm 2\%$	$0 \pm 4\%$
		pseudo c03 vs baseline	$50 \pm 3\%$	$50 \pm 1\%$	$1 \pm 3\%$
		pseudo c04 vs baseline	$53 \pm 4\%$	$52 \pm 3\%$	$1 \pm 5\%$
		pseudo c06 vs baseline	$52 \pm 3\%$	$53 \pm 3\%$	$-1 \pm 4\%$
		pseudo c12 c06 vs c01 c02	$72 \pm 11\%$	$73 \pm 10\%$	$-1 \pm 9\%$
		pseudo c12 c06 vs motor	$90 \pm 9\%$	$84 \pm 9\%$	$6 \pm 7\%$
		pseudo c12 vs baseline	$62 \pm 7\%$	$60 \pm 5\%$	$2 \pm 5\%$
		pseudo motor vs c12 c06	$92 \pm 8\%$	$87 \pm 8\%$	$5 \pm 7\%$
		pseudo nc3 vs baseline	$54 \pm 6\%$	$54 \pm 4\%$	$1 \pm 7\%$
		pseudo nc4 vs baseline	$52 \pm 2\%$	$52 \pm 4\%$	$-0 \pm 3\%$
[29]	Balloon Analog Risk-taking	cash fixed vs baseline	$72 \pm 7\%$	$66 \pm 6\%$	$6 \pm 8\%$
		control fixed vs baseline	$77 \pm 7\%$	$60 \pm 6\%$	$17 \pm 6\%$
		control pumps demean vs baseline	$80 \pm 8\%$	$74 \pm 9\%$	$6 \pm 9\%$
		ctrl demean vs pumps demean	$88 \pm 7\%$	$86 \pm 7\%$	$2 \pm 7\%$
		ctrl fixed vs pumps fixed	$86 \pm 5\%$	$76 \pm 7\%$	$10 \pm 7\%$
		ctrl realrt vs pumps realrt	$73 \pm 9\%$	$66 \pm 9\%$	$7 \pm 8\%$
		explode fixed vs baseline	$83 \pm 6\%$	$81 \pm 7\%$	$2 \pm 7\%$
		pumps demean vs baseline	$82 \pm 8\%$	$71 \pm 9\%$	$11 \pm 8\%$
		pumps demean vs ctrl demean	$73 \pm 8\%$	$71 \pm 12\%$	$2 \pm 8\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		pumps fixed vs baseline	$75 \pm 8\%$	$72 \pm 8\%$	$3 \pm 8\%$
		pumps fixed vs ctrl fixed	$78 \pm 6\%$	$65 \pm 8\%$	$13 \pm 9\%$
		pumps realrt vs ctrl realrt	$68 \pm 8\%$	$60 \pm 7\%$	$8 \pm 10\%$
[30]	Baseline trials	feedback vs baseline	$89 \pm 3\%$	$84 \pm 6\%$	$5 \pm 5\%$
		task vs baseline	$78 \pm 6\%$	$76 \pm 5\%$	$2 \pm 7\%$
		task vs feedback	$76 \pm 6\%$	$70 \pm 8\%$	$5 \pm 6\%$
	Classication learning	deterministic classification vs baseline	$69 \pm 8\%$	$70 \pm 9\%$	$-1 \pm 9\%$
		deterministic vs probabilistic classification	$76 \pm 8\%$	$71 \pm 11\%$	$5 \pm 13\%$
		probabilistic classification vs baseline	$73 \pm 11\%$	$63 \pm 6\%$	$10 \pm 10\%$
		probabilistic vs deterministic classification	$75 \pm 6\%$	$72 \pm 7\%$	$3 \pm 7\%$
[31]	Rhyme judgment	pseudoword vs baseline	$68 \pm 7\%$	$60 \pm 10\%$	$8 \pm 9\%$
		word vs baseline	$64 \pm 8\%$	$58 \pm 10\%$	$6 \pm 10\%$
		word vs pseudoword	$89 \pm 9\%$	$76 \pm 11\%$	$13 \pm 11\%$
[32]	Mixed-gambles	distance from indifference vs baseline	$77 \pm 6\%$	$67 \pm 9\%$	$10 \pm 10\%$
		parametric gain vs baseline	$80 \pm 9\%$	$71 \pm 10\%$	$9 \pm 9\%$
		parametric loss vs baseline	$75 \pm 10\%$	$64 \pm 8\%$	$11 \pm 10\%$
		task vs baseline	$93 \pm 4\%$	$88 \pm 6\%$	$5 \pm 7\%$
[33]	Plain or mirror-reversed text	junk	$58 \pm 6\%$	$55 \pm 5\%$	$3 \pm 8\%$
		mr non-switch vs baseline	$53 \pm 6\%$	$51 \pm 5\%$	$2 \pm 6\%$
		mr switch vs baseline	$50 \pm 3\%$	$50 \pm 2\%$	$0 \pm 3\%$
		mr vs plain	$52 \pm 7\%$	$55 \pm 9\%$	$-4 \pm 11\%$
		pl non-switch vs baseline	$55 \pm 6\%$	$60 \pm 8\%$	$-5 \pm 8\%$
		pl switch vs baseline	$50 \pm 5\%$	$50 \pm 5\%$	$-1 \pm 5\%$
		switch vs nonswitch	$51 \pm 4\%$	$51 \pm 6\%$	$-1 \pm 7\%$
		switch vs nonswitch mronly	$68 \pm 11\%$	$65 \pm 7\%$	$3 \pm 13\%$
		switch vs nonswitch plain-only	$61 \pm 8\%$	$59 \pm 7\%$	$2 \pm 12\%$
[34]	Stop-signal	failed stop vs baseline	$57 \pm 4\%$	$53 \pm 3\%$	$4 \pm 3\%$
		failed vs successful stop	$60 \pm 5\%$	$66 \pm 5\%$	$-6 \pm 5\%$
		go vs baseline	$61 \pm 3\%$	$65 \pm 4\%$	$-3 \pm 4\%$
		junk vs baseline	$52 \pm 3\%$	$51 \pm 2\%$	$1 \pm 4\%$
		successful stop vs baseline	$52 \pm 2\%$	$51 \pm 1\%$	$1 \pm 2\%$
		successful stop vs go	$62 \pm 3\%$	$66 \pm 3\%$	$-5 \pm 5\%$
[35]	Conditional stop-signal	failed stop critical vs baseline	$50 \pm 1\%$	$50 \pm 3\%$	$-1 \pm 3\%$
		failed stop non-critical vs baseline	$51 \pm 3\%$	$51 \pm 3\%$	$1 \pm 5\%$
		failed vs successful stop	$63 \pm 8\%$	$65 \pm 7\%$	$-2 \pm 11\%$
		go crital vs go non-crital	$50 \pm 4\%$	$49 \pm 5\%$	$0 \pm 6\%$
		go critical vs baseline	$52 \pm 5\%$	$50 \pm 1\%$	$3 \pm 5\%$
		go non-critical vs baseline	$53 \pm 6\%$	$55 \pm 6\%$	$-2 \pm 8\%$
		junk vs baseline	$57 \pm 6\%$	$54 \pm 3\%$	$3 \pm 7\%$
		successful stop critical vs baseline	$52 \pm 3\%$	$50 \pm 2\%$	$2 \pm 3\%$
		successful stop vs go	$69 \pm 5\%$	$74 \pm 6\%$	$-5 \pm 8\%$
	Stop-signal	failed stop vs baseline	$51 \pm 3\%$	$52 \pm 3\%$	$-1 \pm 4\%$
		failed vs successful stop	$63 \pm 8\%$	$65 \pm 7\%$	$-2 \pm 11\%$
		go vs baseline	$55 \pm 7\%$	$56 \pm 7\%$	$-0 \pm 7\%$
		junk vs baseline	$57 \pm 6\%$	$54 \pm 3\%$	$3 \pm 7\%$
