nature neuroscience

Article

https://doi.org/10.1038/s41593-022-01224-0

Multitask representations in the human cortex transform along a sensory-to-motor hierarchy

Received: 2 December 2021

Accepted: 28 October 2022

Published online: 19 December 2022

Check for updates

Takuya Ito [1] & John D. Murray [1,2,3]

Human cognition recruits distributed neural processes, yet the organizing computational and functional architectures remain unclear. Here, we characterized the geometry and topography of multitask representations across the human cortex using functional magnetic resonance imaging during 26 cognitive tasks in the same individuals. We measured the representational similarity across tasks within a region and the alignment of representations between regions. Representational alignment varied in a graded manner along the sensory–association–motor axis. Multitask dimensionality exhibited compression then expansion along this gradient. To investigate computational principles of multitask representations, we trained multilayer neural network models to transform empirical visual-to-motor representations. Compression-then-expansion organization in models emerged exclusively in a rich training regime, which is associated with learning optimized representations that are robust to noise. This regime produces hierarchically structured representations similar to empirical cortical patterns. Together, these results reveal computational principles that organize multitask representations across the human cortex to support multitask cognition.

Humans perform a variety of tasks in daily life that involve diverse cognitive functions. What are the neural and computational architectures that facilitate multitask cognition? Current efforts to uncover the neural bases of human cognition typically design carefully controlled experiments that target specific cognitive functions while measuring brain activity[1,2]. While this approach has been fruitful for mapping regional activations by cognitive processes, it is typically unable to reveal the organization of task representations and their transformations across the brain. By contrast, advancements in data analysis have enabled the characterization of representational content and transformations of rich sensory stimuli within the visual cortical hierarchy of individuals[3–5]. However, how the brain's large-scale organization supports the representational capacity that enables its diverse cognitive functions beyond sensory perception remains poorly studied.

Univariate task-driven functional magnetic resonance imaging (fMRI) studies have revealed the spatial organization of cognitive specialization across the cortex[6] by mapping the stimulus response properties of brain areas and voxels. Such studies have identified regional correlates of working memory[7], visual processing[8] and motor function[9], among other cognitive functions. Complementing these studies, meta-analyses of neuroimaging studies have made progress in identifying cortical coactivation patterns across many tasks, affording insight into how brain regions coactivate during tasks[10–12]. Despite these initial advances in describing cortical organization of cognitive processes and their correspondence to intrinsic brain organization[11], univariate task activation studies are limited in their ability to reveal the fine-grained (voxel-wise) representations within brain regions and how these representations transform across the cortex.

[1]Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA. [2]Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA. [3]Department of Physics, Yale University, New Haven, CT, USA. e-mail: john.murray@yale.edu

https://doi.org/10.1038/s41593-022-01224-0

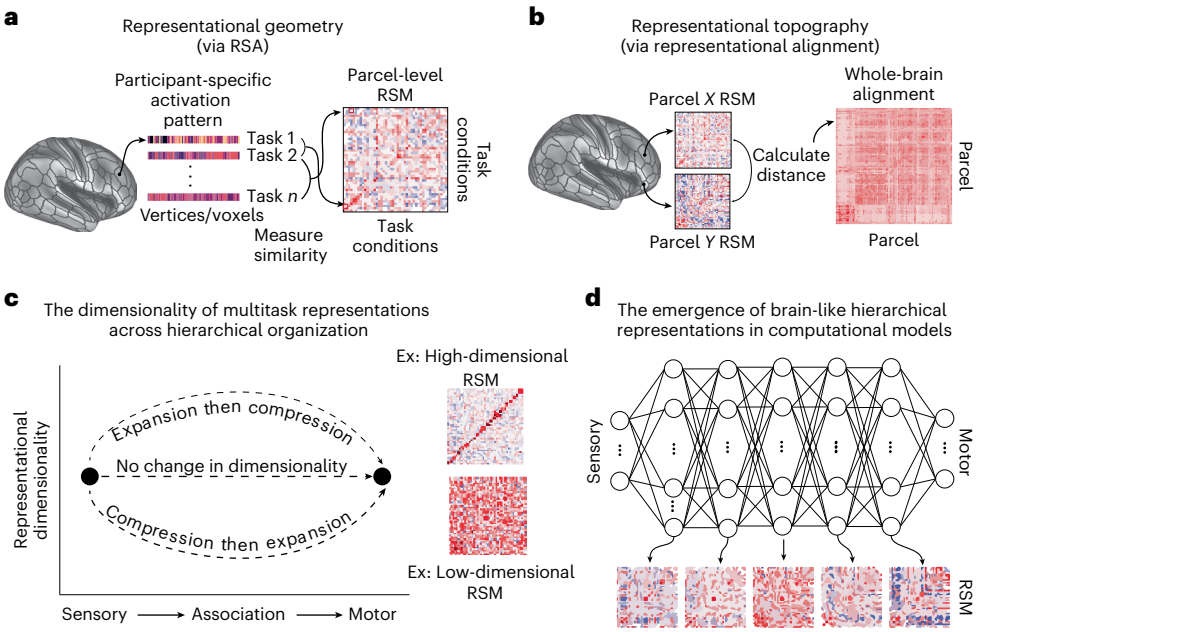


Fig. 1 | Overview of analytic approaches to study the geometry and topography of multitask representations in fMRI data. a, Representational geometry of brain parcels is characterized by RSA (using cosine similarity) applied to individual-specific vertex activation patterns within each parcel[4]. Using individual-specific activation patterns ensures that fine-grained (voxel-wise) representational geometries would not be lost through cross-participant averaging. This enables estimation of an RSM for each brain region using vertices within that region. **b**, Using each region's RSM, we can characterize the topography of representations by measuring the RA (the similarity of regional RSMs) between all pairs of brain regions. **c**, We next asked how the dimensionality of representations changes across the sensory–motor hierarchy. An example (Ex) of a high-dimensional representation is one with a strong diagonal but weak off-diagonal. By contrast, a low-dimensional representation is one with a lack of structure in the RSM and uniform similarity in activation patterns between conditions. **d**, Given the empirical results, we identified the conditions by which similar hierarchical representations emerge in the internal layers of feedforward neural network models trained to produce sensory-to-motor transformations.

One leading approach to investigate the structure of task representations within and across cortical regions is representational similarity analysis (RSA)[4]. RSA measures geometrical properties of task representations within a brain region by comparing the similarity of multivariate activations (for example, multiple voxels in fMRI) across different task conditions. Representational geometries can then be compared between brain regions[4] and between brain data and computational models[5,13]. While this approach can identify task-relevant representational geometries for specific brain regions, most studies have typically been limited to isolated tasks in specific domains (for example, perceptual tasks)[14,15]. This limits the interpretation of representational geometry within and between brain regions because such tasks only recruit a small subset of the diverse cognitive processes of which humans are capable. By contrast, the study of many tasks can clarify the general cognitive principles by which a single brain architecture can implement many diverse functions.

Understanding how the human brain achieves multitask cognition is key to understanding what makes human cognition unique and how to engineer it into model systems[16]. One recent study investigated how different brain areas encode many different tasks across the entire cortex[17]. This study revealed clustering of task representations in the cortex and which brain areas were selective to each task type. We build on that study using RSA to characterize how the topography of representations organize along cortical hierarchies and how the local representational properties of each brain region (such as its dimensionality) change across that cortical hierarchy. Explicitly quantifying representational transformations across cortical hierarchies would provide insight into the computational principles underlying human-like cognitive processing.

Here, we investigated multitask representations across the cortical hierarchy and how artificial neural network models (ANNs) can approximate these representations. To investigate these questions, we used a recently published human fMRI dataset with 26 tasks collected per participant[18]. To characterize the cortex-wide organization of multitask representations, we relate the representational topographies to established large-scale cortical gradients[19] (reflecting hierarchy) derived from resting-state functional connectivity (RSFC). Relating multitask representational organization, which is derived from task fMRI, to RSFC hierarchical gradients, which are derived from task-free MRI, directly grounds regional variations in task representation with the intrinsic organization of the human cortex.

To study how representations transformed across brain regions, we first quantified the multitask representations within each cortical area using RSA. We then measured the alignment of representations between all pairs of cortical areas. This revealed that the axis of greatest representation variation spanned from sensory to association to motor organization. We next quantified the dimensionality of multitask representations across this sensory-to-motor hierarchy, finding that multitask representational dimensionality first compressed from sensory to association areas and then expanded from association to motor areas. This stands in contrast to the expansion then compression of hidden representations documented in task-optimized deep ANNs[20,21]. To investigate the computational principles by which we could reproduce brain-like representations, we trained feedforward ANNs directly on multitask brain activity. Compression-then-expansion organization in ANNs emerged exclusively in a rich feature learning regime, which is associated with learned representations that are robust to noise[22,23]. This regime produced hierarchically structured representations similar to those in the brain. Together, our findings reveal the hierarchical organization of multitask representations in the human brain and establish a framework to produce brain-like representations in computational models.

https://doi.org/10.1038/s41593-022-01224-0

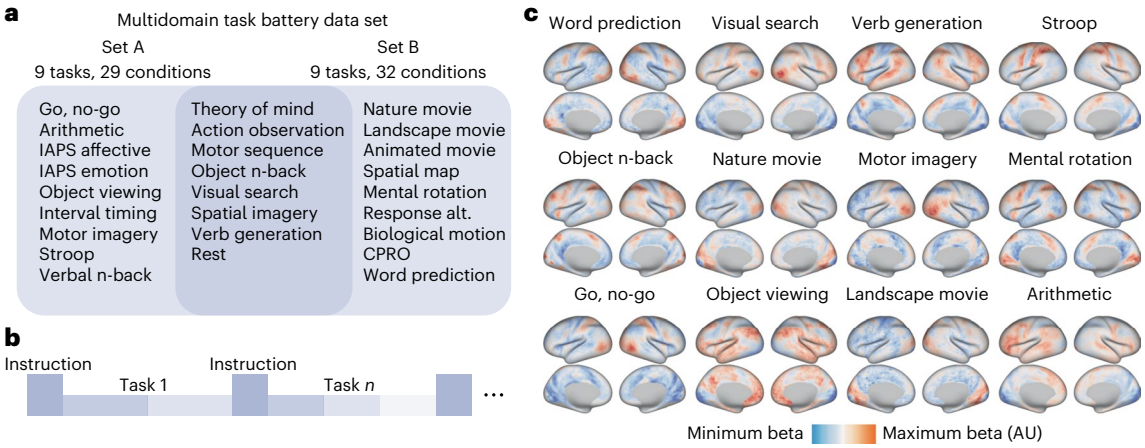


Fig. 2 | Leveraging the MDTB dataset to investigate multitask representations. a, The MDTB dataset consists of 26 distinct tasks with up to 45 unique task conditions per individual and was previously made publicly available[25]. The tasks were split across two sets. Every individual performed each set of tasks twice across different fMRI sessions. (IAPS = International Affective Picture System; CPRO = Concrete Permuted Rule Operations; alt = alternatives). **b**, Task blocks were interleaved across each fMRI session. For each block, instructions were presented for 5 s, followed by a task that was performed continuously for 30 s until the subsequent block. **c**, Whole-cortex group-level activation maps for 12 of 26 cognitive tasks (see Extended Data Fig. 1 for all task activation maps); AU, arbitrary units.

Results

Analytic approach to studying multitask representations

RSA was central to our data analytic approach (Fig. 1)[4]. RSA approximates the representational geometry of a set of multivariate task activations by comparing the similarity of activation vectors across different conditions. By performing RSA on each brain region, we could produce representational similarity matrices (RSMs) for every brain region (Fig. 1a). We used the Glasser parcellation[24], which provides an atlas of 360 cortical brain regions, to measure the RSMs for each cortical region using the vertices (surface voxels) within each predefined region. Critically, RSMs were calculated for each individual, ensuring that fine-grained (vertex-wise) representational geometries would not be lost through cross-individual averaging of activations at the vertex level (Extended Data Fig. 9). These region-specific RSMs enabled us to perform a variety of new analyses, such as comparing RSMs of cortical areas (that is, the representational alignment (RA); Fig. 1b), quantifying the representational dimensionality across cortical areas (Fig. 1c) and identifying the conditions by which computational models can reproduce brain-like representations (Fig. 1d).

A publicly available dataset with 26 cognitive tasks

Characterizing multitask representations across the cortex required a dataset with many tasks per individual. We used the publicly available multidomain task battery (MDTB) human fMRI dataset with 26 cognitive tasks, comprising up to 45 unique task conditions[18] (Fig. 2a). Data were collected across four sessions, enabling within-participant cross-validation analyses across task conditions (see Methods for a list of prior studies using this dataset).

Measuring the RA across brain areas

We used the multitask RSM of each cortical area to investigate how representations varied across the cortex (Fig. 3a). To characterize the regional variation of multitask representations, RA was computed as the cosine similarity between two regions' RSMs (Fig. 3b). This produced a whole-cortex RA matrix (Fig. 3c). Intuitively, RA would be high between two visual regions that both have discriminable visual task representations (that is, highly decodable) but non-discriminable motor task representations. One common way to characterize the large-scale organization of the cortex is through RSFC[25,26]. We found that the similarity of multitask RA and RSFC was moderate ($\rho = 0.37$, $P < 0.0001$; Fig. 3d). This suggested that while the RA matrix appeared to recover overall aspects of intrinsic RSFC organization, the RA matrix offered unique information from RSFC. Critically, characterizing multitask RA in relation to the well-established RSFC organization literature[12,25–27] would clarify how the brain's representational capacity emerges from its resting-state organization.

RA relates to the brain's intrinsic organization

We next characterized RA in the context of the well-established RSFC. RSFC can be used to assign each cortical area to a functional network[25,26,28], and this network organization strongly relates to cognitive task activation patterns[11]. How does the alignment of task representations correspond with the functional networks of the human brain[29]? To address this, we measured the segregation of RA in relation to the segregation of RSFC. Conceptually, segregation measures how clustered a network's (for example, default mode network) representations/FC are in relation to other networks of the brain (Fig. 3g,h)[30]. Thus, if a visual region's representations are highly unique to the visual network, then its segregation would be high. Unimodal regions had significantly higher RA segregation than transmodal regions ($t_{358} = 12.99$, $P < 10 \times 10^{-31}$; Extended Data Fig. 2d). Despite not observing a significant difference in segregation between RA and RSFC across the whole brain ($t_{358} = -1.24$, $P = 0.22$), RA had exaggerated differences in segregation by network; unimodal networks had higher segregation for RA than RSFC ($t_{112} = -3.33$, $P = 0.001$), and transmodal networks had lower segregation for RA ($t_{244} = -4.24$, $P < 10 \times 10^{-4}$; Fig. 3f,g,h). Thus, representations in unimodal regions were more isolated, while transmodal regions shared their representations more broadly with other networks and systems.

