bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cross Atlas Remapping via Optimal Transport (CAROT): Creating connectomes for any atlas when raw data is not available

Javid Dadashkarimi[1,*], Amin Karbasi[1,2], and Dustin Scheinost[3]

[1]Computer Science Department, Yale University, New Haven, CT, 06511, USA
[2]Electrical Engineering, Yale University, New Haven, CT, 06511, USA
[3]Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT, 06511, USA
[*]javid.dadashkarimi@yale.edu

ABSTRACT

Whether using large-scale projects—like the Human Connectome Project (HCP), the Adolescent Brain Cognitive Development (ABCD) study, Healthy Brain Network (HBN), and the UK Biobank—or pooling together several smaller studies, open-source, publicly available datasets allow for unpresented sample sizes and promote generalization efforts. Overall, releasing preprocessing data can enhance participant privacy, democratize science, and lead to unique scientific discoveries. But releasing preprocessed data also limits the choices available to the end-user. For connectomics, this is especially true as connectomes created from different atlases (*i.e.*, ways of dividing the brain into distinct regions) are not directly comparable. In addition, there exist several atlases with no gold standards, and more being developed yearly, making it unrealistic to have processed, open-source data available from all atlases. To address these limitations, we propose Cross Atlas Remapping via Optimal Transport (CAROT) to find a mapping between two atlases, allowing data processed from one atlas to be directly transformed into a connectome based on another atlas without needing raw data. To validate CAROT, we compare reconstructed connectomes against their original counterparts (i.e., connectomes generated directly from an atlas), demonstrate the utility of transformed connectomes in downstream analyses, and show how a connectome-based predictive model can be generalized to publicly available processed data that was processed with different atlases. Overall, CAROT can reconstruct connectomes from an extensive set of atlases—without ever needing the raw data—allowing already processed connectomes to be easily reused in a wide-range of analyses while eliminating wasted and duplicate processing efforts. We share this tool as both source code and as a stand-alone web application (`http://carotproject.com/`).

1 Introduction

Open-source, publicly available datasets have revolutionized modern functional neuroimaging. They allow for unprecedented sample-sizes and statistical power, reduce barriers of entry to new investigators, and provide a treasure trove of data for replication and generalization efforts. Large-scale project (e.g., the Human Connectome Project (HCP), the Adolescent Brain Cognitive Development (ABCD) study[1], Healthy Brain Network (HBN)[2], and the UK Biobank[3]) as well as the sharing of smaller studies have been largely embraced by the neuroimaging community with data from nearly 100,000 individuals available open-source. While sharing raw data is considered best practice, as it imposes little restrictions on how the data is analyzed, releasing preprocessed data—ranging from minimal processing, like de-facing, to completely analyzed statistical maps—has several advantages. First, not all data can be shared in its raw form due to privacy concerns of being able to identify a participant based on unprocessed data. Second, as datasets become larger, fewer and fewer labs will have the resources to process datasets that range in the 1,000-10,000's of participants. Curated releases increase the wider-use of these large datasets, while reducing redundant processing efforts[4]. Third, even completely analyzed statistical maps, like those stored on NeuroVault[5], can be used in meta-analyses[6] or to improve machine learning algorithm[7]. Overall, releasing preprocessing data can enhance participant privacy, democratize science, and lead to unique scientific discoveries.

For connectome-based approaches, lack of access to raw data can be barrier for combining data across open-source dataset. Connectomes are created by parcellating the brain into distinct regions using an atlas (i.e., nodes of a graph or network) and estimating the connections between these regions (i.e., the edges of a graph or network). As different atlases divide the brain into a different number of regions of varying size and topology, connectomes created from different atlases are not directly comparable. In other words, results generated from one atlas cannot be directly applied to processed, open-source data from a different atlas. For example, it is not possible to directly combine the fully processed connectomes provided by two of the largest open-source datasets (the UK Biobank and ABCD studies) as different atlases are used to create these connectomes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Nearly 10,000 participants from the ABCD study would need to be re-processed to match the ∼55,000 participants in the UK Biobank. Further, there exist several atlases with no gold standards, and more are being developed yearly, making it unrealistic to have processed, open-source data available from all atlases. Being able transform to an existing connectome generated from one atlas to a connectome generated from another atlas would allow that existing connectome to be combined with other open-source data for use in pooled analyses or replication and generalization efforts.

To this aim, we propose Cross Atlas Remapping via Optimal Transport (CAROT), which uses optimal transport theory, or the mathematics of converting a probability distribution from one set to another, to find an optimal mapping between two atlases that allows data processed from one atlas to be directly transformed into a connectome based on an unavailable atlas without needing raw data. CAROT is designed for functional connectomes based on functional magnetic imaging (fMRI) data. First, in a training sample with fMRI timeseries data from two different atlases, we find this mapping by solving the Monge–Kantorovich transportation problem[8]. Then, by employing this optimal mapping, timeseries data based on the first atlas (from individuals independent of the training data) can be reconstructed into connectomes based on the second atlas without ever needing to use the second atlas. To validate CAROT, we compare reconstructed connectomes against their original counterparts (i.e., connectomes generated directly from an atlas), demonstrate the utility of transformed connectomes in downstream analyses, and show how a connectome-based predictive model can be generalized to publicly available processed data that was processed with different atlases. Overall, CAROT can reconstruct connectomes from an extensive set of atlases—without ever needing the raw data—allowing already processed connectomes to be easily reused in a wide-range of analyses while eliminating wasted and duplicate processing efforts. We share this tool as both source code and as a stand-alone web application (`http://carotproject.com/`).

2 Results

2.1 Optimal transport

The optimal transport problem solves how to transport resources from one location α to another β while minimizing the cost C[9–12]. It has been used for contrast equalization[13], image matching[14], image watermarking[15], text classification[16], and music transportation[17]. Optimal transport is one of the few methods that provides a well-defined distance metric when the support of the distributions is different. Other mappings approaches such as Kullback–Leibler divergence do not make this guarantee.

The original formulation of the optimal transport problem is known as the Monge problem. Assuming we have some resources $x_1,..,x_n$ in location α and some other resources $y_1,..,y_m$ in location β, we specify weight vectors a and b over these resources and define matrix C as a measure of pairwise distances between points $x_i \in \alpha$ and comparable points $\mathcal{T}(x_i)$. The Monge problem aims to solve the following optimizing problem[18]:

$$\min_{\mathcal{T}} \Big\{ \sum_i C(x_i, \mathcal{T}(x_i)) : \mathcal{T}_\sharp \alpha = \beta \Big\}, \tag{1}$$

where the push forward operator $\sharp$ indicates that mass from α moves towards β assuming that weights absorbed in $b_j = \sum_{\mathcal{T}(x_i)=y_j} a_i$. An assignment problem when the number of elements in the measures are not equal is a special case of this problem, where each point in α can be assigned to several points in β.

As a generalization of the Monge problem, the Kantorvich relaxation solves the mass transportation problem using a probabilistic approach in which the amount of mass located at x_i potentially dispatches to several points in the target[8]. An admissible solution for Kantorvich relaxation is defined as $\mathcal{T} \in \mathbb{R}_+^{n\times m}$ indicating the amount of mass being transferred from location x_i to y_j by $\mathcal{T}_{i,j}$:

$$U(a,b) = \{\mathcal{T} \in \mathbb{R}_+^{n\times m} : \mathcal{T}\mathbb{1}_m = a, \mathcal{T}^T \mathbb{1}_n = b\}, \tag{2}$$

, where $\mathbb{1}$ represents a vector of all 1's. An optimum solution is obtained by solving the following problem for a given $C \in \mathbb{R}^{n\times m}$[19]:

$$L_c(a,b) = \min_{\mathcal{T}\in U(a,b)} < C, \mathcal{T} >= \sum_{i,j} C_{i,j}\mathcal{T}_{i,j}. \tag{3}$$

While a unique solution is not guaranteed[20], an optimal solution exists (see proof in[21,22]). Kantorovich and Monge problems are equivalent under certain conditions (see proof in[23]).

