Generating Synthetic Task-based Brain Fingerprints for Population Neuroscience Using Deep Learning

Emin Serin[1,2,3], Kerstin Ritter[1], Gunter Schumann[4,5], Tobias Banaschewski[6], Andre Marquand[7]*, Henrik Walter[1]*, for the environMENTAL consortium

[1]Department for Psychiatry and Psychotherapy CCM, Charité – Universitätmedizin Berlin, 10117, Germany

[2]Charité – Universitätsmedizin Berlin, Einstein Center for Neurosciences Berlin, 10117, Berlin, Germany.

[3]Bernstein Center for Computational Neuroscience, 10115, Berlin, Germany

[4]Centre of Population Neuroscience and Stratified Medicine (PONS), Department of Psychiatry and Clinical Neuroscience CCM, Charité-Universitätsmedizin Berlin, 10117, Germany.

[5]Centre for Population Neuroscience and Stratified Medicine (PONS), Institute for Science and Technology of Brain-inspired Intelligence (ISTBI), Fudan University, Shanghai, China.

[6]Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany; German Center for Mental Health, Mannheim, Heidelberg, and Ulm, Germany.

[7]Donders Institute for Brain, Cognition and Behavior, Radboud University, Nijmegen, the Netherlands

*shared last authorship

Abstract:

Task-based functional magnetic resonance imaging (tb-fMRI) provides valuable insights into individual differences in the neural basis of cognitive functions because it links specific cognitive tasks to their evoked neural responses. Yet, it is challenging to scale to population-level data due to its cognitive demands, variations in task design across studies, and a limited number of tasks acquired in typical large-scale studies. Here, we present DeepTaskGen, a convolutional neural network (CNN) approach that enables us to generate synthetic task-based contrast maps from resting-state fMRI (rs-fMRI) data. Our method outperforms several benchmarks, exhibiting superior reconstruction performance while retaining inter-individual variation essential for biomarker development. We showcase DeepTaskGen by generating synthetic task images from the UK Biobank cohort, achieving competitive or greater performance compared to actual task contrast maps and resting-state connectomes for predicting a wide range of demographic, cognitive, and clinical variables. This approach will facilitate the study of individual differences and the generation of task-related biomarkers by enabling the generation of arbitrary functional cognitive tasks from readily available rs-fMRI data.

Introduction:

Characterizing inter-individual variability in brain activity patterns has become crucial to understanding individual differences in cognition, behavior, and the risk or resilience to mental disorders. Over the years, many techniques have been developed to discover the neural basis of individual differences. Among these, task-based functional magnetic resonance imaging (tb-fMRI) has emerged as a powerful technique as it can non-invasively measure changes in brain activity associated with a specific cognitive task. For example, the emotional face-matching task has been shown to activate several brain regions, particularly the amygdala[1], which has associations with mental disorders such as depression and anxiety[2]. Researchers can potentially identify individuals at risk of mental disorders by examining individual differences in task-related brain activation. Moreover, an increasing number of studies highlight that task-related brain activity improves prediction performance for individual differences in cognitive and behavioral traits, including fluid intelligence, cognitive flexibility, working memory capacity, and general cognitive ability, compared to task-free (resting-state) brain activity[3–7]. The flexibility of tb-fMRI in eliciting task-specific activations, coupled with its improved predictive power and promising findings regarding reliability[8,9], has established its role as a robust tool for understanding individual differences in brain activity[10].

Despite its advantages, task-based brain imaging has several limitations. Firstly, it is less scalable than other brain imaging techniques like anatomical imaging or resting-state fMRI (rs-fMRI) due to its cognitive demands, the need for subject training and compliance, and because different studies often use different task variants, which complicates aggregation across studies. Therefore, besides large consortia such as the Human Connectome Project, IMAGEN, and ABCD[11–15], which provide comprehensive datasets, functional task data remains limited. This is even more pronounced in populations that have difficulty in performing experimental tasks, such as young children, babies, or individuals with neurological or psychiatric conditions. The second limitation arises from variations in experimental task designs. Reliable and advanced biomarker discovery requires large-scale, multi-site datasets to represent individual variations on a population scale. However, this introduces the challenge of standardization, as researchers must control for differences in scanning sites, parameters, and variations in task designs.

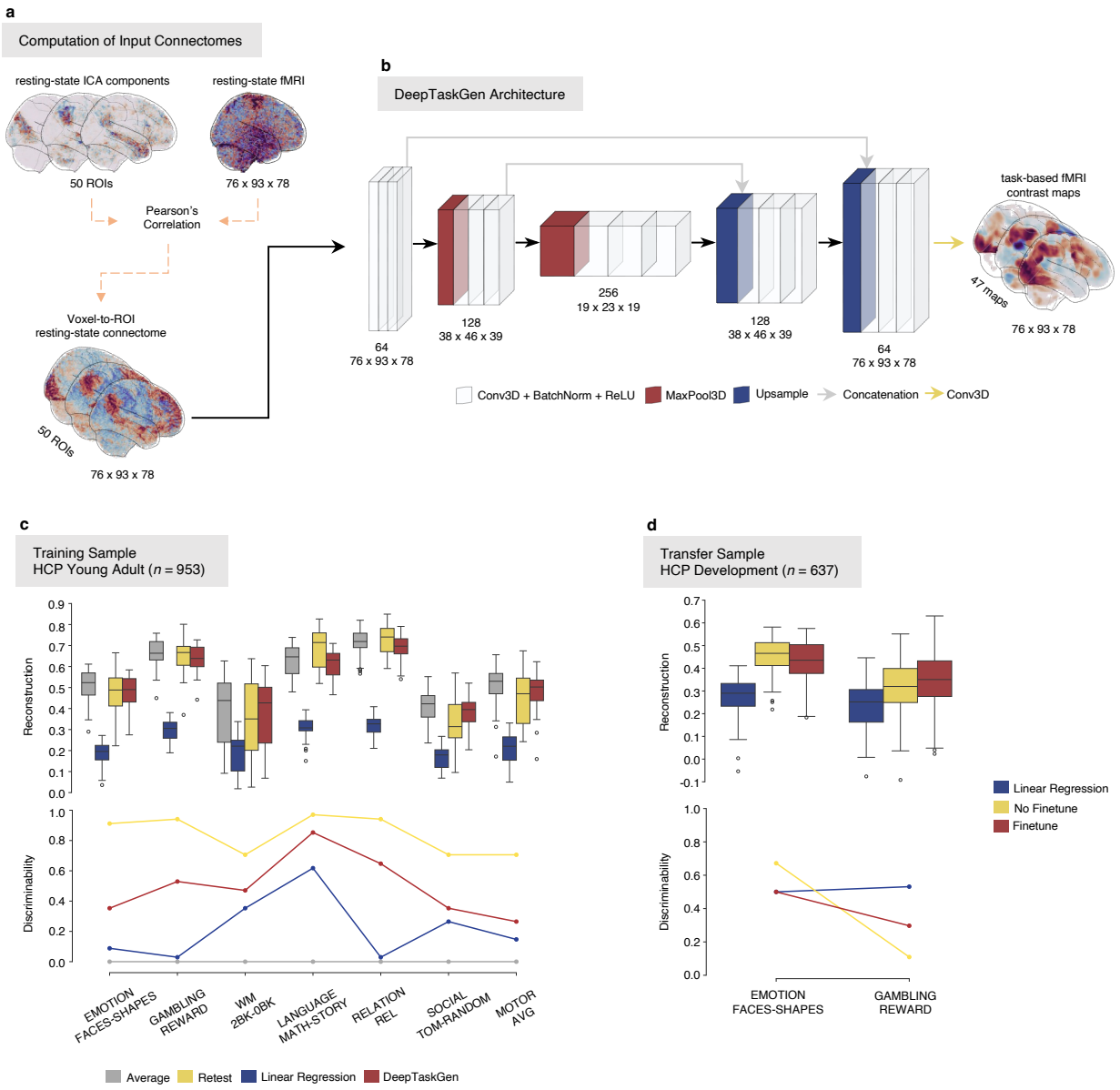
In this paper, we address these limitations by generating *synthetic* task-based brain images from task-free resting-state data. This approach builds upon the discovery that resting-state and task-based brain activity share the same intrinsic network architecture[16–18], which has been utilized by Tavor and colleagues[17]who predicted inter-individual variations in task-induced brain activity from resting-state data. Although this initial work offered valuable

insights into the inherent brain structures underlying brain functions, it has several limitations. It fails to capture the complexity of the network architecture necessary for retaining individual-specific information, lacks effective methods for transferring learned parameters to new datasets, and does not cover the entire brain, overlooking subcortical areas and cerebellum that are critical for several cognitive functions[19–21] and mental disorders[22–24].