Prior work has revealed that the human brain's functional hierarchy can be proxied through identifying topographic cortical gradients using task-free MRI[12,31]. Specifically, extracting the first principal component of the RSFC matrix produces a unimodal–transmodal hierarchy[12] (Fig. 3i), which is highly correlated with the cortical T1-weighted/T2-weighted (T1w/T2w) map (an MRI-contrast correlate of intracortical myelin)[32]. RA segregation was strongly associated with the RSFC principal gradient ($r = 0.39$, $P_{\text{non-parametric}} < 0.001$; Fig. 3i,j)[12] and the T1w/T2w myelin map ($r = 0.36$, $P_{\text{non-parametric}} < 0.001$; Extended Data Fig. 2f,g). While prior studies have studied hierarchical organization in humans using resting-state, transcriptomic or structural imaging techniques, these findings situate multitask representational topography within the intrinsic hierarchical organization.

https://doi.org/10.1038/s41593-022-01224-0

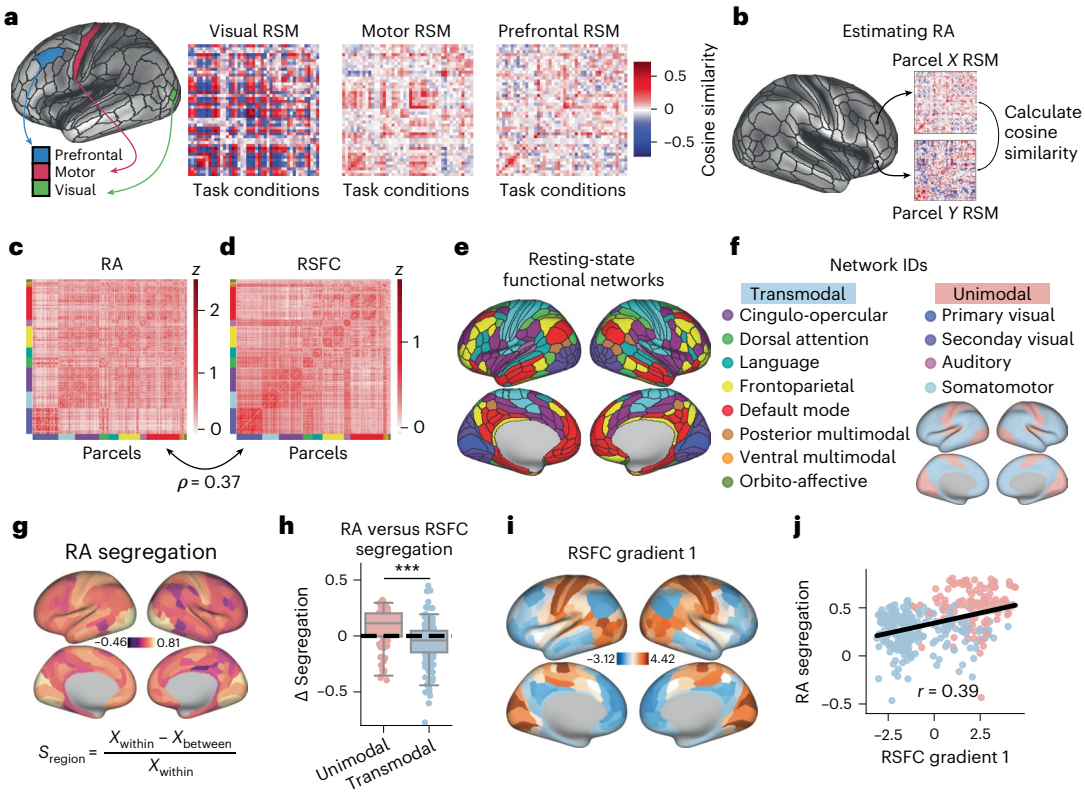


Fig. 3 | Cortical organization of multitask representations. a, Three example RSMs taken from visual, motor and prefrontal areas. RSMs consisted of 45 task conditions that were cross-validated across imaging runs. **b**, RA between pairs of cortical areas is quantified by measuring the cosine similarity between their RSMs. **c**,**d**, RA (**c**) and RSFC (**d**) for all pairs of cortical areas. **e**,**f**, Previously identified RSFC networks. **g**, Segregation (S_{region}) of the RA and RSFC matrices, defined as the difference of within-network (X_{within}) and between-network ($X_{between}$) values divided by within-network values. **h**, Unimodal regions ($n = 114$) have greater segregation for RA than RSFC, and transmodal regions ($n = 246$) have less segregation for RA than RSFC. This was despite no difference in overall segregation between RA and RSFC. Data were analyzed by two-sided t-test ($P < 10 \times 10^{-34}$). **i**,**j**, The cortical topography of RA segregation is correlated with the RSFC principal gradient, a proxy of intrinsic hierarchy. Blue and red dots reflect transmodal/unimodal regions, respectively. Box plot bounds define the first and third quartiles of the distribution, box whiskers indicate the 95% confidence interval, and the center line indicates the median.

Hierarchical organization of representational dimensionality

Recent studies have investigated task representational dimensionality during task performance[33,34]. Note that representational dimensionality refers to the dimensionality of the task space rather than the neural space. However, most prior studies evaluated the representational dimensionality of either a specific task (for example, perceptual task) or within a specific brain region, such as the prefrontal cortex[34]. Here, we evaluated the representational dimensionality across many tasks and across the entire cortex.

We measured the representational dimensionality by measuring the participation ratio of the RSM for each cortical region (Fig. 4a)[35–37]. Intuitively, the participation ratio is a statistical estimate of dimensionality and is related to the flatness of the RSM's eigenspectrum. An implication of this is that regions with low representational dimensionality have stereotyped task responses that are largely shared across many tasks (that is, task responses coexist in a low-dimensional linear subspace). We also estimated the multitask (45-way) decoding as a complementary measure of dimensionality because decoding has been previously used to estimate dimensionality[34].

Representational dimensionality and the multitask decoding accuracy were highly correlated across cortical areas, indicating the reliability of these measures ($r = 0.94$, $P_{non\text{-}parametric} < 0.001$; Fig. 4b,c). Next, we addressed whether representational dimensionality was related to intrinsic hierarchical organization. Indeed, representational dimensionality and multitask decoding were highly correlated with the RSFC principal gradient ($r = 0.49$, $P_{non\text{-}parametric} < 0.001$) and T1w/T2w myelin map ($r = 0.41$, $P_{non\text{-}parametric} < 0.001$; Fig. 4d and Extended Data Fig. 3). This illustrated that like representational segregation, representational dimensionality was also higher in unimodal regions than in transmodal regions ($t_{358} = 6.54$, $P < 10 \times 10^{-9}$; Fig. 4d). These findings are consistent with previous studies that found that higher-order association areas typically have relatively low decoding accuracies, even for tasks that heavily involve those regions[38].

As control analyses, we tested for the effect of parcel size and number of task conditions. After conditioning on parcel size as a covariate using linear regression, the associations between representational dimensionality and hierarchy remained significant (Extended Data Fig. 3c). We tested for robustness to the number of task conditions by randomly sampling subsets of task conditions and found that the hierarchical differences in representational dimensionality were robust with at least 10 task conditions (Extended Data Fig. 4a–c).

Compression then expansion of task representations

We next mapped the axis of greatest RA variation, elucidating how representations vary across the cortex. We performed a principal-component analysis (PCA) to extract the first principal gradient of corticocortical RA. In contrast to the unimodal–transmodal principal gradient exhibited from RSFC (Fig. 3i), RA's principal gradient exhibited a sensory-to-motor gradient and explained 29% of variance in the RA matrix (Fig. 5a,b). This was more than two times the variance relative to the second RA principal component (13%). This analysis was corroborated using non-negative matrix factorization (Extended Data Fig. 10).

https://doi.org/10.1038/s41593-022-01224-0

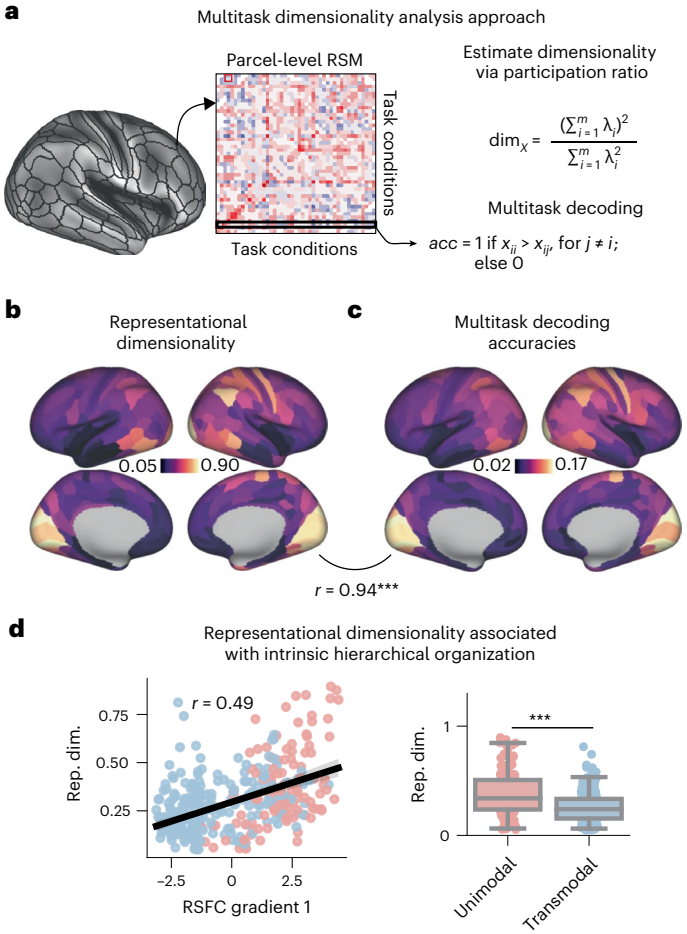


Fig. 4 | The representational dimensionality of task activations follows hierarchical organization. a, We estimated multitask representational dimensionality using two approaches: (1) estimating the dimensionality of the cross-validated RSM, where λ_i refers to the ith eigenvalue of the RSM, and (2) calculating the within-participant decoding accuracy (acc) across all possible task conditions ($n = 45$), with cross-validation across sessions (split-half). **b**,**c**, The representational dimensionality (**b**) and multitask decoding accuracy (**c**) of each cortical parcel. **d**, Across the cortex ($n = 360$), representational dimensionality (rep. dim.) was positively correlated with the first principal gradient of RSFC, with unimodal regions containing higher representational dimensionality than transmodal regions (two-sided unpaired t-test; $P < 10 \times 10^{-9}$). Red dots reflect unimodal regions ($n = 114$), and blue dots reflect transmodal regions ($n = 246$; see Extended Data Fig. 3 for equivalent plots for the decoding approach). Box plot bounds define the first and third quartiles of the distribution, box whiskers indicate the 95% confidence interval, and the center line indicates the median; ***$P = <0.0001$.

This RA sensory-to-motor gradient was highly correlated with the second principal component of RSFC that also reflects a sensory-to-motor cortical organization ($r = 0.59$, $P_{\text{non-parametric}} < 0.001$; Fig. 5c). Thus, the axis of greatest RA variation places sensory and motor area representations on opposite ends (Fig. 5d).

To understand how representations transform from sensory to motor ends, we evaluated the representational dimensionality across the sensory-to-motor hierarchy. We fit several competing statistical models to evaluate how representational dimensionality (dependent variable) changed as a function of the sensory–motor gradient (independent variable), including linear, quadratic and exponential decay models. Using the Akaike/Bayesian information criterion (Extended Data Fig. 5), a convex quadratic fit best explained representational dimensionality as a function of the sensory-to-motor hierarchy (Fig. 5e). We corroborated this result using the RA principal gradient rather than the second RSFC sensory–motor gradient, illustrating the robustness of this phenomenon (Fig. 5f). The quadratic dependence was robust to subsampling the task conditions (Extended Data Fig. 4d–f). This analysis revealed that representational dimensionality compressed then expanded across the sensory-to-motor hierarchy.

To verify compression then expansion from sensory-to-motor systems, we grouped regions by cortical systems. Both sensory and motor systems had greater dimensionality than association regions (sensory versus association, $t_{319} = 7.22$, $P < 10 \times 10^{-11}$; motor versus association, $t_{283} = 2.59$, $P = 0.01$; Fig. 5g). To further establish compression then expansion along the RA hierarchy, we created 10 bins of brain regions sorted by the loadings of the RA principal gradient (Fig. 5h). We fitted a continuous piecewise regression model, varying the breakpoint between the two line segments at every intermediary bin. We selected the model with highest R^2, which resulted in the piecewise model with a breakpoint at bin 3 (Methods). We then tested the statistical significance for the coefficients of the piecewise regression (for a negative slope from bin 1 to 3 and a positive slope from bin 3 to 10). Indeed, a negative slope from bin 1 to 3 ($t_7 = -12.51$, $P < 0.001$) and a positive slope from bin 3 to 10 (two sided; $t_7 = 2.55$, $P = 0.038$) confirmed the compression then expansion of dimensionality across the sensory-to-motor hierarchy.

Compression then expansion of task representations in ANNs

We next investigated the computational mechanisms that produced the compression then expansion of representational dimensionality observed in fMRI data. Interestingly, two recent ANN studies observed expansion then compression of representational dimensionality when trained on image recognition tasks[20,21]. Therefore, we first asked how the compression-then-expansion phenomena of representational dimensionality emerges in ANNs. We used multilayer, feedforward linear ANNs with tied weights to study how fMRI activations in visual areas were successively transformed into motor activations under different learning regimes (ANNs with untied weights are presented in Extended Data Fig. 7 and with other optimization parameters in Extended Data Fig. 8).

Prior research has shown that small alterations to weight initialization parameters can greatly impact the learned hidden representations in ANNs[22,23]. Specifically, those studies found that during a 'rich' training regime (in which network initializations had small weight variances), ANNs learned lower-dimensional and structured representations. By contrast, during a 'lazy' training regime (large variance weight initializations), task performance was achieved by randomly projecting input features into a high-dimensional embedding in hidden layers. Therefore, we examined representations as a function of the rich and lazy training regimen.

Using the sensory-to-motor RSFC gradient 2 (Fig. 5c), we selected two brain regions on opposite ends of this axis (that is, lowest and highest loadings). This resulted in a visual and a motor parcel (Fig. 6a). Note that these sensory and motor parcels had highly similar RSMs to primary visual cortex and primary motor cortex, respectively, suggesting that the gradient-selected parcels were appropriate to model early sensory to late motor transformations in data (Extended Data Fig. 6a–c). We took the visual parcel's fMRI activations of each of the 45 task conditions (that is, a 45 task condition × vertex-wise activations input matrix) and trained an ANN with 10 hidden layers using weight initializations with different standard deviations to predict motor activations (Fig. 6b). We trained 20 random initializations for each weight initialization (ranging from 0.2 to 2.0 in increments of 0.2).