2.2 Cross Atlas Remapping via Optimal Transport (CAROT)

CAROT operates by, first, transforming timeseries data from one atlas (labeled the source atlas) into timeseries from an unavailable atlas (labeled the target atlas). This transformation is spatial mapping between the two atlases. Next, after this

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

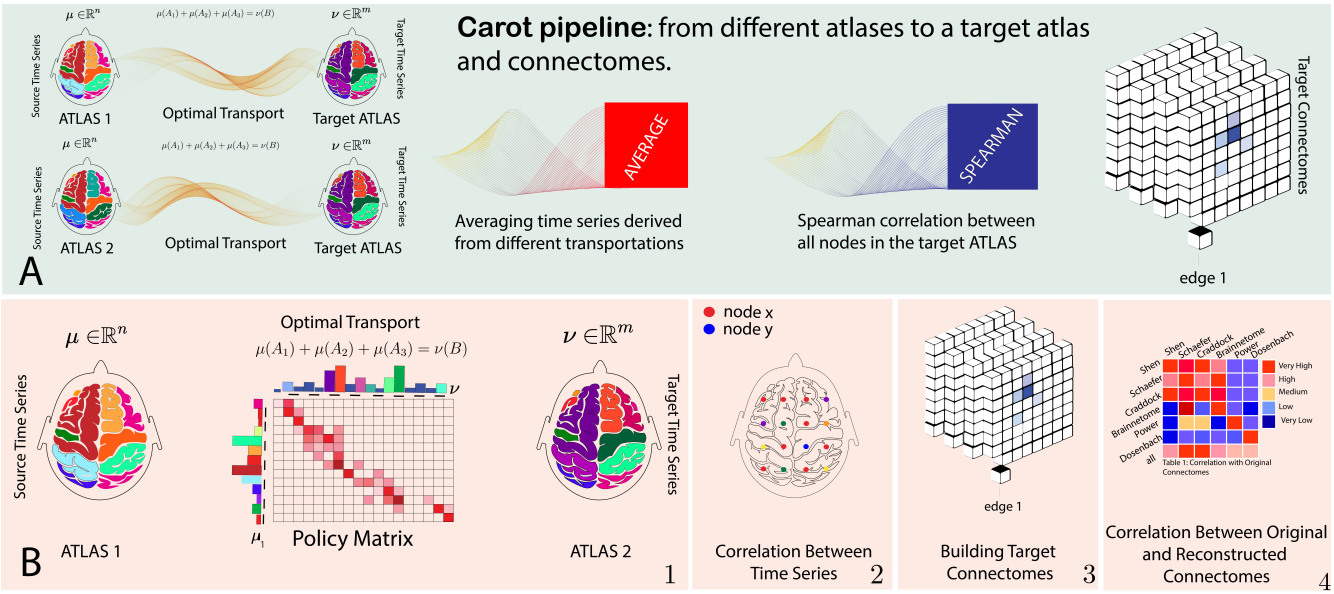


Figure 1. Schematic of CAROT: A) CAROT transforms timeseries fMRI data from multiple source atlases to create a connectome based on a missing target atlas. Mappings between the source and target atlases are found by employing optimal transport. B) For each pair of source and target atlases, we solve the Monge–Kantorovich transportation problem using the Sinkhorn approximation. For a single time point in the timeseries data, the solution provides a transformation that maps the brain activity parcellated using the source atlas to brain activity parcellated based on the target atlas. Results from several pairs of source and target atlases can be combined to improve the quality of final reconstructed connectome.

transformation, the corresponding functional connectomes can be estimated using standard approaches (e.g., full or partial correlation). Transforming the timeseries data rather then connectomes themselves has two benefits. First, this results in a lower dimensional mapping, which are more robust to estimate. Second, connectomes can be constructed wih standard methods (like correlation), guarantee properties like semi-positive definite. A direct mapping between connectomes may not guarantee this property.

Formally, lets assume we have training timeseries data consisting of T timepoints from the same individuals but from two different atlases (atlas $\mathscr{P}_n$ with n regions and atlas $\mathscr{P}_m$ with m regions). Additionally, let $\mu_t \in \mathbb{R}^n$ and $\nu_t \in \mathbb{R}^m$ to be the vectorized brain activity at single timepoint t based on atlases $\mathscr{P}_n$ and $\mathscr{P}_m$, respectively. For a fixed cost matrix $C \in \mathbb{R}^{n\times m}$, which measures the pairwise distance between regions in $\mathscr{P}_m$ and $\mathscr{P}_n$, we aim to find a mapping $\mathcal{T} \in \mathbb{R}^{n\times m}$ that minimizes transportation cost between μ_t and ν_t:

$$L_c(\mu_t, \nu_t) = \min_{\mathcal{T}} C^T \mathcal{T} \textbf{ s.t, } A\underline{\mathcal{T}} = \begin{bmatrix} \mu_t \\ \nu_t \end{bmatrix}, \tag{4}$$

in which $\underline{\mathcal{T}} \in \mathbb{R}^{nm}$ is vectorized version of $\mathcal{T}$ such that the $i+n(j-1)$'s element of $\mathcal{T}$ is equal to $\mathcal{T}_{ij}$ and A is defined as:

$$A = \begin{array}{c} \\ m \\ n \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & n \end{array} \\ \begin{pmatrix} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} & \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} & \dots \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \\ (1 \;\; 1 \;\; \dots \;\; 1) & (1 \;\; 1 \;\; \dots \;\; 1) & \dots (1 \;\; 1 \;\; \dots \;\; 1) \\ \vdots & & \\ (1 \;\; 1 \;\; \dots \;\; 1) & \dots & \dots (1 \;\; 1 \;\; \dots \;\; 1) \end{pmatrix} \end{array}. \tag{5}$$

$\mathcal{T}$ represents the optimal way of transforming the brain activity data from n regions into m regions. Thus, by applying $\mathcal{T}$ to every timepoint from the timeseries data of the source atlas, we can estimate the timeseries data of the target atlas. As solving this large linear program is computationally hard[24], we use the entropy regularization, which gives an approximation solution

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Shen	Schaefer	Craddock	Brainnetome	Power	Dosenbach
Shen		0.46 $\pm$0.01	0.550$\pm$0.001	0.42 $\pm$0.003	0.360$\pm$0.03	0.39 $\pm$0.004
Schaefer	0.31 $\pm$0.001		0.380 $\pm$0.01	0.38$\pm$0.001	0.33 $\pm$0.01	0.34 $\pm$0.002
Craddock	0.48 $\pm$0.01	0.540 $\pm$0.01		0.51 $\pm$0.003	0.43 $\pm$0.002	0.43$\pm$0.001
Brainnetome	0.17$\pm$0.003	0.230$\pm$0.003	0.23 $\pm$0.002		0.19 $\pm$0.004	0.18 $\pm$0.002
Power	0.24$\pm$0.002	0.350 $\pm$0.003	0.32$\pm$0.001	0.29$\pm$0.07		0.32 $\pm$0.002
Dosenbach	0.24 $\pm$0.001	0.320$\pm$0.003	0.28 $\pm$0.02	0.28 $\pm$0.0001	0.28 $\pm$0.01	

Table 1. Spearman correlation between reconstructed connectomes and original connectomes for each source-target pairs. Presented results show as mean $\pm$ standard deviation over 100 random splitting the data into training and testing sets.

with complexity of $\mathscr{O}(n^2\log(n)\eta^{-3})$ for $\varepsilon = \frac{4\log(n)}{\eta}$[20], and instead solve the following:

$$L_c(\mu_t, \nu_t) = \min_{\mathcal{T}} C^T\mathcal{T} - \varepsilon H(\mathcal{T}) \textbf{ s.t, } A\underline{\mathcal{T}} = \begin{bmatrix}\mu_t \\ \nu_t\end{bmatrix}. \tag{6}$$

Specifically, we use the Sinkhorn algorithm—an iterative solution for Equation 6[25]—to find $\mathcal{T}$. For training data with S participants and K time points per participant, first, we estimate the optimal mapping $\mathcal{T}_{s,k}$, independently, for time point k for a given participant $s$using Equation 6. Next, we average $\mathcal{T}_{s,k}$ overall time points and participants to produce a single optimal mapping $\mathcal{T}$ in the training data (*e.g.*, $\mathcal{T} = \frac{1}{|S||K|}\sum_{s=1}^{|S|}\sum_{k=1}^{|K|}\mathcal{T}_{s,k}$).