Building upon previous work[25], here we present a volumetric neural network architecture that (i) addresses all the limitations mentioned above, (ii) provides the ability to generate synthetic task-based images in cohorts where these tasks were not acquired, and (iii) provides superior reconstruction performance in several datasets compared to different baseline methods, while also retaining individual differences that are essential for biomarker development. To demonstrate this approach, we first train DeepTaskGen on the Human Connectome Project Young-Adult (HCP-YA)[11] and demonstrate competitive performance in retaining individual differences while predicting task-based contrast maps (i.e., task-based brain activity). Next, we apply our trained network to the Human Connectome Project Development (HCP-D) dataset[15] to show its generalizability by predicting task-based contrast maps absent from this dataset. Next, we apply our trained network to generate synthetic task images for a comprehensive battery of cognitive functions for one of the largest population-based datasets currently available (UK Biobank).[26] Finally, we demonstrate the utility of these synthetic data as

biomarkers by predicting subjects' age, sex, fluid intelligence, grip strength, and overall health, as well as several clinical measures, including depression status, alcohol use frequency, anxiety, and depressive symptom scores.

Fig. 1: Input, DeepTaskGen architecture, and model evaluation on training and independent samples.



a. We utilized voxel-to-ROI rs-fMRI connectomes as input. A connectome was constructed for each subject by calculating the full correlation between the averaged time series from 50 ICA-based ROIs and the time series of individual voxels. **b.** Task-contrast maps for various tasks were predicted from rs-fMRI connectomes using our proposed DeepTaskGen architecture. DeepTaskGen is a volumetric U-Net model that processes the input resting-state connectome through a series of convolutional blocks, each comprising a 3D convolution layer, batch normalization, and a non-linear activation function. By utilizing max pooling, the model compresses images while preserving task-relevant patterns, and then, it upsamples the images to align with the output task contrast maps. The numbers below each block represent the output shape of each block and the number of feature maps (above). The details of the architecture are presented in Supplementary Table 1. **c.** We trained and evaluated DeepTaskGen on the HCP Young Adult dataset. The figure above shows the reconstruction performance computed by taking Pearson's correlation between predicted and actual contrast maps for representative contrasts from seven distinct tasks. The figure below displays discriminability, an index showing how unique subjects' task contrasts are, computed using a fingerprinting methodology[27]. We compared DeepTaskGen with methods like group-averaged contrast maps, retest scans, and a linear model (each depicted in distinct colors). **d.** We further fine-tuned the trained DeepTaskGen model on the HCP Development dataset using either task contrasts (e.g., GAMBLING REWARD). We predicted the other contrast (e.g., EMOTION FACES-SHAPES). The fine-tuned model was compared to the non-fine-tuned DeepTaskGen and linear models (shown in distinct colors). Reconstruction performance and discriminability were again used to assess the models' performance for each task contrast.

Results:

DeepTaskGen demonstrates overall better reconstruction performance and discriminability on the training dataset.

First, we trained DeepTaskGen on the HCP-YA dataset using a train set (N = 827) and a validation set (N = 92) and assessed its reconstruction performance (i.e., Pearson's correlation between actual and synthetic maps) and its ability to retain individual differences (i.e., inter-individual variance) on a separate test set (N = 34), which is essential for deriving functional biomarkers. Next, we used permutation testing to compare these measures with group-average task contrasts, a linear model[17], and retest scans (which can be seen as a theoretical upper bound on prediction ability). Figure 1c displays the results for seven representative main task contrasts of seven different paradigms, while Supplementary Figure 1 and Supplementary Tables 3-4 provide the performance for all 47 task contrasts and their test statistics, respectively. Additionally, we present the actual and predicted group-average maps for the seven representative main task contrasts in Figure 2a.

Our results showed that DeepTaskGen (μ = 0.497, σ = 0.169; μ and σ: mean and standard deviation scores across 47 contrast maps) achieved significantly greater reconstruction performance than the linear model (μ = 0.203, σ = 0.096) in all 47 task contrasts, indicating an advantage of our network in predicting task contrast maps compared to the linear model. DeepTaskGen also significantly outperformed retest scans (μ = 0.465, σ =

0.208) in 27 task contrasts while achieving worse performance in 6 task contrasts. However, we found no significant difference between DeepTaskGen and retest scans in 14 task contrasts in terms of reconstruction performance. The improved performance of DeepTaskGen over retest scans can be attributed to scan-specific noise present in retest scans or shared session-specific information across both task-based and task-free brain activity retained by DeepTaskGen. While group-average contrast maps (μ = 0.521, σ = 0.170) achieved the highest scores in reconstruction performance in nearly all task contrasts, these do not provide any inter-individual variance whatsoever. This emphasizes that reconstruction performance, while necessary, is insufficient for downstream analyses such as prediction and biomarker development, which require individual-specific information.

DeepTaskGen (μ = 0.413, σ = 0.202) consistently yielded a higher discriminability than the linear model (μ = 0.094, σ = 0.108) in 46 of 47 task contrasts while providing similar discriminability in MOTOR T. This indicates our model's significant advantage in retaining individual differences compared to the linear model. However, retest scans (μ = 0.762, σ = 0.192) showed the highest discriminability, suggesting they provide the most distinct task contrasts. However, this high discriminability may be influenced by individual-specific noise, which could contribute to lower reconstruction performance compared to DeepTaskGen.

In summary, compared to various baseline methods, DeepTaskGen achieves a good balance between reconstruction performance and maintaining individual-specific variation. The test results and discriminability scores for all 47 task contrasts are given in Supplementary Table 3-4.

DeepTaskGen allows for generating contrast maps on an unseen dataset

Next, we evaluated DeepTaskGen's performance in generating task contrasts on a dataset not included in its original training. To do this, we used the HCP-D dataset, which is ideal because it shares multiple tasks with the HCP-YA dataset used for training the model. This, in turn, allows us to assess the quality of the map generation against ground truth. To adapt the pre-trained model to the new dataset, we fine-tuned it using a single task contrast common to both datasets (see Methods). We then used the fine-tuned model to predict the task contrast that we did not use during fine-tuning. For example, the model for the GAMBLING REWARD contrast was fine-tuned using EMOTION FACES-SHAPES, and vice versa for EMOTION FACES-SHAPES. Furthermore, we assessed to what extent the trained model performed on an unseen dataset without fine-tuning, which we referred to as “No Finetune.”

