



**HAL**  
open science

# Extracting Universal Representations of Cognition across Brain-Imaging Studies

Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux

► **To cite this version:**

Arthur Mensch, Julien Mairal, Bertrand Thirion, Gaël Varoquaux. Extracting Universal Representations of Cognition across Brain-Imaging Studies. 2018. hal-01874713v1

**HAL Id: hal-01874713**

**<https://hal.science/hal-01874713v1>**

Preprint submitted on 14 Sep 2018 (v1), last revised 18 May 2021 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Extracting Universal Representations of Cognition across Brain-Imaging Studies

Arthur Mensch<sup>o\*</sup>, Julien Mairal<sup>•</sup>,  
Bertrand Thirion<sup>o</sup> and Gaël Varoquaux<sup>o</sup>

<sup>o</sup>Inria, CEA, Neurospin, Parietal team,  
Univ. Paris Saclay, 91120 Palaiseau, France

<sup>•</sup>Univ. Grenoble Alpes, Inria, CNRS,  
Grenoble INP, LJK, 38000 Grenoble, France

\*`arthur.mensch@m4x.org`, `first.last@inria.fr`

September 17, 2018

## Abstract

The size of publicly available data in cognitive neuro-imaging has increased a lot in recent years, thanks to strong research and community efforts. Exploiting this wealth of data demands new methods to turn the heterogeneous cognitive information held in different task-fMRI studies into common—universal—cognitive models. In this paper, we pool data from large fMRI repositories to predict psychological conditions from statistical brain maps across different studies and subjects. We leverage advances in deep learning, intermediate representations and multi-task learning to learn universal interpretable low-dimensional representations of brain images, usable for predicting psychological stimuli in all input studies. The method improves decoding performance for 80% of studies, by permitting cognitive information to flow from every study to the others: it notably gives a strong performance boost when decoding studies of small size. The trained low-dimensional representation—task-optimized networks—is interpretable as a set of basis cognitive dimensions relevant to meaningful categories of cognitive stimuli. Our approach opens new ways of extracting information from brain maps, overcoming the low power of typical fMRI studies.

## 1 Introduction

Cognitive neuro-imaging uses functional brain imaging to provide information on the brain structures underlying mental processes. The field is accumulating neural activity responses to specific psychological manipulations—tasks or

stimuli. The diversity of studies that probe different mental processes gives a big picture on cognition<sup>1</sup>. Yet, most individual studies suffer from a low statistical power<sup>2</sup>. To address this issue, large-scale projects have been collecting data from many subjects<sup>3,4</sup>. They must however focus on a small number of cognitive tasks due to practical aspects. In contrast, establishing *broad* models of cognition demands data from varied cognitive tasks<sup>5</sup>. In this paper, we pool data across many task-fMRI studies to increase both cognitive coverage and statistical power of decoding models. Critically, our approach bypasses the need of an overarching cognition theory to find commonalities between mental manipulations across studies.

Characterizing the functions of given brain structures requires to analyse brain responses across many cognitive paradigms. As pioneered by Poldrack *et al.*<sup>6</sup>, this functional selectivity can be obtained in a data-driven fashion by a *decoding* model that predicts mental processes from brain activity. In such a large-scale decoding setting, covering a broader set of cognitive paradigms gives a more precise functional description of each brain structure. Bearing this in mind, text-based meta-analyses from the literature capture a broad view of cognitive paradigms<sup>1</sup>, but are limited in their spatial resolution<sup>7</sup>. Open repositories of brain functional images hold the promise of very broad decoding directly at the resolution of the images<sup>8,9</sup>. This endeavor needs new methods to enable decoding across studies without an explicit correspondence in the mental processes that they manipulate. We address this by adapting tools from multi-task learning<sup>10,11</sup> and deep learning<sup>12,13</sup> to extract distributed brain structures which ground decoding across studies. We show that these structures provide universal priors of functional mapping and gather information across paradigms.

An important challenge is to build a model that generalizes in measurable ways to new cognitive paradigms. This is fundamentally difficult as each cognitive study frames a particular question and resorts to specific task oppositions that seldom have any exact counterpart in other studies<sup>14</sup>. In particular, the typical outcome of a cognitive fMRI study is a set of *contrast* brain maps, each of which corresponds to an elementary psychological manipulation, often unique to a given protocol. Working with pooled contrast maps requires to circumvent the undocumented nature of protocols' relationships. For this, labeling common aspects of psychological manipulations across studies has been proposed to build decoders that can describe aspects of unseen paradigms<sup>15,16</sup>. This annotation strategy is however difficult to scale to a large set of studies as it requires expert knowledge on each study. Current cognitive ontologies<sup>17</sup> that decompose psychological manipulations into mental process are also limited<sup>18</sup>.

We develop multi-study decoding models that rely on the original contrasts and their labels in each study. Instead of relabeling data into a common ontology, we let common cognitive dimensions be extracted from data. Our guiding hypothesis is that activation maps may be accurately decomposed into latent atomic components that form the neural building blocks underlying cognitive processes<sup>19</sup>. We capture such latent decomposition as the intermediate representations of a linear neural network. These *interpretable* representations are supported by trained spatial brain networks, associated with common aspects

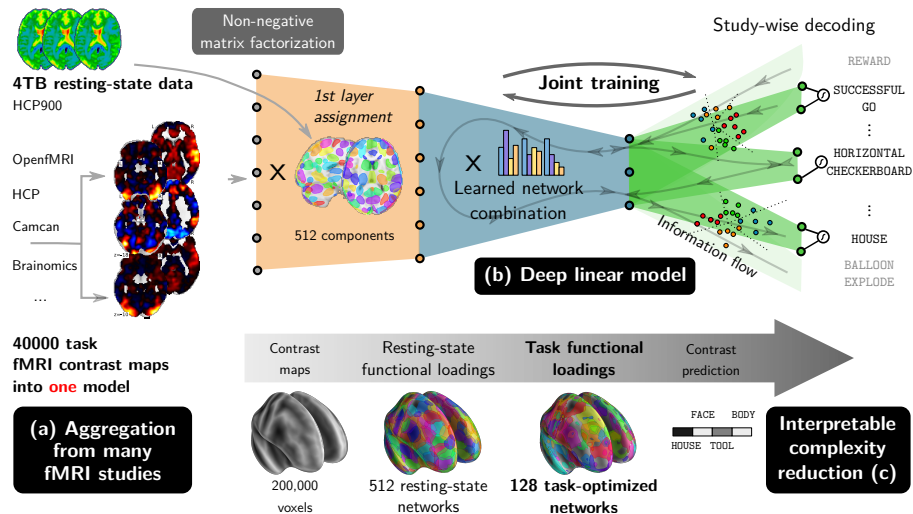


Figure 1: We perform inter-subject decoding using a shared three-layer model trained on multiple studies. An initial layer projects the input images from all studies onto functional networks learned on resting-state data. Then, a second layer combines the functional networks loadings into common meaningful 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.

of psychological manipulations across studies. Our approach overcomes the limitations of single-study cognitive subtraction models<sup>18</sup>. It *quantitatively* improves decoding performance for a vast majority of studies, which shows that the functional information acquired across many studies can help decoding unseen paradigms. In particular, the method gives a stronger boost in statistical power to studies with a small number of subjects.

## 2 Results

### 2.1 Method overview: a deep linear model

First, we briefly describe our approach to multi-study inter-subject decoding (the full technical description is available in [appendix A](#)). 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. Building upon the increasing availability of public task-fMRI data, we gather 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 analy-

sis, computing z-statistics maps for either base conditions, or for contrasts of interest when available. The 35 studies we use provide 40,000 contrast maps.

We cast inter-subject decoding as a machine-learning classification problem. Namely, we associate each brain activity contrast map to 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 512 functional units learned from resting-state data. The second layer performs dimension reduction and outputs an embedding of the brain activity into 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<sup>20</sup>.

With this layered model, the study-specific decoding is performed on a shared universal low-dimensional brain representation. This representation is made of the product of the second layer with the first layer: a linear combination of functional modules identified from resting state. The second layer and the third layer are jointly extracted from the task-fMRI data using regularized stochastic optimization<sup>21,22</sup>—the shared brain representation is optimized simultaneously with the third layer that provides decoding for every study. As we will show, projecting data onto this representation improves across-subject predictive accuracy, removing confounds while preserving the cognitive signal.

Finally, we perform a linear transformation of the second and third layers to express the corresponding dimension reduction on an interpretable small set of stable functional modules (Fig. 1c). For this, we use non-negative sparse decompositions<sup>23</sup> of an ensemble of models trained with different stochastic seeds. Together, the first two layers project input data onto *multi-study task-optimized networks* (MSTONs). These networks capture a general multi-study representation of the cognitive signal contained in statistical maps.

## 2.2 Data and performance metrics

We apply our method on a set of 35 publicly available task-fMRI studies, listed in Table 1; a few are acquired on 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 of those recruit related aspects of cognition (e.g., motor, attention, judgement tasks, object recognition). We measure decoding *accuracy* on left-out subjects for each study, and 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. Details are reported in appendix C.

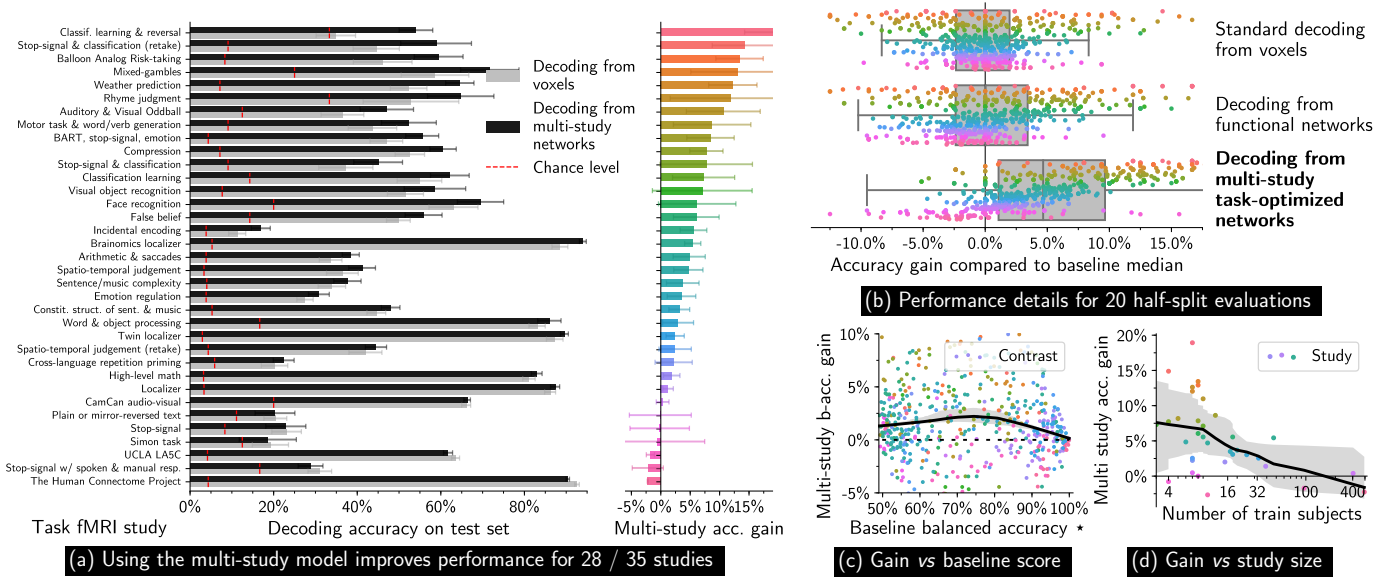


Figure 2: Performing joint training improves the performance of inter-subject decoding for most studies (a). Overall, decoding from task-optimized networks leads to a mean improvement accuracy of 5.8%; improvement is skewed across studies (b). Studies of typical size strongly benefit (d) from transfer learning, whereas little information is gained for larger or easier to decode studies (c). Error bars calculated over 20 random data half-split. \*(c) shows *per-contrast* balanced accuracy (50% chance level), whereas *per-study* classification accuracy is used everywhere else.