For the cost matrix C, we used a distance metric (labeled functional distance) that is based on the similarity of pairs of timeseries from the different atlases:

$$C = 1 - \begin{pmatrix} \rho(U_{1,.}, N_{1,.}) & \dots & \rho(U_{1,.}, N_{n,.}) \\ \vdots & \ddots & \vdots \\ \rho(U_{m,.}, N_{1,.}) & \dots & \rho(U_{m,.}, N_{n,.}) \end{pmatrix} \in \mathbb{R}^{m\times n} \tag{7}$$

where U_x and N_x are timeseries from $\mathscr{P}_m$ and $\mathscr{P}_n$ and $\rho(U_x, N_y)$ is Spearman correlation between them. To increase a reliable estimation of C, we calculate the timeseries correlation independently for each individual in the training data and average over these correlations. Functional distance was used over Euclidean distance between nodes for two mains reasons: *(i)* functional distance does not require having access to the atlas or node locations, which provides greater flexibility should a unknown and unavailable atlas be used, and *(ii)* spatial proximity in the brain does not guarantee similar function. For example, the medial prefrontal nodes of the default mode network are more correlated with nodes in the posterior cingulate cortex than other nodes in the frontal lobe. Finally, comparisons between functional and Euclidean distances suggest that functional distances results in higher similarity between reconstructed and original connectomes (Fig. S1).

2.3 Reconstructed connectomes are similar to original connectomes

To evaluate CAROT, we compared reconstructed connectomes to their original counter part using data from 515 individuals from the Human Connectome Project (HCP)[26]. Data were processed as previously described[27]. Timeseries data and connectomes were created from 6 different atlases: Shen (268 nodes)[28], Schaefer (400 nodes)[29], Craddock (200 nodes)[30], Brainnetome (246 nodes)[31], Power (264 nodes)[32], and Dosenbach (160 nodes)[33] atlases. For each scan, the average time series within each region was obtained. To calculate connectomes, the Pearson's correlation between the mean time series of each pair of regions was calculated and converted to be approximately normally distributed using a Fisher transformation. We partitioned our data into 25/75 split, where 25% of the individuals are to estimate the optimal mapping $\mathcal{T}$ and 75% is used to evaluate the reconstructed connectomes.

As shown in Table 1, the correlation between the reconstructed connectomes and their original counterparts depends on the atlas pairing, with pairs of more similar atlases appearing to having higher correlations. For instance, a strong correlation is seen while transforming data from the Craddock atlas to the Shen atlas ($\rho = 0.48$, $p < 0.001$). Both of these atlases are based on clustering timeseries fMRI data using variants of the N-cut algorithm. In contrast, a weaker correlation exist for transforming data from the Dosenbach atlas (which was constructed based meta-analysis of task activations) to the Shen atlas ($\rho = 0.24$, $p < 0.001$).

Next, we investigated the sensitivity of CAROT to the number of time points and number of participants used to find the mappings, and the value of ε's in the Sinkhorn approximation. No clear pattern of performance change was observed across either parameter, suggesting that CAROT is not affected the number of frames and participants, and the range of ε's (Fig. S2).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

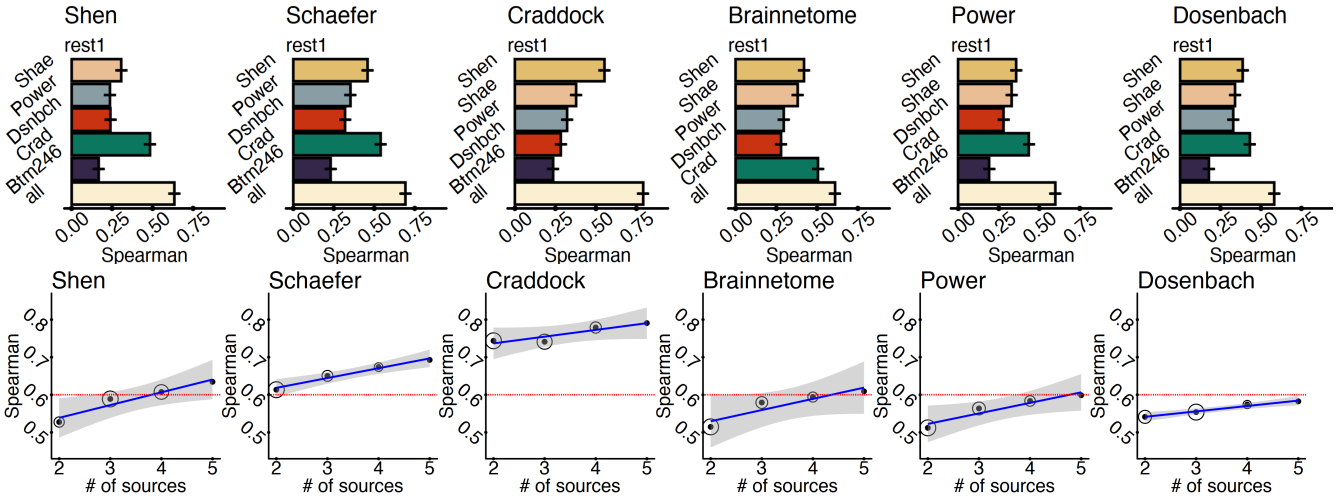


Figure 2. Using multiple source atlases improves the similarity of reconstructed connectomes. A) The Spearman's rank correlation between of the reconstructed connectomes and connectomes generated directly with the target atlases are shown for each pair of source and target atlas as well reconstructed connectomes using all of the source atlases. For each of the target atlases, using all source atlases produces higher quality reconstructed connectomes. Error bars are generated from 100 iterations of randomly splitting the data into 25% for training and 75% for testing. B) For each target atlas, increasing the number source atlases increases the similarly of reconstructed and original connectomes. For most atlases, a Spearman's correlation of $\rho > 0.60$ (red line) can be achieved with using fewer than five source atlases (i.e., all available source atlases). Circle size represents the variability of the correlation over 100 iterations of splitting the data into training and testing sets.

As preprocessed data is often released with timeseries data from multiple atlases[34], we investigated using these additional data to better reconstruct connectomes from an unavailable atlas. To achieve this, we applied CAROT to each pair of atlases, transforming timeseries data for each source atlas into the timeseries data for the target atlas. Next, the transformed timeseries data are averaged across all source atlases and a single connectome for the target atlas is created (Fig. 1). Overall, we observed a large improvement when including data from multiple source atlas and, in every case, using all available data produce more similar connectomes to their original counterparts (all $\rho's > 0.50$; Fig. 2). In fact, for most atlases, explained variance is more than tripled using CAROT with multiple source atlases compare to using a single source atlas. Further, we investigated the impact of using a smaller number of source atlases by only including k random source atlases when creating connectome for the target atlas. This process was repeated with 100 iterations over a range of k =2–6. As shown in Fig. 2, while similarity between reconstructed and original connectomes increases as the number of source atlases increases, strong correlations (e.g., $\rho > 0.6$) can be observed with as little as two or three source atlases, suggesting that a small number of atlases may be sufficient for most applications.