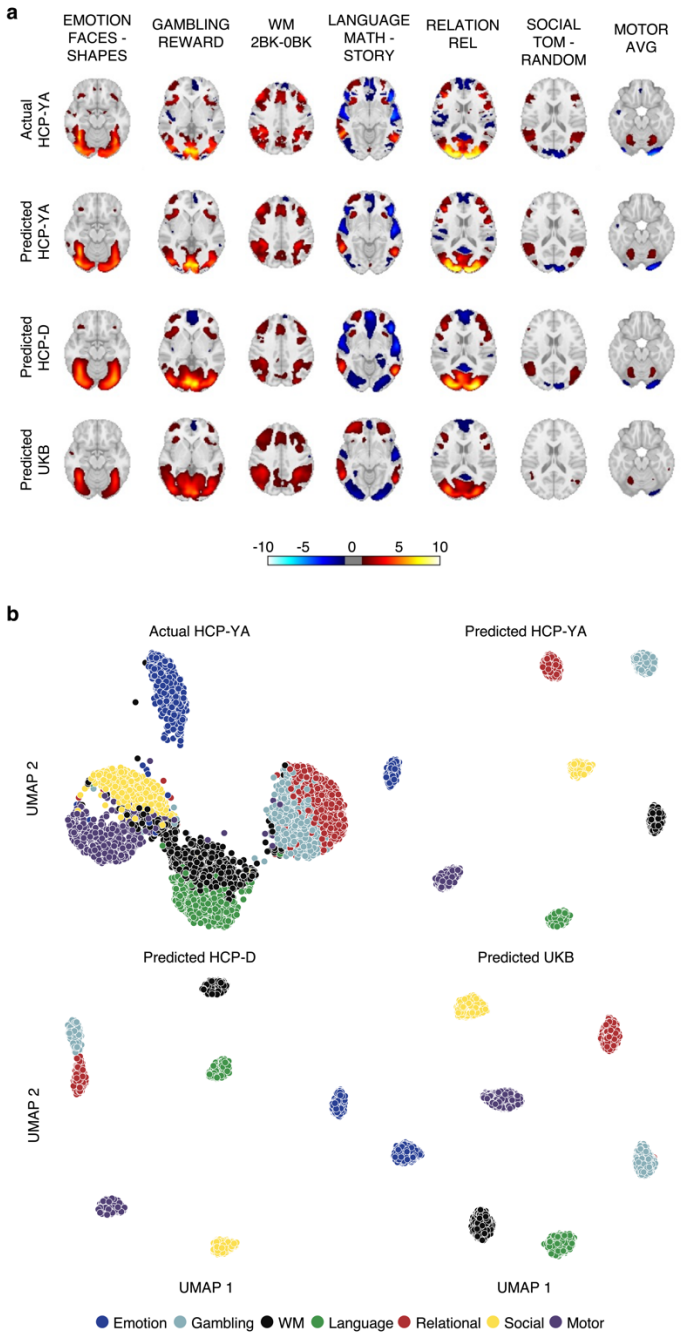
Fig 1d depicts the reconstruction and discriminability performance of the models for the task contrasts mentioned above. DeepTaskGen-based models showed significantly better reconstruction performance than the linear model

in the "EMOTION FACES-SHAPES" and "GAMBLING REWARD" contrasts (Supplementary Table 5). While the fine-tuned model with EMOTION FACES-SHAPES outperformed the non-fine-tuned model in the GAMBLING REWARD contrast ($t = 6.381$, $p = 0.002$), the fine-tuned model with GAMBLING REWARD performed significantly worse in the EMOTION FACES-SHAPES contrast ($t = -5.200$, $p = 0.002$) compared to the non-fine-tuned model. This suggests that the nature of the task is also important, and that fine-tuning is not guaranteed to improve reconstruction performance.

Regarding discriminability, the fine-tuned and non-fine-tuned models showed varying performance compared to the linear model (Supplementary Table 6). We noticed considerable variability in discriminability (and, to a lesser extent, reconstruction performance) among DeepTaskGen-based models across different task contrasts. In contrast, the linear model produced similar discriminability for all task contrasts, suggesting it is insensitive to differences between task contrasts, probably due to the model's simplicity.

In summary, DeepTaskGen-based models enable the prediction of trained task contrast maps on datasets lacking such maps. However, the benefit of fine-tuning varies depending on the specific task contrast.

Fig. 2: Visualization of group-level contrast maps and similarity (or dissimilarity) among subjects.



a. Group-level contrast maps for seven representative tasks were generated for predicted maps on three datasets (HCP Young Adult, HCP Development, and UK Biobank) and compared to actual task contrast maps from HCP-YA. Activations and deactivations are displayed using

scaled z-scores, with colors indicating the magnitude of the effect (shown in the color gradient bar). **b.** To visualize the similarity or dissimilarity between subjects based on their task contrast maps, we employed Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique, which projects the high-dimensional task contrast maps into two dimensions for clear visualization. Each dot represents a subject's task contrast map for the seven tasks, colored accordingly. Similar subjects are positioned closer together, indicating similarity, while the distance between dots represents dissimilarity.

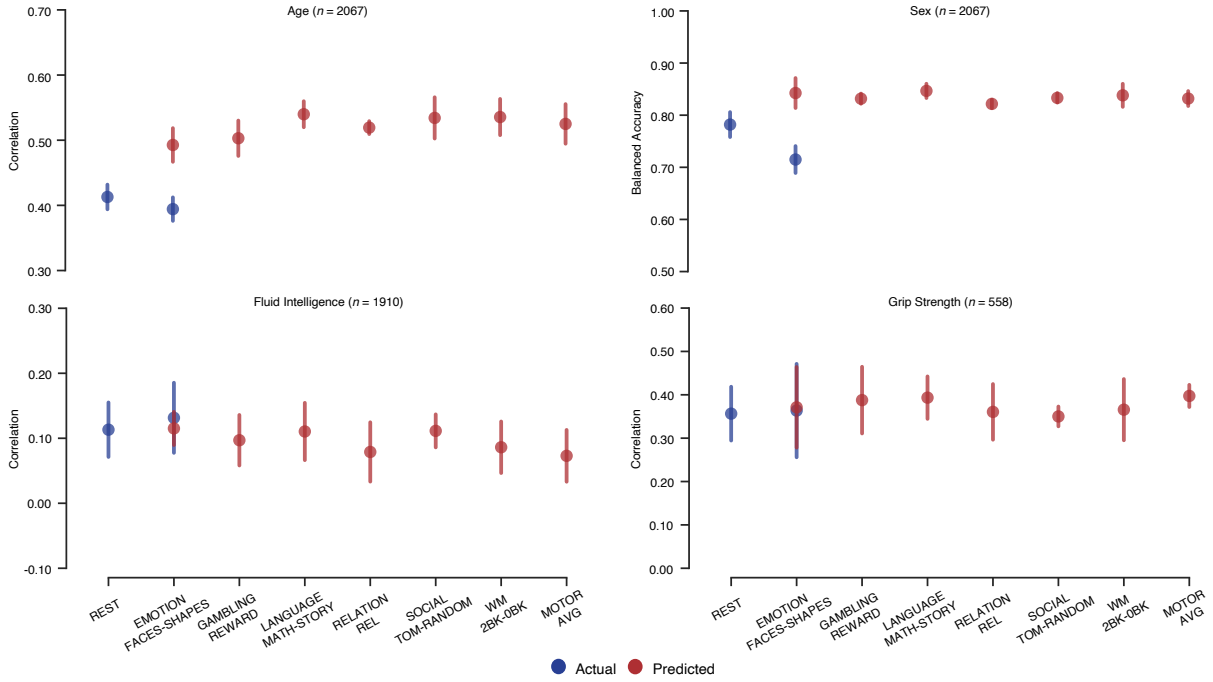
Predicted task contrast maps achieve stronger or comparable performance in predicting demographic, cognitive, and clinical variables

Next, we sought to demonstrate the utility of synthetic task images for downstream prediction tasks within the UK Biobank dataset. To achieve this, we fine-tuned our model on the UK Biobank dataset using EMOTION FACES-SHAPES task contrast ($\mu_{reconstruction}$ = 0.396, $\sigma_{reconstruction}$ = 0.136, discriminability = 0.354) and predicted seven representative contrasts from each task available in HCP-YA (Figs. 3-4).

To ensure that the predicted task contrast maps retain overall network structure and capture individual- and task-related variance for downstream analyses, we visualized them in two ways, along with the actual task contrast maps from the HCP-YA dataset as a reference. First, we generated group-average maps and observed excellent correspondence with the canonical networks for each task[12] and the actual task contrasts (Fig. 2a). Then, we employed Uniform Manifold Approximation and Projection (UMAP)[28] to

visualize high-dimensional differences between individuals and experimental tasks in a more intuitive two-dimensional space (Fig. 2b). We found that the predicted task contrast maps generated by DeepTaskGen contain lower within-task variance yet improved between-task variance compared to actual task contrast maps. This improvement can be attributed to the denoising process performed by our neural network during reconstruction, enhancing the signal-to-noise ratio.