		successful stop vs baseline	$53 \pm 4\%$	$50 \pm 3\%$	$3 \pm 5\%$
		successful stop vs go	$69 \pm 5\%$	$74 \pm 6\%$	$-5 \pm 8\%$
[36]	Balloon analog risk task	accept vs baseline	$75 \pm 7\%$	$74 \pm 6\%$	$1 \pm 7\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		accept vs reject	98 ± 3%	90 ± 6%	8 ± 6%
		explode vs baseline	85 ± 4%	78 ± 6%	6 ± 5%
		reject vs baseline	88 ± 6%	85 ± 7%	3 ± 7%
	Emotion regulation	attend negative vs attend neutral	81 ± 11%	69 ± 12%	12 ± 12%
		attend negative vs baseline	78 ± 7%	68 ± 11%	9 ± 9%
		attend neutral vs baseline	64 ± 10%	63 ± 9%	1 ± 10%
		junk rating vs baseline	57 ± 6%	52 ± 4%	4 ± 5%
		rating all vs baseline	71 ± 12%	67 ± 10%	4 ± 9%
		rating par vs baseline	52 ± 4%	51 ± 4%	1 ± 6%
		suppress negative vs attend negative	56 ± 6%	53 ± 5%	3 ± 7%
		suppress negative vs baseline	61 ± 8%	55 ± 7%	6 ± 7%
	Stop-signal	go vs baseline	85 ± 6%	83 ± 5%	1 ± 7%
		junk vs baseline	72 ± 7%	68 ± 6%	4 ± 6%
		successful stop vs baseline	83 ± 7%	74 ± 8%	9 ± 6%
		successful stop vs go	97 ± 2%	94 ± 5%	3 ± 4%
		unsuccessful stop vs baseline	80 ± 7%	74 ± 9%	6 ± 8%
		unsuccessul vs successful stop	83 ± 7%	76 ± 7%	7 ± 7%
	Temporal discounting task	easy all vs baseline	54 ± 8%	51 ± 7%	2 ± 5%
		easy par vs baseline	50 ± 0%	49 ± 1%	1 ± 1%
		hard all vs baseline	50 ± 0%	54 ± 7%	−4 ± 7%
		hard all vs easy all	50 ± 4%	58 ± 13%	−7 ± 14%
		hard par vs baseline	50 ± 0%	49 ± 6%	1 ± 6%
		junk vs baseline	72 ± 7%	68 ± 6%	4 ± 6%
[37]	Classification probe without feedback	correct dual task classification vs baseline	60 ± 8%	56 ± 7%	4 ± 7%
		correct single task classification vs baseline	59 ± 6%	61 ± 8%	−2 ± 8%
		correct single vs dual task classification	90 ± 3%	76 ± 9%	14 ± 9%
		junk dual task items vs baseline	58 ± 6%	53 ± 5%	5 ± 8%
		junk single task items vs baseline	63 ± 8%	57 ± 7%	6 ± 10%
	Dual-task weather classification	dual task classification vs baseline	84 ± 9%	70 ± 7%	14 ± 8%
		dual task classification vs probe	90 ± 7%	89 ± 7%	1 ± 9%
		dual task probe vs baseline	88 ± 7%	83 ± 7%	5 ± 8%
	Single-task weather classification	single task classification vs baseline	84 ± 9%	80 ± 12%	4 ± 9%
		single task classification vs probe	91 ± 6%	84 ± 8%	7 ± 9%
		single task probe vs baseline	86 ± 6%	73 ± 7%	12 ± 11%
	Tone-counting	tone counting probe vs baseline	88 ± 7%	83 ± 8%	5 ± 8%
		tone counting vs baseline	97 ± 4%	87 ± 8%	10 ± 7%
		tone counting vs probe	96 ± 5%	87 ± 9%	9 ± 7%
[38]	Classification learning	classification vs baseline	85 ± 7%	82 ± 7%	3 ± 7%

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
	Stop-signal	failed stop critical vs baseline	$68 \pm 7\%$	$72 \pm 8\%$	$-4 \pm 8\%$
		failed stop critical vs non-critical	$58 \pm 9\%$	$55 \pm 8\%$	$3 \pm 13\%$
		failed stop critical vs successful	$78 \pm 8\%$	$63 \pm 10\%$	$15 \pm 13\%$
		failed stop non-critical vs baseline	$66 \pm 10\%$	$59 \pm 7\%$	$7 \pm 11\%$
		go critical vs baseline	$63 \pm 13\%$	$65 \pm 11\%$	$-2 \pm 12\%$
		go critical vs non-critical	$62 \pm 10\%$	$56 \pm 6\%$	$5 \pm 10\%$
		go non-critical vs baseline	$84 \pm 9\%$	$79 \pm 17\%$	$5 \pm 16\%$
		junk vs baseline	$62 \pm 11\%$	$52 \pm 6\%$	$10 \pm 12\%$
		successful stop vs baseline	$65 \pm 8\%$	$56 \pm 8\%$	$9 \pm 8\%$
		successful stop vs go non-critical	$75 \pm 10\%$	$81 \pm 8\%$	$-6 \pm 12\%$
	Classification learning	classification vs baseline	$94 \pm 6\%$	$92 \pm 12\%$	$2 \pm 12\%$
	Stop-signal	failed stop critical vs baseline	$82 \pm 13\%$	$85 \pm 9\%$	$-3 \pm 12\%$
		failed stop critical vs non-critical	$84 \pm 6\%$	$71 \pm 11\%$	$13 \pm 10\%$
		failed stop critical vs successful	$91 \pm 10\%$	$64 \pm 9\%$	$27 \pm 11\%$
		failed stop non-critical vs baseline	$62 \pm 12\%$	$56 \pm 8\%$	$6 \pm 10\%$
		go critical vs baseline	$64 \pm 13\%$	$62 \pm 12\%$	$2 \pm 9\%$
		go critical vs non-critical	$78 \pm 14\%$	$64 \pm 10\%$	$14 \pm 14\%$
		go non-critical vs baseline	$80 \pm 8\%$	$77 \pm 10\%$	$3 \pm 7\%$
		junk vs baseline	$62 \pm 11\%$	$53 \pm 11\%$	$9 \pm 14\%$
		successful stop vs baseline	$64 \pm 13\%$	$57 \pm 9\%$	$7 \pm 20\%$
		successful stop vs go non-critical	$89 \pm 7\%$	$83 \pm 6\%$	$6 \pm 11\%$
[39]	Cross-language repetition priming	abstract vs concrete	$94 \pm 5\%$	$84 \pm 8\%$	$10 \pm 10\%$
		english english abstract novel vs baseline	$48 \pm 2\%$	$53 \pm 5\%$	$-4 \pm 5\%$
		english english abstract repeat vs baseline	$51 \pm 3\%$	$50 \pm 3\%$	$0 \pm 4\%$