2.4 Reconstructed connectomes behave like original connectomes in downstream analyses

To further evaluated the reconstructed connectomes created by CAROT, we show that using the reconstructed connectomes in standard connectome-based analyses provide similar results as using their original counterparts. First, we show that meaningful brain-phenotype associations are retained in reconstructed connectomes by predicting fluid intelligence using connectome-based predictive modeling (CPM). We partitioned the HCP dataset into three groupings: g_1, which consisted of 25% of the participants; g_2, which consisted of 50% of the participants; and, g_3, which consisted of the final 25% of the participants. In g_1, we used CAROT to estimate $\mathcal{T}$ for each pair of source and target atlases. We then applied $\mathcal{T}$ on g_2 and g_3 to estimate functional connectomes for each target atlas, resulting in 9 connectomes for each atlas (7 reconstructed connectomes based on a single source atlas, 1 reconstructed connectome based on all source atlases, and the original connectome). Finally, for each set of connectomes, we trained a CPM model of fluid intelligence using g_2 and tested this model in g_3. Spearman correlation between observed and predicted values was used to evaluate prediction performance. This procedure was repeated with 100 random splitting of the data into the three groups. In all cases, connectomes reconstructed using all source atlases performed as well in prediction as the original connectomes (Fig 3A). In fact, the reconstructed connectomes using all source atlases performed better than the original connectomes for the Schaefer and Power atlases. Similar to other analyses, connectomes reconstructed from a single atlas varied in prediction performance, depending on the combination of source and target atlases.

Second, we show that individual uniqueness of connectomes are retained in reconstructed connectomes by identifying

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

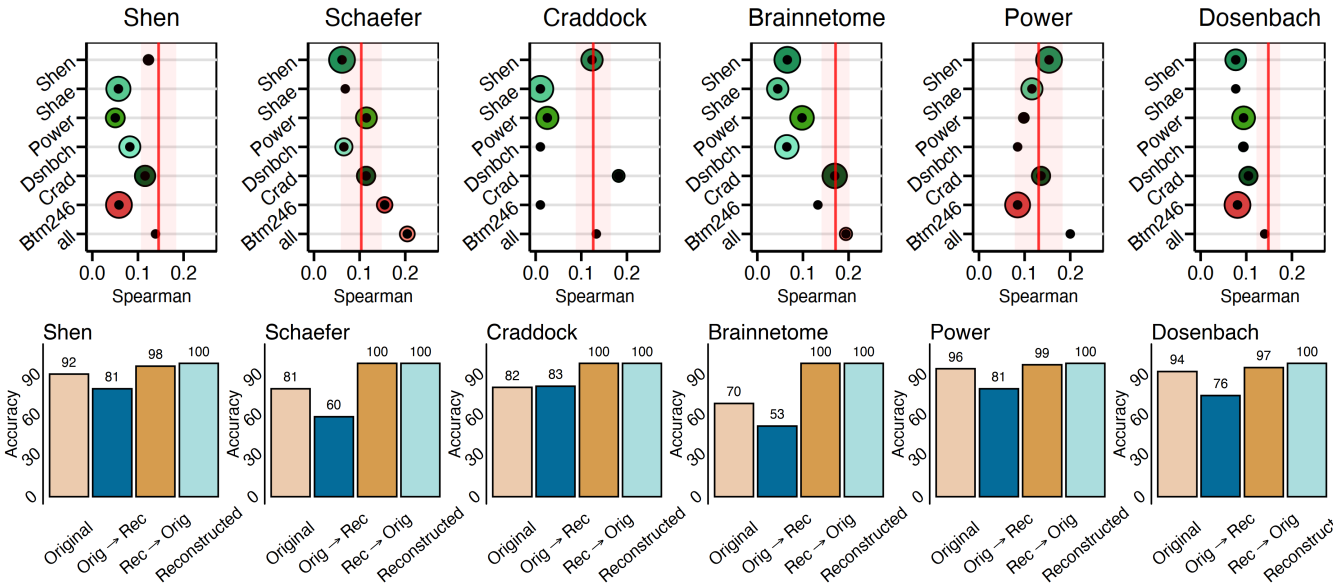


Figure 3. Reconstructed connectomes behave the same as original connectomes in downstream analyses. A) The reconstructed connectomes retain sufficient individual differences to predict IQ using connectome-based predictive modeling. In all cases, reconstructed connectomes based on all available source atlases (bottom circle) predicted IQ as well or better than the original connnectome (red line). Size of circle represents the variance of prediction performance of 100 iteration of 10-fold cross-validation. B) The reconstructed connectomes retain sufficient individual uniqueness to identify individuals using the reconstructed connectomes.

individuals scanned on repeated days[35]. As above, we used the HCP data and a 25%/75% split for creating reconstructed connectomes based on all available source atlases. We performed the identification procedure proposed in[35], where connectomes generated from the day 1 resting-state data was used as the target set and connectomes generated from the day 2 resting-state data was used as the database. Briefly, in an iterative process, one individual's connectome was selected from the target set and compared against each of the connectivity matrices in the database to find the matrix that was maximally similar. Similarity was defined as the Spearman correlation between target connectome and each of the connectomes in the database. A score of 1 was assigned if the predicted identity matched the true identity, or 0, if it did not. Each target connectome was tested against the database in an independent trial. We performed this identification procedure for the original and reconstructed connectomes, independently. For both types of connectomes and all atlases, identification of individuals demonstrated a high rate of success (Fig. 3B). Although reconstructed connectomes performed slightly worse than the originals (original connectomes: mean rate=90%; reconstructed connectomes: mean rate: 79%), these rates are much greater than chance (5%; $p < 0.001$; based on permutation testing). Overall, these results suggest that the reconstructed connectomes retain similar levels of individual differences as their original connectome counterparts and can be used in downstream analyses.

2.5 Mappings with with task-based connectomes are comparable to resting-state connectomes

Next, we investigated the performance of CAROT on task-based connectomes using the 7 tasks (gambling, language, motor, relational, social, working memory (WM), and emotion) collected in the HCP dataset. As above, we used a 25%/75% split of 515 participants in the HCP to train mappings between atlas and evaluate reconstructed connectomes for each task. Overall, CAROT displayed similar performance when using task-based connectomes compared to connectome created from resting-state data (Fig. S3). Given these similar performance, we investigated whether tasks and resting-state data produced the same mapping between atlases (see Fig. S4 for visualizations of these mappings). $\mathcal{T}$'s across different tasks and rest were highly correlated ($\rho's > 0.8$), implying the existence of a summary transportation plan across any condition. Accordingly, we averaged the $\mathcal{T}$ across the 8 conditions (7 tasks and 1 resting-state) for each source-target pair, resulting in a single canonical mapping between each source and target atlas.

2.6 CAROT facilitates external validation of connectome-based predictive models

Finally, we present a real-world example of how CAROT can be used to generalize a preexisting connectome-based predictive model when data from the required atlas is not available. In this example, we generalize a sex classification model (using 100 health adults collected at the Yale School of Medicine and created with the Shen atlas) to the REST-Meta-MDD dataset[34],

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

which only provides preprocessed timeseries data from the Dosenbach, Power, and Craddock atlases. We used the canonical $\mathcal{T}$ (defined above) from HCP dataset, allowing us to show that not only can CAROT facilitate the external validation of a predictive model to other datasets, but also that the mappings estimated by CAROT generalize across datasets. In other words, a canonical set of $\mathcal{T}$ can be estimated once with CAROT (for example, using the HCP) and then be applied to any new datasets without the need to rerun CAROT (for example, the Yale and REST-Meta-MDD datasets).

First, we trained the sex classification model using resting-state data from 100 individuals (50 males) from the Yale dataset. Data were processed and connectomes based on the Shen atlas were created as previously described[36]. We trained ℓ_2-penalized logistic regression model with 10-fold cross-validation to classify self-reported sex. Overall, this model demonstrated significant classification accuracy (Accuracy=60.5% $\pm$ 6%; Naive model accuracy=50%; $\chi^2 = 5.8$; $p = 0.03$). Then, we used canonical mappings estimated from the HCP to transform the publicly available preprocessed data (i.e., timeseires data from the Dosenbach, Power, and Craddock atlases) from the REST-Meta-MDD dataset into the Shen atlas. Data from each source atlas was combined to create a single connectome based on the Shen atlas for the 1005 (585 females) health controls. Finally, the sex classification model created in the Yale dataset was applied to these reconstructed Shen connectomes.