Fig. 3: Prediction of subjects' age, sex, fluid intelligence, and dominant hand grip strength using task contrast maps and resting-state connectome on UK Biobank.

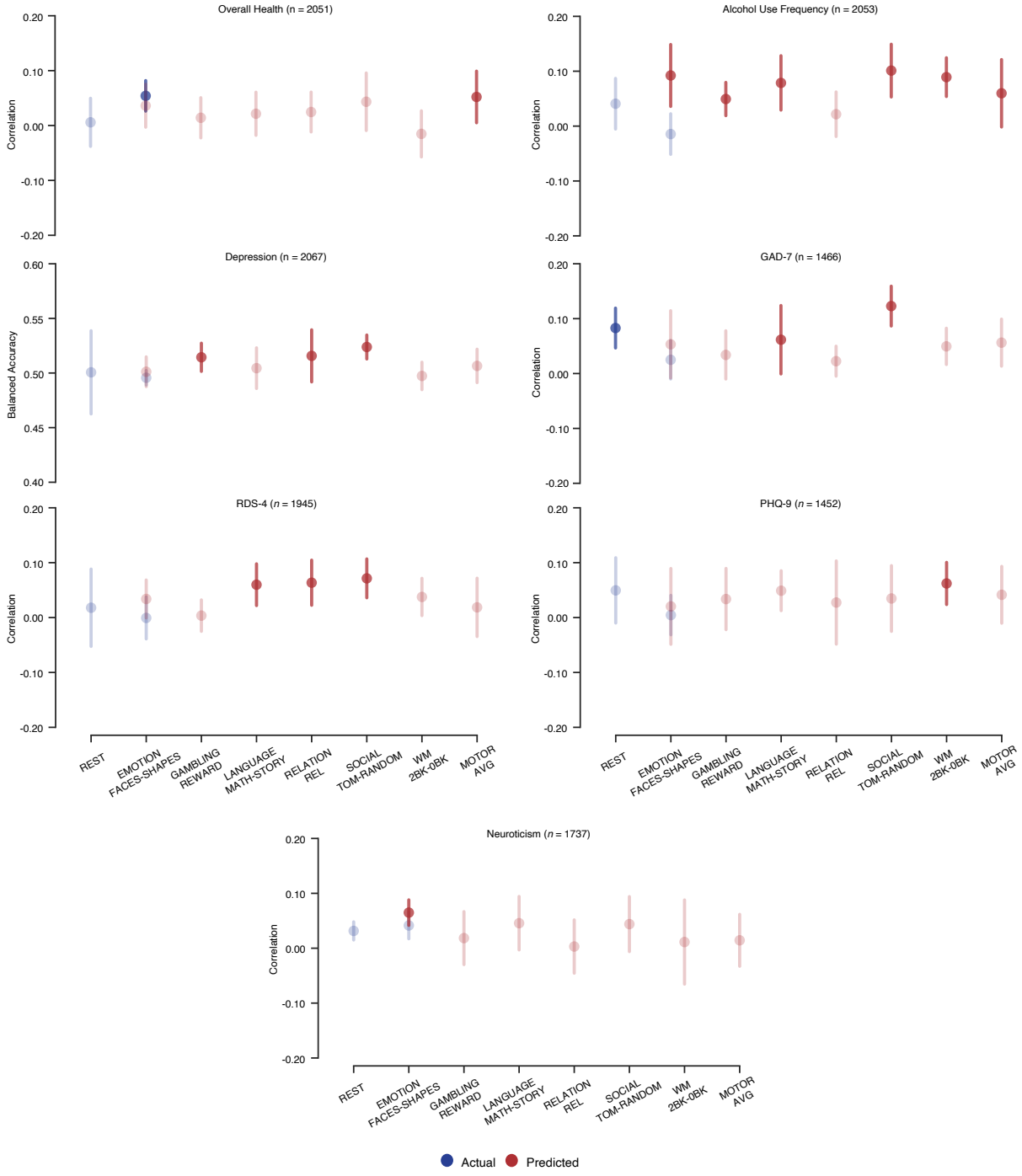


We predicted subjects' demographic and cognitive measures using three modalities: synthetic task contrast maps, actual task contrast maps, and resting-state connectome data. All predictions were made using L2 regularized regression within a 5-fold cross-validation

framework with permutation testing (1000 permutations). Blue represents resting-state connectome and actual task contrast maps, whereas red represents predicted task contrast maps. Significant predictions based on permutation testing are highlighted. Error bars indicate the standard deviation of prediction performance across five CV folds. Balanced accuracy was used to measure sex classification performance, while Pearson's correlation was employed to assess other variables. Sample sizes for all analyses are indicated in each figure. The detailed test statistics are given in Supplementary Tables 8-12.

We then compared the prediction performance of each of the seven predicted task contrast maps to that of actual task contrast maps and resting-state connectome data using permutation testing (1000 permutations). Specifically, we aimed to predict individuals' demographic, cognitive, and clinical measures, including age, sex, fluid intelligence, dominant hand grip strength, overall health, alcohol use frequency, depression status, depressive symptoms (Patient Health Questionnaire [PHQ]-9[29] and Recent Depressive Symptoms [RDS]-4[30]), anxiety symptoms (Generalized Anxiety Disorder Assessment [GAD]-7[31]) and neuroticism from predicted and actual task contrast maps, as well as resting-state connectome data. These were chosen because other studies have reported high prediction accuracy using these tasks[32,33] and because they are critical measures representing individuals' mental health. The exact IDs for the variables are given in Supplementary 7.

Fig. 4: Prediction of subjects' clinical measures using task contrast maps and resting-state connectome on UK Biobank.



We predicted subjects' clinical measures using three modalities: synthetic task contrast maps, actual task contrast maps, and resting-state connectome data. All predictions were made

using L2 regularized regression within a 5-fold cross-validation framework with permutation testing (1000 permutations). Actual and synthetic brain measures are depicted in blue and red colors, respectively. Significant predictions based on permutation testing are highlighted. Error bars indicate the standard deviation of prediction performance across five CV folds. Balanced accuracy was used to measure depression classification performance, while Pearson's correlation was employed to assess other variables. Sample sizes for all analyses are indicated in each figure. The detailed test statistics are given in Supplementary Tables 8,13,14.

The results, presented in Figures 3-4 and Supplementary Tables 8-14, show that the predicted task contrasts achieved greater or comparable prediction performance relative to actual EMOTION FACES-SHAPES task contrasts and resting-state connectome data for all variables. Specifically, we observed a large improvement in sex and age prediction using synthetic task contrasts, indicating that DeepTaskGen retains task activity patterns related to unique biological traits and amplifies their detection due to enhanced signal-to-noise ratios. More notably, for several clinically relevant measures, including RDS-4, PHQ-9, depression status, alcohol use frequency, and neuroticism, synthetic task contrasts achieved significant prediction, while actual brain measures could not survive permutation testing. This enhanced performance indicates that synthetic task contrast maps can retain amplified complex activation patterns related to subjects' clinical characteristics.

Discussion:

In this work, we introduce DeepTaskGen, a robust neural network architecture designed to predict task-based contrast maps using volumetric resting-state connectivity. This architecture enables the prediction of task contrasts on unseen datasets, even when task contrasts are unavailable or have not been acquired. We employed this method to generate task data for a comprehensive battery of cognitive tasks for the UK Biobank dataset and show that these synthetic task images retain individual-specific information essential for biomarker discovery.