### 2.3 Improved decoding performance

Fig. 2 summarizes our quantitative results. For 28 out of the 35 task-fMRI studies that we consider, our MSTON-based decoder brings a significant improvement in prediction accuracy (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). By minimizing a *joint* objective that combines training losses from every study, we extract a second-layer representation that is efficient for many study-specific decoding tasks; the second layer parameters therefore incorporate information from all studies; the joint objective further permits information transfer among the many classification heads of the third layer—predictive accuracy is improved thanks to *transfer learning*. Although we have not specified how experiments are related from a cognitive point of view, our quantitative results show that these relations can be learned during training to improve decoding performance.

Studies with diverse cognitive focus benefit from using multi-study training. Among the highest accuracy gains, we find cognitive control (stop-signal),

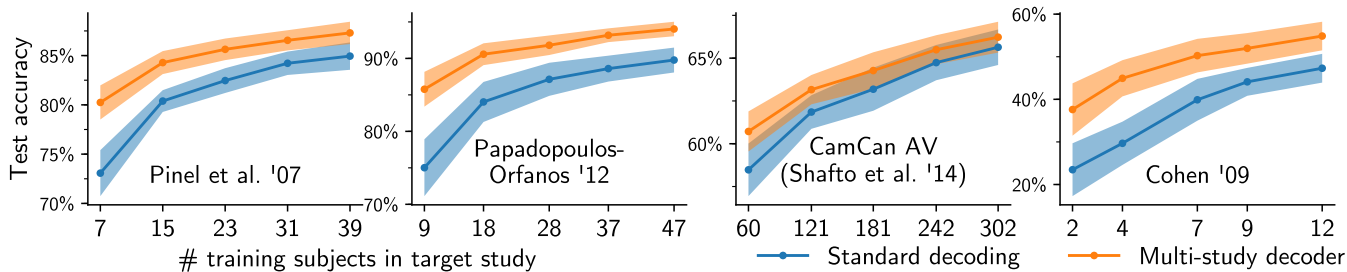


Figure 3: Training a MSTON decoder increases decoding accuracy for many studies (see Fig. 2a). Gains are higher as we reduce the number of training subjects in target studies—pooling multiple studies is especially useful to decode studies performed on small cohorts. Error bars calculated over 20 random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

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, 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 prevents 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. Appending a further *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 training, whereas classes whose balanced-accuracy is around 80% experience the highest decoding improvement (Fig. 2c). Fig. 2d shows that the benefit of multi-study training is higher for smaller studies, confirming that our method boosts their inter-subject decoding performance.

On Fig. 3, we vary the number of training subjects in target studies, and compare the performance of different decoders. We observe that the smaller the training size, the larger the performance gain imputable to multi-study training. Transfer learning in inter-subject decoding is thus more efficient for small cohort studies (e.g., 16 subjects), that still constitute the essential of publicly available task-fMRI data.



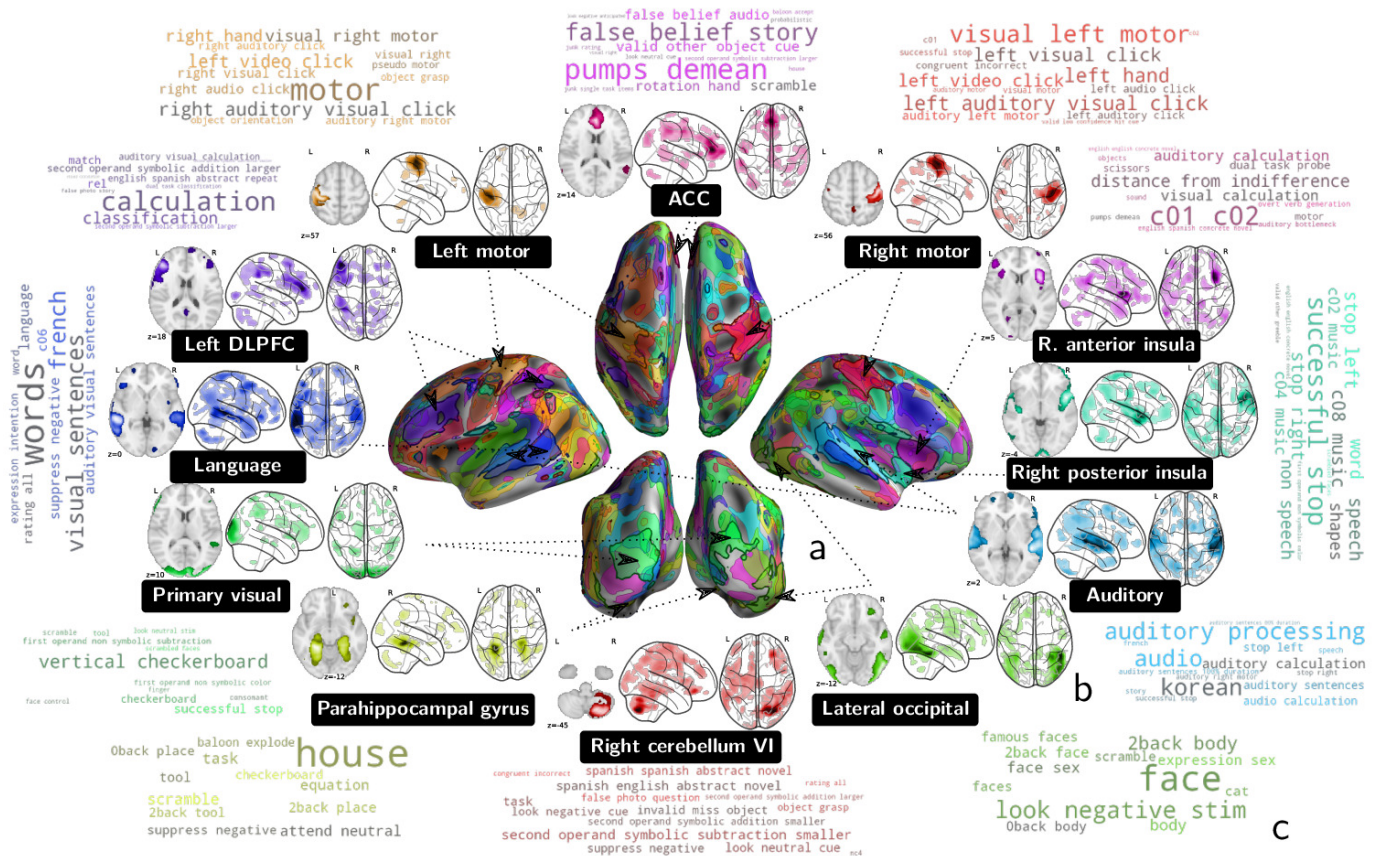


Figure 4: Visualization of some of task-optimized networks. Our approach allows to learn networks that are important for inter-subject decoding across studies. These networks, individually focal and collectively well spread across the cortex, are readily associated with the cognitive tasks that they contribute to predict. We display a selection of these networks, named with the salient anatomical brain region they recruit, along with a word-cloud representation of the stimuli whose likelihood increases with the network activation.

## 2.4 Extracted multi-study networks

The second and third layers of our model identify a *subspace* of the brain images onto which projecting helps decoding. These subspaces prove remarkably stable across runs (see section A.6). Performing non-negative matrix factorization over the parameters of the second layer across multiple runs finds interpretable directions in a consensus subspace. In voxel space, these directions form multi-study task-optimized networks (MSTONs), which constitute the extracted low-dimensional representation of input contrasts.



We outline the contours of the extracted MSTONs in Fig. 4a. The networks almost cover the entire cortex, an expected 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, implemented in very diverse brain localizations. Brain regions that are systematically recruited in task-fMRI protocols, e.g., motor cortex, auditory cortex, and primary visual cortex, are over-segmented by MSTON, i.e., they appear in several different networks. As capturing information in these regions is crucial for decoding many contrasts in our corpus, the model dedicates a large part of its representation capability to it. As decoding requires comparing distributed activation, MSTON are non-connected networks, as outlined in Fig. 4b. For instance, both parahippocampal gyri appear together in the yellow network.

Most importantly, Fig. 4b-c show that extracted MSTON capture cognitive information. Every network plays a significant role to classify a small subset of contrasts. We represent with word-clouds these contrasts’ names, as specified in the original studies (Fig. 4c). MSTON networks capture both low-level and high-level cognitive signal. At a lower level, it identifies 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, it identifies the left DLPFC and the IPS in a single network, used to decode tasks related to calculation and comparison. It successfully delineates the language network and the right posterior insula, important in decoding tasks involving music<sup>24</sup>. Several networks found involve regions of the brains recruited by wide range of tasks, such as the anterior insula and the ACC, a part of the salience network.

## 2.5 Impact of multi-study training on classification maps

To better understand how multi-study training and layered architecture improve decoding performance, we compare classification maps obtained using our model to standard decoder maps in Fig. 5. Those are simple to obtain, as our model remains a linear classifier from voxels to classes. 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<sup>25</sup>. 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 effects on contrast maps, that is already at play when projecting onto simple resting-state functional networks. Second, using multi-study task-optimized networks contributes to finding sharper 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. Our model renders them more distributed, an unfortunate consequence of multi-study training.

