PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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, Palaiseau, France, **2** Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, France

* arthur.mensch@m4x.org

OPEN ACCESS

Citation: Mensch A, Mairal J, Thirion B, Varoquaux G (2021) Extracting representations of cognition across neuroimaging studies improves brain decoding. PLoS Comput Biol 17(5): e1008795. https://doi.org/10.1371/journal.pcbi.1008795

Editor: Daniele Marinazzo, Ghent University, BELGIUM

Received: March 23, 2020

Accepted: February 15, 2021

Published: May 3, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008795

Data Availability Statement: All data files and resulting atlases are available on the website cogspaces.github.io.

Funding: This project has received funding from the European Union's Horizon 2020 Framework

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.

Programme for Research and Innovation under grant agreement N785907 (Human Brain Project SGA2). Arthur Mensch was supported by a grant from the Labex DigiCosme (AMPHI project). Julien Mairal was supported by the ERC grant SOLARIS (N714381) and a grant from ANR (MACARON project ANR-14-CE23-0003-01). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 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 S1 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 (discussed in details in S1 Appendix).

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

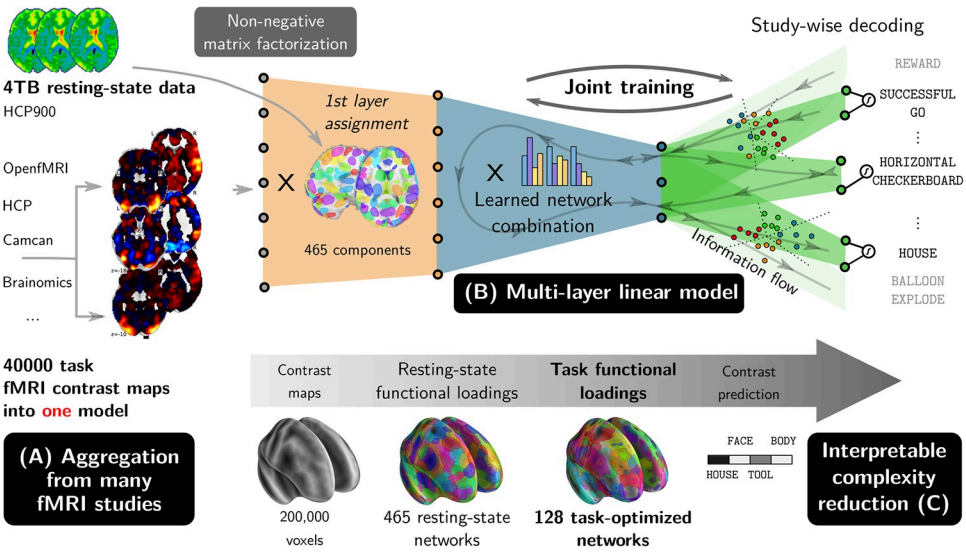

Fig 1. General description of our multi-study decoding approach. We perform inter-subject decoding using a shared three-layer model trained on multiple studies. An initial layer projects the input images from all studies onto functional networks learned on resting-state data. Then, a second layer combines 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.

https://doi.org/10.1371/journal.pcbi.1008795.g001

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.

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. Table C in S1 Appendix 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

https://doi.org/10.1371/journal.pcbi.1008795.t001

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 S1 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 S1 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 S1 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 S1 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 in S1 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 S1 Appendix, 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 neurovault.org. The published MSTON networks hold the representations extracted from the 35 studies that we have considered.

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 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 in S1 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;