We showed that our approach accurately predicts task-based contrast maps from resting-state connectivity in both the training dataset (HCP-YA) and unseen datasets (HCP-D and UKB). DeepTaskGen showed excellent reconstruction performance in that it outperformed retest scans in terms of more than half of the task contrasts and performed consistently better than a linear baseline model. While it may seem surprising that this approach performs better reconstruction than repeat scans, this finding aligns with previous work[25] and may be explained by the fact that session-specific information is present in both task-based and resting-state brain activity, with our model retaining this information during the generation process. More importantly, DeepTaskGen's generated task contrasts preserve individual differences relative to linear models and group-average maps. This indicates that these synthetic contrast images may be better suited as biomarkers for

individual-specific differences. Although group average maps show better reconstruction performance, they do not retain such individual-specific differences.

We performed extensive validation analyses on the UKB dataset to showcase the advantage of using predicted task contrast maps for predicting individual differences in an extensive set of demographics, cognitive, and clinical variables. Our results indicate that these predicted maps offer equivalent or better performance in all the scenarios we evaluated relative to actual EMOTION FACES-SHAPE task contrast images and resting-state connectome data. This is likely because our approach increases the signal-to-noise ratio in the synthetic data by (i) reducing inherent noise in functional MRI data and (ii) allowing an increase in the temporal degrees of freedom from resting-state scans that are frequently longer than task-based scans. Moreover, we observed variability in the predictive performance of various synthetic task contrasts, which indicates DeepTaskGen's capability to retain task-specific variability, also evidenced in the UMAP visualizations (Fig. 2b). This reinforces DeepTaskGen as a robust tool to facilitate predictive neuroscience.

Our method builds directly upon the foundational work of Tavor et al.[17] and Ngo et al.[25], who demonstrated the potential to predict task-based brain activity from task-free (i.e., resting-state brain) activity. The main benefit of our approach is the ability to generate task images in new cohorts where such

366 images were not acquired. However, these previous methods have additional
367 limitations. Tavor et al. employed a linear model, which fails to capture the
368 complexity of the shared network architecture between task-based and task-
369 free brain activity and does not effectively transfer learned parameters to
370 unseen datasets. Ngo et al.[25] introduced a neural network-based architecture,
371 but it is limited to the cortical surface, thereby excluding subcortical areas and
372 cerebellum that are crucial for cognitive functions[19–21] and mental disorders[22–
373 24]. In contrast, our proposed network, DeepTaskGen, overcomes these
374 drawbacks.

375 DeepTaskGen holds a wide range of implications in neuroscience
376 studies. One significant application is the prediction of contrast maps for age
377 groups for whom task data are difficult to acquire, such as children or
378 individuals with neurological and psychiatric disorders. We showed that
379 synthetic contrast maps facilitate predictions and biomarker studies within
380 populations where acquiring task-based fMRI data may be challenging, which
381 creates a significant potential to expand the scope of precision psychiatry.
382 However, the correspondence between resting-state and task-evoked brain
383 activation may vary across different groups, including those with disorders.
384 Therefore, a model trained solely on healthy controls might require careful
385 fine-tuning to ensure accurate predictions in disorder groups, which should be
386 a focus of future studies. Another important application of DeepTaskGen lies
387 in methods that utilize population data, such as normative modeling[34]. A

388 recent study modeling the heterogeneity in the Emotional Face Matching Task
389 using population-wide life-long normative modeling indicates a notable gap in
390 mid-adulthood age groups[35]. In this case, we believe DeepTaskGen can play
391 an important role in this context by augmenting contrast maps for the
392 underrepresented age group in their database. Additionally, while the UKB
393 dataset contributes a large portion of their data due to its sample size, it only
394 includes the Emotion Face Matching Task, limiting the replication of such
395 models for other fMRI tasks. DeepTaskGen can also address this challenge by
396 generating corresponding synthetic task contrasts by requiring only resting-
397 state fMRI data, which can then facilitate the development of more
398 comprehensive normative models. Furthermore, our method can benefit
399 large-scale imaging genetics studies, such as genome-wide association
400 studies (GWAS), where potential associations between genotypes and brain-
401 based phenotypes are studied. For instance, our approach allows pooling data
402 to test for associations between putative genetic variants and specific task-
403 evoked activity patterns. A recent study[36] revealed significant genotype-
404 phenotype associations using multimodal neuroimaging data from UKB, but it
405 still faces the limitations mentioned above. Thus, DeepTaskGen can also
406 enhance GWAS by enabling more specific and task-related associations using
407 synthetic task contrasts.

408 It is important to note the difference in variance distribution between
409 predicted and actual task-based contrast maps. Actual images contain task-

410 related variation (i.e., signal) and epistemic variation (or noise)[34]. During the
411 prediction process, such noise can be effectively removed, amplifying the
412 variance across experimental tasks by retaining only task-related variance.
413 This can be seen in the UMAP visualizations in Fig. 2b, where predicted maps
414 showed lower within-task contrast variance but higher between-task contrast
415 variance than actual maps. Such removal of epistemic noise results in an
416 increased signal-to-noise ratio, leading to improved prediction performance in
417 our validation analyses. While this may not be critical for traditional statistical
418 methods like group comparisons, it becomes more important when these
419 images are used in complex models where noise is explicitly modeled.
420 Therefore, we strongly recommend careful model designs that consider this
421 difference when incorporating both data types within a single model.

422 Several methodological considerations should be noted when
423 interpreting our results. First, the spatial Independent Component Analysis
424 (ICA) and task contrast maps provided in the HCP-YA dataset are surface-
425 based (i.e., fs_LR surface). We projected these surface-based images to the
426 volumetric MNI space using a well-established method, Registration Fusion[37],
427 which, despite its promise, might not be as optimal as direct registration to
428 the coordinate system during preprocessing. Therefore, the projected contrast
429 maps may have potential differences from those originally preprocessed in
430 volumetric space, which might impact prediction performance. Second, we
431 used the mean squared error (MSE) loss function to train our model. Although

DeepTaskGen yielded significantly greater discriminability than the linear model and group-averaged contrast maps, the MSE loss may still bias predictions toward the group average, limiting DeepTaskGen's performance. A previous approach, BrainSurfCNN[25], proposed a reconstructive-contrastive loss to enhance inter-subject variance in predicted contrast maps while minimizing within-subject differences. While the authors observed increased inter-subject variance, they noted its ineffectiveness for fine-tuning, which is the main objective of our study. Additionally, the relationship between reconstructive and contrastive loss can become complex in high-dimensional data, yielding contradictory results[38]. We also observed a similar pattern, where reconstructive performance decreased with increasing inter-subject contrast. This issue may be addressed with hyperparameter optimization of the loss function, but this necessitates expensive hyperparameter optimization and can introduce considerable inter-researcher variance. Another recent approach suggests using residualized connectome and contrast maps (i.e., images subtracted by a mean image), forcing the model to learn only individual differences[38]. Although residualization might simplify the training procedure, it is not feasible for transfer learning as such residualization cannot be effectively done in unseen datasets. Therefore, further exploration of input data and loss function strategies may enhance discriminability.

In summary, we have demonstrated DeepTaskGen, a general-purpose tool for generating arbitrary synthetic task images from resting-state data. We consider that this will facilitate the study of individual differences and the generation of task-related biomarkers.

Methods:

Datasets:

Supplementary Table 2 provides the sample details and function scan acquisition parameters for the datasets used in this study.