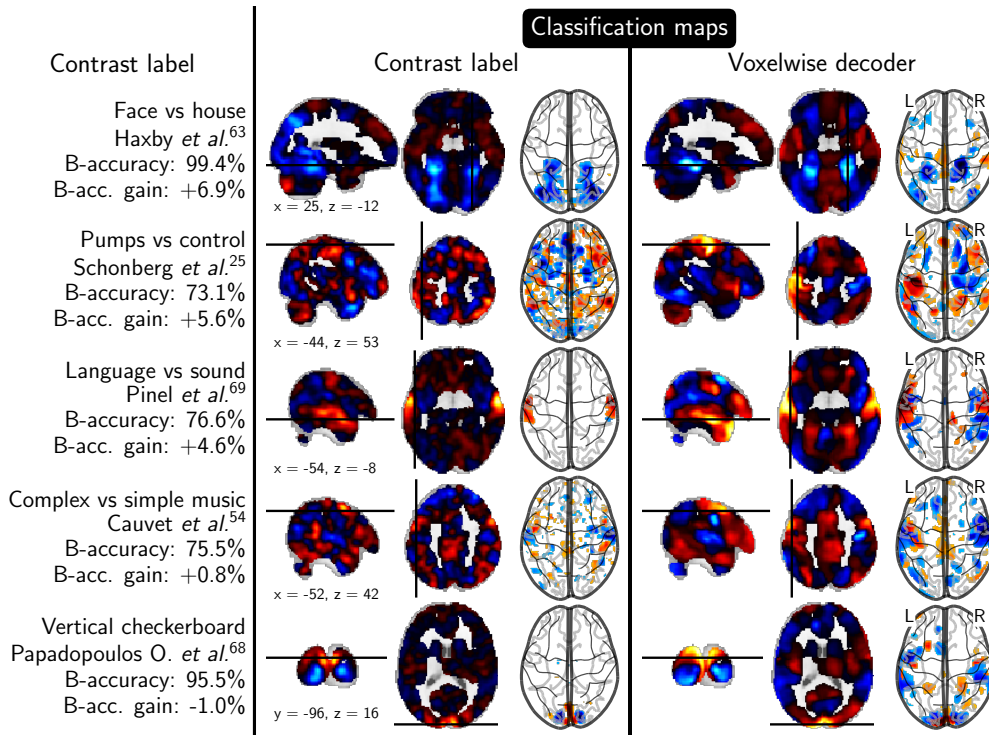


Figure 5: Classification maps obtained from multi-study training of decoding models are smoother and more focal than when decoding from voxels. Relevant brain regions are often better underlined.

In Supplementary 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 structure is a consequence of information transfer between decoding tasks, and partly explains the increase in decoding accuracy. In a dual perspective, we can plot the transformation of input contrast maps defined by the projection on task-optimized networks (Supplementary Fig. 8). Projected data are more focal, i.e., spatial variations that are unlikely to be related to cognition are smoothed. This offers a new angle for understanding quantitative results (Fig. 2), as this smoothing allows decoders to generalize better across subjects than when classifying raw input directly.

### 3 Discussion

Our results outline the power of training multi-layer linear models to decode brain functional images across studies. Such a strategy brings many benefits.

First, in practice, our approach provides a universal way to improve the accuracy of decoding in a newly acquired dataset. Many task-fMRI experiments are performed on groups of less than 30 subjects. In this regime, it is highly likely that decoding performance improves when aggregating existing studies to the new one using a multi-study model as proposed. As the repositories of publicly available data are progressively getting normalized and accessible, multi-study training provides an easy-to-deploy upgrade over simple decoders. We validated our model in a quantitative way, measuring predictive accuracy on left-out subjects. Improvements are also qualitative, as the interpretation of decoding maps is made easier (Fig. 5). Pooling subjects across studies effectively increases the training size of our model, as advocated by Poldrack *et al.*<sup>26</sup>. This increases the statistical power of cognitive modelling, and brings an answer to the reproducibility challenge outlined by Button *et al.*<sup>2</sup>.

Our design choices were driven by the recent successes of deep non-linear models in computer vision and medical imaging. We were not able to increase performance by departing from linear models: introducing non linearities in our models leads to a drastic increase of overfitting and provides no improvement on left-out accuracy. On the other hand, we have shown that pooling many fMRI data sources allows to learn deeper models, although those remain linear. Sticking to linear models has the further advantage of allowing easy interpretation of decoding models. Techniques issued from the deep learning communities prove very useful to train models that generalize well across subjects: using dropout regularization<sup>21</sup> and advanced stochastic gradient techniques<sup>22</sup> proves essential for successful transfer learning and good generalization performance.

Departing from the traditional convex models used for brain decoding makes model interpretation challenging. For this reason, inspired by computer vision work<sup>27</sup>, we transformed trained models (see section A.6) to uncover interpretable cognitive networks that capture information relevant for many decoding tasks. The predictive performance of these networks (Fig. 2 and Supplementary Fig. 6) provides quantitative support for the assumption that the human brain is structured in various basic networks. While extracting a universal basis of cognition is beyond the scope of any single fMRI study, training a joint predictive model on multiple studies finds meaningful approximations of atomic cognitive functions (Fig. 4). This is a step forward to define cognitive processes in a quantitative manner, which remains a fundamental challenge in psychology<sup>8,28</sup>.

Multi-study decoding provides a path towards knowledge consolidation in functional neuro-imaging: our multi-layer model can be further improved by increasing the size of the training corpus. Gathering more task-fMRI data using systematic pipelines should help outlining better task-optimized networks, as they will have to extract the signal needed for decoding more diverse tasks. Increasing standardization<sup>29</sup> and data sharing<sup>26</sup> in neuroimaging will facilitate aggregating bigger corpora. To that end, we have released easy-to-use decoding pipelines and pre-trained models at the address <http://cogspaces.github.io>.

## Acknowledgments

This project has received funding from the European Union’s Horizon 2020 Framework Programme for Research and Innovation under grant agreement N° 785907 (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 (N° 714381) and a grant from ANR (MACARON project ANR-14-CE23-0003-01).

## References

1. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-Scale Automated Synthesis of Human Functional Neuroimaging Data. *Nature Methods* **8**, 665–670 (2011).
2. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. & Munafò, M. R. Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience* **14**, 365–376 (2013).
3. Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M. & Curtiss, S. W. The Human Connectome Project: A Data Acquisition Perspective. *NeuroImage* **62**, 2222–2231 (2012).
4. Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., *et al.* Multimodal Population Brain Imaging in the UK Biobank Prospective Epidemiological Study. *Nature Neuroscience* **19**, 1523–1536 (2016).
5. Gordon, E. M. *et al.* Precision Functional Mapping of Individual Human Brains. *Neuron* **95**, 791–807.
6. Poldrack, R. A., Halchenko, Y. O. & Hanson, S. J. Decoding the Large-Scale Structure of Brain Function by Classifying Mental States Across Individuals. *Psychological Science* **20**, 1364–1372 (2009).
7. Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D. & Nichols, T. E. Meta-Analysis of Neuroimaging Data: A Comparison of Image-Based and Coordinate-Based Pooling of Studies. *NeuroImage* **45**, 810–823 (2009).
8. Poldrack, R. A., Barch, D. M., Mitchell, J., Wager, T. D., Wagner, A. D., Devlin, J. T., Cumba, C., Koyejo, O. & Milham, M. Toward Open Sharing of Task-Based fMRI Data: The OpenfMRI Project. *Frontiers in Neuroinformatics* **7**, 12 (2013).

9. Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat, V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., *et al.* NeuroVault.org: A Web-Based Repository for Collecting and Sharing Unthresholded Statistical Maps of the Human Brain. *Frontiers in Neuroinformatics* **9**, 8 (2015).
10. Ando, R. K. & Zhang, T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Research* **6**, 1817–1853 (2005).
11. Xue, Y., Liao, X., Carin, L. & Krishnapuram, B. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research* **8**, 35–63 (2007).
12. LeCun, Y., Bengio, Y. & Hinton, G. Deep Learning. *Nature* **521**, 436–444 (2015).
13. Collobert, R. & Weston, J. *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning* in *Proceeding of the International Conference on Machine Learning* (2008), 160–167.
14. Newell, A. You Can’t Play 20 Questions with Nature and Win: Projective Comments on the Papers of This Symposium. *Visual Information Processing*, 1–26 (1973).
15. Schwartz, Y., Thirion, B. & Varoquaux, G. *Mapping Paradigm Ontologies to and from the Brain* in *Advances in Neural Information Processing Systems* (2013), 1673–1681.
16. Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W. & Kross, E. An fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of Medicine* **368**, 1388–1397 (2013).
17. Turner, J. A. & Laird, A. R. The Cognitive Paradigm Ontology: Design and Application. *Neuroinformatics* **10**, 57–66 (2012).
18. Poldrack, R. A. & Yarkoni, T. From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure. *Annual Review of Psychology* **67**, 587–612 (2016).
19. Barrett, L. F. The Future of Psychology: Connecting Mind to Brain. *Perspectives on Psychological Science* **4**, 326–339 (2009).
20. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345–1359 (2010).
21. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).
22. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization* in *International Conference for Learning Representations* (2015).
23. Kim, H. & Park, H. Sparse Non-Negative Matrix Factorizations via Alternating Non-Negativity-Constrained Least Squares for Microarray Data Analysis. *Bioinformatics* **23**, 1495–1502 (2007).

24. Hara, N., Cauvet, E., Devauchelle, A. D., Dehaene, S., Pallier, C., *et al.* Neural Correlates of Constituent Structure in Language and Music. *Neuro-Image* **47**, S143 (2009).
25. Schonberg, T., Fox, C., Mumford, J. A., Congdon, C., Trepel, C. & Poldrack, R. A. Decreasing Ventromedial Prefrontal Cortex Activity During Sequential Risk-Taking: An fMRI Investigation of the Balloon Analog Risk Task. *Frontiers in Neuroscience* **6**, 80 (2012).
26. Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E. & Yarkoni, T. Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research. *Nature Reviews Neuroscience* **18**, 115–126 (2017).
27. Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K. & Mordvintsev, A. The Building Blocks of Interpretability. *Distill* **3**, e10 (2018).
28. Uttal, W. R. *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. (The MIT press, 2001).
29. Gorgolewski, K. J. *et al.* The Brain Imaging Data Structure, a Format for Organizing and Describing Outputs of Neuroimaging Experiments. *Scientific Data* **3**, sdata201644 (2016).