		english english concrete novel vs baseline	$55 \pm 6\%$	$53 \pm 5\%$	$2 \pm 7\%$
		english english concrete repeat vs baseline	$71 \pm 11\%$	$67 \pm 9\%$	$4 \pm 6\%$
		english spanish abstract novel vs baseline	$61 \pm 9\%$	$59 \pm 8\%$	$2 \pm 10\%$
		english spanish abstract repeat vs baseline	$56 \pm 6\%$	$56 \pm 7\%$	$-0 \pm 5\%$
		english spanish concrete novel vs baseline	$52 \pm 4\%$	$57 \pm 7\%$	$-5 \pm 7\%$
		english spanish concrete repeat vs baseline	$54 \pm 4\%$	$55 \pm 7\%$	$-2 \pm 8\%$
		spanish english abstract novel vs baseline	$56 \pm 7\%$	$55 \pm 6\%$	$2 \pm 8\%$
		spanish english abstract repeat vs baseline	$53 \pm 5\%$	$52 \pm 5\%$	$0 \pm 5\%$
		spanish english concrete novel vs baseline	$52 \pm 5\%$	$53 \pm 6\%$	$-1 \pm 7\%$
		spanish english concrete repeat vs baseline	$54 \pm 6\%$	$52 \pm 3\%$	$2 \pm 6\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		spanish spanish abstract novel vs baseline	$58 \pm 6\%$	$56 \pm 8\%$	$2 \pm 6\%$
		spanish spanish abstract repeat vs baseline	$58 \pm 7\%$	$56 \pm 6\%$	$1 \pm 8\%$
		spanish spanish concrete novel vs baseline	$63 \pm 8\%$	$58 \pm 4\%$	$5 \pm 7\%$
		spanish spanish concrete repeat vs baseline	$64 \pm 9\%$	$63 \pm 7\%$	$1 \pm 9\%$
[40]	Classification learning	negative feedback vs baseline	$54 \pm 5\%$	$40 \pm 5\%$	$15 \pm 5\%$
		positive feedback vs baseline	$61 \pm 6\%$	$56 \pm 5\%$	$4 \pm 7\%$
		positive vs negative feedback	$81 \pm 5\%$	$57 \pm 4\%$	$24 \pm 6\%$
[41]	Simon task	congruent correct vs baseline	$59 \pm 8\%$	$59 \pm 7\%$	$0 \pm 7\%$
		congruent incorrect vs baseline	$49 \pm 10\%$	$46 \pm 5\%$	$3 \pm 7\%$
		incongruent correct vs baseline	$53 \pm 6\%$	$62 \pm 9\%$	$-10 \pm 12\%$
		incongruent incorrect vs baseline	$51 \pm 5\%$	$55 \pm 11\%$	$-4 \pm 12\%$
		incongruent vs congruent	$56 \pm 14\%$	$48 \pm 2\%$	$8 \pm 14\%$
		incongruent vs congruent correct	$49 \pm 7\%$	$47 \pm 2\%$	$2 \pm 8\%$
		incorrect vs correct	$56 \pm 8\%$	$52 \pm 6\%$	$4 \pm 9\%$
		incorrect vs correct incongruent	$55 \pm 8\%$	$62 \pm 8\%$	$-7 \pm 7\%$
[7]	Visual object recognition	bottle vs baseline	$60 \pm 13\%$	$55 \pm 9\%$	$5 \pm 11\%$
		cat vs baseline	$63 \pm 10\%$	$67 \pm 11\%$	$-3 \pm 10\%$
		chair vs baseline	$71 \pm 15\%$	$73 \pm 16\%$	$-2 \pm 17\%$
		chair vs scramble	$94 \pm 6\%$	$80 \pm 15\%$	$14 \pm 12\%$
		face vs baseline	$66 \pm 14\%$	$62 \pm 14\%$	$4 \pm 8\%$
		face vs house	$99 \pm 1\%$	$91 \pm 9\%$	$8 \pm 9\%$
		face vs scramble	$76 \pm 13\%$	$69 \pm 13\%$	$7 \pm 19\%$
		house vs baseline	$83 \pm 14\%$	$73 \pm 10\%$	$10 \pm 18\%$
		house vs face	$100 \pm 1\%$	$91 \pm 14\%$	$8 \pm 14\%$
		house vs scramble	$80 \pm 18\%$	$83 \pm 9\%$	$-3 \pm 19\%$
		scissors vs baseline	$66 \pm 12\%$	$71 \pm 8\%$	$-5 \pm 13\%$
		scramble vs baseline	$88 \pm 10\%$	$78 \pm 12\%$	$11 \pm 17\%$
		shoe vs baseline	$61 \pm 13\%$	$64 \pm 11\%$	$-3 \pm 8\%$
[42]	Word & object processing	consonant vs baseline	$94 \pm 3\%$	$90 \pm 2\%$	$3 \pm 3\%$
		objects vs baseline	$86 \pm 4\%$	$86 \pm 3\%$	$0 \pm 4\%$
		objects vs scrambled	$97 \pm 1\%$	$96 \pm 2\%$	$2 \pm 2\%$
		scramble vs baseline	$94 \pm 3\%$	$93 \pm 3\%$	$0 \pm 2\%$
		words vs baseline	$84 \pm 4\%$	$81 \pm 5\%$	$4 \pm 5\%$
		words vs consonants	$95 \pm 2\%$	$93 \pm 2\%$	$2 \pm 2\%$
[43]	Emotion regulation	look neg ant vs look neu ant	$52 \pm 3\%$	$51 \pm 3\%$	$1 \pm 3\%$
		look neg cue vs look neu cue	$56 \pm 3\%$	$61 \pm 5\%$	$-5 \pm 6\%$
		look neg rating vs look neu rating	$59 \pm 4\%$	$57 \pm 5\%$	$2 \pm 5\%$
		look neg stim vs look neu stim	$80 \pm 4\%$	$77 \pm 5\%$	$3 \pm 5\%$
		look negative ant vs baseline	$59 \pm 6\%$	$55 \pm 3\%$	$5 \pm 6\%$
		look negative cue vs baseline	$65 \pm 5\%$	$63 \pm 3\%$	$3 \pm 5\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		look negative rating vs baseline	$65 \pm 7\%$	$61 \pm 5\%$	$4 \pm 7\%$
		look negative stim vs baseline	$74 \pm 5\%$	$70 \pm 5\%$	$4 \pm 6\%$
		look neu ant vs look neg ant	$53 \pm 3\%$	$54 \pm 4\%$	$-0 \pm 4\%$
		look neu ant vs reapp neg ant	$54 \pm 4\%$	$53 \pm 4\%$	$1 \pm 3\%$
		look neu cue vs look neg cue	$57 \pm 5\%$	$60 \pm 6\%$	$-3 \pm 6\%$
		look neu rating vs look neg rating	$55 \pm 4\%$	$55 \pm 5\%$	$0 \pm 5\%$
		look neu rating vs reapp neg rating	$61 \pm 6\%$	$57 \pm 5\%$	$4 \pm 5\%$
		look neu stim vs look neg stim	$70 \pm 5\%$	$65 \pm 5\%$	$5 \pm 4\%$
		look neu stim vs reapp neg stim	$80 \pm 5\%$	$78 \pm 6\%$	$2 \pm 4\%$
		look neutral ant vs baseline	$57 \pm 5\%$	$56 \pm 4\%$	$1 \pm 5\%$