Human Connectome Project Young-Adult (HCP-YA): The Human Connectome Project Young-Adult (HCP-YA) 1200 Release was used as a main dataset to train and evaluate models[11]. 953 participants comprising all the task and resting-state sessions were included. Of those 953 subjects, 34 have re-test 3T fMRI data from the second visit, which we used as an evaluation set. Rs-fMRI data were acquired in four 15-minute runs, with 1200 timepoints per run, totaling 4800 timepoints per subject. However, since fewer timepoints (e.g., 300-500) are common in neuroimaging datasets and stable functional connectivity can be computed with shorter resting-state acquisitions[27,39], we divided resting-state data into eight contiguous parts (600 timepoints) as previously shown[25]. Like previous studies[17,25], we used ROIs from the 50-component parcellation derived from spatial ICA released within HCP-YA to compute individual functional connectivity. As tb-fMRI data, we used 47 unique contrasts from 7 main task domains[12]: WORKING MEMORY, EMOTION, SOCIAL, RELATIONAL, LANGUAGE, MOTOR, and GAMBLING. We used minimally preprocessed and FIX-cleaned fMRI data provided by HCP[40]. While HCP also provides volumetric rs-fMRI data, the spatial ICA maps, and z-transformed task-activation maps are only surface-based. Therefore, we

479 projected surface-based data to volumetric space using the registration fusion
480 (RF-ANTs) method[37]. Registration Fusion is a popular method that allows
481 nonlinear mapping between MNI152 and fsaverage coordinate systems using
482 coordinate mapping computed on 745 subjects from the Brain Genomics
483 Superstruct Project (GSP)[41]. The details of the mapping process are given
484 elsewhere[37]. Subsequently, the scan images were downsampled to 2x2x2 mm
485 voxel resolution and cropped to the cortex (resulting image dimension is
486 76x93x78) to reduce the computational costs.

487 ***Human Connectome Project Development (HCP-D):*** The
488 evaluation of DeepTaskGen's generalizability across different populations with
489 various age groups and transferability of learned parameters to unseen
490 datasets was conducted on the HCP-Development (HCP-D) dataset from the
491 HCP-Lifespan project[15]. 637 subjects that had completed four resting-state
492 runs (488 timepoints per run) and Emotion[1] and Guessing[42] tasks were
493 included in the study. The choice of Emotion and Guessing tasks was due to
494 the Emotion task performed in both the main HCP-YA and HCP-D projects,
495 while the Guessing task served as an adaptation of the Gambling task in HCP-
496 YA, with some modifications. Similarly, the minimally preprocessed volumetric
497 resting state (e.g., rfMRI_REST_hp0_clean.nii.gz) and task-fMRI (e.g.,
498 tfMRI_GUESSING_PA_hp0_clean.nii.gz) images provided by the HCP were
499 used in our study. The voxel-to-ROI matrices for each subject within the
500 dataset were computed using the same ICA-based parcellation used in the

training dataset. The task-based contrast maps were calculated using "nilearn.glm.first_level.FirstLevelModel" method from Nilearn[43]. Specifically, the first level model with a high-pass filter (0.01 Hz), "glover" hemodynamic response function, smoothing with 6 mm FWHM Gaussian kernel, and z-transformation was applied to minimally preprocessed task-based images. Nuisance correction was done using the Friston 24 motion parameters.

UK Biobank (UKB): To assess the practical applicability of generated contrast maps and further generalizability of the trained model on a large-scale sample, we used the UK Biobank dataset[26]. A total of 20,792 subjects with resting-state (490 timepoints) and Emotion task[1] MRI images were included in this study. We used the consortium-processed dataset, which was preprocessed using an FSL-based pipeline, which includes unwarping, motion correction, fieldmap correction, registration, normalization, automatic noise selection as implemented in ICA-AROMA[44], and smoothing using a 5 mm FWHM Gaussian kernel. Following preprocessing, the same ICA-based parcellation used in the previous analyses was utilized to extract general-purpose resting-state images and compute voxel-to-ROI time series. Z-transformed contrast maps for the Emotion task were computed using FSL. Subjects with a mean relative root mean square head motion greater than 0.5 mm were excluded.

523 **Voxel-to-ROI functional connectivity:**

524 DeepTaskGen employs resting-state voxel-to-ROI connectivity to
525 predict task contrasts. Voxel-to-ROI connectivity is computed as Pearson's
526 correlation between each voxel's time series and the average signal of the
527 target ROI[45]. The target ROIs are derived from components obtained through
528 group-level independent component analysis (ICA) applied to resting-state
529 fMRI time series[46]. Voxel-to-ROI connectivity provides an increased spatial
530 resolution compared to the conventional ROI-to-ROI approach, which requires
531 deeper models at the generation step. Although voxel-to-voxel connectivity
532 could produce a higher spatial resolution than voxel-to-ROI connectivity, it is
533 computationally infeasible due to excessive GPU memory use. Additionally, as
534 mentioned earlier, the voxel-to-ROI approach allows for modeling the
535 connectivity of the entire cortex, providing a more inclusive solution than the
536 vertex-to-ROI approach.

537

538 **DeepTaskGen:**

539 The Architecture of DeepTaskGen is depicted in Fig 1. DeepTaskGen is
540 a volumetric adaptation of BrainSurfCNN[25], a surface-based convolutional
541 neural network based on U-Net architecture[47] widely used in biomedical image
542 segmentation. Like autoencoders, U-Net consists of encoding (i.e.,
543 downsampling), decoding (i.e., upsampling) blocks, and a bottleneck block.
544 Specifically, DeepTaskGen's network architecture is as follows: each block in

545 the encoding and decoding arms consists of a 3D convolutional layer, batch
546 normalization, and non-linear ReLU activation function. Additionally, max
547 pooling (MaxPool3d) is applied before each encoding block to reduce the
548 dimensions of the input image by half by retaining task-related information.
549 In contrast, compressed images are upsampled before each decoding block to
550 match the output dimensionality with actual maps. The details of the
551 architecture are provided in Supplementary Table 1.
552
553 **Model Training:**
554 DeepTaskGen was trained on HCP-YA (N_{train} = 827, $N_{validation}$ = 92, N_{test}
555 = 34) for 100 epochs using a batch size of 12, mean squared error (MSE) as
556 the loss function, and the Adam optimizer. Unlike the previous study[25], we did
557 not utilize the reconstructive-contrastive loss in our training process as we
558 found no noticeable training advantage from using this loss when transferring
559 the model to unseen datasets, which was also reported in the previous study.
560 The model training was conducted on a single Nvidia Tesla V100-SXM2-32GB
561 GPU and took approximately 8 hours and 30 minutes.
562
563 **Model transfer to HCP-D and UKB:**
564 Our main objective for the HCP-D dataset was to validate the
565 performance of DeepTaskGen to predict unavailable arbitrary task contrast
566 maps on an unseen dataset. For instance, we fine-tuned the model using the

EMOTION FACES-SHAPES contrast to predict the GAMBLING REWARD contrast. This approach relies on the assumption that the weights in the trained model's output layer are specific to a given task contrast (e.g., GAMBLING REWARD), while the backbone layers (i.e., all the layers before the output layer) contain information shared across multiple task contrasts, such as dataset-related features. By adapting only the backbone of the model, the trained model can effectively predict the contrasts from the HCP-YA dataset, where the original training has been completed. To achieve this, we froze the parameters of our trained model's output layer and fine-tuned the backbone of our model on the HCP-D dataset using either GAMBLING-REWARD or EMOTION FACES-SHAPES contrast. Specifically, we fine-tuned the pre-trained model using only a single task contrast (e.g., GAMBLING-REWARD) for 50 epochs. We used 515 and 58 subjects for training and validation during fine-tuning, respectively. Subsequently, we replaced the output layer of the fine-tuned model with one that matched the target task contrast (e.g., EMOTION FACES-SHAPES) and predicted the target task contrast maps for 64 test subjects (i.e., holdout set). The model transfer performance of DeepTaskGen was also compared with that of a non-fine-tuned model and linear regression.

The initially trained model was further fine-tuned on UKB (N_{train} = 16840, $N_{validation}$ = 1872, N_{test} = 2080) to test the practical applicability of predicted task contrast maps. We fine-tuned the backbone of the trained model using the EMOTION FACES-SHAPES task contrast available in UKB for

50 epochs. Then, we replaced the coefficients in the output layer of the fine-tuned model with corresponding model coefficients of the 47 task contrasts available in the training dataset to predict these task contrast maps for 2080 test subjects.

Baselines:

Group-average contrasts: The group-average contrasts were utilized as a reference to represent the features in the task contrast that are shared among individuals. A simple predictive model can effectively capture these common features. By comparing the performance of the predicted task contrasts with the group-average contrasts, we can assess the degree to which the predicted task contrasts capture individual-specific information beyond what is captured by the common features represented in the group-average contrasts. It could also show the degree to which variation would be in each task contrast, e.g., a high correlation between true and group-average task contrasts indicates a reduced level of inter-individual variation.