After ANN training, we measured the representational dimensionality of each ANN's hidden layer. Rich training regimes (for example, weight initialization s.d. = 0.2; Methods) showed compression then expansion across layers, consistent with empirical data (Fig. 6c). Similar to the empirical data, we fit a second-order polynomial regression to model dimensionality as a function of layer depth. In the rich regime, in

https://doi.org/10.1038/s41593-022-01224-0

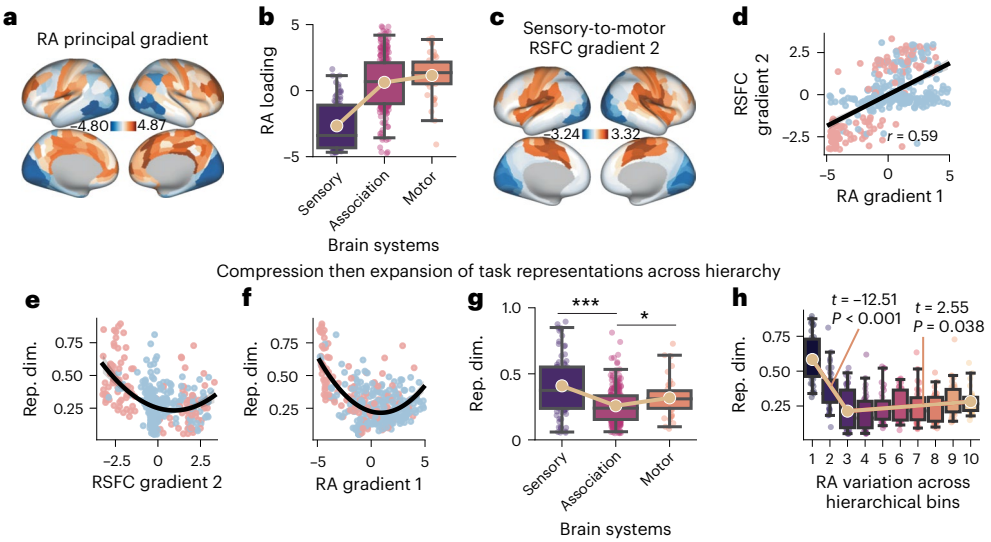


Fig. 5 | Principal component of the RA matrix reveals a sensory-to-motor gradient that compresses then expands task representations. a, Similar to estimating intrinsic RSFC gradients, we extracted the first principal component of the cortical RA matrix. **b**, The average component loadings (for each region) averaged across sensory ($n = 75$), association ($n = 246$) and motor ($n = 39$) regions in the brain. **c**, The RA principal gradient showed striking similarity to the second RSFC (that is, sensory–motor) gradient. **d**, Correlation ($n = 360$) between the sensory–motor RSFC gradient and the principal RA gradient. **e**,**f**, We plotted the representational dimensionality against both the RSFC sensory–motor gradient (**e**) and RA principal gradient (**f**), finding that a second-order convex polynomial model was a better fit than a first-order polynomial model and an exponential decay model (Extended Data Fig. 5). This suggested that representational dimensionality compressed then expanded across the sensory–motor hierarchy. **g**, Same as in **e** and **f** but after grouping together sensory (visual and auditory network; $n = 75$), motor (somatomotor network; $n = 39$) and association (all other networks; $n = 246$) parcels according to network affiliation. Data were analyzed by two-sided, two-sample t-test with a Bonferroni correction. **h**, Same as in **e** and **f** but after placing regions into 10 bins ($n = 36$ each) sorted according to the RA hierarchy (that is, binning regions together with similar loadings). Using a continuous piecewise linear regression, a significant negative-then-positive slope best accounted for dimensionality, consistent with compression then expansion of dimensionality. Data were analyzed by two-sided t-test. Box plot bounds define the first and third quartiles of the distribution, box whiskers indicate the 95% confidence interval, and the center line indicates the median. Error bands reflect a 95% confidence interval in **d**–**f**; $^{***}P < 0.0001$; $^{*}P < 0.05$.

particular for weight initializations starting at an s.d. of 0.2, the quadratic fit was convex and had higher R^2 values (Fig. 6d). Thus, rich regimes produced compression then expansion of hierarchical representations.

The rich training regime could produce a compression then expansion of learned representations in ANNs, reproducing the empirical brain data and in contrast to what is typically found in task-optimized ANNs[20,21]. But do the compressed-then-expanded representations exhibit greater similarity to empirical brain representations? Hidden representations learned in the rich regime (that is, <1.0 s.d. weight initializations) were more similar to those found in empirical data (rich, cosine = 0.42; lazy, cosine = 0.37; rich versus lazy, $t_{198} = 15.28$, $P < 10 \times 10^{-34}$; Fig. 6e). We then partitioned brain parcels into 10 bins of 36 parcels and sorted them according to their loading relative to the sensory–motor RSFC gradient (Fig. 6g). We correlated the RSMs of each bin with each ANN layer according to depth (for example, similarity of RSMs for ANN layer i with fMRI bin i). The later 8/10 bins/layers had higher similarity in the rich regime (for 8/10, false discovery rate (FDR)-corrected P value of 10^{-16}). While the first two fMRI bins had greater correspondence with the lazy learning regime, these first two bins primarily consisted of visual areas. Empirically, visual areas contained high-dimensional representations. Because the lazy learning embeds input features in a high-dimensional space, the higher similarity of lazily trained ANNs with visual regions was unsurprising. Thus, with the exception of early visual areas, richly trained ANNs have greater correspondence with fMRI data in terms of both representational dimensionality and content.

Richly trained ANNs learn structured transformations

Having modeled the successive transformation of fMRI activations from visual to motor regions in feedforward ANNs, we evaluated the properties of ANNs that contributed to better correspondence with brain data. First, we characterized the structure of representations that emerged in ANNs trained under different learning regimes. This was done through a similar analytic strategy as in empirical data. For each layer, we computed its RSM using each of the 45 task activation patterns and computed the cosine similarity of that layer's RSM with the RSMs from all other layers (Fig. 7a). This produced a layer-by-layer RA matrix (Fig. 7d). ANNs trained in the rich regime learned structured representations that were consistent with structured representations in data; adjacent layers had high RA to each other, but distal layers had low RA to each other (Fig. 7b–d). We quantified this by calculating the mean of the RA matrix for each weight initialization (Fig. 7e) and the dimensionality of the RA matrix (Fig. 7f,g). The higher (lazier) the weight initialization, the greater the overall RA across layers and the lower the dimensionality (rich versus lazy cosine difference = −0.12, $t_{18} = -124.70$, $P < 10 \times 10^{-28}$; rich versus lazy variance explained by the first principal component = −15.52%, $t_{18} = -99.10$, $P < 10 \times 10^{-26}$). In contrast to the rich training regime, the lazy regime had nearly no meaningful transformations in the hidden layers for weight initializations, with s.d. > 1.2; outputs were generated from the readout weights only (Fig. 7d). The ANN's internal representations were a direct byproduct of its learned (via training) connectivity structure. Analysis of the optimized network weights revealed that richly trained networks exhibited small-world network structure and low-dimensional connectivity (Extended Data Fig. 6).

Transformational trajectories from visual to motor areas

To provide an intuition of the transformations in rich and lazy ANNs in the task-state space, we characterized the transformational trajectories from visual to motor representations. This involved plotting ANN representations in a two-dimensional space. The y axis reflected the alignment (inner product) with visual (input) RSM, and the x axis

https://doi.org/10.1038/s41593-022-01224-0

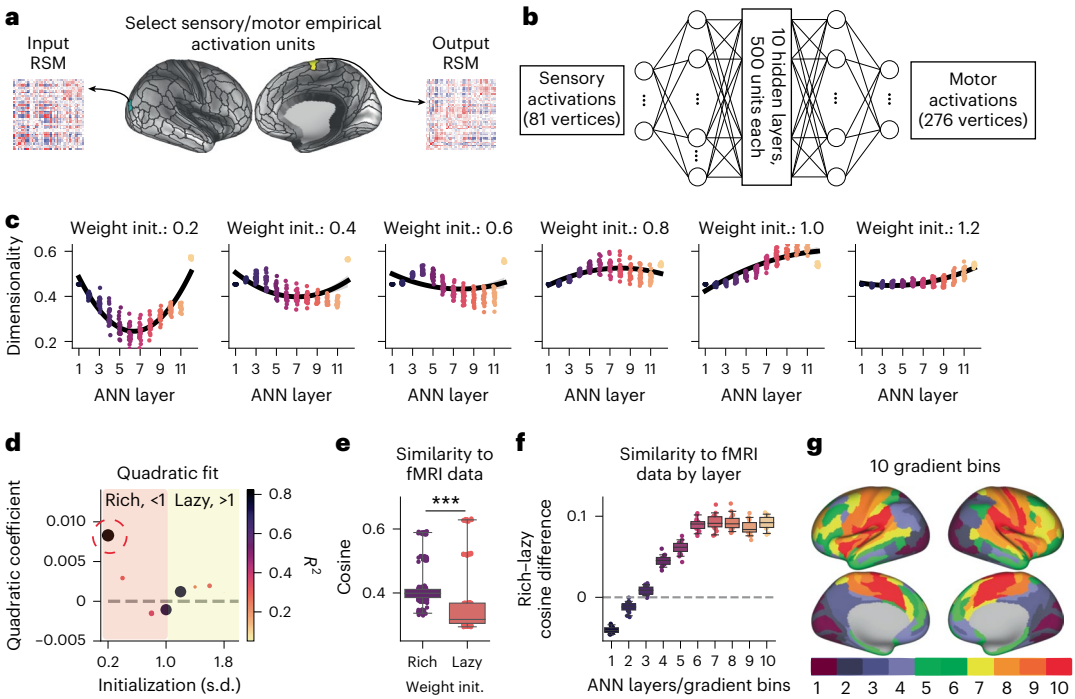


Fig. 6 | Multitask representations in the human cortex were consistent with ANN representations trained in a rich regime. a, We identified the brain parcels at the bottom (visual parcel) and top (motor output parcel) of the sensory–motor RSFC gradient 2 and extracted their vertex-wise task activation patterns. **b**, We trained a feedforward ANN with 10 hidden layers to predict task activations in the motor parcel using vertex-wise activations from the visual parcel. **c**, The representational dimensionality (participation ratio of RSMs) of each layer with different weight initialization (weight init.) standard deviations. We observed compression then expansion of representations in rich training regimes. **d**, We fit a second-order polynomial regression to the dimensionality across weight initializations. We plotted the quadratic coefficient (coeff.; positive for convex) and the overall R^2 fit to assess how dimensionality changed across ANN layers. R^2 peaked during rich training regimes and was consistent with a convex fit. **e**, We compared the overall similarity of the RSMs of ANNs at each layer with the RSMs for every region in the brain, finding stronger similarity when the ANN was trained in the rich regime (<1.0 s.d. initialization); $n = 20$ ANN initializations. Data were analyzed by two-sided t-test ($P < 10 \times 10^{-34}$). **f**, We compared the RSMs of ANNs to the empirical brain RSMs at each bin along the sensory–motor gradient for rich (<1.0) and lazy (>1.0) learning regimes; $n = 20$ ANN initializations. **g**, Empirical gradient bins were defined by partitioning the sensory–motor RSFC gradient 2 into 10 distinct sets of regions sorted by their gradient loadings. Box plot bounds define the first and third quartiles of the distribution, box whiskers indicate the 95% confidence interval, and the center line indicates the median.

reflected the alignment with motor (output) RSM. While a linear transformation would map visual to motor representations directly (Fig. 8a), we hypothesized that compression then expansion would occur by first compressing representations along the visual axis and then expanding along the motor axis (Fig. 8a). This is in contrast to the alternative, where motor representations first expand with minimal loss of visual representations (Fig. 8a). In agreement with this theory, richly trained ANNs first compressed along the visual axis, followed by growth along the motor axis (Fig. 8b). This is consistent with the notion that higher-order brain areas (that is, similar to intermediate layers in an ANN) contain distinct representations from input (visual) and output (motor) representations, are low dimensional and integrate input and output representations. By contrast, lazy ANN representations maintained high similarity to visual input representations, with visual-to-motor representational transformations primarily implemented in the readout weights.

Discussion

Using RSA-based techniques, we mapped the multitask representational organization of the human cortex. Representations in sensory and motor cortices were more isolated from the rest of the cortex yet had higher dimensionality. By contrast, representations in association regions were lower dimensional but were situated between sensory and motor representations. This revealed a representational hierarchy that compressed then expanded from sensory to association to motor areas. To explore the computational mechanisms of hierarchical representations in the brain, we used feedforward ANNs to study how representations compressed then expanded from input to output. During a rich training regime, ANNs learned structured and hierarchical representations that (1) compressed then expanded representations and (2) had greater similarity to representations found in fMRI data. Further analysis of the ANN revealed that this training regime produced low-dimensional weights with a heavy-tailed distribution, consistent with empirical brain networks[39]. Together, these findings characterize the topographic organization of multitask representations in the cortex and provide a framework for understanding how brain-like representations emerge in ANNs.

We combined multitask analyses with RSA. While RSA has been widely used since its original inception more than a decade ago[4], applications of RSA have typically been limited to the sensory domain (that is, where the rows and columns of RSMs are sensory stimuli). The combined approach of leveraging a multitask design with RSA provides a unique opportunity to investigate how different brain regions represent diverse task information. One recent study collected many tasks per participant and, by using individualized encoding models, identified clusters of tasks in a latent cognitive space and how task specialization emerged across the cortex[17]. Here, we expand on that study by explicitly quantifying the transformation of representations between cortical areas, investigating how these variations relate to hierarchical organization and addressing how the dimensionality of task representations changes across this hierarchy.

Our findings of cortical gradients in task representations adds to a growing literature identifying hierarchy as a fundamental principle

https://doi.org/10.1038/s41593-022-01224-0

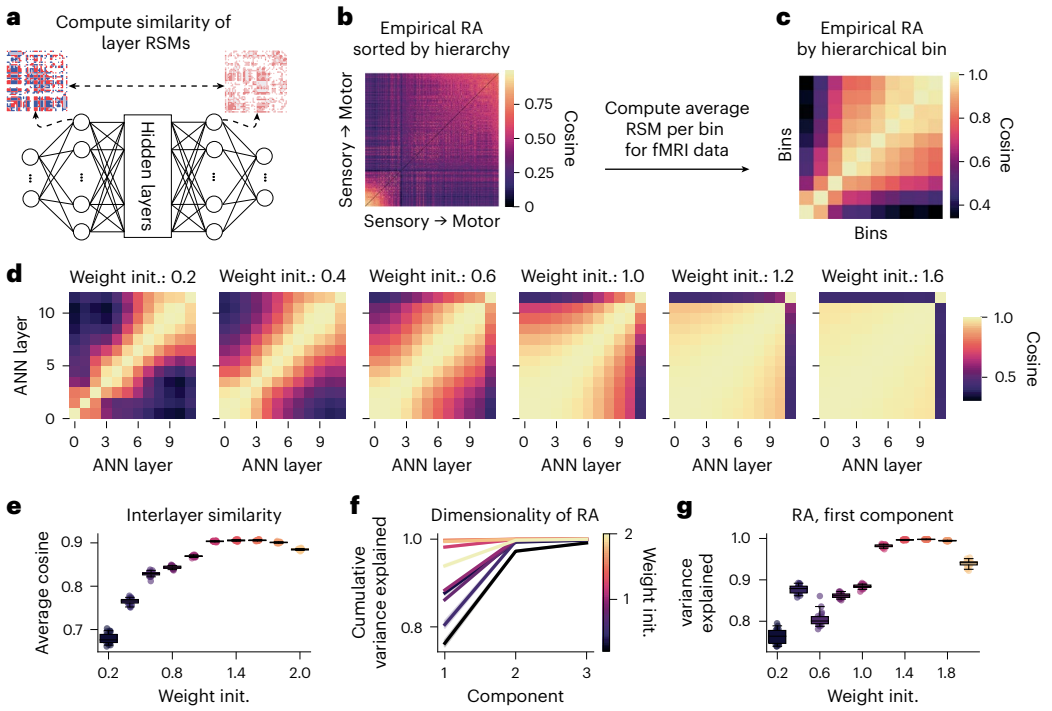


Fig. 7 | Analysis of the ANN revealed that richly trained ANNs learn diverse and structured representations consistent with empirical data. a, We computed the RA between all layers by computing the cosine similarity between the RSMs of each hidden layer. **b**,**c**, For comparison, we sorted the empirical fMRI RA by the RSFC sensory–motor (second) gradient (**b**) and downsampled it into 10 discrete bins for comparison with the ANN analysis (**c**). **d**, The RA for ANNs by layer across weight initializations. ANNs trained in the rich regime (for example, weight initializations of <1) learned differentiated and structured representations. By contrast, ANNs trained in the lazy regime largely produced impoverished representations that only transformed sensory representations in the final layer. **e**, The average cosine similarity of each RA matrix by weight initialization ($n = 20$ per weight initialization). **f**, Cumulative variance explained plot of the first three components of the RA matrix under different weight initializations. **g**, Variance explained of only the first principal component of the RA matrix ($n = 20$), which captures RA dimensionality; the larger the variance explained, the lower the dimensionality. Box plot bounds define the first and third quartiles of the distribution, box whiskers indicate the 95% confidence interval, and the center line indicates the median.