# 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.

|   |           |
|---|-----------|
| <b>A Detailed methods</b>   | <b>15</b> |
| A.1 Inter-subject decoding setting . . . . .                        | 15        |
| A.2 Baseline voxel-space decoder . . . . .                          | 15        |
| A.3 Baseline dimension reduced decoder . . . . .                    | 16        |
| A.4 Three-layer model description . . . . .                         | 16        |
| A.5 Resting-state data . . . . .                                    | 18        |
| A.6 Model introspection with ensembling . . . . .                   | 19        |
| <b>B Discussion on the model design</b>                             | <b>20</b> |
| B.1 Understanding the role of task-optimized networks . . . . .     | 20        |
| B.1.1 Performance of separately trained networks . . . . .          | 20        |
| B.1.2 Distance between classification maps . . . . .                | 21        |
| B.1.3 Effect of brain-map dimension reduction . . . . .             | 21        |
| B.2 How to induce transfer learning ? . . . . .                     | 23        |
| B.2.1 The need for objective coupling . . . . .                     | 23        |
| B.2.2 Dropout as a transfer incentive . . . . .                     | 24        |
| B.2.3 Transfer through low-rank constraints/penalty . . . . .       | 24        |
| B.2.4 Empirical comparison of transfer penalties . . . . .          | 25        |
| B.3 Interpretability incentives . . . . .                           | 26        |
| B.3.1 Consensus model and resting-state initialization . . . . .    | 28        |
| B.3.2 Effect of selecting grey-matter components . . . . .          | 28        |
| B.4 Effect of variational Dropout and batch normalization . . . . . | 29        |
| B.5 Effect of study weights . . . . .                               | 29        |
| <b>C Data corpus and reproduction</b>                               | <b>30</b> |
| C.1 Software and parameters . . . . .                               | 30        |
| C.2 Validation metrics . . . . .                                    | 31        |
| C.3 Task-fMRI studies . . . . .                                     | 32        |

**Notations.** We denote scalars, vectors and matrices using lower-case, bold lower-case and bold upper-case letters, e.g.,  $x$ ,  $\mathbf{x}$  and  $\mathbf{X}$ . We denote the elements of  $\mathbf{X}$  by  $x_{i,j}$  and its rows by  $\mathbf{x}_i$ . 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 Detailed 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 [section C.3](#)), 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 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<sup>1</sup> 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  $1 \leq k \leq c^j$  that corresponds to  $k$ -th contrast/base condition (called contrast in the following). Overall, this produced a set of z-maps  $(\mathbf{x}_i^j)_{i \in [c^j n^j]}$  living in  $\mathbb{R}^p$ , where  $p$  is the number of voxels, associated with a sequence of contrast  $(k_i^j)_{i \in [c^j n^j]}$ . Inter-subject decoding proposes a model  $f_\theta^j : \mathbb{R}^p \rightarrow [1, c^j]$  that predicts contrast from z-maps, i.e.,  $\hat{k}_i^j \triangleq f_\theta^j(\mathbf{x}_i^j)$ , where  $\theta$  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  $\mathbf{x}_i$  in  $\mathbb{R}^p$ , we compute the logits  $\mathbf{l}_i$  in  $\mathbb{R}^c$  as

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

where  $\mathbf{W} \in \mathbb{R}^{c \times p}$  and  $\mathbf{b} \in \mathbb{R}^c$  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.,  $\hat{y}_i = \operatorname{argmax}_{1 \leq y \leq c} l_{i,y}$ . The model is trained on the data  $(\mathbf{x}_i, y_i)_{i \in [n]}$  by minimizing the  $\ell_2^2$  regularized multinomial classification problem

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

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

---

<sup>1</sup>[nistats.github.io](https://nistats.github.io)

### 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

$$l_i(\mathbf{V}, \mathbf{b}, \mathbf{D}) \triangleq \mathbf{V}\mathbf{D}\mathbf{x}_i + \mathbf{b},$$

where  $\mathbf{V}$  in  $\mathbb{R}^{c \times k}$  forms the classifying weights of the model, and the matrix  $\mathbf{D}$  in  $\mathbb{R}^{k \times p}$  is *assigned* during training to functional networks learned on resting-state data, as detailed in [section A.5](#). Multiplying input data by  $\mathbf{D}$  projects statistical images onto meaningful resting-state components, in an attempt to improve classification performance and reduce computation cost, akin to the methods proposed in Smith *et al.*<sup>30</sup>, Yeo *et al.*<sup>31</sup>. The model is trained by solving the convex objective (1) separately for each study, replacing  $\mathbf{W}$  by  $\mathbf{V}$  in  $\mathbb{R}^{c \times k}$ :

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

Our results ([Fig. 2c](#)) show that decoding from functional networks is not significantly better than decoding from voxels directly. For both baselines, the parameter  $\lambda$  is found by half-split cross-validation. Training is performed using a L-BFGS solver<sup>32</sup>. We use non standardized maps  $(\mathbf{x}_i)_i$  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

$$l_i^j(\mathbf{W}^j, \mathbf{b}^j) \triangleq \mathbf{W}^j \mathbf{x}_i^j + \mathbf{b}^j,$$

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

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

The matrix  $\mathbf{D}$  corresponds to the first-layer weights pictured in [Fig. 1](#),  $\mathbf{L}$  to the second-layer weights, and  $(\mathbf{U}^j, \mathbf{b}^j)_j$  to the various classification heads of the third layer. In this work, we choose  $k \approx 512$  and  $l = 128$ . While  $\mathbf{D}$  remains fixed, the second-layer matrix  $\mathbf{L}$  and the  $N$  classification heads  $(\mathbf{U}^j)_{j \in [N]}$  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<sup>10,11</sup>, 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*<sup>2</sup>, described by Pan & Yang<sup>20</sup>.

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:

$$\min_{\substack{\mathbf{L} \in \mathbb{R}^{l \times k} \\ (\mathbf{U}^j, \mathbf{b}^j)_j}} - \sum_{j=1}^N \frac{(n^j)^\beta}{n^j} \sum_{i=1}^{n^j} \left( l_{i, y_i}^j(\mathbf{U}^j, \mathbf{b}^j, \mathbf{L}) - \log \left( \sum_{k=1}^{c^j} \exp l_{i, k}^j(\mathbf{U}^j, \mathbf{b}^j, \mathbf{L}) \right) \right), \quad (4)$$

where the dependence on  $\mathbf{D}$  is left implicit. The scalar  $\beta$  in  $[0, 1]$  is a parameter that regulates the importance of each study in the joint objective, that we further discuss in section B.5. 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  $(\mathbf{x}_i^j)_{i \in B}$  that we use to compute the unbiased objective estimate

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

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

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  $\ell_2$  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*<sup>21</sup> during training alleviates the overfitting issue and allows transfer learning to occur. Dropout is a stochastic regularization method that prevents the weights from each layer to co-adapt by perturbing them with multiplicative noise during training. It ensures that the information is well spread across coefficients rows and columns<sup>33</sup>. In our case, this favors transfer learning, as it ensures that no single row of  $\mathbf{L}$ , 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<sup>34</sup> 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

<sup>2</sup>This case is less studied than the classical multi-task setting where input data are single-source but learning tasks are multiple.

noise matrices

$$\mathbf{M}_D = \text{Diag}([b_{D,t}]_{t \in [k]}), \quad \mathbf{M}_L^j = \text{Diag}([b_{L,t}]_{t \in [l]}),$$

where  $b_{D,t} \sim \mathcal{N}(1, \alpha)$  and  $b_{L,t} \sim \mathcal{N}(1, \alpha^j)$ , with  $\alpha$  fixed and  $\alpha^j$  estimated from data.<sup>3</sup> We then compute the noisy logits

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

and use these to compute the loss (5), to which we add a regularization term that regulates the learning of  $\alpha^j$ , introduced by Molchanov *et al.*<sup>35</sup>. We compute the gradient with respect to  $\mathbf{L}$ ,  $\mathbf{U}^j$ ,  $\mathbf{b}^j$  using the local reparametrization trick<sup>34</sup>. We refer to Molchanov *et al.*<sup>35</sup> for more details on variational Dropout and a Bayesian grounding of this approach.

Optimization is performed using *Adam*<sup>22</sup>, a flavor of stochastic gradient descent that depends less on the step-size. We use batch normalization<sup>36</sup> 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  $\mathbf{D}$  in  $\mathbb{R}^{k \times p}$ , where  $k = 512$ . We consider data from the HCP900 release, and stack all records to obtain a data matrix  $\mathbf{X}$  in  $\mathbb{R}^{n \times p}$ . We then use an online solver<sup>37</sup> to solve the sparse non-negative matrix factorization problem

$$\mathbf{A}, \mathbf{D} \triangleq \underset{\mathbf{D} \in \mathcal{C}, \mathbf{A} \in \mathbb{R}^{k \times n}}{\text{argmin}} \quad \|\mathbf{X} - \mathbf{A}\mathbf{D}\|_F^2 + \lambda \|\mathbf{A}\|_F^2, \quad (6)$$

where the constraint  $\mathcal{C} = \{\mathbf{D} \in \mathbb{R}^{k \times p}, \mathbf{D} \geq 0, \|\mathbf{d}_j\|_1 \leq 1 \forall j \in [k]\}$  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  $\mathbf{D}$  cover the whole brain with as little overlap as possible.

**Second-layer initialization.** To initialize the weights of the second layer, we learn a smaller dictionary  $\mathbf{D}_l$  in  $\mathbb{R}^{l \times p}$  as in (6), where  $l = 128$ . We then compute the initial weights  $\mathbf{L}_l$  so that  $\mathbf{D}_l \approx \mathbf{L}_l \mathbf{D}$  using least-square regression. This way, applying the first two layers initially amount to projecting data onto  $l = 128$  larger functional networks  $\mathbf{D}_l$ , 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.3](#).

**Grey matter restriction.** To help interpreting the obtained model, we found it helpful to remove from  $\mathbf{D}$  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.3.2](#).

<sup>3</sup>This *Gaussian* Dropout has a similar behavior to the more commonly used binary Dropout with parameter  $p = \frac{\alpha}{\alpha+1}$ .