		look neutral cue vs baseline	$61 \pm 5\%$	$59 \pm 4\%$	$2 \pm 4\%$
		look neutral rating vs baseline	$70 \pm 5\%$	$65 \pm 5\%$	$6 \pm 6\%$
		look neutral stim vs baseline	$90 \pm 4\%$	$85 \pm 7\%$	$5 \pm 6\%$
		reapp neg ant vs look neu ant	$55 \pm 4\%$	$52 \pm 4\%$	$3 \pm 5\%$
		reapp neg cue vs look neg cue	$52 \pm 3\%$	$52 \pm 4\%$	$0 \pm 3\%$
		reapp neg stim vs look neg stim	$64 \pm 6\%$	$67 \pm 5\%$	$-3 \pm 5\%$
		reapp negative ant vs baseline	$66 \pm 5\%$	$65 \pm 3\%$	$0 \pm 5\%$
		reapp negative cue vs baseline	$60 \pm 6\%$	$59 \pm 5\%$	$2 \pm 5\%$
		reapp negative rating vs baseline	$71 \pm 6\%$	$68 \pm 6\%$	$2 \pm 7\%$
		reapp negative stim vs baseline	$80 \pm 7\%$	$76 \pm 7\%$	$4 \pm 8\%$
[44]	False belief	false belief question vs baseline	$76 \pm 5\%$	$71 \pm 6\%$	$5 \pm 5\%$
		false belief story vs baseline	$78 \pm 5\%$	$77 \pm 4\%$	$1 \pm 6\%$
		false photo question vs baseline	$82 \pm 6\%$	$72 \pm 5\%$	$9 \pm 5\%$
		false photo story vs baseline	$84 \pm 6\%$	$79 \pm 5\%$	$4 \pm 5\%$
		falsebelief vs falsepicture	$59 \pm 6\%$	$54 \pm 3\%$	$5 \pm 6\%$
		falsebeliefquestion vs falsepicturequestion	$71 \pm 6\%$	$68 \pm 6\%$	$4 \pm 5\%$
		falsebeliefstory vs falsepicturestory	$69 \pm 5\%$	$73 \pm 4\%$	$-4 \pm 4\%$
[45]	Incidental encoding	cue vs fixation	$51 \pm 4\%$	$50 \pm 0\%$	$2 \pm 4\%$
		high confidence hit object vs miss object	$50 \pm 2\%$	$50 \pm 3\%$	$0 \pm 4\%$
		invalid high confidence hit cue vs baseline	$50 \pm 2\%$	$51 \pm 5\%$	$-1 \pm 5\%$
		invalid high confidence hit object vs baseline	$49 \pm 2\%$	$50 \pm 4\%$	$-1 \pm 5\%$
		invalid high confidence hit object vs invalid miss object	$54 \pm 6\%$	$50 \pm 3\%$	$4 \pm 6\%$
		invalid low confidence hit cue vs baseline	$56 \pm 7\%$	$51 \pm 2\%$	$5 \pm 6\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		invalid low confidence hit object vs baseline	57 ± 5%	53 ± 3%	3 ± 4%
		invalid miss cue vs baseline	49 ± 2%	54 ± 4%	−5 ± 5%
		invalid miss object vs baseline	58 ± 6%	57 ± 7%	1 ± 11%
		invalid miss object vs invalid high confidence hit object	54 ± 6%	53 ± 7%	1 ± 6%
		invalid other greeble cue vs baseline	53 ± 6%	51 ± 3%	2 ± 5%
		invalid other greeble vs baseline	57 ± 8%	52 ± 5%	5 ± 9%
		invalidly vs validly cued objects	52 ± 4%	50 ± 2%	2 ± 5%
		miss object vs high confidence hit object	50 ± 2%	50 ± 1%	−0 ± 2%
		valid high confidence hit cue vs baseline	60 ± 5%	57 ± 6%	3 ± 7%
		valid high confidence hit object vs baseline	73 ± 7%	62 ± 8%	11 ± 9%
		valid high confidence hit object vs valid miss object	64 ± 6%	65 ± 9%	−1 ± 11%
		valid low confidence hit cue vs baseline	52 ± 3%	49 ± 1%	2 ± 3%
		valid low confidence hit object vs baseline	52 ± 4%	52 ± 5%	−1 ± 5%
		valid miss cue vs baseline	55 ± 5%	51 ± 3%	4 ± 6%
		valid miss object vs baseline	59 ± 6%	55 ± 5%	4 ± 7%
		valid other greeble cue vs baseline	70 ± 8%	56 ± 5%	13 ± 11%
		valid other greeble vs baseline	76 ± 10%	65 ± 5%	11 ± 8%
		valid other object cue vs baseline	53 ± 4%	50 ± 2%	2 ± 4%
		valid other object vs baseline	56 ± 5%	52 ± 5%	3 ± 6%
		valid valid miss object vs high confidence hit object	67 ± 9%	63 ± 9%	4 ± 9%
[46]	Covert verb generation	covert verb generation vs baseline	86 ± 8%	84 ± 8%	2 ± 6%
	Line bisection	correct bisection vs baseline	73 ± 12%	71 ± 12%	2 ± 16%
		incorrect bisection vs baseline	57 ± 9%	50 ± 5%	7 ± 9%
		no response control vs baseline	68 ± 13%	62 ± 10%	6 ± 15%
		no response task vs baseline	56 ± 9%	51 ± 5%	6 ± 11%
		response control vs baseline	71 ± 8%	68 ± 10%	4 ± 10%
	Motor	finger vs baseline	85 ± 8%	84 ± 8%	1 ± 6%
		foot vs baseline	84 ± 8%	82 ± 8%	2 ± 11%
		lips vs baseline	90 ± 6%	87 ± 9%	3 ± 9%
	Overt verb generation	overt verb generation vs baseline	69 ± 10%	62 ± 8%	6 ± 12%
	Overt word repetition	overt word repetition vs baseline	73 ± 14%	58 ± 7%	15 ± 15%
[47]	Auditory oddball	auditory oddball vs baseline	58 ± 7%	55 ± 5%	4 ± 9%
		auditory oddball vs standard	63 ± 7%	68 ± 10%	−5 ± 10%

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		auditory rt vs baseline	71 ± 6%	55 ± 4%	15 ± 7%
		auditory standard vs baseline	87 ± 8%	80 ± 6%	7 ± 7%
	Visual oddball	visual oddball vs baseline	60 ± 9%	50 ± 4%	10 ± 7%
		visual oddball vs standard	68 ± 9%	69 ± 8%	−1 ± 11%
		visual rt vs baseline	70 ± 9%	60 ± 7%	10 ± 11%
		visual standard vs baseline	82 ± 7%	74 ± 7%	8 ± 9%
[48]	Continuous house vs face	baseline	92 ± 5%	88 ± 6%	4 ± 5%
		continuous house face 100ms frequency vs baseline	58 ± 7%	64 ± 8%	−6 ± 8%