Overall, the sex classification model preformed significantly better than chance in the REST-Meta-MDD dataset when using the reconstructed connectomes (Accuracy=66.5%; Naive model accuracy=52.3%; $\chi^2 = 13.9$; $p = 0.0002$). To better this result into context, we created connectomes for the Dosenbach, Power, and Craddock atlases in the Yale dataset, created a sex classification model for connectome type, and generalized these models to the REST-Meta-MDD dataset. The generalization accuracy of reconstructed connectomes (Shen: 66.5%) was numerically superior to the generalization accuracies based on original connectomes (Dosenbach: 59.6%, Power: 59.0%, and Craddock: 64.5%), suggesting that using CAROT and reconstructed connectomes perform as well as original connectomes in generalizing a preexisting predictive model.

2.7 Software availability and implementation

To facilitate open science and the wider adoption of CAROT, we have created `http://carotproject.com/`. This web application allows end-users to convert timeseries data from the Shen, Schaefer, Craddock, Brainnetome, Power, and Dosenbach atlases to connectomes for any of the other atlases. As a web application, it works without any software installation and across multiple platforms (*e.g.*, Window, Linux, MacOS, Android). The only requirement is a modern web browser, such as Google Chrome. Please note, that any data used on `http://carotproject.com/` remains on the local computer and is never uploaded or stored on a remoted server. In addition, we provide the CAROT software and associated canonical mapping as opensource at `https://github.com/dadashkarimi/carot/`. CAROT is implemented in Python 3, building on top of the Python Optimal Transport (POT) toolbox[37]. Specifically, we provide functionality: *(i)* to generate the cost matrix based on functional distance for timeseries data from two different atlases; *(ii)* to generate the mapping $\mathcal{T}$ between between two atlases based on the cost matrix defined above; and *(iii)* to convert timeseries data from one or more source atlases to connectomes based on a target atlas. In addition, we provide canonical mappings based on the HCP data to map between every pair of the Shen, Schaefer, Craddock, Brainnetome, Power, and Dosenbach atlases. Based on the results present here, these mappings should work in other datasets, saving researchers the need to regenerate these mappings for themselves. We will look to provide mappings between additional atlases as they become available.

3 Discussion and conclusions

Neuroimaging has transitioned to a big-data era, with data from nearly 100,000 individuals available opensource. This transition is critical as the field is becoming increasingly aware of the need for larger samples and external validation datasets[38–40]. Yet, as it is typically not possible for a single lab to collect these, pooling open source data becomes a practical means to increase power and demonstrate the reproducibility of brain-behavior associations[4]. Here, we introduced and validated Cross Atlas Remapping via Optimal Transport (CAROT) to address one limitation of pooling previous processed data: connectomic data processed with different atlases. CAROT allows functional connectomes from different atlases to transformed to a common atlas and combined in downstream analyses. CAROT relies on optimal transport to find a frame-to-frame mapping of fMRI timeseries data used to create functional connectomes for a missing atlas. We show that these reconstructed connectomes are highly like the original ones and perform interchangeablely in downstream analyses. Specially, reconstructed connectomes retain sufficient individual differences to predict IQ and uniqueness to identify individuals. Finally, we provide a real-world example of how a connectome-based predictive model (based on the Shen atlas) can be generalized to opensource, preprocessed data that was processed not with the Shen atlas. We hope that CAROT and `http://carotproject.com/` will save researcher time and effort by not needed to reprocess data and increase the ease of performing mega-analysis and external validation efforts.

Across analyses, we show that CAROT produces reconstructed connectomes that produce similar results in downstream analyses as connectomes created directly from raw data. This observation holds true across a range of atlases that differ in their construction and constituent brain regions (see Methods for a brief of these for each atlas). While atlases pairs that are more similar in terms of their construction and coverage produced better pair-wise mappings (e.g., the Craddock and Shen

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

atlases were created with N-cut algorithms and cover the cortex, sub-cortex, and cerebellum[28,30]), using multiple source atlases minimizes the effect. Likely, combining transformed timeseries averages out the small idiosyncrasies in the individual mappings between atlas pairs, producing more stable results. Overall, when using multiple source atlases, CAROT is robust to differences between the source and target atlases.

While including all available data generated the most similar connectomes, strong correspondence between reconstructed and original connectomes were observed when as little as 2 or 3 different source atlases were used. Together, this suggests that an exhaustive list of every possible atlas does not need to be released, but that only including a few extra atlases could vastly increase the utility of any released preprocessed data. Balancing the utility of released data and the effort it takes to release the data is a delicate task. If the data is not in a convenient form for end-users, it will not be used and, if the effort is too high to share data, data will not be shared. We believe that CAROT can help in balancing these, by increasing the utility of the shared data with only a small increase in effort for sharing the data.

Given that multiple source atlases produce more robust results, it encourages future studies to release preprocessed data from a few atlases. Not only does this increase the chances that the needed atlas is available for an end-user, but also better facilities the data be used when the needed atlas is not available. Some opensource dataset are already release data from multiple atlases (e.g., REST-Meta-MDD). This may also be relevant for large-scale projects, like ABCD and UK Biobank, that share both raw data as well as curated releases. Given that these datasets range in the several thousands of participants, curated data from multiple atlases further facilities smaller labs and research groups being able to use these data with a wider-range of atlases and tools.

Additionally, CAROT may help with connectome-based meta analyses, for which there are few, by allowing results to be pooled across studies. Coordinate based meta analyses are popular for task activation and brain morphometry studies[41–43] and are possible as most neuroimaging studies rely on a common template (*i.e.*, the MNI template). This common template allows for spatial comparisons and pooling of results across different studies. Yet, a similar approach is difficult for connectome-based results as there is a lack spatial correspondence between different atlases. In an analogous manner to spatial transformation between MNI and Talairach space, being able to transform and compare connectomes between different atlases would allow a greater number of studies to be pooled together in such meta-analyses.

There are a few notable strengths and limitations of CAROT. First, CAROT appears to be robust to the choices of algorithmic choices such as the number of frames and individuals used for training, the choice of cost matrix, and the equation used solve to the optimal transport problem. We reason that the robustness to parameter choices is in part due to the large amount of spatial and temporal autocorrelation in fMRI data[44], which allow something complex as a connectome to be compactly parameterized. Future work includes generalizing CAROT to other functional timeseries data—such as electroencephalography (EEG), functional near infrared spectroscopy (fNIRS), or even wide-field CA2+ imaging data in mice[45]—where spatial and temporal autocorrelation patterns will be different. Similar, as CAROT is based timeseries data, it is not appropriate for structural connectomes. Yet, the "missing atlas" problem exist with these data, for which no solution exists. CAROT, based on Euclidean distance rather than functional distance, may be a reasonable approach to map between atlases used to create structural connectome as well as map between different atlas used in morphometric analyses (such as the Desikan-Killiany and Destrieux atlases used in FreeSurfer). While we tested CAROT with a large range of atlases, we could not test CAROT in every functional atlas, for which there are many. Nevertheless, given the range in atlas size (200-500 nodes) and atlas coverage (whole-brain and cortical only), we expect CAROT to work well for modern atlases not tested here and look to update CAROT when a new generation of brain atlases emerges.

In sum, CAROT allows data processed from one atlas to be directly transformed into a connectome based on another atlas without needing raw data. These reconstructed connectomes are similar to and, in downstream analyses, behave like to the original connectomes, created direct from the raw data. Using CAROT on preprocessed opensource data will increase its utility, accelerate the use of big data, and help make generalization and replication efforts easier to perform.

4 Methods

4.1 Datasets:

Three datasets were used in the primary analyses described here: the Human Connectome Project (HCP), the REST-Meta-MDD Consortium and the Yale Low-Resolution Controls Dataset.