## A.6 Model introspection with ensembling

Given any invertible matrix  $\mathbf{M}$  in  $\mathbb{R}^{l \times l}$ , the non-regularized version of the objective (4) is left invariant when transforming  $\mathbf{L}$  into  $\mathbf{ML}$  and each  $\mathbf{U}^j$  into  $\mathbf{U}^j \mathbf{M}^{-1}$ . This prevents us from interpreting the coefficients of  $\mathbf{L}$  at the end of the training procedure, and to retrieve relevant networks by reading the weights of the second weight. The only aspect of  $\mathbf{L}$  that remains unchanged after a linear parameter transformation is its span. Dropout regularization, which favors the canonical directions in matrix space<sup>21</sup>, should break this symmetry, but does not help to uncover meaningful directions in the span of  $\mathbf{L}$  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  $(\mathbf{L}_r)_r$  of the second-layer into a big matrix  $\tilde{\mathbf{L}}$ . We performed a SVD on this matrix, and observed that the first  $l = 128$  components captured 98% of the variance of  $\tilde{\mathbf{L}}$  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  $\mathbf{L}$  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  $(\mathbf{L}_r)_r$  of the second layer for each run, along with the average matrices and biases

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

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  $(\mathbf{L}_r)_r$  into a fat matrix  $\tilde{\mathbf{L}} \in \mathbb{R}^{lr \times k}$  on which we perform sparse non-negative matrix factorization. Namely, we compute  $\bar{\mathbf{L}} \in \mathbb{R}^{l \times k}$ , the new weight matrix for the second layer, solving

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

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

The rows of  $\bar{\mathbf{L}}$  are now interpretable separately, as the non-negative and sparse constraints have broken the inherent parameter invariance of the original model. The rows of  $\bar{\mathbf{L}}$  hold the coefficients for combining resting-state networks



held in  $\mathbf{D}$  into  $l$  multi-study task-optimized networks  $\bar{\mathbf{L}}\mathbf{D}$  in  $\mathbb{R}^{l \times p}$ . We initialize the sparse NMF algorithm with the weights  $\mathbf{L}_l$  computed in section A.5, to inject a small prior regarding final MSTON distribution: before running NMF, those are set to  $\mathbf{L}_l\mathbf{D} \approx \mathbf{D}_l$ , i.e., are close to large resting-state functional networks.

We observed that directly enforcing negativity/sparsity over  $\mathbf{L}$  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.

## 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 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 Performance of separately trained networks

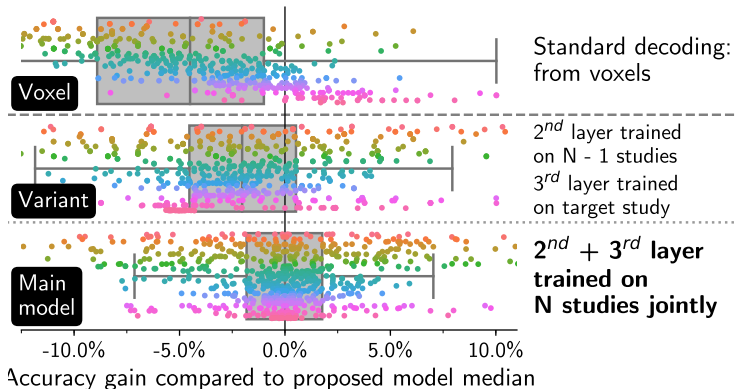


Figure 6: Quantitative improvement linked to training the model on the joint objective (4), versus improvement linked to transfer in the second-layer only. Box plots calculated over 20 random data half-split and all studies.

We argue that using the joint objective (4) improves decoding performance because the data from every study influences 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. 6 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<sup>4</sup>). On the other hand, it performs worse than training 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). 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.1.2 Distance between classification maps

We explore how the model induces structure among classification maps, which partially explains how transfer learning operates. In Fig. 7, we compare correlation between classification maps obtained with our model and the baseline decoder. The absolute correlation between classification maps within and across studies is higher on average. This is because the whole classification matrix is low-rank 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 a smaller contrast in the correlation matrix. Multi-study training thus acts as a regularizer, by forcing maps from each study to be more correlated to maps from other studies.

### B.1.3 Effect of brain-map dimension reduction

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

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

---

<sup>4</sup>Due the fact that half-split folds are overlapping and performance between 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.

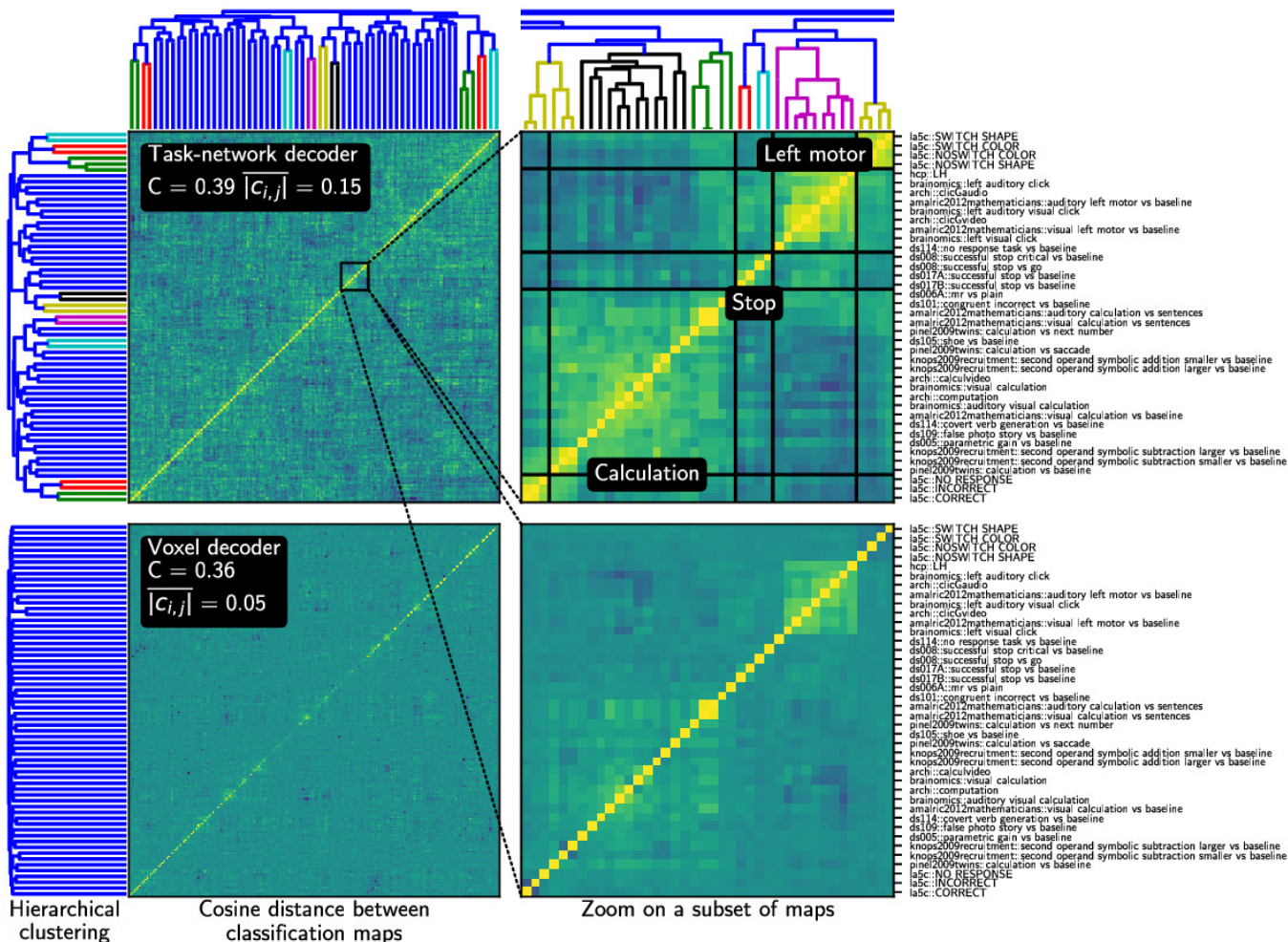


Figure 7: Cosine distances between classification maps, obtained with our multi-study decoder (top) and with decoders learned separately (bottom), clustered using average-linkage hierarchical clustering. The classification maps obtained when decoding from task-optimized networks are more easily clustered into cognitive-meaningful groups using hierarchical clustering—the cophenetic coefficient of the top clustering is thus higher.

$\mathbf{x}_{\text{proj}}$  is thus a denoised, low-dimensional representation of the brain map  $\mathbf{x}$ , held in the span of the  $l$  multi-study task-optimized networks held in matrix  $\mathbf{M}$ . We compare different maps  $\mathbf{x}$  to their projection  $\mathbf{x}_{\text{proj}}$  in Figure 8.

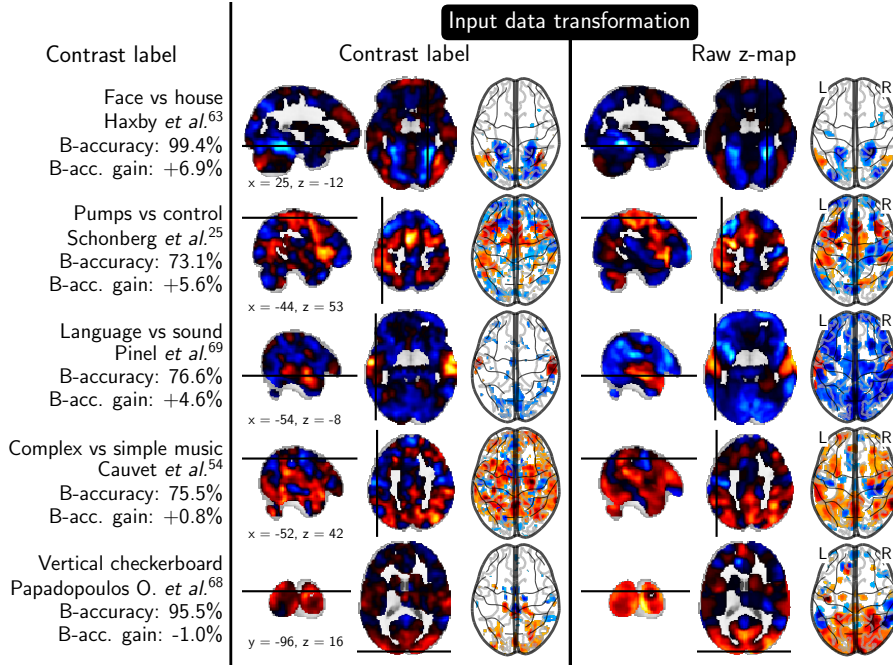


Figure 8: In a dual perspective to Fig. 5, the representation of input data on task-optimized networks is simpler and therefore easier to classify.