		continuous house face 1600ms frequency vs baseline	52 ± 5%	52 ± 5%	−0 ± 6%
		continuous house face 17ms frequency vs baseline	58 ± 6%	58 ± 7%	0 ± 5%
		continuous house face 200ms frequency vs baseline	67 ± 8%	58 ± 7%	9 ± 9%
		continuous house face 3200ms frequency vs baseline	53 ± 7%	52 ± 4%	1 ± 6%
		continuous house face 33ms frequency vs baseline	55 ± 6%	53 ± 6%	2 ± 8%
		continuous house face 400ms frequency vs baseline	51 ± 4%	51 ± 4%	0 ± 4%
		continuous house face 4800ms frequency vs baseline	60 ± 8%	60 ± 9%	1 ± 9%
		continuous house face 50ms frequency vs baseline	57 ± 8%	54 ± 5%	3 ± 6%
		continuous house face 800ms frequency vs baseline	54 ± 6%	51 ± 4%	3 ± 6%
		high vs low frequency	76 ± 6%	73 ± 7%	3 ± 5%
		low vs high frequency	85 ± 5%	78 ± 6%	7 ± 5%
	Discontinuous house (800ms) vs face	baseline	92 ± 5%	88 ± 6%	4 ± 5%
		discontinuous house face 800ms frequency 100ms duration vs baseline	52 ± 5%	50 ± 2%	2 ± 4%
		discontinuous house face 800ms frequency 33ms duration vs baseline	51 ± 3%	52 ± 4%	−1 ± 4%
		discontinuous house face 800ms frequency 400ms duration vs baseline	53 ± 7%	52 ± 4%	1 ± 7%
		discontinuous house face 800ms frequency 50ms duration vs baseline	54 ± 5%	54 ± 5%	−0 ± 6%
		discontinuous house face 800ms frequency 800ms duration vs baseline	53 ± 6%	53 ± 6%	1 ± 5%
		high vs low frequency	76 ± 6%	73 ± 7%	3 ± 5%
		low vs high frequency	85 ± 5%	78 ± 6%	7 ± 5%
		medium vs other frequency	65 ± 6%	65 ± 9%	−0 ± 8%

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
	Discoutinuous house (400ms) vs face	baseline	$92 \pm 5\%$	$88 \pm 6\%$	$4 \pm 5\%$
		discontinuous house face 400ms frequency 100ms duration vs baseline	$55 \pm 7\%$	$51 \pm 3\%$	$4 \pm 7\%$
		discontinuous house face 400ms frequency 200ms duration vs baseline	$51 \pm 5\%$	$56 \pm 7\%$	$-5 \pm 7\%$
		discontinuous house face 400ms frequency 33ms duration vs baseline	$50 \pm 2\%$	$53 \pm 4\%$	$-2 \pm 5\%$
		discontinuous house face 400ms frequency 400ms duration vs baseline	$58 \pm 5\%$	$66 \pm 9\%$	$-7 \pm 9\%$
		discontinuous house face 400ms frequency 50ms duration vs baseline	$57 \pm 7\%$	$55 \pm 6\%$	$2 \pm 7\%$
		high vs low frequency	$76 \pm 6\%$	$73 \pm 7\%$	$3 \pm 5\%$
		low vs high frequency	$85 \pm 5\%$	$78 \pm 6\%$	$7 \pm 5\%$
		medium vs other frequency	$65 \pm 6\%$	$65 \pm 9\%$	$-0 \pm 8\%$
	House vs face	baseline	$92 \pm 5\%$	$88 \pm 6\%$	$4 \pm 5\%$
		face vs baseline	$91 \pm 6\%$	$77 \pm 11\%$	$13 \pm 10\%$
		face vs house	$100 \pm 0\%$	$97 \pm 5\%$	$3 \pm 5\%$
		house vs baseline	$93 \pm 6\%$	$85 \pm 10\%$	$8 \pm 7\%$
		object vs baseline	$88 \pm 6\%$	$87 \pm 12\%$	$2 \pm 10\%$
		object vs scramble	$95 \pm 6\%$	$84 \pm 15\%$	$11 \pm 11\%$
		scramble vs baseline	$81 \pm 12\%$	$81 \pm 10\%$	$0 \pm 5\%$
	Continuous house vs face	baseline	$96 \pm 2\%$	$90 \pm 5\%$	$5 \pm 5\%$
		continuous house face 100ms frequency vs baseline	$57 \pm 4\%$	$58 \pm 5\%$	$-1 \pm 3\%$
		continuous house face 125ms frequency vs baseline	$58 \pm 5\%$	$56 \pm 4\%$	$2 \pm 4\%$
		continuous house face 150ms frequency vs baseline	$53 \pm 3\%$	$57 \pm 7\%$	$-4 \pm 6\%$
		continuous house face 175ms frequency vs baseline	$55 \pm 4\%$	$55 \pm 5\%$	$-0 \pm 3\%$
		continuous house face 200ms frequency vs baseline	$62 \pm 5\%$	$61 \pm 4\%$	$1 \pm 6\%$
		continuous house face 250ms frequency vs baseline	$54 \pm 4\%$	$56 \pm 6\%$	$-1 \pm 6\%$
		continuous house face 400ms frequency vs baseline	$54 \pm 4\%$	$58 \pm 6\%$	$-5 \pm 4\%$
		continuous house face 50ms frequency vs baseline	$55 \pm 4\%$	$56 \pm 5\%$	$-1 \pm 5\%$
		continuous house face 75ms frequency vs baseline	$56 \pm 5\%$	$57 \pm 5\%$	$-1 \pm 6\%$
		continuous house face 800ms frequency vs baseline	$60 \pm 5\%$	$62 \pm 6\%$	$-2 \pm 7\%$
		continuous house face high vs low frequency	$77 \pm 5\%$	$75 \pm 5\%$	$2 \pm 5\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		continuous house face low vs high frequency	$77 \pm 5\%$	$79 \pm 4\%$	$-2 \pm 7\%$
		hits vs baseline	$85 \pm 7\%$	$76 \pm 7\%$	$9 \pm 9\%$
		hits vs misses	$68 \pm 7\%$	$64 \pm 7\%$	$3 \pm 7\%$
		misses vs baseline	$69 \pm 6\%$	$64 \pm 8\%$	$5 \pm 7\%$
		misses vs hits	$72 \pm 9\%$	$70 \pm 7\%$	$2 \pm 9\%$
	House vs face	baseline	$96 \pm 2\%$	$90 \pm 5\%$	$5 \pm 5\%$
		face vs baseline	$98 \pm 3\%$	$88 \pm 9\%$	$10 \pm 8\%$
		face vs house	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