4.1.1 HCP participants

From this data set, we used behavioral and functional imaging data as previously described. We restricted our analyses to those subjects who participated in all nine fMRI conditions (seven task, two rest), whose mean frame-to-frame displacement was less than 0.1mm and whose maximum frame-to-frame displacement was less than 0.15mm, and for whom IQ measures were available (n=515; 241 males; ages 22–36+). The HCP minimal preprocessing pipeline was used on these data, which includes artifact removal, motion correction, and registration to standard space[46]. All subsequent preprocessing was performed

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in BioImage Suite61 and included standard preprocessing procedures, including removal of motion-related components of the signal; regression of mean time courses in white matter, cerebrospinal fluid, and gray matter; removal of the linear trend; and low-pass filtering. Task connectivity was calculated based on the "raw" task timecourses, with no regression of task-evoked activity. After processing, the Shen, Schaefer, Craddock, Brainnetome, Power, and Dosenbach atlases were applied to the preprocessed to create mean timeseries for each node. Connectomes were generated by calculating the Pearson's correlation between each pair of these mean timeseries and then tasking the fisher transform of these correlations.

4.1.2 REST-meta-MDD

Fully processed data was downloaded from http://rfmri.org/REST-meta-MDD. Full details about the dataset have been previously published elsewhere[47]. We used data from 21 of the 24 sites. Two sites were removed due to large imbalance between male and female participants (i.e., $< 30\%$ male or female; sites 2 and 12). One site was removed as self-reported sex was not provide (site 4). Briefly, the data was processed as follows. First, the initial 10 volumes were discarded, and slice-timing correction was performed. Then, the time series of images for each subject were realigned using a six-parameter linear transformation. After realignment, individual T1-weighted images were co-registered to the mean functional image using a 6 degrees-of-freedom linear transformation without re-sampling and then segmented into gray matter, white matter, and cerebrospinal fluid. Finally, transformations from individual native space to MNI space were computed with the Diffeomorphic Anatomical Registration Through Exponentiated Lie algebra (DARTEL) tool. To minimize head motion confounds, the Friston 24-parameter model was to regressed from the data. Scrubbing (removing time points with FD>0.2mm) was also utilized to verify results using an aggressive head motion control strategy. Other sources of spurious variance (global, white matter, and CSF signals) were also removed from the data through linear regression. Additionally, linear trend were included as a regressor to account for drifts in the blood oxygen level dependent (BOLD) signal. Temporal bandpass filtering (0.01-0.1Hz) was performed on all time series. Timeseries data from the Craddock, Dosenbach, and Power were used to generate connectomes using Pearson's correlation. Additionally, these data were transformed into timeseries data from the Shen atlas via CAROT and associated connectoems were generated.

4.1.3 Yale participants

In addition, we used resting-state data collected from 100 participants at the Yale School of Medicine to train a sex classification model with four different atlases Shen, Craddock, Dosenbach, and Power. This dataset included 50 females (age=33.3±12.3) and 50 males (age=34.9±10.1) with eight functional scans (48 minutes total). For details of the dataset as well as details about processing can be found in[36]. Briefly, standard preprocessing procedures were applied to these data. Structural scans were skull stripped using an optimized version of the FMRIB's Software Library (FSL) pipeline. Slice time and motion correction were performed in SPM8. The remainder of image preprocessing was performed in BioImage Suite and included linear and nonlinear registration to the MNI template; regression of mean time courses in white matter, cerebrospinal fluid, and gray matter; and low-pass filtering.

4.1.4 Atlases

We evaluated CAROT on 6 common functional atlases from the literature. The Shen atlas[28] was created using functional connectivity data from 45 adult participants. The 268-node atlas was constructed using a group-wise spectral clustering algorithm (derived from the N-cut algorithm) and covers the entire cortex, sub-cortex, and cerebellum. The Craddock atlas[30] was created using functional connectivity data from 41 adult participants. The 200-node atlas was constructed using an N-cut algorithm and covers the entire cortex, sub-cortex, and cerebellum. The Schaefer atlas[29] was created using functional connectivity data from 744 adult participants from the the Genomics Superstruct Project[48]. The 400-node atlas was constructed using a gradient-weighted Markov Random Field (gwMRF) models and covers only the cortex. The Brainnetome atlas[31] was created using structural connectivity data from 40 adult participants from the HCP. The 246-node atlas was constructed using a tractography-based approach and covers the cortex and sub-cortex. The Dosenbach atlas[33] was created from meta-analyses of task-related fMRI studies and covers the cortex, cerebellum, and a few sub-cortical nodes. The Power atlas[32] was created by combining the meta-analytical approach of the Dosenbach atlas with areal boundary detection based on functional connectivity data. The 264-node atlas covers the cortex, sub-cortex, and cerebellum.

4.2 CAROT implementations and evaluation

CAROT was implemented in Python using the Sinkhorn algorithm as implemented by Python Optimal Transport (POT) toolbox[37]. Connectomes reconstructed by CAROT were generated using Pearson's Correlation. To evaluate the similarity between reconstructed and original connectomes, the upper triangles of the connectomes were vectorized and correlated with Spearman's rank correlation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.3 Connectome-based predictive modeling (CPM)

CPM was conducted to predict each of the IQ in the HCP participants using previously validated custom MATLAB scripts[49]. CPM takes connectivity matrices and phenotypic data (e.g., IQ) from individuals as input to generate a predictive model of the behavioral data from connectivity matrices. Edges and phenotypic data from the training data set are correlated using regression analyses using Pearson's correlation to identify positive and negative predictive networks using an edge-selection threshold of$p < .05$. Positive networks are networks for which increased edge weights (increased connectivity) are associated with the variable of interest, and negative networks are those for which decreased edge weights (decreased connectivity) are associated with the variable of interest. Single-participant summary statistics are then created as the sum of the significant edge weights in each network and are entered into predictive models that assume linear relationships with phenotypic data. The resultant linear equation is then applied to the test data set to predict the phenotypic data. One hundred iterations of 10 fold cross-validation were used to train and evaluate models. The correspondence between predicted and actual values, or model performance, was assessed using Spearman's rank correlation (ρ) to avoid strong distribution assumptions.

4.4 Identification procedure

Briefly, a database is first created consisting of all participants' connectivity matrices from a particular session for a specific dataset (in this case: day 1 resting-state data from the HCP). In an iterative process, a connectivity matrix from a participant is then selected from a different session (in this case: day 2 resting-state data from the HCP) and denoted as the target. Pearson correlation coefficients are calculated between the target connectivity matrix and all the matrices in the database. If the highest Pearson correlation coefficient is between the target participant in one session and the same participant in the second session, this would be recorded as a correct identification. We repeat the identification test such that each participant serves as the target subject once. This process is repeated until identifications have been performed for all participants. To determine if identification rates were achieved at above-chance levels, we used permutation testing to generate a null distribution. Specifically, participants identities were randomly shuffled and identification was performed with these shuffled labels. Identification rates obtained using the correct labels were then compared to this null distribution to determine significance.

4.5 Sex classification

We trained a ℓ_2-penalized logistic regression model to classify self-reported sex in the Yale dataset. One hundred iterations of 10-fold cross-validation was used to train the SVM models. Folds were not balances for sex. "Out-of-the-box" scikit-learn models were used for logistic regression. The median performing model was used in external validation in the REST-meta-MDD.

4.6 Data availability

All datasets used in this study are open-source: HCP (ConnectomeDB database, https://db.humanconnectome.org), REST-meta-MDD (http://rfmri.org/REST-meta-MDD), and Yale dataset (http://fcon_1000.projects.nitrc.org/indi/retro/yale_lowres.html). BioImage Suite tools used for processing can be accessed at (https://bioimagesuiteweb.github.io/). CAROT and associated canonical mappings are on GitHub (https://github.com/dadashkarimi/carot). The Python Optimal Transport (POT) toolbox is available at https://pythonot.github.io/.