## B.2 How to induce transfer learning ?

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

### 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  $(\tilde{\mathbf{V}}^j, \mathbf{b}^j)_j$  be the unique solutions of the  $N$  non-regularized convex problems (2). We let  $\tilde{\mathbf{V}} \in \mathbb{R}^{c \times k}$  be the vertical concatenation of  $(\mathbf{V}^j)_j$ . We then form the matrices

$$\mathbf{L} = \begin{bmatrix} \tilde{\mathbf{V}} \\ \mathbf{0} \in \mathbb{R}^{l-c \times k} \end{bmatrix} \in \mathbb{R}^{l \times k} \text{ and } \begin{bmatrix} \mathbf{U}^1 \\ \vdots \\ \mathbf{U}^N \end{bmatrix} \triangleq [\mathbf{I}_c \in \mathbb{R}^{c \times c}, \mathbf{0} \in \mathbb{R}^{l-c \times l}], \quad (8)$$

where  $\mathbf{I}_c$  is the identity matrix of  $\mathbb{R}^{c \times c}$ .  $\mathbf{L}$  is thus split into row-blocks  $(\tilde{\mathbf{V}}^j)_j$ , dedicated to and learned on *single studies*. It follows from elementary considerations that the matrices  $(\mathbf{L}, (\mathbf{U}^j, \mathbf{b}^j)_j)$  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 modifications of the objective (4) allows 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 appendix A, we can use Dropout between the second layer weight  $L$  and the third layer head weights  $U^j$ . Dropout prevents constructions of block-separated solution of objective (4) similar to the one proposed in (8). Indeed, every reduced sample  $LD\mathbf{x}_i^j$  fed to the third layer classification head  $j$  can see any of his features corrupted by multiplicative noise  $M_L$  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  $l_h D$ , for  $h \in [l]$  should be relevant for decoding every study. This fosters transfer learning as  $L$  carry 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

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

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  $\mathbf{V}$  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  $\hat{\mathbf{V}}$  of the separate objectives (2), as we did in the construction (8). As a consequence, the data from studies truly influence the solutions  $(\mathbf{L}, (\mathbf{U}^j, \mathbf{b}_j)_j)$  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 Srebro *et al.*<sup>38</sup>, we may resort to a convex objective function parameterized by  $\mathbf{V}$  in  $\mathbb{R}^{c \times k}$ , that penalizes the rank of  $\mathbf{V}$ . We learn  $\mathbf{V}^j$  in  $\mathbb{R}^{c^j \times k}$  for all study  $j$  in  $[N]$  solving the joint objective

$$\begin{aligned} \min_{(\mathbf{V}^j, \mathbf{b}^j)_j} & - \sum_{j=1}^N \frac{(n^j)^\beta}{n^j} \sum_{i=1}^{n^j} \left( l_{i, y_i}^j(\mathbf{V}^j, \mathbf{b}^j) - \log \left( \sum_{k=1}^{c^j} \exp l_{i, k}^j(\mathbf{V}^j, \mathbf{b}^j) \right) \right) \\ & + \lambda \left\| \left[ \mathbf{V}^{1\top} \dots \mathbf{V}^{N\top} \right] \right\|_{\star}, \end{aligned} \quad (9)$$

where  $\|\mathbf{V}\|_{\star}$  is the nuclear norm of  $\mathbf{V}$ , defined as  $\sum_{i=1}^{\min(c, k)} \sigma_i(\mathbf{V})$ , where  $(\sigma_i(\mathbf{V}))_i$  are the singular values of  $\mathbf{V}$ . The nuclear norm is a convex proxy for the rank of

matrix  $\mathbf{V}$ . As a consequence, the rank of the solution decreases from  $\min(c, k)$  to 0 as  $\lambda$  increases. The objective (9) is solvable using proximal methods, e.g., FISTA<sup>39</sup>. 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<sup>40</sup>, amenable to stochastic gradient descent<sup>41</sup>, that includes the solution of (9) as a minimizer. It is obtained by setting  $l = \max(x, k)$  and adding  $\ell_2^2$  penalties to the objective (4):

$$\min_{\substack{\mathbf{L} \in \mathbb{R}^{l \times k} \\ (\mathbf{U}^j, \mathbf{b}^j)_j}} - \sum_{j=1}^N \frac{(n^j)^\beta}{n^j} \sum_{i=1}^{n^j} \left( l_{i, y_i}^j(\mathbf{U}^j, \mathbf{b}^j, \mathbf{L}) - \log\left(\sum_{k=1}^{c^j} \exp l_{i,k}^j(\mathbf{U}^j, \mathbf{b}^j, \mathbf{L})\right) \right) + \frac{\lambda}{2} \left( \|\mathbf{L}\|_F^2 + \sum_{j=1}^N \|\mathbf{U}^j\|_F^2 \right), \quad \text{where } \mathbf{U}^j \in \mathbb{R}^{c^j \times l} \forall j \in [N].$$

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  $\mathbf{V}$  to be low-rank—this can then be understood as a regularization technique through feature noising<sup>42</sup>.

#### B.2.4 Empirical comparison of transfer penalties

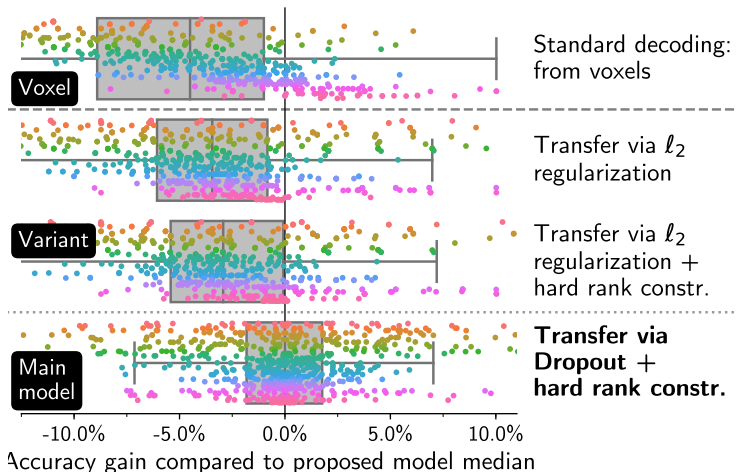


Figure 9: Quantitative comparison of transfer inducing regularization: dropout with hard-rank constraints outperforms  $\ell_2$  regularization with and without hard-rank constraints. Box plots calculated over 20 random data half-split and all studies.

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 the



experiment summarized in Fig. 9. We compare three regularization variants, measuring the improvement due to hard low-rank constraints and the difference between dropout and  $\ell_2$ . The three estimators use input dropout ( $p = 0.25$ ), while dropout between layer 2 and 3 is initialized to  $p = 0.75$  when used.<sup>5</sup> We observe that forcing  $\mathbf{V}$  to be low-rank is beneficial (0.7% mean accuracy gain, 72% experiments with net increase), and that dropout regularization performs significantly better than low-rank inducing  $\ell_2$  penalties (2.7% mean accuracy gain, 79% experiments with net increase). This justifies using dropout and hard-rank constraints for regularization.

### B.3 Interpretability incentives

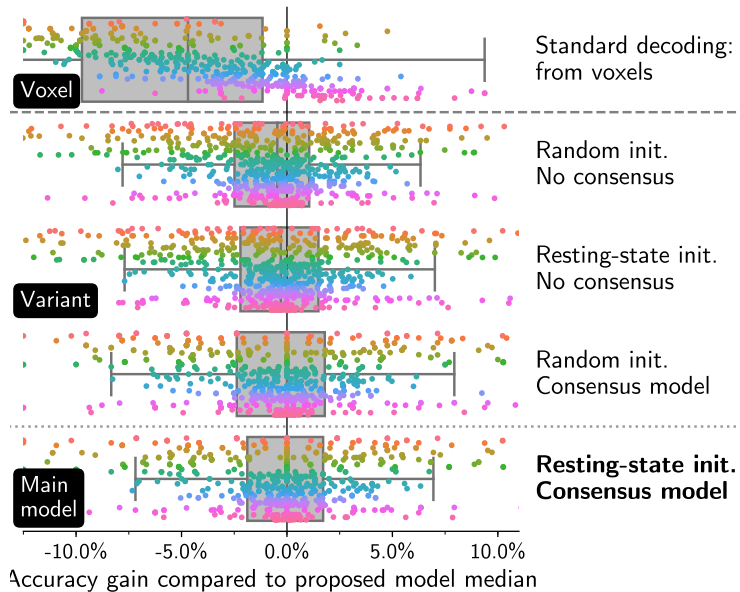


Figure 10: Quantitative improvement linked to ensembling and resting-state initialization in the method. Box plots calculated over 20 random data half-split and all studies.

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  $\mathbf{L}$  are non-negative

<sup>5</sup> The  $\ell_2$  accuracy gain is an upper-bound of its actual performance when  $\lambda$  is set with cross-validation, as we take the highest performing  $\lambda$  on the *test* sets. Symmetrically, we may slightly improve results by setting dropout rates using cross-validation—we choose not to avoid the fragility of cross-validation in neuro-imaging<sup>43</sup>.

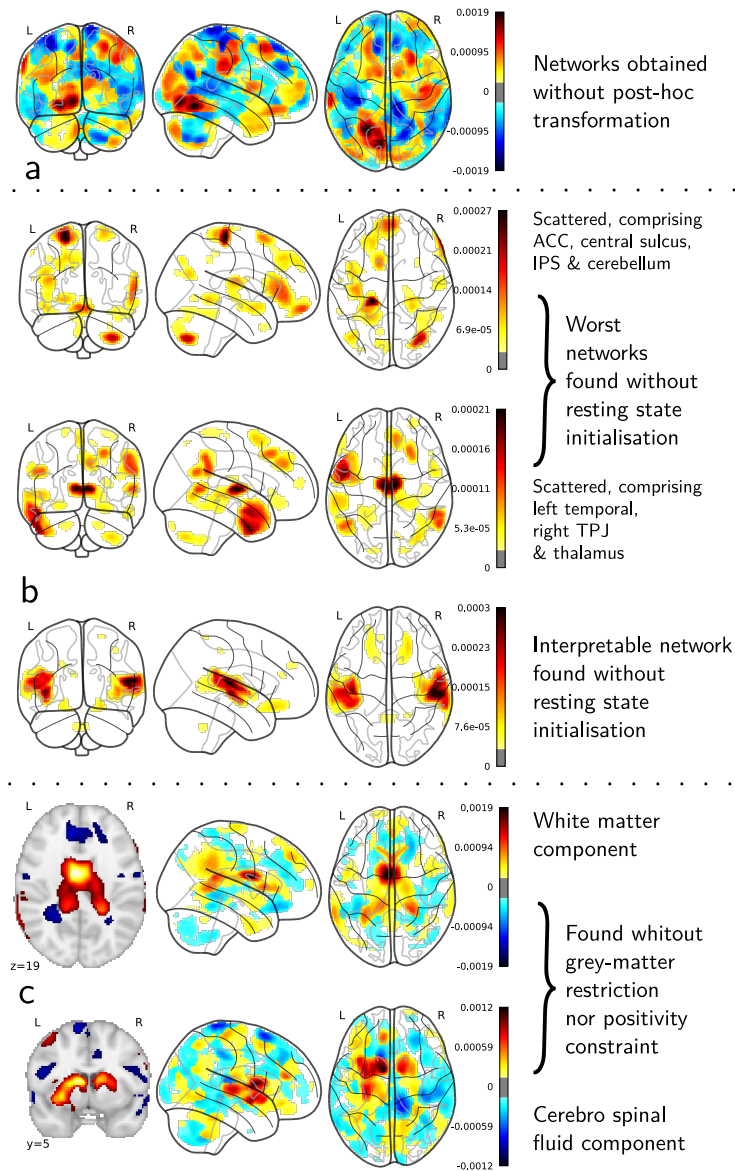


Figure 11: Without interpretability refinements (a), resting-state based initialization (b) and grey matter components selection (c), some task-optimized networks may be hard to interpret/ not relevant from a cognitive perspective.

and interpretable. Third, we initialize the second-layer weights so that  $L_{\text{init}}D$  corresponds to resting-state functional networks  $D_l$ , coarser than  $D$ . This initialization is used both during the training phase and the consensus phase.