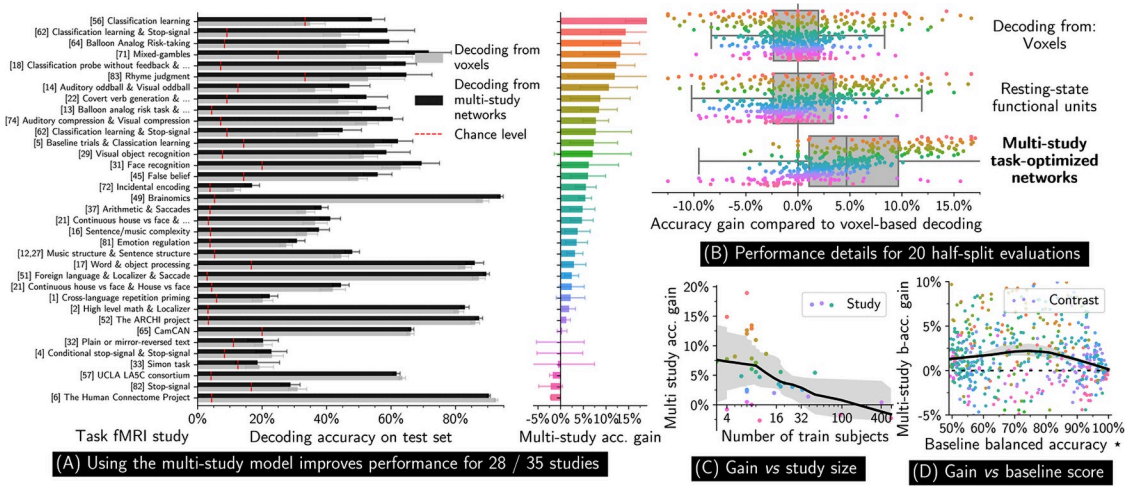

Fig 2. Quantitative performance of multi-study decoding. (A) Multi-study decoding improves the performance of cognitive task prediction across subjects for most studies. **(B)** Overall, decoding from task-optimized networks leads to a mean improvement accuracy of 5.8% compared to voxel or networks based 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 in S1 Appendix.

https://doi.org/10.1371/journal.pcbi.1008795.g002

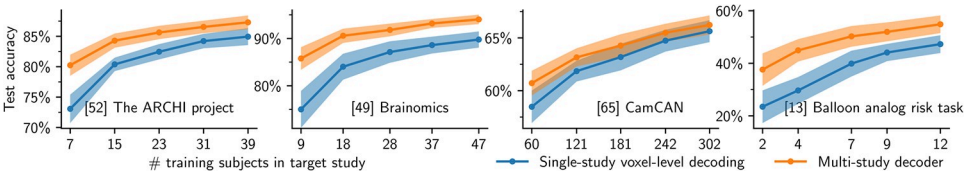

Fig 3. Varying accuracy improvement with study size. Training an MSTON decoder increases decoding accuracy for many studies (see Fig 2A). Gains are higher as we reduce the number of training subjects in target studies—pooling multiple studies is especially useful to decode studies performed on small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

https://doi.org/10.1371/journal.pcbi.1008795.g003

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 S1 Appendix), including those for which negative transfer was observed when using full cohorts.

Finally, we show in Fig B in S1 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 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