5 Acknowledgments

This work was supported by NIMH R01 MH121095, NSF (IIS-1845032), ONR (N00014-19-1-2406), and Tata. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; U54 MH091657) and funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Data was also provided by the REST-meta-MDD Consortium Data Sharing, which was supported by the National Key R&D Program of China (2017YFC1309902), the National Natural Science Foundation of China (81671774, 81630031, 81471740 and 81371488), the 13th Five-year Informatization Plan (XXH13505) of Chinese Academy of Sciences, Beijing Municipal Science & Technology Commission (Z161100000216152, Z171100000117016, Z161100002616023 and Z171100000117012), Department of Science and Technology, Zhejiang Province (2015C03037) and the National Basic Research (973) Program (2015CB351702). The remainder of the data used in this study were provided by the Philadelphia Neurodevelopmental Cohort (Principal Investigators: Hakon Hakonarson and Raquel Gur; phs000607.v1.p1). Support for the collection of the data sets was provided by grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon Hakonarson.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Casey, B. *et al.* The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Dev. cognitive neuroscience* **32**, 43–54 (2018).
2. Alexander, L. M. *et al.* An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. data* **4**, 1–26 (2017).
3. Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779 (2015).
4. Horien, C. *et al.* A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nat. human behaviour* **5**, 185–193 (2021).
5. Gorgolewski, K. J. *et al.* Neurovault.org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Front. Neuroinformatics* **9**, DOI: 10.3389/fninf.2015.00008 (2015).
6. Bolt, T., Nomi, J. S., Yeo, B. T. & Uddin, L. Q. Data-driven extraction of a nested model of human brain function. *The J. Neurosci.* **37**, 7263–7277, DOI: 10.1523/jneurosci.0323-17.2017 (2017).
7. Mensch, A., Mairal, J., Thirion, B. & Varoquaux, G. Extracting representations of cognition across neuroimaging studies improves brain decoding. *PLOS Comput. Biol.* **17**, e1008795, DOI: 10.1371/journal.pcbi.1008795 (2021).
8. Kantorovich, L. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, vol. 37, 227–229 (1942).
9. Tolstoi, A. Methods of finding the minimal total kilometrage in cargo transportation planning in space. *TransPress Natl. Commis. Transp.* **1**, 23–55 (1930).
10. Hitchcock, F. L. The distribution of a product from several sources to numerous localities. *J. mathematics physics* **20**, 224–230 (1941).
11. Koopmans, T. C. Optimum utilization of the transportation system. *Econom. J. Econom. Soc.* 136–146 (1949).
12. Gangbo, W. & McCann, R. J. The geometry of optimal transportation. *Acta Math.* **177**, 113–161 (1996).
13. Delon, J. Midway image equalization. *J. Math. Imaging Vis.* **21**, 119–134 (2004).
14. Li, P., Wang, Q. & Zhang, L. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *Proceedings of the IEEE International Conference on Computer Vision*, 1689–1696 (2013).
15. Mathon, B., Cayre, F., Bas, P. & Macq, B. Optimal transport for secure spread-spectrum watermarking of still images. *IEEE Transactions on Image Process.* **23**, 1694–1705 (2014).
16. Huang, G. *et al.* Supervised word mover's distance. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4869–4877 (2016).
17. Flamary, R., Févotte, C., Courty, N. & Emiya, V. Optimal spectral transportation with application to music transcription. *arXiv preprint arXiv:1609.09799* (2016).
18. Monge, G. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sci. de Paris* (1781).
19. Rubner, Y., Tomasi, C. & Guibas, L. J. The earth mover's distance as a metric for image retrieval. *Int. journal computer vision* **40**, 99–121 (2000).
20. Peyré, G., Cuturi, M. *et al.* Computational optimal transport: With applications to data science. *Foundations Trends Mach. Learn.* **11**, 355–607 (2019).
21. Birkhoff, G. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucuman, Ser. A* **5**, 147–154 (1946).
22. Bertsimas, D. & Tsitsiklis, J. Introduction to linear optimization, athena scientific, 1997. *URL: http://athenasc.com/linoptbook. html* .
23. Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Commun. on pure applied mathematics* **44**, 375–417 (1991).
24. Dantzig, G. B. Reminiscences about the origins of linear programming. In *Mathematical Programming The State of the Art*, 78–86 (Springer, 1983).
25. Altschuler, J., Weed, J. & Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *arXiv preprint arXiv:1705.09634* (2017).
26. Van Essen, D. C. *et al.* The wu-minn human connectome project: an overview. *Neuroimage* **80**, 62–79 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

27. Gao, S., Greene, A., Constable, T. & Scheinost, D. Combining multiple connectomes improves predictive modeling of phenotypic measures. *Neuroimage* **In Press** (2019).

28. Shen, X., Tokoglu, F., Papademetris, X. & Constable, R. T. Groupwise whole-brain parcellation from resting-state fmri data for network node identification. *Neuroimage* **82**, 403–415 (2013).

29. Schaefer, A. *et al.* Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cereb. cortex* **28**, 3095–3114 (2018).

30. Craddock, R. C., James, G. A., Holtzheimer III, P. E., Hu, X. P. & Mayberg, H. S. A whole brain fmri atlas generated via spatially constrained spectral clustering. *Hum. brain mapping* **33**, 1914–1928 (2012).

31. Fan, L. *et al.* The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cereb. cortex* **26**, 3508–3526 (2016).

32. Power, J. D. *et al.* Functional network organization of the human brain. *Neuron* **72**, 665–678 (2011).

33. Cohen, A. L. *et al.* Defining functional areas in individual human brains using resting functional connectivity mri. *Neuroimage* **41**, 45–57 (2008).

34. Yan, C.-G., Wang, X.-D., Zuo, X.-N. & Zang, Y.-F. Dpabi: data processing & analysis for (resting-state) brain imaging. *Neuroinformatics* **14**, 339–351 (2016).

35. Finn, E. S. *et al.* Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. neuroscience* **18**, 1664 (2015).

36. Scheinost, D. *et al.* Sex differences in normal age trajectories of functional brain networks. *Hum. Brain Mapp.* **36**, 1524–1535, DOI: 10.1002/hbm.22720 (2014).

37. Flamary, R. & Courty, N. Pot python optimal transport library (2017).

38. Yarkoni, T. Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—commentary on vul et al. (2009). *Perspectives on Psychol. Sci.* **4**, 294–298, DOI: 10.1111/j.1745-6924.2009.01127.x (2009).

39. Szucs, D. & Ioannidis, J. P. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage* **221**, 117164, DOI: 10.1016/j.neuroimage.2020.117164 (2020).

40. Marek, S. *et al.* Reproducible brain-wide association studies require thousands of individuals. *Nature* DOI: 10.1038/s41586-022-04492-9 (2022).

41. Laird, A. R., Lancaster, J. L. & Fox, P. T. BrainMap: The social evolution of a human brain mapping database. *Neuroinformatics* **3**, 065–078, DOI: 10.1385/ni:3:1:065 (2005).

42. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Essen, D. C. V. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670, DOI: 10.1038/nmeth.1635 (2011).

43. Eickhoff, S. B. *et al.* Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Hum. Brain Mapp.* **30**, 2907–2926, DOI: 10.1002/hbm.20718 (2009).

44. Shinn, M. *et al.* Spatial and temporal autocorrelation weave human brain networks. *bioRxiv* DOI: 10.1101/2021.06.01.446561 (2021). https://www.biorxiv.org/content/early/2021/06/01/2021.06.01.446561.full.pdf.

45. Lake, E. M. R. *et al.* Simultaneous cortex-wide fluorescence ca2 imaging and whole-brain fmri. *Nat. Methods* **17**, 1262–1271, DOI: 10.1038/s41592-020-00984-6 (2020).

46. Glasser, M. F. *et al.* The minimal preprocessing pipelines for the human connectome project. *NeuroImage* **80**, 105–124, DOI: https://doi.org/10.1016/j.neuroimage.2013.04.127 (2013). Mapping the Connectome.