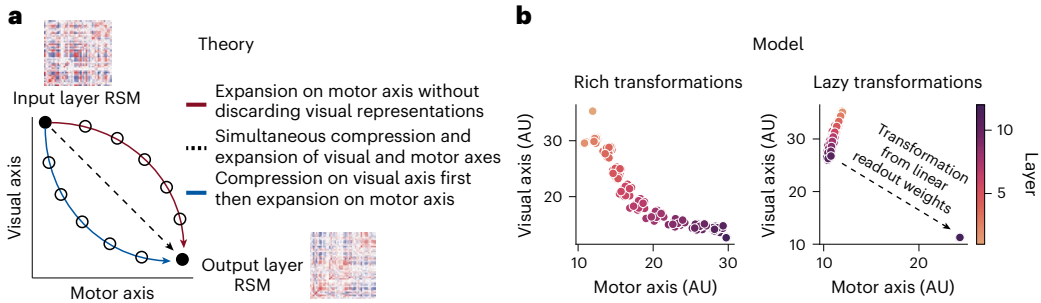


Fig. 8 | Trajectories of representational transformations from visual to motor content. a, A theory of how representations transform across layers/brain areas from visual to motor representations. Axes reflect the similarity (computed as the inner product) to the visual input region's RSM (y axis) and the motor output region's RSM (x axis). Hidden layer representations can then be plotted along these two dimensions by calculating the inner product between the sensory and motor RSMs. **b**, We plotted the ANN's internal representations along these two dimensions and found that rich representations are consistent with compression first along the visual axis and then expansion along the motor axis. By contrast, lazy ANNs preserve visual representations in hidden layers until the final readout weights transform visual into motor representations. Note that the y and x axes are not necessarily orthogonal and are plotted as such for visualization purposes. Each dot in the scatter plots reflects a different ANN initialization and layer.

of cortical organization. Early seminal work using tract-tracing techniques revealed hierarchical connectivity organization in the macaque visual cortex[40]. More recent work has shown that such hierarchical organization can be studied in humans in vivo with MRI and electrophysiology. These studies have focused on identifying structural[32], transcriptomic[31], RSFC[12] and intrinsic timescale signatures of hierarchical organization[41,42]. Most of these hierarchical descriptions used task-free MRI data, and here, we establish an overarching link that bridges multitask representations with fundamental hierarchical organization. Other studies that evaluate the role of connectivity organization[43,44] and shared multitask dynamics[45] have also identified key hub regions in the association cortex. This is consistent with our finding that association areas contain integrative representations that link the sensory and motor representations lying on opposing ends of the sensory-to-motor axis. Future studies can explore how specializations of the association cortex, in long-range anatomical connectivity and local microcircuitry, contribute to the formation of low-dimensional integrative representations.

Compression then expansion of representations emerged across the sensory-to-motor hierarchy in brain data. Surprisingly, this contrasts with task-optimized ANNs that typically exhibit expansion then compression of representational dimensionality across hidden layers[20,21]. Algorithmically, representational expansion then compression in ANNs affords high task performance due to the projection of input features into a high-dimensional embedding in the hidden layers. These high-dimensional representations subsequently allow for easy selection of the few features that are useful for task performance. What might be the algorithmic purpose of the compression then expansion of representations observed in the human brain? Recent work at the intersection of machine learning and neuroscience found that, in contrast to projecting input features into a high-dimensional space, lower-dimensional factorized representations (that is, 'abstract' or 'disentangled') may be useful for generalization because they can easily be recycled in novel contexts[46–48]. Moreover, these lower-dimensional factorized representations can be learned by ANNs through a rich training regime[47]. Here, we expanded on this prior work by (1) leveraging a multitask dataset to demonstrate the generality of these principles (rather than manipulating distinct context representations within a single task paradigm) and (2) revealing the organization of representation transformations across the cortex. Although our findings suggest that the low-dimensional association cortex representations are shared across multiple tasks (which likely aid in out-of-task generalization), the current dataset is unable to evaluate how shared components are used to generalize to novel tasks. This is due to the lack of systematic factorization of task components in this multitask setting, which is required to test whether factorized components can be compositionally reused. Therefore, it will be important for future studies to provide a unified understanding of the contribution of low-dimensional representations for task generalization performance[49].

Our computational modeling results provide a parsimonious framework to study representational transformations in relation to empirical data. There are multiple directions in modeling and analytics that future studies can explore. First, we used a simple feedforward ANN, motivated by our findings of a dominant sensory-to-motor gradient. Future models can examine the impact of more complex and recurrent ANN architectures of internal representations[50–52]. We found that representations depended strongly on the training regime, which we controlled by weight initialization following prior literature[22,23]. Future modeling should explore alternative training methods for ANNs to examine how they alter the similarity to brain representations. Finally, future studies should examine the metrics used to quantify the similarity between empirical and model representations. For instance, inherent constraints on RSMs can be used to define alternative measures of RA[53,54]. Therefore, it will be important for future work to explore the space of biologically relevant strategies that produce feature-rich hierarchical representations in models that can be quantitatively related to neural datasets.

In conclusion, we characterized the multitask representational geometry and topography across the human cortical hierarchy and provide insight into the mechanisms that produce similar representations in ANNs. Overall, analysis of the task representational hierarchy revealed a sensory-to-motor gradient that compressed then expanded task representations. Subsequent modeling of these task activations in ANNs revealed that a rich training regime can reproduce representations that are consistent with brain data. This finding provides a framework to explore how to build ANNs that learn task representations in a brain-like manner. We expect these findings to spur new investigations into how the study of multitask representations in the brain can inform new models of multitask performance in machine learning models.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-022-01224-0.

References

1. Genon, S., Reid, A., Langner, R., Amunts, K. & Eickhoff, S. B. How to characterize the function of a brain region. *Trends Cogn. Sci.* **22**, 350–364 (2018).
2. Poldrack, R. A. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* **72**, 692–697 (2011).
3. Gallant, J., Nishimoto, S., Naslaris, T. & Wu, M. C. K. In *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging* (eds Kriegeskort N. & Krieman G.) Ch. 6 (The MIT Press, 2011).
4. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
5. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
6. Kanwisher, N. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl Acad. Sci. USA* **107**, 11163–11170 (2010).
7. Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **7**, 415–423 (2003).
8. Wandell, B. A. & Winawer, J. Computational neuroimaging and population receptive fields. *Trends Cogn. Sci.* **19**, 349–357 (2015).
9. Arbuckle, S. A. et al. Structure of population activity in primary motor cortex for single finger flexion and extension. *J. Neurosci.* **40**, 9210–9223 (2020).
10. Yeo, B. T. T. et al. Functional specialization and flexibility in human association cortex. *Cereb. Cortex* **25**, 3654–3672 (2015).
11. Smith, S. M. et al. Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl Acad. Sci. USA* **106**, 13040–13045 (2009).
12. Margulies, D. S. et al. Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proc. Natl Acad. Sci. USA* **113**, 12574–12579 (2016).
13. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
14. Huth, A. G., Heer, W. A. D., Griffiths, T. L., Theunissen, F. E. & Jack, L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
15. Naselaris, T., Allen, E. & Kay, K. Extensive sampling for complete models of individual brains. *Curr. Opin. Behav. Sci.* **40**, 45–51 (2021).
16. Yang, G. R., Cole, M. W. & Rajan, K. How to study the neural mechanisms of multiple tasks. *Curr. Opin. Behav. Sci.* **29**, 134–143 (2019).
17. Nakai, T. & Nishimoto, S. Quantitative models reveal the organization of diverse cognitive functions in the brain. *Nat. Commun.* **11**, 1142 (2020).
18. King, M., Hernandez-Castillo, C. R., Poldrack, R. A., Ivry, R. B. & Diedrichsen, J. Functional boundaries in the human cerebellum revealed by a multi-domain task battery. *Nat. Neurosci.* **22**, 1371–1378 (2019).
19. Bernhardt, B. C., Smallwood, J., Keilholz, S. & Margulies, D. S. Gradients in brain organization. *NeuroImage* **251**, 118987 (2022).
20. Ansuini, A., Laio, A., Macke, J. H. & Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems* Vol. 32 (Curran Associates, Inc., 2019).
21. Recanatesi, S. et al. Dimensionality compression and expansion in deep neural networks. Preprint at https://doi.org/10.48550/arXiv.1906.00443 (2019).

22. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Rich and lazy learning of task representations in brains and neural networks. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.23.441128 (2021).
23. Woodworth, B. et al. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory* 3635–3673 (PMLR, 2020).
24. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
25. Power, J. D. et al. Functional network organization of the human brain. *Neuron* **72**, 665–678 (2011).
26. Yeo, B. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
27. Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S. & Petersen, S. E. Intrinsic and task-evoked network architectures of the human brain. *Neuron* **83**, 238–251 (2014).
28. Ji, J. L. et al. Mapping the human brain's cortical–subcortical functional network organization. *NeuroImage* **185**, 35–57 (2019).
29. Huntenburg, J. M., Bazin, P. -L. & Margulies, D. S. Large-scale gradients in human cortical organization. *Trends Cogn. Sci.* **22**, 21–31 (2018).
30. Chan, M. Y., Park, D. C., Savalia, N. K., Petersen, S. E. & Wig, G. S. Decreased segregation of brain systems across the healthy adult lifespan. *Proc. Natl Acad. Sci. USA* **111**, E4997–E5006 (2014).
31. Burt, J. B. et al. Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat. Neurosci.* **21**, 1251–1259 (2018).
32. Glasser, M. F. & Van Essen, D. C. Mapping human cortical areas in vivo based on myelin content as revealed by T1-and T2-weighted MRI. *J. Neurosci.* **31**, 11597–11616 (2011).
33. Badre, D., Bhandari, A., Keglovits, H. & Kikumoto, A. The dimensionality of neural representations for control. *Curr. Opin. Behav. Sci.* **38**, 20–28 (2021).
34. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
35. Abbott, L. F., Rajan, K. & Sompolinsky, H. In *The Dynamic Brain: An Exploration of Neuronal Variability and Its Functional Significance* (eds Ding M. & Glanzman D.) 1–16 (Oxford University Press, 2011).
36. Gao, P. et al. A theory of multineuronal dimensionality, dynamics and measurement. Preprint at *bioRxiv* https://doi.org/10.1101/214262 (2017).
37. Recanatesi, S., Ocker, G. K., Buice, M. A. & Shea-Brown, E. Dimensionality in recurrent spiking networks: global trends in activity and local origins in connectivity. *PLoS Comput. Biol.* **15**, e1006446 (2019).
38. Bhandari, A., Gagne, C. & Badre, D. Just above chance: is it harder to decode information from prefrontal cortex hemodynamic activity patterns? *J. Cogn. Neurosci.* **30**, 1473–1498 (2018).
39. Bassett, D. S. & Bullmore, E. Small-world brain networks. *Neuroscientist* **12**, 512–523 (2006).
40. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
41. Honey, C. J. et al. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron* **76**, 423–434 (2012).
42. Ito, T., Hearne, L. J. & Cole, M. W. A cortical hierarchy of localized and distributed processes revealed via dissociation of task activations, connectivity changes, and intrinsic timescales. *NeuroImage* **221**, 117141 (2020).
43. Cole, M. W. et al. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat. Neurosci.* **16**, 1348–1355 (2013).
44. van den Heuvel, M. P. & Sporns, O. Network hubs in the human brain. *Trends Cogn. Sci.* **17**, 683–696 (2013).
45. Shine, J. M. et al. Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nat. Neurosci.* **22**, 289 (2019).
46. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
47. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron* **110**, 1258–1270 (2022).
48. Ito, T. et al. Compositional generalization through abstract representations in human and artificial neural networks. Preprint at https://doi.org/10.48550/arXiv.2209.07431 (2022).
49. Cole, M. W., Laurent, P. & Stocco, A. Rapid instructed task learning: a new window into the human brain's unique capacity for flexible cognitive control. *Cogn. Affect. Behav. Neurosci.* **13**, 1–22 (2012).
50. van Bergen, R. S. & Kriegeskorte, N. Going in circles is the way forward: the role of recurrence in visual inference. *Curr. Opin. Neurobiol.* **65**, 176–193 (2020).
51. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
52. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X. -J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).https://doi.org/10.1038/s41593-018-0310-2
53. Shahbazi, M., Shirali, A., Aghajan, H. & Nili, H. Using distance on the Riemannian manifold to compare representations in brain and in models. *NeuroImage* **239**, 118271 (2021).
54. Williams, A. H., Kunz, E., Kornblith, S. & Linderman, S. W. Generalized shape metrics on neural representations. Preprint at https://doi.org/10.48550/arXiv.2110.14739 (2021).

Methods

MDTB dataset

Portions of this section are paraphrased from the Methods section of the original dataset[18]. We used the publicly available MDTB dataset[18]. Prior studies with this dataset investigated the topographic and functional boundaries of the human cerebellum[18] and the validity of cortical parcellations (or brain atlases) derived from either resting-state or task fMRI[55]. The present study used this dataset to investigate a distinct topic: the structure and organization of multitask representational transformations across cortical hierarchies.

The MDTB dataset contains task fMRI data for 24 individuals collected at Western University (16 females and 8 males; mean age = 23.8 years, s.d. = 2.6; all individuals were right-handed; see ref. [18] for exclusion criteria). All participants gave informed consent under an experimental protocol approved by the institutional review board at Western University, where the dataset was originally collected. Briefly, the MDTB dataset contains 26 unique cognitive tasks with up to 45 different task conditions per participant. Participants first scanned all tasks in set A and returned for a second session to perform tasks in set B (Fig. 2a). Each task set session consisted of two imaging runs. Half of the individuals had sessions separated by 2–3 weeks, while the other half had sessions separated by 1 year. Of the 24 individuals, a separate resting-state fMRI scan was collected for 18 participants. Resting-state FC analyses presented in Fig. 3 were performed using this subset of participants.

A large battery of tasks was selected to broadly recruit cognitive processes from many functional domains (Fig. 2a). Set A consisted of cognitive, motor, affective and social tasks. Set B contained eight tasks that were also included in set A (for example, theory of mind and motor sequence tasks) and nine unique tasks. Both sets contained 17 tasks each. Additional details regarding the experimental tasks and conditions have been previously reported (https://static-content.springer.com/esm/art%3A10.1038%2Fs41593-019-0436-x/MediaObjects/41593_2019_436_MOESM1_ESM.pdf)[18].