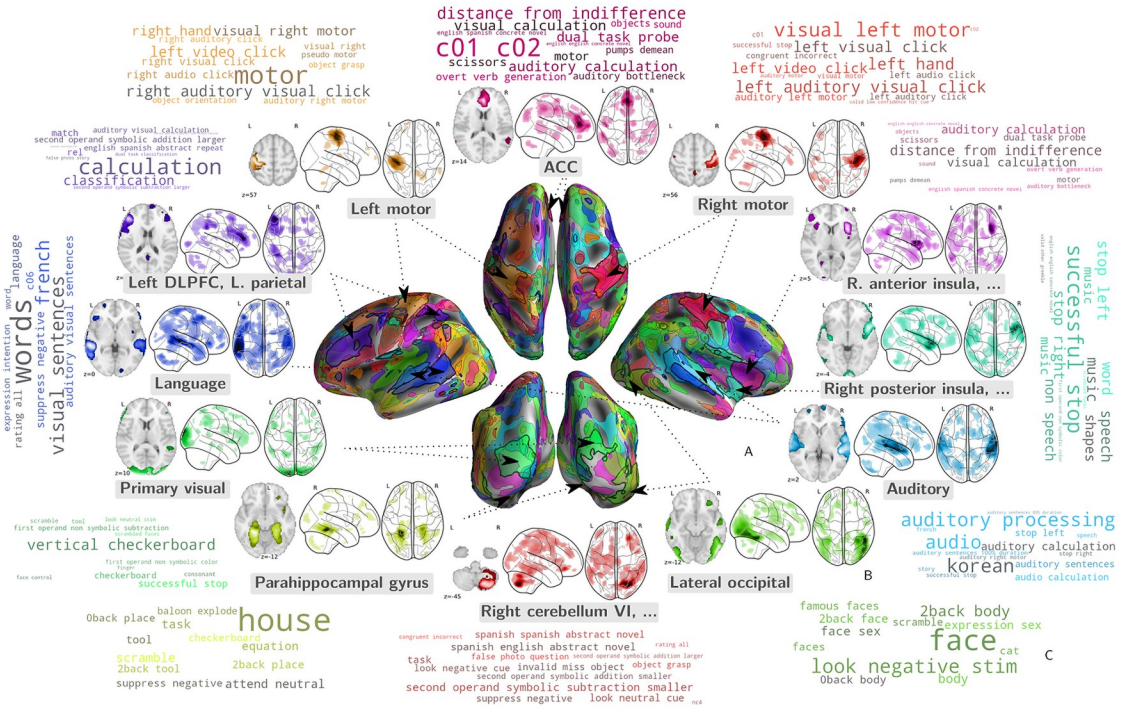

Fig 4. Visualization of some of task-optimized networks. Our approach learns networks that are important for decoding across studies. These networks are individually focal and collectively well spread across the cortex. They are readily associated with the cognitive tasks that they contribute to predict. 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.

https://doi.org/10.1371/journal.pcbi.1008795.g004

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 and 4C 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

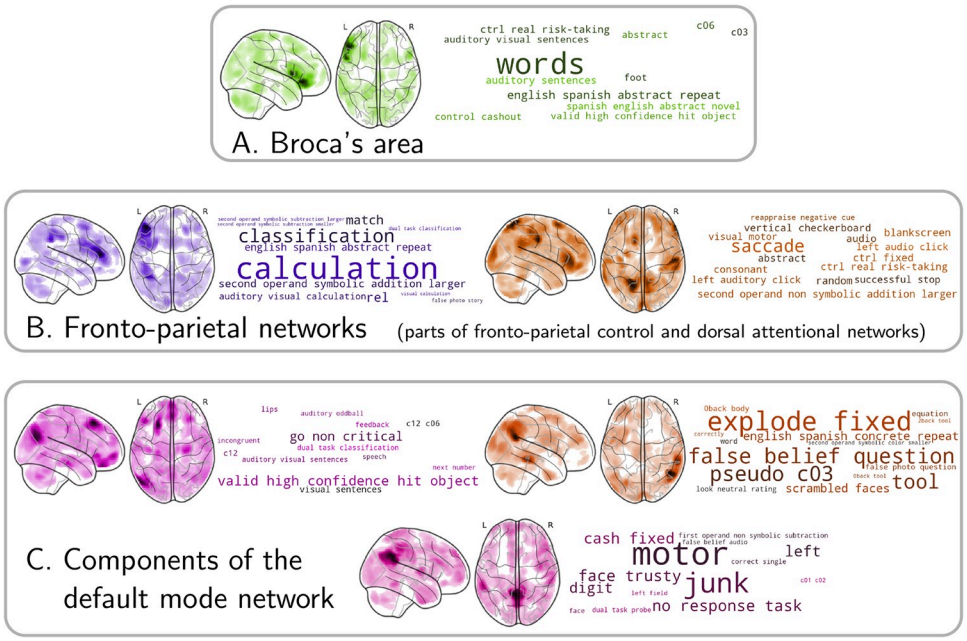

Fig 5. Task-optimized networks associated with high-level functions. Some MSTONs outline brain-circuits that are associated with language, e.g. Broca's area **(A)**, or more abstract functions, e.g. fronto-parietal networks **(B)** or even part of the default mode network **(C)**. Those networks are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

https://doi.org/10.1371/journal.pcbi.1008795.g005

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 in S1 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