		house vs baseline	$100 \pm 0\%$	$96 \pm 4\%$	$4 \pm 4\%$
		object vs baseline	$98 \pm 3\%$	$89 \pm 7\%$	$10 \pm 6\%$
		object vs scramble	$100 \pm 0\%$	$97 \pm 5\%$	$3 \pm 5\%$
		scramble vs baseline	$100 \pm 0\%$	$93 \pm 7\%$	$7 \pm 7\%$
[49]	Emotion	faces	$98 \pm 1\%$	$99 \pm 0\%$	$-1 \pm 1\%$
		shapes	$98 \pm 0\%$	$99 \pm 0\%$	$-1 \pm 0\%$
	Gambling	punish	$88 \pm 2\%$	$92 \pm 1\%$	$-4 \pm 1\%$
		reward	$88 \pm 1\%$	$92 \pm 1\%$	$-4 \pm 1\%$
	Language	math	$99 \pm 0\%$	$99 \pm 0\%$	$0 \pm 0\%$
		story	$99 \pm 0\%$	$99 \pm 0\%$	$-0 \pm 0\%$
	Motor	cue	$100 \pm 0\%$	$100 \pm 0\%$	$0 \pm 0\%$
		lf	$100 \pm 0\%$	$100 \pm 0\%$	$-0 \pm 0\%$
		lh	$100 \pm 0\%$	$100 \pm 0\%$	$-0 \pm 0\%$
		rf	$99 \pm 0\%$	$100 \pm 0\%$	$-1 \pm 0\%$
		rh	$100 \pm 0\%$	$100 \pm 0\%$	$-0 \pm 0\%$
	Relational	match	$91 \pm 1\%$	$94 \pm 1\%$	$-3 \pm 1\%$
		rel	$92 \pm 1\%$	$94 \pm 1\%$	$-3 \pm 1\%$
	Social	random	$98 \pm 0\%$	$99 \pm 0\%$	$-1 \pm 0\%$
		tom	$99 \pm 0\%$	$99 \pm 0\%$	$-0 \pm 0\%$
	Wm	0bk body	$89 \pm 1\%$	$91 \pm 2\%$	$-2 \pm 1\%$
		0bk face	$94 \pm 1\%$	$94 \pm 1\%$	$-0 \pm 0\%$
		0bk place	$92 \pm 0\%$	$92 \pm 1\%$	$0 \pm 1\%$
		0bk tool	$90 \pm 0\%$	$93 \pm 1\%$	$-2 \pm 1\%$
		2bk body	$92 \pm 2\%$	$93 \pm 1\%$	$-1 \pm 1\%$
		2bk face	$95 \pm 1\%$	$96 \pm 1\%$	$-1 \pm 0\%$
		2bk place	$94 \pm 1\%$	$94 \pm 0\%$	$-0 \pm 0\%$
		2bk tool	$91 \pm 1\%$	$93 \pm 1\%$	$-2 \pm 1\%$
[50]	Face recognition	faces vs scrambled faces	$91 \pm 8\%$	$92 \pm 5\%$	$-1 \pm 8\%$
		famous faces vs baseline	$70 \pm 7\%$	$68 \pm 7\%$	$3 \pm 5\%$
		famous vs unfamiliar faces	$95 \pm 3\%$	$82 \pm 9\%$	$12 \pm 9\%$
		scrambled faces vs baseline	$83 \pm 9\%$	$78 \pm 10\%$	$5 \pm 8\%$
		unfamiliar faces vs baseline	$65 \pm 6\%$	$65 \pm 7\%$	$0 \pm 7\%$
[51]	Arithmetic	first operand non-symbolic addition vs baseline	$61 \pm 8\%$	$59 \pm 5\%$	$2 \pm 7\%$
		first operand non-symbolic color vs baseline	$65 \pm 8\%$	$58 \pm 7\%$	$6 \pm 7\%$
		first operand non-symbolic subtraction vs baseline	$64 \pm 8\%$	$57 \pm 6\%$	$8 \pm 9\%$
		first operand symbolic addition vs baseline	$64 \pm 7\%$	$60 \pm 5\%$	$4 \pm 6\%$
		first operand symbolic color vs baseline	$72 \pm 6\%$	$69 \pm 8\%$	$3 \pm 5\%$
		first operand symbolic subtraction vs baseline	$68 \pm 8\%$	$63 \pm 5\%$	$5 \pm 8\%$
		operator addition vs baseline	$60 \pm 5\%$	$59 \pm 5\%$	$1 \pm 7\%$
		operator color vs baseline	$71 \pm 8\%$	$72 \pm 8\%$	$-1 \pm 7\%$
		operator subtraction vs baseline	$66 \pm 7\%$	$57 \pm 6\%$	$9 \pm 8\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		response vs baseline	$100 \pm 0\%$	$93 \pm 6\%$	$7 \pm 6\%$
		second operand non-symbolic addition larger vs baseline	$56 \pm 4\%$	$55 \pm 3\%$	$1 \pm 5\%$
		second operand non-symbolic addition smaller vs baseline	$61 \pm 6\%$	$62 \pm 5\%$	$-1 \pm 7\%$
		second operand non-symbolic color larger vs baseline	$55 \pm 6\%$	$54 \pm 5\%$	$1 \pm 4\%$
		second operand non-symbolic color smaller vs baseline	$56 \pm 7\%$	$55 \pm 6\%$	$1 \pm 5\%$
		second operand non-symbolic subtraction larger vs baseline	$63 \pm 6\%$	$62 \pm 6\%$	$0 \pm 5\%$
		second operand non-symbolic subtraction smaller vs baseline	$56 \pm 6\%$	$56 \pm 3\%$	$-0 \pm 7\%$
		second operand symbolic addition larger vs baseline	$72 \pm 9\%$	$69 \pm 7\%$	$3 \pm 8\%$
		second operand symbolic addition smaller vs baseline	$60 \pm 8\%$	$55 \pm 6\%$	$5 \pm 7\%$
		second operand symbolic color larger vs baseline	$56 \pm 5\%$	$53 \pm 3\%$	$3 \pm 5\%$
		second operand symbolic color smaller vs baseline	$55 \pm 5\%$	$56 \pm 5\%$	$-0 \pm 6\%$
		second operand symbolic subtraction larger vs baseline	$71 \pm 11\%$	$67 \pm 10\%$	$5 \pm 7\%$
		second operand symbolic subtraction smaller vs baseline	$60 \pm 6\%$	$54 \pm 3\%$	$6 \pm 7\%$
	Saccades	left field vs baseline	$74 \pm 7\%$	$75 \pm 5\%$	$-1 \pm 5\%$
		left vs right field	$80 \pm 6\%$	$78 \pm 7\%$	$2 \pm 7\%$
		right field vs baseline	$77 \pm 6\%$	$72 \pm 9\%$	$5 \pm 8\%$
		right vs left field	$81 \pm 5\%$	$80 \pm 5\%$	$0 \pm 3\%$
[52]	Balloon analog risk task	baloon accept	$86 \pm 2\%$	$86 \pm 1\%$	$-0 \pm 1\%$
		baloon cashout	$90 \pm 1\%$	$92 \pm 1\%$	$-2 \pm 1\%$
		baloon explode	$91 \pm 1\%$	$94 \pm 2\%$	$-3 \pm 0\%$
		control accept	$86 \pm 1\%$	$86 \pm 2\%$	$0 \pm 2\%$
		control cashout	$61 \pm 3\%$	$63 \pm 3\%$	$-3 \pm 3\%$
	Breath hold task	hold ons	$80 \pm 2\%$	$82 \pm 2\%$	$-2 \pm 2\%$
		prep ons	$91 \pm 2\%$	$92 \pm 1\%$	$-1 \pm 1\%$
		rest ons	$86 \pm 2\%$	$85 \pm 2\%$	$1 \pm 1\%$