Tasks were performed once per imaging session for 35-s blocks. Task blocks began with a 5-s instruction screen followed by 30 s of continuous task performance (Fig. 2b). While most tasks consisted of 10–15 trials per block, the number of trials per task ranged from 1 to 30 (for example, go/no-go task versus movie watching). Eleven of the 26 tasks were passive, meaning no behavioral responses were required (for example, movie watching). For the remaining tasks, responses were made with left, right or both hands using a four-button box. Responses were made with either index or middle fingers of the assigned hand(s). Performing all tasks within a single imaging run for each participant ensured a common baseline between tasks, enabling fine-grained, voxel-wise multitask analyses.

fMRI preprocessing

Resting-state and task-state fMRI data were minimally preprocessed using the Human Connectome Project preprocessing pipeline within the Quantitative Neuroimaging Environment and Toolbox (QuNex, version 0.61.17)[56,57]. The Human Connectome Project preprocessing pipeline consisted of anatomical reconstruction and segmentation, echo-planar imaging (EPI) reconstruction and segmentation, spatial normalization to the MNI152 template and motion correction. Additional nuisance regression was performed on the minimally preprocessed time series. Consistent with previous reports[58], this included six motion parameters, their derivatives and the quadratics of those parameters (24 motion regressors in total). We also removed the mean physiological time series extracted from the white matter and ventricle voxels. We also included the quadratic, derivatives and the derivatives of the quadratic time series of each of the white matter and ventricle time series (eight physiological nuisance signals). This amounted to 32 nuisance parameters in total and was a nuisance regression model that was previously benchmarked[59]. In addition to nuisance regressors, task fMRI data were also modeled with task regressors to extract activation estimates described below.

fMRI task activation estimation

We performed a single-individual task general linear model (GLM) analysis on fMRI task data to estimate vertex-wise surface activations for each task condition on the Connectivity Informatics Technology Initiative file format (CIFTI) grayordinate space[60,61]. We modeled a separate regressor for every trial within each imaging run, similar to a beta series model[61]. The instruction period for each task was not included in the task regressors. This enabled the estimation of specific task conditions within each task block (for example, congruent versus incongruent conditions for the Stroop task). Each regressor (trial) was modeled as a boxcar function from the onset to the offset of the trial (0 s indicates off and 1 s indicates on) and then convolved with the Statistical Parametric Mapping (SPM) canonical hemodynamic response function to account for hemodynamic lags[62]. Activations for a task condition were then obtained by averaging the activation beta coefficients across trials within each imaging run, resulting in one task condition activation per run. Task GLMs were performed using the LinearRegression function within scikit-learn (version 0.23.2) in Python (version 3.8.5).

RSA and RA

We performed a split-half, cross-validated RSA to characterize the geometry of task representations across the cortex[4]. RSA was performed for each parcel in the multimodal (structural, resting-state and task-based MRI) Glasser et al.[24] atlas using vertices within each parcel[24]. We specifically chose this parcellation due to improved delineation of somatotopic and visuotopic areal organization that are not accounted for in atlases defined solely on resting-state fMRI[63]. In particular, the specific features that constituted an improved delineation in the Glasser parcellation (in contrast to other purely data-driven approaches) was the use of prior knowledge (for example, previously published retinotopic maps[64]) to guide the division of areal/parcel boundaries. Critically, RSA was performed at the participant level to ensure that fine-grained, voxel-wise representations were participant specific and that activations would not be averaged across participants. Group averaging was computed after RSMs were constructed for each participant at every parcel. We used all task conditions, resulting in a 45 × 45 RSM. We used cosine similarity to measure the distances between task activations. Despite many alternative metrics[65,66], we specifically chose the cosine similarity because it also takes into account the overall mean magnitude of activation across a set of vertices (in contrast to Pearson correlation). Cross-validation was achieved by measuring the cosine similarity of activation patterns of the first and second imaging sessions (that is, a split-half cross-validation). This was possible because all tasks (in set A and B) were performed in two separate imaging runs. This ensured a non-trivial diagonal element (that is, not equal to 1), which revealed the test–retest reliability (or similarity) of the activation patterns of the same task condition.

Interregional RA was calculated by measuring the cosine similarity of the upper triangle elements (including the diagonal) of two region's RSMs. Related measures have also been previously introduced under the term 'representational connectivity'[4,53,67].

Network segregation

Network segregation for RSFC and multitask RA was measured as the difference between within-network and between-network FC/RA divided by within-network FC/RA[30]. Networks were defined using a previously published whole-brain resting-state network partition[28]. Networks were composed of a non-overlapping set of parcels (or brain regions). Parcels are a collection of non-overlapping vertices. Network segregation[30] was calculated for each region separately using either

the RA or FC matrix. Specifically, the segregation S_{region} of a region was calculated as

$$S_{region} = \frac{X_{within} - X_{between}}{X_{within}},$$

where X_{within} is the within-network FC/RA for the region of interest, and $X_{between}$ is the out-of-network FC/RA.

Representational dimensionality and multitask decoding. Representational dimensionality was measured as the participation ratio of the multitask RSM. Representational dimensionality refers to the dimensionality of the task space rather than the neural space. That is, feature dimensions are defined by task conditions rather than as voxels/neurons. The participation ratio was calculated as

$$\dim_X = \frac{\left(\sum_{i=1}^{m} \lambda_i\right)^2}{\sum_{i=1}^{m} \lambda_i^2},$$

where $\dim_X$ corresponds to the representational dimensionality of region X, and λ_i corresponds to the eigenvalues of the RSM of region X with m eigenvalues. The flatter the eigenspectrum of region X's RSM, the higher the dimensionality. An alternative approach to intuiting this measure is that the dimensionality of task RSMs is inversely related to the amount of variance explained by the first few eigenvectors; the higher the dimensionality, the more eigenvectors are required to explain the same amount of variance. To complement representational dimensionality, we also measured the multitask decodability (45-way classification) of each region using a minimum-distance classifier. We used the cosine angle as our measure of distance and split-half cross-validation. Thus, a successful classification indicated that the diagonal element of a region's cross-validated RSM was greater (that is, smallest distance) than all other off-diagonal elements for a given row (Fig. 4a). We performed additional control analyses to account for parcel size (that is, the number of vertices) when calculating representational dimensionality and multitask decodability. This was performed by conditioning on (regressing out) the number of vertices from each measure using linear regression (regression was performed across parcels). We then recalculated the correlation across brain maps (for example, myelin map versus representational dimensionality) using the residual values (Extended Data Fig. 3b,c).

Gradient analysis

Cortical gradients were calculated using a PCA on parcellated data. Following prior work[12], resting-state FC gradients were extracted by applying PCA on the cortical FC matrix. For RA gradients, PCA was applied on the cortical RA matrix. This means that the covariance matrices of FC and RA were first calculated, and then the eigenvectors of those matrices were extracted. One intuitive way to think about elements in the cov(RA) or cov(RSFC) matrix is to ask, do two regions have similar patterns of RA (or FC/correlations) as the rest of cortex? If they have similar patterns of RA/FC, then those two regions would have similar loadings. Consistent with previous studies[12], matrices were thresholded to include only the top 20% of values before extracting gradients. All correlation-based statistical tests involving gradients (that is, spatial correlations across the cortex) were performed using spatial autocorrelation-preserving permutation tests that generated random surrogate brain maps[68]. We used the BrainSMASH toolbox to generate 1,000 random surrogate brain maps for each cortical map of interest, and non-parametric P values were calculated from the null distribution. Therefore, the lowest precision non-parametric P value we obtained was 0.001.

Testing for compression then expansion in empirical data

Assessing compression then expansion in empirical data involved fitting representational dimensionality to sensory–motor hierarchy loadings using regression models (Fig. 5e,f and Extended Data Fig. 5). We used RSFC sensory–motor gradient 2 loadings (x variable) as the regressor to predict the representational dimensionality of each parcel (y variable). For model adjudication, we used used several competing regression models, including

$$\text{Linear model}: y = \beta_0 + \beta_1 x + \epsilon$$

$$\text{Quadratic model}: y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

$$\text{Exponential decay model}: y(t) = N_0 e^{-\lambda t} + \epsilon$$

where β_i was the fitted coefficient term, and ϵ was the residual error term. For the second-order quadratic model, a positive second-order coefficient indicated a convex quadratic. Selection of the model was based on the lowest Akaike information criterion and Bayesian information criterion (Extended Data Fig. 5).

To further verify compression then expansion across the sensory–motor hierarchy, we binned together groups of 10 bins of 36 parcels according to their RA principal gradient loading (Fig. 5a). To establish compression then expansion along this gradient, we fit a piecewise linear model with the functional form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

We trained a piecewise linear model for every possible breakpoint (that is, where $x_1 < i$ and $x_2 > i$ for every bin i between 1 and 10; eight possible models). Note that x_1 represented values for $x < i$ and 0 otherwise, and x_2 represented values for $x > i$ and 0 otherwise. After identifying the model with the greatest fit evaluated using R^2, which turned out to be the model with the breakpoint at $i = 3$, we tested the statistical significance for the beta coefficients β_1 and β_2, with the hypothesis that they should be negative and positive, respectively. Negative and positive slopes for β_1 and β_2, respectively, would reflect a compression of representational dimensionality from input to the breakpoint and then an expansion from the breakpoint to the output.

ANN modeling and training

We modeled the transformation from visual fMRI activations to motor activations using a linear feedforward ANN. This enabled the characterization of the transformation as a sequence of linear transformations. fMRI activations were selected based on lying on opposite ends of the RSFC sensorimotor gradient (that is, region with the lowest/highest loadings). Input activations were normalized across vertices before training. Inputs and outputs corresponded to the vertex-level fMRI task activations for each parcel. We used the RSFC sensorimotor gradient rather than the task-based RA gradient to avoid any potential confounds of selecting activations from the same task data. The input and output parcels corresponded to parcels 338 and 235 in the Glasser et al.[24] atlas, respectively. We built the ANN with 10 hidden layers with tied weights (500 units per layer), and the ANN was defined by the equations

$$H_1 = XW_{in} + b_{in}$$

$$H_i = H_{i-1}W_{hid} + b_{hid}$$

$$Y = H_n W_{out} + b_{out} + \epsilon$$

where X was the input fMRI activation from the visual parcel, W_{in} mapped vertex activations into the hidden unit space, b_{in} was the input bias, H_i was the hidden unit activations for layer i up to n (that is, 10), W_{hid} and b_{hid} were the weights and biases for the hidden layers, Y was

the predicted motor fMRI activation in the motor parcel, and ϵ was the residual error term. Using tied weights and a linear model reduced the number of free parameters in the model, thereby constraining the solutions and simplifying the model for subsequent analysis. Using tied weights also increased computational efficiency during training. However, we also ran the model without tied weights (where W_{hid} and b_{hid} were distinct for each layer), yielding computationally similar results (Extended Data Fig. 7).

ANN hidden layer weights were initialized from a Xavier normal distribution, with mean 0 and a scaling factor ranging from 0.2 to 2.0 in increments of 0.2 (ref. [69]). Biases were initialized to be 0. Training was implemented using a mean squared error cost function and the Adam optimizer with an initial learning rate of 0.0001 (ref. [70]). Training was stopped once the mean squared error fell below a threshold of 0.2. We also replicated these core mode results using a standard stochastic gradient descent optimizer with a learning rate of 0.01 (Extended Data Fig. 8). Note that smaller learning rates resulted in intractable training times using stochastic gradient descent.

We fit ANNs for each participant's activations separately. For every participant, we trained 20 networks with different random initializations. For each ANN analysis, statistics and network properties (for example, dimensionality, weight norms and so on) were averaged across participants, and statistical tests were performed on the 20 random initializations.

We note that while there is no strict definition of rich versus lazy training, there are several factors that are good proxy measurements of an ANN's training regime. One such proxy is that training cost/time-rich training is far more computationally costly than lazy training[47]. Empirically, we observe that rich training has weight initializations that are smaller than the default initialization (s.d. = 1.0), while computationally cheap lazy training includes initializations that are greater than the default. Nevertheless, these definitions can change with ANN architectures because weight initializations can impact the vanishing and exploding gradient issue in ANN training.

All models were built using PyTorch version 1.4.0 and Python version 3.8.5.

ANN analysis

Trained ANNs were subject to analysis to characterize both the learned intermediate representations and weight distribution properties. Model RSMs were generated by propagating participant-level task activations through the hidden layers. Cross-validated RSMs were constructed and analyzed identically to fMRI data (for example, cosine similarity and then participation ratio to estimate its dimensionality; Fig. 6c). As in our fMRI analysis, we used a split-half cross-validation where we compared task activations between the first and second imaging sessions of each task set. We fitted the dimensionality across ANN layer depth using a second-order polynomial regression to assess how representational dimensionality changed throughout the network (Fig. 6d). Positive and negative second-order coefficients indicated convex and concave quadratics, respectively.

We compared the representational geometries produced by the ANN with the representational geometries found in empirical fMRI data. To directly compare ANN and empirical RSMs, we partitioned the cortex into 10 bins containing 36 parcels each (Fig. 6g). Cortical bins and their ordering were determined by the RSFC sensory–motor gradient, where parcels with similar loadings were placed in adjacent bins (Fig. 5a). We computed the cosine similarity of each region's RSM with each ANN layer's RSM. To evaluate the correspondence between representations in each cortical bin and each ANN layer, we averaged the cosine values across parcels within each bin (Fig. 6f). This was done for ANNs trained under the rich regime (weight initializations less than 1) and the lazy regime (weight initializations greater than 1).

We assessed the interlayer RA within the ANN for different weight initializations (Fig. 7a), which is similar to interregion RA measured in fMRI data (Fig. 3c). This was defined as the cosine similarity between RSMs between pairs of ANN layers (Fig. 7a). We also analyzed the properties of the trained and initialized ANN weights. This included calculation of the Frobenius norm, Fisher kurtosis and singular value decomposition of the weight matrices under different weight initializations. Dimensionality of the ANN's weights was performed by measuring the participation ratio of the singular values. All statistical analyses were performed in Python version 3.8.5 using the NumPy (version 1.18.5) and SciPy (version 1.6.0) packages.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data in this study have been made publicly available on OpenNeuro by King and colleagues (accession number ds002105 (ref. [18])).

Code availability

All code related to this study is publicly available on GitHub (https://github.com/murraylab/multitaskhierarchy). Analyses and models were implemented using Python (version 3.8.5). Cortical visualizations were implemented using workbench (version 1.5.0).

References

55. Zhi, D., King, M., Hernandez-Castillo, C. R. & Diedrichsen, J. Evaluating brain parcellations using the distance-controlled boundary coefficient. *Hum. Brain Mapp.* **43**, 3706–3720 (2022).
56. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124 (2013).
57. Ji, J. L. et al. QuNex—a scalable platform for integrative multi-modal neuroimaging data processing and analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.03.494750 (2022).
58. Ito, T. et al. Task-evoked activity quenches neural correlations and variability across cortical areas. *PLoS Comput. Biol.* **16**, e1007983 (2020).
59. Ciric, R. et al. Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage* **154**, 174–187 (2017).
60. Glasser, M. F. et al. The Human Connectome Project's neuroimaging approach. *Nat. Neurosci.* **19**, 1175–1187 (2016).
61. Rissman, J., Gazzaley, A. & D'Esposito, M. Measuring functional connectivity during distinct stages of a cognitive task. *NeuroImage* **23**, 752–763 (2004).
62. Friston, K. J. et al. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1994).
63. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex* **28**, 3095–3114 (2018).
64. Abdollahi, R. O. et al. Correspondences between retinotopic areas and myelin maps in human visual cortex. *NeuroImage* **99**, 509–524 (2014).
65. Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A. & Love, B. C. Measures of neural similarity. *Comput. Brain Behav.* **3**, 369–383 (2020).
66. Walther, A. et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* **137**, 188–200 (2016).
67. Basti, A., Nili, H., Hauk, O., Marzetti, L. & Henson, R. N. Multi-dimensional connectivity: a conceptual and mathematical review. *NeuroImage* **221**, 117179 (2020).