### B.3.1 Consensus model and resting-state initialization

In Fig. 10, 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  $\mathbf{L}$ . Without this refinement, the directions we obtain are similar to the one displayed in Fig. 11a, 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<sup>44</sup>, 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. 11b. Two of those are scattered networks, which 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. 11b.

### B.3.2 Effect of selecting grey-matter components

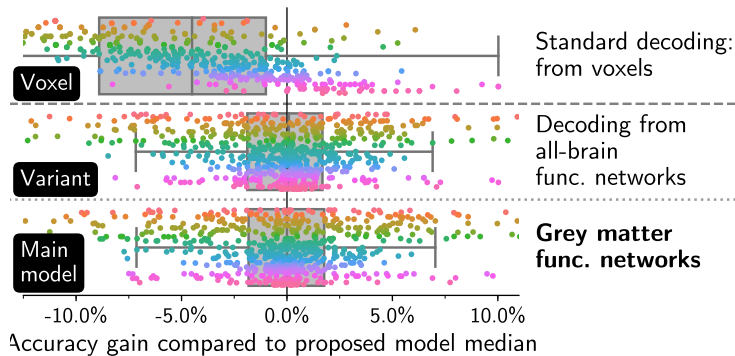


Figure 12: Working with functional networks located in the grey matter only do not have a significant impact on performance. Box plots calculated over 20 random data half-split and all studies.

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 MSTON that are distributed or formed with non grey-matter regions. In Fig. 11c, we show that without those precautions, our model finds networks located in the white matter and the cerebro-spinal fluid zones. Quantitatively (Fig. 12), as expected, performing

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

## B.4 Effect of variational Dropout and batch normalization

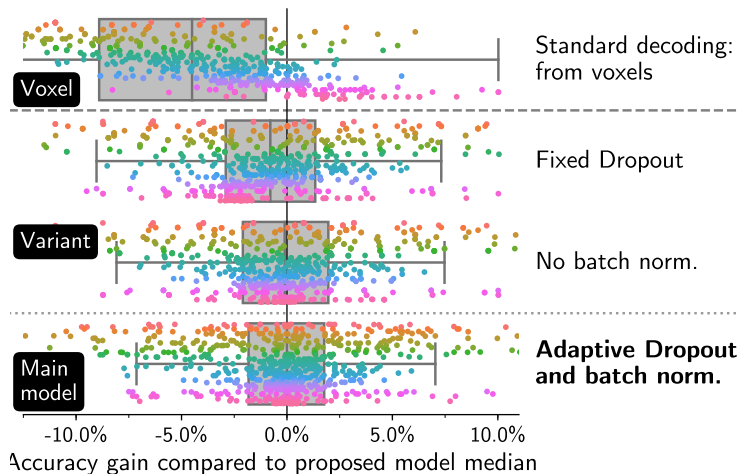


Figure 13: Batch normalization and adaptive variational dropout both have a beneficial impact on classification accuracy of the final learned decoder. Box plots calculated over 20 random data half-split and all studies.

We introduced variational dropout and batch normalization in the training procedure of our algorithm. Fig. 13 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<sup>36</sup>.

## B.5 Effect of study weights

Our model learns the second and third layer weights by solving

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

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. 14b shows, it is beneficial to set an intermediary  $\beta$ , typically  $\beta = 0.6$ . On the one hand, we want to give the smallest study of

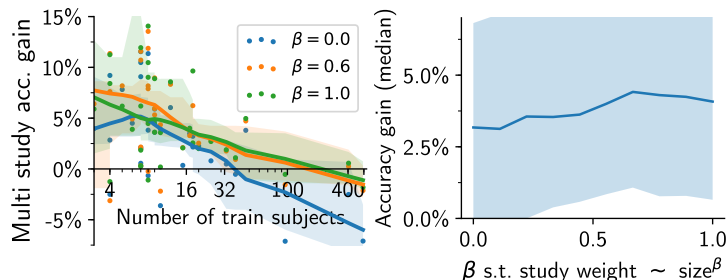


Figure 14: Impact of changing the study weight in the joint objective. Giving more weight ( $\beta \rightarrow 1$ ) to large studies prevent negative transfer learning but may reduce overall performance. Small studies should not be given too much weight ( $\beta \rightarrow 0$ ), as this voids the benefits 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.

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”

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

that is larger than the true sample size for smaller studies and smaller for larger studies. We observe on Fig. 14a that the negative transfer learning endured by large-study decoders such as HCP and LA5C reduces as these studies are given more weight ( $\beta \rightarrow 1$ ). On the other hand, the performance on small datasets slightly reduces for  $\beta > 0.6$ . It also reduces for low  $\beta$ , 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.

## C Data corpus and reproduction

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*<sup>45</sup> and *scikit-learn*<sup>46</sup> in our experiment pipelines, the stochastic solver from Mensch *et al.*<sup>37</sup> to learn resting state dictionaries and *PyTorch*<sup>47</sup> for

model design and training. A Python package<sup>6</sup> is available for reproducibility and reuse. It provides the multi-scale resting-state dictionaries extracted from HCP, as those are costly to learn.

**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<sup>48</sup> 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  $\lambda$  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  $\lambda$  on a grid

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

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

**Consensus phase.** We run the training procedure 100 times with different random seeds. We set  $\lambda = 10^{-4}$ , so as to obtain 80% sparsity. 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%.

## 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 Fig. 9–12 reports the median and 25%, 75% quantiles of

$$\{a_r^j - a_r^j(0), j \in [1, \dots, N], r \in [1, 2, \dots, 20]\},$$

---

<sup>6</sup>[cogspaces.github.io](https://cogspaces.github.io)



where  $r$  is the half-split run index and  $a_r^j(0)$  is the accuracy obtained for study  $j$  and run  $r$  using the baseline method.

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

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

### C.3 Task-fMRI studies

[Table 1](#) recapitulates the various studies used in our corpus and provide their sources. The names corresponds to the ones used in [Fig. 2](#).

| Task name                             | Source study  | # subjects | # contrasts |
|---------------------------------------|---|------------|-------------|
| Cross-language repetition priming     | Alvarez <i>et al.</i> <sup>49</sup>                     | 13         | 17          |
| High-level math                       | Amalric & Dehaene <sup>50</sup>                         | 30         | 31          |
| Classification learning               | Aron <i>et al.</i> <sup>51</sup>                        | 17         | 7           |
| Stop-signal                           | Aron <i>et al.</i> <sup>52</sup>                        | 15         | 12          |
| The Human Connectome Project          | Barch <i>et al.</i> <sup>53</sup>                       | 787        | 23          |
| Constit. struct. of sent. & music     | Hara <i>et al.</i> <sup>24</sup> , Cauvet <sup>54</sup> | 35         | 19          |
| BART, stop-signal, emotion            | Cohen <sup>55</sup>                                     | 24         | 23          |
| Auditory & Visual Oddball             | Collier <i>et al.</i> <sup>56</sup>                     | 17         | 8           |
| Sentence/music complexity             | Devauchelle <i>et al.</i> <sup>57</sup>                 | 40         | 25          |
| Simon task                            | Kelly & Milham <sup>58</sup>                            | 7          | 8           |
| Word & object processing              | Duncan <i>et al.</i> <sup>59</sup>                      | 49         | 6           |
| Weather prediction                    | Foerde <i>et al.</i> <sup>60</sup>                      | 14         | 14          |
| Spatio-temporal judgement             | Gauthier <i>et al.</i> <sup>61</sup>                    | 11         | 30          |
| Spatio-temporal judgement (retake)    | Gauthier <i>et al.</i> <sup>61</sup>                    | 13         | 23          |
| Motor task & word/verb generation     | Gorgolewski <i>et al.</i> <sup>62</sup>                 | 10         | 11          |
| Visual object recognition             | Haxby <i>et al.</i> <sup>63</sup>                       | 6          | 13          |
| Face recognition                      | Henson <i>et al.</i> <sup>64</sup>                      | 16         | 5           |
| Plain or mirror-reversed text         | Jimura <i>et al.</i> <sup>65</sup>                      | 14         | 9           |
| Arithmetic & saccades                 | Knops <i>et al.</i> <sup>66</sup>                       | 19         | 26          |
| False belief                          | Moran <i>et al.</i> <sup>67</sup>                       | 36         | 7           |
| Brainomics localizer                  | Papadopoulos Orfanos <i>et al.</i> <sup>68</sup>        | 94         | 19          |
| Localizer                             | Pinel <i>et al.</i> <sup>69</sup>                       | 78         | 30          |
| Twin localizer                        | Pinel & Dehaene <sup>70</sup>                           | 65         | 34          |
| UCLA LA5C                             | Poldrack <i>et al.</i> <sup>71</sup>                    | 191        | 24          |
| Classif. learning & reversal          | Poldrack <i>et al.</i> <sup>72</sup>                    | 13         | 3           |
| Stop-signal & classification          | Rizk-Jackson <i>et al.</i> <sup>48</sup>                | 8          | 11          |
| Stop-signal & classification (retake) | Rizk-Jackson <i>et al.</i> <sup>48</sup>                | 8          | 11          |
| Balloon Analog Risk-taking            | Schonberg <i>et al.</i> <sup>25</sup>                   | 16         | 12          |
| CamCan audio-visual                   | Shafto <i>et al.</i> <sup>73</sup>                      | 605        | 5           |
| Mixed-gambles                         | Tom <i>et al.</i> <sup>74</sup>                         | 16         | 4           |
| Incidental encoding                   | Uncapher <i>et al.</i> <sup>75</sup>                    | 18         | 26          |
| Compression                           | Vagharchakian <i>et al.</i> <sup>76</sup>               | 16         | 14          |
| Emotion regulation                    | Wager <i>et al.</i> <sup>77</sup>                       | 34         | 26          |
| Stop-signal w/ spoken & manual resp.  | Xue <i>et al.</i> <sup>78</sup>                         | 20         | 6           |
| Rhyme judgment                        | Xue & Poldrack <sup>79</sup>                            | 13         | 3           |
| Gathered data                         |   | 2368       | 545         |

Table 1: Studies used in our corpus.