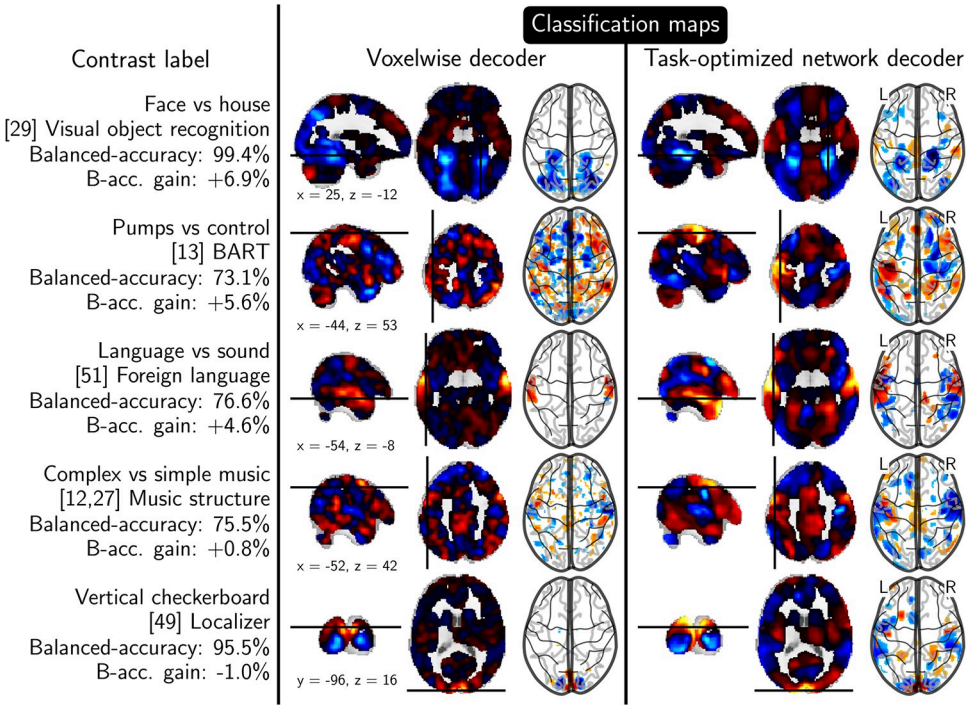

Fig 6. Classification maps obtained from multi-study decoding (right). The maps are smoother and more focused on functional modules than when decoding from voxels (left). For contrasts for which there is a performance boost (top of the figure), relevant brain regions are better delineated, as clearly 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).

https://doi.org/10.1371/journal.pcbi.1008795.g006

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 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 S1 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