68. Burt, J. B., Helmer, M., Shinn, M., Anticevic, A. & Murray, J. D. Generative modeling of brain maps with spatial autocorrelation. *NeuroImage* **220**, 117038 (2020).
69. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* 249–256 (JMLR Workshop and Conference Proceedings, 2010).
70. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://doi.org/10.48550/arXiv.1412.6980 (2015).

Acknowledgements

This project was supported by NIH grant R01MH112746 (J.D.M.), NSF NeuroNex grant 2015276 (J.D.M.) and a Swartz Foundation Fellowship (T.I.). We acknowledge the Yale Center for Research Computing at Yale University for providing access to the Grace cluster and associated research computing resources. We thank M. King, J. Diedrichsen and colleagues for providing public access to the dataset. We also thank W. Pettine, M. Helmer and J. Miller for comments on earlier drafts of the manuscript.

Author contributions

T.I. and J.D.M. conceptualized the project and wrote the paper. T.I. performed the formal analysis and visualization, developed software and wrote the original draft. J.D.M. acquired funding and supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-022-01224-0.

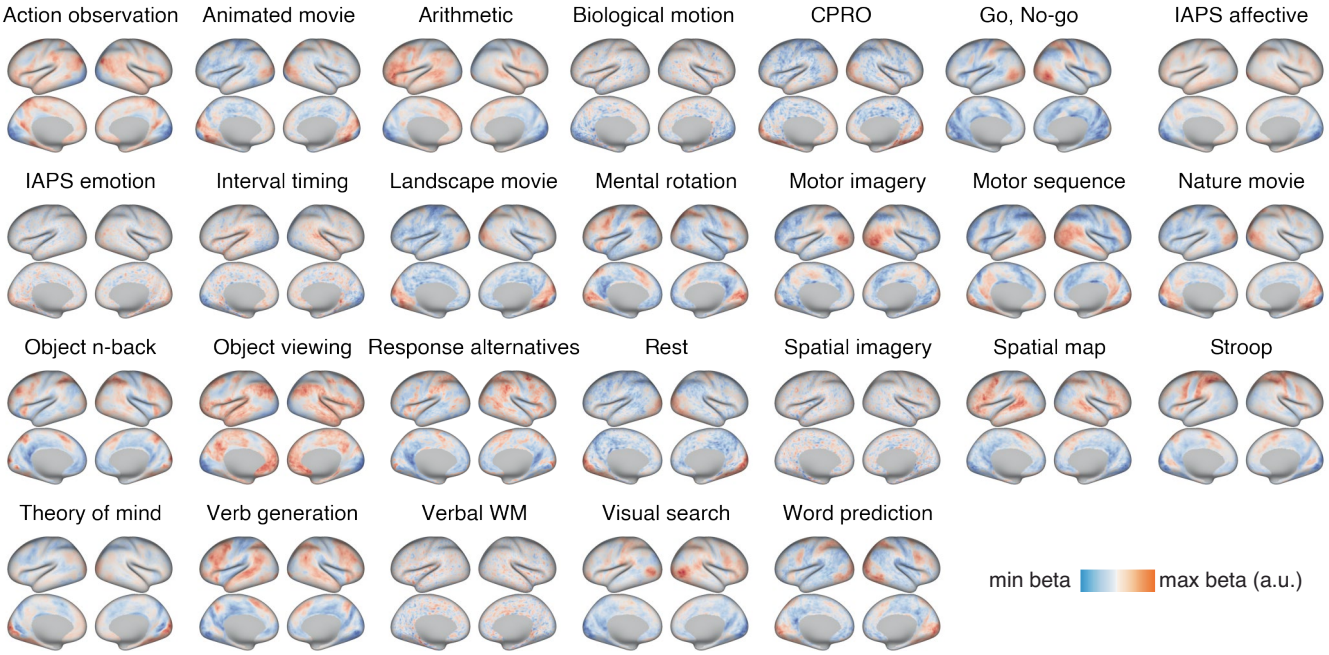
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01224-0.

Correspondence and requests for materials should be addressed to John D. Murray.

Peer review information *Nature Neuroscience* thanks Matthew Farrell, Lucina Uddin, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

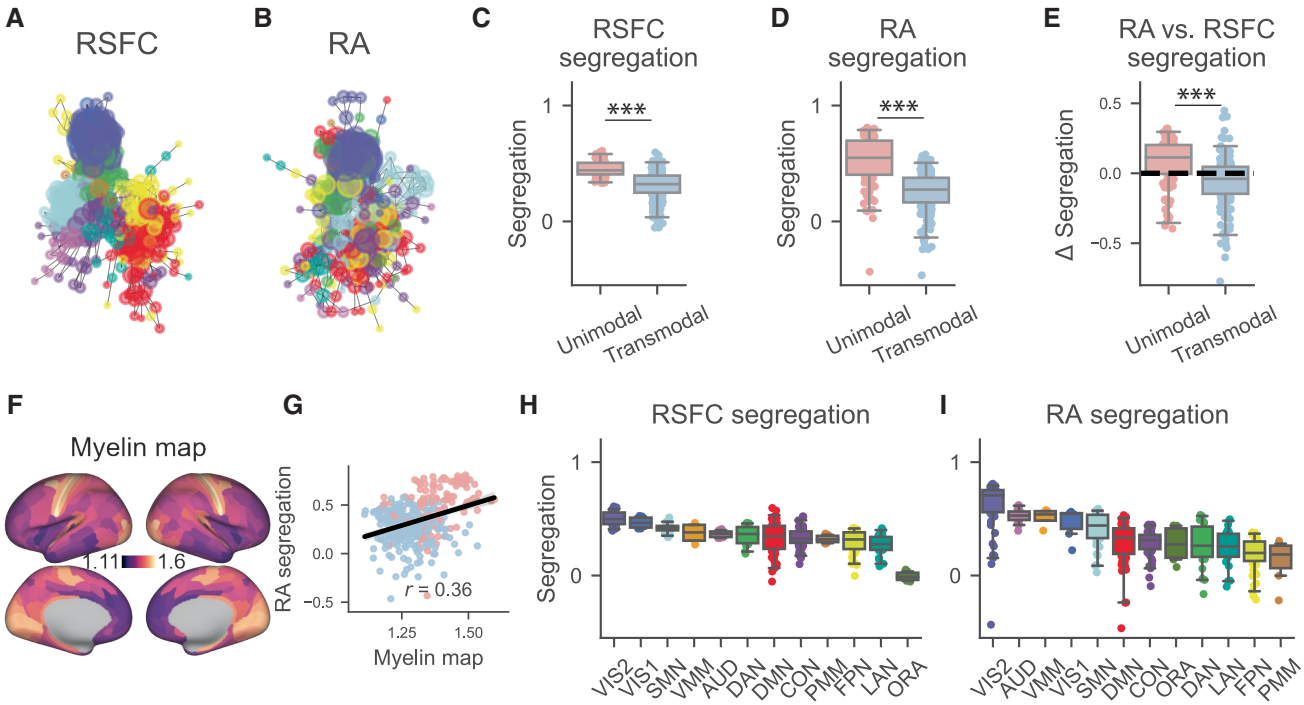
Reprints and permissions information is available at www.nature.com/reprints.

https://doi.org/10.1038/s41593-022-01224-0



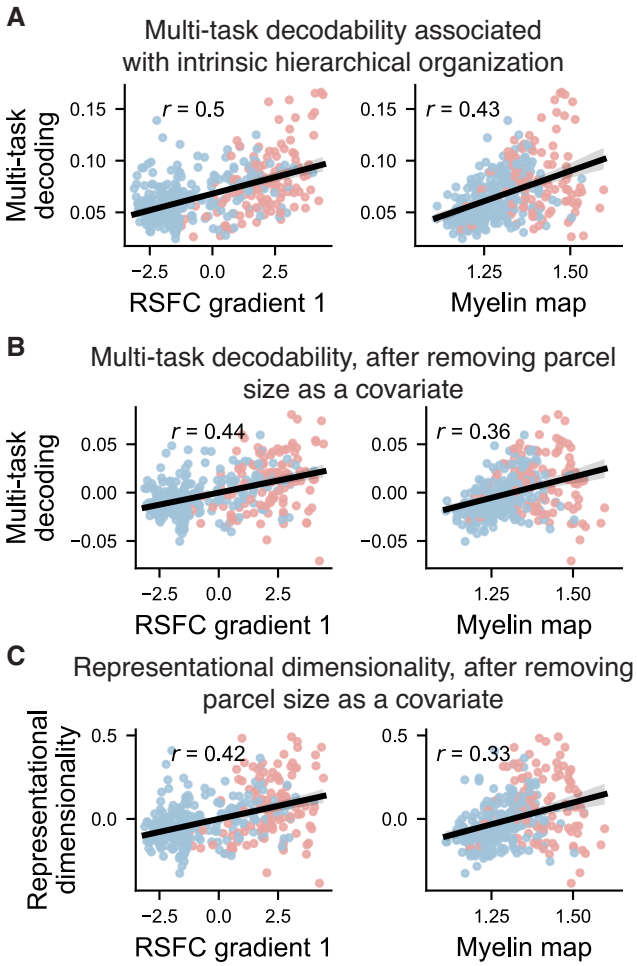
Extended Data Fig. 1 | Whole-cortex group activation maps for all 26 cognitive tasks. Activation maps reflect the GLM beta values and were averaged across conditions within each task.

https://doi.org/10.1038/s41593-022-01224-0



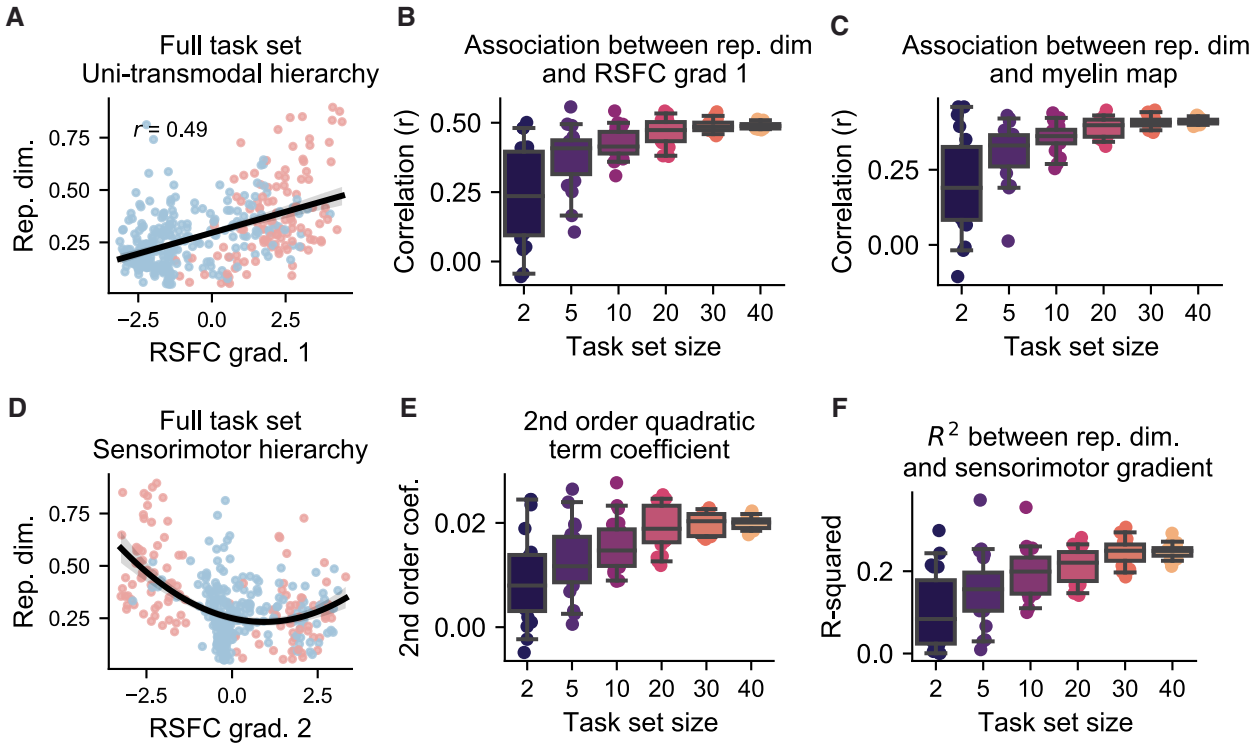
Extended Data Fig. 2 | Comparing segregation of whole-cortex RSFC and RA between unimodal-transmodal areas and functional networks. a, b) Force-directed graphs comparing RSFC and RA community structure (color-coated by functional networks). **c)** Segregation of RSFC and **d)** RA whole-cortex matrices (n = 144 unimodal, n = 246 transmodal). **e)** The direct comparison of differences in segregation between RA and RSFC for unimodal and transmodal regions (same as Fig. 3h). (Panels **c-e** are two-sided t-tests.) **f, g)** Association of regional RA segregation with the cortical myelin map (T1w/T2w structural map). **h)** Segregation of RSFC by functional networks. **i)** Segregation of RA by functional networks. Note that for both RA and RSFC, sensorimotor networks have higher segregation than association networks. Boxplot bounds define the 1st and 3rd quartiles of the distribution, box whiskers the 95% confidence interval, and the center line indicates the median. Network key: VIS1 = Visual 1 (n = 6); VIS2 = Visual 2 (n = 54); SMN = Somatomotor (n = 39); VMM = Ventral multimodal (n = 6); AUD = Auditory (n = 15); DAN = Dorsal attention (n = 23); DMN = Default mode (n = 77); CON = Cingulo-opercular (n = 56); PMM = Posterior multimodal (n = 7); FPN = Frontoparietal (n = 50); LAN = Language (n = 23); ORA = Orbital-affective (n = 4). Colors of each network correspond to colors in panel Fig. 3e. (***p = <0.0001, two-sided t-test.).

https://doi.org/10.1038/s41593-022-01224-0



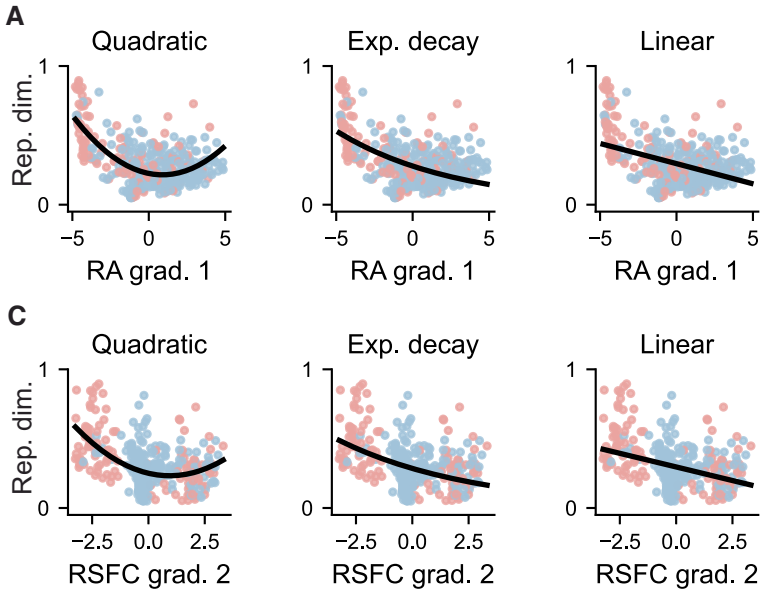
Extended Data Fig. 3 | Representational dimensionality and multi-task decoding produce similar associations with intrinsic hierarchy, even after controlling for parcel size. a) Correlation of multi-task decoding with the principal RSFC gradient and myelin map across regions. **b)** Parcel size (number of vertices within a brain region) and representational dimensionality were positively correlated (r = 0.45, non-parametric p < 0.001). However, after accounting for parcel size (that is, the number of vertices within each parcel) as a covariate (via linear regression), a strong association between decodability and intrinsic hierarchy was maintained. **c)** Same analysis as in panel **b**, but using representational dimensionality rather than decodability. All correlations in a, b, and c resulted in a non-parametric p < 0.001 using surrogate brain maps that accounted for spatial autocorrelation[68]. This suggests that the association between representational dimensionality and intrinsic hierarchy is independent of parcel size. Error bands reflect a 95% confidence interval.