47. Yan, C.-G. *et al.* Reduced default mode network functional connectivity in patients with recurrent major depressive disorder. *Proc. Natl. Acad. Sci.* **116**, 9078–9083, DOI: 10.1073/pnas.1900390116 (2019). https://www.pnas.org/content/116/18/9078.full.pdf.

48. Holmes, A. J. *et al.* Brain genomics superstruct project initial data release with structural, functional, and behavioral measures. *Sci. Data* **2**, DOI: 10.1038/sdata.2015.31 (2015).

49. Shen, X. *et al.* Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *nature protocols* **12**, 506 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplemental Materials

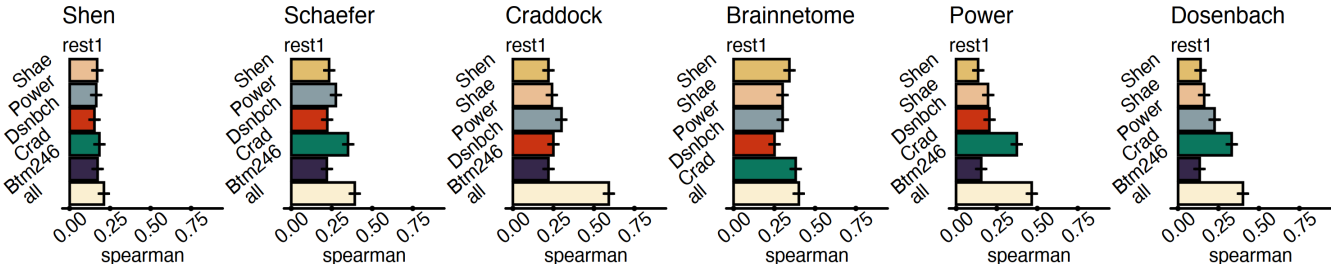
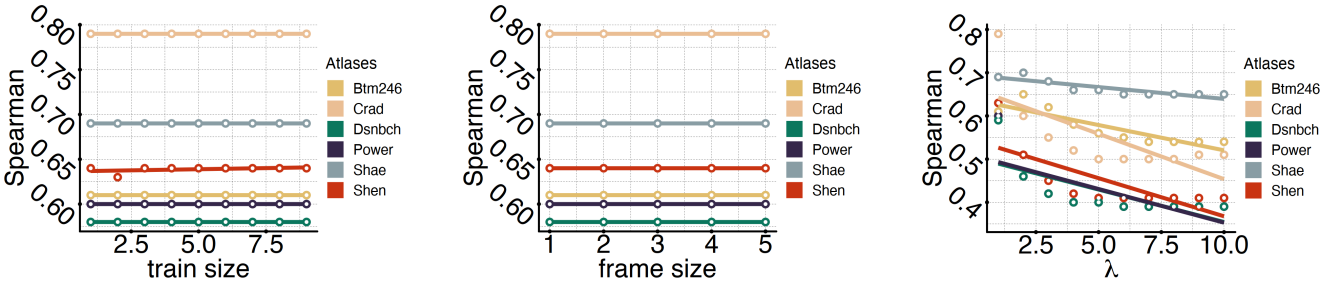


Figure S1. CAROT performance (rest) using euclidean distance as a measure of cost between ROIS.



(a) Frame Size Sensitivity **(b)** Train Size Sensitivity **(c)** Entropy Penalty

Figure S2. Parameter sensitivity of frame size , training data , and entropy regularization ε for different target atlases.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S3. CAROT performance on task connectomes. Across all 7 tasks in the HCP data, the correlation between the reconstructed connectomes and original connectomes are similar to the correlation between the reconstructed connectomes and original connectomes for resting-data data (see Fig. 2). Error bars are generated from 100 iterations of randomly splitting the data into 25% for training and 75% for testing.

bioRxiv preprint doi: https://doi.org/10.1101/2022.07.19.500642; this version posted July 20, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

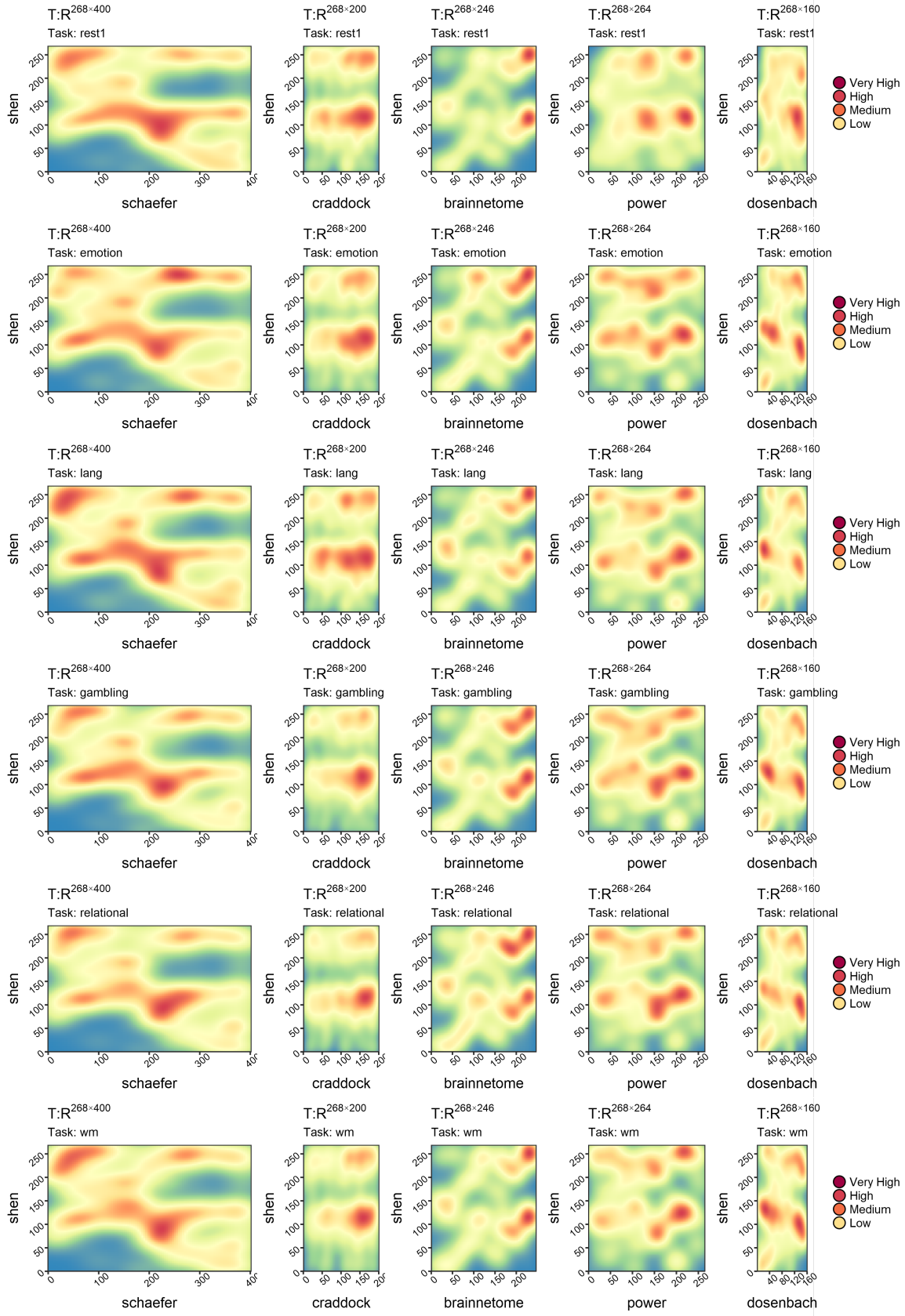


Figure S4. Optimal transport mappings derived from resting-resting and task data from the Shen atlas (source atlas) to each other atlas. Warmer color indicates regions that contribute the most towards mapping between atlases. Horizontal blue areas may indicate locations that are missing in the source atlas. For example, the Schaefer atlas does not include regions in the cerebellum, while the Shen does.