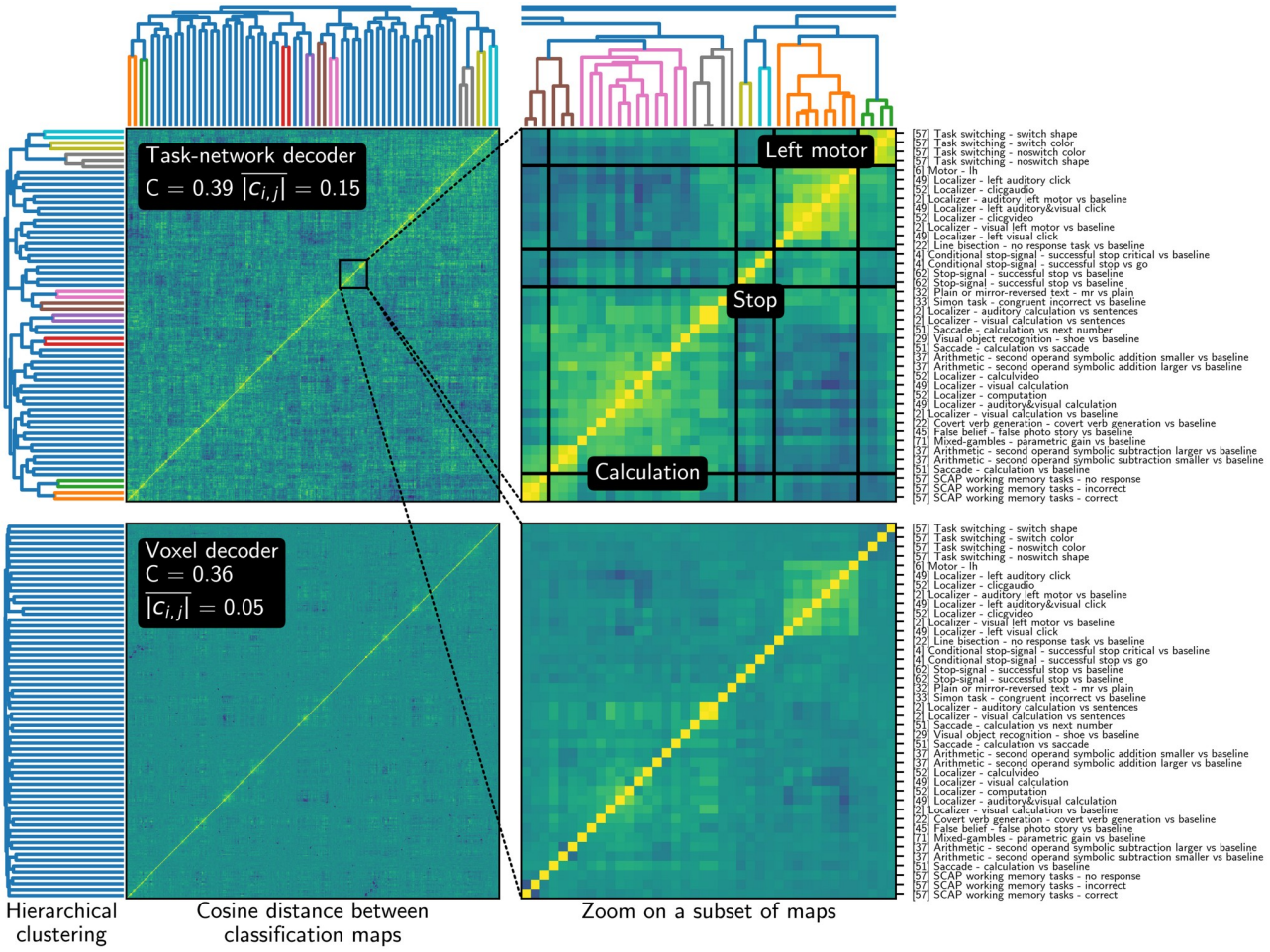

Fig 7. Cosine similarities between classification maps, obtained with our multi-study decoder (top) and with decoders learned separately (bottom), clustered using average-linkage hierarchical clustering. The classification maps obtained when decoding from task-optimized networks are more easily clustered 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.

https://doi.org/10.1371/journal.pcbi.1008795.g007

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 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 in S1 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 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 S1 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 S1 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.

Supporting information

S1 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. **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. **Reproduction details and tables**. In this appendix, we provide implementation details for reproducibility, along with tables with quantitative results per contrast.
(PDF)

S1 Components. This file holds a visualization of all the multi-study task optimized networks that we introduce in this paper.
(ZIP)

Author Contributions

Conceptualization: Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux.

Data curation: Arthur Mensch.

Formal analysis: Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux.

Funding acquisition: Bertrand Thirion, Gaël Varoquaux.

Investigation: Arthur Mensch.

Methodology: Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux.

Project administration: Bertrand Thirion, Gaël Varoquaux.

Resources: Arthur Mensch, Bertrand Thirion, Gaël Varoquaux.

Software: Arthur Mensch, Bertrand Thirion, Gaël Varoquaux.

Supervision: Julien Mairal, Bertrand Thirion, Gaël Varoquaux.

Validation: Arthur Mensch.

Visualization: Arthur Mensch, Gaël Varoquaux.

Writing – original draft: Arthur Mensch.

Writing – review & editing: Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux.

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. https://doi.org/10.1038/nmeth.1635 PMID: 21706013
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. https://doi.org/10.1038/nrn.2016.167 PMID: 28053326
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. https://doi.org/10.1038/nrn3475 PMID: 23571845
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. https://doi.org/10.1016/j.neuroimage.2012.02.018 PMID: 22366334
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. https://doi.org/10.1038/nn.4393 PMID: 27643430
6. Poldrack RA. Can cognitive processes be inferred from neuroimaging data? Trends in cognitive sciences. 2006; 10(2):59–63. https://doi.org/10.1016/j.tics.2005.12.004 PMID: 16406760

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. https://doi.org/10.1126/science.1063736 PMID: 11577229
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. https://doi.org/10.1111/j.1467-9280.2009.02460.x PMID: 19883493
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. https://doi.org/10.3389/fninf.2013.00012 PMID: 23847528
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. https://doi.org/10.3389/fninf.2015.00008 PMID: 25914639
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. https://doi.org/10.1016/j.conb.2018.11.002 PMID: 30513462
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. https://doi.org/10.1056/NEJMoa1204471 PMID: 23574118
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. https://doi.org/10.1371/journal.pcbi.1006565 PMID: 30496171
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. https://doi.org/10.1146/annurev-psych-122414-033729 PMID: 26393866
20. Barrett LF. The Future of Psychology: Connecting Mind to Brain. Perspectives on Psychological Science. 2009; 4(4):326–339. https://doi.org/10.1111/j.1745-6924.2009.01134.x PMID: 19844601
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. https://doi.org/10.1073/pnas.1603205113 PMID: 27071124
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. https://doi.org/10.1186/1471-2202-8-91 PMID: 17973998
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. https://doi.org/10.1016/j.neuroimage.2015.09.052 PMID: 26455807
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. https://doi.org/10.1186/s12883-014-0204-1 PMID: 25412575
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. https://doi.org/10.1016/S1053-8119(09)71443-0
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. https://doi.org/10.1162/jocn.2009.21070 PMID: 18702594

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. https://doi.org/10.3389/fnins.2012.00080 PMID: 22675289

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. https://doi.org/10.1016/j.neuroimage.2005.08.010 PMID: 16139527