https://doi.org/10.1038/s41593-022-01224-0



Extended Data Fig. 4 | Random subsamples of the task set show similar association with both the unimodal-transmodal and the sensorimotor hierarchy. a) The association between representational dimensionality and the principal RSFC gradient (unimodal-transmodal hierarchy) with the entire task set. **b)** We randomly sub-sampled (without replacement) tasks to downsize the RSMs of all parcels, and then measured the correlation between representational dimensionality and RSFC gradient 1. For each sub-sample size, we repeatedly chose (that is, 45 choose n) 20 times to estimate the robustness of the association with arbitrary selection of tasks. The association increased and stabilized as we increased the number of tasks ($n = 20$). **c)** Same as in **b**, but using the myelin map. **d)** The compression-then-expansion fit of representational dimensionality and the sensorimotor (RSFC gradient 2) hierarchy. **e)** We estimated the 2nd-order polynomial fit for randomly sub-sampled tasks, and assessed the coefficient of 2nd-order polynomial fit. The higher (and more positive) the parameter, the more convex the compression-then-expansion was. Increased compression-then-expansion as the number of randomly sampled tasks were included ($n = 20$ random subsamples). **f)** Same procedure as **e**, but measuring the R-squared of the polynomial fit rather than the 2nd-order coefficient term. Boxplot bounds define the 1st and 3rd quartiles of the distribution, box whiskers the 95% confidence interval, and the center line indicates the median. Error bands reflect a 95% confidence interval in panels **a** and **d**.

https://doi.org/10.1038/s41593-022-01224-0



B

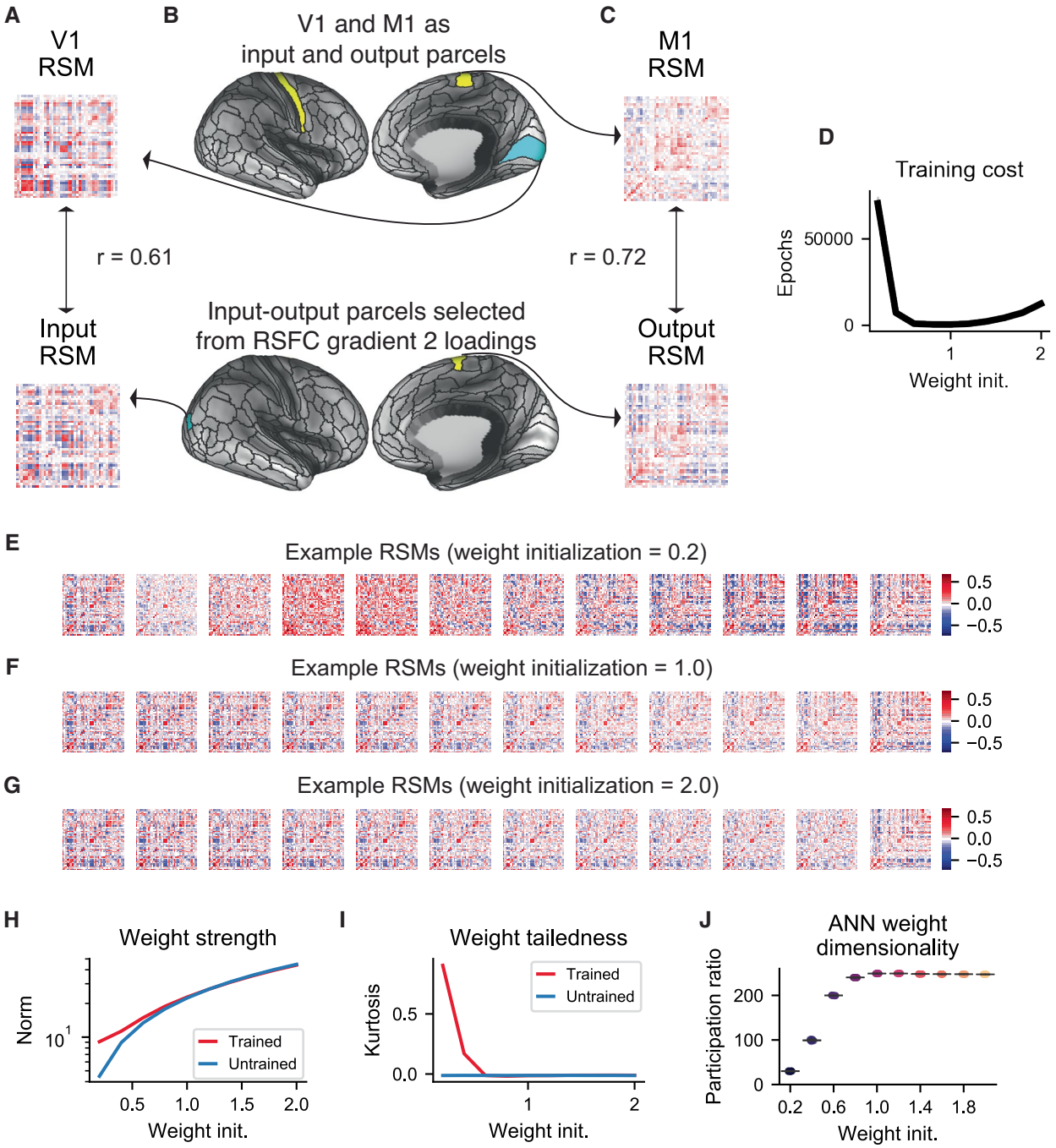
	AIC	BIC
Quadratic	-441.59	-429.93
Exp. decay	-356.28	-348.5
Linear	-326.19	-318.41

D

	AIC	BIC
Quadratic	-356.43	-344.77
Exp. decay	-320.74	-312.96
Linear	-306.02	-298.25

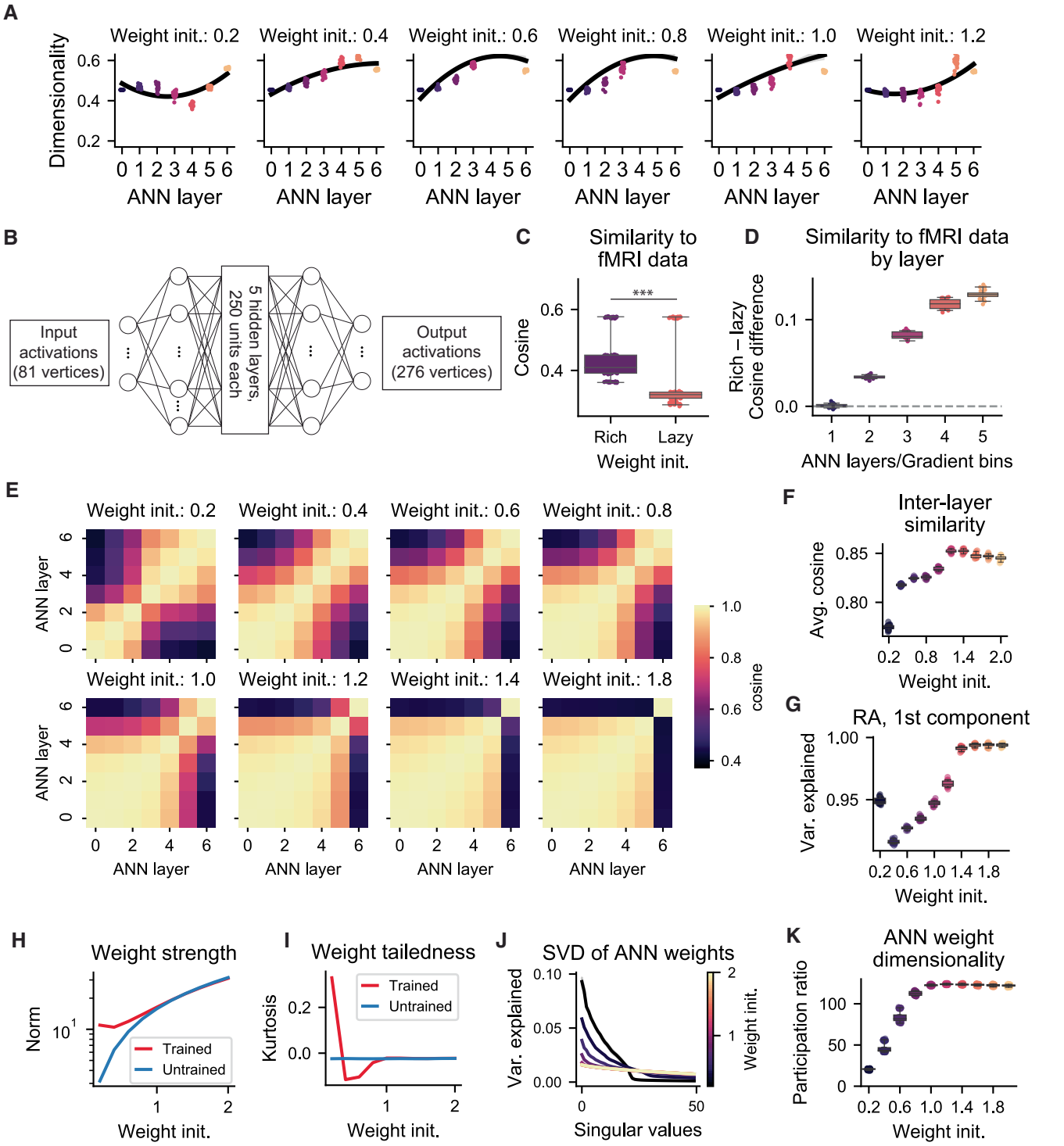
Extended Data Fig. 5 | Establishing compression-then-expansion of representational dimensionality across the sensory-motor hierarchy via model adjudication. a) We fit the representational dimensionality of parcels across the sensory-motor RSFC gradient using three competing models: Quadratic (2nd-order polynomial), linear, and an exponential decay model, where separate models were fit for loadings less than and greater than 0. **b)** The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for all models, which takes into account the maximum likelihood of each model while penalizing the models with more free parameters. Quadratic models had the smallest values for both AIC and BIC. **c,d)** Same as panels **a** and **b**, but using the RA principal gradient. Quadratic models were defined as $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$. Linear models were defined as $y = \beta_0 + \beta_1 x + \epsilon$. Exponential decay models were defined as $y(t) = N_0 e^{-\lambda t} + \epsilon$.

https://doi.org/10.1038/s41593-022-01224-0



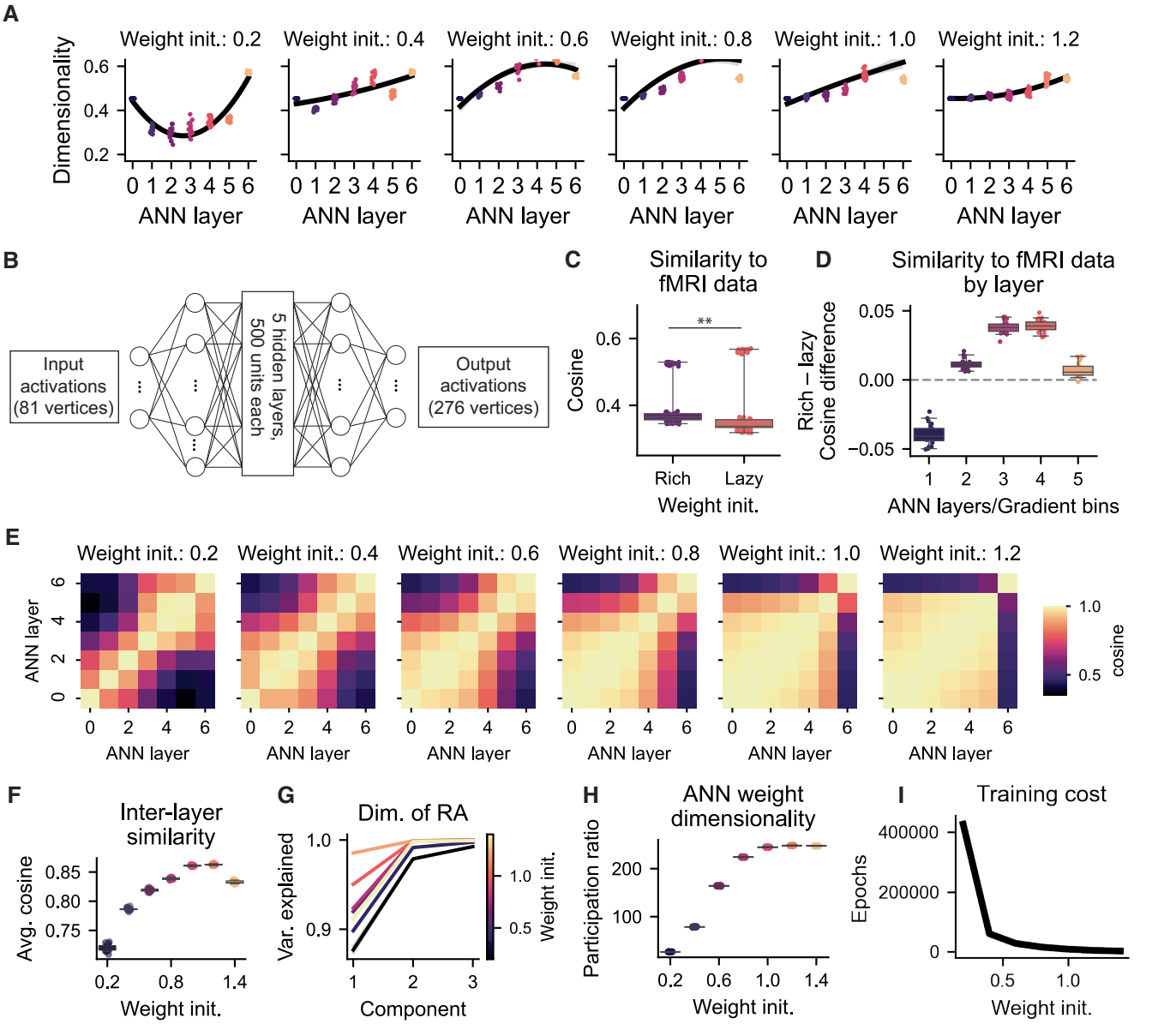
Extended Data Fig. 6 | Supplemental information on ANN modeling during rich and lazy training regimes. The similarity between **a)** the RSMs for V1 and the gradient-identified input parcel for model construction and **b)** the RSMs for M1 and the gradient-selected motor output parcel. Overall, the representational geometries were highly similar between V1 and the input RSM, and M1 and the motor output RSM. **d)** The training cost (that is, number of training epochs required) for different weight initializations. Visualization of RSMs for example ANNs (one initialization each) for **e)** rich, **f)** intermediate (that is, initialization SD = 1.0), and **g)** lazy training regimes. **h-j)** Characterizing the structural network mechanisms that give rise to differences in representational structure across learning regimes in the ANN. **h)** Initialized and trained norm of ANN weights as a function of weight initialization. In line with previous work[47], the Frobenius norm of the trained ANN, which reflects the variability of the hidden weight projections, were significantly smaller in the rich training regime. **i)** The kurtosis of the degree distribution during initialization and after training. The kurtosis of the weight distribution measures the tailedness of the weight distribution. Kurtosis (in terms of connectivity weights) reflects the small-worldness of a network, a well-documented feature in empirical brain networks[39]. The kurtosis of richly trained networks was higher than in lazily trained networks, producing a heavy-tailed weight distribution. **j)** We characterized the dimensionality of the ANN weights to gain insight into the successive representational transformations in the ANN across 20 initializations per weight distribution ($n = 20$). Weight dimensionality was computed by performing a singular value decomposition (SVD) on the weights, and then calculating the participation ratio of the singular values. The dimensionality of the learned weights directly constrains the representations the ANNs produce. Low-dimensionality of the connectivity weights likely aids in cross-task generalization, since low-dimensional connections force the network to extract shared components across tasks. Weight dimensionality was lower in rich training regimes. These findings suggest that across layers, richly trained ANNs with low-dimensional and low-variability weights collectively produced modular patterns of representations across layers, consistent with empirical data. Boxplot bounds define the 1st and 3rd quartiles of the distribution, box whiskers the 95% confidence interval, and the center line indicates the median.