Retest scans: The additional tb-fMRI data acquired in a second run (34 subjects from HCP-YA) was used to evaluate the reliability of the task contrasts.

Linear Model: We implemented this baseline model based on the approach presented by Tavor and colleagues[17]. Briefly, for each ROI (50 in total), the input connectivity and target contrast maps of subjects in the

training set were vectorized, and a linear regression model was fit to map between input and output vectors. This resulted in a weight matrix with dimensions of R x S, where R represents the number of ROIs and S represents the number of subjects in the training set. This weight matrix was averaged across subjects to create a vector containing average weights for each ROI. The task contrast for subjects from the test set was then predicted using this weight vector. This procedure was repeated for each of the 47 task contrasts.

Visualization of group-level contrast maps and inter-subject and inter-task similarity:

To evaluate the structure of the predicted task contrast maps, we visualized both common structural elements across subjects (Figure 2a) and similarities or differences between subjects for various task contrasts (Figure 2b). Figure 2a displays the common brain structure by presenting the voxel-based sample average of the task contrast maps across all subjects.

We employed a non-linear dimensionality reduction method, Uniform Manifold Approximation and Projection (UMAP)[28], to visualize the similarity or variance between subjects and across various tasks. UMAP preserves global and local topological structure while embedding high-dimensional data into a lower-dimensional space. Specifically, UMAP was utilized to project the high-dimensional task-based contrast maps into a two-dimensional space (i.e., two components), where each dot represents a subject's task contrast map. This

633 allows for a straightforward interpretation of similarity or dissimilarity between
634 subjects and fMRI tasks by considering the distance between dots on the two-
635 dimensional plot. We implemented UMAP in Python using the umap-learn
636 package with default hyperparameters (n_neighbors=15, metric=euclidean).
637

638 **Prediction of cognitive and demographic data for validation:**

639 To validate the clinical relevance of our model, we predicted multiple
640 non-imaging variables using the synthetic images derived from the UK
641 Biobank dataset. Specifically, we aimed to predict subjects' cognition- and
642 health-related variables, including age, sex, fluid intelligence, grip strength of
643 the dominant hand, overall health, alcohol use frequency, depression status,
644 depressive symptoms (PHQ-9[29] and RDS-4[30]), anxiety symptoms (GAD-7[31]),
645 and neuroticism. We used a similar L2 regularized linear model and training
646 and evaluation procedure for all these prediction analyses (see below). The
647 exact sample size for each prediction is provided in Figures 3-4.

648 All regression models were built using predicted contrast maps for
649 seven main task contrasts (EMOTION FACES-SHAPES, GAMBLING REWARD,
650 WM 2BK-0BK, LANGUAGE MATH-STORY, RELATION REL, SOCIAL TOM-
651 RANDOM, MOTOR AVG) as well as subjects' actual EMOTION FACES-SHAPES
652 contrast maps and resting-state connectome as comparison. The resting-state
653 connectome for each subject was computed as Pearson’s correlation between
654 each time series extracted using the 50-component ICA parcellation described

earlier (50 x 50 correlation matrix for each individual). Feature matrices in prediction analyses were constructed by taking all voxel values within the brain (235,375 features) or extracting the upper triangle of the correlation matrix (1,225 features) for each subject. We predicted variables using L2 regularized linear regression (i.e., ridge regression). To mitigate the possible effect of the difference in the number of features derived from task-based contrast maps and resting-state connectome, the alpha value, which controls regularization strength, was set to a value computed by dividing the number of features by the training sample size. This resulted in significantly more regularization for task-based features than connectome-based features. Model evaluation was conducted within a 5-fold cross-validation framework and tested for significance using permutation testing (P = 1000). Following predictions, we compared the predictive performance of synthetic contrast maps with actual contrast maps and resting-state connectome using permutation testing. Specifically, we predicted variables using true labels and compared 5-fold CV performance using a paired t-test. To create a null sampling distribution for the t-values, we shuffled labels and repeated the comparisons 1000 times. Then, we computed the p-value as the proportion of permutations resulting in t-values as extreme as or more extreme than the observed t-values. Note that predictions were performed on the 2080 subjects we spared as a test set during the fine-tuning process explained earlier.

677 **Measures of model performance:**

678 ***Reconstruction performance:*** Reconstruction performance is one of
679 the primary measures for evaluating a model's performance. It is determined
680 by calculating Pearson's correlation between the predicted and actual task
681 contrast maps for unseen subjects (i.e., those not used in training) in the test
682 set.

683 ***Discriminability score:*** While reconstruction performance is essential
684 for evaluating a model's capability to generate task contrast maps, it is
685 insufficient for assessing its ability to preserve inter-individual differences in
686 these synthetic maps, which is crucial for downstream biomarker development
687 and individual-specific predictions. Therefore, we employed the fingerprinting
688 score methodology introduced by Finn and colleagues[27], which we refer to as
689 the discriminability score in our study. This score is computed by correlating
690 each subject's predicted task contrast with their actual task contrast and with
691 the actual task contrasts of the remaining subjects and calculating the fraction
692 of subjects whose predicted task contrast exhibits the highest correlation with
693 their actual task contrast (i.e., argmax). Given the correlation between the
694 group-average task contrast and each subject’s actual task contrast map is
695 similar, we expect the discriminability score for the group-average task
696 contrast to be near zero. This allows us to assess to what extent the synthetic
697 task contrast maps can capture individual-specific information necessary for

downstream tasks, such that higher scores indicate that the synthetic maps go beyond average information and contain individual-specific characteristics.

Prediction performance: The predictive power of both actual and predicted task contrast maps was evaluated using Pearson's correlation coefficient for the dominant hand's grip strength, brain age, fluid intelligence, and overall health. For sex classification, balanced accuracy, defined as the arithmetic mean of the true positive rate and true negative rate, was used as the evaluation metric.

Technology use disclosure:

Large language models were used for proofreading (e.g., grammar, typos, integrity, and clarity checks) and as assistance during Python programming. All authors have read, corrected, and verified all information presented in this manuscript and Supplementary Information.

Data availability:

The minimally preprocessed HCP-YA S1200 data can be accessed and downloaded from the following link: https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release. The ICA-based group-averaged parcellations used to construct resting-state connectomes are available at: https://www.humanconnectome.org/study/hcp-young-

adult/document/extensively-processed-fmri-data-documentation. Similarly, preprocessed data for the HCP-D study can be downloaded from https://www.humanconnectome.org/study/hcp-lifespan-development/data-releases. The UK Biobank (UKB) dataset is publicly available at https://www.ukbiobank.ac.uk/.

Code availability:

DeepTaskGen model code can be found on GitHub: https://github.com/eminSerin/DeepTaskGen. Python scripts to reproduce all the results presented in this paper are given at https://github.com/eminSerin/deeptaskgen-paper. Additionally, trained models to generate task contrast maps on various open datasets are publicly available at https://github.com/eminSerin/deeptaskgen-models.

Acknowledgement:

Funded by the European Union. Complementary funding was received by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (10041392 and 10038599). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, the European Health and Digital Executive Agency (HADEA), or UKRI. Neither the European Union nor HADEA nor UKRI can be held responsible for them. This work has also been supported by a Grant from

742 the German Research Foundation to the ENIGMA task-based fMRI Working
743 Group (DFG ER 724/4–1, WA 1539/11–1). Data used in this study were
744 provided in part by the Human Connectome Project, WU-Minn Consortium
745 (principal investigators: D. Van Essen and K. Ugurbil; grant number
746 1U54MH091657), funded by the 16 National Institutes of Health (NIH)
747 institutes and centers supporting the NIH Blueprint for Neuroscience
748 Research, and by the McDonnell Center for Systems Neuroscience at
749 Washington University. Additionally, this research utilized data obtained from
750 UK Biobank, a large-scale biomedical database.