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. https://doi.org/10.1162/jocn.2007.19.10.1643 PMID: 18271738

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. https://doi.org/10.1126/science.1134239 PMID: 17255512

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. https://doi.org/10.3389/fnhum.2014.00339 PMID: 24904378

34. Xue G, Aron AR, Poldrack RA. Common Neural Substrates for Inhibition of Spoken and Manual Responses. Cerebral Cortex. 2008; 18:1923–1932. https://doi.org/10.1093/cercor/bhm220 PMID: 18245044

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. https://doi.org/10.1523/JNEUROSCI.0519-07.2007 PMID: 17409238

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. https://doi.org/10.1073/pnas.0602659103 PMID: 16868087

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. https://doi.org/10.1038/35107080 PMID: 11734855

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. https://doi.org/10.1016/j.neuroimage.2009.03.014 PMID: 19289173

43. Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN. Prefrontal-Subcortical Pathways Mediating Successful Emotion Regulation. Neuron. 2008; 59:1037–1050. https://doi.org/10.1016/j.neuron.2008.09.006 PMID: 18817740

44. Moran JM, Jolly E, Mitchell JP. Social-Cognitive Deficits in Normal Aging. The Journal of Neuroscience. 2012; 32:5553–5561. https://doi.org/10.1523/JNEUROSCI.5511-11.2012 PMID: 22514317

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. https://doi.org/10.1523/JNEUROSCI.0152-11.2011 PMID: 21880922

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. https://doi.org/10.1186/2047-217X-2-6 PMID: 23628139

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. https://doi.org/10.1016/j.schres.2014.06.019 PMID: 25037525

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. https://doi.org/10.1523/JNEUROSCI.2467-12.2012 PMID: 23055513

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. https://doi.org/10.1016/j.neuroimage.2013.05.033 PMID: 23684877

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. https://doi.org/10.3389/fnhum.2011.00076 PMID: 21904527

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. https://doi.org/10.1126/science.1171599 PMID: 19423779

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. https://doi.org/10.1038/sdata.2016.110 PMID: 27922632

53. Pinel P, Dehaene S. Genetic and Environmental Contributions to Brain Activation During Calculation. NeuroImage. 2013; 81:306–316. https://doi.org/10.1016/j.neuroimage.2013.04.118 PMID: 23664947

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. https://doi.org/10.1523/JNEUROSCI.5685-11.2012 PMID: 22745508

55. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191

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. https://doi.org/10.1002/hbm.460020402

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. https://doi.org/10.1109/TSP.2017.2752697

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. https://doi.org/10.1016/j.neuroimage.2020.117126 PMID: 32673748

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. https://doi.org/10.1093/cercor/11.9.825 PMID: 11532888

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. https://doi.org/10.1176/jnp.23.2.jnp121 PMID: 21677237

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. https://doi.org/10.1073/pnas.012470999 PMID: 11756669

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. https://doi.org/10.1073/pnas.98.2.676

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. https://doi.org/10.1162/jocn.2009.21282 PMID: 19580387

68. Gratton C, Sun H, Petersen SE. Control networks and hubs. Psychophysiology. 2018; 55:e13032. https://doi.org/10.1111/psyp.13032

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. https://doi.org/10.1177/1073858411409051 PMID: 21636849

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. https://doi.org/10.1002/hbm.20813 PMID: 19507160

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. https://doi.org/10.1523/JNEUROSCI.5626-10.2011 PMID: 21368033

72. Varoquaux G. Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars. NeuroImage. 2018; 180:68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061 PMID: 28655633

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. https://doi.org/10.1007/978-3-642-22092-0_46 PMID: 21761686

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. https://doi.org/10.1016/j.neuroimage.2013.10.067 PMID: 24239590

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. https://doi.org/10.23915/distill.00010

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. https://doi.org/10.1002/(SICI)1097-0193(1998)6:3<160::AID-HBM5>3.0.CO;2-1 PMID: 9673671

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. https://doi.org/10.1016/j.mri.2006.09.042 PMID: 17222713

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. https://doi.org/10.1016/j.neuroimage.2013.11.046 PMID: 24389422

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. https://doi.org/10.1162/jocn_a_00077 PMID: 21671731

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. https://doi.org/10.1006/brln.1997.1906 PMID: 9593613

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. https://doi.org/10.1093/brain/awn279 PMID: 18971202

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. https://doi.org/10.1016/j.neuron.2011.08.026 PMID: 22017997

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. https://doi.org/10.1038/sdata.2016.44 PMID: 27326542