https://doi.org/10.1038/s41593-022-01224-0



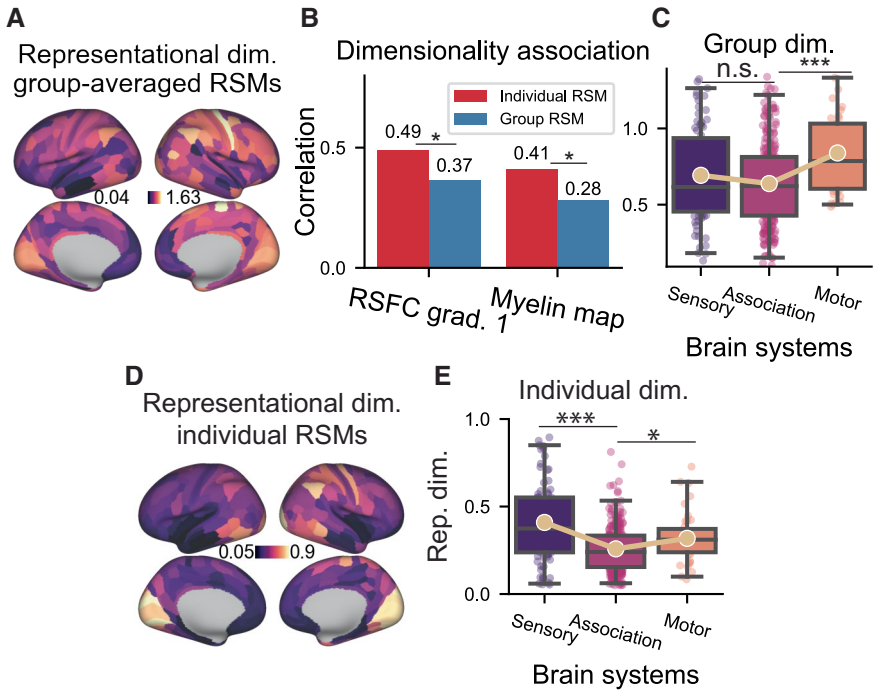
Extended Data Fig. 7 | Training an ANN with untied weights results in qualitatively similar results. We trained a 5-layer ANN with untied weights to produce qualitatively similar results to the ANN in the main manuscript. We reduced the number of layers from 10 to 5 and the number of hidden units from 500 to 250 for computational efficiency. (An ANN with untied weights has significantly greater parameters than one with tied weights.) **a)** Representational dimensionality of ANN layers for different weight initializations. **b)** ANN architecture. **c)** Richly trained ANNs had significantly higher similarity with representations found in empirical data relative to lazily trained ANNs ($n = 20$). **d)** Similarity to fMRI data by layer (rich minus lazy ANNs) ($n = 20$). **e)** Representational alignment of each ANN's layer (cosine similarity between RSMs). **f)** Overall similarity of representations across ANN layers. Greater representational dissimilarity (across layers) is found in richly trained ANNs ($n = 20$). **g)** Variance explained of the first principal component for each of the RA matrices in panel **e** ($n = 20$). **h)** Frobenius norm of the weight distribution across initializations. **i)** The kurtosis (tailedness) of the weight distribution across layers under different weight initialization schemes. **j)** SVD of ANN weights. **k)** Dimensionality (participation ratio) of the weights for different initializations ($n = 20$). Richer training regimes produce low-dimensional weights. Boxplot bounds define the 1st and 3rd quartiles of the distribution, box whiskers the 95% confidence interval, and the center line indicates the median. (***p < 0.0001, two-sided t-test.).

https://doi.org/10.1038/s41593-022-01224-0



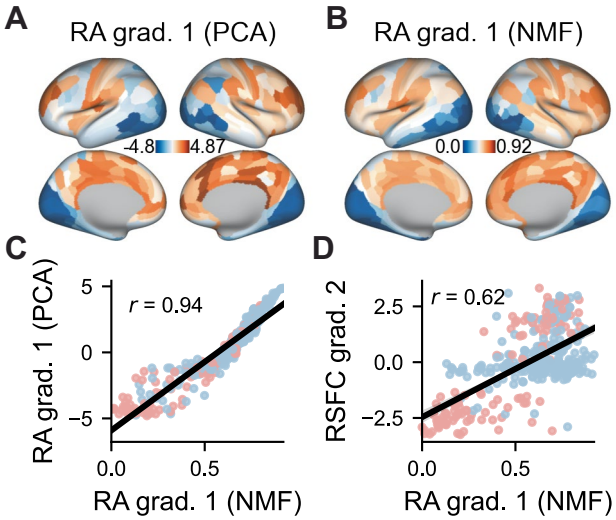
Extended Data Fig. 8 | Using standard stochastic gradient descent (without momentum/weight decay) with tied weights also produces qualitatively similar results. To explore the impact of model optimization and network size on the learned representations in ANNs, we trained a 5-layer ANN to produce qualitatively similar results to the ANN in the main manuscript (Figs. 6, 7). We reduced the number of layers from 10 to 5 for computational efficiency. Instead of using the Adam optimizer (with a learning rate of 0.0001), we used standard stochastic gradient descent with a learning rate of 0.01. (Note that smaller learning rates were highly computationally intractable for learning in the rich training regime.) We did not include model initializations with SD > 1.4 due to exploding gradients. **a)** Representational dimensionality of ANN layers for different weight initializations. **b)** ANN architecture. **c)** Richly trained ANNs had significantly higher similarity with representations found in empirical data relative to lazily trained ANNs (rich>1.0, lazy<1.0) ($n = 20$). **d)** Similarity to fMRI data by layer (rich minus lazy ANNs) ($n = 20$). **e)** Representational alignment of each ANN's layer (cosine similarity between RSMs). **f)** Overall similarity of representations across ANN layers ($n = 20$). Greater representational dissimilarity (across layers) is found in richly trained ANNs. **g)** Cumulative variance explained of the first three principal components for each of the RA matrices in panel **e**. **h)** Dimensionality (participation ratio) of the learned connectivity weights for different initializations ($n = 20$). **i)** Average training cost by weight initialization. Boxplot bounds define the 1st and 3rd quartiles of the distribution, box whiskers the 95% confidence interval, and the center line indicates the median. (**p < 0.001, two-sided t-test.).

https://doi.org/10.1038/s41593-022-01224-0



Extended Data Fig. 9 | The importance of within-subject analyses to capture fine-grained representational patterns. a) Representational dimensionality across the cortical surface when computing dimensionality using the group-averaged RSM (rather than subject-specific RSM). **b)** We computed the correlation between representational dimensionality with two proxies of the unimodal-transmodal hierarchy: RSFC principal gradient and the myelin map (T1w/T2w contrast). We find that when calculating dimensionality from RSMs derived from group-level activation averages, the association with the unimodal-transmodal hierarchy is significantly reduced. **c)** We subsequently measured dimensionality across the sensory-association-motor systems, finding that in contrast to within-subject estimates of representational dimensionality, we no longer observed the dimensionality compression from sensory to association systems in group-derived maps (sensory, $n = 75$; association, $n = 246$; motor, $n = 39$). **d)** Representational dimensionality measured using individual RSMs. (Same as in Fig. 4b, for visual comparison.) **e)** Dimensionality across the sensory-association-motor hierarchy using dimensionality computed from individual RSMs (same as in Fig. 5g, for visual comparison). Boxplot bounds define the 1st and 3rd quartiles of the distribution, box whiskers the 95% confidence interval, and the center line indicates the median. (***$p < 0.0001$, *$p < 0.05$, two-sided t-test.).

https://doi.org/10.1038/s41593-022-01224-0



Extended Data Fig. 10 | Corroborating evidence of the sensory-association-motor axis of hierarchical organization extracted using non-negative matrix factorization (NMF). This revealed that the sensory-to-motor hierarchy was robust to different matrix decomposition algorithms. **a)** The first component extracted using PCA and **b)** NMF. **c)** Correlation between the first components extracted with PCA and NMF. **d)** Correlation of RA gradient 1 (NMF) with the RSFC sensorimotor hierarchy (RSFC gradient 2).

nature portfolio

Corresponding author(s): John D. Murray

Last updated by author(s): 2022/10/19

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected from a previous publication (King et al. 2019, Nat. Neurosci). No software was used for data collection in this study.
Data analysis	QuNex (version 0.61.17) was used to preprocess the raw data. Workbench (version 1.5.0) was used to visualize cortical maps. Custom python code (version 3.8.5) was used for analysis, and will be made publicly available prior to publication. We used NumPy (version 1.18.5) and SciPy (version 1.6.0).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data in this study has been made publicly available on OpenNeuro by King and colleagues (accession number: ds002105). (URL: https://openneuro.org/datasets/ds002105/)

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	We report using biological sex, as reported in the original data resource article (King et al., 2019, Nature Neuroscience).
Population characteristics	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "The final sample consisted of 24 healthy, right-handed individuals (16 females, 8 males; mean age=23.8 years old, SD=2.6) with no self-reported history of neurological or psychiatric illness."
Recruitment	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "Undergraduate and graduate students were recruited (via posters) from the larger student body at Western University. Thus, our sample was biased towards relatively high-functioning, healthy and young individuals. While we don't expect cerebellar organization to be dramatically different in this group, caution needs to be exercised when generalizing the results to the general population."
Ethics oversight	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "The Ethics committee at Western University approved all experimental protocols (Protocol number: 107293)"

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by the previous Data Resource article (King et al. 2019, Nature Neuroscience): "No formal power analysis was conducted, but the number of participants and the amount of data per participant was determined from prior experience. For the within-subject sample size, we required 5.5 hours of fMRI data for each participant. This gave us enough power to reliably estimate activity patterns for the large number of tasks, as well as having enough data to perform out-of-sample prediction tests. A sample size of n=24 participants (after exclusions) was deemed sufficient to both estimate a representative mean organization, as well as obtaining reliable estimates of the between-subject variability." The present study also performed group-level analyses, ensuring similar statistical power to the original study.
Data exclusions	In the current study, no subjects were excluded using the data published in the online repository.
Replication	While no direct replication was performed with an external data set, cross-validation (within subjects) was performed across two independent recording sessions per participant to ensure robustness and reliability of results.
Randomization	N/A -- participants were not allocated into separate groups.
Blinding	No binding was performed, since between-group differences were not assessed.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Magnetic resonance imaging

Experimental design

Design type	Task- and resting-state.
Design specifications	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "Two task sets. 2 fMRI scanning sessions per task set. 8 functional imaging runs per session (10-min each). 17 tasks per imaging run (35 s each)."
Behavioral performance measures	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "Variables recorded: response made, number of correct responses, false alarms, missed responses, response time. Accuracy (% correct) and reaction time (ms) were collected and averaged across tasks per participant"

Acquisition

Imaging type(s)	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "EPI, MPRAGE, and GRE field maps"
Field strength	3T
Sequence & imaging parameters	From the original Data Resource article (King et al., 2019, Nature Neuroscience): "EPI: Gradient echo, multi-band (factor 3, interleaved) with an in-plane acceleration (factor 2). Imaging parameters were: TR=1 sec, FOV=20.8cm, phase encoding direction was P to A, acquiring 48 slices with in-plane resolution of 2.5 mm x 2.5 mm and 3 mm thickness. For anatomical localization and normalization, a 5-min high-resolution scan of the whole brain was acquired (MPRAGE, FOV=15.6 cm x 24 cm x 24 cm, at 1x1x1 mm voxel size)."
Area of acquisition	Whole-brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Resting-state and task-state fMRI data were minimally preprocessed using the Human Connectome Project (HCP) preprocessing pipeline within the Quantitative Neuroimaging Environment & Toolbox (QuNex, version 0.61.17). The HCP preprocessing pipeline consisted of anatomical reconstruction and segmentation, EPI reconstruction and segmentation, spatial normalization to the MNI152 template, and motion correction.
Normalization	Nonlinear spatial normalization was performed to the MNI 152 template using QuNex.
Normalization template	MNI152
Noise and artifact removal	Additional nuisance regression was performed on the minimally preprocessed time series. Consistent with previous reports, this included six motion parameters, their derivatives, and the quadratics of those parameters (24 motion regressors in total). We also removed the mean physiological time series extracted from the white matter and ventricle voxels. We also included the quadratic, derivatives, and the derivatives of the quadratic time series of each of the white matter and ventricle time series (8 physiological nuisance signals). This amounted to 32 nuisance parameters in total, and was a nuisance regression model that was previously benchmarked (Ciric et al. 2017).
Volume censoring	We excluded the first five volumes of each run.

Statistical modeling & inference

Model type and settings	We used univariate GLMs to estimate task activations, and multivariate pattern analyses (RSA and decoding) analyses at the vertex-level.
Effect(s) tested	Mean differences were tested for: Dimensionality; Segregation/Integration Associations (Pearson correlations) were tested between: Dimensionality, RA segregation, and Principal Component Gradient loadings.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	Whole-brain analyses and inferences were made at the parcel-level using pre-defined cortical parcels (defined in Glasser et al., 2016, Nature).
Correction	Since inferences were made at the whole-cortex level (or between systems across cortex), multiple comparisons correction was not necessary.

Models & analysis

n/a	Involved in the study
☐	☒ Functional and/or effective connectivity
☐	☒ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	Pearson correlation was used to estimate functional connectivity matrices.
Graph analysis	Weighted matrices thresholded at 20% were used to estimate principal gradient loadings, consistent with previous reports (Margulies et al., 2016, PNAS). Graph measures, such as network segregation, were applied on weighted and unthresholded matrices.
Multivariate modeling and predictive analysis	Multivariate modeling was performed using multivariate pattern decoding on vertices within each cortical parcel. Additional estimates of multivarate analyses were included, such as representational dimensionality, which measures the participation ratio of the cross-validated representational similarity matrix. Representational similarity matrices were constructed using cross-validated cosine similarity across imaging sessions (within subjects).