## References from appendix

- Ando, R. K. & Zhang, T. A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *Journal of Machine Learning Research* **6**, 1817–1853 (2005).
- Xue, Y., Liao, X., Carin, L. & Krishnapuram, B. Multi-Task Learning for Classification with Dirichlet Process Priors. *Journal of Machine Learning Research* **8**, 35–63 (2007).
- Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345–1359 (2010).

21. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **15**, 1929–1958 (2014).
22. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization in International Conference for Learning Representations* (2015).
30. Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R. & Laird, A. R. Correspondence of the Brain’s Functional Architecture During Activation and Rest. *Proceedings of the National Academy of Sciences* **106**, 13040–13045 (2009).
31. Yeo, T. B. T. *et al.* The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity. *Journal of Neurophysiology* **106**, 1125–1165 (2011).
32. Nocedal, J. Updating Quasi-Newton Matrices with Limited Storage. *Mathematics of Computation* **35**, 773–782 (1980).
33. Neyshabur, B. *Implicit Regularization in Deep Learning* PhD thesis (Toyota Technological Institute at Chicago, 2017).
34. Kingma, D. P., Salimans, T. & Welling, M. *Variational Dropout and the Local Reparameterization Trick in Advances in Neural Information Processing Systems* (2015), 2575–2583.
35. Molchanov, D., Ashukha, A. & Vetrov, D. *Variational Dropout Sparsifies Deep Neural Networks in Proceedings of the International Conference on Machine Learning* (2017), 2498–2507.
36. 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), 448–456.
37. Mensch, A., Mairal, J., Thirion, B. & Varoquaux, G. Stochastic Subsampling for Factorizing Huge Matrices. *IEEE Transactions on Signal Processing* **66**, 113–128 (2018).
38. Srebro, N., Rennie, J. & Jaakkola, T. S. *Maximum-Margin Matrix Factorization in Advances in Neural Information Processing Systems* (2004), 1329–1336.
39. Beck, A. & Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences* **2**, 183–202 (2009).
40. Rennie, J. D. M. & Srebro, N. *Fast Maximum Margin Matrix Factorization for Collaborative Prediction in Proceedings of the International Conference on Machine Learning* (2005), 713–719.
41. Bell, R. M. & Koren, Y. Lessons from the Netflix Prize Challenge. *ACM SIGKDD Explorations Newsletter* **9**, 75–79 (2007).
42. Wager, S., Wang, S. & Liang, P. S. *Dropout Training as Adaptive Regularization in Advances in Neural Information Processing Systems* (2013), 351–359.

43. Varoquaux, G. Cross-Validation Failure: Small Sample Sizes Lead to Large Error Bars. *NeuroImage. New advances in encoding and decoding of brain signals* **180**, 68–77 (2018).
44. Breiman, L. Bagging Predictors. *Machine Learning* **24**, 123–140 (1996).
45. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kos-saifi, J., Gramfort, A., Thirion, B. & Varoquaux, G. Machine Learning for Neuroimaging with Scikit-Learn. *Frontiers in Neuroinformatics* **8**, 14 (2014).
46. Pedregosa, F. *et al.* Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
47. Paszke, A., Gross, S., Chintala, S. & Chanan, G. PyTorch: Tensors and Dynamic Neural Networks in Python with Strong GPU Acceleration (2017).

## Data corpus

24. Hara, N., Cauvet, E., Devauchelle, A. D., Dehaene, S., Pallier, C., *et al.* Neural Correlates of Constituent Structure in Language and Music. *Neuro-Image* **47**, S143 (2009).
25. Schonberg, T., Fox, C., Mumford, J. A., Congdon, C., Trepel, C. & Pol-drack, R. A. Decreasing Ventromedial Prefrontal Cortex Activity During Sequential Risk-Taking: An fMRI Investigation of the Balloon Analog Risk Task. *Frontiers in Neuroscience* **6**, 80 (2012).
48. Rizk-Jackson, A., Aron, A. R. & Poldrack, R. *Classification Learning and Stop-Signal (one Year Test-Retest)* <https://openfmri.org/dataset/ds000017>.
49. Alvarez, R. P., Jaszewski, G. & Poldrack, R. A. *Building Memories in Two Languages: An fMRI Study of Episodic Encoding in Bilinguals in Society for Neuroscience Abstracts* (2002).
50. Amalric, M. & Dehaene, S. Origins of the Brain Networks for Advanced Mathematics in Expert Mathematicians. *Proceedings of the National Academy of Sciences* **113**, 4909–4917 (2016).
51. Aron, A. R., Gluck, M. & Poldrack, R. A. Long-Term Test–Retest Reliability of Functional MRI in a Classification Learning Task. *NeuroImage* **29**, 1000–1006 (2006).
52. Aron, A. R., Behrens, T. E., Smith, S., Frank, M. J. & Poldrack, R. Triangulating a Cognitive Control Network Using Diffusion-Weighted Magnetic Resonance Imaging (MRI) and Functional MRI. *The Journal of Neuro-science* **27**, 3743–3752 (2007).
53. Barch, D. M. *et al.* Function in the Human Connectome: Task-fMRI and Individual Differences in Behavior. *NeuroImage* **80**, 169–189 (2013).
54. Cauvet, E. *Traitement des structures syntaxiques dans le langage et dans la musique* fr. PhD thesis (Paris 6, 2012).

55. Cohen, J. R. *The Development and Generality of Self-Control* PhD thesis (University of the City of Los Angeles, 2009).
56. Collier, A. K., Wolf, D. H., Valdez, J. N., Turetsky, B. I., Elliott, M. A., Gur, R. E. & Gur, R. C. Comparison of Auditory and Visual Oddball fMRI in Schizophrenia. *Schizophrenia research* **158**, 183–188 (2014).
57. Devauchelle, A. D., 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* **21**, 1000–1012 (2009).
58. Kelly, A. & Milham, M. *Simon Task* <https://openfmri.org/dataset/ds000101>.
59. Duncan, K., Pattamadilok, C., Knierim, I. & Devlin, J. Consistency and Variability in Functional Localisers. *NeuroImage* **46**, 1018–1026 (2009).
60. Foerde, K., Knowlton, B. & Poldrack, R. A. Modulation of Competing Memory Systems by Distraction. *Proceedings of the National Academy of Science* **103**, 11778–11783 (2006).
61. Gauthier, B., Eger, E., Hesselmann, G., Giraud, A.-L. & Kleinschmidt, A. Temporal Tuning Properties Along the Human Ventral Visual Stream. *The Journal of Neuroscience* **32**, 14433–14441 (2012).
62. Gorgolewski, K. J., Storkey, A., Bastin, M. E., Whittle, I. R., Wardlaw, J. M. & Pernet, C. R. A Test-Retest fMRI Dataset for Motor, Language and Spatial Attention Functions. *GigaScience* **2**, 6 (2013).
63. Haxby, J. V., Gobbini, I. M., Furey, M. L., Ishai, A., Schouten, J. L. & Pietrini, P. Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science* **293**, 2425–2430 (2001).
64. Henson, R. N., Wakeman, D. G., Litvak, V. & Friston, K. J. A Parametric Empirical Bayesian Framework for the EEG/MEG Inverse Problem: Generative Models for Multi-Subject and Multi-Modal Integration. *Frontiers in Human Neuroscience* **5** (2011).
65. Jimura, K., Cazalis, F., Stover, E. R. S. & Poldrack, R. A. The Neural Basis of Task Switching Changes with Skill Acquisition. *Frontiers in Human Neuroscience* **8** (2014).
66. Knops, A., Thirion, B., Hubbard, E. M., Michel, V. & Dehaene, S. Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. *Science* **324**, 1583–1585 (2009).
67. Moran, J. M., Jolly, E. & Mitchell, J. P. Social-Cognitive Deficits in Normal Aging. *The Journal of Neuroscience* **32**, 5553–5561 (2012).
68. Papadopoulos Orfanos, D., Michel, V., Schwartz, Y., Pinel, P., Moreno, A., Le Bihan, D. & Frouin, V. The Brainomics/Localizer Database. *NeuroImage* **144**, 309–314 (2017).

69. Pinel, P., Thirion, B., Meriaux, S., Jobert, A., Serres, J., Bihan, D. L., Poline, J. B. & Dehaene, S. Fast Reproducible Identification and Large-Scale Databasing of Individual Functional Cognitive Networks. *BMC neuroscience* **8**, 91 (2007).
70. Pinel, P. & Dehaene, S. Genetic and Environmental Contributions to Brain Activation During Calculation. *NeuroImage* **81**, 306–316 (2013).
71. Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K., Mumford, J. A., Sabb, F., Freimer, N., London, E., Cannon, T., *et al.* A Phenome-Wide Examination of Neural and Cognitive Function. *Scientific Data* **3**, 160110 (2016).
72. Poldrack, R. A., Clark, J., Pare-Blagoev, E., Shohamy, D., Creso Moyano, J., Myers, C. & Gluck, M. Interactive Memory Systems in the Human Brain. *Nature* **414**, 546–550 (2001).
73. Shafto, M. A. *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* **14**, 204 (2014).
74. Tom, S. M., Fox, C. R., Trepel, C. & Poldrack, R. A. The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science* **315**, 515–518 (2007).
75. Uncapher, M. R., Hutchinson, J. B. & Wagner, A. D. Dissociable Effects of Top-Down and Bottom-Up Attention During Episodic Encoding. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **31**, 12613–12628 (2011).
76. Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C. & Dehaene, S. A Temporal Bottleneck in the Language Comprehension Network. *The Journal of Neuroscience* **32**, 9089–9102 (2012).
77. Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A. & Ochsner, K. N. Prefrontal-Subcortical Pathways Mediating Successful Emotion Regulation. *Neuron* **59**, 1037–1050 (2008).
78. Xue, G., Aron, A. R. & Poldrack, R. A. Common Neural Substrates for Inhibition of Spoken and Manual Responses. *Cerebral Cortex* **18**, 1923–1932 (2008).
79. Xue, G. & Poldrack, R. A. The Neural Substrates of Visual Perceptual Learning of Words: Implications for the Visual Word Form Area Hypothesis. *Journal of Cognitive Neuroscience* **19**, 1643–1655 (2007).