751 **References:**

752 1. Hariri, A. R., Bookheimer, S. Y. & Mazziotta, J. C. Modulating emotional
753 responses: effects of a neocortical network on the limbic system. *Neuroreport* 11,
754 43–48 (2000).
755 2. Swartz, J. R., Knodt, A. R., Radtke, S. R. & Hariri, A. R. A neural biomarker of
756 psychological vulnerability to future life stress. *Neuron* 85, 505–511 (2015).
757 3. Gao, S. Combining multiple connectomes improves predictive modeling of
758 phenotypic measures. 9 (2019).
759 4. Greene, A. S., Gao, S., Scheinost, D. & Constable, R. T. Task-induced brain
760 state manipulation improves prediction of individual traits. *Nat. Commun.* 9,
761 2807 (2018).
762 5. Gal, S., Coldham, Y., Tik, N., Bernstein-Eliav, M. & Tavor, I. Act natural:
763 Functional connectivity from naturalistic stimuli fMRI outperforms resting-state
764 in predicting brain activity. *NeuroImage* 258, 119359 (2022).
765 6. Sripada, C., Angstadt, M., Rutherford, S., Taxali, A. & Shedden, K. Toward a
766 'treadmill test' for cognition: Improved prediction of general cognitive ability
767 from the task activated brain. *Hum. Brain Mapp.* 41, 3186–3197 (2020).
768 7. Jiang, R. *et al.* Task-induced brain connectivity promotes the detection of
769 individual differences in brain-behavior relationships. *NeuroImage* 207, 116370
770 (2020).
771 8. Shah, L. M., Cramer, J. A., Ferguson, M. A., Birn, R. M. & Anderson, J. S.
772 Reliability and reproducibility of individual differences in functional connectivity
773 acquired during task and resting state. *Brain Behav.* 6, e00456 (2016).
774 9. Tetereva, A., Li, J., Deng, J. D., Stringaris, A. & Pat, N. Capturing brain-

775 cognition relationship: Integrating task-based fMRI across tasks markedly boosts
776 prediction and test-retest reliability. *NeuroImage* 263, 119588 (2022).
777 10. Woo, C.-W., Chang, L. J., Lindquist, M. A. & Wager, T. D. Building better
778 biomarkers: brain models in translational neuroimaging. *Nat. Neurosci.* 20, 365–
779 377 (2017).
780 11. Van Essen, D. C. *et al.* The WU-Minn Human Connectome Project: An overview.
781 *NeuroImage* 80, 62–79 (2013).
782 12. Barch, D. M. *et al.* Function in the human connectome: Task-fMRI and individual
783 differences in behavior. *NeuroImage* 80, 169–189 (2013).
784 13. Schumann, G. *et al.* The IMAGEN study: reinforcement-related behaviour in
785 normal brain function and psychopathology. *Mol. Psychiatry* 15, 1128–1139
786 (2010).
787 14. Casey, B. J. *et al.* The Adolescent Brain Cognitive Development (ABCD) study:
788 Imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* 32, 43–54 (2018).
789 15. Somerville, L. H. *et al.* The Lifespan Human Connectome Project in
790 Development: A large-scale study of brain connectivity development in 5–21
791 year olds. *NeuroImage* 183, 456–468 (2018).
792 16. Cole, M. W., Ito, T., Bassett, D. S. & Schultz, D. H. Activity flow over resting-
793 state networks shapes cognitive task activations. *Nat. Neurosci.* 19, 1718–1726
794 (2016).
795 17. Tavor, I. *et al.* Task-free MRI predicts individual differences in brain activity
796 during task performance. *Science* 352, 216–220 (2016).
797 18. Smith, S. M. *et al.* Correspondence of the brain's functional architecture during
798 activation and rest. *Proc. Natl. Acad. Sci. U. S. A.* 106, 13040–13045 (2009).

19. Koziol, L. F. & Budding, D. E. *Subcortical Structures and Cognition*. (Springer New York, New York, NY, 2009). doi:10.1007/978-0-387-84868-6.
20. Janacsek, K. *et al.* Subcortical Cognition: The Fruit Below the Rind. *Annu. Rev. Neurosci.* 45, 361–386 (2022).
21. Stoodley, C. J. The Cerebellum and Cognition: Evidence from Functional Imaging Studies. *The Cerebellum* 11, 352–365 (2012).
22. Okada, N. *et al.* Subcortical volumetric alterations in four major psychiatric disorders: a mega-analysis study of 5604 subjects and a volumetric data-driven approach for classification. *Mol. Psychiatry* (2023) doi:10.1038/s41380-023-02141-9.
23. Ho, T. C. *et al.* Subcortical shape alterations in major depressive disorder: Findings from the ENIGMA major depressive disorder working group. *Hum. Brain Mapp.* 43, 341–351 (2022).
24. Hoppenbrouwers, S. S., Schutter, D. J. L. G., Fitzgerald, P. B., Chen, R. & Daskalakis, Z. J. The role of the cerebellum in the pathophysiology and treatment of neuropsychiatric disorders: A review. *Brain Res. Rev.* 59, 185–200 (2008).
25. Ngo, G. H., Khosla, M., Jamison, K., Kuceyeski, A. & Sabuncu, M. R. Predicting individual task contrasts from resting-state functional connectivity using a surface-based convolutional network. *NeuroImage* 248, 118849 (2022).
26. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536 (2016).
27. Finn, E. S. *et al.* Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nat. Neurosci.* 18, 1664–1671 (2015).

28. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018) doi:10.48550/ARXIV.1802.03426.
29. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16, 606–613 (2001).
30. Dutt, R. K. *et al.* Mental health in the UK Biobank: A roadmap to self-report measures and neuroimaging correlates. *Hum. Brain Mapp.* 43, 816–832 (2022).
31. Spitzer, R. L., Kroenke, K., Williams, J. B. W. & Löwe, B. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Arch. Intern. Med.* 166, 1092 (2006).
32. Gong, W., Bai, S., Zheng, Y.-Q., Smith, S. M. & Beckmann, C. F. Supervised Phenotype Discovery From Multimodal Brain Imaging. *IEEE Trans. Med. Imaging* 42, 834–849 (2023).
33. He, T. *et al.* Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* 25, 795–804 (2022).
34. Marquand, A. F. *et al.* Conceptualizing mental disorders as deviations from normative functioning. *Mol. Psychiatry* 24, 1415–1424 (2019).
35. Savage, H. S. *et al. Unpacking the Functional Heterogeneity of the Emotional Face Matching Task: A Normative Modelling Approach*. Preprint at http://biorxiv.org/lookup/doi/10.1101/2023.03.27.534351 (2023).
36. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 562, 210–216 (2018).
37. Wu, J. *et al.* Accurate nonlinear mapping between MNI volumetric and FreeSurfer surface coordinate systems. *Hum. Brain Mapp.* 39, 3793–3808 (2018).

38. Zheng, Y.-Q. *et al.* Accurate predictions of individual differences in task-evoked brain activity from resting-state fMRI using a sparse ensemble learner. *NeuroImage* 259, 119418 (2022).
39. Tik, N. *et al.* Generalizing prediction of task-evoked brain activity across datasets and populations. *NeuroImage* 276, 120213 (2023).
40. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* 80, 105–124 (2013).
41. Holmes, A. J. *et al.* Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Sci. Data* 2, 150031 (2015).
42. Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C. & Fiez, J. A. Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* 84, 3072–3077 (2000).
43. Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Front. Neuroinformatics* 8, (2014).
44. Pruim, R. H. R. *et al.* ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *NeuroImage* 112, 267–277 (2015).
45. Khosla, M., Jamison, K., Kuceyeski, A. & Sabuncu, M. R. Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. *NeuroImage* 199, 651–662 (2019).
46. Smith, S. M. *et al.* Resting-state fMRI in the Human Connectome Project. *NeuroImage* 80, 144–168 (2013).
47. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W.

M. & Frangi, A. F.) vol. 9351 234–241 (Springer International Publishing, Cham, 2015).