	Paired associates memory task (encoding)	control	$90 \pm 1\%$	$91 \pm 1\%$	$-1 \pm 1\%$
		task	$94 \pm 1\%$	$93 \pm 1\%$	$1 \pm 1\%$
	Paired associates memory task (retrieval)	control	$90 \pm 1\%$	$91 \pm 1\%$	$-1 \pm 1\%$
		correctly	$75 \pm 3\%$	$77 \pm 2\%$	$-3 \pm 3\%$
		incorrectly	$78 \pm 4\%$	$79 \pm 3\%$	$-1 \pm 3\%$
	SCAP working memory tasks	correct	$85 \pm 2\%$	$86 \pm 2\%$	$-1 \pm 2\%$
		incorrect	$81 \pm 2\%$	$81 \pm 3\%$	$0 \pm 3\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
		no response	$71 \pm 4\%$	$70 \pm 3\%$	$1 \pm 2\%$
	Stop-signal	blankscreen	$95 \pm 1\%$	$93 \pm 1\%$	$1 \pm 1\%$
		go left	$68 \pm 2\%$	$70 \pm 3\%$	$-2 \pm 1\%$
		go right	$73 \pm 2\%$	$74 \pm 2\%$	$-1 \pm 2\%$
		stop left	$71 \pm 2\%$	$74 \pm 2\%$	$-3 \pm 2\%$
		stop right	$75 \pm 3\%$	$75 \pm 3\%$	$-1 \pm 3\%$
	Task switching	noswitch color	$67 \pm 2\%$	$69 \pm 2\%$	$-2 \pm 2\%$
		noswitch shape	$67 \pm 2\%$	$71 \pm 2\%$	$-4 \pm 3\%$
		switch color	$68 \pm 3\%$	$68 \pm 2\%$	$-1 \pm 3\%$
		switch shape	$70 \pm 2\%$	$68 \pm 2\%$	$2 \pm 3\%$
[53]	Foreign language	french vs baseline	$94 \pm 3\%$	$92 \pm 3\%$	$3 \pm 2\%$
		french vs korean	$97 \pm 2\%$	$93 \pm 2\%$	$3 \pm 3\%$
		french vs sound	$87 \pm 4\%$	$79 \pm 4\%$	$8 \pm 3\%$
		korean vs baseline	$91 \pm 3\%$	$92 \pm 2\%$	$-0 \pm 3\%$
		korean vs sound	$93 \pm 3\%$	$91 \pm 4\%$	$2 \pm 4\%$
		language vs sound	$76 \pm 5\%$	$74 \pm 5\%$	$2 \pm 4\%$
		sound vs baseline	$94 \pm 5\%$	$96 \pm 4\%$	$-2 \pm 3\%$
		sound vs french	$97 \pm 1\%$	$95 \pm 3\%$	$2 \pm 3\%$
		sound vs korean	$98 \pm 2\%$	$98 \pm 1\%$	$-1 \pm 2\%$
	Localizer	action vs baseline	$96 \pm 2\%$	$97 \pm 2\%$	$-1 \pm 1\%$
		digit vs baseline	$91 \pm 4\%$	$91 \pm 4\%$	$-0 \pm 3\%$
		digit vs house	$93 \pm 3\%$	$92 \pm 3\%$	$1 \pm 5\%$
		digit vs scramble	$90 \pm 5\%$	$90 \pm 4\%$	$-0 \pm 3\%$
		digit vs words	$99 \pm 1\%$	$96 \pm 3\%$	$3 \pm 3\%$
		face vs baseline	$97 \pm 2\%$	$98 \pm 1\%$	$-1 \pm 2\%$
		face vs house	$99 \pm 1\%$	$98 \pm 2\%$	$1 \pm 2\%$
		face vs scramble	$98 \pm 2\%$	$97 \pm 3\%$	$1 \pm 3\%$
		house vs baseline	$97 \pm 2\%$	$97 \pm 3\%$	$1 \pm 2\%$
		house vs scramble	$100 \pm 0\%$	$99 \pm 2\%$	$1 \pm 2\%$
		scramble vs baseline	$98 \pm 1\%$	$97 \pm 2\%$	$2 \pm 2\%$
		tool vs baseline	$91 \pm 3\%$	$93 \pm 2\%$	$-2 \pm 3\%$
		tool vs house	$99 \pm 1\%$	$99 \pm 1\%$	$-0 \pm 1\%$
		tool vs scramble	$98 \pm 3\%$	$96 \pm 4\%$	$2 \pm 3\%$
		words vs baseline	$91 \pm 4\%$	$88 \pm 4\%$	$3 \pm 4\%$
		words vs digit	$97 \pm 2\%$	$97 \pm 3\%$	$0 \pm 3\%$
		words vs house	$91 \pm 3\%$	$91 \pm 3\%$	$-0 \pm 4\%$
		words vs scramble	$92 \pm 3\%$	$91 \pm 3\%$	$2 \pm 4\%$
	Saccade	calculation vs baseline	$95 \pm 2\%$	$96 \pm 1\%$	$-1 \pm 2\%$
		calculation vs next number	$97 \pm 2\%$	$96 \pm 3\%$	$1 \pm 3\%$
		calculation vs saccade	$96 \pm 3\%$	$94 \pm 4\%$	$2 \pm 2\%$
		next number vs baseline	$99 \pm 2\%$	$98 \pm 2\%$	$1 \pm 1\%$
		saccade vs baseline	$97 \pm 2\%$	$93 \pm 3\%$	$4 \pm 2\%$
		saccade vs calculation	$97 \pm 3\%$	$92 \pm 2\%$	$5 \pm 2\%$
		saccade vs next number	$95 \pm 2\%$	$91 \pm 3\%$	$3 \pm 3\%$
[54]	Auditory compression	auditory bottleneck vs language	$96 \pm 3\%$	$97 \pm 3\%$	$-0 \pm 3\%$
		auditory language vs bottleneck	$90 \pm 6\%$	$78 \pm 9\%$	$12 \pm 9\%$
		auditory sentences 100% duration vs baseline	$74 \pm 10\%$	$69 \pm 10\%$	$5 \pm 9\%$
		auditory sentences 20% duration vs baseline	$84 \pm 5\%$	$81 \pm 7\%$	$3 \pm 8\%$
		auditory sentences 40% duration vs baseline	$84 \pm 7\%$	$83 \pm 8\%$	$0 \pm 5\%$
		auditory sentences 60% duration vs baseline	$67 \pm 7\%$	$61 \pm 9\%$	$6 \pm 6\%$
		auditory sentences 80% duration vs baseline	$61 \pm 7\%$	$67 \pm 8\%$	$-6 \pm 6\%$

Continued on next page

Study	Task	Contrast	Multi-study B-acc	Voxel-level B-acc	B-acc gain
	Visual compression	visual bottleneck vs language	$99 \pm 2\%$	$98 \pm 3\%$	$1 \pm 4\%$
		visual language vs bottleneck	$93 \pm 5\%$	$78 \pm 6\%$	$15 \pm 6\%$
		visual sentences 100% duration vs baseline	$76 \pm 7\%$	$74 \pm 10\%$	$2 \pm 8\%$
		visual sentences 20% duration vs baseline	$76 \pm 8\%$	$64 \pm 7\%$	$12 \pm 10\%$
		visual sentences 40% duration vs baseline	$70 \pm 10\%$	$77 \pm 9\%$	$-7 \pm 10\%$
		visual sentences 60% duration vs baseline	$61 \pm 5\%$	$53 \pm 5\%$	$9 \pm 6\%$
		visual sentences 80% duration vs baseline	$69 \pm 8\%$	$62 \pm 8\%$	$7 \pm 7\%$

Table A. Accuracies per study in multi-study decoding.

Study	Chance level	Multi-task accuracy	Single-task accuracy	Accuracy gain
[22] High level math & Localizer	3%	$83 \pm 1\%$	$81 \pm 1\%$	$2 \pm 1\%$
[23] The ARCHI project	3%	$87 \pm 1\%$	$86 \pm 1\%$	$1 \pm 1\%$
[24] Brainomics	5%	$94 \pm 1\%$	$88 \pm 2\%$	$5 \pm 1\%$
[25] CamCAN	20%	$66 \pm 1\%$	$66 \pm 1\%$	$0 \pm 1\%$
[26, 27] Music structure & Sentence structure	5%	$48 \pm 2\%$	$45 \pm 2\%$	$3 \pm 2\%$
[28] Sentence/music complexity	4%	$38 \pm 3\%$	$34 \pm 3\%$	$4 \pm 3\%$
[29] Balloon Analog Risk-taking	8%	$59 \pm 6\%$	$46 \pm 7\%$	$13 \pm 4\%$
[30] Baseline trials & Classication learning	14%	$62 \pm 5\%$	$55 \pm 5\%$	$7 \pm 5\%$
[31] Rhyme judgment	33%	$65 \pm 8\%$	$53 \pm 11\%$	$12 \pm 10\%$
[32] Mixed-gambles	25%	$72 \pm 7\%$	$59 \pm 8\%$	$13 \pm 8\%$
[33] Plain or mirror-reversed text	11%	$20 \pm 5\%$	$20 \pm 3\%$	$-0 \pm 5\%$
[34] Stop-signal	17%	$29 \pm 3\%$	$31 \pm 3\%$	$-2 \pm 3\%$
[35] Conditional stop-signal & Stop-signal	8%	$23 \pm 5\%$	$23 \pm 4\%$	$-0 \pm 5\%$
[36] Balloon analog risk task & Emotion regulation & Stop-signal & Temporal discounting task	4%	$56 \pm 4\%$	$47 \pm 4\%$	$8 \pm 4\%$
[37] Classification probe without feedback & Dual-task weather classification & Single-task weather classification & Tone-counting	7%	$65 \pm 3\%$	$52 \pm 4\%$	$12 \pm 4\%$
[38] Classification learning & Stop-signal	9%	$45 \pm 6\%$	$37 \pm 6\%$	$8 \pm 8\%$
[38] Classification learning & Stop-signal	9%	$59 \pm 8\%$	$45 \pm 6\%$	$14 \pm 6\%$
[39] Cross-language repetition priming	6%	$22 \pm 3\%$	$20 \pm 3\%$	$2 \pm 3\%$
[40] Classification learning	33%	$54 \pm 4\%$	$35 \pm 5\%$	$19 \pm 5\%$
[41] Simon task	12%	$19 \pm 7\%$	$19 \pm 4\%$	$-1 \pm 8\%$
[7] Visual object recognition	8%	$59 \pm 7\%$	$52 \pm 4\%$	$7 \pm 8\%$
[42] Word & object processing	17%	$86 \pm 3\%$	$83 \pm 2\%$	$3 \pm 3\%$
[43] Emotion regulation	4%	$31 \pm 2\%$	$28 \pm 2\%$	$4 \pm 2\%$
[44] False belief	14%	$56 \pm 4\%$	$50 \pm 3\%$	$6 \pm 4\%$
[45] Incidental encoding	4%	$17 \pm 2\%$	$11 \pm 2\%$	$6 \pm 2\%$
[46] Covert verb generation & Line bisection & Motor & Overt verb generation & Overt word repetition	9%	$52 \pm 7\%$	$44 \pm 6\%$	$9 \pm 7\%$
[47] Auditory oddball & Visual oddball	12%	$47 \pm 6\%$	$36 \pm 5\%$	$11 \pm 6\%$
[48] Continuous house vs face & Discontinuous house (800ms) vs face & Discoutinuous house (400ms) vs face & House vs face	3%	$41 \pm 3\%$	$36 \pm 4\%$	$5 \pm 2\%$
[48] Continuous house vs face & House vs face	4%	$44 \pm 3\%$	$42 \pm 4\%$	$2 \pm 3\%$
[49] The Human Connectome Project	4%	$90 \pm 0\%$	$93 \pm 1\%$	$-2 \pm 0\%$
[50] Face recognition	20%	$69 \pm 6\%$	$63 \pm 6\%$	$6 \pm 7\%$
[51] Arithmetic & Saccades	4%	$38 \pm 2\%$	$34 \pm 3\%$	$5 \pm 3\%$
[52] UCLA LA5C consortium	4%	$62 \pm 1\%$	$64 \pm 1\%$	$-2 \pm 1\%$
[53] Foreign language & Localizer & Saccade	3%	$90 \pm 1\%$	$87 \pm 2\%$	$2 \pm 2\%$
[54] Auditory compression & Visual compression	7%	$61 \pm 3\%$	$53 \pm 4\%$	$8 \pm 3\%$